Machine Learning Applications to Survey Nonresponse

Inaugural Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Social Sciences in the Graduate School of Economic and Social Sciences at the University of Mannheim

By John 'Jack' Collins

November 2025

Full-time Dean of the Faculty of Social Sciences:
Dr. Julian Dierkes

Primary supervisor:
Prof. Dr. Christoph Kern

Secondary supervisor:
Prof. Dr. Florian Keusch

Primary reviewer:
Prof. Dr. Florian Keusch

Secondary reviewer:
Prof. Dr. Frauke Kreuter

Day of the disputation:
November 3rd 2025

Acknowledgments

In the making of this dissertation, much gratitude is owed to many.

Firstly, Jun.-Prof. Dr. Christoph Kern trusted me with his research project through which most of the papers that comprise this thesis were published. Thanks also to Christoph for his excellent mentorship. Almost any PhD-closer does so thanks to their supervisors, but I am especially lucky to have had the benefit of Christoph's thoughtful and ever-constructive advice as well as his co-authorship through which I could learn straight from the master.

Warm thanks also to Prof. Florian Keusch and Prof. Frauke Kreuter for many things. They have agreed to review this dissertation, and moreover provided a superbly stimulating research environment. Florian has always been open and available with excellent advice on the art of research to myself and all Sociologists in the Universität Mannheim school. Thankyou, Florian, for always having an answer to my endless queue of "quick questions." I'm indebted to Frauke for making me a part of her research group through which I've gotten to see some of the best research in this area as well as to see how top-tier academics go to work. Thanks to Frauke also for putting faith in me to be a part of the first 'Data Science for Social Good X' in Munich.

I'm very pleased with all of my papers and that is much in thanks to my teammates at GESIS. It is thanks to this team that I was able to work on the project "Prediction-based Adaptive Designs for Panel Surveys (PrADePS)," which yielded three of this dissertation's four papers. Also, I was able to see how experienced practitioners like Tobias Gummer and Bernd Weiß go about the business of survey research. Not least, it has been amazing to go through my PhD in tandem with teammate Saskia Bartholomäus, not just because of her comradeship, but also because her work taught me much about survey research.

All across the University of Mannheim and the Ludwig Maximilian University of Munich, brilliant people helped me with reviews, feedback, ideas, and even just encouragement and advice. It's been a privilege to have such august peers.

Finally, I would not have had this adventure of a lifetime if not for my wife, who is my best friend and chiefest of all supporters.

Thankyou, Dr. Rajika Kuruwita. මම ඔයාට ආදරෙයි.

Contents

1 Introduction

This dissertation explores how Machine Learning (ML) can help researchers avoid biased inferences due to low response rates in general population surveys. Low response rates can, though do not always, cause these survey studies to fail in one of their primary objectives: making accurate inferences about a population based on a sample. When survey respondents differ systematically from nonrespondents in regards to the topic of the study, the resulting sample will be biased toward the characteristics of those who participate. This effect is called 'nonresponse bias' (Groves, 2006).

Machine learning, a subfield of computer science, focuses on developing algorithms that predict outcomes based on historical data. Given this principle, ML algorithms are a natural choice to learn patterns in survey data and predict individual tendencies to participate, which, as I shall explain, can in turn be leveraged to address nonresponse bias in various ways. The contributions of this dissertation, while varied, follow a common approach: applying ML techniques in novel ways to the challenge of survey nonresponse and demonstrating how survey practitioners can benefit from adopting these innovative methods.

Specifically, this work provides survey practitioners with new methods for evaluating the role of past behavior in predicting future nonresponse behavior (Chapter 2), making earlier predictions in newly commenced panel surveys (Chapter 3), enhancing response rates with model-based incentive targeting (Chapter 4), and improving election predictions (which are often confounded by nonresponse bias) by augmenting poll-based models with ML (Chapter 5). Chapters 2-4 are about techniques to ameliorate nonresponse bias, Chapter 5 is about a technique for correcting inferences despite the presence of nonresponse bias.

This introductory chapter provides context for these four studies. This chapter begins with an introduction to survey research and an explanation of how nonresponse bias presents a critical contemporary challenge in the field. Next, I introduce the relevant principles of ML. Then, I combine these two threads to explain why ML has the potential to alleviate the problem of nonresponse and how it has already been applied to this field. Finally, I detail the four research papers' contributions to this broader research agenda.

1.1 Survey Research and Nonresponse Bias

The primary objective of general population surveys is to infer the characteristics of a given population based on a sample (Groves, 2011a). Surveys can target various kinds of groups but this dissertation focuses primarily on those designed to study a nationwide population (hereafter referred to as "surveys").

Modern survey methodologies emerged in the 1930s and 1940s (Groves, 2011b). In this time, researchers such as Neyman (1934) formalized processes that are now standard means of conducting general population surveys. For example, Neyman (1934) formalized stratified sampling. Also in this period, Likert (1932) introduced the Likert scale, which helped researchers to systematically measure individuals' attitudes toward social, political, and psychological topics. N-point scales like the Likert scale have since become standard in social surveys. The efficacy of structured surveys was further demonstrated in 1936 when the first Gallup poll correctly predicted Franklin D. Roosevelt's victory in the U.S. presidential election (Gallup and Rae, 1940). Developments like these led Groves (2011b) to argue that the 1930-40s marked the formalization of modern survey research. Key methodological principles were established during the period, including the use of structured questionnaires, standardized sampling methods such as fully random, or stratified random sampling, and the calculation of margins of errors around population-wide inferences.

From the 1930s to the 2020s, nations with established survey institutions have faced a common problem: declining response rates (Massey and Tourangeau, 2013). High nonresponse rates create the possibility for nonresponse bias. Nonresponse bias occurs when respondents and nonrespondents differ systematically concerning the variables of interest in the study. In such cases, inferences drawn from the sample may be biased toward the characteristics of the respondents (Groves, 2006). In the following passages, I examine how declining response rates have exacerbated nonresponse bias in contemporary surveys and how this dissertation contributes to mitigating these challenges.

Here I shall describe the severity of the response rate problem. Declining response rates are an issue globally and are unmitigated by the survey mode. In the United States and Europe, national surveys typically achieved response rates above 70% as recently as the 1970s. However, by the 2000s, response rates below 40% had become common (Massey and Tourangeau, 2013). Prominent survey institutions, such as the Pew Research Center, report a decline in response rates across all telephone surveys, from 36% in 1997 to just 9% by 2016 (Mitchell, 2017). Dutwin and Buskirk (2021) examined three other regular telephone surveys and found that only one, the National Health Interview Survey, maintained response rates above 80% into the 2010s, whereas others dropped below 20%. Similarly, Williams and Brick (2018) analyzed five U.S. national-level, face-to-face surveys and found that response rates fell from ranges of 75–95% in 1990 to 60–80% in 2014.

If response rates are declining in both telephone and face-to-face surveys, the trajectory may be even worse for web-based surveys. Daikeler, Bošnjak, and Lozar Manfreda (2020) studied four mixed-mode surveys conducted between the 1990s and 2010s and found that the web-based mode, on average, had 12% fewer respondents than the telephone mode. Jabkowski and Cichocki (2024) reviewed three mixed-mode, Europe-wide studies which fielded surveys regularly between 1999 and 2018. Their analysis concludes that response rates declined by an average of 4% per decade across all modes. Taken together, these studies provide substantial evidence that response rates are falling across survey modes and nations.

The most common explanations for this decline are technological and economic: rising

wages have made face-to-face interviews unscalable. This was particularly the case as better job opportunities arose for women, who made up the bulk of the face-to-face interviewing workforce of earlier decades (Groves, 2011b). Face-to-face surveys eventually gave way to relatively cheaper phone interviews. However, as mobile phones replaced landlines, it became much more difficult to match a participant's location to their number and thereby control geographic stratification (Massey and Tourangeau, 2013; Dutwin and Buskirk, 2021). As home internet proliferated in the 1990s and 2000s, people became less responsive to telephone and face-to-face surveys (Massey and Tourangeau, 2013). However, from the 1990s to 2020s online surveys had issues that other modes did not. Firstly, they excluded those without internet literacy or devices (Groves, 2011b). Secondly, the relatively low cost of fielding a survey led to a saturation of survey invitations, potentially overwhelming the limited pool of the population willing to participate (Leeper, 2019). Whatever the particulars of the causes of this phenomena, response rates have fallen globally and across survey modes, with rates below 40% now very common, and this trend has continued even as recently as 2020 (Jabkowski and Cichocki, 2024).

Here I shall describe the relationship between nonresponse rates and nonresponse bias. Much research has been conducted on the extent of the problems caused by declining response rates and potential solutions. Studies such as Groves (2006) formalized the proposition that low response rates may be benign under certain circumstances. To understand when and why nonresponse bias would arise as a consequence of high nonresponse rates, it helps to examine the formal definition of nonresponse bias. Intuitively, nonresponse bias describes the extent to which a given sample-based inference would deviate from the true population value as a direct consequence of nonresponses. Groves (2006) formalized this in the following equation 1.1.

$$\mathrm{NRB}(\bar{y}) = \mathrm{NRR} \cdot (\bar{y}_r - \bar{y}_{nr}) \tag{1.1}$$

$\mathrm{NRB}(\bar{y})$ represents the nonresponse bias of the given variable y mean, NRR is the nonresponse rate, $\bar{y}_r$ is the mean of the variable for respondents, and $\bar{y}_{nr}$ is the mean of the variable for nonrespondents. Note that $\bar{y}_{nr}$ might practically be impossible to determine because, by definition, these individuals have not provided survey answers.

I will briefly note that Groves (2006) also proposed an alternative, and very popular, definition of nonresponse bias which is the correlation of the variable in question with each individual's 'response propensity', which is their probability of responding (Koch and Blohm, 2016). However, this dissertation does not use this definition because, as I shall detail below, the validity of those propensity estimates is under scrutiny in this dissertation, and so to separate the propensity measures from the bias measures, this research primarily uses the definition in 1.1.

Equation 1.1 has a few implications. Firstly, nonresponse bias is a per-variable phenomena meaning that in the same survey one variable can be very biased and another not so. Secondly, the severity of the bias is a function of the nonresponse rate as well as the differences in means of the respondent and nonrespondent subgroups. Intuitively, this is because sample-based inferences can vary from the true value because respondents and nonrespondents are very dissimilar but this is mitigated by the proportional

size of the nonresponding share. Therefore, if respondents and nonrespondents tend to yield a similar distribution of answers to the survey's questions (the survey variables), even a sample drawn from a low-responding population can yield accurate inferences. Conversely, even if there is a substantial difference in $\bar{y}$ between respondents and nonrespondents, if the nonresponse rate is very small, the subsequent bias in the inference of $\bar{y}$ will be small. For example, if individuals who choose not to respond to a hypothetical voter intention survey are more likely to support a given political party relative to those who respond, then the inferences drawn from that survey will likely under-estimate the amount of support for that party. However, if the survey was instead about some kind of health problem that affects supporters of any political party equally, this selective nonresponse behavior might not lead to a biased inference because the difference in variable means between responders and nonresponders is negligible.

Although low response rates might not always lead to nonresponse bias, survey researchers have developed strategies for what to do when this does occur. The most common solution is ex-post adjustment. Once a sample is collected, researchers typically check the distribution of the kinds of people who responded against some auxiliary data about the population's distribution[1]. For example, if the researchers are aiming to study the general U.S. population via a stratified random sample of U.S. residents they might check that the portion of males aged 50–70 in each sample strata roughly matches the values in the equivalent strata from the U.S. census. If they do not match, the researchers can use a weighting procedure (Massey and Tourangeau, 2013; Särndal and Lundström, 2005). For example, a common implementation of this procedure is model-based weighting, which estimates each individual's probability of being in the sample relative to every other individual in the same strata or population[2] and assigns individual level weights inversely proportional to that probability (Massey and Tourangeau, 2013; Särndal and Lundström, 2005). Subsequently, more under-represented individuals will contribute more to the estimated population-wide averages and distributions.

Several studies have demonstrated the efficacy of weighting procedures. For example, Duffy et al. (2005) found that weighting eliminated differences between face-to-face and telephone survey results. Iachan et al. (2016) demonstrated how different weighting procedures could bring the inferences from telephone surveys about obesity into alignment with results from more comprehensive physical examination-based surveys.

The optimistic perspective on declining response rates is that, as long as nonresponse is not systematically related to the study variables, inferences should be valid. If such a relationship exists, weighting can mitigate inaccuracies (Massey and Tourangeau, 2013). However, in the following passage I describe studies that cast doubt on this optimistic view.

In practice, low response rates have led to significant errors. In order for researchers to

[1] Although auxiliary data can serve many purposes, in this dissertation, I am primarily concerned with auxiliary data's role in ex-post adjustments.

[2] This procedure accounts for the estimated probability that an individual was invited (design weights), responded to the invitation (response propensity weights), and, if the sample still does not align with auxiliary data such as a census, may also apply additional adjustments to ensure alignment (calibration weights) (Haziza and Beaumont, 2017).

adjust biased samples they must anticipate in advance which survey variables to 'weight by,' meaning selecting the census variables to use in calibration weights or which predictor variables to include in their propensity estimation model (Little and Vartivarian, 2003; Kalton and Flores-Cervantes, 2003). Weighting can correct sample inferences when participants are under-represented along certain dimensions (i.e., age, gender, education) and those factors meaningfully affect the variables of interest. Thus, survey researchers must first identify the dimensions along which the sample under-represents the population, which is typically only possible with auxiliary data such as a census. Even when such data are available, it may still be the case that the under-represented dimension is not captured by variables collected in either the survey or the census (Groves, 2006).

For example, consider a survey about physical exercise habits. If work-related stress is a driver of both poor exercise routines and survey nonresponse, but neither the census nor the survey collects information about stress, then calibrating the sample to align with census variables such as age and gender will not correct for the under-representation of high-stress individuals. As a result, this procedure would fail to eliminate the nonresponse bias driven by work-related stress. The practical difficulty of identifying ideal weighting variables was explored in studies such as Kreuter et al. (2009). These researchers examined five large-scale surveys, searching for variables that correlated with both survey invitation refusal and key study variables. Importantly, these variables needed to be available even if the invitee did not respond, so the authors relied on auxiliary data, for example whether the household address was in a multi-unit complex. They found very few variables that strongly correlated with both nonresponse behavior and any of the study variables. In one survey, they were able to compare weighted and unweighted estimates against known population values, showing that variables more strongly correlated with both nonresponse and the target study variable produced better weighted estimates. However, the selection range of such useful variables was so limited that it became clear survey researchers could not reliably identify effective weighting variables for all outcomes of interest. This finding demonstrates how challenging it is, even for well-resourced survey studies, to determine and collect effective weighting variables in practice.

Several incorrect election predictions in the 2010s illustrate how these issues can lead to highly visible errors, undermining trust in survey research. Elections provide a unique opportunity to assess nonresponse bias, among other sources of error, in surveys. As highly publicized events, they allow pollsters to compare sample-based estimates of voting behaviors against the actual election results. Of course there are other reasons for electoral polls to mis-predict elections such as the effect of campaigns, world events, turnout, and swinging voters (Sciarini and Goldberg, 2015). However, as I shall explain shortly, nonresponse is likely a substantial confounder of survey-based election forecasts.

The 2016 United States presidential election was widely predicted to favor the Democrats but was won by the Republicans (Kennedy et al., 2018). Similarly, in the 2019 Australian election, all six major pollsters forecasted a victory for the Labor Party, yet the Liberal-National Coalition prevailed (Pennay, Misson, and Neiger, 2021). In both cases, pollsters were aware of nonresponse bias and applied weighting procedures, yet they still produced inaccurate results because the appropriate weighting variables were not evident until af-

ter the fact. In the 2016 U.S. election, polls over-represented college-educated voters, who tended to favor Democrats. Even pollsters who used weighting did not account for education due to concerns about the reliability of that data (White, 2020; Kennedy et al., 2018). In the 2019 Australian election, the six leading pollsters applied weighting, but, in hindsight, used incorrect variables. Pennay, Misson, and Neiger (2021) conducted a case study of one pollster, demonstrating that weighting by age, geography, and gender actually increased bias relative to the unweighted estimates, even inverting the predicted outcome. The challenge of correcting nonresponse bias is further exemplified in Sciarini and Goldberg (2015), who demonstrate how nonresponse is an issue in post-election surveys which tend to heavily over-represent voters versus nonvoters. If politically engaged individuals are more likely to respond to post-election surveys, this may also imply that less-engaged voters (or nonvoters) also tend to avoid pre-election surveys, thereby biasing polls towards more politically active people. Post-election surveys could use turnout behavior as a weighting variable, but pre-election surveys would need to use turnout intention which only approximates actual behavior, thereby making it harder for polls to account for a political engagement-based nonresponse bias. These electoral examples demonstrate that even when survey researchers are aware of the dangers of nonresponse bias and exercise precautions, these procedures can fail, particularly when it is difficult to identify and collect appropriate weighting variables, and the survey inferences can be incorrect. Furthermore, these errors are only detected at all because there is convenient comparison data available (the actual election outcome). This implies that similar errors are possible in other survey studies, but go undetected for want of this kind of validation data.

Addressing nonresponse bias remains an ongoing challenge for survey methodologists. This dissertation contributes to this broader research agenda by demonstrating how machine learning can help alleviate nonresponse bias in surveys. Three of the four papers in this dissertation focus on propensity modeling. These models are commonly used in weighting procedures. However, beyond weighting, predicting likely nonrespondents enables proactive intervention to induce responses from those who would otherwise not participate. Chapter 2 aims to improve the predictive accuracy of propensity models in longitudinal studies by better accounting for temporal dependencies. Chapter 3 introduces a novel technique for generating propensity estimates earlier in longitudinal surveys. Chapter 4 moves beyond predictive modeling, and explores how propensity estimates can be used to target longitudinal survey participants with various treatments aimed at retaining low-propensity individuals and what effect this has on bias.

The final paper, Chapter 5, shifts focus from propensity estimates to inference correction in the presence of nonresponse bias. Given that election mis-predictions in the 2010s were notable examples of the dangers of nonresponse bias, Chapter 5 concerns Australian voter intention polls. That study demonstrates how, even if polls, weighted or not, provide biased estimates, machine learning techniques can correct the subsequent forecasts.

Each paper in this dissertation applies established machine learning techniques to survey research in novel ways, with the overarching goal of improving nonresponse bias mitigation. To provide necessary context, the next section introduces key machine learn-

ing concepts before detailing each paper.

1.2 What is Machine Learning?

Machine learning is a subfield of computer science focused on algorithms that compute predictions or make decisions based on data without explicit programming for the task at hand. Instead, the model 'learns' how to complete the task (James et al., 2013). This process involves training a model on a dataset, referred to as 'training data,' to identify patterns and relationships. Once trained (a.k.a.'fitted') the model applies this learning to new data to make predictions or inform decisions. ML encompasses distinct paradigms, but this dissertation focuses on 'supervised' machine learning. In supervised learning, predictions are categorized as either regressions or classifications. Regressors are models that predict a continuous quantity, while classifiers output the probability that a given case belongs to a certain class (James et al., 2013). This dissertation is concerned with predicting whether a given survey invitee will nonrespond (Chapters 2-4) or with predicting the party that shall win certain electoral seats (Chapter 5) and so this dissertation is focused on supervised classification machine learning. The technique is called 'supervised' because it is trained on data from past cases in which the outcome to be predicted is already known (labeled) and that is how the model can be fitted to predict those same outcomes but for new cases where the outcome is not known. In the example of propensity modeling, one trains the ML models on individuals' survey data (age, gender, etc) and their past nonresponse behaviors, to guess if other individuals profiled by those same variables will nonrespond in their future. Formally, a supervised learning model can be represented with function 1.2.

$$\hat{y} = f(X) \tag{1.2}$$

where X represents the predictor variables, f is the model, and $\hat{y}$ is the predicted outcome. In the example of using supervised classification as a propensity model, $\hat{y}$ will be a value between zero and one representing the probability that the given individual will nonrespond (where 1.0 represents an estimated 100% chance of nonresponse).

Several key observations follow from equation 1.2. Firstly, model performance is typically evaluated by comparing predictions ($\hat{y}$) with actual outcomes (y), using metrics such as the average amount by which the predicted probability value deviates from the actual binary outcome value (where 1.0 signifies that the actual outcome was a nonresponse). These kinds of metrics typically measure the extent of the inaccuracy of the ML model's predictions, and this is called 'error' in ML parlance, but is distinct from 'error' in the context of surveys. Both 'model error' and 'survey error' refer to the extent to which an estimate deviates from an actual value, but each is calculated in very different ways as shall be elaborated on in this section. Another important consideration in evaluating an ML model is the baseline against which is it compared. Even imperfect models can be valuable so long as they improve over available alternatives in the given context (Hastie, Tibshirani, and Friedman, 2009).

Secondly, the predictor variables X are not limited to tabular data like a matrix of survey invitees and their demographic details. The primary requirement for input data is that it be represented numerically. X may include diverse formats, such as images represented as vectors of pixel values (Hastie, Tibshirani, and Friedman, 2009). The fact that ML models can accommodate complex data structures for X is critical for Chapter 2 where I explore how one can exploit the nature of time series data, that is, a sequence of matrices of survey invitees by survey variables, which are the time series of repeated survey waves in longitudinal studies. Chapter 2 illustrates how machine learning models tailored for time series analysis can uncover insights that only emerge when the order of events is taken into account.

Now I turn to the f term in equation 1.2. I mentioned that models were trained on historical data so as to make predictions on new data. There are a wide variety of algorithms that transform predictors X into predictions $\hat{y}$, and I shall describe some relevant examples shortly. Regardless of the particulars of the algorithm for f, all supervised learning algorithms share a common mechanism for training: they derive optimal model parameters to minimize prediction error. The process for deriving these parameter values can be computationally intensive and so progress in adopting these methods has been significantly accelerated by advances in hardware (Kingma and Ba, 2014). For example, neural networks, which I employ in Chapters 2 and 5, use a pseudo-random process called 'gradient descent' which requires trialing many different parameter values, measuring the subsequent error, and repeating until the optimal values are discovered (Haykin, 1994). Many algorithms have been developed to make this process as efficient as possible, but it is still a computationally intensive procedure which can involve trialing many combinations of parameters (Kingma and Ba, 2014).

It is important context for this dissertation that although linear regression and logistic regression fit the definition of a machine learning model stated above, they are firmly established techniques in survey research already (Allison, 2009; Särndal and Lundström, 2005; Kern, Klausch, and Kreuter, 2019). Therefore, current research on ML applications in survey methodology typically aims to explore more sophisticated ML techniques.

Now I turn to examples of the kinds of algorithms that serve the role of the f term in equation 1.2. Support Vector Machines (SVMs) trial multiple forms for the function f, such as polynomial or radial functions, thereby accounting for nonlinear[3] relationships between predictors and outcomes (Boser, Guyon, and Vapnik, 1992). Classification and Regression Trees (CART) make sequential, threshold-based decisions on predictor values (decision trees) with the aim of maximizing similarity between cases that are sorted into the same endpoints of those decision sequences (leaves), thereby automatically capturing interactions[4] and nonlinear relationships (Breiman et al., 1984). 'Ensemble' methods like random forests and gradient boosting enhance predictive performance by aggregating predictions from multiple CARTs (Breiman, 2001; Geurts, Ernst, and Wehenkel, 2006;

[3] In a simple linear regression, any change in quantity of a given predictor value will correspond to a constant proportional change in the output value as determined by the variable's weight, so a 'nonlinear' relationship is one in which that proportional change can vary over values of the predictor.

[4] An 'interaction' effect between predictor variables is where two or more variables have a combined effect as well as their own individual effects.

Friedman, 2001). Neural networks, loosely inspired by animal brains, consist of layered sequences of artificial neurons that apply linear regression functions to model highly complex relationships in data (Rumelhart, Hinton, and Williams, 1986). Even the basic logistic regression model can be made more sophisticated with "penalties" which are processes that ensure the model's weights are not too inflated such that a few predictors dominate the output and other, more subtle predictors are ignored (Le Cessie and Van Houwelingen, 1992; Tibshirani, 1996).

Given this diversity of options for f, one must decide which to use. Naturally, one could trial every possible type of model and select the one which yields the least predictive error. In Chapter 2, 3, and 5, we conduct such studies in which we retrospectively test how various forms of f would have performed had they been used to make forecasts at the time. However, there are important nuances to this task that will be critical background knowledge for this dissertation.

Firstly, one must determine what measure of model error one aims to minimize. There are numerous ways to calculate prediction error, and selecting the appropriate metric(s) must suit the specific context. In classification tasks, one is typically concerned with minimizing incorrect classifications, but there are different types of misclassifications. For example, if one aims to predict nonrespondents in an upcoming survey to potentially offer them an extra cash incentives, should researchers prioritize minimizing the number of participants who receive a treatment even though they would have responded anyway (false positives) or the number of individuals who do not respond but were not identified as at-risk by the model (false negatives)? For a survey researcher, this decision would likely depend on factors such as the estimated nonresponse rate and the cost of the treatment. Often there is a trade-off between these two concerns as one would train the model to be more strict in categorizing a case as at-risk so as to reduce wasted cash incentives, but more sensitive to the possibility of risk to avoid letting future nonresponders go unattended. Thus, the seemingly simple task of selecting the algorithm for f which yields the least error is complicated by the need to choose an appropriate error measure and to manage trade-offs between competing concerns. Therefore, each paper in this dissertation will discuss the appropriate measure of model error in its respective chapter.

Secondly, the model that achieves the highest accuracy when predicting cases drawn from the same data as that on which it was trained, might not be the best model for predicting new cases. This tendency is often a result of 'overfitting,' meaning the model learns patterns specific to the training set that do not generalize well to new data (Singh, Thakur, and Sharma, 2016). To address this issue, one must ensure they select the model most likely to succeed when applied to new data. The general solution involves withholding a portion of the data during training and subsequently testing the model on that withheld data which is called the 'test' data (James et al., 2013). Therefore, each paper in this dissertation will incorporate a methodology for separating test and training data and evaluating multiple models to select the best performer based on its performance on the holdout data. The goal of this process is to ensure that the models that appear to perform well in our retrospective predictions will likely perform just as well when used to make real forecasts.

Each paper in this dissertation follows this fundamental ML workflow: determining the target variable to be predicted (e.g., future nonresponse behavior), selecting predictors (e.g., past survey variables), and trialing and comparing a range of candidate models using carefully designed procedures for test-train data separation and error metric selection. However, beyond forecasting nonresponse, ML models have additional applications of interest as follows.

Firstly, when an ML model is well-suited for prediction within a particular domain, it can be used not only for prediction but also for understanding the mechanisms driving nonresponse behavior. *Interpretable* machine learning (Linardatos, Papastefanopoulos, and Kotsiantis, 2020) is a subfield of ML that focuses on investigating how the model works to gain deeper insights into the system it represents. This often involves taking a trained model and examining the impact of various predictors (and their combinations) on estimated outcomes to perhaps help derive the underlying causal mechanisms. In Chapter 2, I apply these techniques to understand the role of temporal dependencies in nonresponse predictions. In the case of Chapter 3, I examine what predictors appear to be most predictive of nonresponse so as to speculate on the drivers of nonresponse behavior.

Secondly, Chapter 3 explores applications of 'pre-training' (Devlin et al., 2019). Pre-training involves training an ML model on one domain (e.g., a specific longitudinal survey) and applying it to a different domain (e.g., another longitudinal survey), rather than merely using it for new cases within the same domain. This technique enables training a model on one survey and applying it to another without the need to retrain on the new study. Chapter 3 proposes this technique and explains how it facilitates earlier predictions in longitudinal studies than would otherwise be possible.

1.3 Machine Learning Applications to Nonresponse

This dissertation concerns the applications of ML to alleviating problems created by nonresponse in surveys. Much research has already been conducted in this area. Thus, to contextualize the contributions of my chapters, I provide an overview of this broader research agenda. This work primarily addresses two areas: Firstly, my research concerns propensity modeling, regarding both enhancing predictive accuracy and its use in mitigating nonresponse problems (Chapters 2-4). Secondly, my research concerns methods to correct sample inferences given the inevitability of some nonresponse bias (Chapter 5). In this section, I review previous research on applying ML to issues related to survey nonresponse, focusing particularly on these two applications. The subsequent section explains how my own papers contribute to this broader research agenda.

Machine learning is a powerful tool for making predictions, and researchers have explored numerous applications of ML to address the challenges posed by survey nonresponse. Although this dissertation focuses on propensity modeling and inference correction, there are numerous applications beyond these areas. For instance, if a participant responds only partially to a survey, ML can be used to impute the missing answers, thereby yielding a more complete dataset (Prakash et al., 2024). Argyle et al. (2023) em-

ployed Large Language Models (LLM)[5] to accurately estimate how certain demographic profiles of respondents might have answered certain surveys, although other studies have shown that this approach is not reliably replicable (Heyde, Haensch, and Wenz, 2025). Other researchers have sought to mitigate nonresponse by obtaining equivalent data from more passive sources. For example, researchers have explored the possibility that if individuals do not respond to surveys but instead post their thoughts on social media, then mining such data could replace or augment traditional survey data. In this context, ML could be used to transform raw social media content into population-wide inferences. However, this approach has yet to be demonstrated as more effective than conventional surveys, irrespective of nonresponse issues (Alvi et al., 2023; Amaya et al., 2021; Diaz et al., 2016; Buntain et al., 2016; Amaya et al., 2021). Beyond predicting nonresponse, ML-based modeling can address other adjacent issues such as checking for misreporting from unengaged respondents. For example, Bach, Eckman, and Daikeler (2020) applied a tree-based propensity model to classify low-propensity participants as "reluctant respondents" in order to investigate their tendency for motivated misreporting, such as speeding through interview questions. Their findings showed no evidence that low-propensity participants were more likely to engage in such behavior. Taken together, these studies show the diversity of uses for ML in mitigating nonresponse-related challenges, however this dissertation focuses primarily on propensity models and inference corrections.

With respect to propensity modeling, one might assume that the primary application of ML is to produce more accurate propensity estimates. Research into that possibility is ongoing, but I argue that there is value in exploring other applications for the propensity estimates that are already available. Logistic regression has traditionally been the most common algorithm used for propensity models intended for weighting corrections (Särndal and Lundström, 2005; Kern, Klausch, and Kreuter, 2019; Olson, 2013; Larbi et al., 2024; Buskirk and Kolenikov, 2015; Tourangeau, Groves, and Redline, 2010; Massey and Tourangeau, 2013; Dutwin and Buskirk, 2021; Groves and Peytcheva, 2008). Over the 2010s and 2020s, researchers have examined whether alternative ML algorithms might provide more accurate predictions of participant nonresponse and, consequently, improve the accuracy of model-based weighted inferences. The evidence suggests that while ML models do indeed yield more accurate predictions of nonresponse, that does not always correspond to an improvement in the accuracy of weighted sample inferences. This may be because predictive accuracy is only one of many factors that determine the success of weighting procedures. The variables driving nonresponse, their interactions with study variables, and the trade-offs between reducing bias in one variable versus exacerbating it in another may also play crucial roles (Larbi et al., 2024).

Buskirk and Kolenikov (2015) simulated the efficacy of various model-based weighting strategies and found that tree-based models were the most accurate at predicting future nonresponse, yet often produced less accurate sample-based inferences than basic logistic regressions. This finding was confirmed in a similar study by Larbi et al. (2024). This apparent contradiction, better predictions but worse weighted inferences,

[5] An LLM is a type of ML model that predicts bodies of text given context text as input.

can be explained by the principle that effective weighting procedures must adjust for variables correlated with both response propensity and the target survey variables. In this sense, more accurate propensity estimates may help identify variables that correlate with actual future response behavior. However, this is only one part of the challenge: a variable must also be correlated with the study outcomes to be effective for weighting (Kreuter et al., 2009). Rather than focus on improving nonresponse prediction even further for the sole purpose of weighting, this dissertation instead explores other uses for these predictive models.

Given that ML can yield better predictions, if not better weights, survey researchers have explored many uses for this enhanced prescience. For instance, Phillips et al. (2023) used logistic regression-based propensity models to estimate the number of survey invitations needed to meet a specific response quota. Another example comes from Zhang (2025), who applied linear and tree-based models to estimate the number of call attempts until an invitee either agrees or refuses to participate, aiding in the estimation of phone survey costs.

One of the most prominent applications of these ML-based nonresponse predictions is to inform adaptive survey design (ASD). ASD is a process in which survey managers adapt survey protocols so that specific types of invitees receive a survey experience tailored to them, as opposed to the traditional approach of providing the same experience to everyone. The goal is to more closely align the participation probabilities of under- and over-represented types of participants (Schouten, Peytchev, and Wagner, 2017). Many studies in this area have relied on simulation-based approaches to demonstrate the potential of various ASD implementations (Zhang and Wagner, 2024; Gummer, 2020; Schouten, Cobben, Lundquist, et al., 2016). For example, Watson and Cernat (2023) employed logit models to estimate propensity before and after a follow-up contact, and simulated how various ASD strategies could maintain sample size under various budget constraints for follow-ups. In another study, McCarthy, Wagner, and Sanders (2017) compared U.S. agricultural survey data against equivalent census benchmarks to simulate scenarios in which follow-up resources were diverted from high-propensity to low-propensity cases, examining the consequent effects on response rates and sample composition.

These simulation studies rely on the assumption that the estimated propensity scores are accurate, but research on propensity modeling shows that no model is 100% accurate (Kern, Weiß, and Kolb, 2021; Larbi et al., 2024; Buskirk and Kolenikov, 2015). Therefore, field experiments into how ML can direct ASD targeting are essential. My coauthors and I shall argue in Chapter 4 that there are too few field experiment studies that test ASD strategies for survey managers to review these reports and select an appealing strategy for their own context. An example of one of the few field experiments available is Beste et al. (2023), who ran a treatment-control experiment in which extra cash incentives were targeted at low-propensity households and found that this improved overall response rates.

An important open question in this field is "what is the best criteria to guide the targeted allocation of ASDs?" For example, how low should a given participant's propensity be before it's desirable to issue an extra cash incentive? Are there other criteria than

propensity to consider? A simulation study by Watson and Cernat (2023) suggests that incorporating measures of potential nonresponse bias (Schouten, Cobben, and Bethlehem, 2008) in combination with propensity estimates is better than using only propensity scores. The researchers found that this composite approach was better at reducing sample bias for a given budget for follow-ups.

Taken together, this research on ASD shows that propensity estimates can help survey research beyond just model-based weighting. Instead, propensity models and ASD together can potentially help improve the response rates and boost the presence of otherwise under-represented types of participants. With more of these under-represented persons in the sample, this further enhances the efficacy of weighting as there is then a greater sample size among this sub-group to better represent this particular type of participant (Kalton, 2009; Watson and Cernat, 2023).

1.4 This Dissertation's Contributions to Survey Methodology

Having now reviewed the big picture of ML's application to nonresponse-related issues, I will contextualize the contributions of this dissertation's four papers (two published, one accepted for publication, one under review, details below; Collins and Kern (2024) and Collins (2025)). The appended Section Statement of Contributions details each author's contributions to each paper.

Chapter 2: Longitudinal Nonresponse Prediction with Time-Series Machine Learning[6]

This chapter aims to improve propensity modeling in panel surveys by introducing novel time-series ML models. Traditional propensity models typically include demographic predictors (Särndal and Lundström, 2005). In panel surveys, where the same participants are repeatedly invited, such models can also incorporate past nonresponse behavior. A common approach is to use each participant's overall nonresponse rate across all waves to which they have been invited (Mulder and Kieruj, 2018; Zinn and Gnambs, 2022; Kocar and Biddle, 2022). However, this method implicitly assumes that missing a survey wave several years ago is just as indicative of future nonresponse as missing a more recent one, since each contributes equally to the overall average. In short, this approach ignores the time-series nature of panel data, treating the order and recency of events as irrelevant.

Prior research by Kern, Weiß, and Kolb (2021) demonstrated that accounting for the the time-series nature of panel data can improve predictive performance. Specifically, that study found that incorporating predictors such as rolling average nonresponse rates over the past one, two, three, and all waves improved predictive accuracy compared to using only the overall nonresponse rate.

[6]Published as: Collins, J. and C. Kern (2024). Longitudinal nonresponse prediction with time series machine learning. Journal of Survey Statistics and Methodology 13 (1), 128–159. doi: 10.1093/jssam/smae037.

These findings suggest that capturing both recent and more distant past non-response behavior can yield better predictions. However, using a fixed three-wave rolling average is arguably arbitrary. Why three waves? Is that optimal for all types of participants? Can we develop a more principled approach to modeling how prior nonresponse behavior predicts future nonresponse?

To address these questions, I explored ML techniques designed to model temporal dependency, which is defined here as accounting for how the sequence of events influences outcomes, not just the events themselves. For instance, in Kern, Weiß, and Kolb (2021), nonresponse over the past three waves served as a powerful predictor. In contrast, time-series models capture nuances such as the difference between a participant who missed one wave but then responded to the next two versus a participant who missed the most recent of three waves. Moreover, these models can dynamically determine how many prior waves are relevant, eliminating the need for arbitrary fixed window sizes (Collins and Kern, 2024).

The best-performing time-series model that I trialed, the Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), did not outperform the tree-based model from Kern, Weiß, and Kolb (2021), but it did produce comparable results without relying on manually engineered rolling average predictors. Instead, the LSTM effectively internalized and learned temporal patterns, essentially creating similar predictive features automatically. While the three-wave rolling window used in prior work was derived through trial and error, the LSTM learned such dependencies through its training process.

Thus, the key contribution of this chapter is demonstrating that LSTM models can effectively capture temporal dependencies in panel survey nonresponse. For survey researchers, this means that LSTMs can serve as a validation tool: if an LSTM outperforms a current model, it suggests that the model may not be sufficiently capturing the time-dependent nature of nonresponse. This approach provides a concrete method for assessing whether existing models adequately incorporate both recent and distant response behaviors.

Chapter 3: Pre-Trained Nonresponse Prediction in Panel Surveys with Machine Learning[7]

This chapter presents two notable findings for survey researchers aiming to address nonresponse issues. First, any predictive modeling effort must consider whether the method generalizes to new contexts. For example, would a tree-based model that incorporates past nonresponse behavior perform well across all panel surveys? Or might such an approach be more effective in weekly panels than in annual ones? If so, this would limit the generalizability of lessons learned from a single study to the broader field of survey research. Therefore, my coauthor and I investigated whether the modeling approach described in Chapter 2, developed using the GESIS Panel, would also prove effective when applied to four other panel datasets. We found that it did generalize well.

[7] Accepted for publication in *Survey Research Methods*.

This finding should interest survey researchers considering converting their propensity modeling approach from logistic regression to tree-based models because it shows the predictive technique is ubiquitously effective across a range of contexts.

Second, I introduced a novel possibility: if a given set of predictors and algorithms performs well across a variety of panels, could it be feasible to pre-train a model? This would involve training a model on one panel survey and then applying it to a different panel, expecting good predictive performance despite the model not being trained on the target panel's data. Initially, this prospect seemed unlikely for several reasons: panels often differ in scope, recruitment methods, and timing. Moreover, machine learning models are typically sensitive to differences between training and testing domains. Nevertheless, this study demonstrated that pre-trained models can produce predictions that are comparable in quality to those generated by models trained on the target panel.

The key advantage of this approach is that it allows researchers to generate response propensity estimates from the earliest waves of a new panel, without the need to wait to accumulate training data. This is particularly valuable given that many panels experience the highest attrition in their initial waves (see Chapter 3).

Chapter 4: Prediction-Based Adaptive Designs for Reducing Attrition Rates and Bias in Panel Surveys[8]

Numerous authors have applied ML to propensity modeling. The natural next step is to leverage participants' estimated nonresponse risk scores to reduce both nonresponse rates and bias. Previous studies in this area have typically relied on assumption-driven simulations to explore how targeting at-risk participants with modified survey protocols, such as additional cash incentives or shortened surveys, might affect outcomes (Zhang and Wagner, 2024; Gummer, 2020; Watson and Cernat, 2023; Schouten, Cobben, Lundquist, et al., 2016). These simulations often assume theoretical treatment effects or treat the outputs of a propensity model as true propensities, despite the fact that these are merely estimates. Other studies have implemented examples of targeted adaptive designs, but usually examined only a limited set of intervention strategies (Lynn, 2016; Zhang and Wagner, 2024; Gummer and Blumenstiel, 2018; Wagner et al., 2012).

This chapter presents a framework for conducting ex-post simulations with minimal assumptions, providing robust evidence for the likely effects of various adaptive survey designs and ML-based targeting strategies (e.g., targeting those most at risk of nonresponse or those at moderate risk). We describe this process as "minimal-assumptions-based" because the treatment effects are derived from the actual results of a treatment-control experiment, rather than being estimated through modeling.

The primary contribution of this chapter is the demonstration of a generalizable framework that survey researchers can adopt. In our empirical demonstration, we offer compelling evidence that certain strategies can reduce nonresponse rates by 1–2 percentage points. While the effects of these ASD strategies on nonresponse bias were mixed, this mixed result is to be expected given the variable-specific nature of bias.

[8]Submitted to *Sociological Methods and Research*.

We propose that any panel study can implement this approach to simultaneously evaluate a wide range of ASD strategies. Until now, panel managers have relied either on assumption-heavy simulations or costly one-ASD-at-a-time field experiments to assess potential ASDs. Our framework enables the evaluation of multiple design options concurrently, with minimal reliance on assumptions about how these strategies will perform in practice.

Chapter 5: Predicting Australian Federal Electoral Seats with Machine Learning[9]

If the abstract goal of a general population survey is to estimate the distribution of a given set of variables across a population using a sample, then polls-based forecasting is an implementation of this abstraction. The objective of polls-based election forecasting is to estimate voter behavior at the population level, often regarding how these behaviors are distributed across strata such as geographically contiguous electoral divisions or states. These estimates typically rely on one or more polls, which are usually surveys of prospective voters.

As discussed, electoral mispredictions represent highly visible failures of modern survey practice, often caused by low response rates and challenges in selecting the appropriate weighting variables. Elections are unique in that they come with an auxiliary dataset (the actual election outcome) against which researchers can benchmark the accuracy of their pre-election polls. Of course, this only assists in evaluating errors after the fact.

Since the misforecast of the 2016 U.S. Presidential election, pollsters have worked diligently to improve polling accuracy. These efforts include varying survey protocols to better engage reluctant respondents and incorporating additional weighting variables, particularly education, which was a critical issue in 2016 (Keeter and Kennedy, 2024).

This chapter introduces a complementary technique for adjusting polls-based election forecasts, that is, employing ML to learn patterns in the errors between poll-based predictions and actual outcomes from past elections. These models are then used to correct polls-based estimates and generate seat-level forecasts. Unlike these other initiatives aimed at improving the accuracy of polling inferences, this approach accepts the presence of survey errors and seeks to account for them to make better predictions. In essence, this chapter addresses what can be done given the presence of nonresponse bias, rather than how to reduce it.

In the context of Australian election forecasting, the task is to estimate the electoral outcomes across seats. The traditional baseline for such forecasts in Australia is a polls-based model known as the 'Mackerras Pendulum,' which is described in detail in Chapter 5. I demonstrate that these pendulum forecasts can be significantly improved by using them as predictors within an ML model that incorporates additional variables (electoral division level census data, past electoral outcomes, etc) to adjust the pendulum's predictions.

[9] Published as: Collins, John 'Jack' (2025). "Predicting Australian federal electoral seats with machine learning". In: International Journal of Forecasting. doi: 10.1016/j.ijforecast.2025.02.002.

1.5 References

Allison, Paul D. (2009). *Fixed Effects Regression Models*. SAGE Publications, Inc.

Alvi, Quratulain et al. (2023). "On the frontiers of Twitter data and sentiment analysis in election prediction: a review". In: *PeerJ Computer Science* 9, p. 1517.

Amaya, Ashley et al. (2021). "New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data". In: *Social Science Computer Review* 39.5, pp. 943–960.

Argyle, Lisa P. et al. (2023). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis*. Publisher: Cambridge University Press, pp. 1–15.

Bach, Ruben L, Stephanie Eckman, and Jessica Daikeler (2020). "Misreporting Among Reluctant Respondents". In: *Journal of Survey Statistics and Methodology* 8.3, pp. 566–588.

Beste, Jonas et al. (2023). "Case Prioritization in a Panel Survey Based on Predicting Hard to Survey Households by Machine Learning Algorithms: An Experimental Study". In: *Survey Research Methods* 17.3, pp. 243–268.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. Association for Computing Machinery, New York, pp. 144–152.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Breiman, Leo et al. (1984). *Classification and regression trees*. 1st Ed. Chapman and Hall, New York.

Buntain, Cody et al. (2016). "Comparing Social Media and Traditional Surveys Around the Boston Marathon Bombing". In: *Proceedings of the 25th International Conference on World Wide Web*. Montreal, Canada.

Buskirk, Trent and Stanislav Kolenikov (2015). "Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification". In: *Survey Methods: Insights from the Field*.

Collins, John and Christoph Kern (2024). "Longitudinal Nonresponse Prediction with Time Series Machine Learning". In: *Journal of Survey Statistics and Methodology* 13.1, pp. 128–159.

Collins, John 'Jack' (2025). "Predicting Australian federal electoral seats with machine learning". In: *International Journal of Forecasting*.

Daikeler, Jessica, Michael Bošnjak, and Katja Lozar Manfreda (2020). "Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates". In: *Journal of Survey Statistics and Methodology* 8.3, pp. 513–539.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Diaz, Fernando et al. (2016). "Online and Social Media Data As an Imperfect Continuous Panel Survey". In: *PLOS ONE* 11.1. Ed. by Cédric Sueur, e0145406.

Duffy, Bobby et al. (2005). "Comparing Data from Online and Face-to-face Surveys". In: *International Journal of Market Research* 47.6, pp. 615–639.

Dutwin, David and Trent D Buskirk (2021). "Telephone Sample Surveys: Dearly Beloved or Nearly Departed? Trends in Survey Errors in the Era of Declining Response Rates". In: *Journal of Survey Statistics and Methodology* 9.3, pp. 353–380.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232.

Gallup, George and Saul Rae (1940). *The Pulse of Democracy*. Simon and Schuster, New York.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel (2006). "Extremely randomized trees". In: *Machine Learning* 63.1, pp. 3–42.

Groves, Robert M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys". In: *Public Opinion Quarterly* 70.5, pp. 646–675.

— (2011a). *Survey Methodology*. Ed. by Floyd J. Fowler et al. 2nd ed. Wiley Series in Survey Methodology. Wiley, Somerset.

— (2011b). "Three Eras of Survey Research". In: *The Public Opinion Quarterly* 75.5. Publisher: Oxford University Press on behalf of the American Association for Public Opinion Research, pp. 861–871.

Groves, Robert M. and Emilia Peytcheva (2008). "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis". In: *Public Opinion Quarterly* 72.2, pp. 167–189.

Gummer, Tobias (2020). "Adaptive and Responsive Survey Designs". In: *SAGE Research Methods Foundations*. Ed. by P. Atkinson et al. 2020.

Gummer, Tobias and Jan Eric Blumenstiel (2018). "Experimental Evidence on Reducing Nonresponse Bias through Case Prioritization: The Allocation of Interviewers". In: *Field Methods* 30.2. Publisher: SAGE Publications Inc, pp. 124–139.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

Haykin, Simon (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Haziza, David and Jean-François Beaumont (2017). "Construction of Weights in Surveys: A Review". In: *Statistical Science* 32.2, pp. 206–226.

Heyde, Leah von der, Anna-Carolina Haensch, and Alexander Wenz (2025). "Vox Populi, Vox AI? Using Large Language Models to Estimate German Vote Choice". In: *Social Science Computer Review* 0.0.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Iachan, Ronaldo et al. (2016). "National weighting of data from the Behavioral Risk Factor Surveillance System (BRFSS)". In: *BMC Medical Research Methodology* 16.1, p. 155.

Jabkowski, Piotr and Piotr Cichocki (2024). "Survey response rates in European comparative surveys: a 20-year decline irrespective of sampling frames or survey modes". en. In: *Quality & Quantity*.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. Springer, New York.

Kalton, G. and I. Flores-Cervantes (2003). "Weighting methods". In: *Journal of Official Statistics* 19, pp. 81–97.

Kalton, Graham (2009). "Methods for oversampling rare subpopulations in social surveys". In: *Survey Methodology* 35, pp. 125–141.

Keeter, Scott and Courtney Kennedy (2024). *Key things to know about U.S. election polling in 2024*. Pew Research Center URL: `https://www.pewresearch.org/short-reads/2024/08/28/key-things-to-know-about-us-election-polling-in-2024/`.

Kennedy, Courtney et al. (2018). "An Evaluation of the 2016 Election Polls in the United States". In: *Public Opinion Quarterly* 82.1, pp. 1–33.

Kern, Christoph, Thomas Klausch, and Frauke Kreuter (2019). "Tree-based Machine Learning Methods for Survey Research". In: *Survey research methods* 13.1, pp. 73–93.

Kern, Christoph, Bernd Weiß, and Jan-Philipp Kolb (2021). "Predicting Nonresponse in Future Waves of a Probability-Based Mixed-Mode Panel with Machine Learning*". In: *Journal of Survey Statistics and Methodology* 11.1, pp. 100–123.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980*.

Kocar, Sebastian and Nicholas Biddle (2022). "The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design". In: *Quality & Quantity*.

Koch, Achim and Michael Blohm (2016). "Nonresponse Bias". In: *GESIS Survey Guidelines*. GESIS-Leibniz-Institut Für Sozialwissenschaften, Mannheim, Germany.

Kreuter, F. et al. (2009). "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 173.2, pp. 389–407.

Larbi, Khaled et al. (2024). "On the use of Machine Learning methods for the treatment of unit nonresponse in surveys". In: *2nd Workshop on Methodologies for Official Statistics: Proceedings*. Ed. by Orietta Luzi et al. Istituto Nazionale di Statistica, Roma, p. 1.

Le Cessie, S. and J. C. Van Houwelingen (1992). "Ridge Estimators in Logistic Regression". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41.1, pp. 191–201.

Leeper, Thomas J (2019). "Where Have the Respondents Gone? Perhaps We Ate Them All". In: *Public Opinion Quarterly* 83 (S1), pp. 280–288.

Likert, Rensis (1932). "A technique for the measurement of attitudes". In: *Archives of Psychology* 22.140.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis (2020). "Explainable AI: A Review of Machine Learning Interpretability Methods". In: *Entropy* 23.1.

Little, R. J. and S. Vartivarian (2003). "On weighting the rates in non-response weights". In: *Statistics in Medicine* 22, pp. 1589–1599.

Lynn, Peter (2016). "Targeted Appeals for Participation in Letters to Panel Survey Members". In: *Public Opinion Quarterly* 80.3, pp. 771–782.

Massey, Douglas S. and Roger Tourangeau (2013). "Where Do We Go from Here? Nonresponse and Social Measurement". In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 222–236.

McCarthy, Jaki, James Wagner, and Herschel Lisette Sanders (2017). "The Impact of Targeted Data Collection on Nonresponse Bias in an Establishment Survey: A Simulation Study of Adaptive Survey Design". In: *Journal of Official Statistics* 33.3, pp. 857–871.

Mitchell, Travis (2017). *What Low Response Rates Mean for Telephone Surveys*. Pew Research Center URL: `https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/`.

Mulder, J and N Kieruj (2018). *Preserving Our Precious Respondents: Predicting and Preventing Non-Response and Panel Attrition by Analyzing and Modeling Longitudinal Survey and Paradata Using Data Science Techniques*. 2018.

Neyman, Jerzy (1934). "On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection". In: *Royal Statistical Society* 97.4, pp. 541–724.

Olson, Kristen (2013). "Paradata for Nonresponse Adjustment". In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 142–170.

Pennay, D., S. Misson, and D. Neiger (2021). *The impact of weighting by educational attainment and past vote on estimates of pre-election voting intentions: A case study using Australian polling data*. Working Paper 2/2021. CSRM & SRC Methods Paper No. 2/2021. Centre for Social Research and Methods (CSRM), Social Research Centre (SRC), 2021.

Phillips, Benjamin et al. (2023). *Australian Comparative Study of Survey Methods: Technical Report*. Melbourne: Social Research Centre, 2023.

Prakash, Preethi et al. (2024). "Benchmarking machine learning missing data imputation methods in large-scale mental health survey databases". In: *Artificial Intelligence in Health* 2.1, pp. 81–92.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088, pp. 533–536.

Särndal, Carl-Erik and Sixten Lundström (2005). *Estimation in Surveys with Nonresponse*. Wiley, Chichester.

Schouten, Barry, Fannie Cobben, and Jelke Bethlehem (2008). "Indicators for the Representativeness of Survey Response". In: *Survey Methodol* 35.

Schouten, Barry, Fannie Cobben, Peter Lundquist, et al. (2016). "Does more balanced survey response imply less non-response bias?" In: *Journal of the Royal Statistical Society Series A* 179.3. Number: 3 Publisher: Royal Statistical Society, pp. 727–748.

Schouten, Barry, Andy Peytchev, and James Wagner (2017). *Adaptive Survey Design*. Chapman and Hall.

Sciarini, Pascal and Andreas C. Goldberg (2015). "Lost on the Way: Nonresponse and its Influence on Turnout Bias in Postelection Surveys". In: *International Journal of Public Opinion Research* 29.2, pp. 291–315.

Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma (2016). "A review of supervised machine learning algorithms". In: *Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development*, pp. 1310–1315.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Tourangeau, Roger, Robert M. Groves, and Cleo D. Redline (2010). "Sensitive Topics and Reluctant Respondents: Demonstrating a Link Between Nonresponse Bias and Measurement Error". In: *The Public Opinion Quarterly* 74.3, pp. 413–432.

Wagner, James et al. (2012). "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection". In: *Journal of Official Statistics*, p. 23.

Watson, Nicole and Alexandru Cernat (2023). "Simulating the Consequences of Adaptive Survey Design in Two Household Panel Studies". In: *Journal of Survey Statistics and Methodology* 11.4, pp. 806–828.

White, Laurel (2020). *Polls Missed The Mark In 2016. But Experts Say Things Are Different In 2020.* WPR URL: `https://www.wpr.org/politics/polls-missed-mark-2016-experts-say-things-are-different-2020`.

Williams, Douglas and J Michael Brick (2018). "Trends in U.S. Face-To-Face Household Survey Nonresponse and Level of Effort". In: *Journal of Survey Statistics and Methodology* 6.2, pp. 186–211.

Zhang, Shiyu and James Wagner (2024). "The Additional Effects of Adaptive Survey Design Beyond Post-Survey Adjustment: An Experimental Evaluation". In: *Sociological Methods & Research* 53.3, pp. 1350–1383.

Zhang, Xinyu (2025). "Dynamic Time-to-Event Models for Future Call Attempts Required Until Interview or Refusal". In: *Journal of Official Statistics*.

Zinn, Sabine and Timo Gnambs (2022). "Analyzing nonresponse in longitudinal surveys using Bayesian additive regression trees: A nonparametric event history analysis". In: *Social Science Computer Review* 40.3, pp. 678–699.

2 Longitudinal Nonresponse Prediction with Time Series Machine Learning

Abstract

Panel surveys are an important tool for social science researchers, but nonresponse in any panel wave can significantly reduce data quality. Panel managers then attempt to identify participants who may be at risk of not participating using predictive models to target interventions before data collection through adaptive designs. Previous research has shown that these predictions can be improved by accounting for a sample member's behavior in past waves. These past behaviors are often operationalized through rolling average variables that aggregate information over the past two, three, or all waves, such as each participant's nonresponse rate. However, it is possible that this approach is too simple. In this paper, we evaluate models that account for more nuanced temporal dependency, namely recurrent neural networks (RNNs) and feature-, interval-, and kernel-based time series classification techniques. We compare these novel techniques' performances to more traditional logistic regression and tree-based models in predicting future panel survey nonresponse. We apply these algorithms to predict nonresponse in the GESIS Panel, a large-scale, probability-based German longitudinal study, between 2013 and 2021. Our findings show that RNNs perform similarly to tree-based approaches, but the RNNs do not require the analyst to create rolling average variables. More complex feature-, interval-, and kernel-based techniques are not more effective at classifying future respondents and nonrespondents than RNNs or traditional logistic regression or tree-based methods. We find that predicting nonresponse of newly recruited participants is a more difficult task, and basic RNN models and penalized logistic regression performed best in this situation. We conclude that RNNs may be better at classifying future response propensity than traditional logistic regression and tree-based approaches when the association between time-varying characteristics and survey participation is complex but did not do so in the current analysis when a traditional rolling averages approach yielded comparable results.

Statement of Significance

Panel survey practitioners increasingly use predictive modeling to anticipate participant nonresponse. This paper provides the results from a large-scale application of a novel set of predictive techniques. Traditional approaches to predicting participant nonresponse use each participant's most recent or historical average variable values. An example variable is the average of a participant's last three survey satisfaction ratings.

However, information is lost in that process: perhaps a participant suddenly expresses a much lower satisfaction than usual as opposed to an overall low average. By aggregating values over a time range, the information of the sudden drop is lost. The benefit of time series techniques is that the model can account for this information.

Our results show that this innovation could improve predictive accuracy in the proper contexts. This research indicates that RNNs could be a useful technique for panel managers developing their own predictive models when temporal dependencies are complex, but otherwise, rolling average predictors might suffice. Our study also shows the impact of sample refreshments. When new participants are recruited into a panel, modelers may predict their nonresponse propensities using data from pre-existing participants. We find that predicting nonresponse of newly recruited participants is more difficult than predicting future nonresponse of existing panel members, and basic RNN models and penalized logistic regression performed best in this situation.

Statements

This study design and analysis was not preregistered.

2.1 Introduction

Panel surveys, also called longitudinal surveys, are an irreplaceable method for data collection. Panel studies require significantly more operational costs and skilled management than cross-sectional surveys (Pforr and Schröder, 2016). Because they are a relatively significant investment and a critical resource to researchers, sources of error must be controlled as much as possible (Pforr and Schröder, 2016). Adaptive Design (AD) was developed for panel survey managers to intervene before a survey's collection period and reduce nonresponse bias. There are two parts to AD: first, panel managers identify which participants are at risk of nonresponding, and second, the panel managers adapt survey protocols to induce those at-risk participants to respond (Chun, Heeringa, and Schouten, 2018; Coffey, Reist, and Miller, 2020; Groves, 2006; Hoel, Sobel, and Weiss, 1975; Lynn, 2017; Peytchev, Pratt, and Duprey, 2022; Wagner, 2008). This paper contributes to exploring more effective approaches to anticipating nonresponse.

In the search for new ways to predict which participants in a panel are at risk of nonresponding, survey research has turned to machine learning (ML). In the ML approach (see James et al. (2013)), modelers use predictive algorithms that automatically "learn" how to predict a participant's likelihood of nonresponding based on data from previous survey waves. This process is called model training or fitting in ML terminology. For each participant for whom we input their predictor variables into the model, the output is the predicted probability that they will nonrespond in the next wave. Modelers do not know in advance which algorithm will yield the best predictions, so they undertake broad model comparisons. In this approach, the retrospective prediction quality of many different algorithms, and many parameterizations of those algorithms, are tested and compared so that the modeler can select the best one by some quality measurement.

Previous research has demonstrated that ML can accurately predict participant nonresponse. However, this field of study has not progressed beyond trialing various (traditional) ML models and reporting performance (Behr, Bellgardt, and Rendtel, 2005; Gummer, Roßmann, and Silber, 2021; Roßmann and Gummer, 2016; Trappmann, Gramlich, and Mosthaf, 2015; Uhrig, 2008; Zinn and Gnambs, 2022). An under-researched matter is how to utilize the temporal nature of panel data. A participant's likelihood of nonresponding may be discernible from the timeline of their behavior. For example, a decline in survey satisfaction or increasing break-off rates might precede disengagement with the panel. Most studies that consider temporal dependencies when predicting nonresponse do so by aggregating past participation behavior or reports to survey questions over waves (Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022; Roßmann and Gummer, 2016; Trappmann, Gramlich, and Mosthaf, 2015; Uhrig, 2008; Zinn and Gnambs, 2022). An example of such a predictor is the rolling average of a survey satisfaction rating on a five-point scale over the past three waves (Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022).

Predictors such as rolling averages exclude information about the order of events. For example, if a participant had a period of high survey satisfaction followed by a sudden drop, this could have the same rolling average value as a participant who always had a steady medium average satisfaction. Predictive algorithms that discriminate between these two sequences might yield better predictive performance. Another issue with rolling average predictors is that the practitioner must select the time range (i.e., average over three, six, or all waves). Determining the best time range value can only be done through trial and error, which increases the modeling effort. Also, deriving multiple rolling averages with different time ranges from the same predictor adds multicollinearity to the model, which can confound certain models (James et al., 2013).

Time series machine learning models are a broad set of algorithms that can automatically model time series data and, unlike rolling average variables, can account for the sequence of events. We explore the potential of time series techniques by applying them to the GESIS Panel. The GESIS Panel is a German general population omnibus survey that has collected waves of data every two to three months since 2013. We guide our inquiry through the following research questions.

- **Q1:** Can time series machine learning techniques account for the sequence of events in panel data instead of rolling average predictors?
- **Q2:** Do these techniques outperform traditional models that use rolling average predictors?

We also address a secondary question in this field. Many long-running panel surveys conduct periodic sample refreshments, in which new participants are recruited to restore a diminished sample size to its initial quantity. Therefore, our research question is as follows:

- **Q3:** When a machine learning model is trained on panel waves that predate the refreshment intake, can the model make accurate predictions about fresh participants, who were not present in the training data?

To address research questions one and two, we test two sets of novel time series classification techniques. First, we apply recurrent neural networks (RNNs), specifically the simple RNN (Rumelhart, Hinton, and Williams, 1986), gated recurrent unit (GRU; Cho et al. (2014)), and long short-term memory (LSTM; Hochreiter and Schmidhuber (1997)) techniques. RNNs are machine learning algorithms that make predictions using time series data, which naturally suits a panel survey. RNN-based models are commonly used for tasks such as speech recognition from audio data or language translation from text sequences (Graves, Mohamed, and Hinton, 2013; Liu et al., 2014). Second, we investigate kernel-, distance-, and feature-based time series classification algorithms. These techniques transform time series data into variables that describe the time series characteristics. Example variables include the frequency of occurrences across binned value ranges or the highest number of consecutive declines in value over a time series. These variables are then used as predictors to forecast nonresponse. We expect these approaches to yield better predictive performance than previous studies that used standard ML techniques and rolling average predictors.

Regarding research question three, examining the effect of refreshment intakes can tell us whether a model trained on participants from one recruitment wave can be successfully applied to individuals from a new intake. We address this question by training our models on participants from one recruitment wave and then using them to predict participants from a freshly recruited sample. Then, we compare the prediction performance to models trained on and applied only to the same set of participants. This comparison allows us to assess the decline in performance caused by introducing new participants.

In the following sections, we discuss the details of previous research, specifically the types of ML algorithms already investigated, the predictors used, and the achieved predictive performances. We then describe the time series algorithms we intend to apply and why we expect them to outperform these previous approaches.

2.2 Background

2.2.1 Nonresponse Prediction in Previous Research

The most common ML model type used in prior research on nonresponse prediction is logistic regression (Bach, Eckman, and Daikeler, 2020; Hill et al., 2020; Jacobsen et al., 2021; Jankowsky, Steger, and Schroeders, 2022; Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022; Lemay, 2009; Lipps, 2007; Lugtig, 2014; Mulder and Kieruj, 2018; Roßmann and Gummer, 2016; Siegers, Steinhauer, and Dührsen, 2021; Uhrig, 2008; Voorpostel and Lipps, 2011). One major limitation of logistic regression is its limited flexibility and, thus, the need for careful model specification. This limitation means that when the impact of one variable on the outcome is highly non-linear or dependent on another variable (for example, the impact of income on nonresponse propensity could be lower for participants over the retirement age), a simple logistic regression will not account for this dynamic. Models that can automatically account for interaction effects without the need to explicitly define interaction terms, such as random forests, gradient boosting, Bayesian additive regression trees (BART), and feed-forward neural networks

(FNNs), have also been analyzed by previous research on panel nonresponse (Bach, Eckman, and Daikeler, 2020; Kern, Weiß, and Kolb, 2021; Mulder and Kieruj, 2018; Zinn and Gnambs, 2022).

Past research predicting nonresponse have used indicators of survey engagement, respondent demographics, paradata, and information about past survey participation behaviors. Survey engagement is the extent to which the participant is motivated or committed to continuing the survey. Examples include survey break-offs, unanswered phone calls, rejected invitations, self-reported survey satisfaction, and previous nonresponse (Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022; Mulder and Kieruj, 2018; Olson, 2013; Roßmann and Gummer, 2016). Demographic predictors commonly include gender, age, education level, employment status, and income (Burkam and Lee, 1998; Richter, Körtner, and Saßenroth, 2014; Uhrig, 2008; Zinn and Gnambs, 2022). Previous studies have shown that participants who are male, younger, less educated, or unemployed are more prone to unit nonresponse (Becker, 2017; Kocar and Biddle, 2022).

Many studies have found that paradata can be used to derive useful predictors, including indicators of survey engagement or difficulty with completion (Kocar and Biddle, 2022; Olson, 2013; Roßmann and Gummer, 2016; Sarndal and Lundquist, 2014; Struminskaya and Gummer, 2022; Tienda and Koffman, 2021). Examples include survey completion times, browser type, whether the survey was conducted online or by mail, and the interviewer's identity.

Other studies have also used predictors that reflect changes in a characteristic over time, sometimes called time-varying predictors (Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022). In this approach, we expect to detect signs of participant disengagement in the trends of their behaviors. For example, a participant with declining self-reported survey satisfaction may be at risk of attrition. An example predictor is the average response rate of a participant over the previous three waves, six waves, and so on.

We aim to answer research questions one and two by comparing novel time series techniques with the approaches of previous studies. However, it is difficult to compare all studies on predicting nonresponse due to the range of contexts and different metrics used to report model performance. Many papers that use logistic regression report only odds ratios, pseudo R2, and robustness results (Hill et al., 2020; Lipps, 2007; Lugtig, 2014; Roßmann and Gummer, 2016; Siegers, Steinhauer, and Dührsen, 2021; Uhrig, 2008). However, this paper aims to build a predictive model, so we seek studies that publish metrics of predictive performance with which we can make comparisons. In the context of panel nonresponse, predictive performance is commonly evaluated on the basis that most of the participants predicted to nonrespond do so (called precision), and that most of those who nonrespond are correctly identified by the model (called recall).

Another common metric in previous studies is the area under the receiver operator curve (AUROC), which quantifies how well the model makes trade-offs between yielding too many false positives (participants predicted to nonrespond but then did not) and too few true positives (participants predicted to nonrespond who then did so). An AUROC of 1.0 is a perfect classifier, while an AUROC of 0.5 is the worst possible classifier (James et al., 2013). Mulder and Kieruj (2018), working with the Longitudinal Internet Studies for the Social Sciences (LISS) panel, report AUROC scores of 0.65–0.79 after

implementing logistic regression, neural networks, support vector machine (SVM), and tree-based models. Of these models, random forest performed the best. Bach, Eckman, and Daikeler (2020) applied logistic regression and gradient-boosted models to two waves of the LISS Panel and achieved an AUROC of 0.88–0.89. Kern, Weiß, and Kolb (2021) applied logistic regression, tree-based models, and gradient-boosted models to the GESIS Panel (the same survey used in this paper) up to mid-2017 and achieved an AUROC of 0.86–0.89.

Not all studies report their results in AUROC. Zinn and Gnambs (2022) report that their BART model was 95 to 97% accurate (meaning the portion of correct response/nonresponse predictions out of all predictions) across the first five waves of the National Educational Panel Study (NEPS). However, they do not report the recall, precision, or AUROC scores. Kocar and Biddle (2022, pp. 17), using the Life in Australia survey, claim: "With our models, we could correctly identify 90% (or more) of all nonrespondents (recall = 0.9), but for a high price of about five false positives for one true positive (precision = 0.17)." Each of these studies provide context for our own results. Specifically, we expect our own modeling efforts to achieve similar scores when we apply logistic regression and tree-based models fitted with equivalent predictors. Once we achieve a successful baseline set of traditional ML models, we can then evaluate the value added by time-series modeling. We summarize these baseline studies in Table 2.1.

2.2.2 Our Novel Approach: Time Series Models

Recurrent Neural Networks

An RNN is an ML algorithm suited for making predictions based on time-variant data (Zargar, 2021). Examples include predicting the next word that should appear in an incomplete sentence based on the preceding word sequence. In the case of predicting nonresponse, we could compare RNNs with a traditional model, such as a logistic regression. Suppose we train a logistic regression model that predicts whether a given participant will respond in the next wave based only on whether the participant responded in the current wave. If the participant did not respond, the model would predict that the participant also would not respond in the next wave. However, before the current wave, this participant had been extremely reliable. An RNN could account for this participant's previous response history and estimate a higher chance that they will respond in the next wave. In this simplified example, the RNN had access to information that the traditional model did not.

As previously discussed, the rolling average predictors described above allow traditional models to include information from multiple past waves. However, the advantage of the RNNs over traditional models with rolling average predictors is that they can learn how much to allow past events to influence a prediction. For example, it may be that the response status from four to five waves ago is less important than the response status from two to three waves ago, or it may be more or equally important. RNNs can learn these nuances automatically (DiPietro and Hager, 2020; Graves, Mohamed, and Hinton, 2013; Kumar et al., 2018; Ribeiro et al., 2020; Salman et al., 2018; Shewalkar,

Table 2.1: *Summary of previous studies predicting panel nonresponse.*

Dataset	Models	Predictors	Performance (worst- to best-performing predictions in any wave)	Reference
LISS	Logistic regression, neural networks, SVM, and tree-based models	Nonresponse history, demographics	0.65–0.79 AUROC	Mulder and Kieruj (2018)
LISS	Logistic regression and gradient-boosted models	Nonresponse history, demographics	0.88–0.89 AUROC	Bach, Eckman, and Daikeler (2020)
GESIS Panel	Logistic regression, tree-based models, and gradient-boosted models	Nonresponse history, demographics, survey evaluation scores, rolling average nonresponse history, rolling average survey evaluation scores	0.86–0.89 AUROC	Kern, Weiß, and Kolb (2021)
Life in Australia	Logistic regression	Demographics, nonresponse history, paradata	0.9 recall, 0.17 precision	Kocar and Biddle (2022)
NEPS	BART, penalized logistic regression	Demographics and education-specific information such as school grades and the number of books at home	95 to 97% accuracy	Zinn and Gnambs (2022)

2018; Wagner, 2008; Zargar, 2021). The details of the differences between traditional and time series models and the type of data they each use are detailed in Appendix Sections 2.6.2 and Data Formatting.

In this study, we focus on the most common types of RNNs (Sarker, 2021), which are the simple RNN, long short-term memory (LSTM), and gated recurrent unit (GRU; DiPietro and Hager (2020), Shewalkar (2018), and Zargar (2021)). GRUs and LSTMs are similar algorithms, and researchers typically evaluate both model types to determine which performs best (Shewalkar, 2018). The benefit of RNN models is that if the interactions between predictors are complex, neural networks can model these complex dependencies. Consider the following hypothetical example: high-income participants may be more likely to respond to a survey. However, the effect is smaller for men than for women, and the effect size is diminished if the participant increased their income only recently (because they recently started a new job, for example). In this case, the effect of one predictor depends on the values of other predictors, including a temporal factor (the difference is made by the timing of the change in income, not the level). RNNs can

account for these types of dependencies.

Simple RNNs are known to perform poorly on long time sequences (Kumar et al., 2018; Shewalkar, 2018). This tendency is because, in some cases, events from the distant past are essential for a prediction. For example, perhaps nonresponse around the previous year's winter holiday season is a good indicator of nonresponse during the coming year's break. However, in other scenarios, events from the past may not be so important in predicting the future. While simple RNNs cannot account for these differences, GRUs and LSTMs are designed to address this issue (Cho et al., 2014; Hochreiter and Schmidhuber, 1997). We test simple RNNs as a comparison point for the LSTMs and GRUs to evaluate the benefits of this more sophisticated handling of long time series. If LSTMs and GRUs perform substantially better than simple RNNs, this would indicate that long-term temporal dependencies are important for making accurate nonresponse predictions. For a detailed description of our architecture for the simple RNN, GRU, and LSTM, see Appendix Section 2.6.2 (Cho et al., 2014; DiPietro and Hager, 2020; Hochreiter and Schmidhuber, 1997; Ribeiro et al., 2020; Rumelhart, Hinton, and Williams, 1986; Shewalkar, 2018; Zargar, 2021).

Time Series Classification Techniques

Time series classification techniques (TSCTs) are diverse techniques for solving time series classification problems. The TSCTs considered in this paper are of the variety that convert long-format time series input (participants by waves by predictors) into a set of statistics describing each participant's time series (participants by time series predictors; see Appendix 2.6.3 for details; see Faouzi (2024) for an overview). These derived predictors are inputted into an ML model, such as a random forest, which classifies the time series. Unlike simple rolling averages, distinct sequences of events are distinguished by differences in the descriptive predictors (Abanda, Mori, and Lozano, 2018; Fawaz et al., 2019; Fulcher, 2017; Lubba et al., 2019).

There are many different approaches to time series classification. This study tests one model type from a variety of techniques as follows.

Feature-Based Methods Feature-based TSCTs transform a time series dataset into a set of summary statistics that describe that time series (Lubba et al., 2019). To test a common and successful example of this approach, we use the 22 canonical time-series characteristics (Catch22) algorithm (Lubba et al., 2019). Catch22 derives 22 descriptive statistics for each time series variable. An example statistic might be the modal value for a scaled and binned variable or the length of the longest period of successive incremental decreases in the time series. These descriptive statistics (22 for each time series variable) are then used as predictors in a classification model. The concept of Catch22 is that these derived predictors describe a time series of any length well enough to have considerable predictive power. Catch22 is effective in many diverse time series classification scenarios (Christ et al., 2018; Fulcher, 2017). However, training times are long for Catch22 classifiers. Therefore, we only test one classification model. We selected a random forest with the same parameters as the best-performing random forest model

in Kern, Weiß, and Kolb (2021). Those parameters are described in Appendix Section 2.6.4.

We expect the feature-based time series approach to outperform the traditional, rolling average–based approach because the 22 derived predictors should both include and go beyond the same information revealed by these rolling average predictors.

Interval-Based Methods Interval-based TSCTs are similar to feature-based techniques, but instead of deriving predictors from the whole time series, they split the series into intervals and derive the predictors from each interval (Middlehurst, Large, and Bagnall, 2020). This study uses diverse representation canonical interval forest (DrCIF). DrCIF separates each time series variable by random intervals, transforms each separated time series with the Catch22 algorithm, and then uses a random forest classifier to make the final predictions (Middlehurst, Large, and Bagnall, 2020). We compare this model with Catch22 to examine the value added by the random interval approach. The interval approach avoids losing information from aggregating the whole time series in one block. Random intervals are preferable over determined intervals because calculating the optimum intervals for the time series by comprehensive search would be computationally implausible. Instead, random searching is a preferable trade-off between accuracy and computation time (Middlehurst, Large, and Bagnall, 2020).

Kernel-Based Methods The intuition behind kernel-based techniques is that the time series of a variable can have a distinctive "shape" that precedes nonresponse (Dempster, Petitjean, and Webb, 2020). For example, a sudden collapse of survey satisfaction after a long period of high values can indicate sudden irritation with the survey and therefore that the participant will nonrespond. In the kernel-based approach, we make 10,000 random time series sequences (called kernels) and then, for each variable, calculate a similarity score (specifically cosine similarity) between each random time series and the observed variable's timelines (see Figure 2.1 for an illustration). For example, a random time series that we would use in conjunction with participant survey satisfaction values would be a sequence of random numbers from one to five for the same length as there are waves in that participant's time series data. The similarity score is then a number, with 1 indicating that the two time series are identical and 0 that they are completely dissimilar. The similarity score between each time series variable and each random time series is then inputted into a classification model.

This analysis uses a particular implementation of the kernel-based approach called random convolutional kernel transform (ROCKET; Dempster, Petitjean, and Webb (2020)). ROCKET generates 10,000 random kernels, calculates similarity scores with each time series variable, and then uses these scores as predictors in a logistic regression. This technique allows the ROCKET model to detect if the "shapes" of certain time series in certain variables correspond to nonresponse behavior (Dempster, Petitjean, and Webb, 2020).

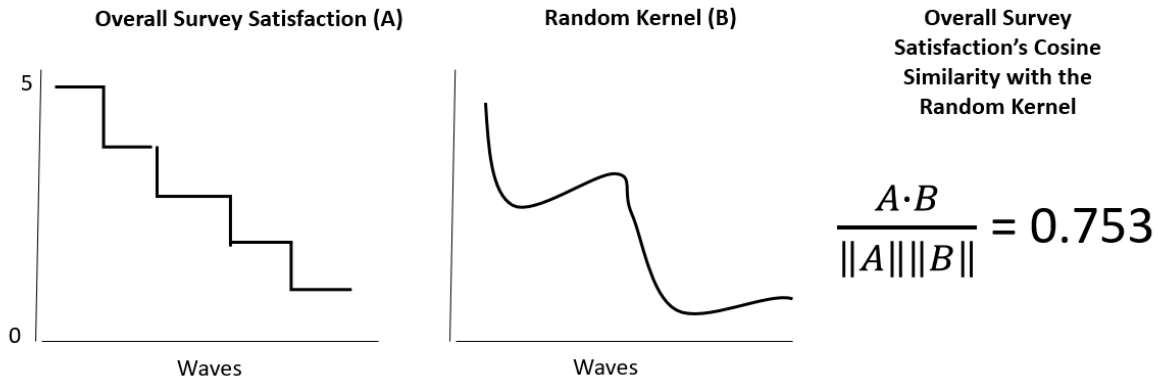


Figure 2.1: *An example of how the ROCKET algorithm derives a similarity score between a time series variable and a randomly generated kernel.*

2.3 Methodology

We aim to answer research questions one and two by comparing traditional ML models to the above-mentioned time series modeling techniques. We train and evaluate models using data from a single, exemplary panel survey: the GESIS Panel (detailed in Section 2.3.1). However, we alter the predictors available in the models to compare their effects. Specifically, we compare the traditional ML models, which use rolling averages as predictors, to time series models without those predictors. Suppose the time series models perform equally well with or without the rolling average variables. This outcome would indicate that the time series algorithms automatically account for the temporal dependencies without needing precalculated rolling averages. We expect the novel time series techniques to make superior predictions over the traditional approach due to this more sophisticated accounting for temporal dependencies.

We seek to answer research question three by examining the effect of sample refreshments. The GESIS Panel recruited an initial sample of roughly 5,000 participants in 2013, which we will call Cohort One. However, the sample diminished over several years, so new participants were recruited in 2016, which we will call Cohort Two. A third refreshment sample was recruited in 2018, which we will call Cohort Three. The effect of these refreshment intakes is that at certain time points, nonresponse prediction models are trained on participants from one cohort and applied to people from a fresh cohort. Therefore, these models would predict nonresponse for participants not present in the training data. It is of practical interest to panel survey practitioners whether models estimated on past panel participants can make useful predictions when examining new potential participants for which past panel behavior is not available. To examine the effect of these refreshment intakes, we fit each model using data from each of the three cohorts separately and again all together (see Appendix Section 2.6.4). We then calculate the loss in predictive performance attributable to the introduction of fresh participants as the difference in AUROC scores when predicting nonresponse of fresh

participants in comparison to pre-existing participants.

2.3.1 Data

Our data source is the GESIS Panel up to May 2021 (GESIS, 2023). The GESIS Panel, managed by the German Leibniz Institute for the Social Sciences, is a longitudinal survey that commenced in October 2013 and continues today (GESIS, 2023). The survey has two modes: web and mail. Since 2013, roughly 75% of participants have used the web-based option, and the rest have submitted their responses by mail (Bosnjak et al., 2018; GESIS, 2023).

The GESIS Panel randomly sampled participants from the German population register. Anyone permanently residing in Germany and aged between 18-70 was eligible for recruitment. For full details of the sampling methodology see (GESIS, 2023).

Variables

Thousands of variables are collected across all GESIS Panel waves. We select variables that align with previous research: demographics, indicators of survey engagement, and paradata. We also derive rolling average predictors. A summary of these variables is provided in Table 2.2 and a full description of all variables in Appendix Table 2.5.

The dependent variable is each participant's binary outcome for nonresponse in the next wave (1 = nonresponse). We follow the American Association for Public Opinion Research (AAPOR) response rate formula RR6, which counts completed and partially completed submissions as responses. Appendix Table 2.8 indicates the categories that we consider nonresponses, including the AAPOR response codes (AAPOR, 2016). Over all waves included in this analysis, the lowest nonresponse rate (1 – RR6) was 6.27% and the highest was 24.96% (see Figure 2.2). A critical matter about the GESIS Panel is the procedure for eliminating participants from the panel: participants cease to be invited to the survey if they either explicitly ask to exit the panel or nonrespond for three consecutive waves. In either scenario, the participant is sent an exit questionnaire, at the end of which they can voluntarily re-enter the panel (Bosnjak et al., 2018). This policy means a given participant can nonrespond in several waves without exiting the panel. In our analysis, we filter out participants who have exited the panel. This means that if the data for a participant at a given wave is included in either the training waves, or the validation wave, then as of that point in time the participant had not yet made three consecutive nonresponses (or they had, but explicitly asked to re-enter the panel in their exit survey) or had not explicitly asked to exit.

Figure 2.2 presents the nonresponse rates (defined as the portion of invitations that yielded a nonresponse) and sample sizes of the GESIS Panel (defined as the count of invitees) by cohort over our study period. After each cohort recruitment, the nonresponse rates spiked before reaching a relatively stable level. There was a spike in nonresponse rates in early 2020, likely attributable to the coronavirus disease 2019 (COVID-19) pandemic in Europe. Also, as each new cohort was commencing, for the first few waves the GESIS Panel managers invited participants as they were recruited, rather than wait un-

til the whole cohort was recruited. The stated reason for this policy was to avoid leaving early-joining participants without any contact for a long time while cohort recruitment was finalized. The outcome is that the sample size for each cohort is relatively small in those early waves (Bosnjak et al., 2018).

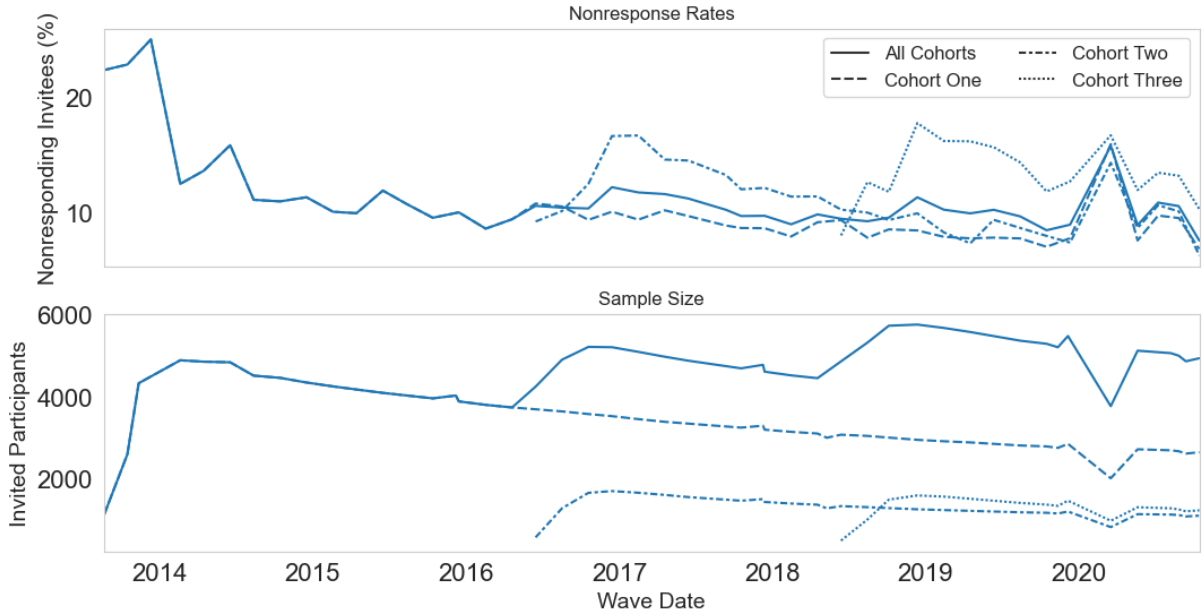


Figure 2.2: *Timeline of nonresponse rates and sample sizes of the GESIS Panel.*

In this analysis, we train our models with only time-variant predictors, only time-invariant predictors, and again with all predictors. This comparison will reveal whether time series models make their forecasts by accounting for temporal dependencies in the time-variant data or just use time-invariant predictors or rolling average predictors like the baseline models. In that case, the time series techniques do not add value over the traditional techniques.

Table 2.2 provides an overview of the different predictor sets. Time-variant predictors include nonresponse or participation mode in a given wave. Time-invariant predictors include demographic characteristics, because they are updated only periodically. Although rolling average predictors are a type of time-variant predictor, we exclude them from that category because we want to compare traditional models with rolling average predictors to time series models without them. See Appendix Table 2.5 for the complete list of predictor variables.

Note that we commence our analysis from the third survey wave onwards to allow the "rolling average over the last three waves" predictors to register valid values.

Table 2.2: *Categorization of each predictor into sets.*

Predictor Set	Type of Predictor	Examples
Time-Variant Predictors	Indicators of survey engagement	Respondents are asked (translated from German into English), "How was the questionnaire?" The scale items are "Interesting, Diverse, Important for Science, Long, Difficult, Too Personal, Overall."
	Paradata	These predictors include whether the survey mode is online or by mail and whether there are any detected survey breaks in the online version.
	Nonresponse in the current wave	-
Time-Invariant Predictors	Demographics	These predictors include age, income, education, and gender.
	Recruitment interviewer assessments	At the recruitment wave, interviewers gave a one-to-five scale rating of the participants' prior experience with surveys and how cooperative they were when interviewed.
Aggregate Predictors	Rolling average predictors	For each survey engagement item, we derive rolling averages for each value over the previous two, three, and all waves. We also derive the rolling average nonresponse rates for each value over the previous two, three, and all waves.

2.3.2 Validation

We want to know how well each model performs when applied to the GESIS Panel. For this objective, we want to simulate what would have been the outcome at each wave had our models been trained only on data available at that time. To accomplish this, we use temporal cross-validation (TCV; see Figure 2.3), where we predict for each panel wave which participants nonrespond, using only data available in all preceding waves (Hyndman and Athanasopoulos, 2021). This procedure means that at some point, each wave is both the wave being predicted (the "test" data) and a wave used to train the predictive model (the "train" data).

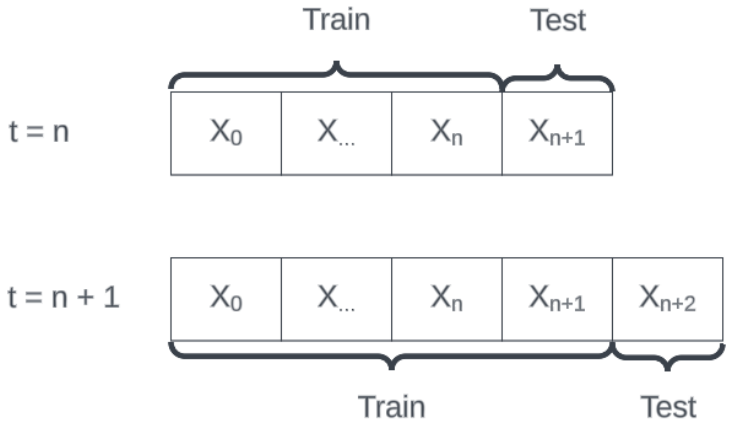


Figure 2.3: *Diagram of temporal cross-validation. For each wave (X), we aim to predict a "test" wave using all waves that precede it (the "train" waves).*

An important critique of TCV used this way, raised by Jankowsky, Steger, and Schroeders (2022), is that in the panel, the same participant is most often present in the training data and in the test data. Furthermore, Kern, Weiß, and Kolb (2021), who also used TCV, found that prior nonresponse status strongly predicted future nonresponse. This finding means that, to some extent, predictive performance is inflated because the models can adapt to the data of the nonrespondents, who go on to repeatedly nonrespond, and also to the respondents who continue to respond. This is of particular concern given that, across all the GESIS Panel waves included in this study, almost half of all cases of nonresponse are cases in which the participant also nonresponded in the preceding wave (12,296 out of 23,456 cases of nonresponse). We use temporal cross-validation because it accurately simulates how GESIS Panel managers would utilize predictive analysis in practice. Furthermore, we also assess predictive performance when applying our models to newly recruited participants, providing a robust assessment of model generalizability. Still, we caveat our results with the possibility that our techniques may yield a lower AUROC when applied to different panels.

2.3.3 Model Types and Tuning

We compare many different models, and each time we train a model, we vary "model settings," i.e., model parameters that affect the fit to the data. This process, called model tuning, determines the best settings for each model by experimentation. Table 2.3 describes our models and the relevant settings we experiment with. The full details of the settings we tune are described in Appendix Section 2.6.4.

2.3.4 Predictor Importances

Although many variables are used as predictors in our models, some variables may contribute more to the quality of the predictions than others. To answer research question one, we want to evaluate whether the RNNs and TSCTs account for temporal dependencies in the data and do not simply rely on the time-invariant predictors. To evaluate this, we examine whether the time-variant predictors are improving predictions from the time series models to a greater extent than the traditional models.

To estimate the extent to which the variables most impact the predictions, we use the permutation feature importance metric (PFI; Altmann et al. (2010)). The typical process for PFI involves taking a certain predictor, shuffling the values randomly, and then assessing the model with that single predictor scrambled. This process is repeated with different random shuffles. The average loss (or gain) in predictive performance (in our case, AUROC) measures how much that predictor contributed to the prediction (James et al., 2013).

To adjust this procedure for a time series context, instead of shuffling a given variable in the test data, we replace it with random values drawn from a normal distribution with the same mean, standard deviation, minimum, and maximum values as that variable. Binary predictors are replaced with random binary values with approximately the same mean value as the corresponding predictor.

Furthermore, the typical PFI procedure will be confounded when many predictors covary. For example, if a participant nonresponds in a given wave, not only will the variable for nonresponse show this, but also their survey satisfaction scores will be zero (i.e., missing, see Appendix Section Data Formatting). This issue means the same information is present across several predictors. If one covariate is withheld from the model, the same information will be available through another predictor, leading to an underestimation of the PFI. We therefore group predictors into blocks and scramble the entire block so that the underlying information is not available through any other predictor. However, it is only possible to eliminate covariation partially. For example, we want to compare the importance of nonresponse history and survey satisfaction scores, so we must keep them as separate groups although they can covary as described above. We detail the groups in Appendix Table 2.6.

Calculating PFI requires long computation time, so we limit our analysis to one example test wave. We select the wave of August 2019 because it preceded the COVID-19 pandemic waves but was at a time when all cohorts had accumulated several waves. By choosing this wave, we ensure that the models' performances are not inhibited by a lack

Table 2.3: *Categorization and description of each model type included in this study.*

Category	Model Type	Description	Settings
Traditional ML models	Baseline logistic regression (James et al., 2013)	This model is a standard logistic regression used as a baseline.	-
	Penalized logistic regression (Le Cessie and Van Houwelingen, 1992; Tibshirani, 1996)	These models are ridge and lasso penalized logistic regressions.	We vary the type and intensity of the penalty.
	Random forest (Breiman, 2001; James et al., 2013)	This model is a popular tree-based ensemble method. This was the best-performing model in Kern, Weiß, and Kolb (2021), who used the same data as this paper.	To save computation time, we use the best discovered model parameters from Kern, Weiß, and Kolb (2021) and do not tune this model in our case.
RNNs	Simple RNN (Rumelhart, Hinton, and Williams, 1986)	This model is a standard RNN as described in Section 2.2.1.	In each of these RNN models, we will vary the "width," which is the number of neurons in each layer, and the "depth," which is the number of layers in addition to the single recurrent layer we always add. Deeper and wider RNNs can model more complex interactions (see Appendix Section 2.6.2 for details; (Salman et al., 2018)).
	LSTM (Hochreiter and Schmidhuber, 1997)	This model is a type of RNN with functionality for weighting the effects of information from the far or recent past. See Appendix Section 2.6.2 for details.	
	GRU (Cho et al., 2014)	This model is similar to an LSTM, but it is implemented differently. For comprehensiveness, it is common practice to evaluate both GRU and LSTM (Shewalkar 2018; Zargar 2021). See Appendix Section 2.6.2 for details.	
TSCT	Catch 22 (Lubba et al., 2019)	This model derives 22 descriptive statistics for a given time series and then uses that information as predictors in a random forest.	-
	DrCIF (Middlehurst, Large, and Bagnall, 2020)	This model is similar to Catch 22 but derives the descriptive statistics for each of a set of random time intervals.	-
	ROCKET (Dempster, Petitjean, and Webb, 2020)	This model implements a technique involving randomly generated kernels. See Section Data Formatting for details.	-

of training data or by the exceptional circumstances of the pandemic. We repeat each feature block's random shuffling four times and present the average loss of AUROC as the PFI value.

2.4 Results

2.4.1 Prediction Performance

Figures 2.4, 2.5, 2.6, and 2.7 present our main findings. The lines in Figure 2.4 depict the AUROC score of each model over the timeline of the GESIS Panel waves. Figure 2.4 features a line for each model per cohort disaggregation. The lines are the median scores across the various model settings and the shaded areas represent the range of best and worst scores. The vertical lines indicate the waves in which cohorts two and three entered the panel. The predictions for the cohort-specific model commence from the wave after each of those cohorts enters the panel, so there is at least one training wave.

In the all predictors set, the models with all cohorts combined each perform within a relatively narrow AUROC range (maximum variance is GRU at 0.68-0.88; see Figure 2.4). The random forest models achieved an average of 0.857 AUROC over all waves, which is roughly as successful as in previous studies (0.79-0.88 AUROC, see Table 2.1). The LSTM and GRU only marginally improved on this score if at all. This narrow band of scores may have several explanations: First, the models may quickly accumulate enough training data to make accurate predictions, and additional data points may only marginally improve performance. Second, after the first few waves of a new cohort, low-propensity participants are filtered out, leaving only consistent respondents and occasional nonrespondents, which the models can easily predict from their historic nonresponse rates. The only instances of sudden drops in performance occurs when the set of waves considered include those during the COVID-19 pandemic. Notably, the random forest model with all predictors drops in performance the least around the pandemic.

Regarding research questions one and two, Figure 2.4 shows that the RNN-based models outperform random forest and logistic regression using only time-variant predictors. GRU and LSTM perform highly at 0.8 to 0.9 AUROC (excluding COVID-19 waves). Meanwhile, random forest and logistic regression perform more poorly with only time-variant data, with an AUROC between 0.7 and 0.8. With only time-invariant predictors available, most models perform almost equally poorly with an AUROC around 0.65, except for random forest with scores between 0.65 and 0.8. When all predictors are available and all cohort groups are included, RNNs and random forest perform roughly equally well, with random forest performing slightly better, with an AUROC between 0.8 and 0.9.

The TSCTs are present only in the last 12 waves (since December 2018) due to the much higher computation required to fit these models. However, we can see that they never outperform either of the other model types, which justifies excluding them from further investigation given their long processing times.

Regarding research question three, in Figure 2.5 we provide a filtered version of Figure

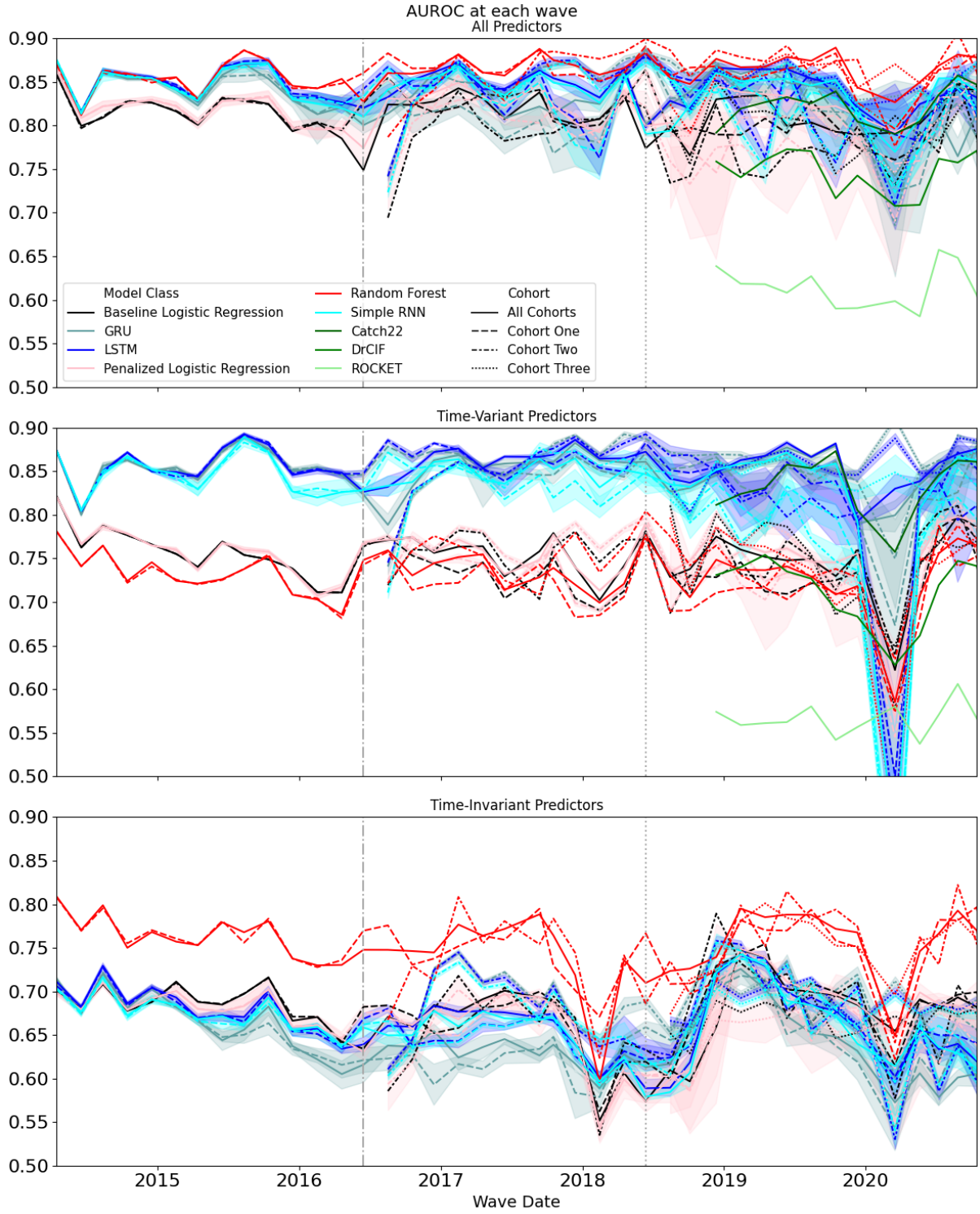


Figure 2.4: *Timeline of AUROC scores at each wave across the different predictor groups. Vertical lines indicate the waves in which cohorts two and three entered the panel.*

2.4 that only shows the two best-performing models (random forest and LSTM) to see the effect of refreshment intakes more clearly. The vertical lines indicate the waves in which cohorts two and three entered the panel. The individual cohort predictions commence from the wave after each of those cohorts enters the panel, so there is at least one training wave. For example, in 2016, cohort two was recruited. So, the model predicting the first wave of cohort two was trained on the roughly 5,000 participants from cohort one (of which only 4,000 remained by that time) but applied to a test set of roughly 4,500 participants (the 4,000 from cohort one and the first 500 fresh participants from cohort two). We can see that when a new cohort is recruited (indicated by the vertical lines in Figure 2.5) and all predictors are available in the models, the "all cohorts" line falls below the line representing the models trained only on a single cohort (excepting the time-variant random forest). This observation indicates that performance drops when a model is trained on participants from one cohort but applied to participants from a fresh cohort. However, the fall is less than 0.03 AUROC, except for LSTM when cohort three commenced which was roughly 0.07 AUROC. To illustrate the implications of a 0.03 reduction in AUROC, we elaborate upon the example of the random forest model at the wave when cohort two commenced. In this example, we classify the participants with the highest 10% nonresponse propensity as likely nonrespondents. The model would then result in a 39.1% instead of 37.5% false positive rate and a 63.2% instead of 61.7% false negative rate when forecasting nonresponse of all participants instead of just cohort one.

To investigate this drop further, we evaluate the predictions for the roughly 500 respondents from cohort two and compare them against the forecasts made for the participants from cohort one who were present in the training data. We repeat this process for cohort three (see Figure 2.6). To simplify the role of model settings, we select settings for each model that yielded the highest AUROC over all waves up until the commencement of the respective new cohort. When the model predicts participants from cohorts two and three who have just entered the panel (the waves indicated by the vertical lines in Figure 2.5), no model is consistently better at predicting fresh participants as opposed to pre-existing participants, with some caveats as follows. On the commencement of cohort two the simple RNN, penalized and unpenalized logistic regression models each performed equally or even slightly better when predicting nonresponse of fresh participants. However, at the commencement of cohort three, every model performed worse when forecasting nonresponse of new participants. The simple RNN (time-variant predictors) and penalized logistic regression (all predictors) provide the best nonresponse predictions for fresh participants in both waves. The implication is that fresh participants and those who have stayed in the panel for several waves behave differently enough that different models are better suited to each class. This result is of interest to panel survey managers because it indicates that different models could be employed for fresh and pre-existing participants.

Figure 2.7 shows the overall average AUROC for each model's best setting across test waves by predictor groups. Note that we only average the performance before August 2019 so as to avoid the effect of the COVID-19 pandemic on the averages. The RNNs perform almost equally well when they use only the time-variant predictors as when they use all predictors. However, random forest performs much worse when only time-variant

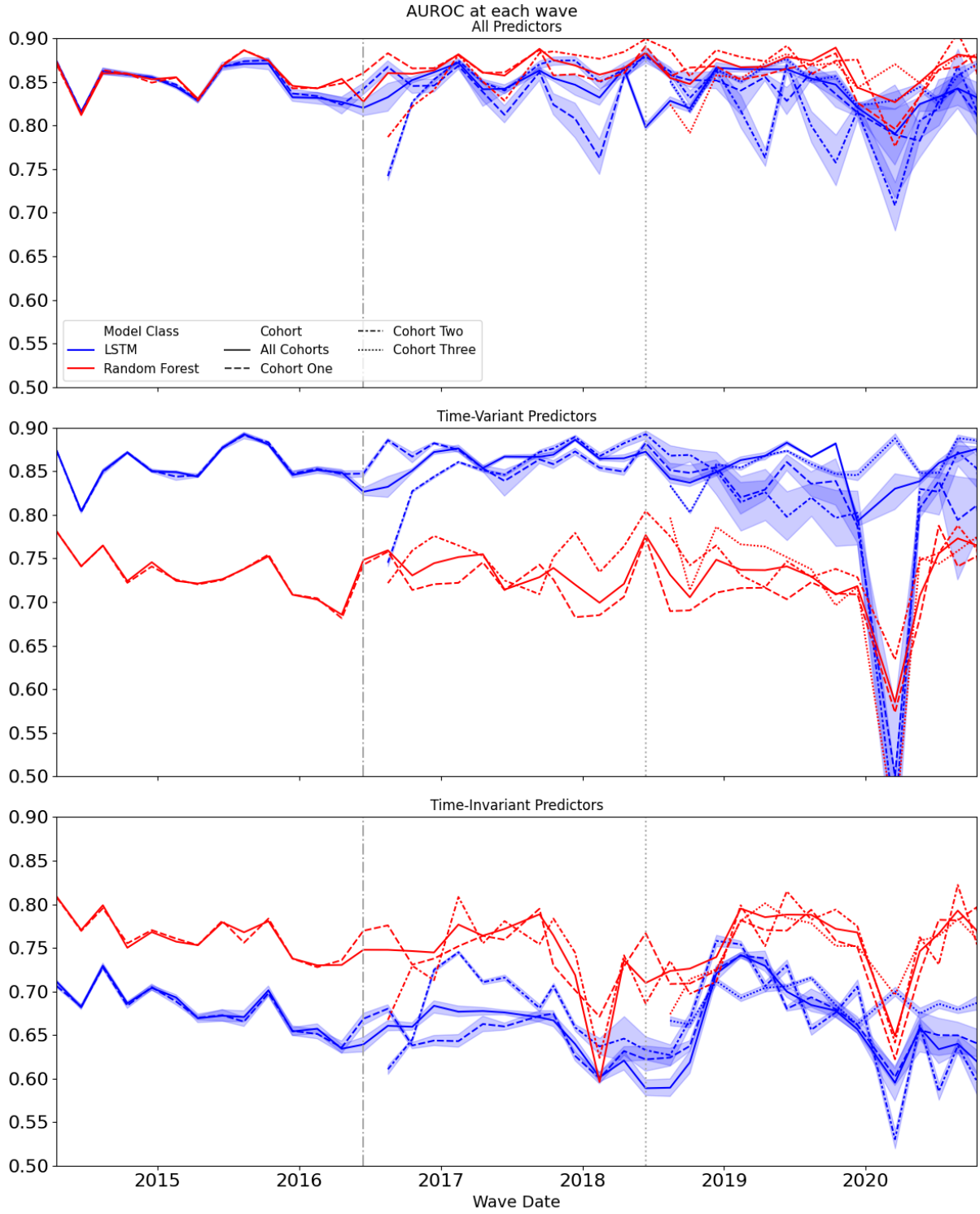


Figure 2.5: *Timeline of AUROC scores of random forest and LSTM models, which highlight the effect of refreshment intakes. Vertical lines indicate the waves in which cohorts two and three entered the panel.*

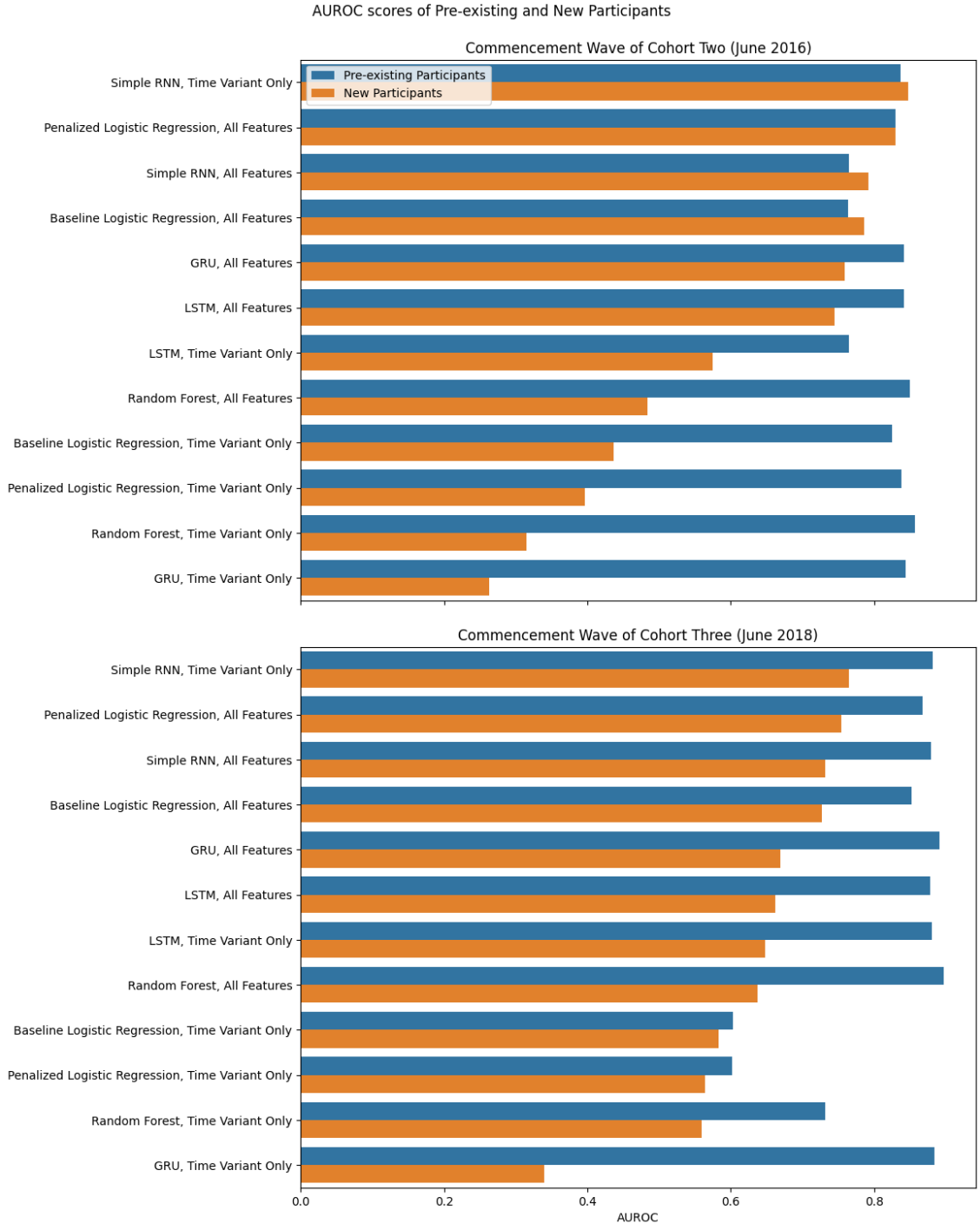


Figure 2.6: *AUROC scores of each model when forecasting the next-wave-nonresponse of newly recruited participants compared to participants who were already in the panel.*

predictors are available. All models perform more poorly when given only time-invariant predictors. This observation indicates that the RNNs effectively engineer the same information as the rolling average predictors within the model's neural network instead of relying on the practitioner to engineer those predictors manually. This conclusion is further supported when we examine the PFIs below. ROCKET is an outlier in poor performance, which indicates that the time series data does not have any indicative "shape" that the kernel-based approach can exploit. The poor performance of models with only time-invariant predictors indicates the critical improvement made by incorporating indicators of the participant's recent behavior instead of only the information collected at the recruitment interview.

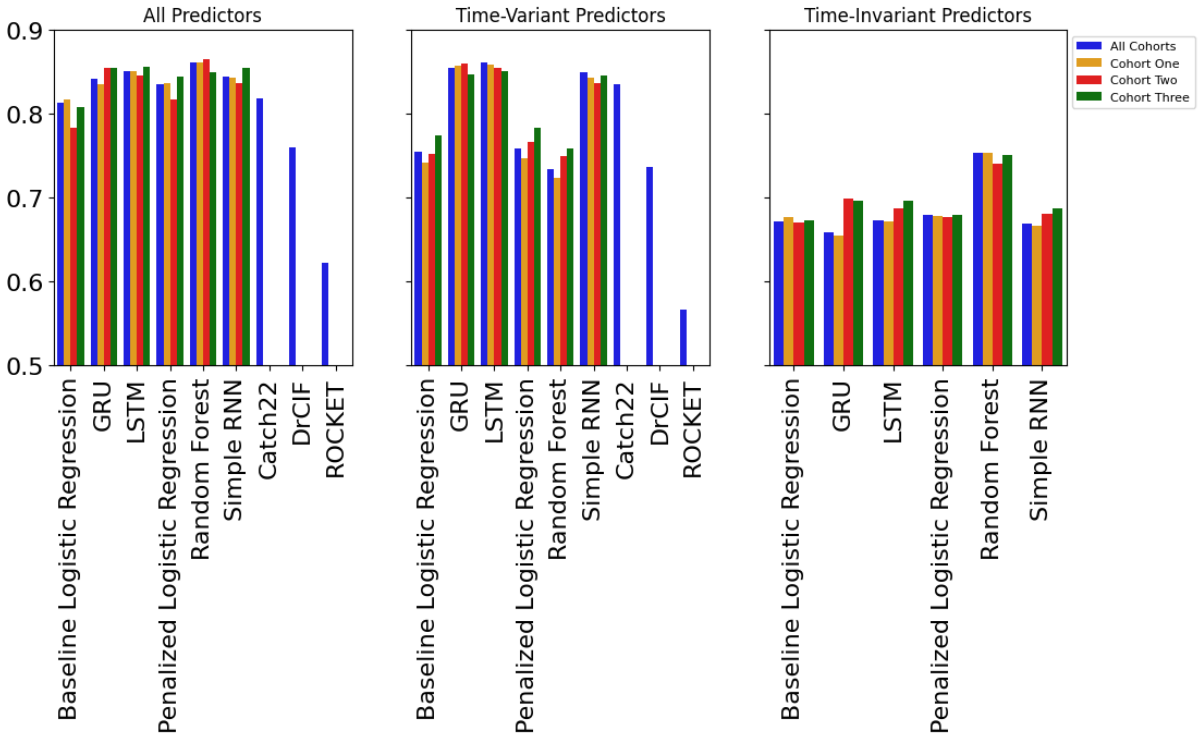


Figure 2.7: *Overall mean AUROC scores (up to August 2019) for each model's best setting.*

Our best-performing model results are comparable to those from previous studies. The random forest with all predictors and the LSTM with only time-variant data are the best-performing models in this study, with an average AUROC over the 40 waves of 0.857 and 0.856 respectively (see Appendix Table 2.12). The best AUROC scores among previous studies are between 0.79 and 0.89 (see Table 2.5).

2.4.2 Permutation Feature Importance

Figure 2.8 shows the average permutation feature importances for each block of predictors (see Appendix Table 2.6). Note that rolling averages are not included in the "time-variant predictors" set. We aim to test how each model performs when these predictors are withheld and the model receives only the unit-level variables instead (i.e., survey satisfaction in a single wave instead of the rolling average). When all predictors are available in the models, the most important predictors for GRU and LSTM models are those that indicate nonresponse in a current wave, while rolling averages of nonresponse history are much less important. Traditional models (logistic regression and random forest), by contrast, rely upon a mixture of rolling average predictors and the equivalent wave-specific predictors whenever they are available.

These observations answer research question one: RNNs can adaptively model temporal dependencies in the data, so RNNs do not strongly benefit from including the rolling average predictors. However, the comparison of RNNs and traditional models indicates that the rolling averages accurately reflect the temporal dependency in the GESIS Panel. A participant's historic response rate is a good indicator of future nonresponse, and the rolling averages are sufficient to provide this information.

Although the RNNs can accomplish the same task automatically, the rolling averages are sufficient to achieve the same result. Therefore, the answer to research question two is that LSTMs can equal the performance of random forest models with rolling average predictors but not exceed their performance in the GESIS Panel (0.856 and 0.857 average AUROC respectively). The best penalized logistic regression model achieved an average AUROC score of 0.82, so LSTM and GRU did outperform these baseline techniques.

2.5 Discussion

In this paper, we proposed the use of time series machine learning techniques to predict panel nonresponse. The novel techniques we applied were recurrent neural networks (simple recurrent neural network, long-short term memory, and gated recurrent unit) and time series classification techniques (feature-, interval-, and kernel-based methods).

We highlight our main results in Table 2.4. In summary, these novel techniques do not necessarily outperform baseline models, but some can automatically model temporal dependencies. Our random forest models, which achieved an average of 0.857 AUROC over all waves, were roughly as successful as in previous studies (0.79-0.88 AUROC, see Table 2.1), but the time series models only marginally improved on this score if at all. However, consider the case in which a modeler has applied rolling averages and achieved a certain level of predictive performance. How would they determine whether a more sophisticated accounting for temporal dependencies would improve the predictions? Our paper demonstrates that RNNs can adapt to temporal dependencies in nonresponse prediction, so the modeler can use RNNs to check whether their current rolling averages are sufficient or whether more sophisticated techniques would yield better results. RNNs are not prohibitively longer to compute than traditional models. When we calculate the average computation time of each model type, we find that average logistic regression

Figure 2.8: *Block-wise permutational feature importances across each predictor set.*

Table 2.4: *A summary of the research questions, the analyses conducted to answer them, and the main results.*

Research Question	Analytical Strategy	Model Combinations	Results
1. Can time series machine learning techniques account for the sequence of events in panel data instead of rolling average predictors?	We compare traditional ML models with rolling average predictors to RNNs and TSCTs without them.	4,512: 37 uniquely parameterized models over 40 waves and three predictor sets, plus three TSCTs over 12 waves and two predictor sets.	RNNs can automatically account for temporal dependencies because they achieve equivalent performance to random forest (and exceed logistic regression) without the need for rolling average predictors.
2. Do these techniques outperform traditional models that use rolling average predictors?	We determine which of the traditional models or the time series models make better predictions.	1,516: 10 uniquely parameterized traditional models versus 27 uniquely parameterized RNNs over 40 waves, plus three TSCTs over 12 waves using all available predictors.	LSTM and GRU performed equally as well as random forest with rolling average predictors. Both RNNs outperform logistic regression with rolling average predictors.
3. When a machine learning model is trained on panel waves that predate the refreshment intake, can the model make accurate predictions about fresh participants, who were not present in the training data?	We compare predictions about newly recruited participants with predictions about participants who were present in the training data.	13,320: 37 uniquely parameterized models across three predictor sets. We fitted these models to over four cohort groupings, although different cohorts had different numbers of waves.	The simple RNN and penalized logistic regression models were best suited to predicting nonresponse of fresh participants, while random forest and LSTM were better for pre-existing participants.

and random forest take around ten seconds to fit, and RNNs take around one minute to fit. TSCTs take between 10 and 20 minutes to fit. Since the use of RNNs for nonresponse prediction is a key contribution of this paper, we provide a detailed demonstration of how they are implemented in the replication material (see Appendix Section 2.6.1). Finally, we tested the effect of sample refreshment on predictive performance. We found that although random forest and LSTM were the best models for forecasting nonresponse for the whole panel, Simple RNN (with time-variant predictors) and penalized logistic regression (with all predictors) were the best at predicting future nonresponse for fresh participants.

The primary limitation of this research is that we only applied RNNs to a single panel study. In the case of the GESIS Panel, RNNs were found to equal the performance of rolling average approaches. There may be other contexts in which RNNs do better or worse than these traditional approaches. The GESIS Panel fields a survey wave every two or three months, so the predictive power of participant behavior in preceding waves

may differ greatly from that of a survey with annual or weekly waves. Also, the GESIS Panel is a general population survey that collects ample demographic data during the recruitment phase and periodically updates that information. In contrast, other surveys target subpopulations, record demographic data more frequently, or perhaps collect fewer demographic data. For these reasons, temporal dependencies are probably very different between panels. Nonresponse in a preceding wave likely signifies very different events for individual panelists in a three-monthly panel survey compared to an annual household survey. In specific cases, such as panels with more frequent waves, RNNs may be able to exploit subtle temporal dependencies better than traditional approaches.

Another consequence of our chosen validation method is that we test all model setting combinations for the entire timeline and present those results. This process could inflate the apparent predictive performance because, in a real prediction scenario, modelers would have had to select a specific set of parameters for their model and then use it without knowing in advance whether other parameters might have performed better. However, when examining the average AUROC scores across model setting combinations, we find that the choice of settings had little impact (see Appendix Table 2.12). Regardless of the settings, all LSTM and GRU models achieved similar overall AUROC scores, and random forest had only one setting in our case. For this reason, little would have changed if we had altered how our model tunings were handled.

Another limitation of this study is that our time series algorithms are not exhaustive. For example, we could construct a "wide-format" dataset such that the wave values of different variables are in separate columns and explore models that account for temporal dependencies this way (Suresh, Severn, and Ghosh, 2022). There are also more RNN architectures to explore and other implementations of the feature-, interval-, and kernel-based techniques (Shewalkar, 2018). This paper covered examples of a broad range of very different and novel approaches to time series classification, so that the most promising techniques could be explored more deeply in later research. However, the range of techniques we have explored is still not exhaustive.

Another limitation of this study is that RNNs can be more difficult to interpret than logistic regression or random forest. Unlike logistic regression models, which have interpretable predictor coefficients, RNNs do not assign a single effect magnitude to a given predictor. When trying to understand what factors affect participant nonresponse, it can be difficult to derive these relationships from RNN models. Random forests also do not assign a single effect to each predictor, but instead consist of a set of decision trees, which consist of a sequence of simple logical tests, which may be easier to understand than the many sets of weights across an RNN's neurons. It is worth noting that penalized logistic regression, while not the best performer, still yielded high AUROC scores. It was also one of the best models for predicting nonresponse of fresh participants. Therefore, although less predictive, penalized logistic regression may offer practitioners more value than RNNs and tree-based models in specific settings because of its convenient interpretability. We nonetheless conclude that novel time series classification techniques such as RNNs are worth considering when achieving high predictive accuracy in complex panel settings is the main objective.

2.6 Appendices

2.6.1 Replication

This project can be replicated by obtaining the dataset, downloading our code, and placing the data file into the code's project directory. You will need to contact GESIS to request access to GESIS Panel data:

> GESIS (2023), "GESIS Panel - Standard Edition." Published: GESIS, Cologne. ZA5665 Data file Version 44.0.0, https://doi.org/10.4232/1.13931 DOI: 10.4232/1.13931

All code for replication is available at the following link:
`https://osf.io/kngdj/?view_only=79dc3aa2d0f947a18a5f95d4ed97c0a0`
The demonstration of an RNN implementation is available at the file:

src/RNN_Demonstration.ipynb

The data from GESIS may be downloaded as a .zip file. Simply place the .zip file in the directory '\data\sensitive_GESIS_raw\' and then follow the instructions in the README.md file. Note that if you use the same hardware as we specify in the Section 'Hardware, software and Computational Resources', it may take up to three weeks to run the full project.

Hardware, Software and Computational Resources
The following describes the specifications of the computing resources used in this project.

- OS: Windows Version 10.0.17763, Build 17763
- Processor: Intel(R) Core (TM) i5-10310U CPU @ 1.70GHz, 2208 MHz, 4 Core(s), 8 Logical
- Installed Physical Memory (RAM): 32 GB
- Total computation time: Up to 21 days to compute all fittings, including the feature importances.

2.6.2 Model Details

The purpose of this Section is to provide the details, including the equations, of the RNN architectures used in this paper. For a general overview of RNNs and neural networks see Shewalkar (2018).

Firstly, all neural networks, including RNNs are composed of networks of nodes and edges (see Appendix Figure 2.9). Each node (also called a 'neuron') represents a function, which shall be detailed below. Each edge represents a connection whereby the

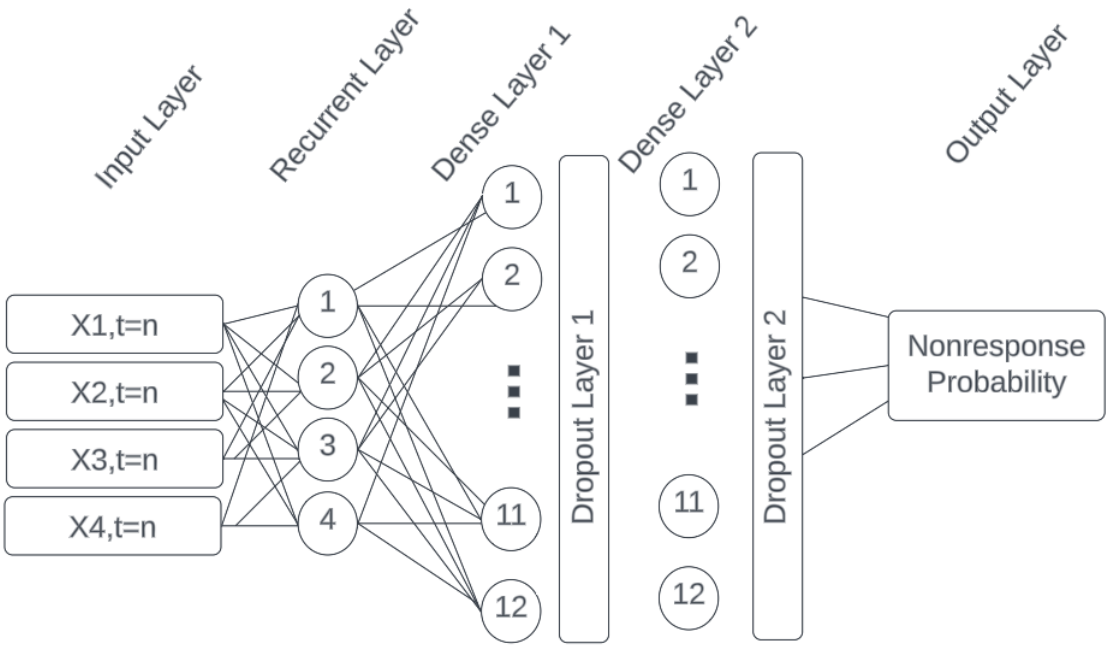


Figure 2.9: *Diagram of example RNN architecture.*

output of one node becomes the input to another node. The following process takes data about participants in a panel survey as input (the precise predictors are detailed in Section 2.3.1) and outputs the estimated probability that the given participants will nonrespond in the next wave.

Simple RNN

1) Firstly, let us describe the input for the RNN. This input is a three-dimensional data format: We have a set of participants, and for each participant, we have several waves (although different participants may have a different number of waves) and in each wave we have a constant set of predictors (i.e., a true/false value for nonresponse in that wave, their age at the time of that wave, etc).

2) The 'input layer' represents the time series predictors which shall be inputted into the 'recurrent layer,' starting from the first wave (t=0). Therefore $x_{1,t=n}$ indicates the value of predictor x_1 (i.e., age at the time of the wave) for wave n. For each wave, the value of each predictor at that wave is inputted into every node in the 'recurrent layer.' Then, for each recurrent neuron, we calculate its 'hidden state value' for that time step. Each neuron retains its hidden state value, then the values from the next wave are inputted into the recurrent neurons, and these values, as well as the hidden state values from the previous time step, are used to

calculate the updated hidden state value for this new wave. This process repeats until all time steps have been processed and the output of each recurrent neuron is the last hidden state value.

The equation for the hidden state value in a recurrent neuron is:

$$\mathrm{h}_{t=n} = \alpha(W_{hx} \bullet X_{t=n} + W_{hh} \bullet h_{t=n-1} + \beta_h) \tag{2.1}$$

Where X is the array of input variables, W_{hx} is a weight array fitted to be multiplied by X, $W_{hh}$is a weight array to be multiplied by the hidden state value from the previous time step (which is 0 as of the first wave) and β_h is a bias number with a value determined by the fitting process. α is an activation function, for example a sigmoid, tanh, or ReLU function (Shewalkar, 2018). The details of these activation functions are not important here, but their purpose is to scale the function's output to an output within a certain range, such as zero to one.

3) The output of the recurrent layer (X) is passed to the neurons in the first 'dense layer.' These neurons do not implement recurrence so the equation for the output of each neuron in that layer is:

$$\mathrm{h} = \alpha(W_x \bullet X + \beta_h) \tag{2.2}$$

4) Each value from each neuron in the first dense layer then goes through a 'dropout layer,' which is an array of zeros and ones, the composition of which is determined by a fitting process. The purpose of the dropout layer is to select specific neurons from the previous layer to convert their output to zero. This is intended to avoid overfitting. Overfitting is where an ML model is fitted to perform very well on data it was trained on but fails to make good predictions on new data (Shewalkar, 2018). The function for the dropout layer is as follows:

$$\mathrm{L}_d = L \odot D \tag{2.3}$$

Where D is the dropout layer (a binary array), and L is the array of the outputs of the preceding layer of neurons.

5) We experiment with different numbers of neurons and layers in our architecture (see Section 2.3.3). If there is more than one dense layer in the network, we repeat steps three and four for each additional layer.

6) The final dense-and-dropout layer then sends its output values to the last layer in the neural network, which is the 'output layer,' which consists of only one neuron. That neuron outputs the final estimated probability of nonresponse in the next wave as a number between zero and one. Note that σ is a Sigmoid function. which outputs a value close to zero when its input is significantly negative and a value close to one when its input is significantly positive. i is the number of neurons in the layer that precedes the output layer.

$$y = \sigma(\sum_{i=1}^{i} [W_i \bullet x_i + \beta_h]) \tag{2.4}$$

Gated Recurrent Unit (GRU)

The process for the GRU is similar to the simple RNN. However, the simple RNN has only a single value for W_{hh} in each recurrent neuron, which means each previous event is attributed the same importance, regardless of whether it occurred far into the past or not. GRU implements a process to learn how to differently weight events in the distant or proximate past.

1) In a GRU, the hidden state values of each neuron in the recurrent layer are calculated with a different process to that of the simple RNN. Before we can calculate the hidden state value, we must first calculate some component values. These are called the 'reset gate,' 'update gate,' and 'candidate hidden layer' values, which are described as follows. Note that many of these functions involve a term for the hidden state value from the previous wave ($h_{t=n-1}$) and that for the first wave (t=0), the value of that term is zero.

2) For each recurrent node, at each time step, we calculate a value called the 'reset gate' ($r_{t=n}$). The reset gate value will determine how much influence the values from previous timesteps should have on the output of this node, which in turn will affect the ultimate output of the GRU model. The formula for the reset gate is:

$$\mathrm{r}_{t=n} = \sigma\left(w_{xr} \bullet x_{t=n} + u_r \bullet h_{t=n-1} + \beta_r\right) \tag{2.5}$$

3) Where $x_{t=n}$ is the array of predictors from the participant at wave n. w_{xr} and u_r are both arrays of weights derived from a fitting process.

4) We calculate the 'candidate hidden state' value ($\widetilde{h}$), which is the maximum possible value of the hidden state value that will be retained by this node at this time step. The formula is:

$$\mathrm{h}_{t=n}^{\sim} = \tanh(w_{xh^{\sim}} \bullet x_{t=n} + w_{h\tilde{h}}(r_{t=n} \odot h_{t=n-1}) + \beta_{\tilde{h}}) \tag{2.6}$$

5) We calculate the 'update gate' value (z). The update gate value determines how much of the value from the current time step should be passed onto the new hidden state value.

$$\mathrm{z}_{t=n} = \sigma\left(w_{xz} \bullet x_{t=n} + u_z \bullet h_{t=n-1} + \beta_z\right) \tag{2.7}$$

6) Finally, we calculate the hidden state value. The value of the update gate will be used to reduce (or leave unchanged if the value is 0) the value of the candidate hidden state with the following formula.

$$\mathrm{h}_{t=n} = z_{t=n} \odot h_{t=n-1} + (1 - z_{t=n}) \odot h_{t=n}^{\sim} \tag{2.8}$$

7) Once the very last hidden state value is calculated in the recurrent layer, the output of the recurrent layer is sent to the first dense layer. The remainder of the process for the GRU is the same as the simple RNN (steps 3 – 6 in Appendix Section 2.6.2 above).

Long Short-Term Memory (LSTM)

The LSTM is similar to the GRU, but the nodes in the recurrent layer apply a different algorithm as follows.

1) Calculate the value for the input gate.

$$\mathrm{i}_{t=n} = \sigma(w_i \bullet [h_{t=n-1}, x_{t=n}] + \beta_i) \tag{2.9}$$

2) Calculate the value for the forget gate.

$$\mathrm{f}_{t=n} = \sigma(w_f \bullet [h_{t=n-1}, x_{t=n}] + \beta_f) \tag{2.10}$$

3) Calculate the 'cell state' value.

$$\mathrm{c}_{\widetilde{t=n}} = \tanh(w_{c\sim} \bullet [h_{t=n-1}, x_{t=n}] + \beta_{c\sim}) \tag{2.11}$$

$$\mathrm{c}_{t=n} = \ f_{t=n} \ \bullet c_{t=n-1} + \ i_{t=n} \ \bullet \ c_{\widetilde{t=n}} \tag{2.12}$$

4) Calculate the hidden state value.

$$\mathrm{h}_{t=n} = \sigma(w_o \bullet [h_{t=n-1}, x_{t=n}] + \beta_o) \ \bullet \tanh(c_{t=n}) \tag{2.13}$$

5) Just like GRUs and simple RNNs, the final hidden state value calculated for the last wave is then the output for that recurrent neuron. The remainder of the process for the LSTM is the same as the simple RNN as per steps 3 - 6 in Appendix Section 2.6.2 above.

2.6.3 Data Details

Additional Information

This section provides additional details about the GESIS Panel data. This includes detailed descriptions of each predictor and the dependent variable (Appendix Table 2.5); how each predictor was categorized into a block for feature importance analysis (Appendix Table 2.6); the dates of each wave (Appendix Table 2.7); and the American Association for Public Opinion Research (AAPOR) response codes which we categorize as nonresponse (Appendix Table 2.6).

Note that for each variable in Appendix Table 2.5, the original survey questions as asked at each wave can be traced from the file in the project files './data/results/glossary.csv' that provides the GESIS question ID for every survey item used to make up each variable. The original question wording can then be found in the codebook provided in the GESIS data.

Table 2.5: *Description of predictors used in machine learning models. Note that predictors without a predictor set are included only in the 'all predictors' predictor set.*

Variable Label	Type	Predictor Set	Values	Description/Comments
Average Nonresponse All Previous Waves	Continuous	-	[0,5]	Rolling average nonresponse rate over the current and all previous waves.
Average Nonresponse Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Average Nonresponse Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Cooperation Panel	Continuous	time invariant	0..5	This value is inputted by the recruitment interviewer. A separate dummy variable flags missingness. Collected for cohort one only.
Cooperation Survey	Continuous	time invariant	0..5	This value is inputted by the recruitment interviewer. A separate dummy variable flags missingness. Collected for cohort one only.
Cooperation Wave	Continuous	time invariant	0..5	This value is inputted by the recruitment interviewer. A separate dummy variable flags missingness. Collected for cohort one only.
Social Trust	Continuous	time invariant	0..5	Collected at recruitment. A separate dummy variable flags missingness. Collected for cohort one only.
Subjective Well Being	Continuous	time invariant	0..5	Collected at recruitment. A separate dummy variable flags missingness. Collected for cohort one only.
Survey Evaluation Difficult	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Difficult All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.
Survey Evaluation Difficult Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Evaluation Difficult Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Survey Evaluation Diverse	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Diverse All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.

Continued on next page...

Variable Label	Type	Predictor Set	Values	Description/Comments
Survey Evaluation Diverse Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Evaluation Diverse Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Survey Evaluation Important	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Important All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.
Survey Evaluation Important Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Evaluation Important Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Survey Evaluation Interesting	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Interesting All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.
Survey Evaluation Interesting Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Evaluation Interesting Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Survey Evaluation Long	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Long All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.
Survey Evaluation Long Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Evaluation Long Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Survey Evaluation Overall	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Personal	Continuous	time variant	0..5	0 is nonresponse.
Survey Evaluation Personal All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.
Survey Evaluation Personal Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Evaluation Personal Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.

Continued on next page...

Variable Label	Type	Predictor Set	Values	Description/Comments
Survey Satisfaction Overall All Previous Waves	Continuous	-	[0,5]	Rolling average for all hitherto waves.
Survey Satisfaction Overall Previous One Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous wave.
Survey Satisfaction Overall Previous Three Waves	Continuous	-	[0,5]	Rolling average of current wave and the previous three waves.
Age	Categorical	time invariant	17.9 to 31.2, 31.2 to 44.4, 44.4 to 57.6, 57.6 to 70.8, 70.8 to 84.0, missing	The ages of all participants were categorized into five bins plus one bin to indicate missing value. The lower value of each bin is exclusive.
Cooperation Responsiveness	Categorical	time invariant	0,5	Inputted by the recruitment interviewer. A separate dummy variable flags missingness.
Disposition Code	Categorical	-	Missing by mode, Nonresponse, Completed, Completed after break, Suspended	Not included in time variant predictor set because it contains mostly the same information as 'participation mode.'
Education Category	Categorical	time invariant	College, No College	Derived from more detailed categories in the original GESIS data. A separate dummy variable flags missingness.
Email Provided	Categorical	time invariant	0,1	A separate dummy variable flags missingness. Collected for cohort one only.
Employment	Categorical	time invariant	Employed, Unemployed	Derived from more detailed categories in the original GESIS data. A separate dummy variable flags missingness.
Household Condition	Categorical	time invariant	0,5	Inputted by the recruitment interviewer. 0 is nonresponse. A separate dummy variable flags missingness. Collected for cohort one only.
Household Income	Categorical	time invariant	[< EUR 900/month, > EUR 5,000/month], Missing	

Continued on next page...

Variable Label	Type	Predictor Set	Values	Description/Comments
Household Size	Categorical	time invariant	1,2,3,4,5+, Missing	
Household Type	Categorical	time invariant	Family House, Residential,	A separate dummy variable flags missingness. Collected for cohort one only.
			Large, Missing	
Invitation Mode	Categorical	-	Online, Offline	Not included in time variant predictor set because it contains mostly the same information as the variable 'participation mode.'
Job Type	Categorical	time invariant	self-employed, blue-collar, white-collar, other, Missing	Derived from more detailed categories in the original GESIS data. A separate dummy variable flags missingness.
Marital Status	Categorical	time invariant	Married, Single, Unknown, Missing	Derived from more detailed categories in the original GESIS data.
Participation Mode	Categorical	time variant	Online, Offline, Nonresponse	
Personal Income	Categorical	time invariant	[< EUR 900/month, > EUR 5,000/month], Missing	
Phone Provided	Categorical	time invariant	0,1	A separate dummy variable flags missingness. Collected for cohort one only.
Response Status	Categorical	-	complete, non-participation, partial	We use participation mode rather than response status as the time variant predictor, because both contain a category for nonresponse and using both introduces unnecessary multicollinearity.
Sex	Categorical	time invariant	Male, Female, Other	A separate dummy variable flags missingness.
Social Status	Categorical	time invariant	0,5	A separate dummy variable flags missingness. Collected for cohort one only.
Survey Break	Categorical	time variant	0,1	A separate dummy variable flags missingness. Note that this only applies to online submissions. Postal submissions (as indicated participation mode) are always 0.

Continued on next page...

Variable Label	Type	Predictor Set	Values	Description/Comments
Survey Experience	Categorical	time invariant	0,5	0 is nonresponse. Collected for cohort one only.
Survey Place	Categorical	time variant	Home, Not Home	A separate dummy variable flags missingness.

Table 2.6: *Matching predictors from Appendix Table 2.5 predictor blocks used in Figure 2.8.*

Variable	Group
Age (17.934, 31.2]	Age
Age (31.2, 44.4]	Age
Age (44.4, 57.6]	Age
Age (57.6, 70.8]	Age
Age (70.8, 84.0]	Age
Age Missing	Age
Average Nonresponse Last Two Waves	Rolling Response Status
Average Nonresponse All Previous Waves	Rolling Response Status
Average Nonresponse Last Three Waves	Rolling Response Status
Cooperation Panel	Cooperation
Cooperation Responsiveness Bad	Cooperation
Cooperation Responsiveness Good	Cooperation
Cooperation Responsiveness Initially bad, later not so bad	Cooperation
Cooperation Responsiveness Moderately	Cooperation
Cooperation Responsiveness Missing	Cooperation
Cooperation Survey	Cooperation
Cooperation Wave	Cooperation
Disposition Code Completed	Response Status
Disposition Code Completed after break	Response Status
Disposition Code Missing by Mode	Response Status
Disposition Code Suspended	Response Status
Disposition Code Unit nonresponse	Response Status
Education Category 1.0	Education
Education Category 2.0	Education
Education Category Missing	Education
Education Level High	Education
Education Level Lower	Education
Education Level Medium	Education
Education Level Missing	Education
Email Provided Don't have E-Mail address	Provided Contact Details
Email Provided Missing	Provided Contact Details
Email Provided No	Provided Contact Details
Email Provided Not asked	Provided Contact Details
Email Provided Yes	Provided Contact Details
Employment full time	Income and Employment
Employment in training	Income and Employment
Employment marginal	Income and Employment

Continued on next page...

Variable	Group
Employment Missing	Income and Employment
Employment not employed	Income and Employment
Employment part time	Income and Employment
Household Condition Bad condition	Household
Household Condition Good condition	Household
Household Condition Missing	Household
Household Condition Satisfactory condition	Household
Household Condition Very bad condition	Household
Household Condition Very good condition	Household
Household Income 0.0	Income and Employment
Household Income 1100.0	Income and Employment
Household Income 1300.0	Income and Employment
Household Income 1500.0	Income and Employment
Household Income 1700.0	Income and Employment
Household Income 2000.0	Income and Employment
Household Income 2300.0	Income and Employment
Household Income 2600.0	Income and Employment
Household Income 3200.0	Income and Employment
Household Income 4000.0	Income and Employment
Household Income 5000.0	Income and Employment
Household Income 6000.0	Income and Employment
Household Income 700.0	Income and Employment
Household Income 900.0	Income and Employment
Household Income Missing	Income and Employment
Household Size 1.0	Household
Household Size 2.0	Household
Household Size 3.0	Household
Household Size 4.0	Household
Household Size 5.0	Household
Household Size Missing	Household
Household Type building big	Household
Household Type Missing	Household
Household Type One Two Family House	Household
Household Type other	Household
Household Type Residential Medium	Household
Invitation Mode Online	Mode
Job Type Employee	Income and Employment
Job Type Missing	Income and Employment
Job Type Other	Income and Employment
Job Type Self-employed	Income and Employment
Job Type Worker	Income and Employment
Marital Status Married	Martial Status
Marital Status Missing	Martial Status
Marital Status Single	Martial Status
Marital Status Unknown	Martial Status
Participation Mode Not participated	Response Status
Participation Mode Offline	Mode
Participation Mode Online	Mode
Personal Income 0.0	Income and Employment
Personal Income 1100.0	Income and Employment

Continued on next page...

Variable	Group
Personal Income 1300.0	Income and Employment
Personal Income 1500.0	Income and Employment
Personal Income 1700.0	Income and Employment
Personal Income 2000.0	Income and Employment
Personal Income 2300.0	Income and Employment
Personal Income 2600.0	Income and Employment
Personal Income 300.0	Income and Employment
Personal Income 3200.0	Income and Employment
Personal Income 4000.0	Income and Employment
Personal Income 500.0	Income and Employment
Personal Income 5000.0	Income and Employment
Personal Income 700.0	Income and Employment
Personal Income 900.0	Income and Employment
Personal Income Missing	Income and Employment
Phone Provided Missing	Provided Contact Details
Phone Provided No	Provided Contact Details
Phone Provided Yes	Provided Contact Details
Response Status complete interview	Response Status
Response Status nonparticipation	Response Status
Response Status partial interview	Response Status
Sex Ambiguous answer	Sex
Sex Female	Sex
Sex Male	Sex
Social Status Indistinguishable	Income and Employment
Social Status Lower class	Income and Employment
Social Status Middle class	Income and Employment
Social Status Missing	Income and Employment
Social Status Upper class	Income and Employment
Social Status Upper middle class	Income and Employment
Social Status Working class	Income and Employment
Social Trust	Social Trust
Subjective Well Being	Subjective Well Being
Survey Break Ambiguous answer	Response Status
Survey Break Item nonresponse	Response Status
Survey Break Missing	Response Status
Survey Break No, participated in one piece	Response Status
Survey Break Not reached	Response Status
Survey Break Unit nonresponse	Response Status
Survey Break Yes, I have interrupted participation for X Minutes.	Response Status
Survey Evaluation Difficult	Survey Evaluation
Survey Evaluation Difficult All Previous Waves	Rolling Survey Evaluation
Survey Evaluation Difficult Last Two Waves	Rolling Survey Evaluation
Survey Evaluation Difficult Last Three Waves	Rolling Survey Evaluation
Survey Evaluation Diverse	Survey Evaluation
Survey Evaluation Diverse All Previous Waves	Rolling Survey Evaluation
Survey Evaluation Diverse Last Two Waves	Rolling Survey Evaluation
Survey Evaluation Diverse Last Three Waves	Rolling Survey Evaluation
Survey Evaluation Important	Survey Evaluation
Survey Evaluation Important All Previous Waves	Rolling Survey Evaluation

Continued on next page...

Variable	Group
Survey Evaluation Important Last Two Waves	Rolling Survey Evaluation
Survey Evaluation Important Last Three Waves	Rolling Survey Evaluation
Survey Evaluation Interesting	Survey Evaluation
Survey Evaluation Interesting All Previous Waves	Rolling Survey Evaluation
Survey Evaluation Interesting Last Two Waves	Rolling Survey Evaluation
Survey Evaluation Interesting Last Three Waves	Rolling Survey Evaluation
Survey Evaluation Long	Survey Evaluation
Survey Evaluation Long All Previous Waves	Rolling Survey Evaluation
Survey Evaluation Long Last Two Waves	Rolling Survey Evaluation
Survey Evaluation Long Last Three Waves	Rolling Survey Evaluation
Survey Evaluation Overall	Survey Evaluation
Survey Evaluation Personal	Survey Evaluation
Survey Evaluation Personal All Previous Waves	Rolling Survey Evaluation
Survey Evaluation Personal Last Two Waves	Rolling Survey Evaluation
Survey Evaluation Personal Last Three Waves	Rolling Survey Evaluation
Survey Experience Don't know	Survey Experience
Survey Experience Missing	Survey Experience
Survey Experience No	Survey Experience
Survey Experience Yes	Survey Experience
Survey Place home	Survey Place
Survey Place Missing	Survey Place
Survey Place not home	Survey Place
Survey Satisfaction Overall All Previous Waves	Rolling Survey Evaluation
Survey Satisfaction Overall Last Two Waves	Rolling Survey Evaluation
Survey Satisfaction Overall Last Three Waves	Rolling Survey Evaluation

Table 2.7: *Index of waves in the GESIS panel (GESIS 2023). Note that waves a11, a12, d11, d12, f11, f12 are recruitment waves and are not included in the analysis.*

Wave	Start	End
a11	8-Jun-13	1-Dec-13
a12	26-Jun-13	31-Jan-14
aa	21-Aug-13	14-Oct-13
ab	16-Oct-13	10-Dec-13
ac	11-Dec-13	19-Feb-14
ba	19-Feb-14	15-Apr-14
bb	16-Apr-14	17-Jun-14
bc	18-Jun-14	12-Aug-14
bd	13-Aug-14	14-Oct-14
be	15-Oct-14	16-Dec-14
bf	17-Dec-14	17-Feb-15
ca	18-Feb-15	14-Apr-15
cb	15-Apr-15	16-Jun-15
cc	17-Jun-15	11-Aug-15
cd	12-Aug-15	14-Oct-15
ce	14-Oct-15	15-Dec-15
cf	15-Dec-15	16-Feb-16
d11	2-May-16	23-Sep-16
d12	2-May-16	23-Sep-16

Continued on next page...

Wave	Start	End
da	17-Feb-16	19-Apr-16
db	20-Apr-16	14-Jun-16
dc	15-Jun-16	16-Aug-16
dd	17-Aug-16	18-Oct-16
de	19-Oct-16	13-Dec-16
df	14-Dec-16	14-Feb-17
ea	15-Feb-17	18-Apr-17
eb	19-Apr-17	13-Jun-17
ec	14-Jun-17	15-Aug-17
ed	12-Sep-17	23-Sep-17
ee	18-Oct-17	12-Dec-17
ef	13-Dec-17	13-Feb-18
f11	1-Apr-18	1-Sep-18
f12	1-Apr-18	1-Sep-18
fa	14-Feb-18	17-Apr-18
fb	18-Apr-18	12-Jun-18
fc	13-Jun-18	14-Aug-18
fd	15-Aug-18	16-Oct-18
fe	5-Oct-18	11-Dec-18
ff	12-Dec-18	12-Feb-19
ga	13-Feb-19	16-Apr-19
gb	17-Apr-19	11-Jun-19
gc	12-Jun-19	13-Aug-19
gd	14-Aug-19	15-Oct-19
ge	16-Oct-19	10-Dec-19
gf	11-Dec-19	11-Feb-20
hb	20-May-20	7-Jul-20
hc	8-Jul-20	23-Aug-20
hd	26-Aug-20	13-Oct-20
he	14-Oct-20	8-Dec-20
hf	9-Dec-20	9-Feb-21
ia	24-Feb-21	20-Apr-21
ib	26-May-21	20-Jul-21

Table 2.8: *How the nonresponse variable is defined. For each invited participant, GESIS categorizes their response to the invitation in the variable 'response category.' If a given response category is any of the following values, we consider that participant to have nonresponded in that wave.*

AAPOR disposition category code	AAPOR disposition category label
319	Nothing ever returned
21121	Explicit refusal
3311	Post: Attempted - Addressee not known at place of address
212	Break-off: questionnaire too incomplete to process / break-off or partial with insufficient information
211211	Explicit refusal with incentive
2112	Known respondent-level refusal

Continued on next page...

AAPOR disposition category code	AAPOR disposition category label
211221	Logged on to survey did not complete any items
2113	Blank questionnaire mailed back implicit refusal
33112	Postal box full
21122	Implicit refusal
33114	Email Bouncer: Mailbox unknown
2111	Other person refusal
33115	Email Bouncer: Postbox full
231	Death (including Post: Deceased)
33113	Email Bouncer: Delivery problem
232	Physically or mentally unable/incompetent
332	Post: Moved left no address
21131	Blank questionnaire with incentive returned
2332	Respondent language problem
211212	Explicit refusal no incentive
331	Post: Undeliverable as addressed
3253	Post: No Mail Receptacle
211	Refusal
2113	Blank questionnaire with no incentive returned
391	Returned from an unsampled person
330	Invitation returned undelivered (Email Bouncer)

Data Formatting

Each of the three types of models (traditional, RNNs, and TSCTs) requires different input data formats. For all models, we first transform the 'wide' format of raw GESIS Panel data to 'long' format (see Appendix Figure 2.10).

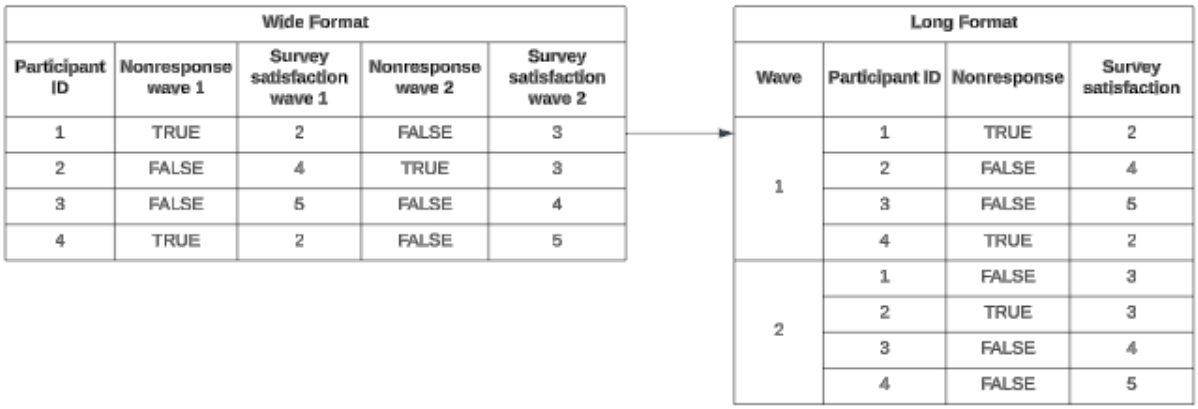


Figure 2.10: *Exemplary diagram of the first stage of transforming raw GESIS Panel data into the format required for the analysis.*

Traditional Models Logistic regression and random forest were given inputs of the

format shown in Appendix Figure 2.11. Appendix Figure 2.11 shows an example scenario in which we are creating a model which will predict nonresponse in the sixth wave of the GESIS survey. We can train the model with predictors drawn from any wave up to wave four and then use nonresponses in wave five as the training dependent. We then test how well our fitted model performs by inputting data from wave five and outputting predictions for the nonresponses in wave six. Those predictions can then be compared with the actual outcomes in wave six to calculate an AUROC score for model performance at wave 6.

In the training data, to reduce computation time, we filter out duplicate participants by including only the row for the participant's second latest wave (because we use their latest wave for the dependent variable). In the test data, we include only participants invited to the wave we are predicting. In the example from Appendix Figure 2.11, the participant with ID '3' is excluded from the test set because they are not invited to wave six. Still, participant number three can be included in the training set because they were present in wave four, so we can use their data up to wave three, and nonresponse outcome at wave four, to train the model.

Train

Metadata		Predictors				Dependent
Participant ID	Wave	Gender	Nonresponse this wave	Survey satisfaction	Rolling Average nonresponse over last three waves	Nonresponse next wave
1	4	Male	TRUE	2	0.6	TRUE
2	4	Female	FALSE	4	0.8	FALSE
3	3	Male	FALSE	5	0.3	FALSE
4	4	Female	TRUE	2	0.6	TRUE

Test

Participant ID	Wave	Gender	Nonresponse this wave	Survey satisfaction	Rolling Average nonresponse over last three	Nonresponse next wave
1	5	Male	FALSE	3	0.5	FALSE
2	5	Female	FALSE	5	0.2	FALSE
4	5	Female	TRUE	1	0.4	TRUE

Figure 2.11: *Diagram of input for the logistic regression and random forest ML models.*

RNN The format for the RNNs is similar to that of the traditional models, except each predictor is a timeline of values instead of a single value. In Appendix Figure 2.12, we can see that the training predictors are three-dimensional: one axis for the participant, another axis is the variables, and the third axis is the temporal axis. For example, for the participant with ID '2,' for the variable 'survey satisfaction,' the RNN model would be trained on an array of the values which are the participant's survey

satisfaction responses for waves four, three, two, and one. In the case of participant 3, they were only in the panel until wave three (and nonresponded in wave four), so we received an array of their values for waves three, two, and one. The test data only consists of participants invited to wave six, so every timeline of values would be every wave each participant had been in up to that point. That timeline would include all survey waves for a participant recruited from the beginning of the GESIS panel but fewer for participants from a later cohort. Note that even waves the participant was invited to, but did not respond to, would still have an entry in the timeline of values.

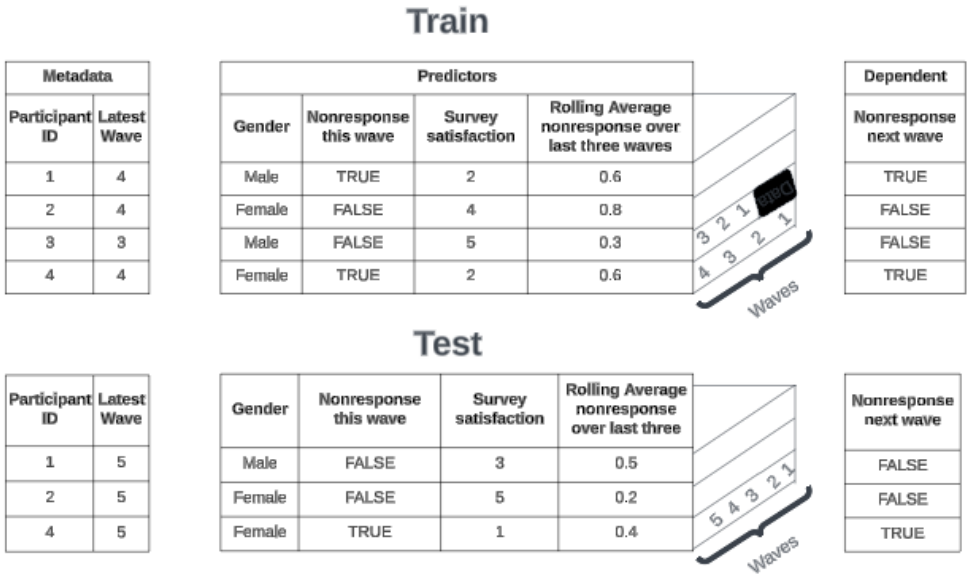


Figure 2.12: *Diagram of example input for the traditional ML models.*

Time Series Classification Techniques The format for these models is the same as that of RNNs but with two exceptions. Firstly, every time series must be the same length, so where participants in the training set were in fewer waves than others (such as participant number three), the values of the missing waves are imputed with zeros. Secondly, the time series in the test data must be the same length as the training waves, even though one additional wave is available in the test set. Therefore, we exclude the earliest wave for each time series in the test data to truncate the time series to be the same length as those in the training set. Appendix Figure 2.13 shows how each TSCT algorithm converts the three-dimensional input data into a two-dimensional Table which is then inputted into its component classifier to make the final prediction.

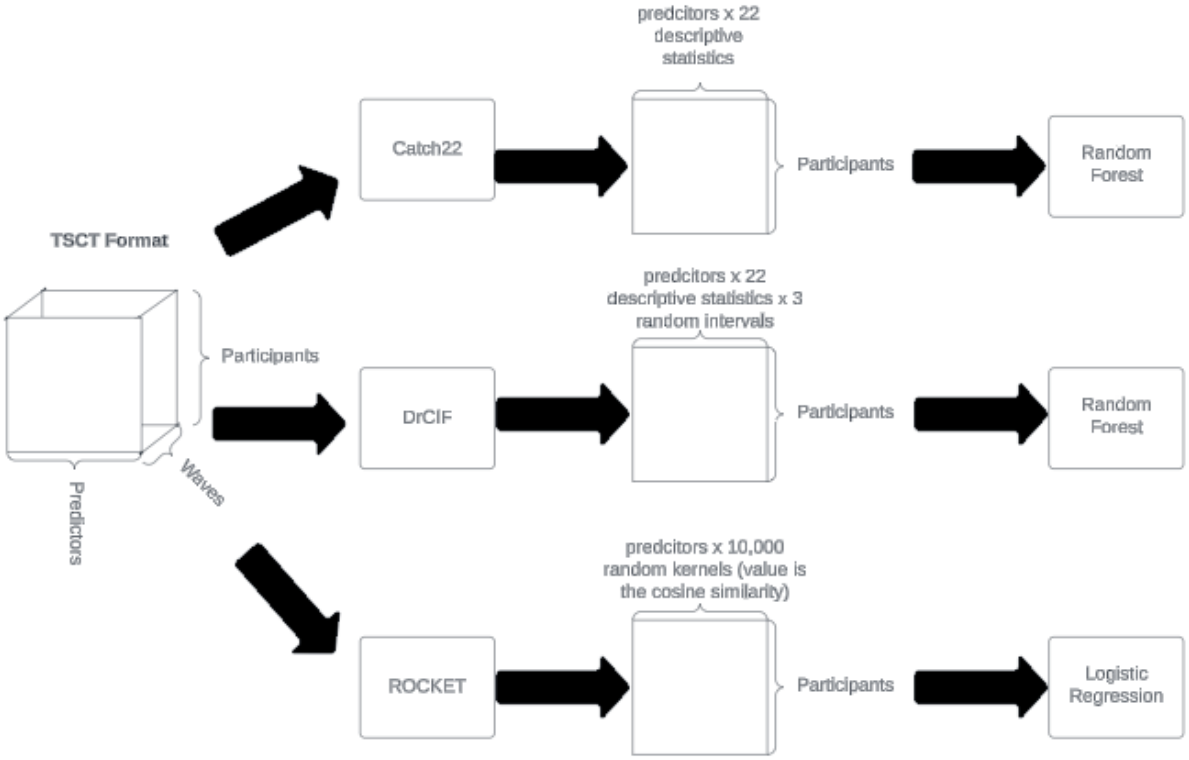


Figure 2.13: *An illustration of how the TSCTs convert the three-dimensional input back into a two-dimensional input, which is then inputted into traditional classifiers. Note that the two-dimensional matrices derived here should not be confused with the long format matrices in Appendix Figures 2.10 and 2.11.*

Missing Values and Scaling
The primary sources of missing values are as follows.

- Participants may refuse to answer a question about their demographics in the recruitment interview. In these cases, we 'dummy encode' that variable, which means to convert a categorical variable (i.e., gender is male, female, other, no answer) into a set of binary variables (i.e., 'is male,' 'is female,' 'is missing'). One of these dummy variables will represent a missing value. If the demographic variable is continuous (i.e., age) we separate the values into bins to make it categorical, see Appendix Table 2.5 for details.

- Demographic values are missing in waves other than the recruitment wave and the periodic waves where these questions are repeated. This is because GESIS only asks the participant for these details in particular waves, instead of every time. We fill the missing value with the last known value in these cases.

- Some values are missing because the participant did not respond to the wave. This affects predictors such as the survey satisfaction questions and whether they took a break during the survey. The only continuous variables this issue introduces error into are the survey satisfaction questions, where we input zero for a missing value (note that when these questions are answered the lowest valid value possible is 1). Therefore, rolling-averages based on these survey evaluation items could be low because the respondent was dissatisfied, or because an item was missing, thereby adding some error into what that variable indicates. This error is mitigated by using the 'nonresponse in this wave' variable as a missingness flag for these survey evaluation indicators. Categorical predictors are dummy encoded so that they have a category for missing values.

- The survey break variables only apply to online respondents. In these cases, the mode variables would indicate the reason for missingness.

- Some values are missing because certain questions were asked in the first cohort recruitment wave, but not in subsequent ones. These are: Household type, household condition, survey cooperation, social status, social trust, prior survey experience, flags for not providing contact details. For these features, we provide a missingness flag.

We scale continuous variables by dividing each element by that variable's maximum possible value. This is appropriate because the maximum value is known in advance for these survey questions, and extreme outliers are therefore not possible. Appendix Table 2.5 indicates which variables are continuous.

2.6.4 Hypertuning Details

This section provides details on the model training procedures. We explain the combinations of all model parameters we experimented with (Appendix Table 2.9); the waves,

cohort groupings, and predictor sets we repeated these fittings over (Appendix Table 2.10); and the particular case of the TSCT models (Appendix Table 2.11). In total, we fit 13,392 (72 of which were TSCT) models to predict nonresponse in the subsequent wave.

Table 2.9: *Model settings for each model when predicting nonresponse in the next wave. For replication purposes, note that all RNNs were fitted with the 'Adaptive Moment Estimation (ADAM)' optimization algorithm, all activation functions were sigmoid functions, and the learning rate was 0.001 (Kingma and Ba, 2017).*

Model Type	Setting	Values	Number of Unique Settings
Logistic Regression	penalty	L1, or L2 Regularization, or No Penalty	9
	Optimization solver	'Liblinear' for Penalized and 'Limited Memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS)' for Unpenalized	
	Fitting stopping tolerance	01	
	C (only applies to those with L1 or L2 penalty)	0.05, 0.1, 1, 1000	
Random Forest	Number of trees in the forest	500	1
	The function to measure the quality of a split	Gini impurity	
	Minimum number of samples for a split	2	
	Minimum number of samples for a leaf	1	
	Number of predictors considered at each split	Square root of number of all predictors	
Simple RNN	depth	8, 32, 128	9
	width	0,1,2	
	dropout	0.6	
LSTM	depth	8, 32, 128	9
	width	0,1,2	
	dropout	0.6	
GRU	depth	8, 32, 128	9
	width	0,1,2	
	dropout	0.6	
Total			37

Table 2.10: *Other variables permutated at each model fitting following from Appendix Table 5.*

Dimension	Parameters	Values
Uniquely Parameterized Models	See Appendix Table 5	37
Waves per Cohort	All cohorts Cohort one Cohort two Cohort three	40 40 26 14
Sub Total Predictor Sets	 Time-variant predictors only Time-invariant predictors only All predictors	120 3
Total		13,320

Table 2.11: *Summary of TSCT model fittings in the project.*

Predictors	Model Class	Waves fitted
All Predictors	Catch22 DrCIF ROCKET	12 12 12
Time Variant Only	Catch22 DrCIF ROCKET	12 12 12
Total		72

2.6.5 Additional Results

Over 13,000 models were fitted for this study. The main text summarizes these results, and in this appendix section, we provide the precise AUROC values for each model setting averaged over all waves. These results allow us to precisely measure the rankings of different settings and examine the impact of hypertuning on model performance.

Table 2.12: *AUROC for each model setting when predicting nonresponse in the next wave. Note that RNNs with "num_layers" = 0 have no layers other than the recurrent and output layers.*

Model Class	Features	Parameters	min	max	mean	std
Random Forest	All Predictors	'max features': 'sqrt', 'n estimators': 500	0.678	0.890	0.857	0.034
LSTM	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.668	0.895	0.856	0.036

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
LSTM	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.673	0.888	0.855	0.036
LSTM	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.611	0.892	0.853	0.045
LSTM	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.673	0.896	0.853	0.035
LSTM	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.639	0.894	0.852	0.040
LSTM	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.684	0.893	0.852	0.035
LSTM	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.654	0.891	0.852	0.038
LSTM	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.668	0.894	0.851	0.037
LSTM	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.658	0.893	0.850	0.041
GRU	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.655	0.893	0.849	0.040
GRU	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.655	0.892	0.849	0.039
GRU	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.656	0.887	0.849	0.039
GRU	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.659	0.891	0.847	0.040
GRU	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.662	0.895	0.847	0.041

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
GRU	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.657	0.889	0.847	0.038
GRU	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.652	0.889	0.846	0.041
GRU	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.661	0.888	0.846	0.042
GRU	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.649	0.890	0.846	0.040
LSTM	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.652	0.882	0.843	0.037
LSTM	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.639	0.878	0.842	0.039
LSTM	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.673	0.880	0.842	0.036
Simple RNN	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.676	0.888	0.840	0.036
LSTM	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.660	0.880	0.839	0.038
LSTM	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.613	0.875	0.839	0.044
Simple RNN	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.620	0.878	0.839	0.043
LSTM	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.647	0.879	0.838	0.040
LSTM	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.642	0.883	0.838	0.045

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
LSTM	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.645	0.877	0.837	0.040
Simple RNN	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.629	0.878	0.837	0.042
Simple RNN	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.632	0.880	0.837	0.040
Simple RNN	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.559	0.884	0.837	0.060
Simple RNN	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.620	0.874	0.837	0.042
LSTM	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.600	0.878	0.837	0.046
Simple RNN	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.585	0.890	0.837	0.057
Simple RNN	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.670	0.878	0.836	0.039
Simple RNN	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.616	0.876	0.835	0.043
Simple RNN	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.622	0.877	0.834	0.045
Simple RNN	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.588	0.875	0.833	0.047
Simple RNN	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.657	0.875	0.831	0.041
Penalized Logistic Regression	All Predictors	'C': 1, 'penalty': 'l1'	0.651	0.868	0.831	0.033

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
GRU	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.659	0.880	0.831	0.040
Simple RNN	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.304	0.892	0.831	0.094
Simple RNN	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.301	0.892	0.829	0.093
Simple RNN	Time Variant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.296	0.894	0.829	0.096
GRU	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.630	0.873	0.829	0.046
GRU	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.633	0.874	0.828	0.045
GRU	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.626	0.873	0.827	0.042
GRU	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.585	0.876	0.827	0.048
Simple RNN	Time Variant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.330	0.892	0.826	0.092
GRU	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.648	0.876	0.826	0.047
GRU	All Predictors	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.653	0.877	0.826	0.040
Simple RNN	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.315	0.884	0.825	0.092
GRU	All Predictors	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.633	0.874	0.824	0.051

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
Penalized Logistic Regression	All Predictors	'C': 0.1, 'penalty': 'l1'	0.636	0.870	0.824	0.037
Simple RNN	Time Variant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.282	0.890	0.824	0.098
GRU	All Predictors	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.629	0.882	0.823	0.058
Catch22	Time Variant Only		0.679	0.873	0.822	0.053
Penalized Logistic Regression	All Predictors	'C': 1000, 'penalty': 'l2'	0.644	0.862	0.817	0.037
Penalized Logistic Regression	All Predictors	'C': 1, 'penalty': 'l2'	0.642	0.865	0.817	0.036
Penalized Logistic Regression	All Predictors	'C': 1000, 'penalty': 'l1'	0.643	0.859	0.816	0.036
Penalized Logistic Regression	All Predictors	'C': 0.1, 'penalty': 'l2'	0.639	0.867	0.813	0.037
Catch22	All Predictors		0.686	0.858	0.812	0.043
Penalized Logistic Regression	All Predictors	'C': 0.05, 'penalty': 'l2'	0.638	0.868	0.811	0.037
Penalized Logistic Regression	All Predictors	'C': 0.05, 'penalty': 'l1'	0.620	0.870	0.810	0.039
Baseline Logistic Regression	All Predictors	'penalty': 'none'	0.640	0.858	0.809	0.035
Penalized Logistic Regression	Time Variant Only	'C': 1, 'penalty': 'l2'	0.583	0.823	0.753	0.042
Penalized Logistic Regression	Time Variant Only	'C': 1, 'penalty': 'l1'	0.580	0.823	0.753	0.042
Penalized Logistic Regression	Time Variant Only	'C': 0.1, 'penalty': 'l2'	0.582	0.823	0.753	0.042
Penalized Logistic Regression	Time Variant Only	'C': 1000, 'penalty': 'l2'	0.583	0.821	0.752	0.042

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
Penalized Logistic Regression	Time Variant Only	'C': 0.1, 'penalty': 'l1'	0.575	0.818	0.752	0.043
Penalized Logistic Regression	Time Variant Only	'C': 0.05, 'penalty': 'l2'	0.581	0.821	0.751	0.043
Penalized Logistic Regression	Time Variant Only	'C': 0.05, 'penalty': 'l1'	0.576	0.815	0.751	0.043
Random Forest	Time Invariant Only	'max features': 'sqrt', 'n estimators': 500	0.596	0.809	0.750	0.045
Baseline Logistic Regression	Time Variant Only	'penalty': 'none'	0.583	0.821	0.749	0.041
Penalized Logistic Regression	Time Variant Only	'C': 1000, 'penalty': 'l1'	0.382	0.821	0.746	0.068
DrCIF	All Predictors	'time limit in minutes': 2.0	0.614	0.773	0.737	0.044
Random Forest	Time Variant Only	'max features': 'sqrt', 'n estimators': 500	0.585	0.781	0.728	0.039
DrCIF	Time Variant Only	'time limit in minutes': 2.0	0.601	0.754	0.705	0.049
Penalized Logistic Regression	Time Invariant Only	'C': 1, 'penalty': 'l1'	0.549	0.756	0.680	0.042
Penalized Logistic Regression	Time Invariant Only	'C': 1000, 'penalty': 'l2'	0.550	0.754	0.678	0.042
Penalized Logistic Regression	Time Invariant Only	'C': 1000, 'penalty': 'l1'	0.550	0.754	0.678	0.043
Penalized Logistic Regression	Time Invariant Only	'C': 1, 'penalty': 'l2'	0.545	0.753	0.677	0.043
Penalized Logistic Regression	Time Invariant Only	'C': 0.1, 'penalty': 'l2'	0.544	0.747	0.673	0.042
Baseline Logistic Regression	Time Invariant Only	'penalty': 'none'	0.552	0.748	0.672	0.042
Penalized Logistic Regression	Time Invariant Only	'C': 0.05, 'penalty': 'l2'	0.547	0.742	0.669	0.041
LSTM	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.579	0.747	0.667	0.037

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
Penalized Logistic Regression	Time Invariant Only	'C': 0.1, 'penalty': 'l1'	0.535	0.749	0.664	0.045
LSTM	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.579	0.740	0.664	0.041
LSTM	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.579	0.745	0.664	0.039
LSTM	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.579	0.743	0.663	0.042
Simple RNN	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.573	0.746	0.663	0.038
LSTM	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.566	0.745	0.663	0.040
LSTM	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.553	0.750	0.661	0.046
LSTM	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.559	0.734	0.661	0.044
Simple RNN	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.570	0.745	0.660	0.042
LSTM	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.565	0.741	0.660	0.042
LSTM	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.565	0.745	0.659	0.046
Simple RNN	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.566	0.741	0.659	0.040
Simple RNN	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.561	0.737	0.659	0.038

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
Simple RNN	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.525	0.743	0.658	0.043
Simple RNN	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.570	0.742	0.658	0.041
Simple RNN	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.576	0.750	0.657	0.042
Simple RNN	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.567	0.737	0.656	0.042
Simple RNN	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.563	0.745	0.656	0.046
GRU	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 32	0.539	0.736	0.650	0.042
Penalized Logistic Regression	Time Invariant Only	'C': 0.05, 'penalty': 'l1'	0.537	0.735	0.647	0.045
GRU	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 128	0.545	0.725	0.645	0.044
GRU	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 128	0.559	0.727	0.645	0.041
GRU	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 8	0.543	0.730	0.644	0.047
GRU	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 8	0.541	0.731	0.643	0.048
GRU	Time Invariant Only	'num layers': 1, 'recurrent dropout': 0.6, 'recurrent units': 32	0.539	0.727	0.642	0.042
GRU	Time Invariant Only	'num layers': 2, 'recurrent dropout': 0.6, 'recurrent units': 128	0.538	0.733	0.642	0.046

Continued on next page...

Model Class	Features	Parameters	min	max	mean	std
GRU	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 32	0.546	0.723	0.642	0.042
GRU	Time Invariant Only	'num layers': 0, 'recurrent dropout': 0.6, 'recurrent units': 8	0.559	0.740	0.638	0.045
ROCKET	All Predictors		0.528	0.657	0.608	0.033
ROCKET	Time Variant Only		0.514	0.606	0.562	0.023

2.6.6 PRICSSA Checklist

In this section we present the Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA) checklist. This information provides transparency and assists in replication for the study. The checklist is presented in Appendix Table 2.13.

Table 2.13: *PRICSSA Checklist (Seidenberg, Moser, and West, 2023).*

PRICSSA item	Description	Response
1.1 Data collection dates	Describe the survey's data collection dates (e.g., range) to provide historical context that could affect survey responses and nonresponse.	See Appendix Table 2.7.
1.2 Data collection mode(s)	Describe the survey's data collection mode(s). Data collection mode can affect survey responses (e.g., to sensitive questions), including nonresponse, and a survey's data collection mode may change over time (e.g., during the COVID-19 pandemic).	See Section 2.3.1. Survey is mixed mode offering web and mail options. Survey waves are two to three monthly.
1.3 Target population	State the target population the survey was designed to represent and describe all weighted estimates with respect to this target population.	A detailed description of the population and sampling method is found in Section 2.3.1 with a reference to the full documentation. The GESIS panel's sampling method was to randomly sample from the German population register. Anyone permanently residing in Germany between the ages of 18-70 was eligible for recruitment. Participants were interviewed at their homes to be recruited into the panel.

Continued on next page...

PRICSSA item	Description	Response
1.4 Sample design	Describe the survey's sample design, including information about stratification, cluster sampling, and unequal probabilities of selection.	A detailed description of the population and sampling method is found in Section 2.3.1 with a reference to the full documentation. Sampling method was random selection from population register of Germany-residents between 18-70.
1.5 Survey response rate(s)	State the survey's response rate and how it was calculated.	The definition of nonresponse based on RR6 and the selection of AAPOR response type codes is now provided (Section Validation) as well as the subsequent calculation of nonresponse rates at each wave (Figure 2.2).
2.1 Missingness rates	Report rates of missingness for variables of interest and models, and describe any methods (if any) for dealing with missing data (e.g., multiple imputation).	See Appendix Sections 2.6.3.
2.2 Observation deletion	State whether any observations were deleted from the dataset. If observations were deleted, provide a justification. Note: It is best practice to avoid deleting cases and use available subpopulation analysis commands no matter what variance estimation method is used.	No individual participant was removed from the dataset. See Appendix Section Data Formatting. which describes a procedure for using only the latest wave of data for a given participant, but this does not omit the participant from the analysis.
2.3 Sample sizes	Include unweighted sample sizes for all weighted estimates.	See Figure 2.2. No weighted estimated used.
2.4 Confidence intervals/standard errors	Include confidence intervals or standard errors when reporting all estimates to inform the reliability/precision of each estimate.	In our case, we do no significance tests but rather provide probabilistic predictions and the related goodness-of-fit information (Sections 2.3 and 2.4).
2.5 Weighting	State which analyses were weighted and specify which weight variables were used in analysis.	No sample weighting used.
2.6 Variance estimation	Describe the variance estimation method used in the analysis and specify which design variables (e.g., PSU/stratum, replicate weights) were used.	Not Applicable
2.7 Subpopulation analysis	Describe the procedures used for conducting subpopulation analyses (e.g., Stata's "subpop" command, SAS's "domain" command).	Not Applicable

Continued on next page...

PRICSSA item	Description	Response
2.8 Suppression rules	State whether or not a suppression rule was followed (e.g., minimum sample size or relative standard error).	Not Applicable
2.9 Software and code	Report which statistical software was used, comprehensively describe data management and analysis in the manuscript, and provide all statistical software code.	See Appendix Section 2.6.1.
2.10 Singleton problem (as needed)	Taylor Series Linearization requires at least two PSUs per stratum for variance estimation. Sometimes an analysis is being performed and there is only a single PSU in a stratum. There are several possible fixes to this problem, which should be detailed if the singleton problem is encountered.	Not Applicable
2.11 Public/restricted data (as needed)	If applicable, state whether the public use or restricted version of the dataset was analyzed.	See Appendix Section 2.6.1.
2.12 Embedded experiments (as needed)	If applicable, provide information about split sample embedded experiments (e.g., mode of data collection or varying participant incentives) and detail whether experimental factors were accounted for in the analyses.	Not Applicable

2.7 References

AAPOR (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9th edition. The American Association for Public Opinion Research.

Abanda, Amaia, Usue Mori, and Jose A. Lozano (2018). *A review on distance based time series classification*. 2018.

Altmann, André et al. (2010). "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10, pp. 1340–1347.

Bach, Ruben L, Stephanie Eckman, and Jessica Daikeler (2020). "Misreporting Among Reluctant Respondents". In: *Journal of Survey Statistics and Methodology* 8.3, pp. 566–588.

Becker, Rolf (2017). "Gender and Survey Participation An Event History Analysis of the Gender Effects of Survey Participation in a Probability-based Multi-wave Panel Study with a Sequential Mixed-mode Design". In: *methods* data. Artwork Size: 29 Pages Publisher: methods, data, analyses, 29 Pages.

Behr, Andreas, Egon Bellgardt, and Ulrich Rendtel (2005). "Extent and Determinants of Panel Attrition in the European Community Household Panel". In: *European Sociological Review* 21.5. Number: 5, pp. 489–512.

Bosnjak, Michael et al. (2018). "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel". In: *Social Science Computer Review* 36.1. Number: 1, pp. 103–115.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Burkam, David T. and Valerie E. Lee (1998). "Effects of Monotone and Nonmonotone Attrition on Parameter Estimates in Regression Models with Educational Data: Demographic Effects on Achievement, Aspirations, and Attitudes". In: *Journal of Human Resources* 33.2. Publisher: University of Wisconsin Press, pp. 555–574.

Cho, Kyunghyun et al. (2014). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014.

Christ, Maximilian et al. (2018). "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)". In: *Neurocomputing* 307, pp. 72–77.

Chun, Asaph, Steven Heeringa, and Barry Schouten (2018). "Responsive and Adaptive Design for Survey Optimization". In: *Journal of Official Statistics* 34, pp. 581–597.

Coffey, Stephanie, Benjamin Reist, and Peter V Miller (2020). "Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates". In: *Journal of Survey Statistics and Methodology* 8.4, pp. 726–747.

Dempster, Angus, François Petitjean, and Geoffrey I. Webb (2020). "ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels". In: *Data Mining and Knowledge Discovery* 34.5, pp. 1454–1495.

DiPietro, Robert and Gregory D. Hager (2020). "Chapter 21 - Deep learning: RNNs and LSTM". In: *Handbook of Medical Image Computing and Computer Assisted Intervention*. Ed. by S. Kevin Zhou, Daniel Rueckert, and Gabor Fichtinger. The Elsevier and MICCAI Society Book Series. Academic Press, London, United Kingdom, 2020, pp. 503–519.

Faouzi, Johann (2024). "Time Series Classification: A Review of Algorithms and Implementations". In: *Time Series Analysis - Recent Advances, New Perspectives and Applications*. Ed. by Jorge Rocha, Cláudia M. Viana, and Sandra Oliveira. IntechOpen, 2024.

Fawaz, Hassan Ismail et al. (2019). "Deep learning for time series classification: a review". In: *Data Mining and Knowledge Discovery* 33.4, pp. 917–963.

Fulcher, Ben D. (2017). *Feature-based time-series analysis*. 2017.

GESIS (2023). *GESIS Panel - Standard Edition*. Published: GESIS, Cologne. ZA5665 Data file Version 44.0.0, https://doi.org/10.4232/1.13931 DOI: 10.4232/1.13931. 2023.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. ISSN: 2379-190X, pp. 6645–6649.

Groves, Robert M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys". In: *Public Opinion Quarterly* 70.5, pp. 646–675.

Gummer, Tobias, Joss Roßmann, and Henning Silber (2021). "Using Instructed Response Items as Attention Checks in Web Surveys: Properties and Implementation". In: *Sociological Methods & Research* 50.1, pp. 238–264.

Hill, Craig A. et al., eds. (2020). *Big Data Meets Survey Science: A Collection of Innovative Methods*. 1st ed. Wiley, New Jersey.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Hoel, David G., Milton Sobel, and George H. Weiss (1975). "2 - A Survey of Adaptive Sampling for Clinical Trials". In: *Perspectives in Biometrics*. Ed. by ROBERT M. Elashoff. Academic Press, 1975, pp. 29–61.

Hyndman, Rob and George Athanasopoulos (2021). *Forecasting: Principles and Practice (3rd ed)*. O Texts.

Jacobsen, Erin et al. (2021). "Predictors of attrition in a longitudinal population-based study of aging". In: *International Psychogeriatrics* 33.8. Publisher: Cambridge University Press, pp. 767–778.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. Springer, New York.

Jankowsky, Kristin, Diana Steger, and Ulrich Schroeders (2022). *Predicting Lifetime Suicide Attempts in a Community Sample of Adolescents Using Machine Learning Algorithms*. 2022.

Kern, Christoph, Bernd Weiß, and Jan-Philipp Kolb (2021). "Predicting Nonresponse in Future Waves of a Probability-Based Mixed-Mode Panel with Machine Learning*". In: *Journal of Survey Statistics and Methodology* 11.1, pp. 100–123.

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. 2017.

Kocar, Sebastian and Nicholas Biddle (2022). "The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design". In: *Quality & Quantity*.

Kumar, Sumit et al. (2018). "Energy Load Forecasting using Deep Learning Approach-LSTM and GRU in Spark Cluster". In: *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)*. 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), pp. 1–4.

Le Cessie, S. and J. C. Van Houwelingen (1992). "Ridge Estimators in Logistic Regression". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41.1, pp. 191–201.

Lemay, Michael (2009). *Understanding the Mechanism of Panel Attrition*. University of Maryland.

Lipps, Oliver (2007). "Attrition in the Swiss Household Panel". In: *Methoden, Daten, Analysen (mda)* 1.1, pp. 45–68.

Liu, Shujie et al. (2014). "A Recursive Recurrent Neural Network for Statistical Machine Translation". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2014. Association for Computational Linguistics, Baltimore, Maryland, pp. 1491–1500.

Lubba, Carl H. et al. (2019). *catch22: CAnonical Time-series CHaracteristics*. 2019.

Lugtig, Peter (2014). "Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers". In: *Sociological Methods & Research* 43.4, pp. 699–723.

Lynn, Peter (2017). "From standardised to targeted survey procedures for tackling nonresponse and attrition". In: *Survey Research Methods* 11.1. Number: 1, pp. 93–103.

Middlehurst, Matthew, James Large, and Anthony Bagnall (2020). "The Canonical Interval Forest (CIF) Classifier for Time Series Classification". In: *2020 IEEE International Conference on Big Data (Big Data)*. 2020 IEEE International Conference on Big Data (Big Data). IEEE, Atlanta, GA, USA, pp. 188–195.

Mulder, J and N Kieruj (2018). *Preserving Our Precious Respondents: Predicting and Preventing Non-Response and Panel Attrition by Analyzing and Modeling Longitudinal Survey and Paradata Using Data Science Techniques*. 2018.

Olson, Kristen (2013). "Paradata for Nonresponse Adjustment". In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 142–170.

Peytchev, Andy, Daniel Pratt, and Michael Duprey (2022). "Responsive and Adaptive Survey Design: Use of Bias Propensity During Data Collection to Reduce Nonresponse Bias". In: *Journal of Survey Statistics and Methodology* 10.1, pp. 131–148.

Pforr, Klaus and Jette Schröder (2016). "Why Panel Surveys?" In: *GESIS Survey Guidelines*. In collab. with GESIS-Leibniz-Institut Für Sozialwissenschaften. Publisher: SDM-Survey Guidelines (GESIS Leibniz Institute for the Social Sciences) Version Number: 2.0.

Ribeiro, Antônio H et al. (2020). "Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, pp. 2370–2380.

Richter, David, John L. Körtner, and Denise Saßenroth (2014). "Personality has minor effects on panel attrition". In: *Journal of Research in Personality* 53, pp. 31–35.

Roßmann, Joss and Tobias Gummer (2016). "Using Paradata to Predict and Correct for Panel Attrition". In: *Social Science Computer Review* 34.3. Publisher: SAGE Publications Inc, pp. 312–332.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323.6088. Number: 6088 Publisher: Nature Publishing Group, pp. 533–536.

Salman, Afan Galih et al. (2018). "Single Layer & Multi-layer Long Short-Term Memory (LSTM) Model with Intermediate Variables for Weather Forecasting". In: *Procedia Computer Science* 135, pp. 89–98.

Sarker, Iqbal H. (2021). "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions". In: *SN Computer Science* 2.6, p. 420.

Sarndal, Carl-Erik and Peter Lundquist (2014). "Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation". In: *Journal of Survey Statistics and Methodology* 2.4, pp. 361–387.

Seidenberg, Andrew B, Richard P Moser, and Brady T West (2023). "Preferred Reporting Items for Complex Sample Survey Analysis (PRICSSA)". In: *Journal of Survey Statistics and Methodology* 11.4, pp. 743–757.

Shewalkar, Apeksha Nagesh (2018). *Comparison of RNN, LSTM and GRU on Speech Recognition Data*. North Dakota State University.

Siegers, Rainer, Hans Walter Steinhauer, and Lennart Dührsen (2021). "SOEP-Core v36 – Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2019)". In: SOEP Survey Papers 1106: Series C. Berlin: DIW/SOEP.

Struminskaya, Bella and Tobias Gummer (2022). "Risk of Nonresponse Bias and the Length of the Field Period in a Mixed-Mode General Population Panel". In: *Journal of Survey Statistics and Methodology* 10.1, pp. 161–182.

Suresh, Krithika, Cameron Severn, and Debashis Ghosh (2022). "Survival prediction models: an introduction to discrete-time modeling". In: *BMC Medical Research Methodology* 22.1.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Tienda, Marta and Dawn Koffman (2021). "Using Paradata to Evaluate Youth Participation in a Digital Diary Study". In: *Social Science Computer Review* 39.4, pp. 666–686.

Trappmann, Mark, Tobias Gramlich, and Alexander Mosthaf (2015). "The effect of events between waves on panel attrition". In: *Survey Research Methods* 9.1, pp. 31–43.

Uhrig, S C Noah (2008). "The Nature and Causes of Attrition in the British Household Panel Survey". In: *ISER Working Paper Series*.

Voorpostel, Marieke and Oliver Lipps (2011). "Attrition in the Swiss Household Panel: Is Change Associated with Drop-out?" In: *Journal of Official Statistics* 22.2, pp. 301–318.

Wagner, James R (2008). "Adaptive Survey Design to Reduce Nonresponse Bias". In: *University of Michigan*.

Zargar, Sakib (2021). *Introduction to Sequence Learning Models: RNN, LSTM, GRU*. ResearchGate.

Zinn, Sabine and Timo Gnambs (2022). "Analyzing nonresponse in longitudinal surveys using Bayesian additive regression trees: A nonparametric event history analysis". In: *Social Science Computer Review* 40.3, pp. 678–699.

3 Pre-Trained Nonresponse Prediction in Panel Surveys with Machine Learning

Abstract

While predictive modeling for unit nonresponse in panel surveys has been explored in various contexts, it is still under-researched how practitioners can best adopt these techniques. Currently, practitioners need to wait until they accumulate enough data in their panel to train and evaluate their own modeling options. This paper presents a novel "cross-training" technique in which we show that the indicators of nonresponse are so ubiquitous across studies that it is viable to train a model on one panel study and apply it to a different one. The practical benefit of this approach is that newly commencing panels can potentially make better nonresponse predictions in the early waves because these pre-trained models make use of more data. We demonstrate this technique with five panel surveys which encompass a variety of survey designs: the Socio-Economic Panel (SOEP), the German Internet Panel (GIP), the GESIS Panel, the Mannheim Corona Study (MCS), and the Family Demographic Panel (FREDA). We demonstrate that nonresponse history and demographics, paired with tree-based modeling methods, make highly accurate and generalizable predictions across studies, despite differences in panel design. We show how cross-training can effectively predict nonresponse in early panel waves where attrition is typically highest.

3.1 Introduction

Panel surveys are an irreplaceable source of data for social scientists. These surveys require more skilled management and resources than one-time surveys, so sources of error must be controlled as much as possible (Pforr and Schröder, 2016). Nonresponse is one of the more severe sources of survey error, and panel survey managers are increasingly under pressure to ameliorate nonresponse rates (Fuchs, Bossert, and Stukowski, 2013; Luiten, Hox, and Leeuw, 2020). This paper is exclusively concerned with 'unit nonresponse,' that is, when a participant is invited to a panel wave and, for any reason, does not submit any usable data. This variety of nonresponse is distinct from permanent dropout from a panel or item nonresponse.

A promising approach to reducing nonresponse bias is the application of predictive modeling to forecast nonresponse in panel surveys. In this approach, practitioners build models that output each participant's estimated nonresponse propensity. These estimates could then help to target the most at-risk participants with interventions aimed at mitigating their risk of nonresponse (Jacobsen et al., 2021; Jankowsky, Steger, and

Schroeders, 2022; Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022; Mulder and Kieruj, 2018). Machine Learning (ML) is an approach to predictive modeling. In this application, data collected about participants in a panel and their nonresponse history is used to train (i.e., 'fit') an ML model to predict future nonresponse behavior based on historic patterns (Hastie, Tibshirani, and Friedman, 2009; James et al., 2013).

Many research papers have explored this approach to predicting participant nonresponse (Bach, Eckman, and Daikeler, 2020; Cheng, Zamarro, and Orriëns, 2016; JSSAM, 2022; Kern, Klausch, and Kreuter, 2019; Kocar and Biddle, 2022; Mulder and Kieruj, 2018; Olson, 2013; Zinn and Gnambs, 2022). However, most of these papers typically focus on one specific panel study, train a range of prediction models, and compare the results obtained for that panel. What is absent in this literature is an understanding of how well the findings about one panel study transfer to another panel. No two longitudinal studies are alike; they differ widely in their 'survey design,' including characteristics such as the target population, the unit of study (i.e., individuals or household respondents), the mode, topics, and wave frequency. When a particular prediction approach is highly effective in one survey context, it is still an open question whether that technique will also be effective in another context.

Practitioners developing a new panel survey and interested in using predictive modeling are left uncertain as to which modeling approach from the literature to adopt. Practitioners could wait until they accumulate enough survey waves to train various models on their panel and select the best performer for future use. However, this requires the panel to accumulate many panel waves and potentially lose panelists during that time. In this paper, we present the possibility of "cross-training," that is, using data from pre-existing panels to train a nonresponse prediction model and apply it to a new panel. Our underlying assumption is that ML models fitted on nonresponse history and demographic data are consistently effective so that these models can be transported across different contexts. Therefore, our first guiding research question is as follows.

1. **What is the predictive performance of a model trained on one panel but applied to another study?**

Assuming that transporting models is viable, we want to understand when and why these models can (or cannot) be interchanged between panels. For example, suppose nonresponse history and demographics like age and income are the key indicators of nonresponse across many different panels. In that case, it explains how cross-training would be effective because this predictive process is ubiquitous across contexts. Alternatively, cross-training between panels would likely fail if, for example, nonresponse history were more predictive in panels with monthly versus annual waves. To understand when cross-training may or may not be successful, we aim to analyze the consistency of the efficacy of the algorithms and predictors across different panel contexts. Therefore, our second research question is as follows.

2. **Across different survey contexts, is there a difference in what predictors and algorithms are effective in predicting nonresponse?**

We test these questions by gathering data from five different panel surveys in Germany, each with a different survey design: the Socio-Economic Panel (SOEP), the German Internet Panel (GIP), the GESIS Panel, the Mannheim Corona Study (MCS), and the Family Demographic Panel (FREDA). These panels were selected to compare common differences between panel surveys. These studies encompass various sampling methods, recruitment methods, data collection modes, units of study, and wave frequencies. We derive equivalent features (i.e., predictors) across all datasets in each survey. For each dataset, we train a set of models and compare the predictive performance of each model for predicting nonresponse *in each panel*. Our study is the first to systematically (cross-)train and evaluate machine learning models for nonresponse prediction across multiple panels at scale.

Our study design allows us to identify which differences in survey characteristics cause certain algorithms to be specifically effective or reduce performance. We also compare the 'permutation feature importance' (PFI; Altmann et al. (2010)), which measures how much each feature contributes to predictive performance. We then evaluate whether specific features are always helpful for prediction-making in any context or whether certain features are more or less powerful under certain contexts.

We propose a process for exploring how ML models can be trained on one panel and applied to another. However, which algorithms and predictors should we explore? Furthermore, how would we evaluate them? In the Background section (3.2), we establish that previous research has most often considered logistic regression or tree-based models trained on demographic and past-nonresponse behavior data. In the Methods section (3.3), we introduce how we implement those modeling approaches using data from the five panel surveys. We also compare and contrast the design characteristics of each of these five panels. To answer research question one, we introduce a framework for 'cross-training' models, that is, training models on one panel's data and then making nonresponse predictions in a another panel.

To answer research question two, we examine whether certain predictors are more or less important in different panels. In the Results section (3.4), we show that cross-training can accurately predict nonresponse in the second wave of a given panel. The ubiquity of nonresponse history and demographics as effective predictors explains this outcome. In the Discussion section (3.5), we consider what these findings should mean for survey practitioners and the limitations of this research.

3.2 Background

Many studies explore forecasting nonresponse in panel surveys (Bach, Eckman, and Daikeler, 2020; Hill et al., 2020; Jacobsen et al., 2021; Jankowsky, Steger, and Schroeders, 2022; Kern, Weiß, and Kolb, 2021; Kocar and Biddle, 2022; Kreuter and Jäckle, 2008; Lipps, 2007; Lugtig, 2014; McLauchlan and Schonlau, 2016; Minderop and Weiß, 2023; Mulder and Kieruj, 2018; Plewis and Shlomo, 2017; Roßmann and Gummer, 2016; Siegers, Steinhauer, and Dührsen, 2021; Uhrig, 2008; Voorpostel and Lipps, 2011). In this section, we identify that nonresponse history and demographics are often the most

powerful predictors of future nonresponse and that logistic regression and tree-based models are highly successful in many studies. This paper will take the extra step of demonstrating that these features and techniques are consistently effective across contexts and that this is why our proposed cross-training approach is viable.

For this study, we are only interested in discussing research that aims to predict future nonresponse instead of explaining it. Also, we are interested in predicting nonresponse propensity in the next wave, as opposed to other possible prediction units like survival time (Lemay, 2009). This decision is because those units require more waves to assess the outcome, and we are interested in models that can be fitted as early as possible to reflect the survey practitioner's need for timely forecasts.

Of the previous studies that aimed to predict future unit nonresponse, only one paper evaluated several surveys (although only one was a panel study) and systematically compared the results: Bach, Eckman, and Daikeler (2020). In this paper, the authors applied a common set of algorithms across three surveys: the Longitudinal Internet Studies for the Social Sciences (LISS), the Survey on Free Time (SOFT), and the Employment and Purchase Behavior in Germany (EPBG). LISS is a household panel survey with around 5,000 households sampled by geographic clustering across the Netherlands. LISS recruited households by mail, telephone, or face-to-face interviews. Regular surveys about topics concerning internet usage have been conducted online and monthly since 2007. SOFT and EPBG are cross-section surveys. SOFT was a 2013, US-based telephone survey with around 300 household respondents sampled by random selection of ZIP codes from the postal service registry. EPBG was a 2011 telephone survey of 12,400 Germans sampled from the federal administrative labor force records. Each survey collected different data, and the researchers used different covariates across the same set of models. Demographic data was available across all three surveys. The models for LISS used information about previous nonresponse history, while SOFT and EPBG used information from the recruitment process, such as the number of missed invitation calls. The two implemented prediction methods were logistic regression and gradient boosting (tree-based). The gradient-boosted models performed best, with very high Area Under Receiver Operator Curve (AUROC[1]) scores of 0.84 for EPBG, 0.88 for LISS, and 0.94 for SOFT. This study demonstrates the efficacy of tree-based models with demographics and nonresponse history, yielding 0.88 AUROC when forecasting nonresponse in the LISS panel.

Zinn and Gnambs (2022) trained models to predict next-wave nonresponse in the National Educational Panel Study (NEPS). NEPS is a panel survey, with waves running every six to twelve months, starting in 2009. The sample of over 40,000 German residents was drawn through cooperation with educational institutions. In each of the six cohorts recruited since 2009, there is a mixture of newborns, kindergarteners, primary schoolers, high schoolers, post-high schoolers, and post-tertiary adults. Zinn and Gnambs (2022) experimented with two models: Bayesian Additive Regression Trees (BART) and logistic

[1]For a classifier that outputs the probability of a given case belonging to a certain class, AUROC is a metric that measures the trade-off between sensitivity (true positive rate) and specificity (true negative rate). AUROC values range from 0 to 1, where 1.0 represents a perfect classifier, and 0.5 represents random guessing (the worst possible classifier).

regression. Because NEPS is focused on education-related topics, many of the model's features were substantive information like the number of books a child has at home, the number of sick days taken, and demographics like migration background and federal state. Zinn and Gnambs report their results in terms of accuracy (the portion of correct predictions) as 89-99% for the first five waves of NEPS with both models. This study is an example of another successful implementation of tree-based and logistic regression models.

Kocar and Biddle (2022) predicted next-wave nonresponse in the Life in Australia (LIA) panel survey. LIA has run roughly monthly waves since May 2018. The sample was recruited by random digit dialing of registered numbers amongst the general Australian population. Interviews were conducted online. Kocar and Biddle used demographics, past nonresponse behavior, and online paradata such as browser type and page-click behavior. Kocar and Biddle fitted these features with a logistic regression model and achieved a recall score of over 0.9 and a (considerably lower) precision of 0.2. This study also shows the viability of logistic regression with demographic and past-nonresponse predictors.

Mulder and Kieruj (2018) predicted next-wave nonresponse in the LISS panel. They used features such as demographics, past nonresponse, physical/mental health, personality measures, and incentive sizes. Mulder and Kieruj used these features to build various prediction models: logistic regression, support vector machines, random forest, gradient boosted, and neural networks. The resultant Area Under Receiver Operator Curve (AUROC) scores ranged from 0.65 for the neural network to 0.79 for the random forest.

Kern, Weiß, and Kolb (2021) predicted next-wave nonresponse with the GESIS Panel. The GESIS Panel is a general German population panel with online/postal options, and the wave frequency is between two and three months. The authors used demographics, past nonresponse behavior, and rolling-average nonresponse rates with varying window sizes (i.e., average nonresponse over the past two waves, three waves, etc.). These researchers applied these predictors to various models, including logistic regression, random forest, and extra tree classifiers. Over the GESIS Panel waves from late 2013 to mid-2017, these models achieved average AUROC scores ranging from 0.86 with penalized logistic regression to 0.89 with random forest.

These studies show that panel study practitioners are interested in predictive modeling to intervene with at-risk participants preemptively. However, panel studies have different techniques for alleviating nonresponse bias. Numerous studies analyze the characteristics of responders and nonresponders to evaluate the risk of nonresponse bias and the effectiveness of nonresponse weights to mitigate such bias. Some examples of this analysis were carried out under the University of Michigan's Panel Study of Income Dynamics (PSID) (Fitzgerald, Gottschalk, and Moffitt, 1998) and the United Kingdom's Understanding Society panel (Lynn, Cabrera-Álvarez, and Clarke, 2023). Such explanatory (rather than predictive) modeling similarly indicates that a core set of individual characteristics can consistently differentiate between responders and nonresponders: Durrant and Steele (2008) analyze nonresponse in six United Kingdom Government surveys and report that only selected predictor variables (such as self-employment, household type, region) exhibit survey-specific effects while many demographic characteristics are impor-

tant predictors of nonresponse for all six surveys.

This literature review shows that nonresponse history and demographics used in a logistic regression or tree-based model have often been effective in predicting nonresponse. Across these studies, AUROC values in the 0.8-0.9 range have been achievable with these techniques. However, not all implementations were equivalent in that model parameters varied, and the exact method for deriving each variable differed across studies. This paper aims to apply the same technique to various panels to understand which approaches are ubiquitously effective.

3.3 Methods

3.3.1 Data

We selected five panel surveys which cover a range of common panel survey designs. These panels target the general German population but vary widely in other respects. We have a wide range of maturities, with SOEP being a "traditional" and widely used panel study commenced in 1984, whereas FREDA is extremely recent, starting in 2021. The purpose of the surveys varies from FREDA, which is focused on family affairs, to the GESIS Panel, which is an omnibus survey. MCS is focused on the COVID-19 pandemic. Survey modes have been evolving over the past several decades, with face-to-face, phone, mail, and online modes all varying in prominence over time. Throughout its lifetime, SOEP has employed many different survey modes, including face-to-face and mail, compared to the GESIS Panel and FREDA, which focus on mail and online data collection. GIP and MCS are entirely online panels. By comparing these surveys, we evaluate how prediction techniques in one era and with one given study objective can generalize to another context. In the following sections, we describe each panel in detail before summarising their similarities and differences in Table 3.1.

The Socio-Economic Panel (SOEP)

The Socio-Economic Panel (SOEP) is a German general-population household survey (Liebig et al., 2022). SOEP collects data about economic matters, political attitudes, and psychological factors, among other topics. It has been running annually since 1984. In this paper, we follow the initial recruitment intake of 15,000 participants, which has steadily declined to around 2,500 as of 2020 (see Figure (Appendix) 3.5). Over the years, survey modes have included face-to-face, phone, mail, and online. (German Institute for Economic Research (DIW Berlin), 2023; Goebel et al., 2019; Siegers, Steinhauer, and Dührsen, 2021). The initial sampling method selected households by random walks across geographic regions to provide a representative sample of Germany at the time (i.e., pre-reunification). For each household, every resident over the age of 16 was invited to provide an individual response. Also, a 'head of household' provides information about the whole household. SOEP panelists can exit the survey by explicit request, death, or moving abroad.

German Internet Panel (GIP)

The German Internet Panel (GIP) is a general German population survey concerning politics and economics, among other topics (Blom, Gathmann, and Krieger, 2015; Blom, Gonzalez Ocanto, et al., 2022). The panel commenced in 2012 and runs waves every two months. We follow the initial recruitment intake of roughly 1,500 participants (see Figure (Appendix) 3.5). The survey mode is online only. Initial sampling was based on geographic stratified clustering, in which regions of roughly equal populations were selected to be representative of Germany's distribution of federal states and urbanity. German residents aged 16 to 75 were eligible to participate. Participants were recruited by face-to-face interviews, and subsequent waves were conducted online. Households without sufficient internet or computer access were provided with devices and support.

One issue with GIP data is that the published dataset does not include whether participants have asked to exit the panel. As a result, we cannot distinguish between temporary nonresponders and permanent dropouts. In other panels, we can exclude exited participants and analyze only temporary unit nonresponse. This matter has the effect of making the apparent GIP active panel size (the number of participants invited to each wave) stay at roughly 1,500 over time, whereas other panels attrite invitees (see Figure (Appendix) 3.5).

GESIS Panel

The GESIS Panel is an omnibus survey of the general German population, covering topics such as politics, time use, and well-being (GESIS, 2023). It commenced in October 2013 and ran in two-monthly waves until February 2021, when the wave frequency became three-monthly.

We follow the initial recruitment intake, which commenced with roughly 5,000 participants and steadily declined to around 2,500 by 2021 (see Figure (Appendix) 3.5). The survey has two modes: Web (roughly 75%) and mail (Bosnjak et al., 2018; GESIS, 2021; GESIS, 2023). The GESIS Panel's sampling method randomly selected invitees from the German population register. The recruitment criteria allowed German residents between the ages of 18-70 to participate. Recruitment interviews were conducted face-to-face. Panelists exit the study either by explicit request or by nonresponding to three consecutive waves.

There is a peculiarity regarding the GESIS Panel's first two post-recruitment interview waves (waves 3 and 4). Recruitment took many months, but the Panel managers were concerned about losing participants if they were not contacted for a long time. Therefore, only the participants recruited by that time were invited in these early waves. The result is a substantially smaller sample in those early waves (Bosnjak et al., 2018).

Mannheim Corona Study (MCS)

The Mannheim Corona Study (MCS) was a survey of individuals concerning how COVID-19 affected the daily lives of the general German population. The panel ran *weekly* waves for 16 weeks from 20th March to 10th July 2020. All waves were administered online.

The same team managed the Mannheim Corona Study as the GIP, and the participants were a randomly selected subset of GIP participants as of 2020, which was larger than its initial recruitment size of 1,500 (because of additional intakes in 2014 and 2018). Therefore, unlike all other surveys in this study, MCS did not start with a typical recruitment survey because the participants had already been recruited (Blom, Cornesse, et al., 2021). Because the survey only ran for 16 weeks, participants who committed to the study were invited every week. Cases of requests to exit were minimal, and no data is available on those requests. Therefore, the apparent sample size of eligible panelists for MSC, like GIP, stays constant at 4,400 invitees (see Figure (Appendix) 3.5).

The German Family Demography Panel Study (FREDA)

The German Family Demography Panel Study (FREDA) is a panel survey that aims to study family life and relationships (including singles) in Germany (Bujard et al., 2023). The waves are annual, consisting of three sub-waves three months apart each year. Starting in 2021, in each subwave, around 38,000 participants were invited to respond. The modes were online and mail. Initial sampling was random sampling from the population register. German residents between 18-45 years of age were eligible to participate. As of this paper, only the first three sub-waves of data have been published (Federal Institute for Population Research, 2022). Therefore, we can evaluate how predictive techniques perform when applied to a freshly commenced panel survey. Currently, participants who completed the first wave are all invited to the second and third wave, so none have yet exited the panel.

3.3.2 Design Comparison

Table 3.1 summarizes the above panel design aspects. We can see that all panels target the German population. However, they differ in various aspects: One of the most substantial differences is the wave frequency, ranging from annual to weekly waves. We expect that the period between waves would impact the drivers of nonresponse because the frequency leads to very different commitments of time and discipline. Another important consideration when comparing panel surveys is the treatment of the recruitment interviews. Each panel, except for MCS, starts with a recruitment interview, and we can only access data about those participants who responded because those who did not participate did not agree to have their data shared. The result is that when predicting nonresponse in the first post-recruitment wave, the models trained on data from recruitment waves are missing nonresponse history, which we expect to be a very important predictor. This issue needs to be kept in mind when we review our results.

3.3.3 Modeling Setup

Outcome

The dependent variable that each model aims to predict is each participant's nonresponse at the next wave in a given panel. We provide the American Association for

Table 3.1: *Comparison of survey designs. F2F: Face-to-Face, HH: Households, I: Individuals.*

Characteristic	SOEP	GIP	GESIS Panel	MCS	FREDA
Started	1984	2012	2013	2020	2021
Modes	F2F, Phone, Post, Online	Online	Online/Post	Online	Online/Post
Wave Frequency	Annual	Two months	Two-three months	Weekly	Three months
Unit of study	HH/I	HH/I	I	I	Family/singles
Sampling Method	Regionally clustered, multi-stage random samples	Regionally clustered, multi-stage random samples	Probabilistic sample of the German-speaking population	Regionally clustered, multi-stage random samples	Probabilistic sample of the German-resident population
Recruitment age	16+	16-75	18-70	16-75	18-49
Recruitment method	F2F/Phone	F2F	F2F	F2F	Phone/Post
Main Topics	Economics, politics, psychology	Attitudes, politics, economics	Omnibus	COVID-19	Family and relationships

Public Opinion Research (AAPOR) response codes we consider nonresponses in Table (Appendix) 3.5. Where possible, we aim to follow AAPOR's definition of nonresponse 'RR6,' which includes partial responses, failure to make contact, implicit and explicit refusal, and the participant's incapacity or death. However, AAPOR response codes are only available for the GESIS Panel and FREDA. We attempted to derive similar response codes for the SOEP, which predates the AAPOR standard and adopted the system only in later waves. Furthermore, nonresponse in the GIP can only be inferred based on whether a given participant ID is not present in the wave. MCS records only a binary 'participation' variable, so we cannot infer the specific type of nonresponse. See Table (Appendix) 3.5 for the data used to derive nonresponse in each panel.

We filter data only to include members of each panel's first recruitment intake to avoid any effect of sample refreshment. Finally, each survey wave is given an individual date to compare panels over time. We date each survey from the start of the data collection period as many of them do not publish a specific end date of data collection. Figure (Appendix) 3.6 shows the timeline of nonresponse rates at each wave for each panel we are analyzing. Nonresponse in GESIS starts high (20-25%) and falls gradually (10%) as low-propensity participants exit the panel, leaving only "reliable" participants. In GIP and MCS, participants are never removed from the panel for consecutive nonresponses, so the subsequent nonresponse rate climbs over time, from 20% to 40% and 18% to 24%, respectively. SOEP maintains a steady average nonresponse rate between 8-12%, likely because the managers maintain a target response quota and have a year to meet it. Nonresponse rates were 41% and 45% across the second and third FREDA waves.

Predictors

To predict each participant's propensity of nonresponse in the next wave, we input the data we have about each participant as of a given wave into an ML algorithm. We use 'Temporal Cross-Validation,' meaning we iterate over waves in which we predict nonresponse using only data available up until that time (Bergmeir and Benítez, 2012; Kern, Weiß, and Kolb, 2021).

Table 3.2 details the variables we derived from each panel to make predictions. The predictor variables are selected to cover common types of predictors used in past research as long as those covariates can be derived from all five panels of our study. Following previous studies in the literature review, we focus on socio-demographic characteristics and nonresponse history. To account for the concept of survey fatigue (Lugtig, 2014), we additionally include a variable for the number of waves each participant has been invited to thus far. These are variables that all of the panel surveys collect despite their different topics of focus. For each demographic variable, we also include a binary variable indicating missingness. Also, the different panels refresh demographic data at different intervals: GESIS, SOEP, and GIP periodically update demographic data, but FREDA and MCS have such short running times that these variables are, in practice, time-invariant in those cases.

Note that we scale (standardize) each continuous variable using only data available at the time of prediction. This way, our retrospective models are fitted the way they could have been at the time. Table (Appendix) 3.3 shows the descriptive statistics of the unscaled predictor variables across all panels.

Prediction algorithms

In this study, we test prominent models representing the main types of classification algorithms explored in past research.

- **Logistic regression.** Regression models are often successful when classification can be made by additively summarising the effects of the covariates. We evaluate penalized and unpenalized logistic regressions (Le Cessie and Van Houwelingen, 1992; Tibshirani, 1996). Although logistic regression can be specified to account for feature interactions by deliberately building in interaction terms, we use this method with only main effects as we use other algorithms that can algorithmically account for interactions in this study.

- **Random forest.** Tree-based models are often successful in cases where there are complex interactions between variables. A random forest is a set ('ensemble') of decision trees tuned to maximize the homogeneity of cases at the endpoint of each decision path. The final prediction is based on the portion of decision trees that 'vote' for each classification (Breiman, 2001; James et al., 2013).

- **Gradient Boosted Classifier (GBC).** This algorithm is similar to random forest, except that trees are built sequentially rather than independently. Compared

Table 3.2: *Predictors derived for each panel*

Type	Variable	Value range	Description
Socio-demographics	Is Married	0, 1	The respondent positively self-identifies as married.
	Age	0 - infinity	Derived by the survey date and year of birth.
	Household Size	0 - infinity	Count of people residing in the participant's residence.
	Household Income	0 - infinity	Monthly combined income of the participant's household in Euros.
	Personal Income	0 - infinity	Monthly personal income in Euros.
	Is Female	0, 1	The participant indicated a 'female' sex.
	Is Unemployed	0, 1	The participant self-identifies as unemployed. We treat part-time, full-time, and parental leave as employment. Unemployment includes studying, retraining, or being retired.
Response history	Invited Waves	1 - infinity	Count of the number of waves this respondent had ever been invited to.
	Nonresponse This Wave	0, 1	Indicates if the participant did not respond in the current wave.
	Historic Nonresponse Rate	0 - 1	The participant's average nonresponse rate over all of their invited waves.

to random forests, boosting may achieve better performance when predicting nonresponse but needs more careful model tuning because small changes in the ensemble setup can greatly impact the results (Friedman, 2001; James et al., 2013).

For each algorithm, we repeat the training process with different parameter settings. This process is a common part of ML modeling, called 'hyperparameter tuning,' and is intended to discover, by experimentation, which parameters (in this context called 'hyperparameters' [2]) are the best algorithm settings (Feurer and Hutter, 2019). We will trial parameters as described in Table (Appendix) 3.4. For completeness, we present the results of all hyperparameter settings.

Model Comparison

We limit the maximum number of preceding waves used in the training set to avoid long-fitting times and adverse impacts from using training data from too far in the past to be relevant. Therefore, the maximum number of training waves for all panels is up to the ten most recent waves for each test wave. For the GESIS Panel, SOEP, and GIP, we test our models on the second through to the 20th survey wave. This limitation is

[2]Hyperparameters are constant values in a machine learning algorithm that are set before training. Examples include the penalty rate in regularized logistic regression or the choice of a homogeneity measure in a classification tree. Tuning hyperparameters involves repeatedly training the model with different hyperparameter values and comparing outcomes using a specific performance metric to select the best settings.

also to avoid long computation times and also we are concerned with nonresponse in the earliest stages of a panel. We test our models for MCS on the second through to the sixteenth wave, which is all available data. For FREDA, we test on the second wave, as only three waves are currently available. We predict nonresponse in each of the outlined waves and calculate the AUROC, recall, and precision scores. Recall is the proportion of positive cases that the model correctly identifies[3]. Precision is the proportion of predicted positive cases that are true positives[4]. AUROC is a value between 0 and 1, indicating the trade-offs between false positives and false negatives. An AUROC of 0.5 represents the worst possible binary classifier, and 1.0 is the best score.

Further, we compare models through Permutation Feature Importance (PFI)[5]. PFI measures how much a given predictor contributes to a model's predictive performance (Altmann et al., 2010; Oh, 2022; Saarela and Jauhiainen, 2021). Feature importance is calculated by taking a trained model and then scrambling each predictor's values by randomly shuffling values in that column for each predictor in the test data. That test data, with a single scrambled predictor, is inputted into the trained model, and the AUROC score of those predictions is calculated. We repeat this process ten times for each predictor with a different random shuffle. Each predictor's PFI is the average loss in AUROC compared to the original performance in the test dataset. Because the shuffling neutralized the predictive power of the scrambled predictor, the loss in AUROC indicates how much predictive performance is contributed by that predictor. We calculate the PFI for each wave and report the average for each variable across each panel. However, this method is vulnerable to covariation. That means that when two predictors are correlated, withholding one predictor will not substantially reduce AUROC because the same information is still available to the model through the other covariate. This issue means that correlated predictor pairs will have their relative PFI understated. When we examine PFI, we must remember that selected pairs of variables may have their importances understated. Missing value flags, for instance, covary with nonresponse.

3.3.4 Cross-Training

In this paper, we train models on one panel and then use them to make nonresponse predictions in another panel. To make the most use of our data, we conduct cross-training, in which the training panel (i.e., the panel survey used for model training) both predates and post-dates the test panel. For example, we will show how SOEP data

[3] I.e., of those who nonrespond in the next wave, recall is the proportion that was correctly predicted.

[4] I.e., of those who were predicted to nonrespond in the next wave, precision is the proportion that did so.

[5] To measure PFI, we take a fitted machine learning model and input a set of cases from the test set to measure the model's baseline AUROC scores. Next, we repeatedly input the same cases but shuffle the values of a given predictor column, effectively removing that feature's predictive power, and measure the average AUROC scores with these "scrambled" test sets. The PFI is then calculated as the difference between the baseline AUROC score and the average AUROC across these repetitions. A higher PFI value indicates a greater drop in AUROC score when the given predictor is withheld from the model.

from 1985 can predict nonresponse in GIP in 2012 but also apply a model trained on GIP's 2012 data to predict nonresponse in SOEP in 1985. We provide two different methods for deriving the training data, and evaluate both. These two methods are as follows. Figure 3.1 provides an illustrated example of each method.

Latest Data Available

In this method, for each wave in the test panel (i.e., the panel study used for model evaluation), we train a model on a fixed number of waves (in our case, five) from the training panel, which precede the start date of that target wave. For example, the first wave of the GESIS Panel was administered in 2013. We can train a model on five SOEP waves from 2007 to 2012 and then use that model to predict nonresponse in the first GESIS Panel wave in 2013. This method aims to train a model using waves that are close to the target wave in time because we expect the contexts to be most similar when they are close together in time.

Note that we can only report results with this method for cases where the training waves predate the target wave. Therefore, we cannot, for example, predict nonresponse in the starting waves of SOEP with GESIS Panel data. Instead, we predict later SOEP waves once data from other panels becomes available. Also, we do not cross-train between GIP and MCS because they are drawn from a common set of individuals.

Equivalent In Lifecycle

For each wave we predict nonresponse in the target panel, we train a model on all waves in the training panel available at the equivalent point in the survey's lifetime. For example, the fourth GESIS Panel survey wave takes place 12 months after the first survey wave. We thus can train a model on SOEP data using waves that took place up to 12 months from the start of SOEP (which would be only the first SOEP wave because it is an annual survey). Conversely, the third SOEP survey wave takes place 24 months after the start of SOEP, and we can predict nonresponse in this wave using a model trained on the 12 GESIS Panel waves that took place within 24 months of the start of the GESIS Panel. This cross-training approach aims to compare equivalent periods in the survey's lifetime by, for example, applying a model trained on the early period of one panel to the equivalent period of the other panel.

3.4 Results

3.4.1 Model Comparison

We commence our results analysis by establishing a baseline of prediction models' performances. Figure 3.2 shows the performance results from training each model type with data of the same panel study, using information available as of each given target wave starting from the second wave of each survey. In the early waves, AUROC is rather low (<0.8) across all panels except MCS, the second wave of the SOEP (in which there was

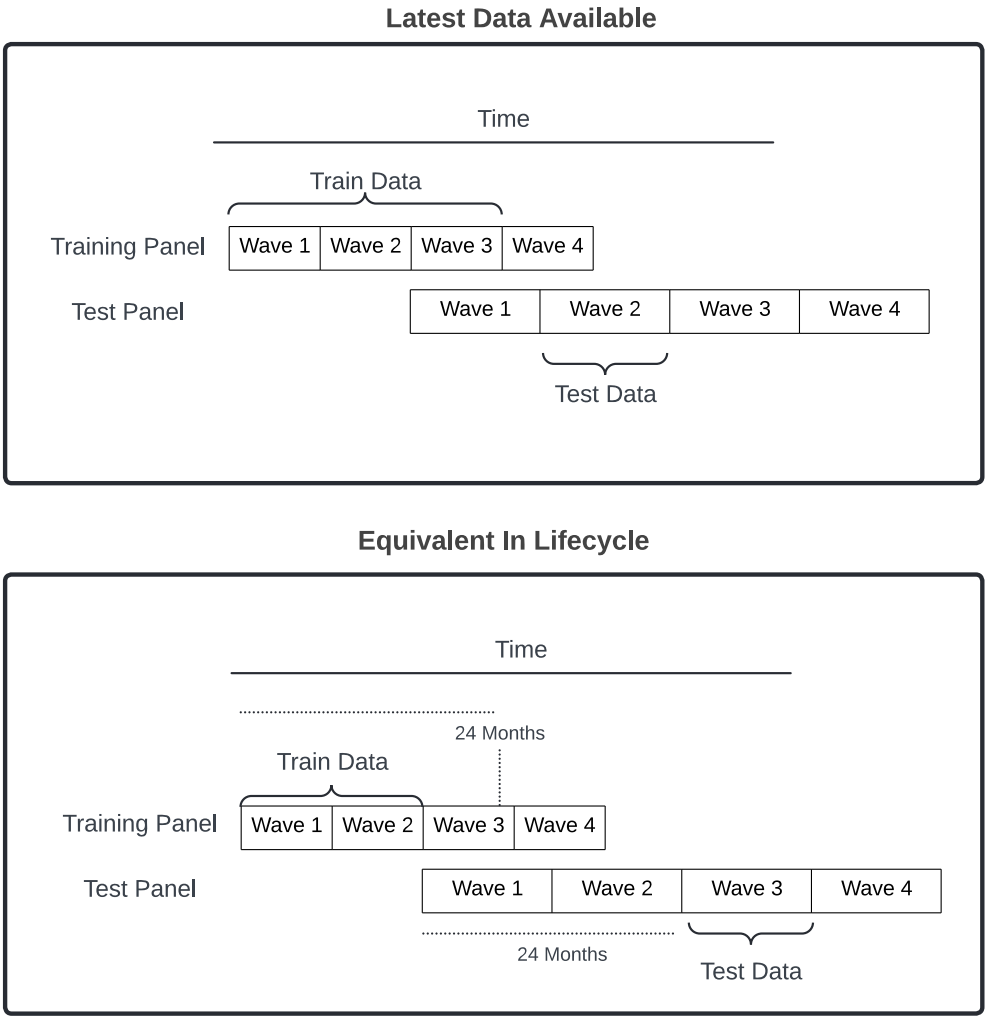


Figure 3.1: *Examples of the two methods for cross-training models. Here, we have a training and a test panel, each with different lengths of time between waves. In the 'latest data available' approach, we can train a model using only data available by the start of the test wave. Because wave four of our training panel was still in its fieldwork period when the test wave started, we can only use data from up to wave three of the training panel. In the 'equivalent in lifecycle' approach, we calculate that our test wave commenced 24 months into the panel's lifetime. Therefore, we train a model on any data that was available in the training panel within 24 months of its respective lifecycle. Because wave three of the training panel was still in fieldwork as of 24 months, we use up to wave two for the training data.*

substantial nonresponse, making it easy to anticipate correctly), and the random forest models in the GESIS Panel. This limited early performance may be because insufficient training data had accumulated at that point to build effective models.

In addition, the recruitment waves may be detrimental as training data. Each panel, except for MCS, starts with a recruitment interview, and the GESIS Panel commences with a two-stage recruitment (the same participants are interviewed across two recruitment waves). Predicting nonresponse following a recruitment interview is a fundamentally different process than predicting nonresponse from a regular panel wave, and models trained on the former might not reliably predict the latter. However, once the recruitment waves are over and more training data accumulates, we can see substantial improvement in AUROC across all panels.

Comparing performance across panels (Figure 3.2), there is a trend that shows higher prediction performance for panel studies with more frequent panel waves. SOEP, an annual survey, has an average AUROC of just below 0.8, while GIP and the GESIS Panel, both two-monthly surveys in this period, are around 0.9 and 0.8, respectively. MCS, which features weekly surveys, exceeds 0.9. Predicting nonresponse in FREDA, which, as of writing, has only one recruitment wave and one regular wave for which we know the dependent variable values, performs poorly at a high score of 0.6.

Each type of model performs equally well. Aside from the results for MCS, however, tree-based models perform better than logistic regression in the earliest one or two waves of a panel study. This outcome indicates that flexible models have a slight advantage in early waves, but in later waves and established panels, main effects models may be sufficient to achieve good performance. Equivalent figures providing recall and precision scores are provided in Figures (Appendix) 3.7 and (Appendix) 3.8.

Figure 3.2: *Comparing model performance across panels where each model is trained using up to 10 preceding waves of the same panel. Auras around the lines are the range of scores across different hyperparameter values. However, models with different hyperparameter settings have such close values that these auras are hardly visible. In the FREDA survey, we can only predict nonresponse in wave three based on the data from wave two, with a model trained on wave one. In that wave, all models achieved roughly 0.6 AUROC.*

Figure 3.3: *Heatmap comparison of permutation feature importances across panels.*

Figure 3.3 shows which features were most predictive across different panels. As survey wave frequency increases (GIP, GESIS Panel, MCS), historic nonresponse becomes more important for all models. As survey waves become less frequent (SOEP, FREDA), demographic features become relatively more important. However, nonresponse history remains important across all models and panels, except for FREDA (because data on nonresponders in the first recruitment interview is omitted, meaning there is no nonresponse history to exploit in the first training wave).

Survey mode has little impact. GESIS Panel is a mixed mode panel study, while GIP and MCS are both online-only, yet they all have similar feature importance profiles. Age is a relevant predictor across all panels, although often more important to random forest models, indicating that age may have an interactive or non-linear effect. This outcome corresponds to other research, which shows that very young and very old participants are particularly at-risk groups for nonresponse (Lipps, 2009).

From this analysis, we can address our second research question. Nonresponse history and demographics are ubiquitously effective across all of the panels analyzed in this paper. AUROC scores after the first few waves of data had accumulated converged across all panels at around 0.75-0.85. Tree-based models are usually better than logistic regression, but logistic regression is often almost as good and sometimes slightly better.

3.4.2 Model Cross-Training

Figure 3.4 shows the result of training nonresponse prediction models on each of our five panel studies and applying them to the GESIS Panel. The results of all other cross-training exercises are detailed in the Appendix section 3.6.3. Except for MCS as the target panel, all cross-trained models start with low AUROC when applied to predict next-wave nonresponse in the respective first wave of a different panel, with a high of 0.65 when nonresponse in SOEP is predicted with a model trained on GESIS Panel data (Figure (Appendix) 3.10). However, when predicting next-wave nonresponse based on data from the second wave, for all models except those trained on FREDA data, the performance of the cross-trained models is often the same or better than the baseline models' performances (which use training data from the same panel). The results show that pre-trained models can achieve AUROC values over 0.75. This performance is seen when nonresponse in the GIP is predicted with models trained on SOEP or GESIS Panel data (Figure (Appendix) 3.11); when models predict nonresponse in the GESIS Panel trained on SOEP, GIP or MCS (Figure 3.4); when nonresponse in the MCS is predicted based on models trained on SOEP or GESIS Panel data (Figure (Appendix) 3.12); or when nonresponse in FREDA is predicted by models trained on any other panel (Table (Appendix) 3.7 and (Appendix) 3.8).

Critically, when cross-trained models predict next-wave nonresponse based on the second wave of a target panel, using the 'Latest Data Available' approach, AUROC was always the same or higher than the baseline approach. Also, although the baseline approach could have been conducted in practice, it would have required training the model as soon as the data collection period ended for a given wave and applied immediately to the next wave, which is potentially a short time window. The pre-trained model could

have been ready beforehand, and predictions about participants could have been made as their responses became available. Using a pre-trained model could be a valuable innovation for newly commencing panels. The strong performance of pre-trained models is likely because they benefit from more training data than the baseline approach.

However, not all cross-training applications are successful. Firstly, using training data that was available as close as possible to the date of the test wave ('Latest Data Available' method) was much more successful than using the 'Equivalent in Lifecycle' method. This result implies a temporal effect, such that training data is more effective when it is closer by date to the target wave, even when the training data is from a different panel. Models trained on FREDA data often performed poorly, likely because of limited training data. Pre-trained tree-based models outperformed logistic regression models on average across all panels, indicating that flexible models have advantages over main effect models in this context.

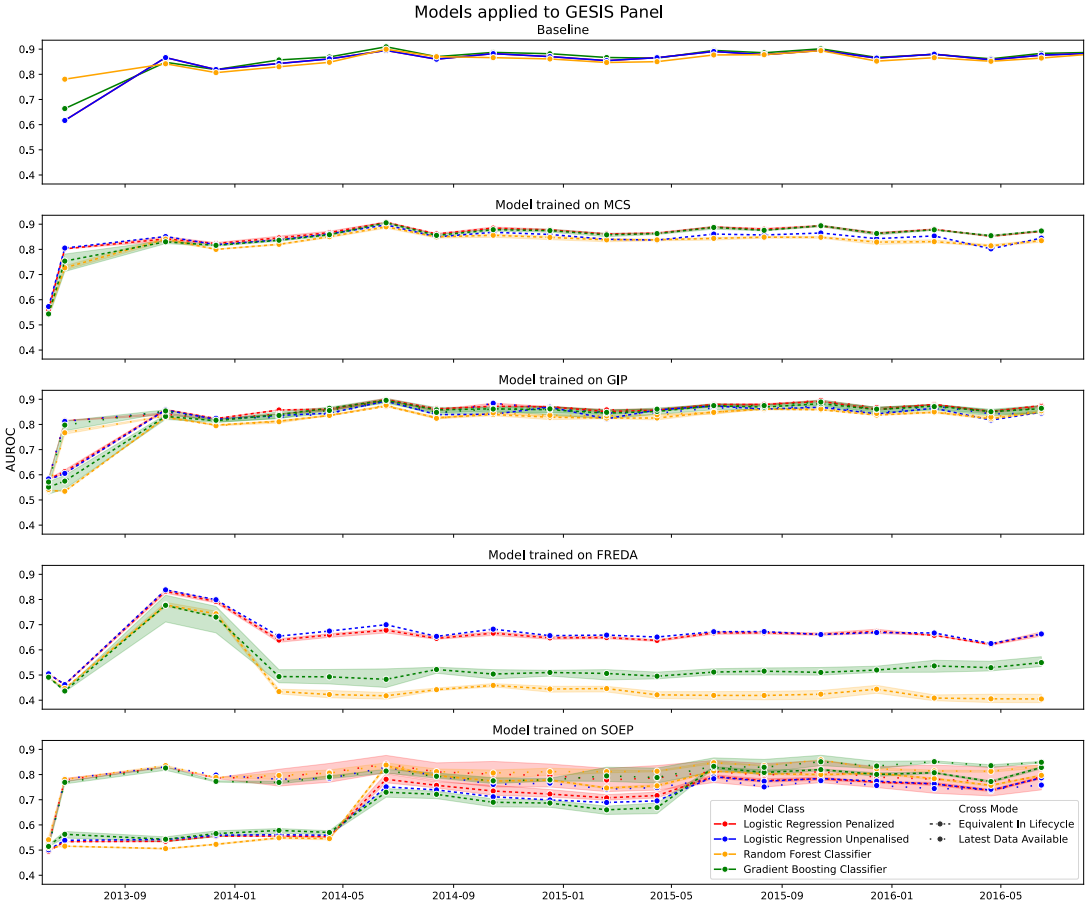


Figure 3.4: *Models trained on other surveys but applied to the GESIS Panel. The 'Baseline' subplot shows performance results when models are trained using training data of the same panel as the target wave. Auras around the lines indicate the range of performance values across different hyperparameter settings.*

3.5 Discussion

This paper presents the first demonstration of 'cross-training' for nonresponse prediction in panel surveys. We show that predictors of nonresponse are so consistently effective across diverse contexts that it is possible to predict nonresponse effectively in the second wave of a panel study using models trained with data from a different panel. This finding is important because a pre-trained model would be available to make predictions sooner for waves one and two than a panel-specific model, which can only be developed once the required training data is available. This timeliness can be critical in the early waves of a panel study, where attrition is often highest.

However, not all applications of cross-training were successful, with some cases performing worse than baseline models and exhibiting low-performance scores. Predicting

the very first wave with a pre-trained the model was generally unsuccessful. However, in the second wave, nonresponse can be predicted accurately, with AUROC scores of 0.75 to 0.85, and pre-training can outperform baseline models as they 'borrow' training data from multiple waves of another panel study.

The main limitation of this research is the number of panels we could compare. A considerable effort is required to process the raw survey data of multiple panels into a common set of features. Including more panel surveys would risk certain surveys not collecting all the same features. Another limitation of this study is that each panel we compare aims to study the general German population. This limitation means we could not compare the effect of different population frames.

The main contribution of this study has been to show that the processes driving panel nonresponse can be very similar between panels despite different survey designs. In our comparisons, only the frequency of survey waves stood out as a factor that influences nonresponse predictability, such that more frequent panels are more predictable. Overall, our findings imply that modeling techniques proven effective in one panel should interest managers of similar panels when deciding their modeling approach. Also, it is possible to pre-train models on one survey and apply them to another with high predictive accuracy. This novel technique could allow survey managers to target and intervene with low-propensity participants in the earliest, most critical waves of a panel study, thereby reducing attrition.

How should panel managers commencing a new panel make use of pre-trained nonresponse models? The suggested method, based on this paper's results, is as follows. Firstly, the best type of panel to use as pre-training data is one that targets the same population of interest. It is also beneficial to use training data that was collected close in time to the target waves. In such cases, fit the model to predict next-wave nonresponse using up to five waves that commenced closest to the start date of the new panel. During the first field period of the new panel, the pre-trained model will not make accurate predictions about who will nonrespond in wave two, so attempting to do this is not recommended. Instead, during the second field period, as responses come in, the model can be used to estimate nonresponse propensity for each participant based on their behavior in the first and second waves.

3.6 Appendices

3.6.1 Software and replication

Code for replication, including instructions on how to import the panel data and run the code, is available at the following URL and on the website of Survey Research Methods. Instructions for accessing the necessary data is detailed in the ReadMe.md file in the replication documents.

`https://osf.io/n4y6w/?view_only=18eb6d46900e4c7d84175042072ff1eb`

The data used in this study for each panel is referenced in the bibliography and cited as

follows: The Socio-Economic Panel (SOEP) (Liebig et al., 2022), German Internet Panel (GIP) (Blom, Gonzalez Ocanto, et al., 2022), GESIS Panel (GESIS, 2023), Mannheim Corona Study (MCS) (Blom, Cornesse, et al., 2021), German Family Demography Panel Study (FREDA) (Bujard et al., 2023).

3.6.2 Supplementary items

This section provides additional details about this study. We provide descriptive statistics about each of the panel survey datasets (Figure (Appendix) 3.5, Figure (Appendix) 3.6, Table (Appendix) 3.3); details about the modeling (Table (Appendix) 3.4), details about our definition of nonresponse (Table (Appendix) 3.5), a data quality checklist (Table (Appendix) 3.6); and further results (Figure (Appendix) 3.7, Figure (Appendix) 3.8, Figure (Appendix) 3.9, Figure (Appendix) 3.10, Figure (Appendix) 3.11, Figure (Appendix) 3.12, Table (Appendix) 3.7, Table (Appendix) 3.8).

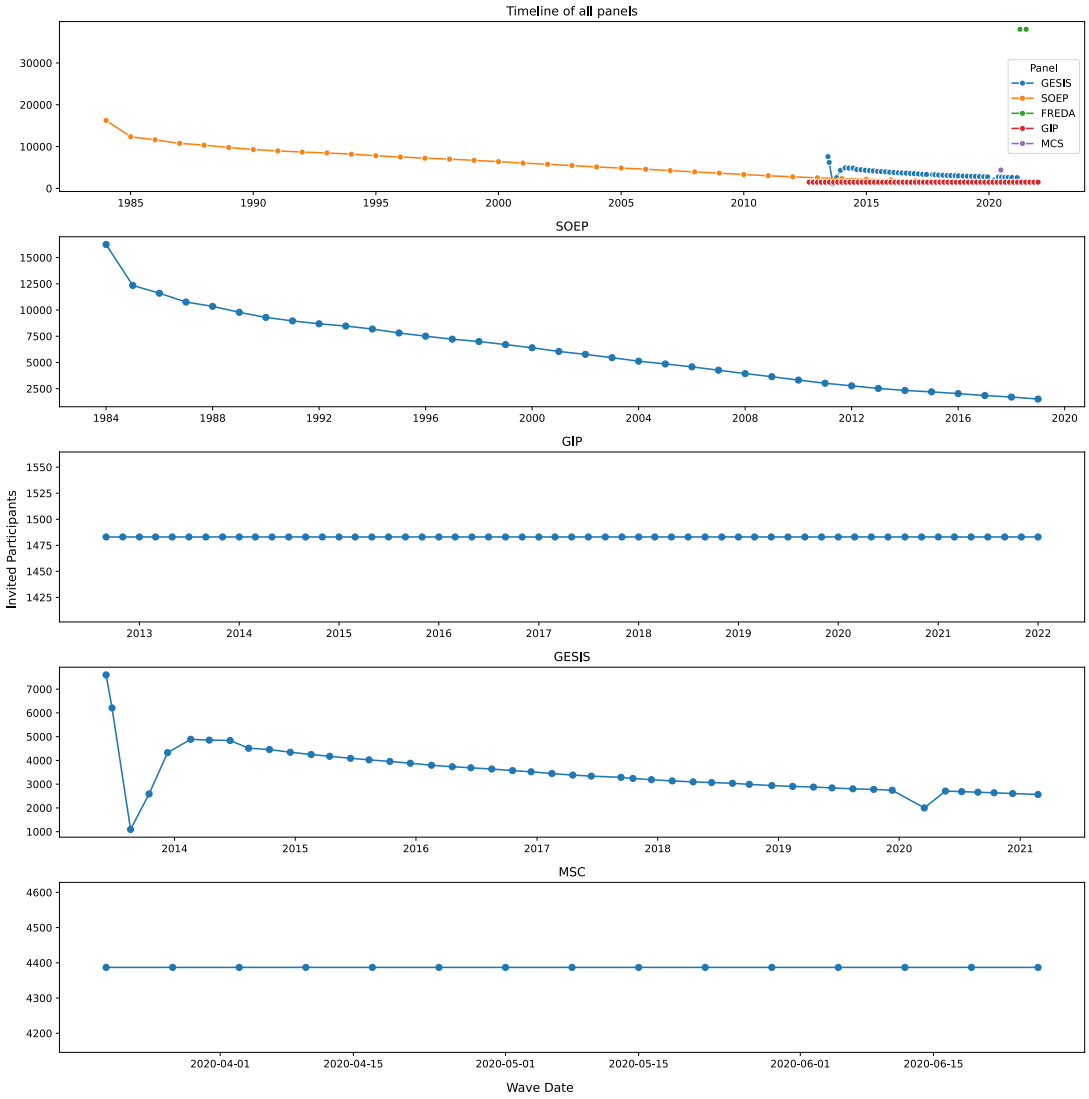


Figure (Appendix) 3.5: *Timeline of the number of invited participants for each panel. Note that we include only those participants who were invited as of the first wave, so these values do not include any participants recruited since then. FREDA had only accumulated three waves by the time of this study, and 38,056 individuals were invited to each wave.*

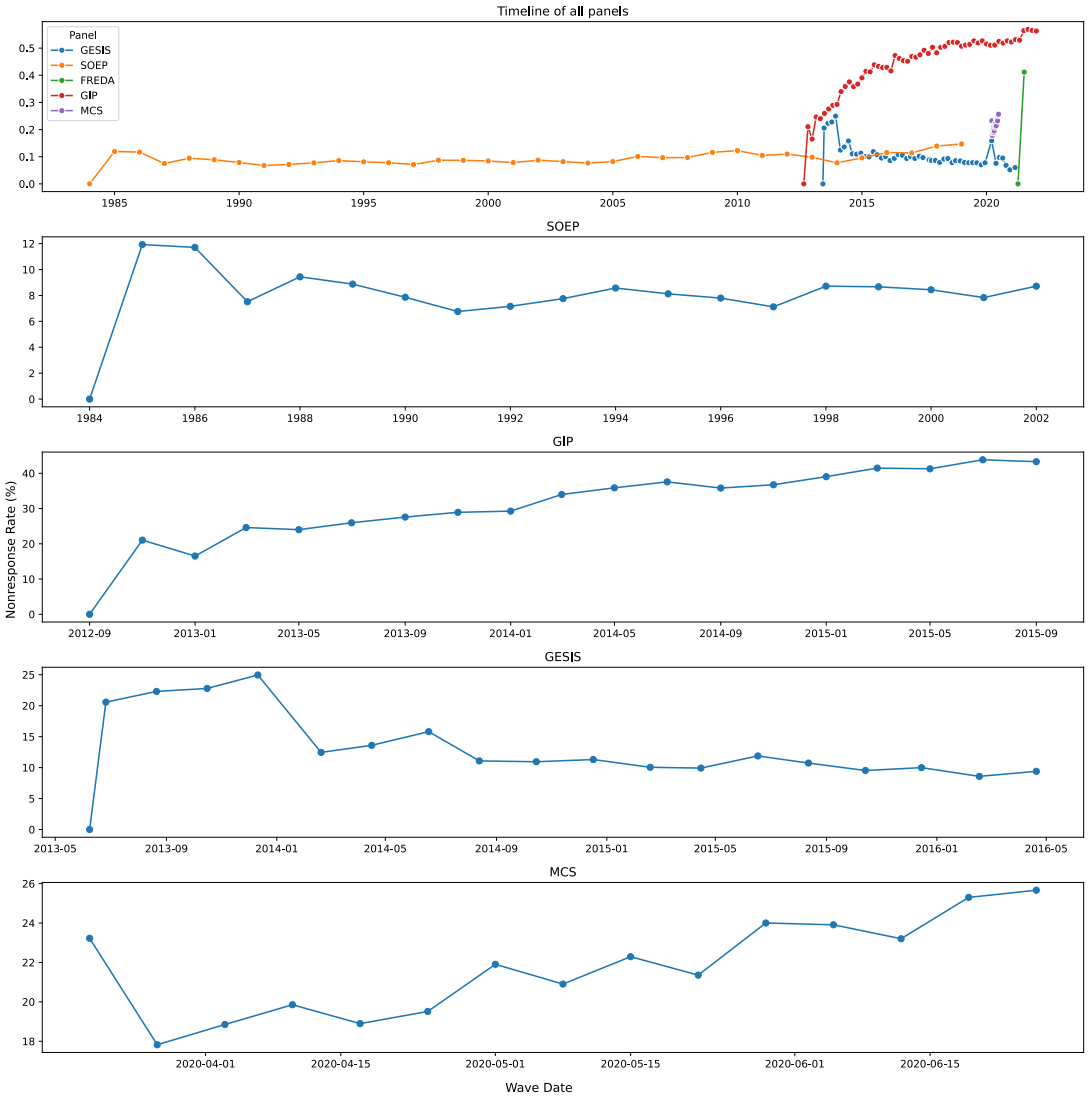


Figure (Appendix) 3.6: *Timelines of each of the panels. FREDA is not included because only the first two waves are included in our analysis. The first wave has a nonresponse rate of zero because no nonrespondent data is retained. Nonresponse rates were 41% and 45% across the second and third FREDA waves.*

Table (Appendix) 3.3: *Distributions of predictive features across each panel*

Variable	Value	SOEP	GIP	GESIS	MCS	FREDA
Age	mean	46.380	52.058	49.416	51.661	33.418
Age	std	18.409	15.605	14.632	15.862	10.161
Age	min	0.000	0.000	0.000	0.000	0.000
Age	max	102.000	87.000	78.000	85.000	68.000
Household Size	mean	2.957	2.530	2.622	2.354	2.897
Household Size	std	1.454	1.140	1.152	1.111	1.438
Household Size	min	1.000	0.000	0.000	0.000	0.000
Household Size	max	17.000	6.000	5.000	6.000	20.000
Household Income	mean	1,224.505	2,277.657	1,989.130	2,609.574	999.193
Household Income	std	1,658.368	1,784.550	1,624.572	2,043.399	2,765.467
Household Income	min	0.000	0.000	0.000	0.000	0.000
Household Income	max	29,000.000	7,500.000	6,000.000	7,500.000	250,000.000
Personal Income	mean	1,396.277	1,468.971	1,498.441	1,752.576	0.000
Personal Income	std	1,491.114	1,309.104	1,149.260	1,391.522	0.000
Personal Income	min	0.000	0.000	0.000	0.000	0.000
Personal Income	max	51,128.000	7,500.000	5,000.000	7,500.000	0.000
Invited Waves	mean	11.984	29.000	20.455	8.000	1.500
Invited Waves	std	8.984	16.452	13.529	4.321	0.500
Invited Waves	min	1.000	1.000	1.000	1.000	1.000
Invited Waves	max	36.000	57.000	48.000	15.000	2.000
Nonresponse This Wave	mean	0.084	0.438	0.104	0.218	0.206
Historic Nonresponse Rate	mean	0.025	0.330	0.059	0.207	0.103
Historic Nonresponse Rate	std	0.081	0.363	0.122	0.315	0.202
Historic Nonresponse Rate	min	0.000	0.000	0.000	0.000	0.000
Historic Nonresponse Rate	max	0.857	0.982	0.857	1.000	0.500
Is Married	mean	0.581	0.101	0.618	0.099	0.124
Missing Is Married	mean	0.000	0.005	0.000	0.002	0.000
Sex Female	mean	0.511	0.498	0.518	0.486	0.431
Missing Sex Female	mean	0.000	0.000	0.000	0.002	0.000
Is Unemployed	mean	0.395	0.355	0.305	0.333	0.017
Missing Is Unemployed	mean	0.000	0.006	0.001	0.017	0.000
Missing Age	mean	0.000	0.000	0.000	0.002	0.022
Missing Household Size	mean	0.000	0.007	0.000	0.028	0.020
Missing Household Income	mean	0.516	0.227	0.066	0.247	0.752
Missing Personal Income	mean	0.232	0.092	0.000	0.067	0.000
Missing Employment Status	mean	0.000	0.000	0.000	0.016	0.012

Table (Appendix) 3.4: *Parameters we hypertune in the fitting process. "N settings" refers to the number of different settings for each hyperparameter. LBFGS: Limited-memory Broyden–Fletcher–Goldfarb–Shanno.*

Model Type	Hyperparameter	Values	*N settings*
Logistic Regression	Penalty	L1, L2 Regularization, No Penalty	5
	Optimization solver	Liblinear for Penalized, LBFGS for Unpenalized	
	Fitting stopping tolerance	0.0001	
	C (applies to penalized)	0.5, 1	
Random Forest	Number of trees in the forest	50, 100, 500	3
	Function to measure split quality	Gini impurity	
	Minimum samples for a split	2	
	Minimum samples for a leaf	1	
	Number of features considered at each split	Square root of all features	
Gradient Boosted Classifier	Number of trees in the forest	50, 100, 500	3
	Function to measure split quality	Gini impurity	
	Minimum samples for a split	2	
	Minimum samples for a leaf	1	
	Number of features considered at each split	Square root of all features	

Table (Appendix) 3.5: *For each panel, these are the types of responses or other information used to define a given case as a nonresponse.*

Panel	Nonresponse if coded as
SOEP	Currently not available Cannot be found Explicit Refusal Currently not available Cannot be found Deceased
GIP	Implied when no response data for that participant is published
GESIS Panel	Nothing ever returned Explicit refusal Post: Attempted - Addressee not known at place of address Break-off: questionnaire too incomplete to process / break-off or partial with insufficient information Explicit refusal with incentive Known respondent-level refusal Logged on to survey did not complete any items Blank questionnaire mailed back implicit refusal Postal box full Implicit refusal Email Bouncer: Mailbox unknown Other person refusal Email Bouncer: Postbox full Death (including Post: Deceased) Email Bouncer: Delivery problem Physically or mentally unable/incompetent Post: Moved left no address Blank questionnaire with incentive returned Respondent language problem Explicit refusal no incentive Post: Undeliverable as addressed Post: No Mail Receptacle Refusal Blank questionnaire with no incentive returned Returned from an unsampled person Invitation returned undelivered (Email Bouncer)
MCS	Binary response/nonresponse variable
FREDA	No response Moved unknown Refused Not surveyable/deceased/permanently ill/not surveyable during field time

Table (Appendix) 3.6: *PRICSSA Checklist (Seidenberg et al. 2023).*

PRICSSA item	Description	Response
1.1 Data collection dates	Describe the survey's data collection dates (e.g., range) to provide historical context that could affect survey responses and nonresponse.	See Figure (Appendix) 3.5 and Figure (Appendix) 3.6.
1.2 Data collection mode(s)	Describe the survey's data collection mode(s). Data collection mode can affect survey responses (e.g., to sensitive questions), including nonresponse, and a survey's data collection mode may change over time (e.g., during the COVID-19 pandemic).	See Section 3.3.1.
1.3 Target population	State the target population the survey was designed to represent and describe all weighted estimates with respect to this target population.	See Table 3.1 and Section 3.3.1. We use only unweighted data.
1.4 Sample design	Describe the survey's sample design, including information about stratification, cluster sampling, and unequal probabilities of selection.	See Table 3.1 and Section 3.3.1.
1.5 Survey response rate(s)	State the survey's response rate and how it was calculated.	See Figure (Appendix) 3.6 and Table (Appendix) 3.5.
2.1 Missingness rates	Report rates of missingness for variables of interest and models, and describe any methods (if any) for dealing with missing data (e.g., multiple imputation).	See Table (Appendix) 3.3.
2.2 Observation deletion	State whether any observations were deleted from the dataset. If observations were deleted, provide a justification. Note: It is best practice to avoid deleting cases and use available subpopulation analysis commands no matter what variance estimation method is used.	We included only cases from the first recruitment wave to avoid any impact on model results caused by the introduction of fresh participants to the training data.
2.3 Sample sizes	Include unweighted sample sizes for all weighted estimates.	See Figure (Appendix) 3.5.
2.4 Confidence intervals/standard errors	Include confidence intervals or standard errors when reporting all estimates to inform the reliability/precision of each estimate.	Significance tests are not applicable to our models, but instead, we provide predictive performance metrics (See Section 3.4).
2.5 Weighting	State which analyses were weighted and specify which weight variables were used in analysis.	Not applicable.
2.6 Variance estimation	Describe the variance estimation method used in the analysis and specify which design variables (e.g., PSU/stratum, replicate weights) were used.	Not applicable.
2.7 Subpopulation analysis	Describe the procedures used for conducting subpopulation analyses (e.g., Stata's "subpop" command, SAS's "domain" command).	Not applicable.
2.8 Suppression rules	State whether or not a suppression rule was followed (e.g., minimum sample size or relative standard error).	Not applicable.

2.9 Software and code	Report which statistical software was used, comprehensively describe data management and analysis in the manuscript, and provide all statistical software code.	See Section 3.6.1.
2.10 Singleton problem (as needed)	Taylor Series Linearization requires at least two PSUs per stratum for variance estimation. Sometimes an analysis is being performed and there is only a single PSU in a stratum. There are several possible fixes to this problem, which should be detailed if the singleton problem is encountered.	Not applicable.
2.11 Public/restricted data (as needed)	If applicable, state whether the public use or restricted version of the dataset was analyzed.	See Section 3.6.1.
2.12 Embedded experiments (as needed)	If applicable, provide information about split sample embedded experiments (e.g., mode of data collection or varying participant incentives) and detail whether experimental factors were accounted for in the analyses.	Not applicable.

3.6.3 Additional Results

Figure (Appendix) 3.7: *Model performance over time, but with Recall instead of AUROC.*

Figure (Appendix) 3.8: *Model performance over time, but with Precision instead of AUROC.*

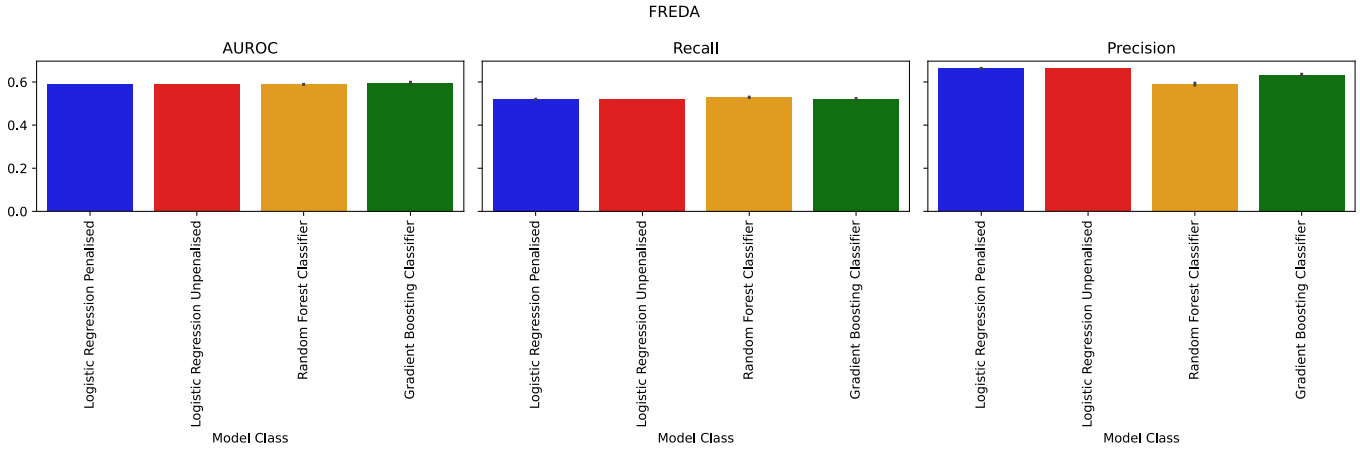


Figure (Appendix) 3.9: *Performance metrics for the second wave of FREDA for which we can make predictions with a model trained on the first FEDA wave.*

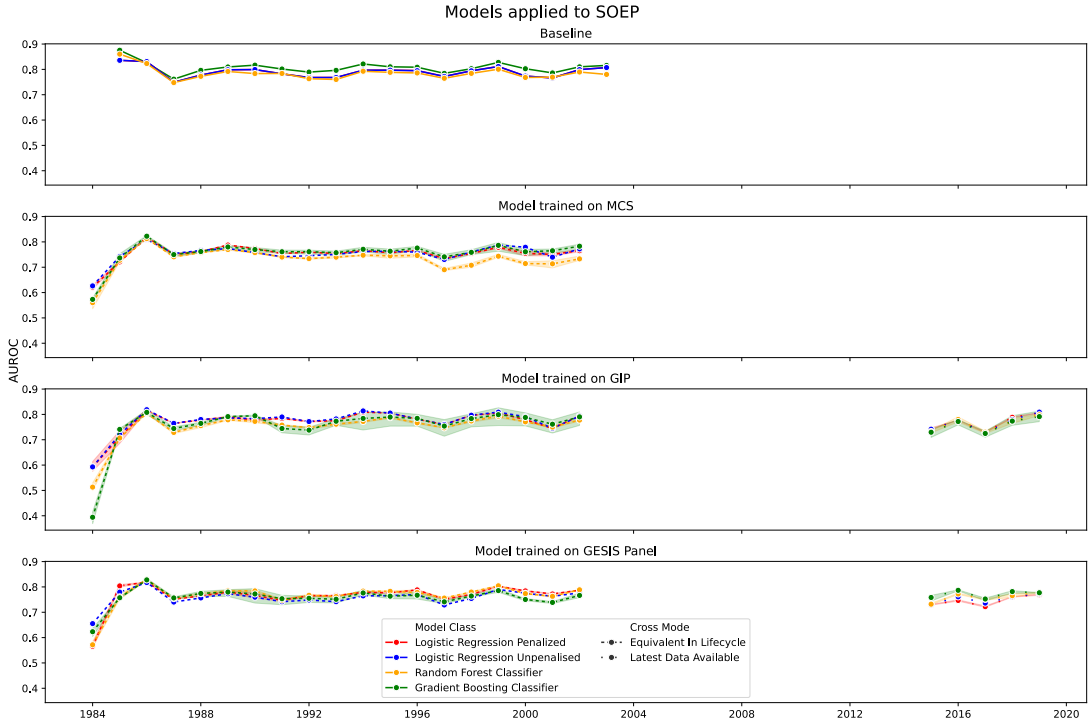


Figure (Appendix) 3.10: *Models trained on other surveys but applied to the SOEP Panel. The 'Baseline' subplot shows performance results when models are trained using training data of the same panel as the target wave. Auras around the lines indicate the range of performance values across different hyperparameter settings.*

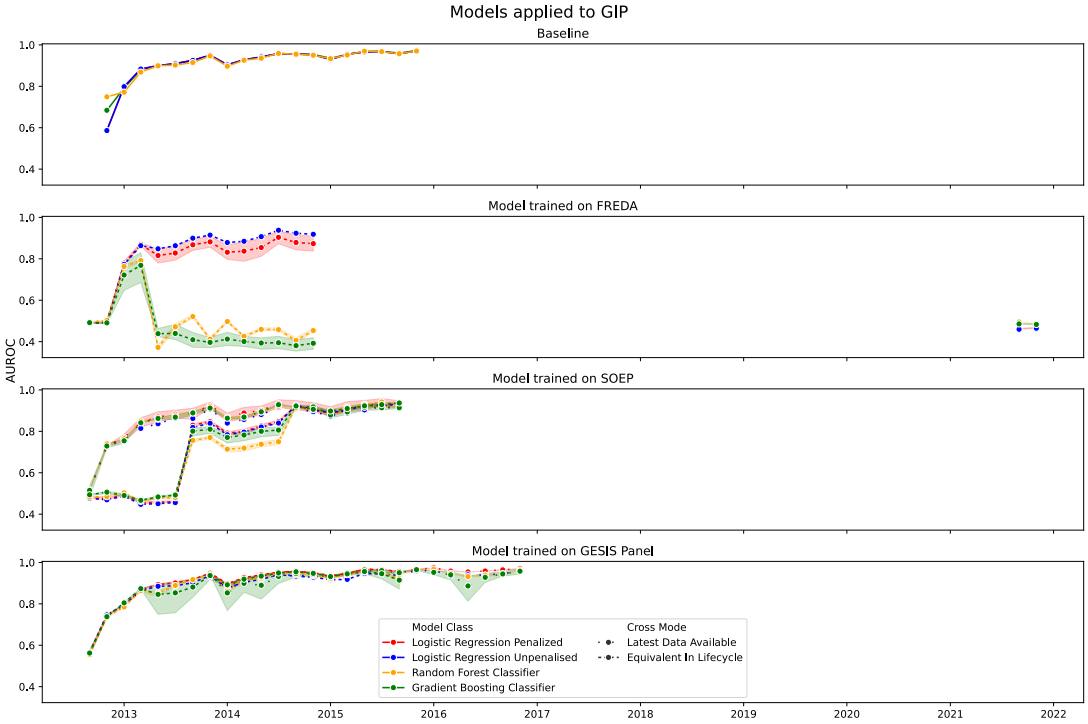


Figure (Appendix) 3.11: *Models trained on other surveys but applied to the GIP Panel. The 'Baseline' subplot shows performance results when models are trained using training data of the same panel as the target wave. Auras around the lines indicate the range of performance values across different hyperparameter settings.*

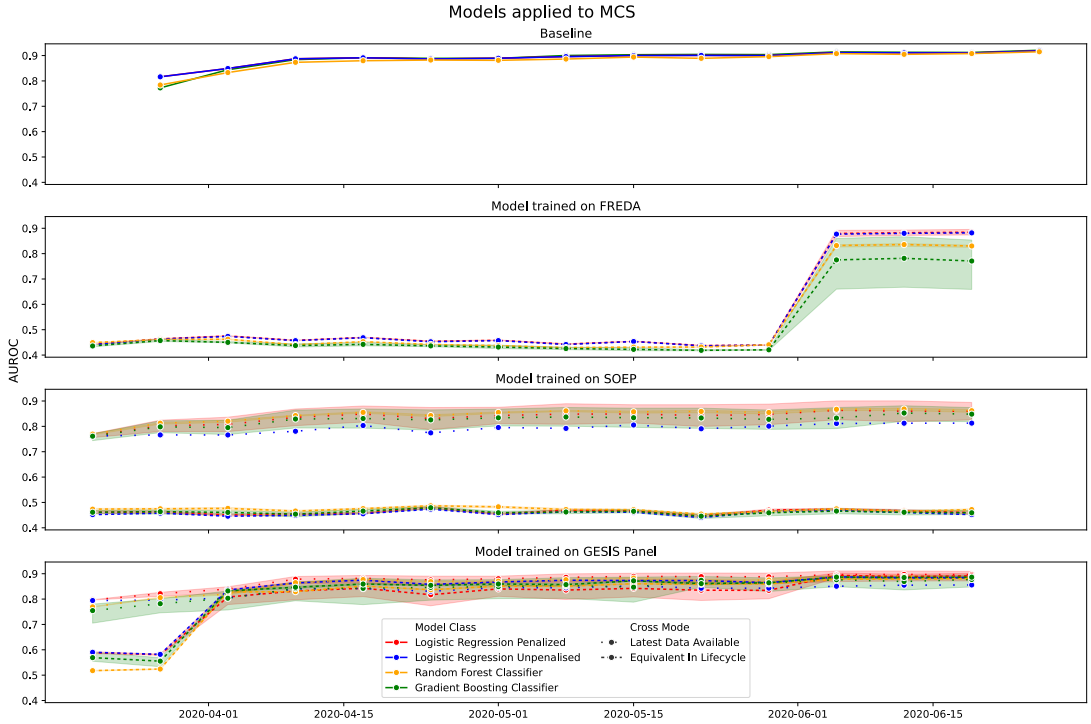


Figure (Appendix) 3.12: *Models trained on other surveys but applied to the MCS. The 'Baseline' subplot shows performance results when models are trained using training data of the same panel as the target wave. Auras around the lines indicate the range of performance values across different hyperparameter settings.*

Table (Appendix) 3.7: *Models trained on other surveys but applied to FREDA Panel. Part one: Latest Data Available method.*

Cross Mode	Test Wave	Model Class	Train Data	AUROC
Baseline	7/07/2021	Gradient Boosting Classifier	-	0.60
		Logistic Regression Penalised	-	0.59
		Logistic Regression Unpenalised	-	0.59
		Random Forest Classifier	-	0.59
Equivalent In Lifecycle	7/04/2021	Gradient Boosting Classifier	GESIS Panel	0.51
			GIP	0.49
			MCS	0.50
			SOEP	0.53
		Logistic Regression Penalized	GESIS Panel	0.53
			GIP	0.52
			MCS	0.49
			SOEP	0.51
		Logistic Regression Unpenalised	GESIS Panel	0.49
			GIP	0.52
			MCS	0.50
			SOEP	0.51
		Random Forest Classifier	GESIS Panel	0.51
			GIP	0.51
			MCS	0.51
			SOEP	0.54
	7/07/2021	Gradient Boosting Classifier	GESIS Panel	0.87
			GIP	0.80
			MCS	0.86
			SOEP	0.74
		Logistic Regression Penalized	GESIS Panel	0.88
			GIP	0.88
			MCS	0.87
			SOEP	0.57
		Logistic Regression Unpenalised	GESIS Panel	0.88
			GIP	0.88
			MCS	0.87
			SOEP	0.53
		Random Forest Classifier	GESIS Panel	0.87
			GIP	0.86
			MCS	0.86
			SOEP	0.75

Table (Appendix) 3.8: *Models trained on other surveys but applied to FREDA Panel. Part two: Equivalent In Lifecycle.*

Cross Mode	Test Wave	Model Class	Train Data	AUROC
Latest Data Available	7/04/2021	Gradient Boosting Classifier	GESIS Panel	0.51
			GIP	0.51
			MCS	0.49
			SOEP	0.49
		Logistic Regression Penalized	GESIS Panel	0.49
			GIP	0.51
			MCS	0.48
			SOEP	0.51
		Logistic Regression Unpenalised	GESIS Panel	0.49
			GIP	0.51
			MCS	0.49
			SOEP	0.51
		Random Forest Classifier	GESIS Panel	0.51
			GIP	0.53
			MCS	0.53
			SOEP	0.50
	7/07/2021	Gradient Boosting Classifier	GESIS Panel	0.87
			GIP	0.87
			MCS	0.86
			SOEP	0.87
		Logistic Regression Penalized	GESIS Panel	0.86
			GIP	0.88
			MCS	0.87
			SOEP	0.87
		Logistic Regression Unpenalised	GESIS Panel	0.86
			GIP	0.87
			MCS	0.87
			SOEP	0.87
		Random Forest Classifier	GESIS Panel	0.87
			GIP	0.87
			MCS	0.86
			SOEP	0.87

3.7 References

Altmann, André et al. (2010). "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10, pp. 1340–1347.

Bach, Ruben L, Stephanie Eckman, and Jessica Daikeler (2020). "Misreporting Among Reluctant Respondents". In: *Journal of Survey Statistics and Methodology* 8.3, pp. 566–588.

Bergmeir, Christoph and José M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191. Publisher: Elsevier, pp. 192–213.

Blom, Annelies G., Carina Cornesse, et al. (2021). *Mannheim Corona Study*. GESIS Data Archive, Cologne. ZA7745 Data file Version 1.0.0, https://doi.org/10.4232/1.13700. 2021.

Blom, Annelies G., Christina Gathmann, and Ulrich Krieger (2015). "Setting Up an Online Panel Representative of the General Population: The German Internet Panel". In: *Field Methods* 27.4. Publisher: SAGE Publications Inc, pp. 391–408.

Blom, Annelies G., Marisabel Gonzalez Ocanto, et al. (2022). *German Internet Panel, Wave 58 (March 2022)*. GESIS, Cologne. ZA7878 Data file Version 1.0.0, https://doi.org/10.4232/1.14054. 2022.

Bosnjak, Michael et al. (2018). "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel". In: *Social Science Computer Review* 36.1. Number: 1, pp. 103–115.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Bujard, Martin et al. (2023). *FReDA – The German Family Demography Panel*. In collab. with Infas Institute Für Applied Social Science, Bonn. Version Number: 3.0.0. 2023.

Cheng, Albert, Gema Zamarro, and Bart Orriëns (2016). "Personality as a Predictor of Unit Nonresponse in Panel Data: An Analysis of an Internet-Based Survey". In: *EDRE Working Paper* 2016.

Durrant, Gabriele B. and Fiona Steele (2008). "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence From Six UK Government Surveys". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 172.2, pp. 361–381.

Federal Institute for Population Research (2022). *FReDA* URL: `https://www.bib.bund.de/EN/Research/Family/Projects/FReDA-Family-Research-and-Demographic-Analysis.html`.

Feurer, Matthias and Frank Hutter (2019). "Hyperparameter Optimization". In: *Automated Machine Learning*. Ed. by Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Series Title: The Springer Series on Challenges in Machine Learning. Springer International Publishing, Cham, 2019, pp. 3–33.

Fitzgerald, John, Peter Gottschalk, and Robert Moffitt (1998). *An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics*. t0220. Cambridge, MA: National Bureau of Economic Research, 1998, t0220.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232.

Fuchs, Marek, Dayana Bossert, and Sabrina Stukowski (2013). "Response Rate and Nonresponse Bias - Impact of the Number of Contact Attempts on Data Quality in the European Social Survey". In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 117.1. Publisher: SAGE Publications Ltd, pp. 26–45.

German Institute for Economic Research (DIW Berlin) (2023). *DIW Berlin: SOEP-Core v37eu (Data 1984-2020, EU-Edition)* URL: https://www.diw.de/en/diw_01.c.838578.en/edition/soep-core_v37eu__data_1984-2020__eu-edition.html.

GESIS (2021). *GESIS - Leibniz Institute for the Social Sciences* URL: https://www.gesis.org/en/gesis-panel/documentation.

— (2023). *GESIS Panel - Standard Edition*. Published: GESIS, Cologne. ZA5665 Data file Version 44.0.0, https://doi.org/10.4232/1.13931 DOI: 10.4232/1.13931. 2023.

Goebel, Jan et al. (2019). "The German Socio-Economic Panel (SOEP)". In: *Jahrbücher für Nationalökonomie und Statistik* 239.2, pp. 345–360.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

Hill, Craig A. et al., eds. (2020). *Big Data Meets Survey Science: A Collection of Innovative Methods*. 1st ed. Wiley, New Jersey.

Jacobsen, Erin et al. (2021). "Predictors of attrition in a longitudinal population-based study of aging". In: *International Psychogeriatrics* 33.8. Publisher: Cambridge University Press, pp. 767–778.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. Springer, New York.

Jankowsky, Kristin, Diana Steger, and Ulrich Schroeders (2022). *Predicting Lifetime Suicide Attempts in a Community Sample of Adolescents Using Machine Learning Algorithms*. 2022.

JSSAM (2022). *Special Virtual Issue on Nonresponse Rates and Nonresponse Adjustments* URL: https://academic.oup.com/jssam/pages/special-virtual-issue-on-nonresponse-rates-and-nonresponse-adjustments.

Kern, Christoph, Thomas Klausch, and Frauke Kreuter (2019). "Tree-based Machine Learning Methods for Survey Research". In: *Survey research methods* 13.1, pp. 73–93.

Kern, Christoph, Bernd Weiß, and Jan-Philipp Kolb (2021). "Predicting Nonresponse in Future Waves of a Probability-Based Mixed-Mode Panel with Machine Learning*". In: *Journal of Survey Statistics and Methodology* 11.1, pp. 100–123.

Kocar, Sebastian and Nicholas Biddle (2022). "The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design". In: *Quality & Quantity*.

Kreuter, F and A Jäckle (2008). "Are contact protocol data informative for potential nonresponse and nonresponse bias in panel studies? A case study from the Northern Ireland subset of the British Household Panel Survey". In: *Panel Survey Methods Workshop, Colchester*, pp. 14–15.

Le Cessie, S. and J. C. Van Houwelingen (1992). "Ridge Estimators in Logistic Regression". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41.1, pp. 191–201.

Lemay, Michael (2009). *Understanding the Mechanism of Panel Attrition*. University of Maryland.

Liebig, Stefan et al. (2022). *Socio-Economic Panel, data from 1984-2020, (SOEP-Core, v37, EU Edition)*. doi: 10.5684/SOEP.CORE.V37EU. 2022.

Lipps, Oliver (2007). "Attrition in the Swiss Household Panel". In: *Methoden, Daten, Analysen (mda)* 1.1, pp. 45–68.

— (2009). "Attrition of Households and Individuals in Panel Surveys". In: *SSRN Electronic Journal*.

Lugtig, Peter (2014). "Panel Attrition: Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers". In: *Sociological Methods & Research* 43.4, pp. 699–723.

Luiten, Annemieke, Joop Hox, and Edith de Leeuw (2020). "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys". In: *Journal of Official Statistics* 36.3, pp. 469–487.

Lynn, Peter, Pablo Cabrera-Álvarez, and Paul Clarke (2023). "Sample composition and representativeness on Understanding Society". In: *Fiscal Studies* 44.4, pp. 341–359.

McLauchlan, Cynthia and Matthias Schonlau (2016). "Are Final Comments in Web Survey Panels Associated with Next-Wave Attrition?" In: *Survey Research Methods* 10, pp. 211–224.

Minderop, Isabella and Bernd Weiß (2023). "Now, later, or never? Using response-time patterns to predict panel attrition". In: *International Journal of Social Research Methodology* 26.6, pp. 693–706.

Mulder, J and N Kieruj (2018). *Preserving Our Precious Respondents: Predicting and Preventing Non-Response and Panel Attrition by Analyzing and Modeling Longitudinal Survey and Paradata Using Data Science Techniques*. 2018.

Oh, Sejong (2022). "Predictive case-based feature importance and interaction". In: *Information Sciences* 593, pp. 155–176.

Olson, Kristen (2013). "Paradata for Nonresponse Adjustment". In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 142–170.

Pforr, Klaus and Jette Schröder (2016). "Why Panel Surveys?" In: *GESIS Survey Guidelines*. In collab. with GESIS-Leibniz-Institut Für Sozialwissenschaften. Publisher: SDM-Survey Guidelines (GESIS Leibniz Institute for the Social Sciences) Version Number: 2.0.

Plewis, Ian and Natalie Shlomo (2017). "Using Response Propensity Models to Improve the Quality of Response Data in Longitudinal Studies". In: *Journal of Official Statistics* 33.3, pp. 753–779.

Roßmann, Joss and Tobias Gummer (2016). "Using Paradata to Predict and Correct for Panel Attrition". In: *Social Science Computer Review* 34.3. Publisher: SAGE Publications Inc, pp. 312–332.

Saarela, Mirka and Susanne Jauhiainen (2021). "Comparison of feature importance measures as explanations for classification models". In: *SN Applied Sciences* 3.2.

Siegers, Rainer, Hans Walter Steinhauer, and Lennart Dührsen (2021). "SOEP-Core v36 – Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2019)". In: SOEP Survey Papers 1106: Series C. Berlin: DIW/SOEP.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Uhrig, S C Noah (2008). "The Nature and Causes of Attrition in the British Household Panel Survey". In: *ISER Working Paper Series*.

Voorpostel, Marieke and Oliver Lipps (2011). "Attrition in the Swiss Household Panel: Is Change Associated with Drop-out?" In: *Journal of Official Statistics* 22.2, pp. 301–318.

Zinn, Sabine and Timo Gnambs (2022). "Analyzing nonresponse in longitudinal surveys using Bayesian additive regression trees: A nonparametric event history analysis". In: *Social Science Computer Review* 40.3, pp. 678–699.

4 Prediction-Based Adaptive Designs for Reducing Attrition Rates and Bias in Panel Surveys

Abstract

Machine learning-based nonresponse prediction in panel surveys enables targeted preemptive interventions. However, the optimal use of ML predictions in Adaptive Survey Design (ASD) remains uncertain. Prior research relies on assumption-heavy simulations or experimental ASDs with limited generalizability. This paper proposes a method that integrates field experiment results on incentives with ML-based propensity models to ex-post simulate ASD strategies with minimal assumptions. Using German panel data, we show that targeting the 15% of lowest-propensity panelists with cash incentives or a survey module on their preferred topic reduces nonresponse rates by 1-2 percentage points. Findings on nonresponse bias are mixed, with some variables showing reduced bias while others remain unchanged or worsen. We present this method as a framework for panel managers to test the expected outcomes of many ASD options simultaneously with minimal assumptions. The framework helps select the criteria for targeting participants and the change to the survey protocol.

4.1 Introduction

Nonresponse in panel surveys limits the potential of longitudinal analysis and can bias substantive insights (Groves, 2006). Survey managers can address nonresponse either ex-post, by applying longitudinal survey weights, or ex-ante, by encouraging low-propensity participants to respond. Here, we focus on the latter approach. The first step is identifying panelists likely to nonrespond. Methodological research has demonstrated the potential of machine learning (ML) models to predict "wave nonresponse," which is when an active panelist is invited to a survey wave but does not participate (Kern, Weiß, and Kolb, 2021; Koch and Blohm, 2016; Bach, Eckman, and Daikeler, 2020; Mulder and Kieruj, 2018; Zinn and Gnambs, 2022; Trappmann, Gramlich, and Mosthaf, 2015; Felderer, Kueck, and Spindler, 2023). Using these predictions, participants at risk of nonresponse can be identified and, as a second step, targeted by adaptive survey designs (ASDs) that adjust survey protocols across participant subgroups to reduce their nonresponse risk (Wagner, 2008; Schouten, Peytchev, and Wagner, 2017). While combining flexible ML models with ASDs is conceptually promising, critical questions remain regarding the practical utility of ML predictions in targeting interventions to

reduce panel nonresponse and bias.

The interplay between prediction quality and an intervention's treatment effects is critical to the performance of prediction-based ASDs. Predicted response propensities estimate each participant's likelihood of responding, but these are imperfect—some predicted nonrespondents will respond, and some predicted to respond will not. To build effective prediction-based ASDs, we are not interested in the model's accuracy per se, but in how the resulting predictions can help to select targets for treatment protocols that lead to reduced nonresponse rates and bias. It is thus crucial to understand whether certain interventions are more or less effective for (predicted) low-propensity participants compared to the average participant. If so, greater reductions in nonresponse bias may be achieved by targeting the most at-risk participants with an intervention that is particularly effective for them. Alternatively, it may be preferable to target moderately at-risk participants, assuming that certain interventions could be more effective in that subgroup and yield higher retention rates. In order to optimally combine nonresponse predictions with potential interventions, the expected outcomes of these decisions need to be assessed in advance.

We present a framework for survey researchers seeking to evaluate their options for prediction-based adaptive survey designs. Specifically, we show how to integrate results from a survey experiment with an ML-based propensity model to conduct a series of ex-post simulations, assessing how predicted response propensities could have been used to target at-risk participants with different interventions. We then systematically quantify the impact on wave nonresponse rates and bias across a diverse set of survey variables. This framework is applicable to any panel study and enables evaluating multiple ASD options simultaneously under realistic conditions, rather than testing a single intervention at a time.

We advance prior research by (i) demonstrating a principled approach to designing prediction-based ASDs, (ii) showing how treatment effects can vary across propensity groups, and (iii) showcasing the effectiveness of allocating treatments based on ML predictions using German panel data. To that end, we examine a set of interventions that are widely replicable in panel surveys: varying cash incentives, survey length, and survey content. By studying the optimal use of predictions for treatment allocation and selection, we shed light on the "missing link" in developing effective prediction-based ASDs, i.e. the combination of predictions and treatment decisions.

Our paper is structured as follows. First, we evaluate the state of the research and present the research questions we aim to address (see Background). Second, we introduce data and methods that answer these research questions (see Methods). Third, we present the results (see Results) and end with a discussion on how survey research and practice can build on these findings (see Discussion).

4.1.1 Background

Panel surveys offer valuable data to study changes in attitudes and behavior within individuals over time and thus can help uncover causal effects (Allison, 2009; Andreß, Golsch, and Schmidt, 2013; Lynn, 2009). In contrast to cross-sectional surveys, panel surveys

have to represent the population of interest over several waves to ensure the validity of results based on the data. Panel surveys, however, face challenges in selective wave nonresponse of panel members, which may result in biased estimates if response propensities correlate with outcomes of interest (Groves, 2006). Considering these different response propensities, ASDs adjust survey protocols between different groups of respondents to reduce the nonresponse bias of a panel (Wagner, 2008; Schouten, Peytchev, and Wagner, 2017). In contrast to traditional survey designs that provide the same survey protocol for each respondent, participants with varying risks of nonresponse receive different survey protocols in ASD to align their response probabilities (Gummer, 2020). A successful ASD has two requirements: (i) correctly predicting those respondents who are at risk of nonresponse and (ii) effectively allocating a treatment to these respondents that improves their response probabilities.

i. Nonresponse Prediction

Survey practitioners have long employed statistical modeling to estimate each participant's probability of responding to a given survey wave (e.g., Bethlehem, 1988; Trappmann, Gramlich, and Mosthaf, 2015; Roßmann and Gummer, 2016; Kocar and Biddle, 2022). Recent research has adopted ML techniques to improve predictive accuracy. Numerous studies have successfully employed ML-based nonresponse prediction (Bach, Eckman, and Daikeler, 2020; Collins and Kern, 2024; Kern, Weiß, and Kolb, 2021; Mulder and Kieruj, 2018; Zinn and Gnambs, 2022). For example, Kern, Weiß, and Kolb (2021) found that a random forest model would predict over half of a given wave's nonrespondents correctly while yielding only a 20% false positive rate in the GESIS Panel (which is also used in this study). These authors reported that ML approaches outperform traditional and frequently used methods such as logistic regression. While no prediction model is completely accurate, the degree of predictive performance of the ML models shows that their predicted response propensities have a degree of validity in estimating each participant's true probability to (non)respond. They thus provide value for targeting adaptive survey designs.

ii. Treatment Allocation

Despite their conceptual appeal, previous research only covers a limited range of possible Adaptive Survey Design (ASD) implementations. Earlier studies on implementing ASD are strongly based on simulation studies (Zhang and Wagner, 2024; Gummer, 2020; Watson and Cernat, 2023; Schouten, Cobben, et al., 2016). For example, Watson and Cernat (2023) fitted a logit model to estimate both nonresponse propensity and the change in that propensity after an in-field follow-up using data from an Australian and United Kingdom-based panel survey. They then used these models to estimate the effect of different follow-up allocation strategies to evaluate how they could maintain sample balance even with a lower budget for follow-ups. McCarthy, Wagner, and Sanders (2017) compared data from an agricultural survey with a corresponding census, allowing them to know in advance the true population values. The authors simulated scenarios where response propensity estimates were used to divert follow-up resources away from high-propensity cases to low-propensity cases and derived the impact on response rates and

sample composition. Each of these simulation studies prove the concept of various ASD techniques, but rely on strong assumptions about how they would fare in practice. Wagner (2008) minimized the assumptions in their simulation by using observational data from existing panels to estimate how call times and modes affected different participants' response behavior. The study then used simulations to show what the outcome would have been had these modes and call schedules been targeted to maximize responses. Our paper goes a step further by executing a field experiment to trial specific interventions deliberately, as opposed to relying on data that naturally occurs through the normal operations of a panel.

When it comes to implementing ASD in practice, more knowledge on which treatments to use and how to use them is required. Experimentation in large-scale population surveys is costly and often conflicts with project goals, as Zhang and Wagner (2024) have argued. Consequently, empirical studies on the effects of different treatments on different risk groups are limited, often only comparing a single treatment against a control group. For instance, Lynn (2016) and Zhang and Wagner (2024) investigated tailored invitation letters, whereas Gummer and Blumenstiel (2018) and Wagner et al. (2012) tested the impact of allocating extra interviewer effort to respondents classified as 'high priority' by supervisors or the interviewer team. Zhang, West, et al. (2024) conducted a field experiment in which participants from areas with a high Hispanic population were provided Spanish invitation letters, and households from rural areas or with elderly members were provided paper surveys. Wagner (2013) fitted models that predicted the best times to call certain participants and used these to recommend a call list to the telephone interviewers. Beste et al. (2023) targeted the lowest-propensity households with extra cash incentives and improved their response rates. However, since these studies focus on how specific ASD implementations fared in practice, they provide only limited guidance on how to best utilize nonresponse predictions to allocate interventions in prediction-based ASDs.

iii. Interplay of Propensity and Interventions

Even with overall accurate nonresponse predictions and effective treatments, it remains an open question how to select the at-risk respondents who should be treated differently from the remainder of the sample. Researchers must set a threshold of nonresponse risk at which participants should be targeted and choose the appropriate survey protocol (i.e., treatment). Yet, there is a lack of studies investigating the role of different cutoff points in ASD. Beste et al. (2023) simply used predicted response propensities by targeting the 50% most at-risk households for extra cash incentives. Some studies explore the use of expert opinion to allocate extra call attempts or follow-ups. For example, Wagner et al. (2012) tested a multi-stage process of case prioritization during the field period, using both propensity models and expert opinions to identify the 50% most at-risk cases. Similarly, Coffey et al. (2020) tested modifying propensity models with expert opinions to improve the estimated response propensities.

These experiments show that targeting the most at-risk participants with certain interventions was beneficial, but they did not explore the possibility that other techniques could perform even better. Simulation studies, although assumption-dependent, allow

researchers to explore alternative targeting regimes. Watson and Cernat (2023) simulated various targeting mechanisms, specifically using either the predicted response propensity, a measure of how much that participant would contribute to nonresponse bias if they did not participate (R-indicator), or a combination of both. The simulation showed that by incorporating the R-indicator, the panel could reduce the number of follow-ups needed to reach a given level of sample balance. In the context of our study, this would have been equivalent to targeting the 75% highest priority (prioritized based on estimated propensity and R-indicator) participants with extra follow-ups, as they reduced the number of follow-ups by 25% by ignoring the lowest-priority participants. This simulation highlights the value of exploring more possibilities than just targeting the lowest-propensity participants. One possibility that has not been explored, however, is targeting participants with a moderate to high risk of nonresponse but not the highest risk, as the latter group may be most difficult to persuade to stay in the panel. If the given intervention is significantly less effective on the most at-risk participants but still effective on moderately at-risk participants, then we might expect a higher response rate for the same cost of the intervention if we target ASD designs in this way.

iv. Research Questions

Previous research has proposed various methods for accurately predicting nonresponse in panel studies and explored a range of intervention strategies for balancing response rates and decreasing nonresponse bias. However, there is limited guidance on how to best combine these two key ingredients – nonresponse predictions and treatments – to build effective prediction-based ASDs. Simulation studies prove the conceptual value of ASD but are too assumptions-dependent to provide certainty about the outcomes of specific ASD strategies. At the same time, experimenting with ASD can conflict with study objectives, so survey researchers are rightly cautious about undertaking ASD without some evidence for the likely outcomes. As a result, ASD research is progressing slowly despite the need for innovation and actionable insights.

Our study addresses this problem by presenting ex-post simulations that directly draw from a survey experiment and, therefore, minimize the assumptions in the simulations. In other words, we base our results on what different groups of actual panelists did after receiving a specific treatment, rather than inferring what would have happened based on hypothetical/assumed treatment effects. We argue that this simulation technique allows us to compare the likely outcomes of a wide variety of prediction-based ASD strategies. Our approach thereby allows researchers to explicitly model the interplay between ML-predictions and treatment allocation and how it will impact the performance of different ASD strategies. For example, if certain interventions are more (or less) effective for low propensity (i.e., high-risk) participants, then this might affect which group is optimal to target with interventions for the goal of reducing overall nonresponse. This study, therefore, provides a rich resource to systematically investigate the possible impact of different ASD strategies.

We structure our investigation around the following research questions (RQs) that are situated in the interplay between prediction and treatment allocation:

- RQ1. How do treatment effects vary with predicted response propensity? Are certain treatments more/less effective at specific propensity levels?
- RQ2. Does an adaptive design targeted towards panelists based on their predicted response propensity decrease nonresponse rates and bias?
- RQ3. What is the optimal threshold for predicted risk to use when deciding which participants to target with treatments in an prediction-based ASD?

We exemplify how ex-post simulations can be used to answer these questions by drawing on a survey experiment and prediction models that we developed and implemented in a German probability-based panel study, the GESIS Panel.

4.2 Methods

We will simulate several scenarios in which ML-based predictions are used to select which participants should receive a different survey protocol than the rest of the sample to lower their likelihood of wave nonresponse. These scenarios are derived from data collected in a randomized treatment-control experiment conducted within the GESIS Panel. The response propensities were estimated using an ML model trained with data from earlier waves of the same panel (see Nonresponse Prediction Model).

In this section, we first provide context by describing the details of the GESIS Panel (see Data). After introducing the panel study, we outline the survey data quality indicators we will evaluate in the simulation scenarios (see Nonresponse Rates and Nonresponse Bias). Next, we describe the design of our simulation, starting with the two key components: the randomized treatment-control experiment (see Experimental Design) and the ML prediction model (see Nonresponse Prediction Model). Finally, we explain how these two elements are used together to conduct ex-post simulations that estimate the impact of various adaptive design strategies on survey data quality (see Analytical Strategy).

4.2.1 Data

The dataset for our paper is the GESIS Panel up to January 2024 (GESIS, 2024), totaling 54 panel waves.

The GESIS Panel is a probability-based omnibus survey that began in October 2013 and ran bi-monthly waves until February 2021, when it switched to quarterly waves. It operates as a self-administered, mixed-mode survey, allowing participants to choose between a web-based or paper-based mail format, with approximately 75% of participants opting for the web-based option and the rest responding by mail (GESIS, 2024). Each wave's questionnaire takes up to 25 minutes to complete and includes diverse content from various social science disciplines. For each wave, respondents receive an unconditional, prepaid incentive of EUR 5.

The GESIS Panel's initial sampling method randomly selected invitees from the German population register. Residents of Germany aged 18-70 were eligible for recruitment. Subsequent refreshment samples were based on a piggybacking approach utilizing the

German General Social Survey (ALLBUS) and the German parts of the International Social Survey Programme (ISSP). With these refreshments, the target population's age range slightly changed, now including the adult population (18+). For more details on the sampling methodology, we refer readers to the study documentation (GESIS, 2022). An initial sample of roughly 5,000 participants was recruited in 2013, with subsequent refreshment samples in 2016, 2018, and 2023 to maintain the total sample size around 5,000 (see Figure 4.1).

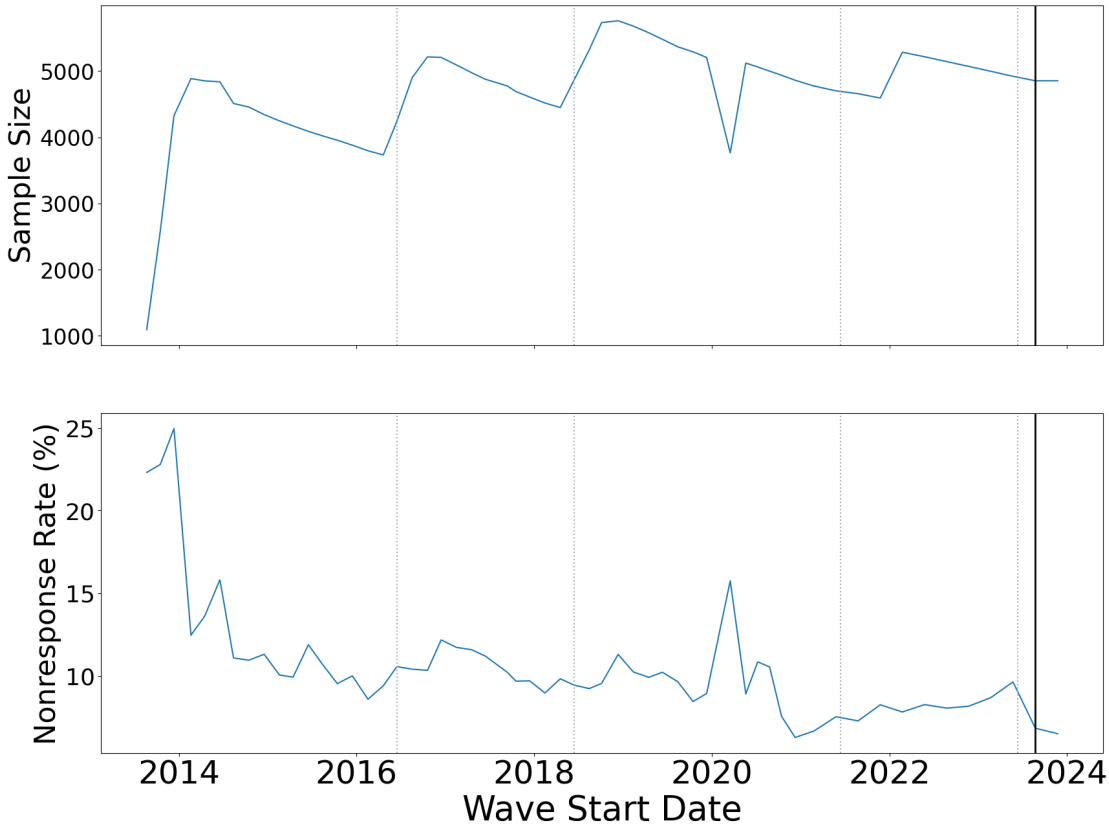


Figure 4.1: *Timeline of sample sizes, nonresponse rates and recruitment intakes for the GESIS Panel. Dotted lines mark the commencement of refreshment intakes, and the solid line marks the wave when the experiment was conducted (see Experimental Design.*

4.2.2 Nonresponse Rates and Nonresponse Bias

The indicators of survey data quality we present are (1) wave-level nonresponse rate and (2) nonresponse bias in specific survey variables of interest, which are discussed below.

The nonresponse rate for a given wave refers to the portion of invited participants who did not respond within the field period. We follow the RR6 definition (AAPOR, 2016), which counts unusable partial responses as nonresponses. The GESIS Panel stops inviting participants who have not provided any submission (partial or complete) for

three consecutive waves. The participants are unaware of this rule (Bosnjak et al., 2018). The development of nonresponse rates over the course of the GESIS Panel is shown in Figure 4.1. We can see that nonresponse rates change over time such that nonresponse rates are high when the panel first commences. Low-propensity participants attrite from the panel, leading to a reduced sample size, but also lower, stable nonresponse rates. This pattern continues with the exception of a spike in nonresponses around the COVID-19 pandemic in 2020.

There are several ways to conceptualize nonresponse bias (Koch and Blohm, 2016). In the following, nonresponse bias is calculated for a given variable at each wave as the difference between the mean among respondents and nonrespondents (appendix equation 4.3). For robustness, we present alternative measures in the appendix (see appendix section Additional Results). Because our study draws on panel data, we can use the participant's last known value for these calculations, even if they did not respond in a given wave. Every variable we examine in our study was refreshed roughly once a year as of the time of the experiment. Therefore, even if we rely on a value from a past wave to fill in the missing value of a nonrespondent to calculate the nonresponse bias, this data is almost always no more than one year out of date.

A longitudinal study aims to understand a population's attitudes and behaviors through inferences drawn from a sample (Groves, 2011). Therefore, we have selected a range of variables that cover diverse attitude and behavioral themes as well as some of the socio-demographics typically used to weight samples. For socio-demographics, we include age, college education, and (monthly) household income. We also include self-rated physical health on a six-point scale. For behaviors, we include variables that raise the expectation that participation might correlate with the given type of behavior, namely, the status of COVID-19 vaccination (Soeder et al., 2024) and political participation (Gummer and Blumenstiel, 2018). Note that we collect COVID-19 vaccination status since 2021, but demographics were collected since 2013 and all other variables since 2014–2015. Political participation was measured using four items on a five-point scale asking how frequently the respondent participates in a social movement, political party, labor union, or charity work. We use the highest value reported among those four items to indicate political participation. A higher value indicates a higher frequency of participation. This simple measure assumes all four areas are of equal importance to political participation. For attitudes, we include a seven-point measure of how serious they feel is the threat of climate change. A higher value indicates a higher seriousness. We also include how they rate themselves on an 11-point scale of politically left- or right-wing (1 and 11, respectively). Details of how these variables are derived, including the original survey items, are presented in the appendix section Replication.

4.2.3 Experimental Design

Our study uses results from an experiment conducted within the GESIS Panel to simulate an adaptive design. This section describes the experiment and how it is used in the simulation is covered in section Analytical Strategy. During the data collection wave of August 2023, participants were randomly assigned to either a control group or one

of three treatment groups. Henceforth, the wave in which the experimental treatments were administered shall be called wave "W_0,", the subsequent wave (November 2023) shall be "W_1" and the previous wave (May 2023) "W_{-1}." The experimental groups were as follows:

Control Group: The control group received the questionnaire for W_0 as it would have been without the field experiment. The questionnaire contained three core (constant) modules on "Media and Social Networks", "Work and Occupation" , and "Flight and Immigration" as well as an additional module on "Political Attitudes and Behavior". The questionnaire took an average of 19 minutes and 15 seconds to complete. Each respondent received an unconditional EUR 5.00 incentive.

Shortened Survey: Participants received an abbreviated version of the survey, with an average completion time of 15 minutes and 47 seconds instead of the control group's 19 minutes and 15 seconds. The survey was shortened by removing half of the additional module "Political Attitudes and Behavior" discussed above.

Interesting Survey Topic: Each survey in the GESIS Panel consists of several core topic modules, and additional modules that change each wave. For this treatment, we determined which of those topics in the GESIS omnibus was the most popular overall. Participants rated their preferences for the variable topics in wave W_{-1}. The most popular topic was "Nature and Environment," which was then selected as the additional topic in the treatment group in wave W_0. Participants in the control group received a question module on the topic "Political Attitudes and Behavior" instead (which respondents had rated as less preferred). Both questionnaires were structurally equivalent.

Extra Cash Incentive: Participants in this group received a EUR 20.00 cash incentive instead of the control group's EUR 5.00. All cash incentives are unconditional, prepaid, and sent directly to the participant with the invitation letter.

4.2.4 Nonresponse Prediction Model

We study adaptive design strategies in which interventions are targeted at participants based on their predicted risk of nonresponse. Each participant's risk of nonresponse in W_0 is predicted using a machine learning approach following Kern, Weiß, and Kolb (2021) and Collins and Kern (2024), in which random forest models provided the highest predictive accuracy overall across a range of model types. We re-build their random forest model using past nonresponse behavior and socio-demographic factors to predict each participant's probability of response in the next wave using a binary response outcome (response [1] vs. nonresponse [0]; following RR6 definition) as the prediction target. We selected these predictors to be comparable to any other panel study that would typically collect demographic information and record past nonresponse behavior. The specifications of this model are presented in appendix section Methods Details. Note that the distribution of the predicted response propensities is skewed towards "response"

with a mean value of 93% likely to respond.

We explore two techniques for using the predicted response propensities to target interventions, and we compare them against a baseline in which interventions were assigned randomly instead of using ML predictions. By using random assignment, we keep the portion of the sample that receives an intervention the same, but vary only the manner in which we select recipients, thereby testing the effect of the targeting mechanism.

Targeting High-Risk Participants: In this scenario, we explore how nonresponse rates and biases could have been improved by targeting participants with the lowest likelihood of response. We simulate targeting the lowest 15%, 25%, and 35% of predicted response propensities.

Targeting Moderate-Risk Participants: Here, we consider what would happen if we targeted participants with a mild, but not the highest, probability of nonresponse. Participants at the highest risk may still not respond even after interventions, leading to an inefficient use of resources. Thus, we target participants with a middle-tier risk, focusing on the 15%, 25%, and 35% below the wave's mean response propensity value.

Random Allocation Baseline: Finally, as a baseline, we simulate a targeting regime where 15%, 25%, and 35% of participants were selected randomly (called "random allocation") and use this to compare the effectiveness of targeted interventions versus applying treatments without a prediction-based targeting strategy.

4.2.5 Analytical Strategy

This section describes our approach to answering RQ1 by evaluating how the various treatments interact with the predicted response propensities. Next, we detail our process for the ex-post simulations that demonstrate the expected outcomes of various adaptive design strategies (RQ2 and RQ3).

Analyzing the Treatment Effects

To answer RQ1, we evaluate the treatment effects of the interventions described in section Experimental Design. We study treatment effects concerning differences in the nonresponse rate in both waves W_0 and W_1 between the control and treatment group (see appendix equation 4.1). We examine how treatment impacts nonresponse in both the same wave that the treatments were administered (W_0) and the following wave (W_1) because some treatments may not have an effect until after they are experienced (for example, a shorter survey might not affect the respondent until after they have had the experience of completing the survey in less time). We analyze the treatment effects with logistic regression models predicting the binary nonresponse outcome of each participant (i.e., wave nonresponse in wave W_0 and W_1), and the independent variable is the binary treatment term. The specifications of that model are in appendix section Methods Details.

We examine the treatment effects specifically among predicted moderate and high-

risk participants with different threshold values that may be used for targeting (see Nonresponse Prediction Model), compared to the overall average treatment effect. We also fit two logistic regression models predicting each participant's binary nonresponse outcome (one for wave W_0 and another for W_1) using treatment status, predicted response propensity, and an interaction term between the two as predictor variables. The coefficient of the interaction term would indicate the rate at which the efficacy of the treatment increases/decreases by the ML-predicted response propensity.

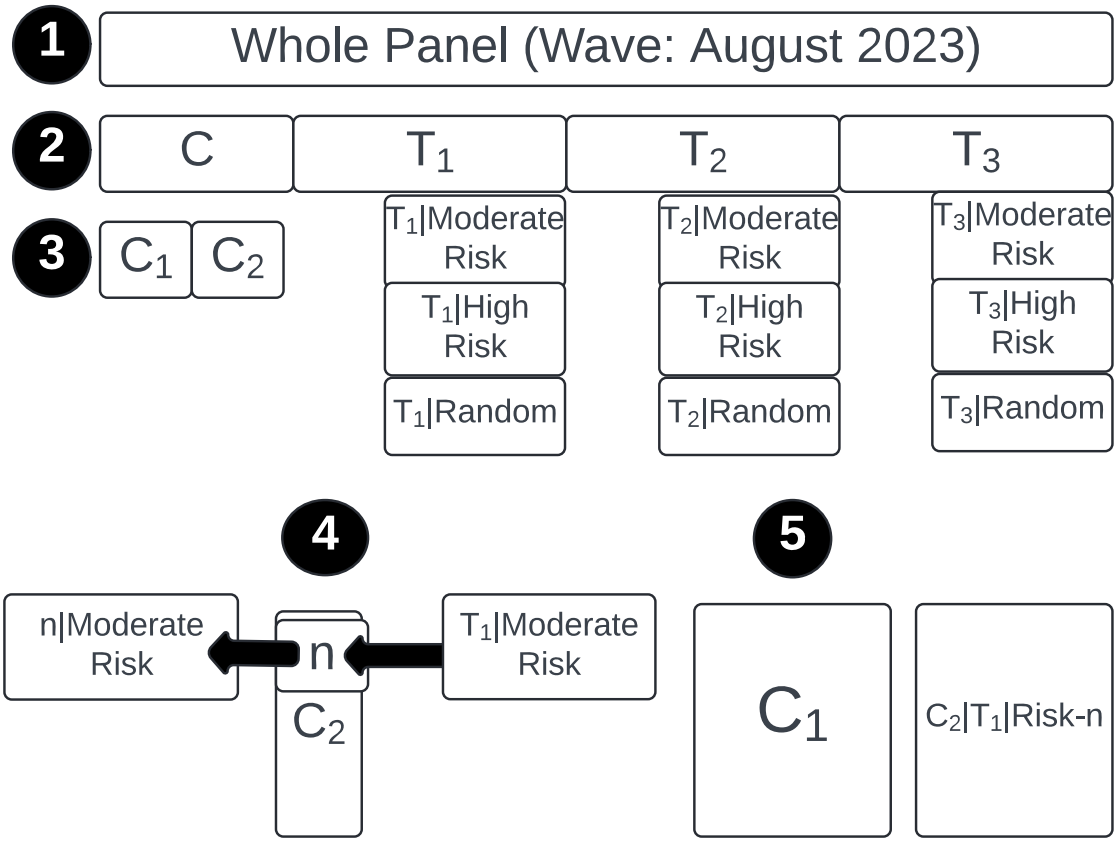


Figure 4.2: *Diagram describing the process for the ex-post simulation of an adaptive design implementation.*

Simulating Adaptive Designs

We address RQ2 and RQ3 through ex-post simulations. Adaptive designs usually use some fixed threshold to differentiate propensity groups and in most cases treatments are applied to the "low propensity group." In our experimental design, however, respondents are randomly allocated to experimental groups, independently from their response propensity. Thus, our design allows us to create multiple scenarios ex-post, based on different criteria for treatment allocation (i.e., different ways of implementing an adaptive survey design). Figure 4.2 presents a diagram of how we conduct the simulation. We

begin with the entire panel as of the experiment wave, W_0 (step 1). We randomly divide the panel into the three treatment groups and one control group for the experiment described in section Experimental Design (step 2). In our simulation, we randomly split the control group into two halves (C_1 and C_2). For each treatment group, we derive three sub-groups based on the targeting strategies (see Nonresponse Prediction Model) (step 3): The k% highest risk participants from each treatment group (T|High-Risk), the k% moderately at-risk (T|Moderate-Risk), and a group of k% randomly selected participants (T|Random); given that k is a value of 15%, 25%, or 35%. Next, we simulate a scenario where the panel survey institution had the capacity to target k% of participants in the panel with one of the treatments using one of the targeting regimes to select recipients. We simulate this by taking group C_2 and removing the k% highest-risk (or moderately at-risk, or a random selection) participants and replacing them with one of the groups (T|High-Risk), (T|Moderate-Risk), or (T|Random) (step 4). For example, in one simulation, we would derive a sub-sample that is based on C_2 except that we replace the 25% most at-risk members of group C_2 with the 25% most at-risk panelists from the cash-incentive treatment group.

Finally, we compare group C_1, which represents the scenario where no intervention was conducted, with the modified version of C_2, which means the scenario where k% of panelists received a given treatment and were targeted based on one of the strategies (step 5). By comparing C_1 and the modified C_2, we derive the nonresponse rates in wave W_0 and W_1 and nonresponse biases for the outlined substantive variables in wave W_1 (see Nonresponse Rates and Nonresponse Bias). This strategy of selecting and matching control and treatment groups allows us to study the outcomes of different prediction-based ASD regimes under highly realistic conditions. We repeat this process for each treatment group, each targeting regime, and for different values of k. Each permutation is repeated one hundred times to account for the randomness introduced by splitting the control group. We then present the average change in nonresponse rates and nonresponse bias caused by each adaptive design execution.

Finally, we conduct t-tests across the pooled samples' outcomes to calculate the probability that the observed 100 simulation outcomes (the change in nonresponse rate or bias value) could have come from a distribution with a true mean of 0 (meaning the specific ASD strategy had no effect on bias or nonresponse rate).

Robustness Checks (RC)

To ensure that our simulation results are robust, we trial variations on the design presented in section Analytical Strategy. These variations are presented in appendix section Additional Results are as follows:

- RC_1. To check if nonresponse bias would accumulate over several waves, we use the 100 repeated simulations as if they were ten sequences of ten waves and then calculate nonresponse bias based on overall responses and nonresponses across those groups of ten. Specifically, we run 100 simulations as described above. However, rather than calculating the change in nonresponse rates and biases in each individual simulation, we group the simulations into sets of ten. Within each group,

we pool all participants from their respective C_1 and modified C_2 conditions, then calculate the nonresponse rates and biases across the pooled participants. Finally, we present the distribution of outcomes across the ten groups. This approach simulates the aggregate bias and nonresponse rates across ten waves, with ten simulation repetitions, to test whether the accumulation of nonresponses over several waves affects the outcomes for bias and nonresponse rates.

- RC_2. By randomly splitting C into C_1 and C_2, we may be creating errors by reducing the sample size. To check if this is substantially affecting the result, we conduct the simulation so that C_1 and C_2 are both duplicates of C.
- RC_3. In this variation, instead of defining nonresponse bias as mean difference between respondents and nonrespondents means, we define it as the difference between the respondents' mean value and the whole sample's mean value (i.e., the variables mean value across all active panelists as of W_1).

4.3 Results

This section presents the results of the survey experiment (see Experimental Design) and, in particular, whether the treatment effects vary between panelists with different response propensities (see Survey Experiment Results). Next, we simulate how those treatment effects can reduce nonresponse bias when employed in an adaptive design strategy that targets participants based on their predicted response propensity (see Simulated Adaptive Designs).

4.3.1 Survey Experiment Results

Examining the differences in average nonresponse rates across treatment groups and predicted risk levels reveals several interesting findings (Figure 4.3). Note that we present a moderate and high-risk group selected according to the rules defined in section Nonresponse Prediction Model with a k value of 25%. Descriptive statistics of both the treatment and risk groups, including results from other k values, are presented in the Appendix Table 4.3. Each treatment-control group had 1,213–1,214 participants. In W_0, there were 94 (7.7%) nonresponders in the control group, 97 (8%) in the short survey treatment group, 85 (7%) in the interesting survey topic treatment group, and 56 (4.6%) in the extra cash incentive treatment group.

Figure 4.3 shows that the shorter survey does not reduce nonresponse in the unfiltered sample in either the experiment or post-experiment wave. In contrast, the extra cash incentive reduces nonresponse in both W_0 (by 3.2 percentage points) and W_1 (by two percentage points). This effect appears to be larger for the lower predicted propensity groups. Finally, the results for the 'interesting survey' treatment are noteworthy. Although the treatment effect in the experiment wave is minor, the effect is more pronounced in the subsequent wave, especially for the moderate and high risk groups.

We aim to validate these observations with results from the logistic regressions in Appendix Table 4.4. The models show that, when considering the whole sample instead

of the groups filtered by risk (where k = 25%), cash has a significant ($p < 0.05$) effect of reducing nonresponse rates in W_0 and W_1. The interesting survey is effective only in W_1. In all other cases, the effect is not significant. Appendix Table 4.4 shows no significant effects in any of the high or moderate risk-groups. However, this may be due to small sample sizes among those risk groups, and so to explore the connection between treatment effects and propensity further, we examine results from the logistic regressions with interaction terms, presented in Appendix Table 4.5. These results further support the efficacy of cash and interesting survey treatments as follows. In W_0, the interaction terms indicate that both the cash incentive and the interesting survey become increasingly effective at reducing nonresponse rates as predicted response propensity decreases ($p < 0.05$). In W_1, however, the interaction term for interesting survey loses significance, while the cash incentive still shows a substantial interaction with the predicted response propensities. The shortened survey is effective at reducing nonresponse rates in W_0 as nonresponse risk increases, but this effect does not persist in W_1. These results should be interpreted cautiously due to the limited number of nonrespondents in the experiment and the small sample size for each treatment-control group. However, the critical observation is that cash and the interesting survey interventions are, to some extent, more effective for participants with lower predicted response propensity.

With regard to RQ1, these results suggest that both increased cash incentives and the interesting survey version are effective interventions. The treatment effect of the cash incentive appears stronger among higher-risk participants in both the wave in which the treatment was administered (W_0) and the subsequent wave (W_1). The shortened survey version seems to encourage responses in W_0 among moderately at-risk participants, though it is less effective among high-risk participants and may even increase nonresponse rates in W_1. By contrast, the interesting survey appears to be more effective in W_1 than W_0 (and more effective as nonresponse risk increases in W_0).

4.3.2 Simulated Adaptive Designs

Figures 4.4, 4.5, 4.6, and 4.7 present the distribution of changes in nonresponse rates and variable-wise biases between the intervention scenario and the no-intervention scenario (see Analytical Strategy). Note that each adaptive design approach is simulated 100 times to account for the randomness introduced by splitting the control group into C_1 and C_2, and so the distribution of outcomes represents the uncertainty around a given ASD outcome. Higher values on the vertical axis indicate that the intervention yielded an improvement: either nonresponse rates declined by the value on the vertical axis, or the nonresponse bias fell, meaning that the mean value for nonresponders became closer to that of responders by the value on the vertical axis. A value below the horizontal zero line indicates that the intervention increased nonresponse rates or biases, which means this ASD was worse than doing nothing. Vertical axis values are in their original scale.

The simulations show that in an ASD scenario in which 25% of the lowest predicted propensity participants (high-risk) were given the interesting survey version in W_0, then in wave W_1 the wave nonresponse rate would have been, on average, 6.2% instead of the (on average) 7.3% yielded from the 'no intervention' simulated scenarios (see Figure 4.4).

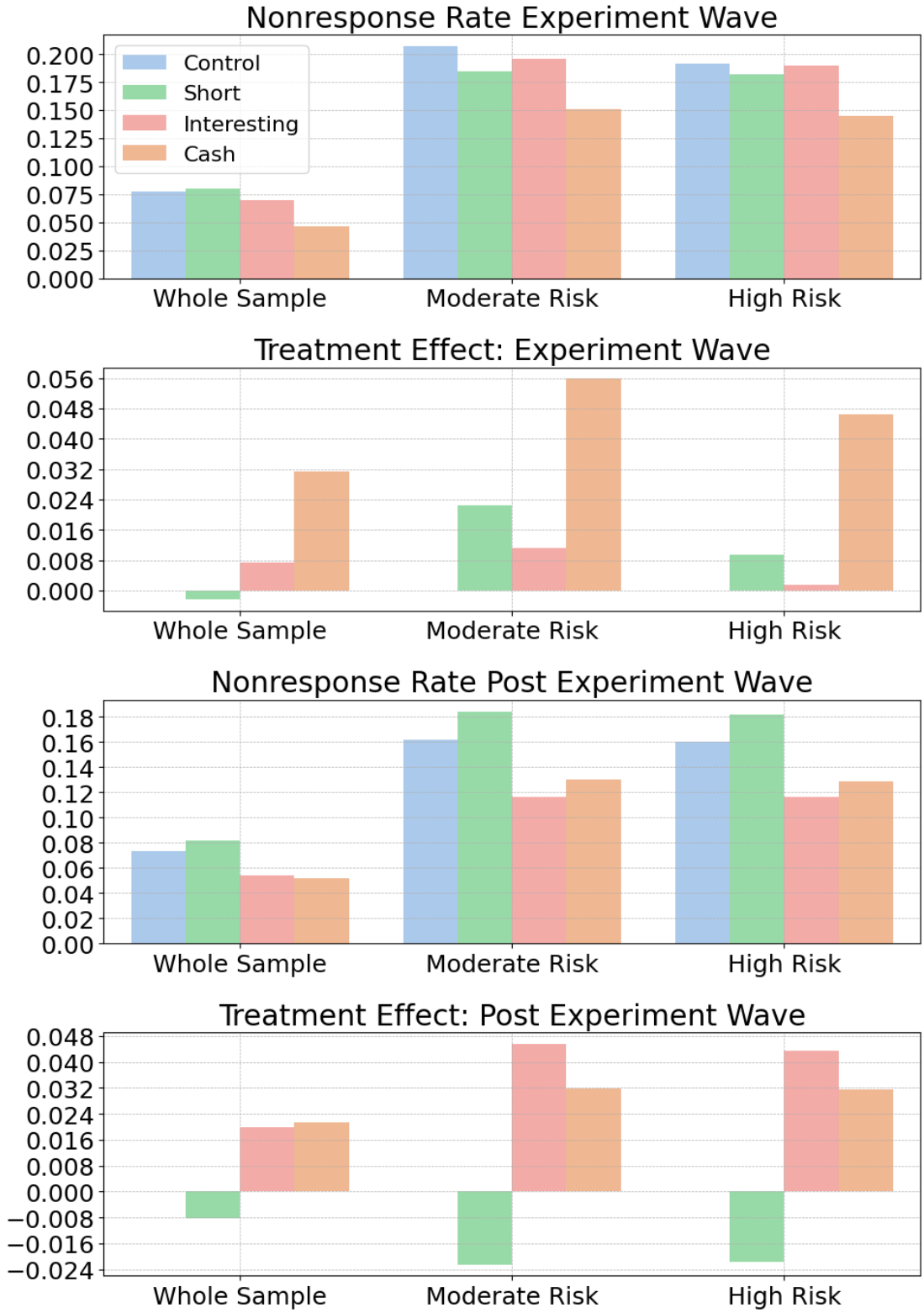


Figure 4.3: *Differences between treatment and control groups in the experiment wave (August 2023) across risk groups (k=25%).*

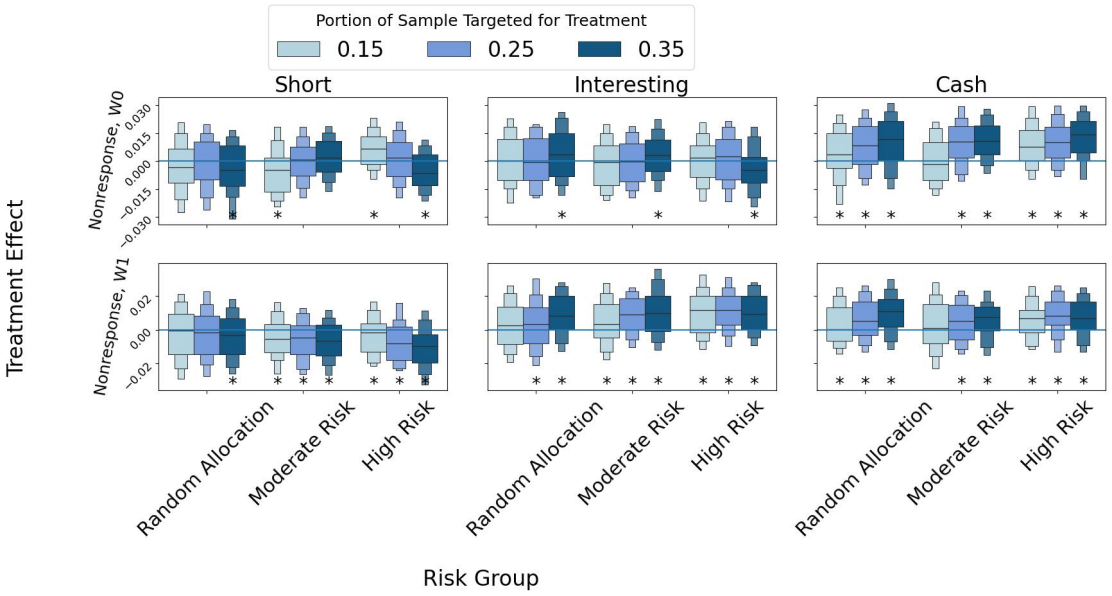


Figure 4.4: *Changes in nonresponse rates between simulated scenarios with and without adaptive designs. The range of values comes from repeating the simulation 100 times to account for the randomness introduced by splitting the control group into C_1 and C_2. Positive values indicate improvements in nonresponse rates due to the intervention.* $*p < 0.05$.

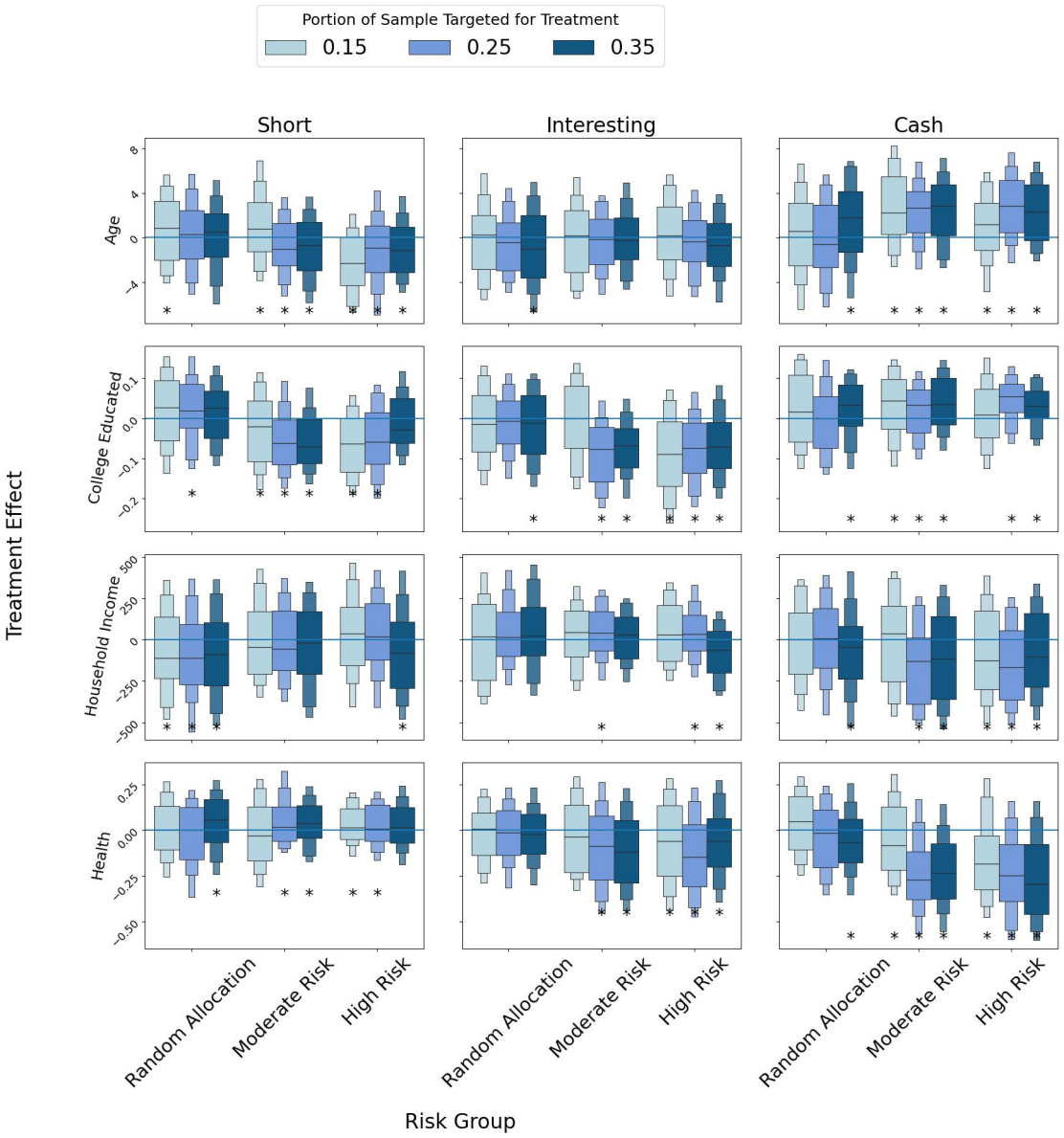


Figure 4.5: *Changes in nonresponse bias (demographic variables) between simulated scenarios with and without adaptive designs. The range of values comes from repeating the simulation 100 times to account for the randomness introduced by splitting the control group into C_1 and C_2. Positive values indicate improvements in nonresponse bias due to the intervention.* $*p < 0.05$.

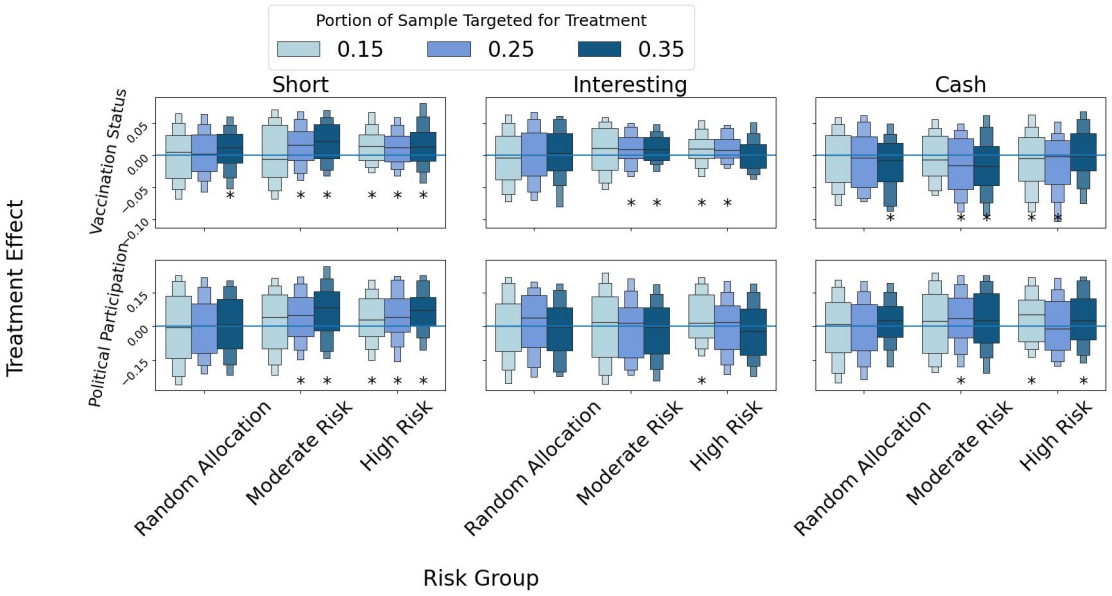


Figure 4.6: *Changes in nonresponse bias (behavioural variables) between simulated scenarios with and without adaptive designs. The range of values comes from repeating the simulation 100 times to account for the randomness introduced by splitting the control group into C_1 and C_2. Positive values indicate improvements in nonresponse bias due to the intervention.* $*p < 0.05$.

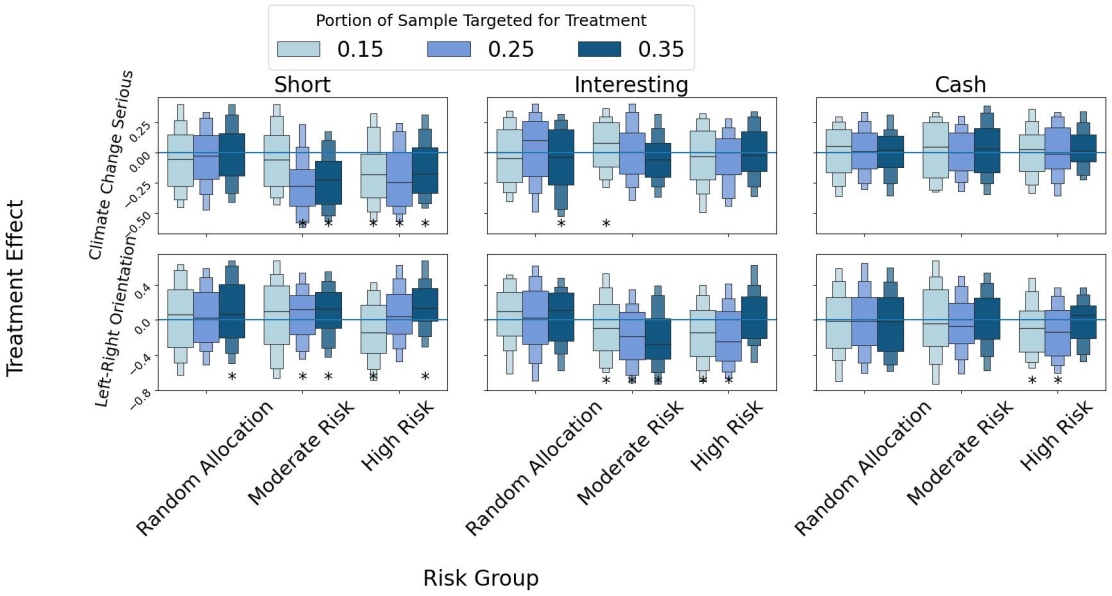


Figure 4.7: *Changes in nonresponse bias (attitudinal variables) between simulated scenarios with and without adaptive designs. The range of values comes from repeating the simulation 100 times to account for the randomness introduced by splitting the control group into C_1 and C_2. Positive values indicate improvements in nonresponse bias due to the intervention.* $*p < 0.05$.

Notably, the interesting survey version also reduced nonresponse by one percentage point when targeting just the 15% most at-risk participants, whereas random targeting of 35% of panelists reduced the nonresponse rate by about 0.5 percentage points. Improvements can be observed for the ASD with extra cash incentive, which yielded a one percentage point decrease in nonresponse rates in multiple scenarios (e.g., high-risk group at W_0 and W_1). In the wave following the ASD implementation (W_1), targeting the increased cash incentive at the 15% most at-risk participants was roughly as effective as targeting 35% of participants randomly. The ASD with a shorter questionnaire, in contrast, did not substantially reduce nonresponse rates across the different scenarios and showed opposing effects in some cases.

The results are mixed when examining changes in variable-wise nonresponse bias (see Figure 4.5, 4.6, and 4.7). When extra cash is applied to highly or moderately at-risk participants, nonresponse bias for age decreases. Specifically, when $k = 15\%$, the no-ASD scenario group was 606 participants of which, on average across repeated random splits of C_1 and C_2, 47 nonresponded (nonresponse rate = 7.5%), but in the mean ASD scenario, nonresponse rates were 0.6 percentage points lower (nonresponse rate = 6.9%). As participants transferred from the (small) nonrespondent to the (large) respondent subgroup, the result was that the average age of nonresponders increased from 51 to 53, while the average age of responders remained around 58. Therefore, cash incentives with k = 15% targeted at high-risk participants on average decreased the nonresponse bias value from 7 to 5 years, which made the variable less biased by two years. Bias in college education improved with the cash incentive, though household income bias did not consistently improve under any ASD strategy (see Figure 4.5). Physical health bias worsened under the same ASD strategies that reduced age bias, possibly because few participants rated their health below four out of six, making it difficult to reach the rare participants with poorer health (see Appendix Table 4.3). The panel also includes few unvaccinated participants (approximately 10%), but targeting the 15% moderately at-risk participants with the interesting survey lowers the bias for this variable. The cash incentive consistently reduces bias in the political participation variable, although only slightly (up to 0.2 on a five-point scale) (see Figure 4.6). No ASD substantially changes bias in the climate change seriousness variable, and none improves bias in the left-right political scale variable (see Figure 4.7). While there are clear ASD strategies that increase overall response rates, there are no particular ASD strategies that always improve bias across all the variables of interest.

To answer RQ2, although the nonresponse rate was reduced by targeting the interesting survey and the increased cash incentive to a certain percentage of panelists based on their predicted response propensities, the variable-wise nonresponse bias was not consistently reduced. In answer to RQ3, for a given k value, targeting the most at-risk participants with cash or an interesting topic was always better than targeting moderately at-risk participants. Higher k values yielded better response rates, although at a diminishing return. The impact on nonresponse bias across the variables of interest was mixed, highlighting how different cut-offs and targeting strategies can have different impacts on nonresponse bias. These result underline the value of a simulation setup as proposed in our paper to first test the possible outcomes of different ASD designs before

implementing them.

4.3.3 Robustness Checks

In Appendix Figure 4.9, Appendix Figure 4.10, and Appendix Figure 4.11 we present RC_1, in which we use 100 repeated simulations as if they were ten sequences of ten waves, calculating nonresponse bias based on overall responses and nonresponses across those waves. This approach allows us to examine whether the accumulation of nonresponses over several waves affects bias. We find similar results to those in Simulated Adaptive Designs, indicating that the passage of time does not affect nonresponse rates or bias in our simulation application.

Appendix Figure 4.12, Appendix Figure 4.13, Appendix Figure 4.14, and Appendix Figure 4.15 show RC_2, in which we duplicate rather than split the control group (see Analytical Strategy), with similar results as the main findings.

Appendix Figure 4.16, Appendix Figure 4.17, and Appendix Figure 4.18 show RC_3, in which we define nonresponse bias as the difference between the responder's mean value and the whole sample's mean value. In this lens, we see no improvements to bias in most indicators except for some ASD strategies on age and college education.

Finally, Appendix Figure 4.19, Appendix Figure 4.20, and Appendix Figure 4.21 show the same information as the main results Figures, but the y-axis is scaled around each variable's respective overall sample mean. This consistency across robustness checks suggests that changes in the simulation approach do not substantially alter the findings.

4.4 Discussion

Nonresponse is a major challenge for panel surveys, which need to maintain data quality over several waves to ensure the validity of results based on the data. In order to reduce biases due to wave nonresponse in panel surveys, adaptive survey designs need to (i) accurately predict panelists who are at risk of nonresponding and (ii) allocate treatments that improve those panelists' response propensities. The success of ASDs thereby critically depends on the effective use of predictions to target interventions and thus on understanding the interplay between those two components.

Our ex-post simulation approach allows survey researchers to study the impacts of different ASD strategies and thus how predictions may be best utilized to target interventions. Importantly, ex-post simulations that are informed by survey experiments can be used to realistically pre-screen the potential outcomes of various ASD strategies simultaneously. This overcomes the limitations of both assumptions-heavy synthetic simulations and single ASD trials that test only one specific ASD strategy at a time. Our approach instead allows estimating the outcomes of various interventions and targeting mechanisms with regard to response rates and variable-wise bias and to select the approach that best suits a given study's objectives.

The survey experiment and ex-post simulations in our paper demonstrate that using machine learning predictions that identify panel members at risk of nonresponse and targeting different treatments accordingly can be useful to reduce a panel's nonresponse

rate but not necessarily to address nonresponse bias. Prediction-based adaptive designs for panel surveys may thus be most effective in contexts with high nonresponse rates but low nonresponse bias. Applying this technique, e.g., after panel refreshment or in new panels, could encourage panelists with a high risk of nonresponse to respond to survey invitations.

Our study provides evidence for the interplay between prediction and treatment allocation: it matters which respondents are selected to receive a treatment, which treatment is used to enhance survey participation, and which treatment is allocated to which respondent group. We show how ex-post simulations, informed by survey experiments, allow researchers to assess these questions to build effective prediction-based ASDs tailored to their application context. Based on our findings, we encourage future research on ASDs to not only focus on the prediction or treatment part of ASDs but to take a comprehensive perspective. This perspective is what is most important for practical implementations of ASD and until now has not received attention by research. Finally, our findings show that prediction, treatment, and use of predictions to allocate treatments need to be adjusted to the goal of the ASD, as each decision will impact the ASD's performance. In other words, whether the overall nonresponse rate, group-specific nonresponse rates, or nonresponse bias (and in which variables), or a combination of these is the goal of the researchers will require adjustments of the ASD.

Coming to our specific findings, relying on a survey experiment, we showed that providing a high monetary incentive or an interesting survey, especially to panelists at risk of nonresponding, increased their response propensity either in the treatment or the subsequent wave. In contrast, shortening the survey had no effect (RQ1). RQ2 concerns what ASDs were effective and RQ3 concerns which thresholds and targeting mechanisms were most effective. In that regard, our ex-post simulations suggest that targeting just 15% of the most at-risk participants with a survey version that aligns with their expressed interests is as effective as randomly distributing extra cash to 35% of the sample. Specifically, nonresponse rates were, on average, one percentage point lower in the wave following this adaptive design compared to what would otherwise have been the case. Therefore, we suggest reviewing topic popularity or paying a higher monetary incentive to reduce nonresponse rates. Building on this, we encourage further research that examines the specifics of enhancing a survey's interestingness and the associated trade-offs with substantive research. Our findings suggest that these interventions should target respondents most at risk of wave nonresponse, in our simulation, the 15% lowest-propensity.

However, using machine learning predictions to target different treatments was less useful for reducing nonresponse bias. While the bias decreased for some variables, it increased for others. This finding aligns with the nature of nonresponse bias; as it is inherently a variable-specific bias, targeting based on a generic nonresponse prediction model is unlikely to mitigate biases across a diverse range of substantive variables. Simply applying the approaches we tested here thus will not remedy all biases due to wave nonresponse, only for some variables. Researchers should consider their specific variables of interest and whether they would be positively affected by the adaptive design. Optimizing the machine learning model to identify panelists at risk of nonresponding

and who are likely to increase nonresponse bias for specific variables of interest may reduce the variability. However, even if those respondents are identified, they may react differently to the chosen treatments. A prediction model that both identifies participants and assigns the optimal treatment to them may thus help reduce nonresponse bias. We welcome research that pursues this approach further and discusses trade-offs between optimizing response rates and response biases in panel surveys.

The limitations of our study yield possibilities for future research. First, we would welcome replications of the experiment in other panels, especially those with different panel designs and higher wave nonresponse. Our experiment was conducted in a mature panel with a high participation rate. Thus, low-propensity panelists had likely already attrited by this stage in the panel lifecycle, leaving limited room for improvement with the presented treatments. Second, while we acknowledge that the possibility of conducting our experiment in a large-scale probability-based panel helps generalize our findings, the number of low-propensity cases across our experimental groups limited statistical power for our analyses. In line with this, we welcome implementing different adaptive survey design strategies in survey practice to analyze their impact on response rates and bias. Third, we encourage replication of our experiment in other countries as survey climate and contexts might differ. We especially encourage cross-national studies to compare treatment effects of different interventions across countries.

4.5 Appendices

4.5.1 Replication

Materials for replicating this study are available at the following URL:

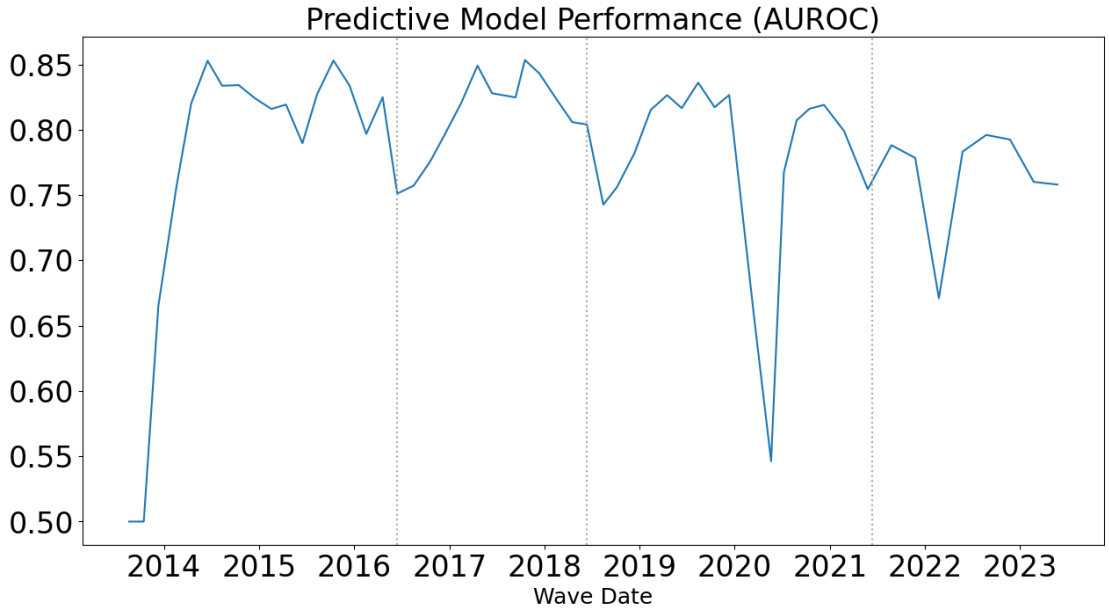
```
https://osf.io/h8k7u/?view_only=fb329f73a3cd4320bf882bbc21723273
```

Details of the precise GESIS Panel survey items that are used in each of the variables presented in this paper are located in:

```
src\settings\variable_settings.py
```

4.5.2 Methods Details

This section provides further information about the methodology of this study. This includes descriptions of the predictors used to fit the ML model (Appendix Table 4.1); descriptive statistics for the experiment sample sizes (Appendix Table 4.2) and the variables of interest across risk groups (Appendix Table 4.3); the predictive performance of the propensity model (Appendix Figure 4.8); the equations for the treatment effects and the specifications of the logistic regression models used to calculate p-values for treatment effects and the interaction effect between propensity and treatments (Appendix Table 4.4, Appendix Table 4.5).



Appendix Figure 4.8
Timeline of Area Under Receiver Operator Curve values for the predictive model across waves preceding the experiment.

Prediction Model

The propensity model was a random forest model with 100 trees of unlimited depth. Purity was measured by the Gini criterion. The maximum number of features used in each tree was the square root of the total number of predictors (Appendix Table 4.1). The minimum number of cases for a split was two and the minimum number of cases allowed in a leaf was one. The training data comprised of all cases of each participant-at-wave for all waves from October 2013 to May 2023. The predictors used to train the model are described in Appendix Table 4.1.

Measures

The formal definition of treatment effect in this context is as per appendix equation 4.1. The treatment effect is the difference in the mean of the outcome variable (in our case, nonresponse) between the group that received the treatment and the control group. We will calculate this treatment effect for both the experiment wave and the wave after the experiment in case of delayed effects.

$$E = \bar{X}_C - \bar{X}_T \tag{4.1}$$

Logistic Regression Models

We wish to analyze the robustness of any inferences we make about treatment effects. For this reason, we also conduct a logistic regression to estimate the effect of each treatment on nonresponse. We then assess the model's p-values to confirm the statistical significance of our results. The logistic regression model is unpenalized, without an

Appendix Table 4.1

Details of the predictors used in the propensity estimation model.

Variable	Value Range	Description
Sex Female	0, 1	-
Nonresponse This Wave	0, 1	-
Personal Income	0 - Infinity	Monthly
Household Income	0 - Infinity	Monthly
Is Married	0, 1	-
Household Size	0 - 5	The maximum allowed response is 5.
Survey Evaluation Interesting	0 - 5	-
Survey Evaluation Diverse	0 - 5	-
Survey Evaluation Important	0 - 5	-
Survey Evaluation Long	0 - 5	-
Survey Evaluation Difficult	0 - 5	-
Survey Evaluation Personal	0 - 5	-
Survey Evaluation Overall	0 - 5	-
Is Unemployed	0, 1	-
Participation Mode Online	0, 1	-
Survey Place Not Home	0, 1	Meaning the participant did not select the 'at home' option for where they filled out the survey.
Rolling Average Nonresponse Rate	0 - 1	Average nonresponse rate over all past invited waves.
Missing Sex Female	0, 1	-
Missing Personal Income	0, 1	-
Missing Household Income	0, 1	-
Missing Is Married	0, 1	-
Missing Age	0, 1	-
Missing Household Size	0, 1	-
Missing Survey Evaluation Interesting	0, 1	-
Missing Survey Evaluation Diverse	0, 1	-
Missing Survey Evaluation Important	0, 1	-
Missing Survey Evaluation Long	0, 1	-
Missing Survey Evaluation Difficult	0, 1	-
Missing Survey Evaluation Personal	0, 1	-
Missing Survey Evaluation Overall	0, 1	-
Missing Is Unemployed	0, 1	-
Missing Participation Mode Online	0, 1	-
Missing Survey Place Not Home	0, 1	-
Age: (17.938 30.4]	0, 1	Indicates that age, derived from the year of birth and the survey date, is within this range.
Age: (30.4 42.8]	0, 1	-
Age: (42.8 55.2]	0, 1	-
Age: (55.2 67.6]	0, 1	-
Age: (67.6 80.0]	0, 1	-

Appendix Table 4.2

Details of the sample sizes and number of nonrespondents in each group presented in the main text.

	Treatment_Group	Control	Short	Interesting	Cash
Sample Size	Whole Sample	1,213	1,214	1,214	1,214
	Moderate Risk	266	304	276	285
	High Risk	287	324	300	303
Nonrespondents in W_0	Whole Sample	94	97	85	56
	Moderate Risk	55	56	54	43
	High Risk	55	59	57	44
Nonrespondents in W_1	Whole Sample	89	99	65	63
	Moderate Risk	43	56	32	37
	High Risk	46	59	35	39

Appendix Table 4.3

Descriptive statistics for each risk group across the variables of interest. See Appendix Table 4.2 for details on the number of supports in each group.

Targeting Strategy	K	Metric	Age	College	HH Income	Health	Vaccination Status	Political Participation	Climate Change Serious	Left-Right
All	-	Mean	57.535	0.512	3016.313	4.895	0.914	1.625	5.087	5.527
All	-	Std	14.408	0.500	1917.967	1.439	0.281	0.932	1.495	1.885
High	0.15	Mean	51.570	0.548	2914.286	4.805	0.882	1.525	4.935	5.283
High	0.15	Std	15.741	0.498	2025.603	1.536	0.323	0.934	1.720	2.035
High	0.25	Mean	52.474	0.557	2898.847	4.837	0.886	1.556	4.960	5.375
High	0.25	Std	15.698	0.497	2020.651	1.505	0.319	0.926	1.657	2.005
High	0.35	Mean	52.647	0.552	2945.733	4.856	0.886	1.576	5.025	5.436
High	0.35	Std	15.562	0.497	2029.174	1.493	0.317	0.935	1.616	1.970
Moderate	0.15	Mean	53.216	0.558	2827.473	4.815	0.887	1.571	4.982	5.471
Moderate	0.15	Std	15.541	0.497	2015.228	1.472	0.316	0.924	1.588	1.943
Moderate	0.25	Mean	52.309	0.556	2884.792	4.828	0.884	1.555	4.974	5.377
Moderate	0.25	Std	15.638	0.497	2011.361	1.505	0.320	0.930	1.656	2.001
Moderate	0.35	Mean	52.309	0.556	2884.792	4.828	0.884	1.555	4.974	5.377
Moderate	0.35	Std	15.638	0.497	2011.361	1.505	0.320	0.930	1.656	2.001

intercept term and features only a single predictor which is the binary treatment presence (treatment = 1, control = 0). We fit a logistic regression model for each of the three treatments. We fit the model with cases only for that respective treatment group and the control group. The coefficient of the treatment term will allow us to calculate the odds ratio for the change in nonresponse probability with the treatment. The p-value of the coefficient will indicate the probability that the observed treatment effect is due to random chance. In appendix equation 4.2 we fit a logistic regression, but now we include a term for each participant's response propensity (R) and the interaction between the treatment and the propensity (R:T). The coefficient of the interaction term shall indicate the rate at which the treatment effect varies with the propensity level.

$$P(Y = 1) = \frac{e^{R \cdot \beta_R + T \cdot \beta_T + (R \cdot T) \cdot \beta_{R:T}}}{1 + e^{R \cdot \beta_R + T \cdot \beta_T + (R \cdot T) \cdot \beta_{R:T}}} \tag{4.2}$$

The nonresponse bias is given by the absolute difference in means between responders and nonresponders (appendix equation 4.3)

$$\text{Nonresponse Bias} = |\mu_{\text{responders}} - \mu_{\text{nonresponders}}| \tag{4.3}$$

4.5.3 Additional Results

This section provides supplementary results to those presented in the main text. This includes the outputs of the logistic regression models (Appendix Table 4.4, Appendix Table 4.5); the mean values for the variables of interest across responder-nonresponder groups (Appendix Table 4.6); the subsequent nonresponse bias values (Appendix Table 4.7); and the treatment effects (Appendix Table 4.8). The correlation coefficients with p-value decorators from a Pearson test between the variables of interest and nonresponse behavior are presented in Appendix Table 4.9. Finally, the results of the variations on the simulation procedure, which are used as robustness checks, are presented across the remaining figures in this section.

Appendix Table 4.4

Details of the treatment effect coefficient in the logistic regression models fitted to each treatment-control group pair. Models are fitted on nonresponse in both the wave of the experiment and the wave afterwards. See Appendix Table 4.2 for details on the number of supports in each group.

Target	Treatment	Coefficient	Odds Ratio	Standard Error	P-Value	K
Experiment Wave	Cash Incentive	-0.552	0.576	0.174	0.001	Whole Sample
Experiment Wave	Cash Incentive	-0.343	0.709	0.218	0.116	High Risk
Experiment Wave	Cash Incentive	-0.356	0.701	0.220	0.106	Moderate Risk
Experiment Wave	Interesting Survey	-0.110	0.896	0.156	0.481	Whole Sample
Experiment Wave	Interesting Survey	-0.053	0.949	0.208	0.800	High Risk
Experiment Wave	Interesting Survey	-0.061	0.941	0.208	0.770	Moderate Risk
Experiment Wave	Short Survey	0.033	1.034	0.151	0.826	Whole Sample
Experiment Wave	Short Survey	-0.145	0.865	0.209	0.488	High Risk
Experiment Wave	Short Survey	-0.144	0.866	0.212	0.498	Moderate Risk
Post Experiment Wave	Cash Incentive	-0.369	0.691	0.170	0.030	Whole Sample
Post Experiment Wave	Cash Incentive	-0.275	0.760	0.234	0.239	High Risk
Post Experiment Wave	Cash Incentive	-0.231	0.794	0.236	0.327	Moderate Risk
Post Experiment Wave	Interesting Survey	-0.336	0.714	0.168	0.046	Whole Sample
Post Experiment Wave	Interesting Survey	-0.390	0.677	0.240	0.104	High Risk
Post Experiment Wave	Interesting Survey	-0.372	0.689	0.241	0.122	Moderate Risk
Post Experiment Wave	Short Survey	0.115	1.121	0.152	0.451	Whole Sample
Post Experiment Wave	Short Survey	0.190	1.209	0.218	0.384	High Risk
Post Experiment Wave	Short Survey	0.158	1.171	0.223	0.478	Moderate Risk

Appendix Table 4.5

Coefficients of the logistic regression models which include interaction terms between treatment and predicted response propensities. Models are fitted on nonresponse in both the wave of the experiment and the wave afterwards. See Appendix Table 4.2 for details on the number of supports in each group.

Target	Treatment	Variable	Coefficient	P-value	Standard Error	Odds Ratio
Experiment Wave	Cash Incentive	Cash Incentive	1.574	0.001	0.489	4.826
Experiment Wave	Cash Incentive	Propensity	-2.844	0.000	0.126	0.058
Experiment Wave	Cash Incentive	Cash × Propensity	-2.528	0.000	0.604	0.080
Experiment Wave	Interesting Survey	Interesting Survey	1.703	0.001	0.494	5.489
Experiment Wave	Interesting Survey	Propensity	-2.844	0.000	0.126	0.058
Experiment Wave	Interesting Survey	Interesting × Propensity	-1.993	0.001	0.580	0.136
Experiment Wave	Short Survey	Short Survey	1.272	0.004	0.438	3.568
Experiment Wave	Short Survey	Propensity	-2.844	0.000	0.126	0.058
Experiment Wave	Short Survey	Short × Propensity	-1.386	0.008	0.520	0.250
Post Experiment Wave	Cash Incentive	Cash Incentive	0.815	0.084	0.471	2.260
Post Experiment Wave	Cash Incentive	Propensity	-2.861	0.000	0.127	0.057
Post Experiment Wave	Cash Incentive	Cash × Propensity	-1.392	0.014	0.565	0.249
Post Experiment Wave	Interesting Survey	Interesting Survey	0.632	0.210	0.504	1.882
Post Experiment Wave	Interesting Survey	Propensity	-2.861	0.000	0.127	0.057
Post Experiment Wave	Interesting Survey	Interesting × Propensity	-1.078	0.068	0.591	0.340
Post Experiment Wave	Short Survey	Short Survey	0.688	0.112	0.433	1.990
Post Experiment Wave	Short Survey	Propensity	-2.861	0.000	0.127	0.057
Post Experiment Wave	Short Survey	Short × Propensity	-0.649	0.203	0.510	0.523

Appendix Table 4.6

Mean values of the variables of interest. N = 1,214 except for the Control group which is 1,213.

Variable	Nonresponse In Post Experiment Wave	Cash	Control	Interesting	Short
Age	0.000	57.474	58.221	57.783	57.910
Age	1.000	55.429	52.236	51.831	53.192
College Educated	0.000	0.517	0.499	0.521	0.497
College Educated	1.000	0.492	0.584	0.646	0.525
Household Income	0.000	2997.567	3021.619	3078.764	3023.587
Household Income	1.000	2687.302	2848.315	3435.385	2452.525
Health	0.000	4.861	4.926	4.909	4.902
Health	1.000	4.651	4.809	4.923	4.919
Vaccination Status	0.000	0.922	0.918	0.904	0.914
Vaccination Status	1.000	0.873	0.888	0.908	0.929
Political Participation	0.000	1.632	1.644	1.626	1.611
Political Participation	1.000	1.683	1.517	1.569	1.566
Climate Change Serious	0.000	5.052	5.106	5.130	5.102
Climate Change Serious	1.000	5.159	4.944	4.877	4.848
Left-Right	0.000	5.578	5.628	5.452	5.531
Left-Right	1.000	5.143	5.191	5.231	5.343

Appendix Table 4.7

Nonresponse Bias Values of the Variables of Interest.

Variable	Cash	Control	Interesting	Short
Age	2.046	5.985	5.953	4.718
Climate Change Serious	-0.107	0.162	0.253	0.254
College Educated	0.025	-0.085	-0.125	-0.028
Health	0.210	0.117	-0.014	-0.017
Household Income	310.266	173.305	-356.620	571.062
Left-Right	0.435	0.437	0.221	0.188
Political Participation	-0.051	0.127	0.057	0.045
Vaccination Status	0.049	0.031	-0.003	-0.015

Appendix Table 4.8

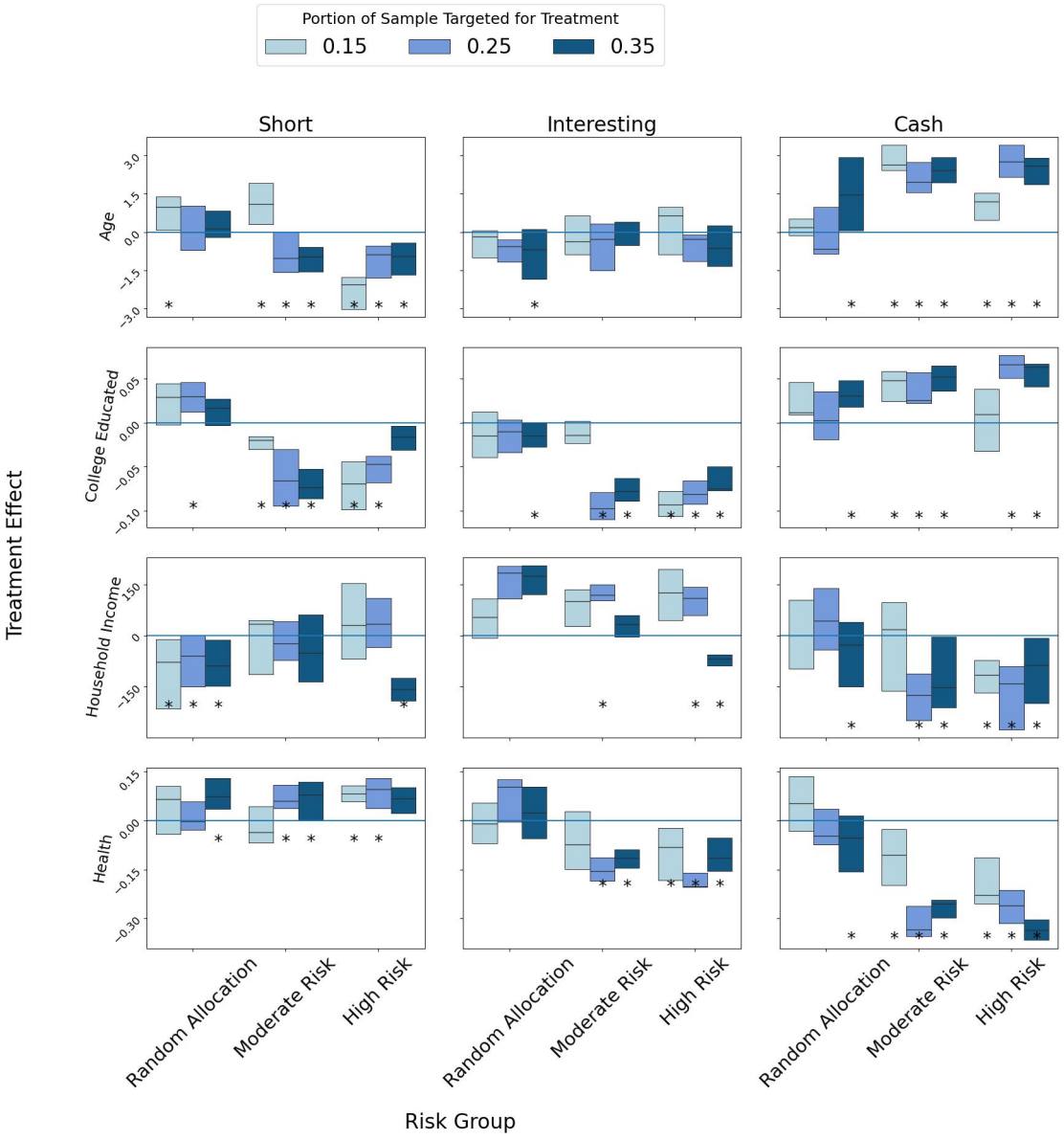
Treatment effects for each of the variables of interest.

Variable	Interesting	Cash	Short
Age	0.032	3.939	1.266
Climate Change Serious	-0.091	0.269	-0.092
College Educated	0.040	-0.110	-0.057
Health	0.131	-0.093	0.134
Household Income	529.925	-136.961	-397.758
Left-Right	0.216	0.002	0.250
Political Participation	0.071	0.178	0.082
Vaccination Status	0.034	-0.018	0.046

Appendix Table 4.9

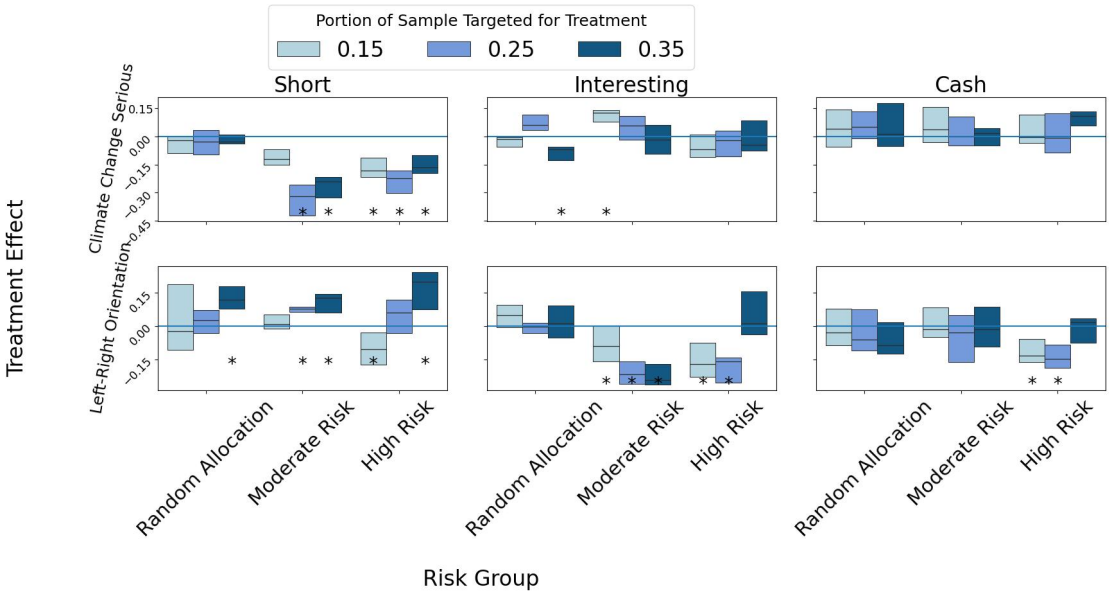
How the substantive survey variables of interest correlate (Pearson) with Nonresponse in the Next Wave. $*p < 0.05 * *p < 0.01$. *N = 2,428.*

Variable	Correlation with nonresponse in the next wave
Age	-0.14**
Health	-0.10**
Left-Right	-0.10**
Political Participation	-0.08**
Household Income	-0.06**
Vaccination Status	-0.05**
Climate Change Serious	-0.05**
College Educated	-0.02**

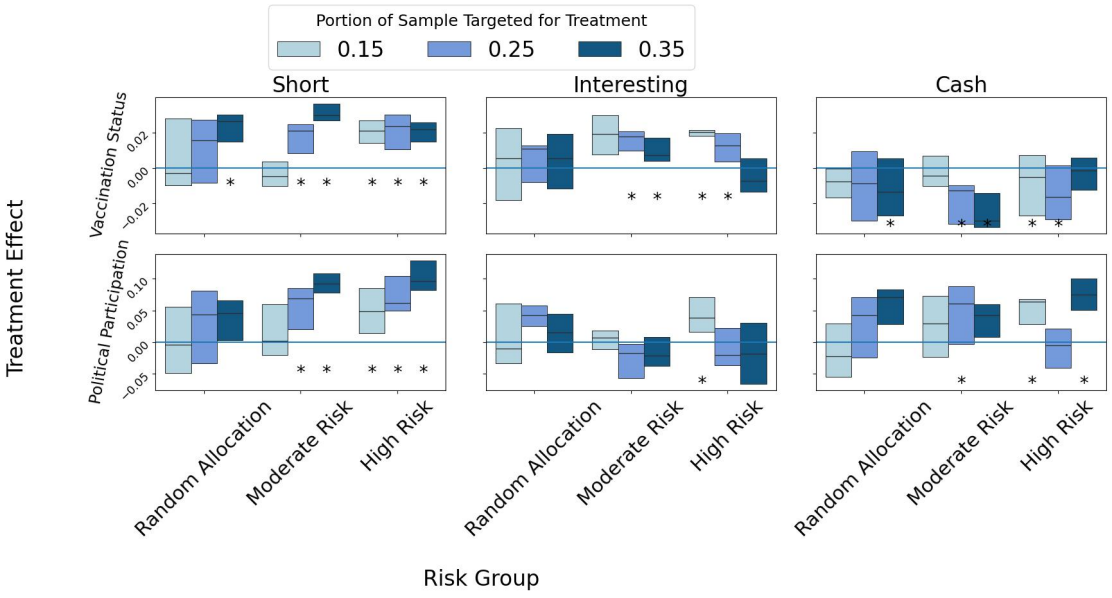


Appendix Figure 4.9

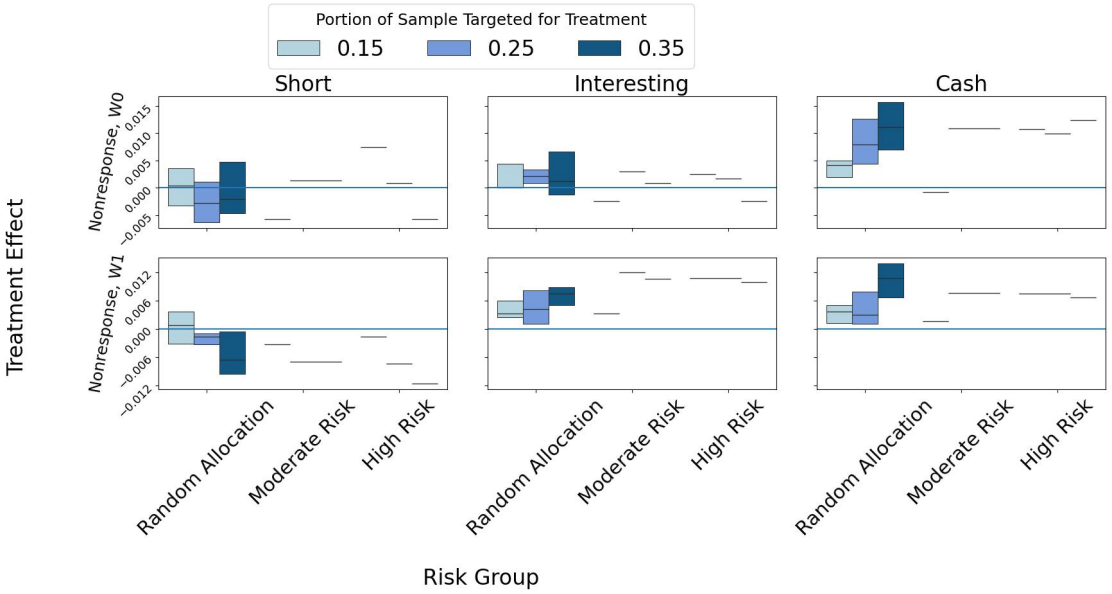
Long term changes part one: A variation of the simulation where we simulate a sequence of ten waves using the method described to simulate a single wave, We then calculate variable-wise nonresponse bias for each variable of interest across all responses/nonresponses across the ten waves. Each sequence of ten waves is repeated 10 times to account for randomness in the simulation. $*p < 0.05 * *p < 0.01$.



Appendix Figure 4.10
Long term changes part two. $*p < 0.05 * *p < 0.01$.

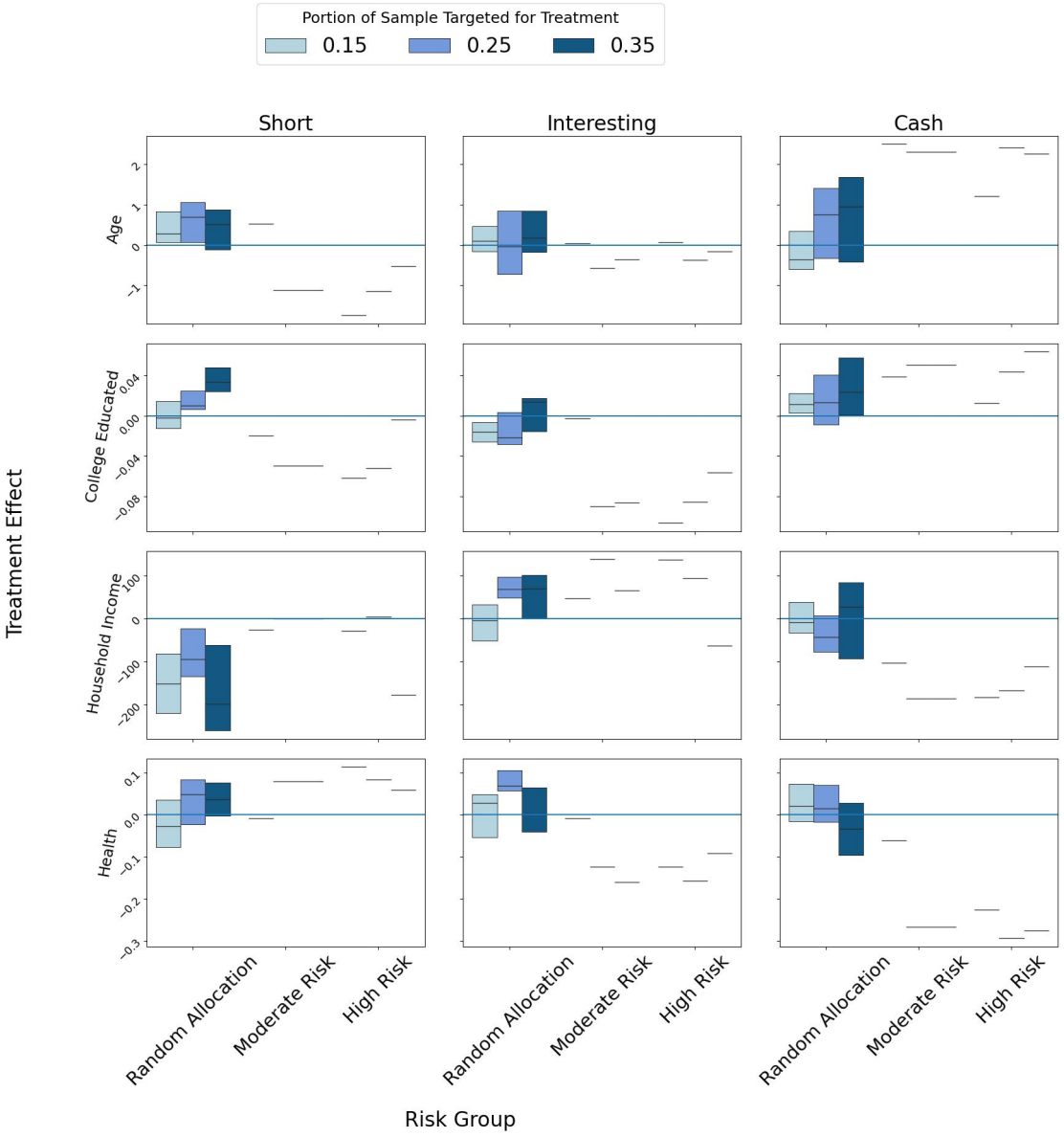


Appendix Figure 4.11
Long term changes part three. $*p < 0.05 * *p < 0.01$.

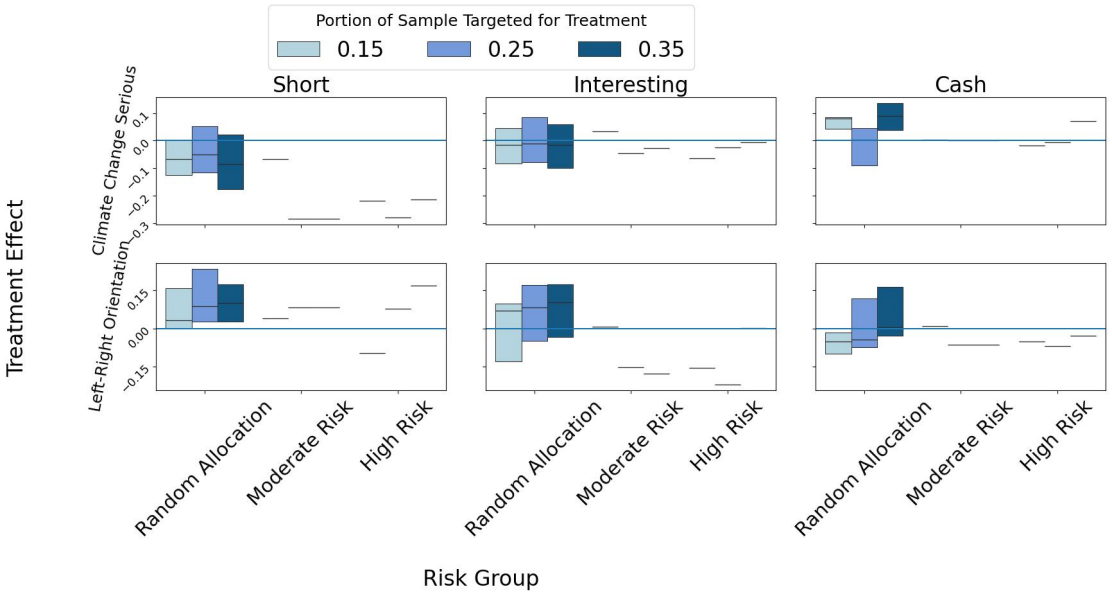


Appendix Figure 4.12

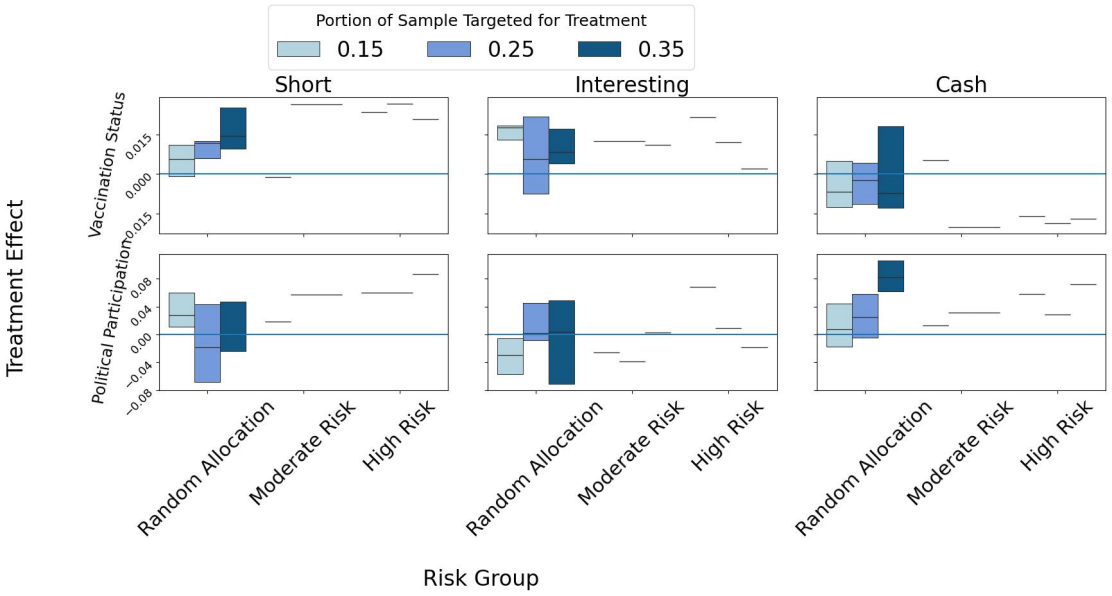
Alternative simulation part one: A variation of the simulation where the control group was not split, but instead the k% of treated participant replaces the equivalent k% of control participants according to the targeting strategy. This means that there is only an element of randomness in the random allocation baselines, ad he outcome of the ML-based targeting strategies is deterministic. We make 10 repetitions of the random allocation baseline. because the outcome is no long a distribution, there are no p-values.



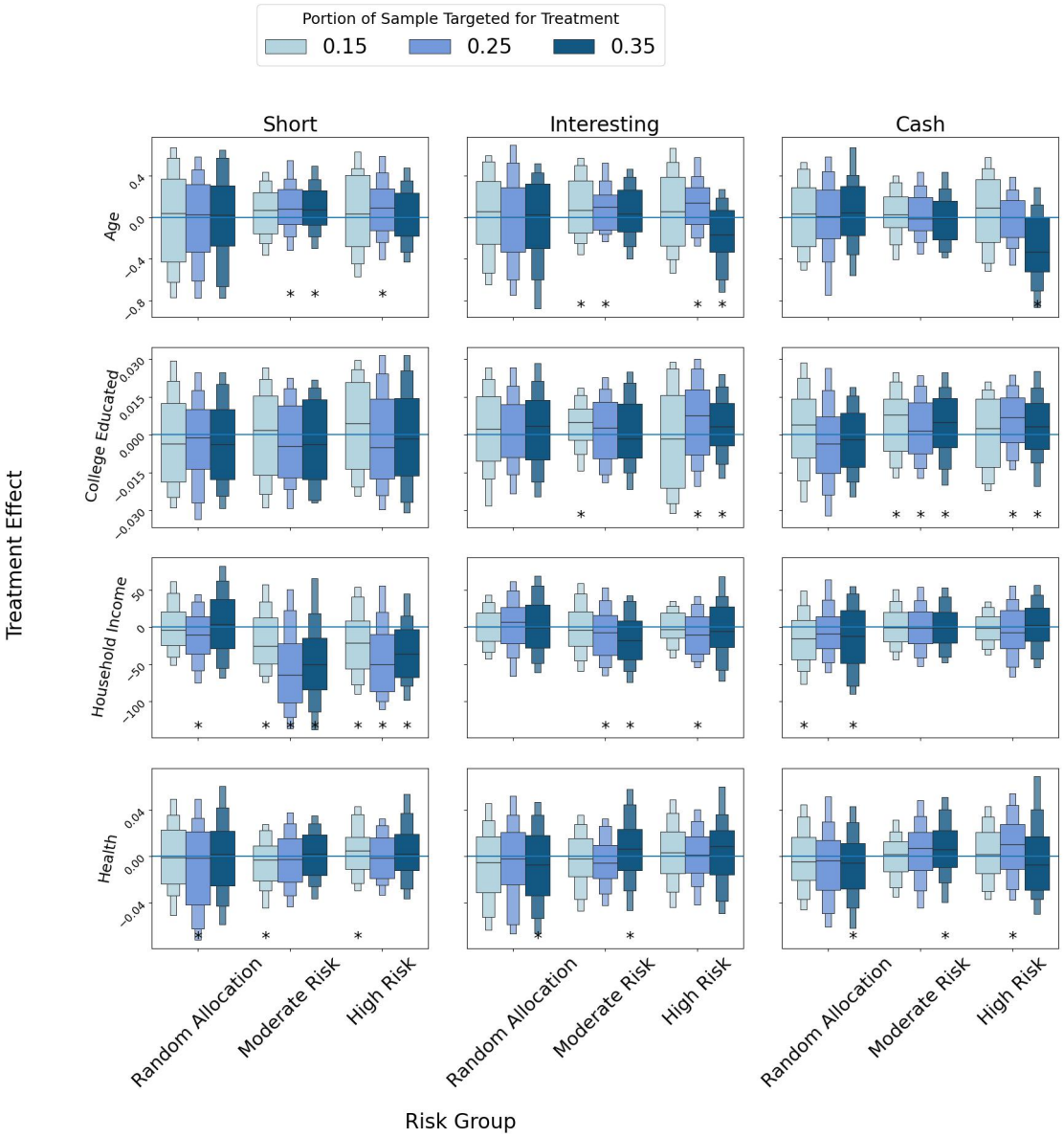
Appendix Figure 4.13
Alternative simulation part two.



Appendix Figure 4.14
Alternative simulation part three.

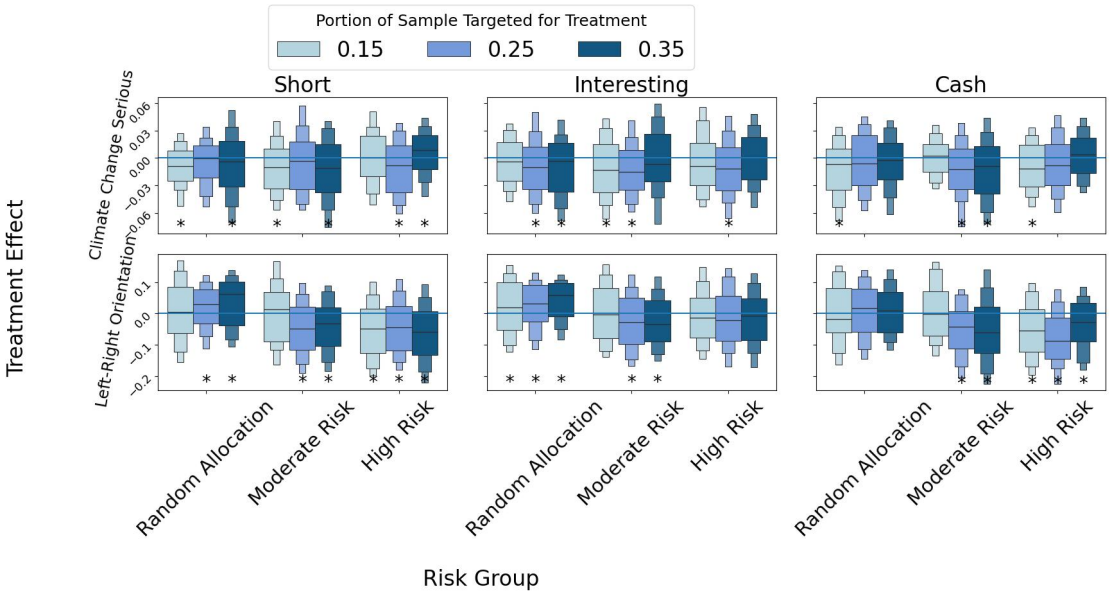


Appendix Figure 4.15
Alternative simulation part four.



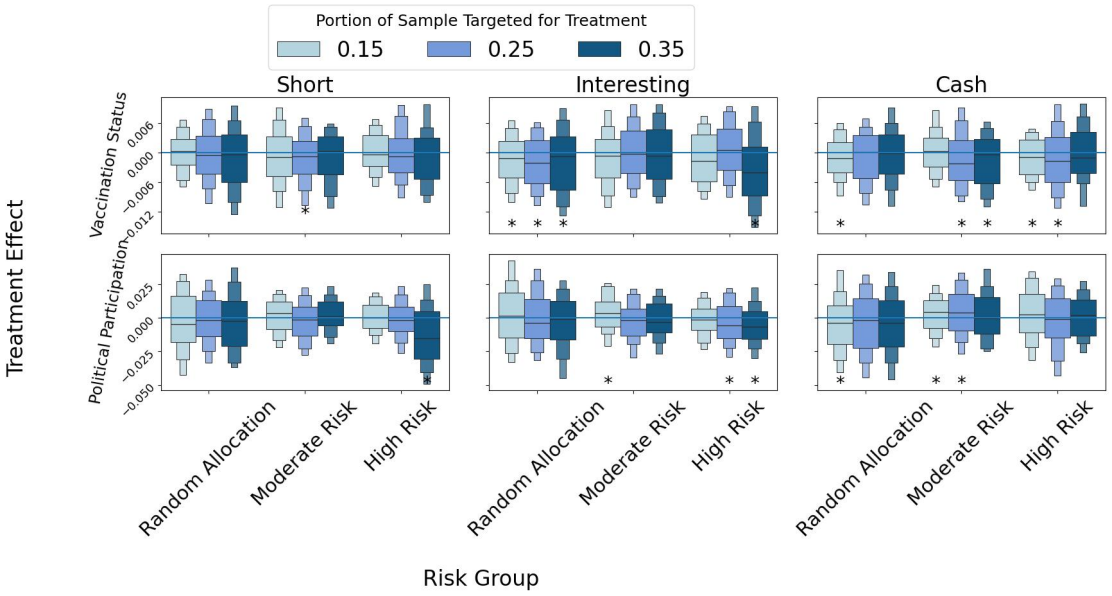
Appendix Figure 4.16

Alternative bias measurement part one: Ranges of changes in nonresponse bias values between simulated scenarios with and without adaptive design. Here, nonreponse bias is calculated as the difference between the overall sample mean as of the experiment wave (benchmark value) and the mean of the respondents in the ASD scenario. The y-axis values in the plot are the amount by which the benchmark value and 'observed' means converged due to the ASD. $*p < 0.05 * *p < 0.01$.



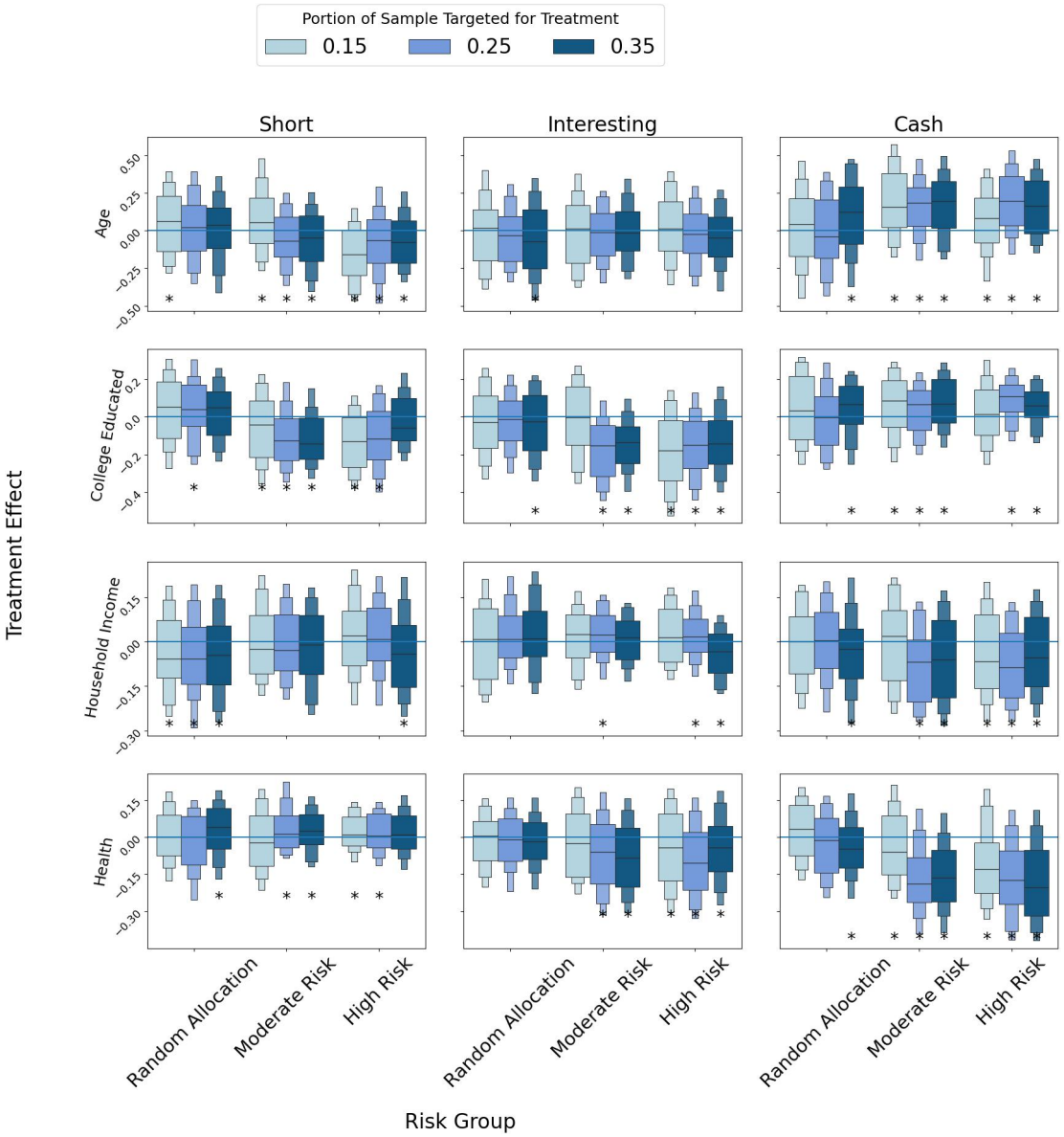
Appendix Figure 4.17

Alternative bias measurement part two. $*p < 0.05 * *p < 0.01$.



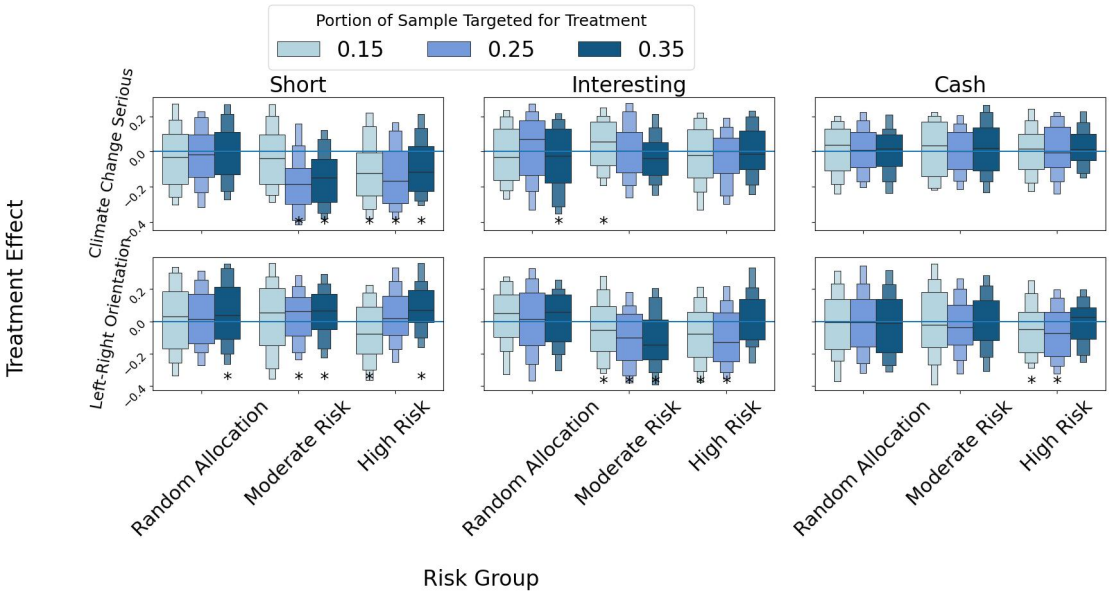
Appendix Figure 4.18

Alternative bias measurement part three. $*p < 0.05 * *p < 0.01$.



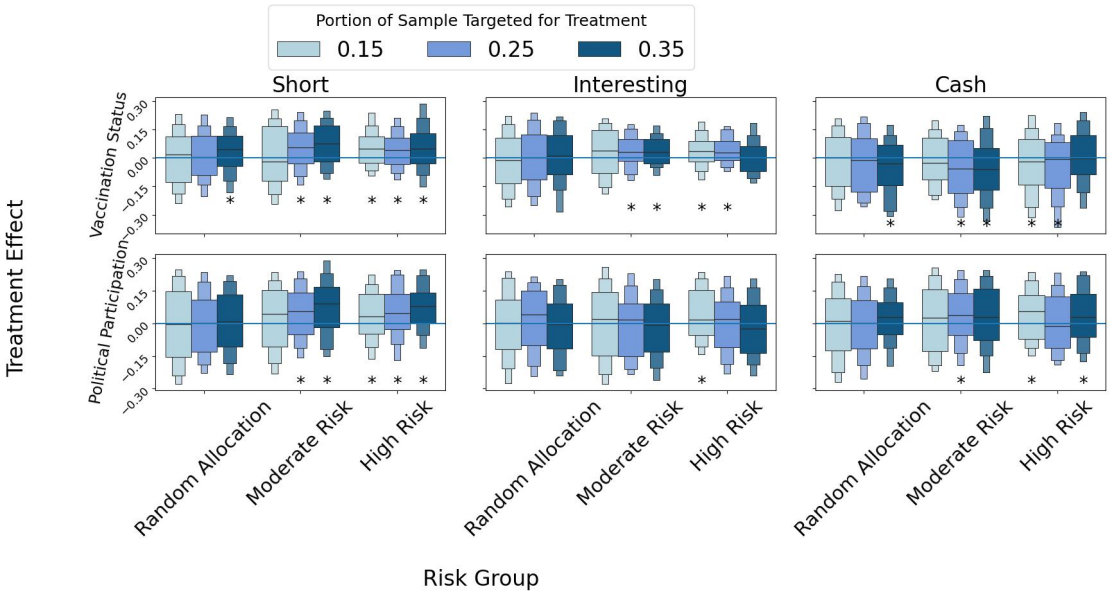
Appendix Figure 4.19

Scaled simulation results part one: A revised version of the simulation results in which every variable has been scaled around the mean so that the y-axis is the number of standard deviations from the overall variable mean changed due to the given ASD. $*p < 0.05 * *p < 0.01$.



Appendix Figure 4.20

Scaled simulation results part two. $*p < 0.05 * *p < 0.01$.



Appendix Figure 4.21

Scaled simulation results part three. $*p < 0.05 * *p < 0.01$.

4.6 References

AAPOR (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9th edition. The American Association for Public Opinion Research.

Allison, Paul D. (2009). *Fixed Effects Regression Models*. SAGE Publications, Inc.

Andreß, Hans-Jürgen, Katrin Golsch, and Alexander W. Schmidt (2013). *Applied Panel Data Analysis for Economic and Social Surveys*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Bach, Ruben L, Stephanie Eckman, and Jessica Daikeler (2020). "Misreporting Among Reluctant Respondents". In: *Journal of Survey Statistics and Methodology* 8.3, pp. 566–588.

Beste, Jonas et al. (2023). "Case Prioritization in a Panel Survey Based on Predicting Hard to Survey Households by Machine Learning Algorithms: An Experimental Study". In: *Survey Research Methods* 17.3, pp. 243–268.

Bethlehem, Jelke (1988). "Reduction of Nonresponse Bias through Regression Estimation". In: *Journal of Official Statistics* 4, pp. 251–260.

Bosnjak, Michael et al. (2018). "Establishing an Open Probability-Based Mixed-Mode Panel of the General Population in Germany: The GESIS Panel". In: *Social Science Computer Review* 36.1. Number: 1, pp. 103–115.

Coffey, Stephanie et al. (2020). "What do you think? Using expert opinion to improve predictions of response propensity under a Bayesian framework". In: *Methoden, daten, analysen* 14.2. Publisher: NIH Public Access.

Collins, John and Christoph Kern (2024). "Longitudinal Nonresponse Prediction with Time Series Machine Learning". In: *Journal of Survey Statistics and Methodology*, smae037.

Felderer, Barbara, Jannis Kueck, and Martin Spindler (2023). "Using Double Machine Learning to Understand Nonresponse in the Recruitment of a Mixed-Mode Online Panel". In: *Social Science Computer Review* 41.2, pp. 461–481.

GESIS (2022). *GESIS - Leibniz Institute for the Social Sciences* URL: `https://www.gesis.org/en/gesis-survey-guidelines/statistics/weighting-overview/weighting`.

— (2024). *GESIS Panel Standard Edition. GESIS, Cologne. ZA5665 Datafile Version 54.0.0, doi: 10.4232/1.14386*. 2024.

Groves, Robert M. (2006). "Nonresponse Rates and Nonresponse Bias in Household Surveys". In: *Public Opinion Quarterly* 70.5, pp. 646–675.

— (2011). *Survey Methodology*. Ed. by Floyd J. Fowler et al. 2nd ed. Wiley Series in Survey Methodology. Wiley, Somerset.

Gummer, Tobias (2020). "Adaptive and Responsive Survey Designs". In: *SAGE Research Methods Foundations*. Ed. by P. Atkinson et al. 2020.

Gummer, Tobias and Jan Eric Blumenstiel (2018). "Experimental Evidence on Reducing Nonresponse Bias through Case Prioritization: The Allocation of Interviewers". In: *Field Methods* 30.2. Publisher: SAGE Publications Inc, pp. 124–139.

Kern, Christoph, Bernd Weiß, and Jan-Philipp Kolb (2021). "Predicting Nonresponse in Future Waves of a Probability-Based Mixed-Mode Panel with Machine Learning*". In: *Journal of Survey Statistics and Methodology* 11.1, pp. 100–123.

Kocar, Sebastian and Nicholas Biddle (2022). "The power of online panel paradata to predict unit nonresponse and voluntary attrition in a longitudinal design". In: *Quality & Quantity*.

Koch, Achim and Michael Blohm (2016). "Nonresponse Bias". In: *GESIS Survey Guidelines*. GESIS-Leibniz-Institut Für Sozialwissenschaften, Mannheim, Germany.

Lynn, Peter, ed. (2009). *Methodology of Longitudinal Surveys*. Wiley Series in Survey Methodology. John Wiley & Sons, New York, NY.

— (2016). "Targeted Appeals for Participation in Letters to Panel Survey Members". In: *Public Opinion Quarterly* 80.3, pp. 771–782.

McCarthy, Jaki, James Wagner, and Herschel Lisette Sanders (2017). "The Impact of Targeted Data Collection on Nonresponse Bias in an Establishment Survey: A Simulation Study of Adaptive Survey Design". In: *Journal of Official Statistics* 33.3, pp. 857–871.

Mulder, J and N Kieruj (2018). *Preserving Our Precious Respondents: Predicting and Preventing Non-Response and Panel Attrition by Analyzing and Modeling Longitudinal Survey and Paradata Using Data Science Techniques*. 2018.

Roßmann, J. and T. Gummer (2016). "Using Paradata to Predict and Correct for Panel Attrition". In: *Social Science Computer Review* 34.3, pp. 312–332.

Schouten, Barry, Fannie Cobben, et al. (2016). "Does more balanced survey response imply less non-response bias?" In: *Journal of the Royal Statistical Society Series A* 179.3. Number: 3 Publisher: Royal Statistical Society, pp. 727–748.

Schouten, Barry, Andy Peytchev, and James Wagner (2017). *Adaptive Survey Design*. Chapman and Hall.

Soeder, Jana et al. (2024). "Changes in Attitudes Toward Infection Control Measures in the Workplace During the COVID-19 Pandemic—Longitudinal Data From Employees in Germany". In: *Occupational Medicine* 74 (Supplement_1).

Trappmann, Mark, Tobias Gramlich, and Alexander Mosthaf (2015). "The effect of events between waves on panel attrition". In: *Survey Research Methods* 9.1, pp. 31–43.

Wagner, James (2013). "Adaptive contact strategies in telephone and face-to-face surveys". In: *Survey research methods*. Vol. 7. Issue: 1, pp. 45–55.

Wagner, James et al. (2012). "Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection". In: *Journal of Official Statistics*, p. 23.

Wagner, James R (2008). "Adaptive Survey Design to Reduce Nonresponse Bias". In: *University of Michigan*.

Watson, Nicole and Alexandru Cernat (2023). "Simulating the Consequences of Adaptive Survey Design in Two Household Panel Studies". In: *Journal of Survey Statistics and Methodology* 11.4, pp. 806–828.

Zhang, Shiyu and James Wagner (2024). "The Additional Effects of Adaptive Survey Design Beyond Post-Survey Adjustment: An Experimental Evaluation". In: *Sociological Methods & Research* 53.3, pp. 1350–1383.

Zhang, Shiyu, Brady T. West, et al. (2024). "Incorporating Adaptive Survey Design in a Two-Stage National Web or Mail Mixed-Mode Survey: An Experiment in the American Family Health Study". In: *Journal of Survey Statistics and Methodology* 12.3. Publisher: Oxford University Press, pp. 578–592.

Zinn, Sabine and Timo Gnambs (2022). "Analyzing nonresponse in longitudinal surveys using Bayesian additive regression trees: A nonparametric event history analysis". In: *Social Science Computer Review* 40.3, pp. 678–699.

5 Predicting Australian Federal Electoral Seats with Machine Learning

Abstract

I expand the international literature on election forecasting with the first application of machine learning (ML) to the Australian context. I apply these models to five elections from 2010 to 2022 and compare them against the dominant forecasting tool in Australia, the Mackerras pendulum. I evaluate these models' accuracy in predicting the winning party for each electoral seat and estimating the total number of seats won by each party. Pendulum forecasts corrected with an extra trees model that incorporates state effects, seat-level unemployment rate, and vote share history, predicts up to 15 additional seats correctly six to three months before each election. The traditional pendulum is increasingly strained by polling errors and a larger crossbench. New modeling techniques will only emerge through experimentation. This study demonstrates the potential for ML-based election forecasting in Australia and provides a starting point for further efforts to surpass the pendulum.

5.1 Introduction

Australia's dominant federal election forecasting model is the Mackerras pendulum (Mackerras, 1976; Browne, 2022), which predicts the number of seats won by each major party in the House of Representatives. While there are many variants of this model, the typical pendulum method calculates the change in the "two-party preferred" (TPP) poll[1] for each major party since the previous election (also known as the swing). It assumes a corresponding, uniform change in vote share across all seats relative to the previous election (Mackerras, 1976). The Australian Broadcasting Corporation's (ABC) pendulum has correctly predicted the party which forms government in 17 out of the 20 elections held since 1972, miscalling only 1998, 2010, and 2019 (Goot, 2022; Green, 2019; Green, 2022; Green, 2016). However, the pendulum has several critical issues. Firstly, seats held by minor parties (also called crossbenchers) are, by default, always predicted to stay with their incumbent. This naive assumption led to only small inaccuracies in the past when minor parties won few seats. However, this changed in 2022, when the crossbench grew from 6 to 16 seats out of 151 (Green, 2019; Green, 2022). Secondly, the pendulum relies heavily on accurate polling, which is why the 2019 election was

[1]This is a poll in which participants are asked to specify only which of the two major parties they would most prefer and all other parties are ignored.

miscalled, given a year of unusually large polling errors (Goot, 2021). Thirdly, while the pendulum successfully predicts the number of seats won by each party, it is imprecise in predicting the winner of each seat. When a seat changes party, the pendulum predicts the correct winner less than half the time (Browne, 2022). The pendulum achieves good seat count accuracy because its opposing misclassifications balance out. The pendulum's primary function is to estimate seat count. Seat-level winner prediction is not its intended use. However, the pendulum is still Australia's leading method for seat outcome forecasting because nothing has yet been developed that outperforms it (Browne, 2022).

This paper presents machine learning (ML) models that aim to improve over the pendulum in these areas. Specifically, I present "ML-based synthetic" models (Lewis-Beck and Dassonneville, 2015). This approach combines historical polls and past electoral outcomes with economic and demographic predictors (also called fundamentals; for an overview, see (Hummel and Rothschild, 2014) to fit ML models that predict seat-level outcomes. As well as adding to the international literature on ML in election forecasting, I demonstrate a novel technique: correcting pendulum predictions with ML. Firstly, for the benefit of international readers, I will provide a summary of the Australian electoral system (Section 5.1.1). Next, I will describe other attempts to improve Australian electoral forecasting and argue that there has yet to be any major improvement over the pendulum (5.1.2). I will then review the literature on ML-based electoral forecasting in other democracies and show that this technique has potential and is worth trialing in Australia (5.1.3). I will close this section by listing the criteria against which I will evaluate the ML-based electoral models presented in this paper (5.1.4).

5.1.1 The Australian Electoral System

Australia has two parliamentary houses. This paper focuses on the House of Representatives, which had 150-151 seats in the period of study, which are comprised of roughly equally populated, geographically contiguous zones. Australia has a multiparty system, although dominated by two major parties. Government terms are a maximum of three years, but the sitting government may select the exact election date. In a federal election, eligible voters receive a fine if they fail to go to a polling site (but do not necessarily have to vote), which is why voter turnout is generally above 90% (Australian Electoral Commission, 2011). Australia has preferential voting, which means that when no candidate has a majority in a seat after all first preferences are counted, the least popular candidate is eliminated. That party's votes are redistributed according to the second preferences. This process repeats until a candidate reaches 50% of that seat's votes (Australian Electoral Commission, 2019b).

5.1.2 Previous Attempts to Improve Forecasting

The Mackerras pendulum dominates Australian electoral forecasting despite several attempts to outperform it. Researchers have explored logistic regression models that incorporate economic indicators and betting markets, finding reasonable predictive performance in anticipating the winning party (Greenop-Roberts, 2022; Jackman and Marks,

1994; Jackman, 2005; Leigh and Wolfers, 2006; Wolfers and Leigh, 2002). Some of these models successfully predicted the 2010 election-winning party, although they were incorrect about other elections (Greenop-Roberts, 2022). However, none of this research has attempted to improve the classification of seat winners. Kefford (2021) interviewed many campaigners from major Australian parties and learned that they conduct internal electoral modeling, but the results of these exercises are not public. YouGov conducted a multilevel regression with poststratification (MRP) model for the 2022 election, which used a survey of around 18,000 respondents to model the outcome of each federal seat (YouGov, 2021). The results were released two weeks before election day, but the most likely outcome (80 seats for the winning Labour Party) was slightly worse than if the same data had been used in a traditional pendulum (a predicted 79 seats for the Labour Party, with the actual outcome being 77; (Pack, 2023). The MRP model misclassified nine seats and counted six more as too close to call (Bowe, 2021a; Bowe, 2021b; Lewis-Beck and Dassonneville, 2015; Pack, 2023; Rustika, 2021; YouGov, 2021). This paper will compare that performance with the ML models for the 2022 election, but it is important to note that the lead time was only two weeks and only for one election, which limits how well that approach can be compared to this study until the MRP is trialed in more elections.

5.1.3 ML-Based Electoral Forecasting

I expect ML models to perform better than the pendulum because they can learn from past elections to correct for historical polling errors and to vary the seat-by-seat effect of national polls by local features such as the unemployment rate or median income. Other democracies have trialed ML-based synthetic models for predicting seat-level (or the equivalent electoral unit) outcomes (Argandoña-Mamani et al., 2024; Fisher, 2016; Graefe, 2019; Gschwend et al., 2022; Kang and Oh, 2023; Linzer, 2013; Mackerras, 1976; Magalhães, Aguiar-Conraria, and Lewis-Beck, 2012; Montalvo, Papaspiliopoulos, and Stumpf-Fétizon, 2019; Theis, Hense, and Damrath, 2005; Turgeon and Rennó, 2012; Umeda, 2023). However, few models show consistent predictive accuracy. In the USA, forecasters predicted 100% of state races in the 2012 federal election, though with poorer performance in 2016 (Jackman, 2014; Kennedy, Wojcik, and Lazer, 2017; Linzer, 2013; Wezerek, 2019). In Germany, Munzert (2017) used a novel procedure in which one regression model made predictions and another corrected anticipated errors. This system predicted 92% of the 299 districts in the 2013 German federal election. Gschwend et al. (2022) used a synthetic model that made 96% accurate constituency predictions in 2017 and 78.6% in 2021, one week from the election (Gschwend, 2017; Gschwend et al., 2022). The most outstandingly consistent result is from Neunhoeffer et al. (2020), who used a neural network to estimate the outcomes of the constituency votes for the 2009, 2013, and 2017 federal elections. This model correctly classified 89% to 92% of constituencies with a three month lead. In the United Kingdom (UK), the constituency-level accuracy in predicting the House of Commons varies considerably. The predictions of uniform swing models were 100% accurate for the elections in 2001 and 2017 but only between 2–52% accurate for the elections from 1987 to 2015 (Murr,

Stegmaier, and Lewis-Beck, 2021). Synthetic models in the UK also had a highly variable constituency-level accuracy between 57 and 100% for the elections from 1987 to 2017 (Curtice and Firth, 2008; Fisher, 2016; Murr, Stegmaier, and Lewis-Beck, 2021). These studies show that seat winner prediction remains challenging worldwide, with few models attaining consistently satisfactory results over several elections. The neural network model by Neunhoeffer et al. (2020) demonstrates the most consistently excellent results, underlining the potential benefits of utilizing ML techniques over swing models.

5.1.4 Evaluation Criteria

This paper explores whether ML-based synthetic electoral modeling can outperform the Mackerras pendulum in Australia. To this end, I will develop various ML models and compare them against several versions of the pendulum. Here, I describe my criteria for evaluating these models against the pendulum.

1. **Accuracy for seat counts:** Predicting each party's share of the House of Representatives is critical because it informs us which party will rule the country and whether it will have a majority or minority government. I expect to find an ML model that will predict seat outcomes more accurately than the pendulum, thereby predicting seat counts more closely.

2. **Accuracy for seat winners:** An ML model is better than the pendulum if it yields fewer misclassifications when predicting which party will win a seat. Greater accuracy in predicting minor party seats or in anticipating which of the most contested seats shall change party would be of particular value as these are noted weaknesses of the pendulum.

3. **Consistent accuracy:** If a model has high average accuracy over many elections but is prone to occasionally miscalling some elections very badly, it cannot be trusted if users cannot anticipate when it will suddenly fail (Gelman et al., 2020; Rothschild, 2015). The best model may not be the one with the highest average accuracy but the one that is consistently reasonably accurate.

4. **Lead time:** The longer in advance of election day a model can predict an outcome, the more time it affords the contestants to change the outcome (Jennings, Lewis-Beck, and Wlezien, 2020; Rothschild, 2015). To evaluate this, I test several lead times to show how each technique would have performed had it been used at that retrospective inference time (i.e., two months before the election of 2013).

5. **Uncertainty calibration:** Forecasting models typically output a range of possible outcomes for an election. Often, no model predicts the exact result. Rather, an ML model is better than the pendulum if its range of possibilities falls closer to the actual outcome (Gelman et al., 2020; Xie et al., 2023). Also, models with wide ranges of possible outcomes are not helpful (i.e., predicting that a party could win between 50 and 100 seats).

6. **Parsimony:** If two models have roughly the same performance, the simpler model is preferable. ML models are necessarily more complex than pendulum models. The reliance on historical data adds new sources of potential error (Lewis-Beck, 2005). Therefore, to prefer an ML model over the pendulum, its enhanced performance should justify its greater complexity.

5.2 Methods

This paper evaluates how ML models could improve Australian election forecasting over the pendulum. To achieve this, I will compare a range of ML models and versions of the pendulum. Additionally, I will explore different sets of predictors to determine which information, such as economic indicators or polls, is most useful for making accurate predictions. In this section, I describe the data sources (Section 5.2.1), the ML and pendulum models, and the process for fitting, evaluating, and comparing these models (5.2.2).

5.2.1 Data

ML models are trained on data from past elections to predict future ones. The following section details the predictors and the dependent variable in this study. For a list of all predictors see Appendix Table 5.6.

Dependent Variable: Seat Winner Party

Australia has a multiparty system, but only one of the two major parties, the Australian Labor Party (ALP) or the Liberal–National Coalition Party (LNP), has won the vast majority of federal seats since 1949 (see Figure 5.1). Therefore, I categorize all non-major parties as "other party" (OTH) and aim to predict which of the three categories each seat will fall into at each election.

The Australian Electoral Commission does not conveniently provide detailed data about election preferences before 2007 (Australian Electoral Commission, 2001; Australian Electoral Commission, 2007; Australian Electoral Commission, 2010; Australian Electoral Commission, 2013; Australian Electoral Commission, 2016; Australian Electoral Commission, 2019a; Australian Electoral Commission, 2022). So, because at least one election is required for training an ML model, I examine the five elections between 2010 and 2022.

Polling

Polling averages are inputs for ML models and pendulums (see Section Pendulums). In Australia, many polls on voting intention are conducted every few months, with frequency increasing as election day approaches. These polls typically fall into two categories: two-party preferred, where respondents choose between the two major parties, or multiparty, where respondents indicate their first preference among all parties.

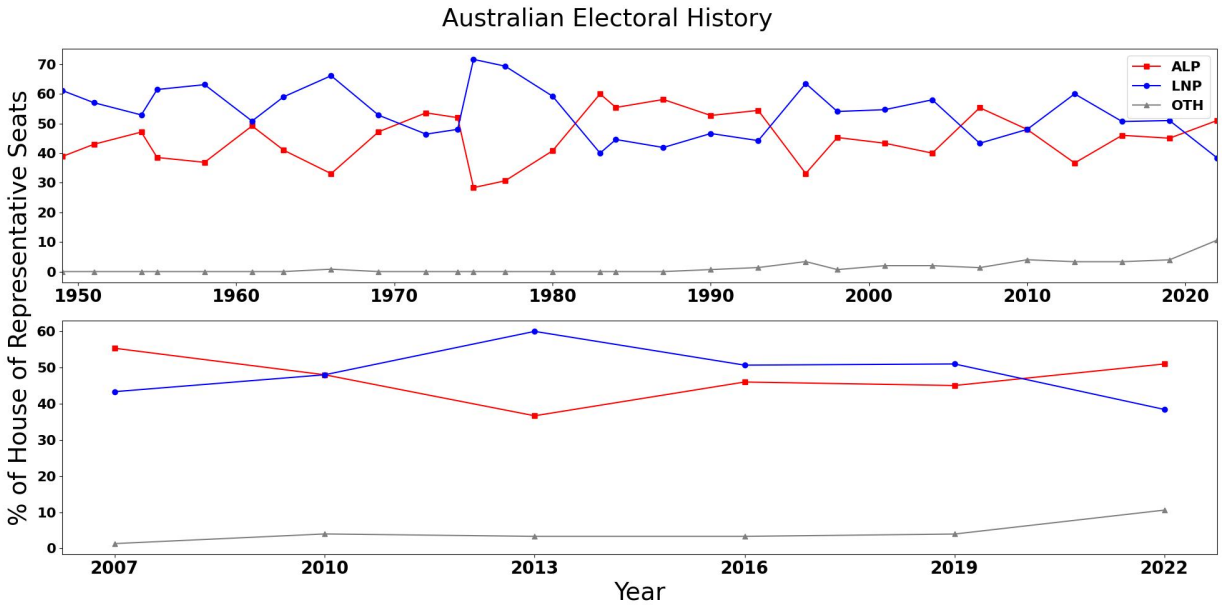


Figure 5.1: *Timeline of election outcomes. The top figure shows Australian federal elections since 1949. The bottom figure shows only the elections tested in this study.*

Although many pollsters operate in Australia, only a few publicly available sources compile polling data. The website The Poll Bludger has aggregated polling data since 2019 (Bowe, 2023), but for earlier years, I rely on other public sources (PhantomTrend, 2016). Unfortunately, the detailed results of many original polls are not public (Bowe, 2023; PhantomTrend, 2016). I have only basic details for each poll consistently, such as party-wise percentages, pollster names, and field period dates. Appendix Table 5.7 lists all data sources. Details such as nonresponse or undecided rates are often unavailable in these aggregate sources.

When predicting seat-level outcomes, the challenge is that opinion polling is typically conducted at the state or national level, not the seat level. Although respondents may provide their postcode (and thus their electoral seat), sample sizes are usually around 1,000, randomly drawn from the entire population, resulting in small sample sizes for individual seats (Bowe, 2023; PhantomTrend, 2016). Furthermore, aggregated poll results, not individual respondent data, are consistently available, making it impossible to disaggregate polls by seat (Bowe, 2023). Pollsters conduct swing seat-specific surveys, but their results are not consistently publicly available (Goot, 2023). Therefore, for each seat at each election, I derive features for national and state polling averages over a rolling four-week window.

To address the lead time criterion, I will refit each model and pendulum using polling averages available six, three, two, one month(s), and one week before election day. ML models can utilize both the most recent and the preceding polling averages, allowing the model to account for temporal trends. For example, an ML model forecasting an

election two months before election day will use the polling averages two months from election day as well as three and six months out as predictors.

If state-level polling is available at a given lead time before a given election, then I also use that four-week average as a predictor for seats in that state (See Appendix Section 5.5.2). Furthermore, the ML models incorporate polling figures from the previous election to account for past polling errors. Finally, I include the standard deviation of each polling average (see Appendix Table 5.6). Figure 5.2 shows the results of each national poll and the time windows in which I collected the polls to derive an average value. Appendix Section 5.5.2 provides the equivalent figure at the state level, at which far fewer polls were taken.

A final consideration regarding this paper's polling averages is that not all published pendulum models use a four-week polling aggregate as I have. Other election analysts select specific polls taken as close as possible to a given lead time or take averages among a small set of polls (Browne, 2022; Goot, 2023; Green, 2016; Green, 2019; Green, 2022). This requires intuition on the part of the analyst to select some polls over others. Rather than selectively filtering polls, which adds an element of retrospective decision-making, this paper takes the averages as they were available from public, aggregated sources at the inference time. This may result in this paper's pendulums deviating from forecasts conducted by other analysts for the same election.

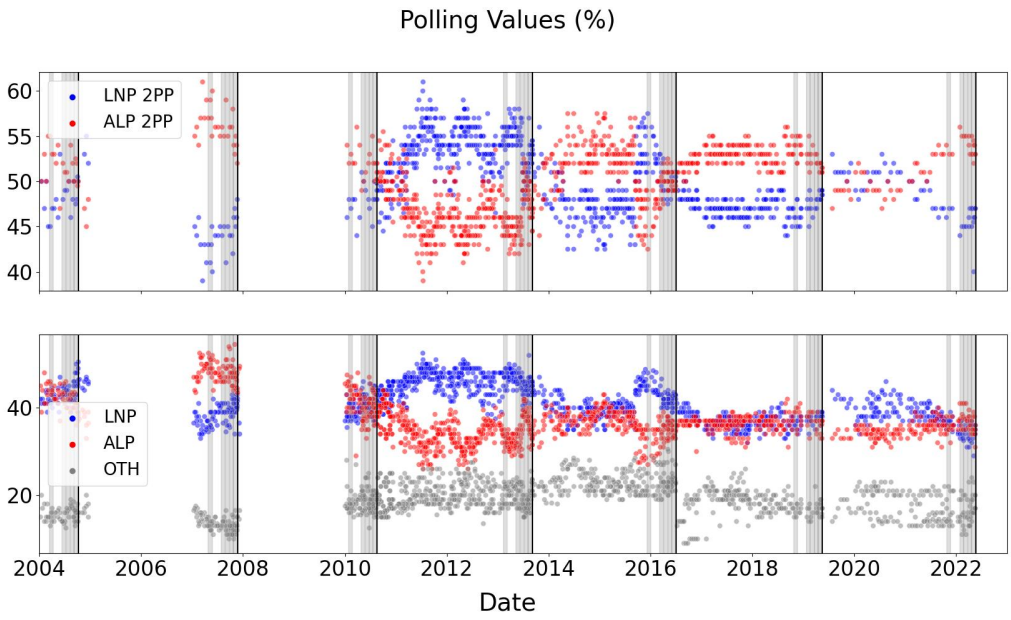


Figure 5.2: *Polling values and the four-week time windows (gray boxes) in which polling averages are calculated. The black lines indicate an election day.*

Election Tallies

For each election, I use seat-level results from the previous election to make predictions. Specifically, I consider the highest percentage share of votes each party received before a winner was declared. I also use a binary indicator to show whether the seat flipped in the previous election. Previous election tallies are also used to calculate the margins used in pendulum models (see Section 5.2.2).

State Effects

These predictors indicate the federal state of each seat. These variables allow the ML models to account for state-level effects, such as how national polling averages correspond to different vote shares in specific states.

Pendulum Predictions

In some feature sets (see Section Feature Sets), I use seat-level predictions from a TPP proportional swing model (see Section Pendulums) as inputs for the ML models. The ML model alters the pendulum's predictions in these trials by accounting for other variables, such as socioeconomics or state effects.

Fundamentals

Fundamentals are predictors based on the proposition that voters punish incumbents for poor economic performance and are reluctant to change the government otherwise (Hummel and Rothschild, 2014). For this reason, I use macroeconomic variables and seat-level socioeconomics to estimate each seat's condition and, consequently, the voting behavior. Supporters of the fundamental approach to election prediction emphasize that these variables are theoretically more predictive than polls at long lead times (Hummel and Rothschild, 2014; Jennings, Lewis-Beck, and Wlezien, 2020; Lewis-Beck, 2005). In addition to indicators of economic performance, I also include demographics, specifically each seat's median age, homeownership rate, and portion of indigenous and overseas-born population. I include these variables to capture the effect of generational, renter-landowner, and racial social divides in Australia.

The Australian Bureau of Statistics provides various socioeconomic statistics disaggregated at the electoral seat level (Statistics, 2021). These data come from the 2006, 2011, 2016, and 2021 censuses. I use the latest available census figures for each seat at each election in the dataset. The details of the variables are available in Appendix Section 5.5.2. I include median income, rent, mortgage payments, share of employment status, type of job, migrant history, and Indigenous heritage to represent each seat's economic and social composition.

In addition to seat-level socioeconomic indicators, I consider national macroeconomic factors in each election. I use real GDP growth and inflation rates from the year before the election (which avoids using data that would have been unavailable at the retrospec-

tive inference time) to estimate how voters feel about the macroeconomy in the run-up to the election.

Missing Values

Missing values are imputed with zeros. The sources of missing values in the data are as follows. New seats are created at certain elections due to Australia's periodic redistricting. Therefore, no valid values exist regarding the previous election. State-level polling variables are missing for some time windows (see Appendix Section 5.5.2).

5.2.2 Modeling Setup

I will make retrospective predictions using both pendulums and ML models for each election from 2010 to 2022, each for six months, three months, two months, one month, and one week before election day. I selected these time frames based on (Jennings, Lewis-Beck, and Wlezien, 2020), who concluded in a cross-national study that most accurate electoral forecasts become valid around two to three months before an election.

For each ML algorithm (Section Machine Learners), I will repeat each trial with different sets of predictors to evaluate the effectiveness of different features. For example, will a given ML model perform better with only fundamentals or will including polls improve election forecasts? I explore these combinations to identify the best possible ML algorithm and predictors.

I will also provide a "no change" baseline, which predicts that every seat in a given election will remain with its incumbent. I include this baseline because most seats stay with their incumbent in any election, and so even this naive model will achieve a certain level of accuracy. I expect any good model to outperform this baseline. In any baseline model, if a seat is newly formed after redistricting, the predicted winner is whatever party leads in national TPP polls.

The following section will first provide details on the different ML algorithms and pendulums (Sections Machine Learners & Pendulums). Next, I will present the sets of features which I shall vary to test the value of different types of predictors (see Feature Sets). Then, I will explain how each ML model is fitted to replicate how predictions could have been made at the inference time (i.e., two months before the 2019 election) and how this approach prevents overfitting (see Temporal Cross-Validation and Hyperparameters). Additionally, I will describe how I measure uncertainty in each model (see Uncertainty Calibration). Finally, I will explain how I will use permutation feature importance (PFI; Altmann et al. (2010)) to explain how the most successful ML models make their predictions (see Model Interpretation).

Machine Learners

Machine learning covers a vast range of diverse algorithms. It would be unfeasible to trial every possible ML algorithm, so I focus on a set of the most common types according to a literature review by Singh, Thakur, and Sharma (2016). I trial main effect models, namely logistic/linear regression, and models that automatically account for interaction

effects, namely tree-based models. I also include a neural network based on Neunhoeffer et al. (2020) because that method yielded outstanding results (see Section 5.1.3). Specifically, I experiment with neural network architectures with the same dimensions as that study (two dense layers of 128 and 64 neurons) and I vary those parameters (see Appendix Section 5.5.3). The algorithms I have selected are unpenalized and penalized logistic regression (ULR, PLR), extra trees (ET), gradient boost (GB), and multilayer perceptron (MLP; a type of neural network). Details of these models are presented in Appendix Section 5.5.3 I discuss hyperparameters in Section Temporal Cross-Validation and Hyperparameters.

I trial both classifiers, which predict the probability of victory for each party category (ALP, LNP, OTH), and regression models, which predict each party's seat-level vote share with an uncertainty range. In regression trials, I substitute logistic regression with linear regression (LR), as the former is only suitable for classification.

Each ML model is trained on tabular data derived from past elections. Each row is a seat at a given election. The data about each seat are the predictors, which will vary according to the feature sets described in Section Feature Sets. The dependent variable for the classifiers is the category of the candidate who wins that seat at that election. For the regressors, the dependent variable is the vote share won by each party, for which the largest will be taken as that regressor's predicted seat winner. I iterate over elections from 2010 to 2022, each time training the models on data available before inference time (i.e., three months before the 2022 election).

Pendulums

There are several versions of the pendulum, which I will present for comprehensiveness. The details of how I calculate the predictions for each pendulum variant are in Appendix Section 5.5.3. Table 5.1 provides an overview of each variation.

Feature Sets

I aim to compare different ML algorithms and different sets of features to determine the best approach for forecasting elections. Using all available features can often confound models (Hastie, Tibshirani, and Friedman, 2009). Instead, I test different sets of features for each ML model that reflect various forecasting strategies, as detailed in Table 5.2. To avoid lengthy model fitting times, I only test some possible combinations of sets. I never test polls-only models, as they offer little improvement over crudely using the TPP leader. I always include state effects and a variable indicating whether a seat flipped in the previous election in every feature set to limit the number of permutations to test. The regressors are trained only on the "polls and fundamentals" set so as to reduce the number of models fitted and because the results will show that this technique is not promising enough to warrant deeper investigation (see Section 5.3.1). Appendix Table 5.1 provides the complete list of predictors and their respective feature sets.

Table 5.1: *Summary of pendulum models.*

Pendulum variant	Description
Two-party preferred uniform (TPP uniform)	This model is the typical Mackerras pendulum which the ABC uses (Green, 2022). I use the preceding election's national two-party vote shares and the target election's TPP polls to calculate the swing for each major party. I then assume that each seat's vote share from the previous election will change by the same swing. In some seats, this shift in vote share would be enough to unseat the incumbent party. In this case, the opposing major party is predicted to take the seat. Seats held by OTH candidates are assumed to remain with their incumbent. A known flaw in this model is that it cannot predict a seat held by a major party to be won by an OTH candidate or vice-versa.
Two-party preferred proportional (TPP proportional)	This model is the same as the TPP uniform, except that when a state-level TPP polling value is available for that particular four week window (there must be at least one poll published in that period), that value is used to calculate the swings for seats in that state.
Two-candidate preferred uniform (TCP uniform)	This variation of the TPP uniform is based on the Reed pendulum proposed by Resolve Strategic (Reed, 2022). This model aims to better predict seats won by OTH candidates. It calculates swings for major and minor parties using multiparty polls instead of TPP. If the swing against an incumbent party is enough to lose the seat, that seat is predicted to go to the party that received the second-highest votes in the previous election.

Table 5.2: *Summary of feature sets.*

Feature set name	Description
Fundamentals	These are the socioeconomic variables described in Section Fundamentals, previous election vote shares, and state effects.
Polls and fundamentals	The same as the fundamentals set but includes polling averages described in Section Polling.
Pendulum, polls, and fundamentals	In this feature set, I also include seat-level predictions from a TPP proportional pendulum (see Section Pendulums). This approach aims to improve the pendulum's predictions by adjusting them based on these additional variables.
Pendulum and fundamentals	This feature set excludes polls to test whether those predictors confound pendulum-based ML models.
Pendulum and polls	This feature set excludes fundamentals to test whether those predictors confound pendulum-based ML models.

Temporal Cross-Validation and Hyperparameters

To ensure my forecasts reflect out-of-sample estimations as they could have been made at the time, I use temporal cross-validation (TCV; Bergmeir and Benítez (2012)). With TCV, I predict each seat in each election using models trained on data from all preceding elections. Each ML model fitting involves selecting hyperparameters (Arnold et al., 2024). Typically, practitioners train models on data and then validate them on randomly selected hold-out data to identify the hyperparameter settings expected to perform best on new data (Hastie, Tibshirani, and Friedman, 2009; James et al., 2013). However, in this context, validating models against randomly withheld seats from the same elections as the training set would lead to overfitting those models to those elections. Instead, I want to select models and their hyperparameters based on how well they can predict elections that occurred after the training elections (Bergmeir and Benítez, 2012). To do so, I will find the hyperparameter settings for each model with the lowest overall predictive error (least misclassifications for classifiers and lowest mean absolute error across vote shares for regressors) when applied to the target (out-of-sample) election. I then present the outputs of these settings for each model in the Results section and the outputs for all other settings in the reproduction materials (Appendix Section 5.5.4).

An important aspect of this procedure is that, in the results section, I present the hyperparameters that performed best, averaged across the elections from 2010 to 2022. If practitioners had been fitting models at the time, they might have selected the best hyperparameters known to them at that point. While this procedure identifies the best models that can be discovered using data from these five elections, it may not perfectly reflect how modeling would have been conducted at the time in this regard.

Another parameter I explore is the number of elections to include in the training set. Like any democracy, Australia's political climate changes over time, so data from the 2007 election might confound models predicting the 2022 election. To evaluate this effect, I will refit all models using either all previous elections or just the most recent two as training data. I will then determine which approach yields the highest accuracy scores. I limit the analysis to the past five or two elections to reduce the number of model fittings required.

Uncertainty Calibration

This paper will present the uncertainty around each estimated election outcome. Pendulums, classifiers, and regressors handle forecast probabilities differently. For each model, I will provide the most likely forecast and the range of outcomes within the 95% most likely scenarios (i.e., the 95% confidence interval (CI)). This section explains how I calculate these ranges for each model type.

For each seat in the predicted election, each ML classifier model outputs the probability that a party will win (e.g., an 80% likelihood of the ALP winning the seat of Adelaide in 2010, 15% for the LNP, and 5% for OTH). These probabilities represent the model's confidence intervals. For instance, among seats assigned an 80% likelihood of ALP victory, approximately 80% were won by the ALP in the training data. I simulate a

hypothetical election for each model, using each seat's class probabilities as weights. For example, if a seat has a 20% chance of an ALP victory, it should be predicted to go to the ALP in 20% of simulations. Due to the computational expense across multiple models, I simulate 20 elections for each ML classifier. For each simulation, I calculate the probability of that particular outcome using the model probabilities and discard the 5% least likely outcomes. I then present the range of seat counts and accuracy scores from the remaining simulations which represent the 95% most likely election results. Finally, I include a scenario in which the most likely winner of each seat is assumed to be the winner and present this as the single most likely election result according to that classifier.

Regressors do not produce probability estimates around their vote share predictions. Instead, I use bootstrapping, stratified by the three classes, to randomly withhold 10% of the training data. I create ten randomly bootstrapped variants of the training data, retrain the model on each one, and take the different forecasts as the uncertainty range (Thai et al., 2013). I also calculate the result without bootstrapping and include that in the distribution. I present the 95% confidence interval of classification accuracy scores and seat shares across these variations. I present the mean seat count and accuracy score as the most likely outcome.

For the pendulum models, I simulate swing values, assuming polling averages are up to two standard deviations above and below the mean. This approach accounts for cases where pollsters published a wide range of values, resulting in greater uncertainty for the pendulum. I present the pendulum outputs when using the mean polling value as each pendulum's most likely predicted outcome.

Model Interpretation

I aim to select the best ML model based on the evaluation criteria (Section 5.1.4) and understand how it makes predictions and which features are most influential. To do this, I use permutation feature importance (PFI), which measures how much predictive accuracy is lost when a given feature is withheld from the model (Altmann et al., 2010). A well-known issue with PFI is that when two features are correlated (i.e., TPP polls for ALP and LNP), withholding one does not lower the model's accuracy because the same information is available through another predictor. Therefore, I withhold predictors in blocks of related features to evaluate the importance of these feature blocks rather than of the individual predictors. Appendix Section 5.5.2. details which features are categorized into which blocks. I fit the model for this PFI analysis using data available at three months before election day because I want to know how the model functions at a substantial lead time. I will not calculate PFI for models other than the best selected one because PFI requires long calculation times.

5.3 Results

5.3.1 Comparing Accuracy Scores

This section presents only the models fitted on a maximum of two preceding elections, because that approach yielded the most accurate results for both seat counts and seat classifications (see Appendix Section 5.5.4 for full details). Training on a limited number of past elections improved accuracy for the most accurate model (ET as explained below), suggesting that distant past elections are unreliable guides for later ones.

Figure 5.3 shows each model's classification accuracy scores (the portion of correct classification predictions out of all seats). The error bars represent the range of scores across the 95% most likely election results according to each model, and the dots represent each model's single most likely predicted outcome (see Section Uncertainty Calibration). Table 5.3 presents a tabular summary of Figure 5.3 for the best-performing ML model and pendulum (as discussed below). The ML classifiers' error bars are often skewed to the left of the most likely outcome. This skewness is because these ML techniques often decrease in accuracy if anything other than the most likely prediction is fielded, indicating that each probability estimate was well-calibrated, and so deviating from the most likely class usually leads to misclassification. In other words, for these ML classifier models, the most likely predicted scenario is usually its most accurate forecast out of any of the simulated elections (see Section Uncertainty Calibration).

In contrast, the most likely outcome, according to each of the pendulum models, is often roughly in the middle of their error bars. This distribution is because I estimate the pendulum's uncertainty by tilting the polling margins by up to two standard deviations for or against the incumbent (see Section Uncertainty Calibration). For someone using a pendulum to predict an election, this would be like sensitivity analyzing how the pendulum's forecasts change as they try varying the uniform swing value. Unlike the ML classifiers, this method for deriving the 95% most probable election outcomes can often yield more accurate forecasts than the mean polling average available at the inference time.

In Australia, safe seats rarely change party (which is why the no-change baseline achieves 87% accuracy on average). Therefore, it is especially critical to predict marginal seats in Australia. Also, as discussed, predicting crossbench victories is a notable weakness of the pendulum. By comparing ML models with the no-change baseline I can evaluate performance regarding flip seats. I will examine crossbench seat prediction in Section 5.3.2. Additionally, Appendix Section 5.5.4 presents the seat classification accuracies for marginal and crossbench seats. Table 5.3 presents the marginal seat classification accuracies for the best performing ML and pendulum models (discussed below).

Before discussing the ML models, I shall first determine which of the baseline models was the best so as to compare ML models against that technique. In the 2019 and 2022 elections forecasts, the TPP uniform pendulum performed much worse than in previous elections (roughly 83% accuracy, down from 88% averaged across all lead times). With more state polls available in those years, the TPP proportional pendulum deviated from

the TPP uniform pendulum. The proportional model outperformed the uniform model at six months and one week before the 2019 election (each by roughly 6%) but underperformed in the forecast one month before the 2022 election (by 5%). The predictions of the TCP pendulum did not exceed those of the TPP uniform pendulum for the elections from 2010 to 2016. Overall, TPP proportional is preferable to uniform because it performed as well or better at a substantial (three month) lead time. It is important to note that the proportional pendulum will be identical to the uniform pendulum unless state polls are available.

Regarding the ML models, in any given election, at least one ML model appears to outperform the TPP proportional pendulum in at least one lead time but rarely consistently across elections. ET is almost always the better performing algorithm among any feature set (see Figure 5.3). Tree-based models incorporating fundamental features outperform the pendulums in 2019, achieving over 90% accuracy on average, however they perform much worse in elections 2010-2016. Regression models perform better than the pendulums in 2019 but worse in all other elections, which is why I do not investigate these techniques further. However, I present the regression outputs versus the actual vote share results in Appendix Section 5.5.4.

For seat classification accuracy, I evaluate that ET with pendulum, polls, and fundamental features performed best overall. This model is consistently roughly as good, or slightly better than the best pendulum at any election at one month before election day. This model is also as good or slightly better than the best pendulum at any election at both six and three months out from each election except for 2013 (see Table 5.3). The ET model is roughly 10% better than the TPP proportional pendulum from three months lead time in 2019 and 3% in 2022. With the exception of 2013, ET either equals or outperforms the pendulum (sometimes by over 10%) when classifying the marginal seats. This model appears to perform roughly as well as the pendulum, but avoids the pendulum's collapses in accuracy in 2019 and 2022.

Figure 5.4 presents the forecasts of seat shares for the ALP, with equivalent figures for LNP and OTH in Appendix Section 5.5.4. The most consistently accurate model is the GB algorithm, which incorporates pendulum and polling features. Aside from 2019, this model's 95% confidence interval falls within three seats of the actual seat count every election from three months lead time. Table 5.4 provides a precise comparison of error scores, showing the values for this GB model, the TPP proportional pendulum, and the ET model, which performed as the best overall classifier.

An ideal model would outperform the pendulums in both seat count and classification accuracy, but unfortunately, the ET with all predictors is the best classifier, while the GB with pendulum and polling predictors is the best seat count forecaster. However, at one month lead time, the pendulum is only closer to the actual result than the ET model in 2016 (by 12 seats). Furthermore, the ET model never miscalls which major party forms government as of three months lead: It overestimates ALP's share in 2010, incorrectly predicting a majority for the ALP. In 2013 and 2016, ET is far off the seat count but correctly calls the election for LNP. In 2019, ET forecasts either major party having a chance of forming minority government, which is much closer to the actual outcome (a slim two seat majority for LNP) than the proportional pendulum (which predicted

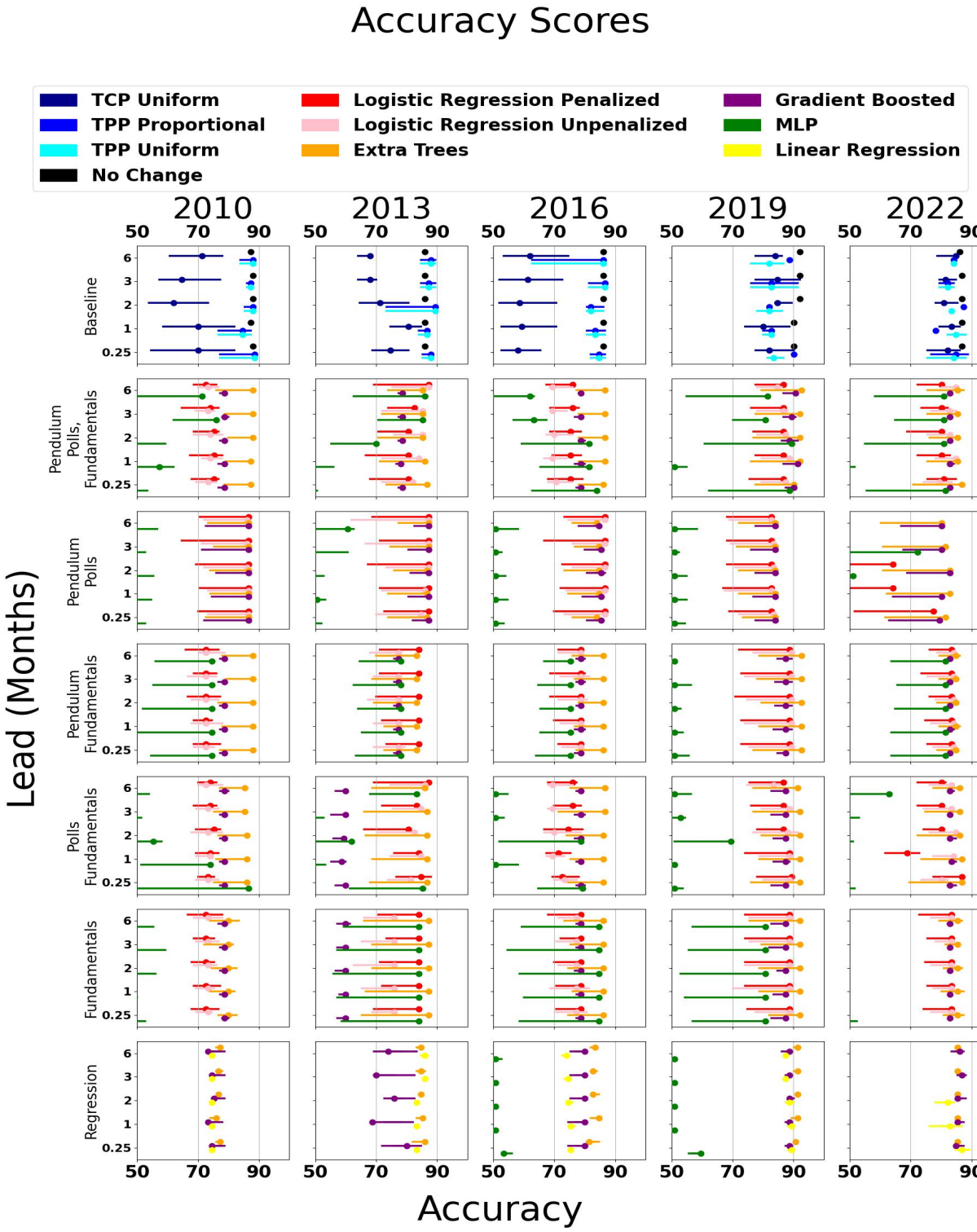


Figure 5.3: *Classification accuracy scores of each ML model and pendulum across every election and lead time. Note that any variation in the "no change" models is caused by newly redistricted seats.*

Table 5.3: *An abridged set of accuracy scores for the most likely predicted outcome, along with the lower and upper bounds of the 95% confidence interval (see Section Uncertainty Calibration for details; refer to Appendix Section 5.5.4 for the complete table). For instance, if the lower bound accuracy for an ML model is 80%, it indicates that among all accuracy scores generated from the simulated elections, the value at the 95% confidence interval's lower bound was 80%. Also, if a pendulum model shows 90% accuracy at the upper bound, it suggests that when polling values are adjusted to be two standard deviations higher for the election winning party than the mean, the pendulum's classification accuracy reaches 90%.*

			All Seats			Marginal Seats		
Lead (Months)	Year	Model	Most probable outcome	Lower bound (95% CI)	Upper bound (95% CI)	Most probable outcome	Lower bound (95% CI)	Upper bound (95% CI)
1	2010	Extra Trees Classifier	87.33	79.33	87.33	74.58	59.32	74.58
		TPP Proportional	84.67	76.67	87.33	67.80	47.46	74.58
	2013	Extra Trees Classifier	86.00	71.33	86.00	73.21	53.57	73.21
		TPP Proportional	86.00	84.00	87.33	76.79	71.43	80.36
	2016	Extra Trees Classifier	86.67	78.00	86.67	78.13	59.38	78.12
		TPP Proportional	83.33	80.67	86.67	70.31	64.06	78.12
	2019	Extra Trees Classifier	92.05	76.16	92.05	82.76	55.17	82.76
		TPP Proportional	82.78	80.13	83.44	58.62	51.72	60.34
	2022	Extra Trees Classifier	85.43	75.50	85.43	72.92	54.17	79.17
		TPP Proportional	78.15	78.15	78.15	70.83	70.83	70.83
3	2010	Extra Trees Classifier	88.00	78.67	88.00	76.27	62.71	76.27
		TPP Proportional	86.67	86.00	87.33	72.88	71.19	74.58
	2013	Extra Trees Classifier	85.33	72.00	85.33	71.43	48.21	75.00
		TPP Proportional	87.33	84.67	89.33	80.36	75.00	85.71
	2016	Extra Trees Classifier	86.67	78.00	86.67	78.13	56.25	78.12
		TPP Proportional	85.33	81.33	86.67	75.00	65.63	78.12
	2019	Extra Trees Classifier	92.05	77.48	92.05	82.76	55.17	82.76
		TPP Proportional	82.78	76.16	91.39	58.62	46.55	81.03
	2022	Extra Trees Classifier	85.43	75.50	85.43	72.92	62.50	79.17
		TPP Proportional	82.12	79.47	84.11	70.83	70.83	70.83
6	2010	Extra Trees Classifier	88.00	76.00	88.00	76.27	55.93	77.97
		TPP Proportional	87.33	84.00	88.00	74.58	66.10	76.27
	2013	Extra Trees Classifier	85.33	74.00	85.33	71.43	50.00	78.57
		TPP Proportional	88.00	84.67	89.33	82.14	75.00	85.71
	2016	Extra Trees Classifier	86.67	77.33	86.67	78.13	62.50	78.12
		TPP Proportional	86.00	62.67	86.00	76.56	21.88	76.56
	2019	Extra Trees Classifier	92.72	79.47	92.72	84.48	60.34	84.48
		TPP Proportional	88.74	88.74	88.74	74.14	74.14	74.14
	2022	Extra Trees Classifier	85.43	75.50	87.42	72.92	60.42	77.08
		TPP Proportional	84.11	84.11	84.11	68.75	68.75	68.75

over 90 seats for ALP). In 2022, the ET model predicted a slim minority for ALP which is very close to the actual two seat majority that eventuated, and a better prediction than the anticipated 100 ALP seats forecasted by the proportional pendulum. Although not the most accurate for seat count, the ET model is still correct about which party forms government, and unlike the more accurate GB model, ET is the better classifier. Furthermore, GB's accuracy declines sharply in 2019, whereas the ET model did not, making it more consistent in that sense. For these reasons, I select this ET model as the best model to explore more thoroughly in the next section.

5.3.2 Presenting the Best Models

This section details the best models identified in this study. The ET model with the pendulum, polls, and fundamentals performed roughly as well or much better than the pendulum in seat classification accuracy three months before each election (Figure 5.3). However, focusing solely on misclassification rates or seat count errors overlooks important details for evaluating these models. This section compares this ET model with the TPP proportional pendulum to explore how each method would have predicted each election differently. I make the comparison at three months before each election which is the longest lead time at which the ET model reaches or nears maximum classification accuracy in most election forecasts (see Figure 5.3). I chose the TPP proportional model for its higher accuracy at this lead time than the other pendulum models (Appendix Section 5.5.4 presents this analysis using the uniform pendulum instead).

Figure 5.5 shows the confusion matrix for the ET model's and the TPP proportional pendulum's most likely predictions. The ET model's classification probabilities against actual vote shares are presented in Appendix Section 5.5.4. The confusion matrix reveals that the ET model misclassifies one fewer seat than the TPP proportional pendulum in 2010, one additional seat in 2013, and the same overall number in 2016. Therefore, up to this point, the ET model has offered no major improvement over the simpler pendulum model. This changed in 2019, when the pendulum misclassifies a substantial 26 seats, whereas the ET model misclassifies only 11. The ET continues to outperform the pendulum in 2022, where it miscalls 22 seats, compared to the pendulum's 27. Looking at the composition of misclassifications, it seems the ET model favors the incumbent more than the pendulum. Hence, ET gave the LNP better chances of victory than the pendulum in 2019 and misclassifies in favor of the ruling major party (counting the term 2010-13 as 'ruled' by ALP even though they were in minority) in every other election. Critically, the ET model does not improve over the pendulum in predicting seats won by OTH parties.

YouGov's MRP model made only nine misclassifications in 2022, and six were too close to call (Goot, 2023). Even if I consider all of those seats as failures, the MRP still beats ET 15 to 22, and so MRP is the better overall seat classifier. MRP also performs better for seat count predictions, forecasting 80 for ALP, versus the ET's 71 and an actual result of 77. However, the MRP was released two weeks before election day, whereas these predictions were made three months before. Also, the actual result of 77 was within the 95% confidence interval for the ET model even at three months from

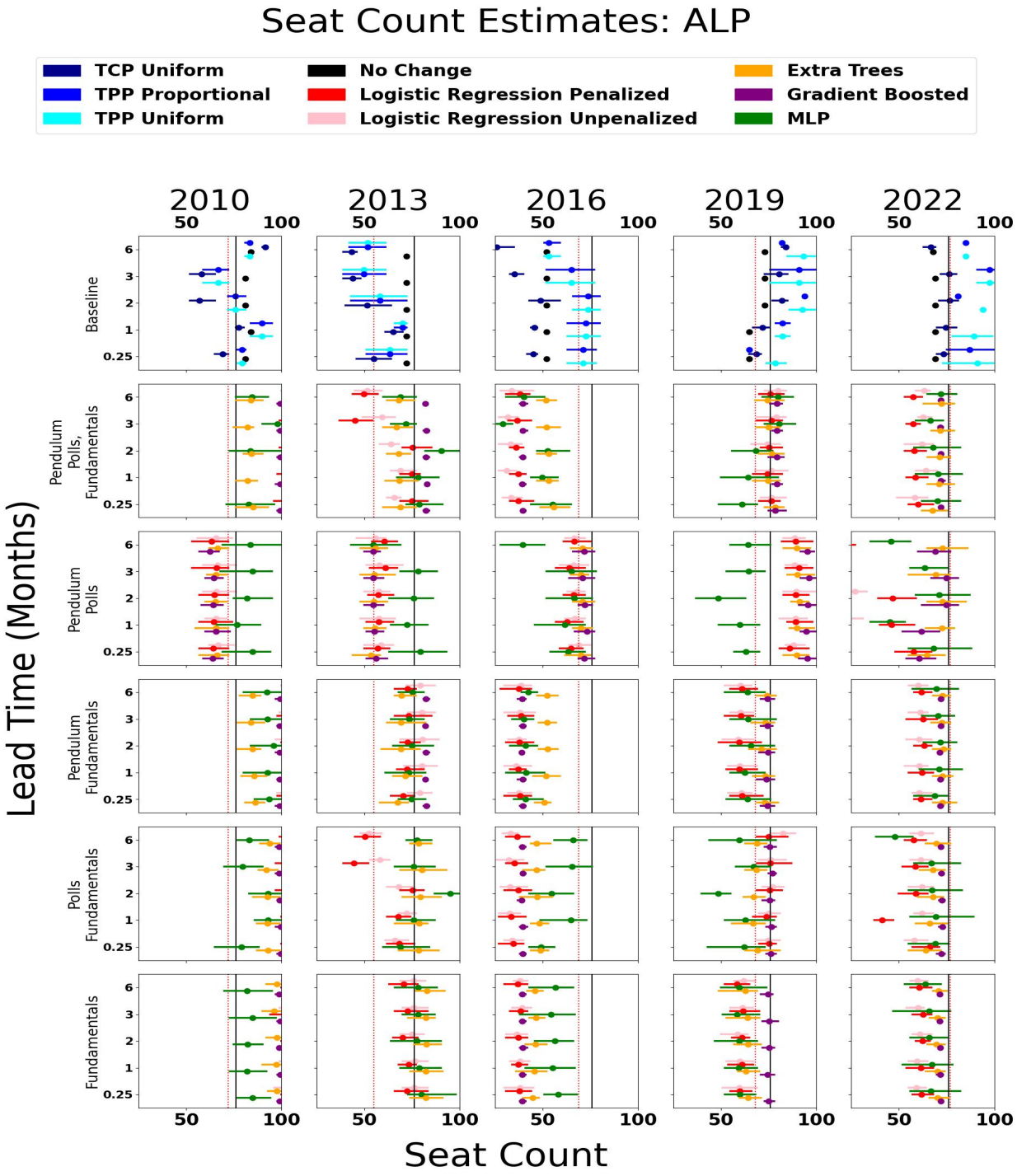


Figure 5.4: *The count of seats for the ALP as predicted by each model. The dotted line indicates the actual seat count for that election, and the solid line represents the 76 seats required to form a majority government.*

Table 5.4: *An abridged set of ALP seat count error scores for the most likely predicted outcome, along with the lower and upper bounds of the 95% confidence interval (see Section Uncertainty Calibration for details; refer to Appendix Section 5.5.4 for the complete table). For example, an error of -2 at the lower bound indicates that the actual ALP seat count was two seats higher than the lowest estimate of seat count within the 95% confidence interval.*

Lead (Months)	Year	Feature Set	Model			
1	2010	Baseline	TPP Proportional	18	12	23
		Pendulum and Polls	Gradient Boosted Classifier	−6	−12	1
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	10	4	15
	2013	Baseline	TPP Proportional	15	11	17
		Pendulum and Polls	Gradient Boosted Classifier	0	−4	5
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	13	6	23
	2016	Baseline	TPP Proportional	4	−6	11
		Pendulum and Polls	Gradient Boosted Classifier	4	−2	8
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	−16	−22	−11
	2019	Baseline	TPP Proportional	14	11	18
		Pendulum and Polls	Gradient Boosted Classifier	27	24	32
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	7	−1	13
	2022	Baseline	TPP Proportional	29	29	29
		Pendulum and Polls	Gradient Boosted Classifier	−15	−25	−6
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	−6	−12	2
3	2010	Baseline	TPP Proportional	−5	−13	0
		Pendulum and Polls	Gradient Boosted Classifier	−8	−12	−3
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	10	3	13
	2013	Baseline	TPP Proportional	−5	−16	6
		Pendulum and Polls	Gradient Boosted Classifier	−1	−5	5
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	12	5	20
	2016	Baseline	TPP Proportional	−4	−17	8
		Pendulum and Polls	Gradient Boosted Classifier	2	−5	8
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	−17	−22	−10
	2019	Baseline	TPP Proportional	23	8	35
		Pendulum and Polls	Gradient Boosted Classifier	28	24	32
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	7	1	13
	2022	Baseline	TPP Proportional	21	14	26
		Pendulum and Polls	Gradient Boosted Classifier	−2	−10	4
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	−5	−10	2
6	2010	Baseline	TPP Proportional	11	9	12
		Pendulum and Polls	Gradient Boosted Classifier	−10	−15	−5
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	12	4	18
	2013	Baseline	TPP Proportional	−3	−13	6
		Pendulum and Polls	Gradient Boosted Classifier	0	−5	3
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	13	7	21
	2016	Baseline	TPP Proportional	−16	−18	−10
		Pendulum and Polls	Gradient Boosted Classifier	3	−3	8
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	−17	−22	−12
	2019	Baseline	TPP Proportional	14	14	14
		Pendulum and Polls	Gradient Boosted Classifier	27	24	31
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	6	0	15
	2022	Baseline	TPP Proportional	8	8	8
		Pendulum and Polls	Gradient Boosted Classifier	−8	−17	0
		Pendulum, Polls, and Fundamentals	Extra Trees Classifier	−5	−14	2

election day 2022 (Figure 5.4). In tight elections like 2010, 2016, 2019, and 2022, better precision in predicting even two seats can be crucial with sufficient lead time, which the MRP demonstrated. However, it was only a short lead time and in only one election.

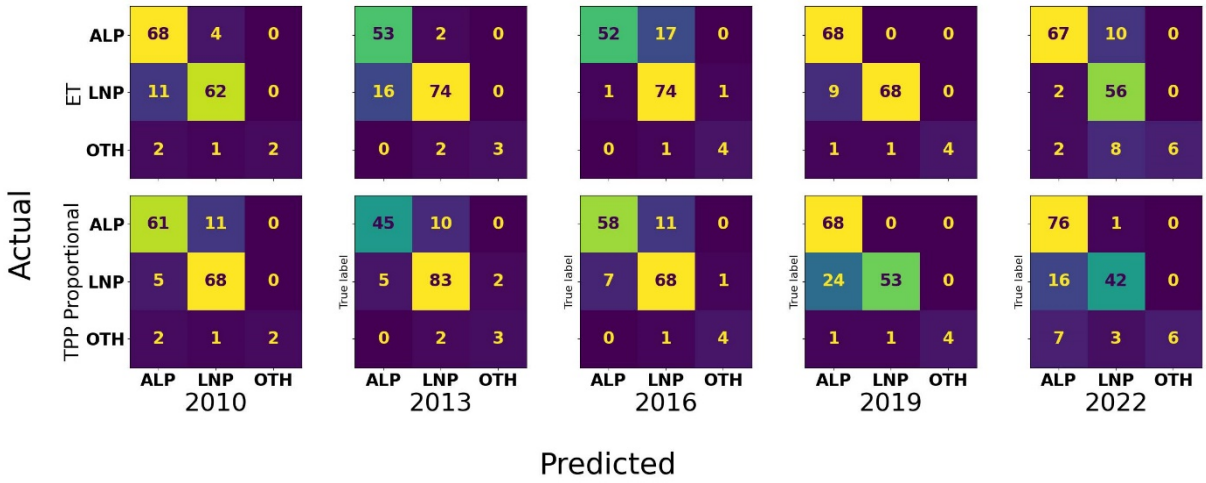


Figure 5.5: *Confusion matrices for the best model in this study (ET with all predictors) versus the best pendulum (TPP proportional). Each model's most likely predictions were made three months before election day. For example, the top leftmost number (68) indicates that in 2010, the ET model predicted 68 seats would go to the ALP and did so. However, the number 11 underneath shows that 11 seats were also predicted to go to the ALP but were instead won by the LNP.*

Permutation Feature Importance

Figure 5.6 shows the permutation feature importance (PFI) for each feature block in the best-performing ET model. The model primarily relies on the pendulum and each party's vote share from the previous election, suggesting that the ET model adjusts pendulum predictions based on electoral margins. Other factors contribute inconsistently to predictive accuracy, improving it in some elections while decreasing it in others. This observation suggests that pruning certain socioeconomic factors might improve performance. State effects and unemployment rate are the most consistently important predictors outside of vote shares and pendulum features.

Interestingly, polls have little impact, likely because they are correlated with the pendulum predictors. However, Figure 5.3 shows that including polls slightly improves predictions in some cases (e.g., six months before the 2013 election), indicating some value in incorporating polls.

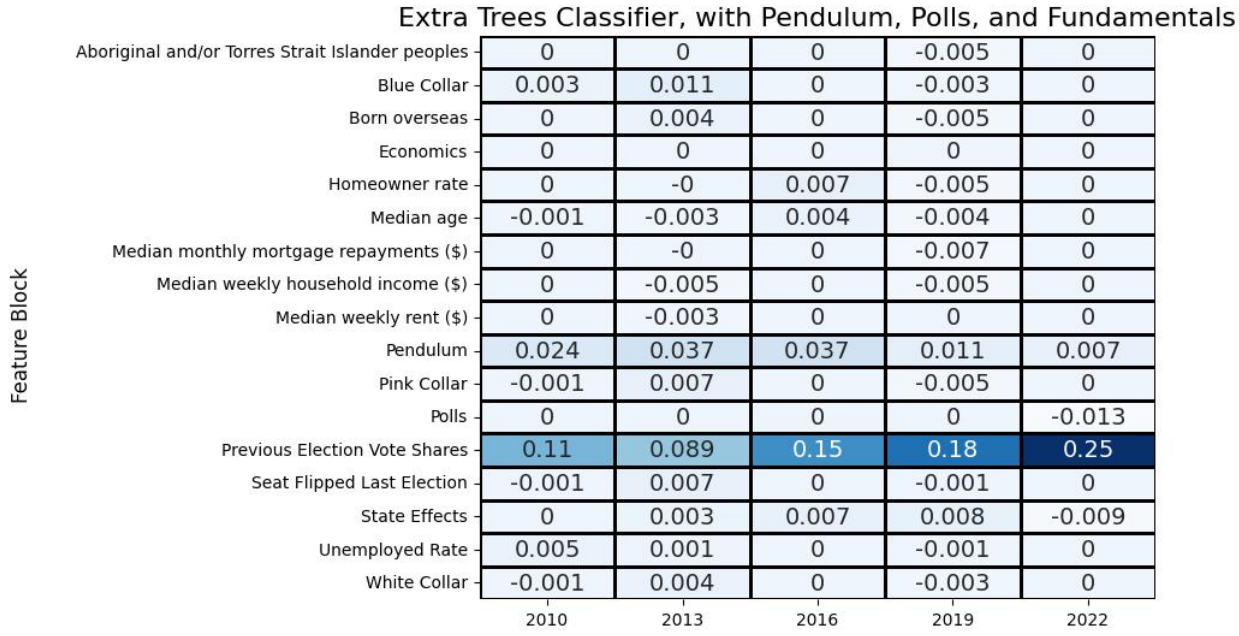


Figure 5.6: *The PFI of each block of predictors for the best-performing model (ET with pendulum, polls, and fundamental predictors). Higher positive values indicate that including the feature block improves accuracy, negative values indicate that withholding this block improves accuracy.*

5.3.3 Evaluation

Table 5.5 compares the ET model against the evaluation criteria outlined in the Introduction. Based on these criteria, ET appears to be a good technique to continue to trial in future elections although with caveats. Since 2019, the pendulum has been less accurate in both seat classification and seat counts. The ET model appears to classify seats either roughly as well as the pendulum (from 2010 to 2016) or 5-15 seats better (2019-2022) with a three month lead time. However, seat count error was often higher than the pendulum. Although ET never miscalled the major party winner, the actual seat count rarely fell within the 95% confidence interval, meaning the model was too sure of an incorrect outcome. Another possibility could be to use GB for seat count estimation (accepting that GB may be vulnerable to polling error the same as the pendulum is) and ET for seat classification.

5.4 Discussion

The Mackerras pendulum has been a reliable predictive model in Australian politics, accurately forecasting the winning party in all but three elections since 1972. However, as polling response rates decline and errors increase, the pendulum's effectiveness may diminish, particularly if the crossbench expands. This will strain the model because it relies on accurate polls and is naive about predicting minor party seats. In order to develop models that outperform the pendulum, it is necessary to experiment with new approaches. This paper is the first to apply machine learning techniques to the Australian context, following successful applications in other democracies.

This paper trialed a wide range of algorithms and types of predictors and revealed two notable models. Firstly, an extra trees classifier with pendulum, polling, and fundamental features was the best at classifying seat outcomes, outperforming the pendulum by 15 seats in 2019 and five in 2022. However, in elections 2010-2016, the ET model performed roughly as well as the pendulum, which would not have justified adopting a more complex model as of that time. Secondly, a gradient boosted classifier with pendulum and polling predictors forecasted seat counts with greater accuracy than the pendulum in every election from 2010 to 2022 except 2019, in which the pendulum also performed poorly. Each model achieved this performance as early as three or six months before election day. If the pendulum continues to perform as poorly as in 2019 and 2022, then these ML techniques may be very valuable in accurately calling both the overall winner and that of seat contests. Even an improvement of a few seats could help better anticipate the outcomes of tight elections like 2010, 2016, 2019, and 2022.

Further improvements can be made to these models. The PFI analysis shows that the extra trees model corrects pendulum forecasts by factoring in margins, state effects, and unemployment. Simplifying the model by pruning less impactful features could make it simpler and more accurate. Another key improvement would be combining the seat count accuracy of the gradient boosted model with the classification accuracy of the extra trees model, resulting in a single model that excels at both tasks. This might be achieved by ensembling the two techniques. Finally, tree-based models are inherently more complex

Table 5.5: *Comparing the best discovered ML model to the evaluation criteria.*

Criteria	Extra trees with pendulum, polls, and fundamentals
Seat count accuracy (see Figure 5.4, Table 5.4, and Appendix Section 5.5.4).	Taking each model's most likely predictions at three months' lead time, ET was closer to the actual seat share for the major parties than the proportional pendulum by up to 16 seats in the 2019 and 2022 elections but farther away by 5-13 seats in other elections.
Seat classification accuracy (see Figure 5.3, Table 5.3, and Appendix Section 5.5.4).	ET is overall a much better seat classifier than any pendulum model. The worst performance by ET was one seat additional misclassification than the TPP proportional pendulum in 2013, but ET misclassified 15 fewer seats in 2019 and five fewer in 2022. Marginal seat and crossbench seat prediction is a particular weaknesses of the pendulum. In this regard, ET was a better marginal seat classifier than the TPP proportional pendulum at every election except 2013 by up to 24% accuracy (Table 5.3). ET is roughly the same as the pendulum for predicting crossbench winners.
Consistency	ET was consistently either roughly as accurate or better than the pendulum at classification and never worse at seat count forecasting to the extent that it would miscall the winning major party. In comparison, the proportional and uniform pendulum models miscalled 2019.
Lead time	In all elections except 2013, ET was equally as or more accurate (classification) than the pendulum by six months lead time (Figure 5.3). Although less accurate than the pendulums at seat counting, ET correctly predicted the winning major party of each election by six months lead time.
Uncertainty calibration	ET had smaller uncertainty ranges than the pendulum in both seat count and classification accuracy. However, the actual seat count rarely fell within the 95% confidence interval of the ET model, meaning the model was over-confident of a wrong prediction.
Parsimony	This ET model required many different kinds of indicators, including polls, fundamentals, and pendulum predictions. As a result, the model was prone to losing accuracy due to fitting to features in the training elections that did not generalize to the test election (Figure 5.6). This indicates that some features could be pruned from the model to make it simpler and more accurate. The ET model is much more complex than the TPP pendulum, so it is only justifiable to use this ML model if the pendulum continues to perform at its 2019-2022 level as opposed to its higher accuracy scores in 2010-16 (Figures 5.3 and 5.4).

than pendulum models. Developing a logistic regression model incorporating interaction and nonlinear effect terms could produce a model as accurate as the tree-based ones but simpler and easier to interpret.

International electoral forecasters may find value in using machine learning (ML) to improve uniform swing models rather than building synthetic models based on fundamentals, margins, and polls. Incorporating pendulum predictions enhanced the extra trees model. Even when the pendulum performed poorly (2019-2022), the other features were sufficient to correct the misclassifications.

A key limitation of this study is that all predictions were retrospective. The true test of any electoral model is its ability to make prospective forecasts. Although I structured the modeling to closely simulate how the model would have forecasted at inference time, I selected the best hyperparameters retrospectively. As a result, there is a risk that I may have chosen a model that coincidentally performed well over five elections (and at five lead times each). While consistent performance across all five elections reduces the likelihood of overfitting, the real test will be applying the extra trees and gradient boosted models to future elections. Another limitation is the reliance on publicly available datasets. Pollsters collect rich data, such as undecided rates, nonresponse rates, and respondent postcodes for geo-location, but this data is rarely made available across multiple elections and lead times. Novel data sources could be explored, especially if they help predict swing seats, which are typically the hardest to forecast. Marginal seat-specific polls often fail to accurately predict swing seats (Goot, 2023) but could still perhaps improve ML-based forecasts. Another possible source of data could come from projects such as Evershed and Nicholas (2022), which use social media and news data to detect and categorize announcements of public works projects targeted at marginal seats near election time. If these public announcements have an electoral impact, a synthetic model could exploit this data to anticipate which marginals will flip more precisely.

5.5 Appendices

5.5.1 Replication Material

Reproduction code, including all data used in the modeling, can be found at the following link. All code originally run in python version 3.10.

`https://osf.io/rvc2f/?view_only=a6f7c8360a3d4312ab6681dea0e1500e`

5.5.2 Additional Information about the Data

List of Predictors

This section details all predictors used in the study. Note that some predictors do not have a feature set because they are always included in each model.

Table 5.6: *Details of every variable used in the study.*

Feature Set	Feature Block	Variable Name	Values	Description
Fundamentals	Aboriginal and, or Torres Strait Islander peoples	Aboriginal and, or Torres Strait Islander peoples	[0, 1]	The portion of the seat's population that identifies as Aboriginal and, or Torres Strait Islander.
	Blue collar	Blue collar	[0, 1]	The portion of the workforce in this seat is classified as "technicians and trades workers, machinery operators and drivers, or laborers.
	Born overseas	Born overseas	[0, 1]	The portion of the seat's population born overseas.
	Median age	Median age	[0, inf)	The median age among the seat's population.
	Pink collar	Pink collar	[0, 1]	The portion of the workforce in this seat classified as community and personal service workers", clerical and administrative workers, or sales workers.
	White collar	White collar	[0, 1]	The portion of the workforce in this seat classified as managers or professionals.
	Unemployment	Unemployment	[0, 1]	The portion of the seat's population identified as unemployed and looking for work (part-time or full-time) on the census.
	Homeowner	Homeowner	[0, 1]	Portion of the seat's population who own their residence.
	Median monthly mortgage repayments ($)	Median monthly mortgage repayments ($)	[0, inf)	The median monthly mortgage repayment among the seat's population.
	Median weekly household income ($)	Median weekly household income ($)	[0, inf)	The median weekly household income among the seat's population.
	Median weekly rent ($)	Median weekly rent ($)	[0, inf)	The median weekly rent among the seat's population.
	Economics	Real GDP growth from the previous year	(-inf, inf)	For example, in 2019, I used the real GDP growth rate for 2018 as a predictive feature to estimate voter sentiment about the macro economy.
	Economics	Inflation rate (consumer prices) previous year	(-inf, inf)	For example, in 2019, I use the inflation rate of consumer prices for 2018 as a predictive feature to estimate voter sentiment about the macro economy.

Continued on next page...

Feature Set	Feature Block	Variable Name	Values	Description
-	Flipped last election	Flipped last election	0, 1	This binary variable is true if the seat voted out its incumbent in the previous election.
-	Previous Election Vote Share	Highest share of preferences previous election [ALP, LNP, OTH]	[0, 1]	For each party category (ALP, LNP, or OTH), the maximum portion of preferences reached by that party in the previous election before it was either eliminated in a preference-round or a winner passed 50
Pendulum	Pendulum	Pendulum [ALP, LNP, OTH]	[0, 1]	A TPP proportional pendulum model's prediction for the given seat.
Polling	Polls	National- and state-level [ALP, LNP, OTH] polling average and standard deviation at [six months, three months, two months, one month, one week] lead	[0, 1]	For each party (ALP, LNP, or other), their respective polling average and standard deviation over all polls published in the four weeks ending at the given lead time before election day. There is a national- and state-level version of each variable.
	Polls	National- and state-level [ALP, LNP] TPP polling average and standard deviation at [six months, three months, two months, one month, one week] lead	[0, 1]	For each major party (ALP and LNP), their respective TPP polling average and standard deviation over all polls published in the four weeks ending at the given lead time before election day. There is a national- and state-level version of each variable.
-	State Effects	State [ACT, NSW, NT, QLD, SA, VIC, WA]	[0, 1]	Indicates the state or territory that the given seat is in.

State Polls

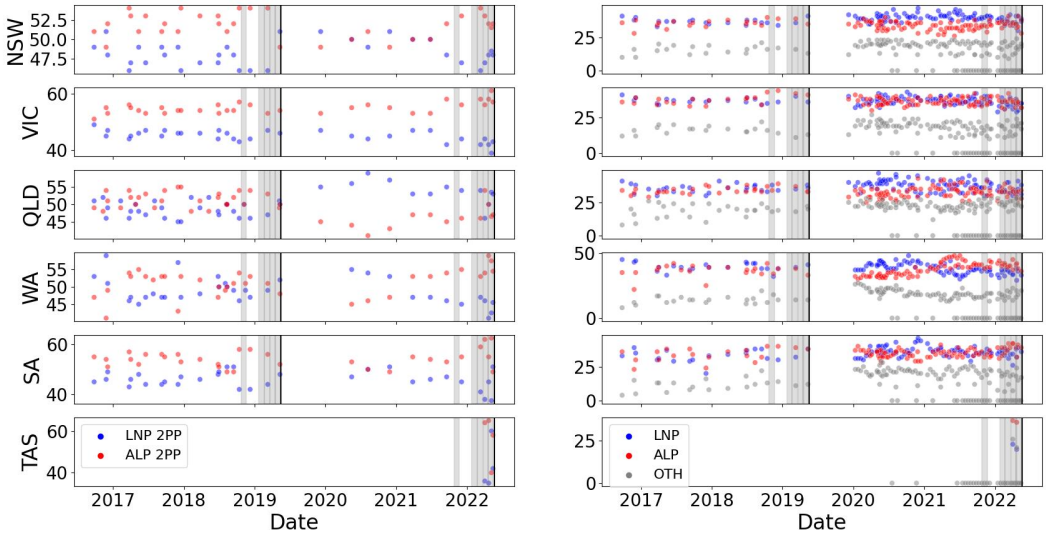


Figure 5.7: *State-level polling values and the time window over which the polling aggregates were taken. Black lines show election days.*

Economic Indicators

Figure 5.8: *Timeline of macroeconomic variables.*

Data Sources

A copy of all original data used in this study are available with the replication material (Appendix Section 5.5.1). Those data are drawn from the sources indicated in Appendix Table 5.7.

Table 5.7: *Sources for each variable category.*

Category	Period	Source
Seat-level socioeconomics (Fundamentals)	2006–2021	Statistics (2021)
Electoral history	2001	Australian Electoral Commission (2001)
	2004	Australian Electoral Commission (2004)
	2007	Australian Electoral Commission (2007)
	2010	Australian Electoral Commission (2010)
	2013	Australian Electoral Commission (2013)
	2016	Australian Electoral Commission (2016)
	2019	Australian Electoral Commission (2019a)
	2022	Australian Electoral Commission (2022)
Polling	2004	PhantomTrend (2016)
	2007	PhantomTrend (2016)
	2010	PhantomTrend (2016)
	2013	PhantomTrend (2016)
	2016	PhantomTrend (2016)
	2019	Bowe (2023)
	2022	Bowe (2023)
National macroeconomic variables (fundamentals)	2006–2021	International Monetary Fund (2024)

5.5.3 Additional Information about the Models

Overview of Machine Learning Models

Table 5.8: *Summary of the selected models in this study and a comparison of their strengths and weaknesses.*

Model	Description	Strengths	Weaknesses
Logistic regression.	I trial logistic regression models without penalties and with lasso and ridge penalties, which protect against overfitting (Le Cessie and Van Houwelingen, 1992; Tibshirani, 1996).	Excels when cases can be classified by summarizing a set of variables multiplied by weights. Successful in a diverse range of applications and is easy to interpret.	It can only account for nonlinearity or interaction effects if the modeler specifically builds the model to handle these (which I will not explore in this study). It may require large training data to attain good results and can be confounded when predictors are correlated.
Extra trees classifier	This algorithm is a tree-based model that is often successful in cases where there are complex interactions between variables. Extra trees builds an ensemble of classification trees that are tuned to maximize homogeneity of the leaves and then takes the portion of "votes" from the trees as the classification probabilities. Unlike logistic regression, extra trees can automatically account for interaction effects (Geurts, Ernst, and Wehenkel, 2006).	Automatically handles interaction effects. More robust against overfitting than neural networks but more vulnerable than logistic regressions.	Performance can be slow as the number of trees increases. They cannot handle interaction effects as complex as neural networks can.
Gradient boost	This algorithm is similar to extra trees; however, the ensemble of trees is built additively, potentially reducing error more effectively (Friedman, 2001).	More robust against outliers in the data than extra trees.	Same as an extra trees model, but it requires longer fitting times.

Continued on next page...

Model	Description	Strengths	Weaknesses
Multilayer perceptron	This algorithm is a type of neural network. Neunhoeffer et al. (2020) successfully applied neural networks to seat-level predictions in Germany, so I evaluate this same type of algorithm here (Haykin, 1994).	Highly effective in fitting to complex interaction effects.	Long computation times and vulnerable to overfitting when interaction effects are not sufficiently complex.

Hyperparameters

Table 5.9: *Hyperparameter grid.*

Model	Parameter	Values	Count
Extra trees	Number of estimators	50,100,1000	3
	Criterion function for measuring split quality	'gini, 'entropy', 'log loss'	3
	Sub total		9
Penalized logistic regression	Penalty function	lasso (L1), ridge (L2)	2
	Penalty rate (C)	0.1, 0.5, 1.0, 2.0	4
	Sub total		8
Unpenalized logistic regression			1
	Sub total		1
Gradient boost	Number of estimators	50, 100, 1000	3
	Sub total		3
Multilayer perceptron	Number of hidden layers	1, 2	2
	Number of neurons in each layer	64, 128	2
	Activation functions in each neuron	sigmoid, relU, tanh	3
	Sub Total		12
Total settings			33

Pendulums

Two-Party Preferred Uniform and Proportional Swing

This is an implementation of the Mackerras pendulum (Mackerras, 1976) used by the Australian Broadcasting Company (Green, 2016, 2019, 2022) to predict election outcomes. The algorithm is based on the idea that each seat in an election will shift

its two-party vote share by similar amounts, clustering around an average (the uniform swing). Although the actual changes in each seat may vary, they tend to be close enough to the average that this system will make good predictions. This allows for predicting the overall number of seats for each major party by estimating the national average swing. In this implementation, I estimate that swing by taking the two-party vote share from the previous election as the true TPP (two-party preferred) support at the time of that election and then using the TPP polling average to estimate the current two-party support. The estimated swing is the difference between those two numbers.

In this system, if a given seat is held by a non-major party (OTH) before an upcoming election, then I predict that its incumbent will retain that seat. Therefore, the seats predicted to be won by OTH parties will always be the same as those held by OTH parties before the election. If the seat is held by either ALP or LNP, I determine my predicted outcome for that seat as follows (see Equations 5.1 – 5.2). First, I calculate each party's national two-party-preferred vote share (TPV) at each election as the percentage of total votes that were counted towards that party. If neither of the two major parties won that seat, I counted only their votes before they were eliminated from the voting rounds and their preferences were transferred. TPV of each major party will sum up to 100%, but the total number of votes may not equal the number of votes cast because of the seats won by minor parties. Next, for each election, I calculate the swing for each party as the national two-party preferred polling average (TPP) minus the TPV from the previous election. If ALP or LNP held the seat then for each seat at each election if the incumbent's margin plus the swing is less than or equal to zero, I predict that the other major party will win the seat. Otherwise, I predict the incumbent major party will hold the seat. In the proportional version of this pendulum, if a state TPP poll value is available, that value is used instead.

$$Swing_{2022}^{Party} = TPPPollingAverage_{2022}^{Party} - TPV_{2019}^{Party} \tag{5.1}$$

$$\text{PartyVictory}_{2022} = SeatMargin_{2019}^{Party} + \hat{Swing}_{2022}^{Party} > 0 \tag{5.2}$$

If the party is OTH, the prediction is OTH. If the incumbent is not the predicted winner, the other major party is the expected winner.

Two Candidate Vote Uniform Swing (TCV Uniform)

This system is based on the Reed pendulum proposed by Resolve Strategic (Reed, 2022). This version provides a more sophisticated treatment of minor party seats, which is becoming more important as non-major parties hold more seats. This system calculates swing based on the two-candidate votes share (TCV), where I calculate the votes won by each party in each seat and the votes won by whichever party won the second most

votes in that seat. All votes for each party category (ALP, LNP, and OTH) are summed up, and the percentage for each category is calculated. In this system, every vote is counted in the percentages, and each category's total percentage is 100%. I calculate each party's national swing for each election, that party's polling average minus their TCV from the previous election. Next, if the incumbent party's margin plus their party's national swing is less than or equal to zero for each seat, I predict the party with the second most votes in the previous election will win this seat (see Equation 5.3).

$$Swing_{2022}^{party} = PollingAverage_{2022}^{Party} - TCV_{2019}^{Party} \tag{5.3}$$

5.5.4 Additional Results

For an expanded version of Table 5.3 which presents the accuracy scores for all models across all elections and lead times, see the reproduction materials' file:

output/all_models_uncertainty_summary_all_and_marginal_seats.csv

For an expanded version of Table 5.4 which presents the seat count error for all models across all elections, parties, and lead times, see the reproduction materials' file:

output/all_seat_count_errors_summary.csv

Alternative Training Sizes

Appendix Figures 5.9 and 5.10 show that when the ML models are fitted on all preceding training elections, no single ML model consistently, or in overall average accuracy at three months lead time, exceeds the pendulum in either classification or seat count accuracy. For example, ET with pendulum, polls and fundamental predictors now underperforms against (instead of equaling) the proportional pendulum in 2016 at three months lead time (Appendix Figure 5.9). This is why the main paper presents results for ML models trained on up to two preceding elections.

Raw Prediction Data

The predictions made by every model for every trial can be found in the replication material at file location:

output\pickles\raw.pkl

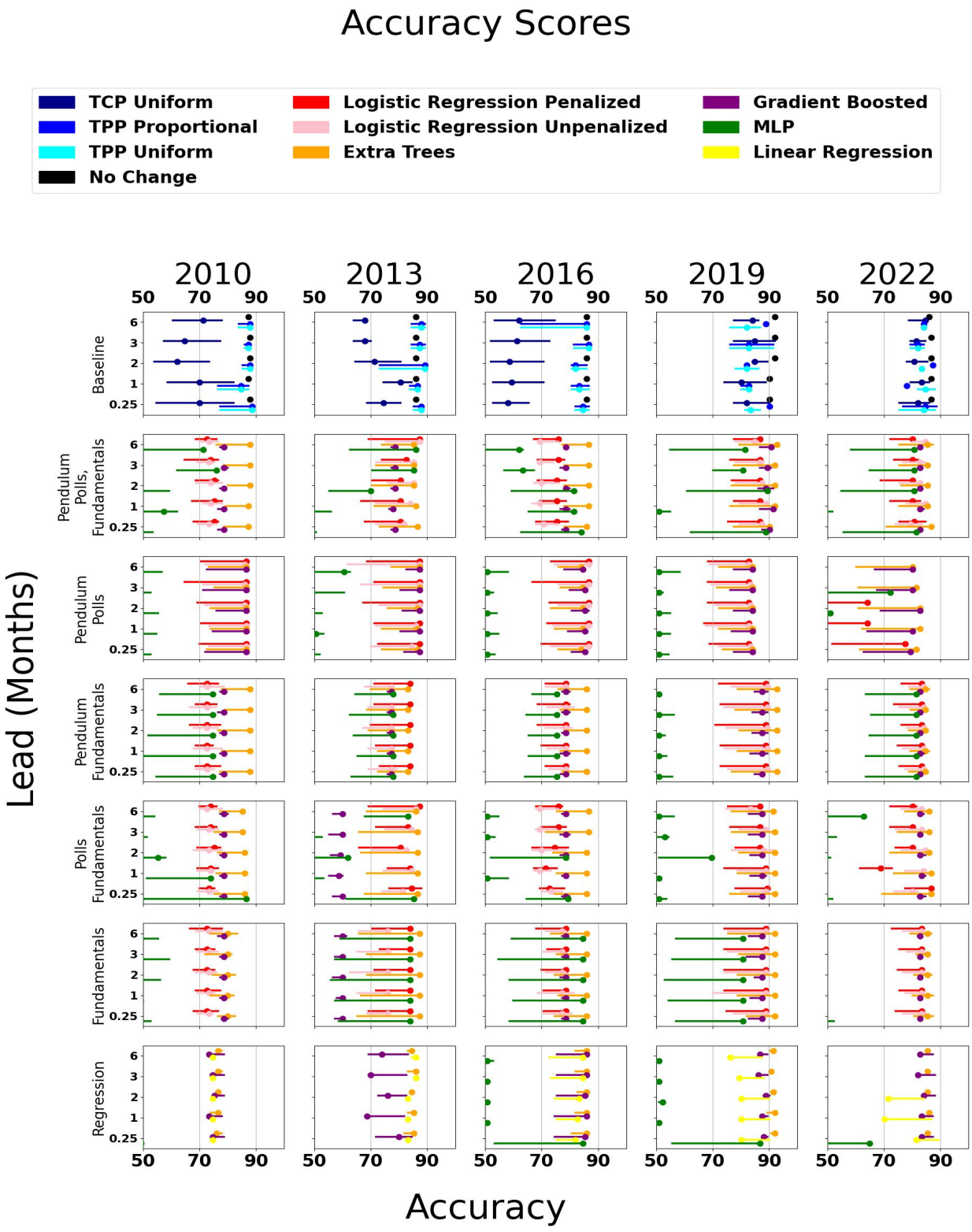


Figure 5.9: *Classification accuracy across models and feature sets when the models are trained on all elections since 2007 instead of the most recent two.*

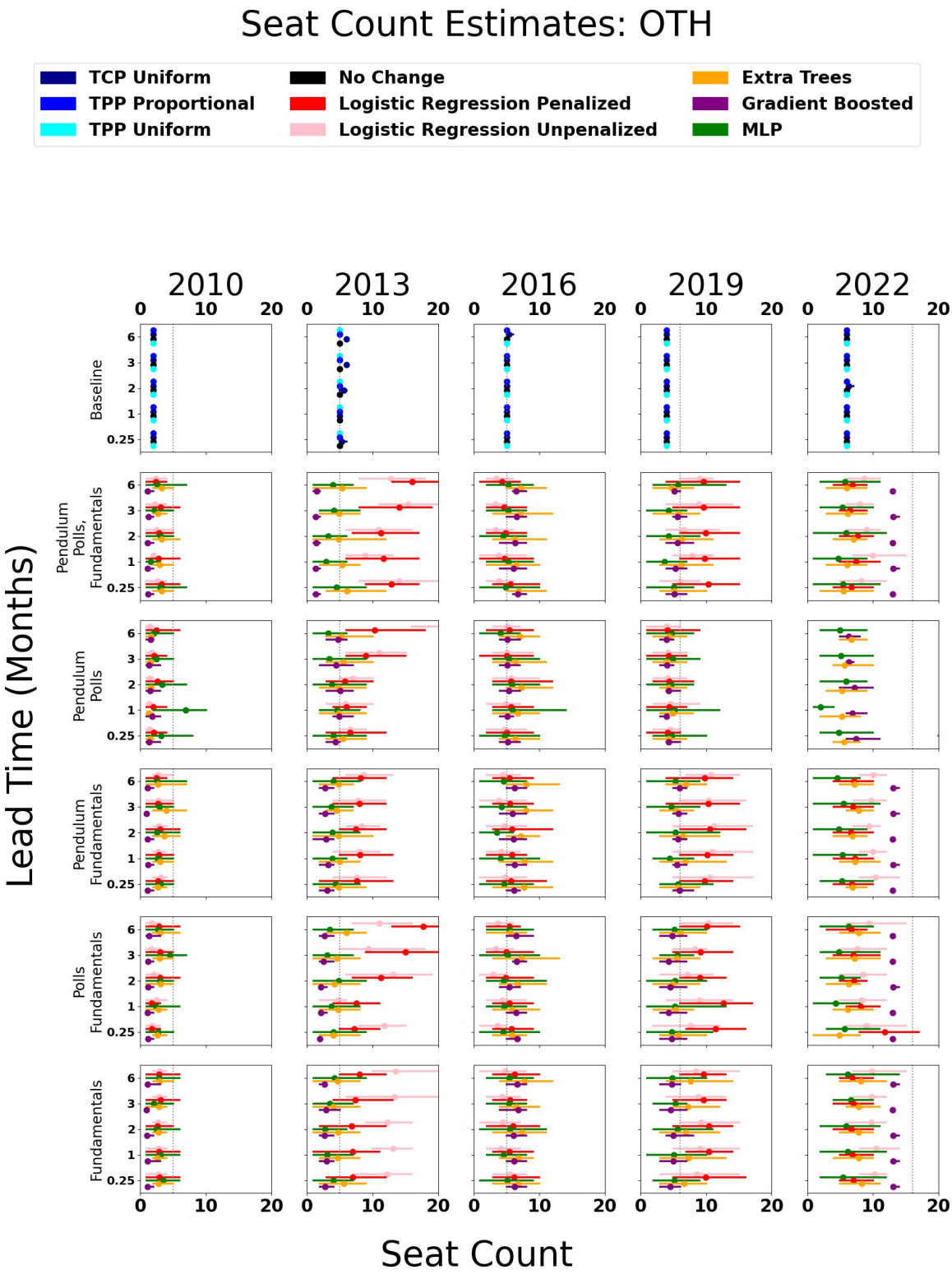


Figure 5.10: *Seat count accuracy when the models are trained on all elections since 2007 instead of the most recent two.*

Additional Seat Classification Results

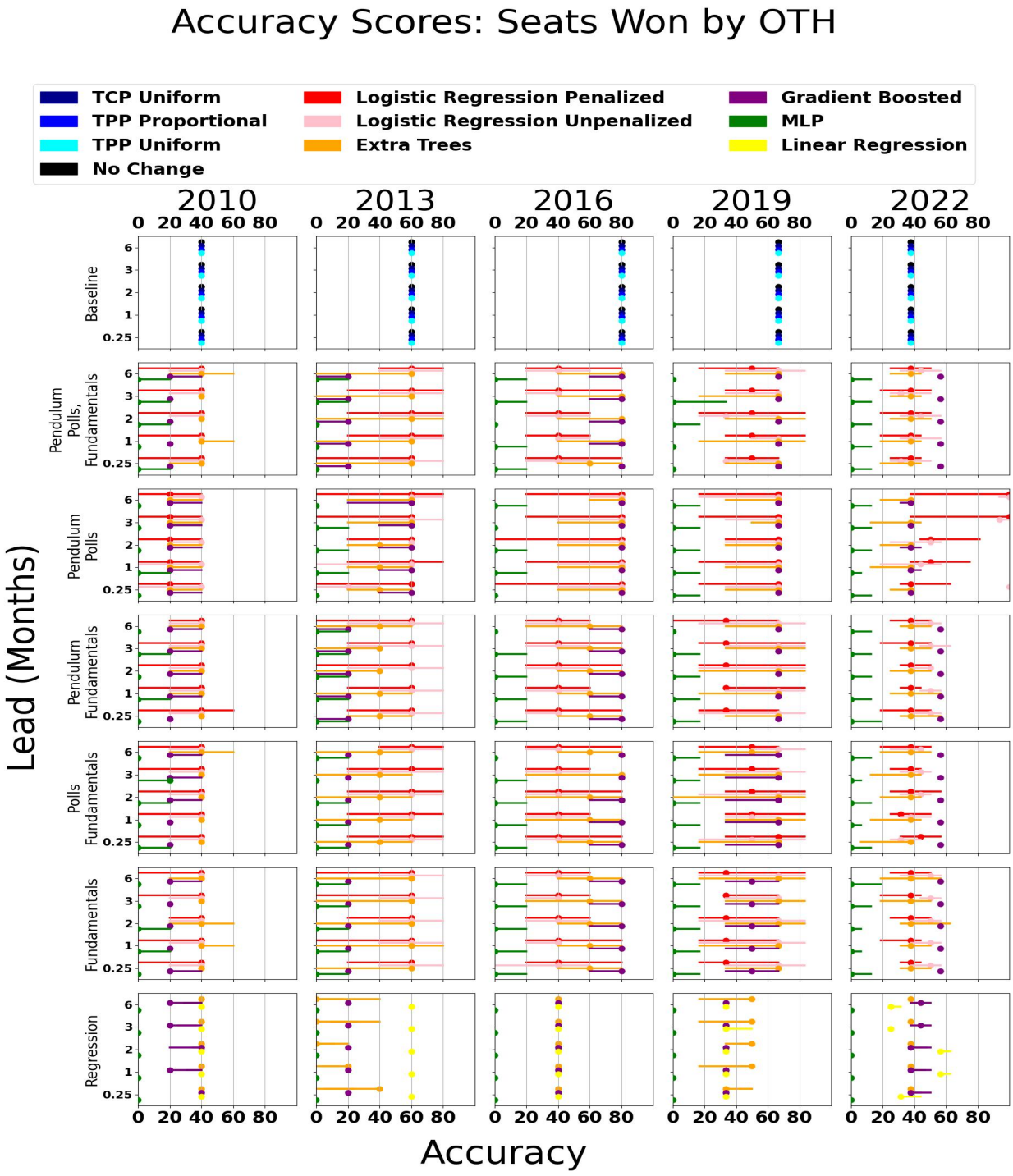


Figure 5.11: *Classification accuracy across models and feature sets when summarized only for seats that were won by a minor party.*

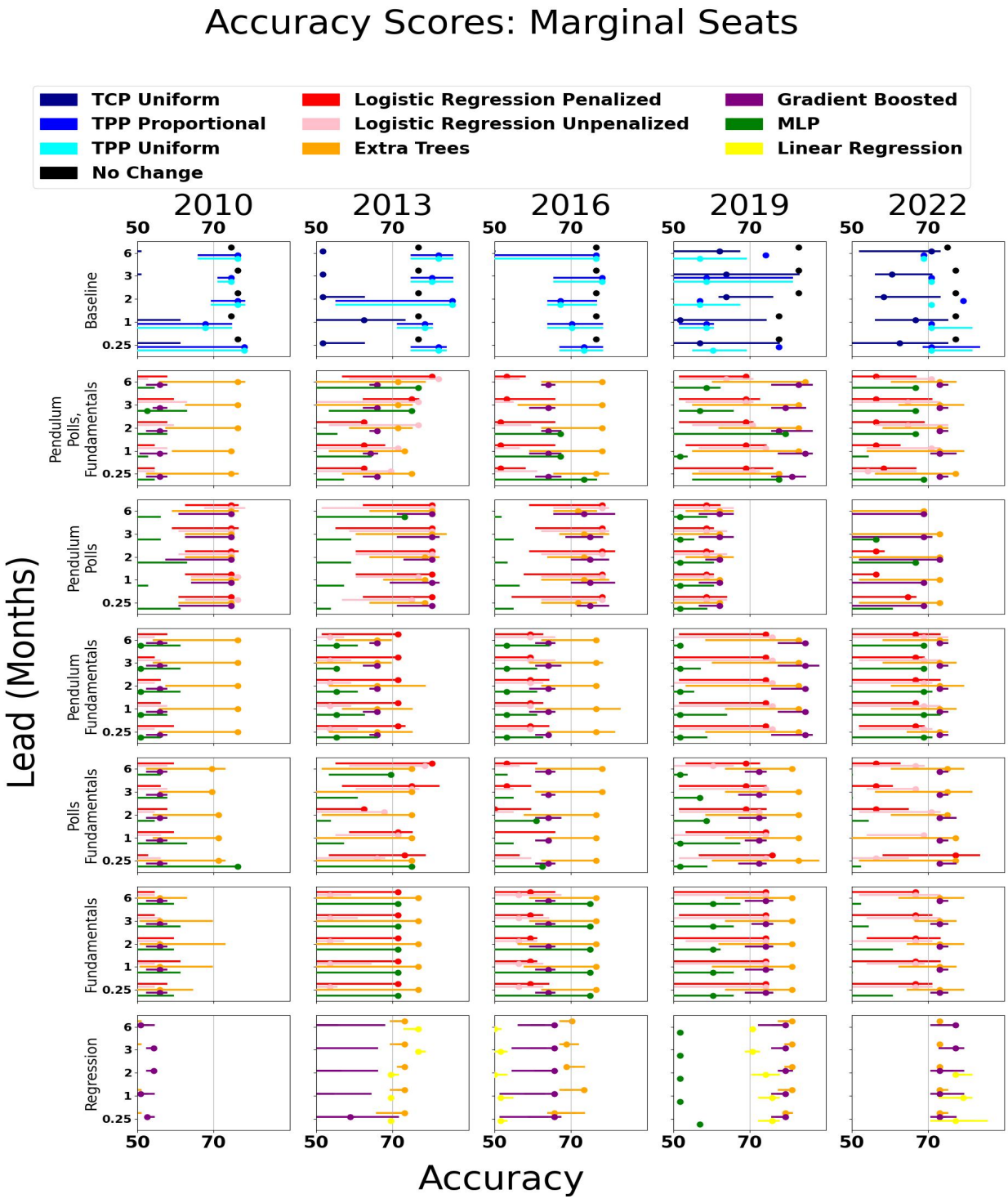


Figure 5.12: *Classification accuracy across models and feature sets when summarized only for seats with a pre-election margin less than 6% which is the Australian Electoral Commission's standard for the 'marginal' categorization.*

Additional Seat Count Results

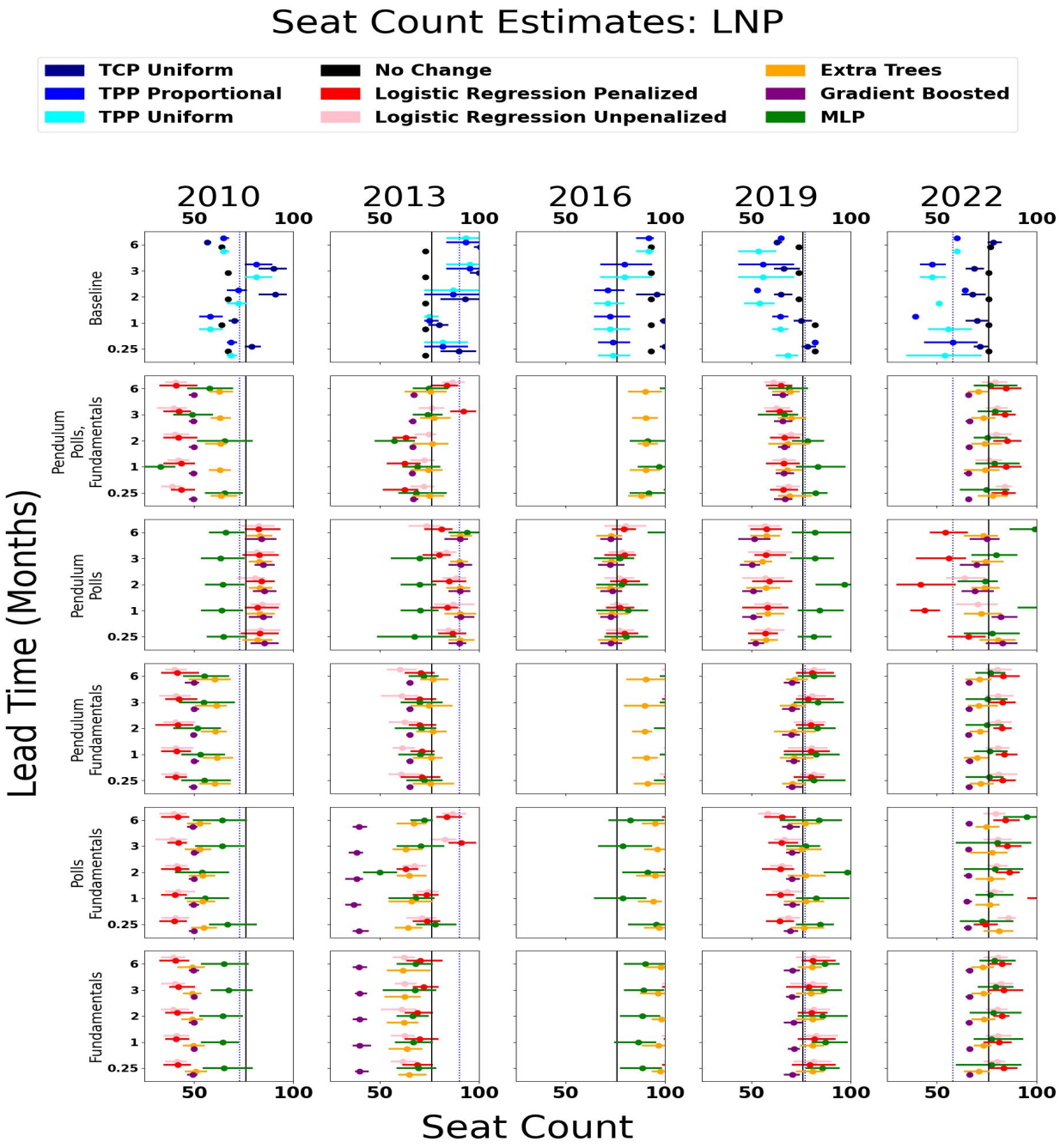


Figure 5.13: *The count of seats for the LNP as predicted by each model. The dotted line indicates the actual seat count for that election, and the solid line represents the 76 seats required to form a majority government.*

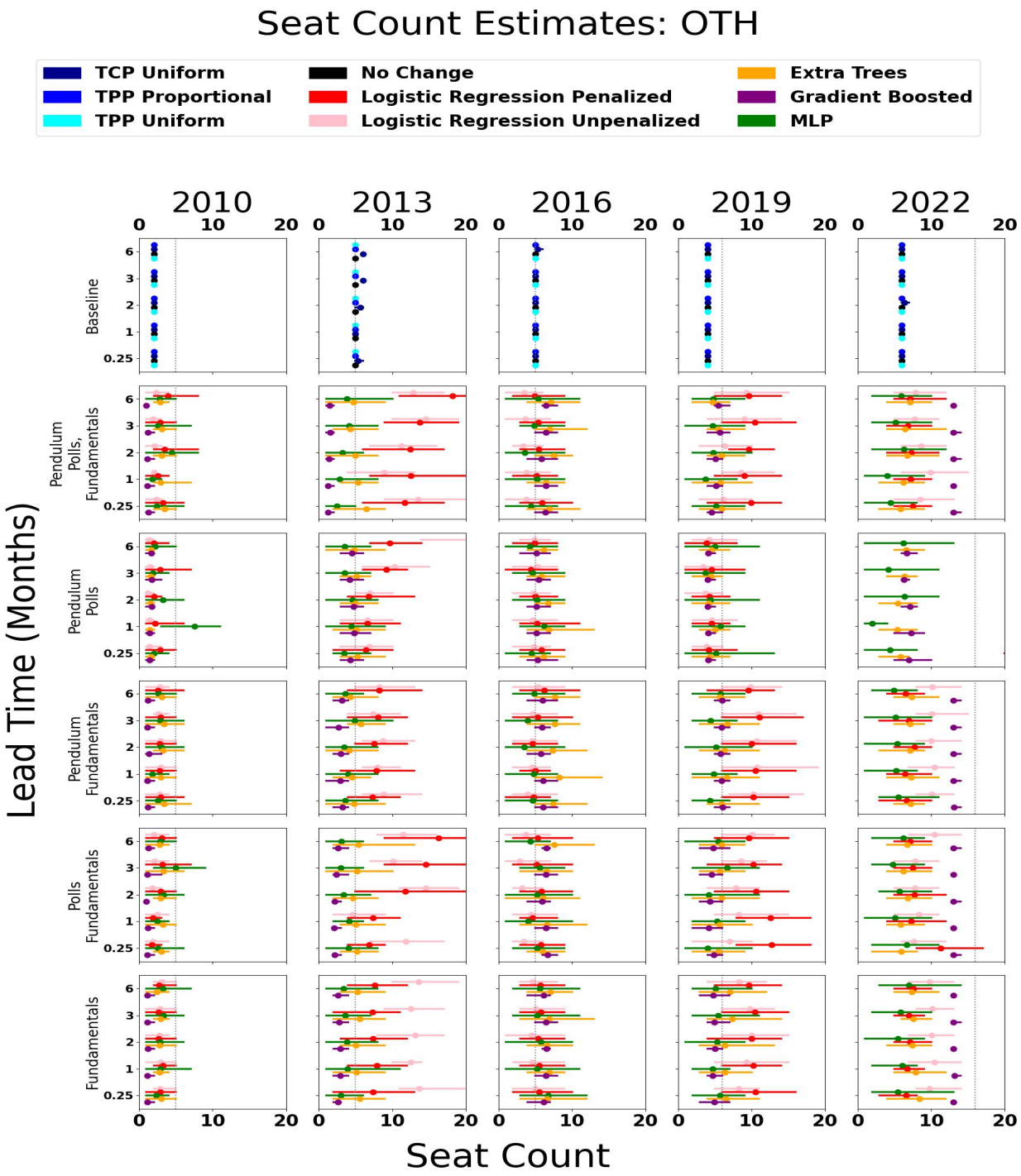


Figure 5.14: *The count of seats for the minor parties as predicted by each model. The dotted line indicates the actual seat count for that election, and the solid line represents the 76 seats required to form a majority government.*

Post-Election Seat Margin versus Vote Share Estimate

For each regression model, at three months before each election, Appendix Figure 5.13 shows the predicted versus actual vote shares for every seat. The variance in predictions comes from the bootstrapping technique discussed in Section 5.2.2. MLP has very wide 95% confidence intervals whereas other models are relatively consistent. Figure 5.16 shows that OTH vote margins in seats won by the crossbench are typically underestimated by every model.

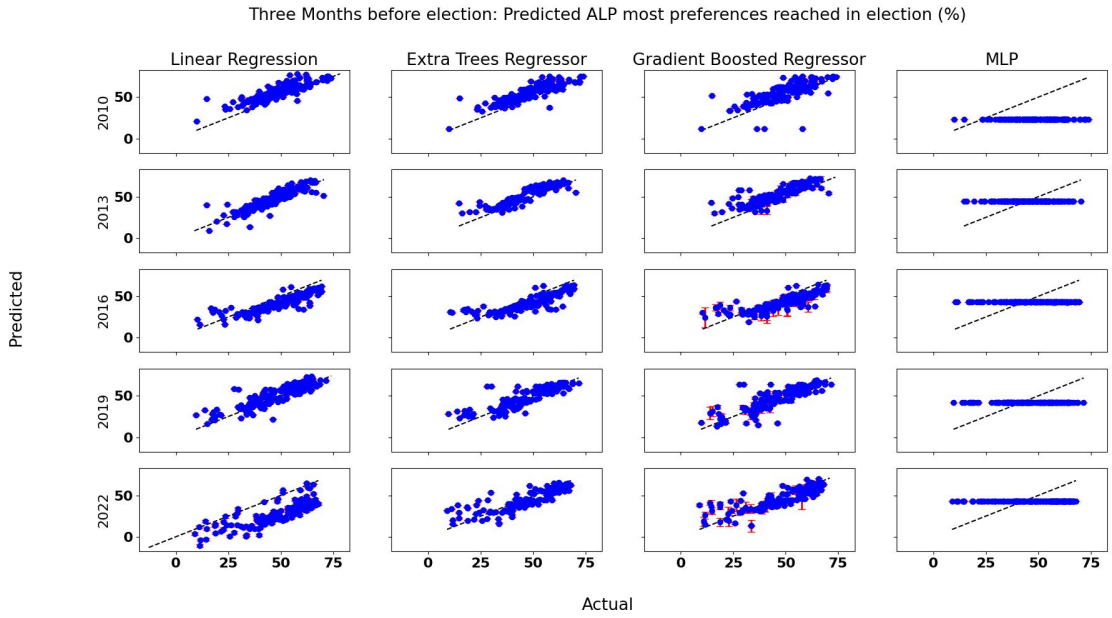


Figure 5.15: *Part 1: A scatter plot showing how each regression model's predicted versus actual vote shares. The dotted line indicates where perfect predictions would fall. The red bars are the 95% confidence interval of each regressor across bootstrapped predictions.*

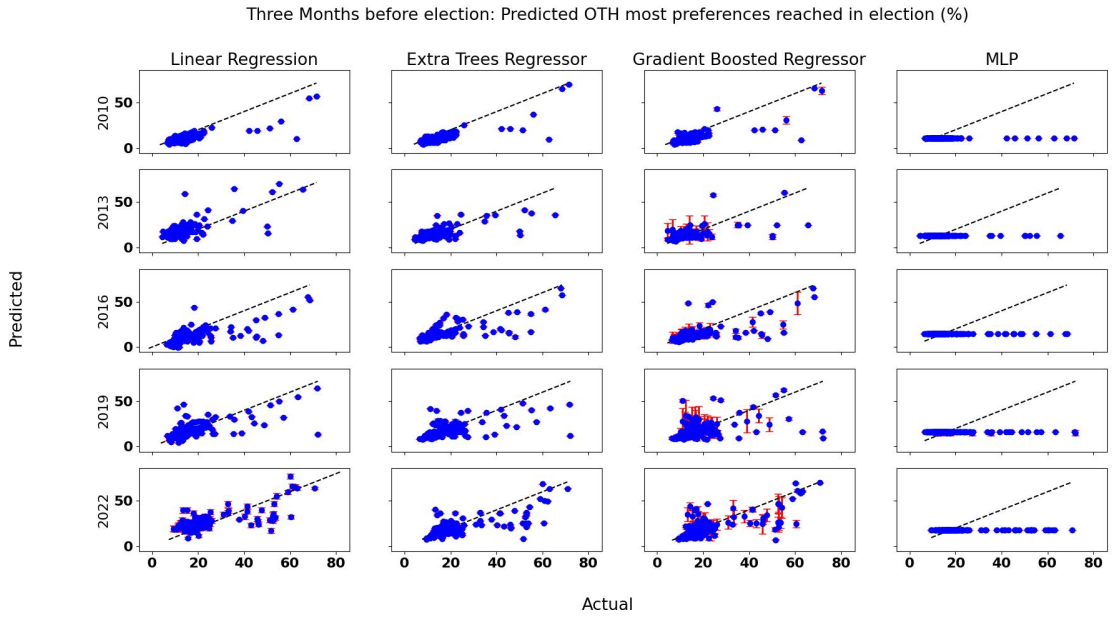


Figure 5.16: *Part 2: A scatter plot showing how each regression model's predicted versus actual vote shares. The dotted line indicates where perfect predictions would fall. The red bars are the 95% confidence interval of each regressor across bootstrapped predictions.*

Post-Election Seat Margin versus Classification Certainty

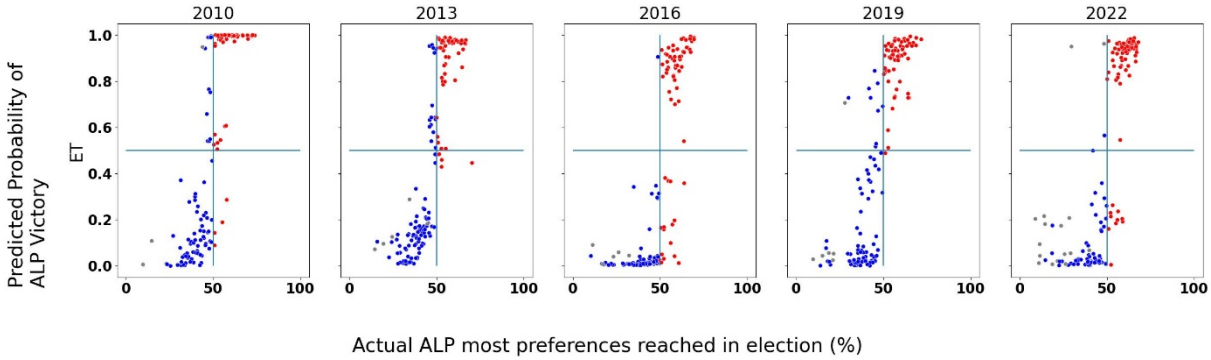


Figure 5.17: *A scatterplot showing the ET with pendulum, polls, and fundamentals model's classification probability for the ALP versus the actual vote share that the ALP reached in that election. The horizontal solid line represents a 50% certainty of an ALP classification, and the vertical line represents 50% of the seat's votes.*

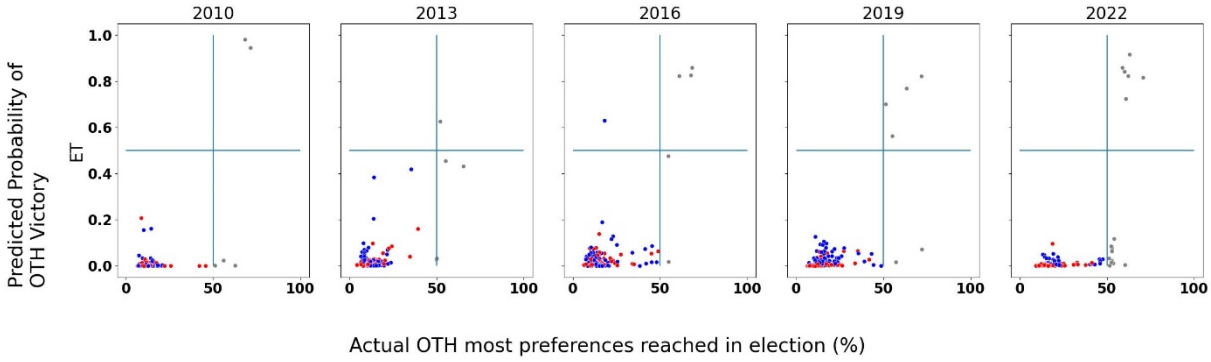


Figure 5.18: *A scatterplot showing the ET with pendulum, polls, and fundamentals model's classification probability for a minor party versus the actual vote share that any minor party reached in that election. The horizontal solid line represents a 50% certainty of an OTH classification, and the vertical line represents 50% of the seat's votes.*

Uniform Pendulum Confusion Matrices

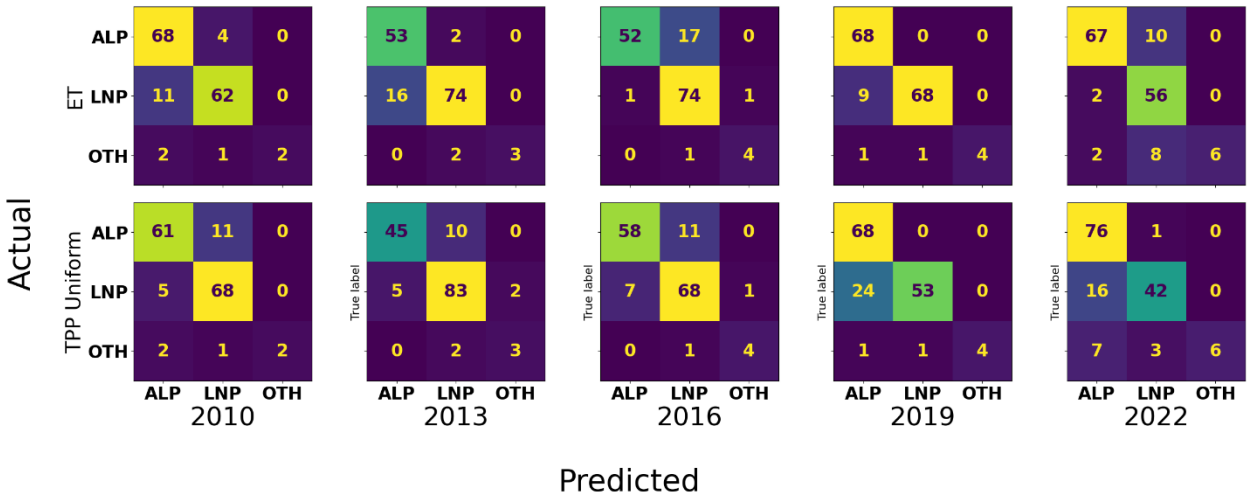


Figure 5.19: *Confusion matrices in which a uniform instead of proportional TPP pendulum was used for the pendulum predictions both as a model and inputs for the ML models.*

5.6 References

Altmann, André et al. (2010). "Permutation importance: a corrected feature importance measure". In: *Bioinformatics* 26.10, pp. 1340–1347.

Argandoña-Mamani, Alexander et al. (2024). "Predicting Election Results with Machine Learning—A Review". In: *Proceedings of Eighth International Congress on Information and Communication Technology*. Ed. by Xin-She Yang et al. Lecture Notes in Networks and Systems. Springer Nature, Singapore, pp. 989–1001.

Arnold, Christian et al. (2024). "The Role of Hyperparameters in Machine Learning Models and How to Tune Them". In: *Political Science Research and Methods* 12.4, pp. 841–848.

Australian Electoral Commission (2001). *AEC: When: Past Electoral Events* URL: `https://results.aec.gov.au/10822/Website/index.html`.

— (2004). *Election 2004: the official election report and results* URL: `https://results.aec.gov.au/12246/default.htm`.

— (2007). *2007 Federal Election Results* URL: `https://results.aec.gov.au/13745/`.

— (2010). *Australian Electoral Commission Virtual Tally Room* URL: `https://results.aec.gov.au/15508/Website/Default.htm`.

— (2011). *Compulsory voting in Australia*. Australian Electoral Commission URL: `https://www.aec.gov.au/About_AEC/publications/voting/`.

— (2013). *Australian Electoral Commission Virtual Tally Room* URL: `https://results.aec.gov.au/17496/Website/Default.htm`.

— (2016). *2016 Federal Election*. Australian Electoral Commission URL: `https://results.aec.gov.au/20499/Website/HouseDefault-20499.htm`.

— (2019a). *2019 Federal Election*. Australian Electoral Commission URL: `https://results.aec.gov.au/24310/Website/HouseDefault-24310.htm`.

— (2019b). *AEC Fact-Sheet*. 2019.

— (2022). *2022 Federal Election*. Australian Electoral Commission URL: `https://results.aec.gov.au/27966/Website/HouseDefault-27966.htm`.

Bergmeir, Christoph and José M. Benítez (2012). "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191. Publisher: Elsevier, pp. 192–213.

Bowe, William (2021a). *YouGov MRP poll (part one) and more – The Poll Bludger* URL: `https://www.pollbludger.net/2022/05/11/yougov-mrp-poll-part-one-and-more/`.

— (2021b). *YouGov MRP poll (part two): Labor 80 seats, Coalition 63, others 8 – The Poll Bludger* URL: `https://www.pollbludger.net/2022/05/12/yougov-mrp-poll-part-two-labor-80-seats-coalition-63-others-8/`.

— (2023). *The Poll Bludger – Analysis and discussion of elections and opinion polls in Australia* URL: `https://www.pollbludger.net`.

Browne, Bill (2022). *Between Sense and Nonesense: the predictive power of the electoral pendulum*. The Australia Institute, Canberra.

Curtice, John and David Firth (2008). "Exit polling in a cold climate: the BBC–ITV experience in Britain in 2005". In: *Journal of the Royal Statistical Society: Series A*

(Statistics in Society) 171.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-985X.2007.00536.x, pp. 509–539.

Evershed, Nick and Josh Nicholas (2022). "How we used machine learning to cover the Australian election". In: *The Guardian*.

Fisher, Stephen D. (2016). "Piecing it all together and forecasting who governs: The 2015 British general election". In: *Electoral Studies* 41, pp. 234–238.

Friedman, Jerome H. (2001). "Greedy function approximation: A gradient boosting machine." In: *The Annals of Statistics* 29.5, pp. 1189–1232.

Gelman, Andrew et al. (2020). "Information, incentives, and goals in election forecasts". In: *Judgment and Decision Making* 15.5, pp. 863–880.

Geurts, Pierre, Damien Ernst, and Louis Wehenkel (2006). "Extremely randomized trees". In: *Machine Learning* 63.1, pp. 3–42.

Goot, Murray (2021). "How good are the polls? Australian election predictions, 1993–2019". In: *Australian Journal of Political Science* 56.1. Number: 1, pp. 35–55.

— (2022). *Over the last 30 years, a fifth of polls have called the wrong winner. Here are 3 things poll-watchers need to understand*. The Conversation URL: http://theconversation.com/over-the-last-30-years-a-fifth-of-polls-have-called-the-wrong-winner-here-are-3-things-poll-watchers-need-to-understand-182594.

— (2023). "Seat-by-seat polling versus the pendulum". In: *Watershed*. Ed. by Anika Gauja, Marian Sawer, and Jill Sheppard. 1st ed. The 2022 Australian Federal Election. ANU Press, 2023, pp. 383–412.

Graefe, Andreas (2019). "Accuracy of German federal election forecasts, 2013 & 2017". In: *International Journal of Forecasting*. Forecasting issues in developing economies 35.3, pp. 868–877.

Green, Anthony (2016). *Pendulum* URL: https://abc.net.au/news/elections/federal/2016/guide/pendulum.

— (2019). *Antony Green's swing calculator* URL: https://www.abc.net.au/news/2019-04-01/federal-election-2019-antony-green-house-of-reps-calculator/10872122.

— (2022). *Antony Green's election calculator - Federal election 2022* URL: https://abc.net.au/news/elections/federal/2022/guide/calculator.

Greenop-Roberts, Hamish (2022). "Forecasting Federal Elections: New Data From 2010–2019 and a Discussion of Alternative and Emerging Methods". In: *Australian Economic Review* 55.1. Number: 1, pp. 25–39.

Gschwend, Thomas (2017). *zweitstimme.org — Die Prognose zur Bundestagswahl 2017* URL: http://2017.zweitstimme.org/.

Gschwend, Thomas et al. (2022). "The Zweitstimme Model: A Dynamic Forecast of the 2021 German Federal Election". In: *PS: Political Science & Politics* 55.1. Publisher: Cambridge University Press, pp. 85–90.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY.

Haykin, Simon (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hummel, Patrick and David Rothschild (2014). "Fundamental models for forecasting elections at the state level". In: *Electoral Studies* 35, pp. 123–139.

International Monetary Fund (2024). *IMF Data* URL: `https://www.imf.org/en/Data`.

Jackman, Simon (2005). "Pooling the polls over an election campaign". In: *Australian Journal of Political Science* 40.4, pp. 499–517.

— (2014). "The Predictive Power of Uniform Swing". In: *PS: Political Science & Politics* 47.2, pp. 317–321.

Jackman, Simon and Gary N. Marks (1994). "Forecasting Australian elections: 1993, and all that". In: *Australian Journal of Political Science* 29.2, pp. 277–291.

James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. Springer, New York.

Jennings, Will, Michael Lewis-Beck, and Christopher Wlezien (2020). "Election forecasting: Too far out?" In: *International Journal of Forecasting* 36.3, pp. 949–962.

Kang, Seungwoo and Hee-Seok Oh (2023). "Forecasting South Korea's presidential election via multiparty dynamic Bayesian modeling". In: *International Journal of Forecasting*.

Kefford, Glenn (2021). *Political Parties and Campaigning in Australia: Data, Digital and Field*. Political Campaigning and Communication. Springer International Publishing, Cham.

Kennedy, Ryan, Stefan Wojcik, and David Lazer (2017). "Improving election prediction internationally". In: *Science* 355.6324. Publisher: American Association for the Advancement of Science, pp. 515–520.

Le Cessie, S. and J. C. Van Houwelingen (1992). "Ridge Estimators in Logistic Regression". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41.1, pp. 191–201.

Leigh, Andrew and Justin Wolfers (2006). "Competing approaches to forecasting elections: economic models, opinion polling and prediction markets". In: NBER Working Paper No. 12053.

Lewis-Beck, Michael S. (2005). "Election Forecasting: Principles and Practice". In: *The British Journal of Politics and International Relations* 7.2. Publisher: SAGE Publications, pp. 145–164.

Lewis-Beck, Michael S. and Ruth Dassonneville (2015). "Forecasting elections in Europe: Synthetic models". In: *Research & Politics* 2.1.

Linzer, Drew A. (2013). "Dynamic Bayesian Forecasting of Presidential Elections in the States". In: *Journal of the American Statistical Association* 108.501, pp. 124–134.

Mackerras, Malcolm (1976). "Uniform swing: Analysis of the 1975 election". In: *Politics* 11.1, pp. 41–46.

Magalhães, Pedro C., Luís Aguiar-Conraria, and Michael S. Lewis-Beck (2012). "Forecasting Spanish elections". In: *International Journal of Forecasting*. Special Section: Election Forecasting in Neglected Democracies 28.4, pp. 769–776.

Montalvo, José G., Omiros Papaspiliopoulos, and Timothée Stumpf-Fétizon (2019). "Bayesian forecasting of electoral outcomes with new parties' competition". In: *European Journal of Political Economy* 59, pp. 52–70.

Munzert, Simon (2017). "Forecasting elections at the constituency level: A correction–combination procedure". In: *International Journal of Forecasting* 33.2, pp. 467–481.

Murr, Andreas E., Mary Stegmaier, and Michael S. Lewis-Beck (2021). "Vote Expectations Versus Vote Intentions: Rival Forecasting Strategies". In: *British Journal of Political Science* 51.1, pp. 60–67.

Neunhoeffer, Marcel et al. (2020). "Ein Ansatz zur Vorhersage der Erststimmenanteile bei Bundestagswahlen". In: *Politische Vierteljahresschrift* 61.1, pp. 111–130.

Pack, Mark (2023). *How did MRP do in Australia?* The Week in Polls URL: `https://theweekinpolls.substack.com/p/how-did-mrp-do-in-australia`.

PhantomTrend (2016). *PhantomTrend* URL: `https://github.com/PhantomTrend/ptcode/blob/32db0c38d7d622364b8e385eb90bc390460ae6ec/PollingData/NationalData.csv`.

Rothschild, David (2015). "Combining forecasts for elections: Accurate, relevant, and timely". In: *International Journal of Forecasting* 31.3, pp. 952–964.

Rustika, Kevin (2021). *Successes and lessons from the YouGov MRP — YouGov* URL: `https://au.yougov.com/politics/articles/42933-successes-and-lessons-from-yougov-mrp`.

Singh, Amanpreet, Narina Thakur, and Aakanksha Sharma (2016). "A review of supervised machine learning algorithms". In: *Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development*, pp. 1310–1315.

Statistics, Australian Bureau of (2021). *TableBuilder — Australian Bureau of Statistics* URL: `https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder`.

Thai, Hoai-Thu et al. (2013). "A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models". In: *Pharmaceutical Statistics* 12.3, pp. 129–140.

Theis, S. E., A. Hense, and U. Damrath (2005). "Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach". In: *Meteorological Applications* 12.3, pp. 257–268.

Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Turgeon, Mathieu and Lucio Rennó (2012). "Forecasting Brazilian presidential elections: Solving the N problem". In: *International Journal of Forecasting*. Special Section: Election Forecasting in Neglected Democracies 28.4, pp. 804–812.

Umeda, Michio (2023). "Aggregating qualitative district-level campaign assessments to forecast election results: Evidence from Japan". In: *International Journal of Forecasting* 39.2, pp. 956–966.

Wezerek Gus, Jay Boice (2019). *How Good Are FiveThirtyEight Forecasts?* URL: `https://projects.fivethirtyeight.com/checking-our-work/`.

Wolfers, Justin and Andrew Leigh (2002). "Three Tools for Forecasting Federal Elections: Lessons from 2001". In: *Australian Journal of Political Science* 37.2, pp. 223–240.

Xie, Yuying et al. (2023). "An overview of deterministic and probabilistic forecasting methods of wind energy". In: *iScience* 26.1.

YouGov (2021). *YouGov's first Australian MRP Poll with the Australian Conservation Foundation — YouGov* URL: https://au.yougov.com/politics/articles/37812-yougov-first-australian-mrp-poll.

6 Conclusion

Declining response rates imperil the reliability of general population surveys around the world. The effort to address this issue is a broad undertaking, with many researchers approaching the problem in different ways. This dissertation represents a set of studies that each aim to contribute to this endeavor.

Up until the studies that comprise this dissertation, several gaps persisted in this area of research. Firstly, prior research had established that accounting for the time-series nature of panel data was helpful in improving propensity modeling, but the best practices to date required model engineering by trial and error. The time-series models presented in Chapter 2 demonstrated a tool for automatically learning temporal dependency which is then used to check the sufficiency of other, simpler models. Secondly, practitioners seeking to adopt an ML-based propensity model from past studies faced a common problem: they had no way to know what would apply in their own context, as every panel is different. Chapter 3 shows that tree-based models using demographic and past nonresponse behavior predictors are consistently effective across diverse panel designs. Therefore, survey researchers can adopt this technique with confidence. Thirdly, propensity modeling in panel studies faced the difficulty that these models are not effective until several waves of training data have accumulated, meanwhile participants are attriting. The pre-trained models demonstrated in Chapter 3 provide a solution. Fourthly, survey researchers implementing Adaptive Survey Design (ASD) had to rely on assumptions or limited field experiment reports to estimate how various options might affect their sample. The framework presented in Chapter 4 offers researchers a way to test several ASD options at once and gather strong evidence of how those strategies will affect sample composition and response rates. Finally, if nonresponse is to remain a persistent issue for the foreseeable future, Chapter 5 presents an example of how ML can be used to correct inference errors and allow researchers to continue making accurate predictions despite biased samples.

Taken together, this body of research can be summarized as follows. My coauthors and I developed tools to improve predictions of unit nonresponse in panel surveys, both by enhancing accuracy through more sophisticated handling of temporal dependencies and by enabling more timely forecasts via pre-training. We demonstrated that these ML-based techniques are highly generalizable, which should encourage broader adoption among survey researchers. Rather than applying response propensity models to ex-post weighting, as other researchers have done, we explore their use in an ex-ante solution to nonresponse bias: ASD. By empowering researchers to better control targeted, adaptive protocols, we support the achievement of more balanced samples. Finally, given that nonresponse bias will remain a challenge for the foreseeable future, we propose leveraging ML to correct sample-based predictions with a demonstrative application to electoral

forecasting.

What is the eventual goal of this research direction, and what are the next steps to approach it? In my view, assuming that the general population cannot be induced to respond to surveys more diligently, the next best outcome is this: samples are balanced enough, and our understanding of the factors that drive nonresponse is apt enough that we can ex-post adjust sample inferences effectively and consistently. Ideally, survey managers would be able to rapidly anticipate their sample balance, adjust protocols to induce typical nonrespondents to participate, and reach samples with a sufficient mix of respondents such that weighting can be sufficiently effective for whatever purpose the study aims for. ASD, as discussed in Chapter 4, is a core technique for this targeted protocol adjustment so as to induce more typical nonrespondents to participate. Additionally, we would ideally have the explanatory modeling capacity to connect the variables we collect with the causes of nonresponse and thereby weight by the right variables.

I believe that this dissertation has made some progress toward that goal by helping to mature the practice of ML modeling. Institutions adopt predictive models at various levels of maturity: first, one proves that a given model has a certain degree of merit in concept. Next, one typically demonstrates that the model is ready to be adopted in real conditions. This step typically involves simulating what would have happened if the model had been used and then assessing the various hypothetical outcomes for error, cost, ethical concerns, etc, with the goal of showing the safety and benefits of this change in practice. Finally, the model is adopted in practice and the performance and value of that model are constantly monitored and it is adjusted (or even decommissioned) as needed.

In this maturity framework, Chapters 2, 3 and 5 aimed to prove a concept for new modeling techniques: Time-series models in Chapter 2, pre-trained models in Chapter 3, and corrected pendulums in Chapter 5. Chapters 3 and 4 fit into the adoption-readiness step: Chapter 3 demonstrated the consistent success of tree-based propensity models. Chapter 4 provided a framework for establishing whether a prospective ASD strategy is fit to be deployed or not. Studies like these help survey managers address various concerns about a new technique and assist in establishing new practices.

The next step for each of these strands of research is to reach its next level of maturity. Long Short-Term Memory models could be tested in new panel surveys with the goal of assisting the researchers to discover the optimum way to handle temporal dependency. Pre-trained models could be applied in real early panel waves. The ASD options identified in Chapter 4 could be applied in practice to observe how closely the estimated outcomes match actual outcomes in terms of bias and response rates. This would be the next logical step in supporting the notion that this framework is suitable for broader adoption. Finally, the same ML models presented in Chapter 5 should be applied to the 2025 Australian election. Regardless of the outcome, that study would establish whether the efficacy of these retrospective predictions carries over to a prospective application. Taking these logical next steps in each strand of research will bring us closer to enhancing the control that survey researchers have over their samples and the accuracy of their inferences.