International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Predicting Australian federal electoral seats with machine learning[☆]

John 'Jack' Collins [*]

University of Mannheim, Germany

ARTICLE INFO

Article history:
Dataset link: https://osf.io/rvc2f/?view_only=a6f7c8360a3d4312ab6681dea0e1500e

Keywords:
Australian federal election
Machine learning
Australian politics
Electoral forecasting
Electoral prediction
Synthetic electoral modeling

ABSTRACT

I expand the international literature on election forecasting with the first application of machine learning (ML) in the Australian context. I apply these models to five elections from 2010 to 2022 and compare them with 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 predict 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.

1. Introduction

Australia's dominant federal election forecasting model is the Mackerras pendulum (Browne, 2022; Mackerras, 1976), 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 for each major party since the previous election (also known as the swing).[1] 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 that formed government in 17 out of the 20 elections held since 1972, miscalling only 1998, 2010, and 2019 (Goot, 2022; Green, 2016, 2019, 2022). However, the pendulum has several critical issues. First, seats held by minor parties (also called crossbenchers) are, by default, always predicted to stay with their incumbents. 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, 2022). Second, the pendulum relies heavily on accurate polling, which is why the 2019 election was miscalled, given a year of unusually large polling errors (Goot, 2021). Third, 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

☆ The numerical results presented in this manuscript were reproduced by the Editor-in-Chief on 1 February 2025.

* Correspondence to: 59 Halfey CCT, Wanniassa, ACT, 2903, Australia.

E-mail address: john.collins@mzes.uni-mannheim.de.

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.

https://doi.org/10.1016/j.ijforecast.2025.02.002

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 & Dassonneville, 2015). This approach combines historical polls and past electoral outcomes with economic and demographic predictors (also called fundamentals; for an overview, see Hummel & 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. First, for the benefit of international readers, I provide a summary of the Australian electoral system (Section 1.1). Next, I describe other attempts to improve Australian electoral forecasting and argue that there has yet to be any major improvement over the pendulum (Section 1.2). I 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 (Section 1.3). I close this section by listing the criteria against which I will evaluate the ML-based electoral models presented in this paper (Section 1.4).

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 represent roughly equally populated, geographically contiguous zones. Australia has a multiparty system, although it is 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 second preferences. This process repeats until a candidate reaches 50% of that seat's votes (Australian Electoral Commission, 2019b).

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, 2005; Jackman & Marks, 1994; Leigh & Wolfers, 2006; Wolfers & Leigh, 2002). Some of these models predicted the 2010 election-winning party successfully, 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 a 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, 2021b; Lewis-Beck & 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 with this study until the MRP is trialed in more elections.

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 & Oh, 2023; Linzer, 2013; Mackerras, 1976; Magalhães et al., 2012; Montalvo et al., 2019; Munzert, 2017; Nadeau & Lewis-Beck, 2020; Neunhoeffer et al., 2020; Theis et al., 2005; Turgeon & Rennó, 2012; Umeda, 2023). However, few models show consistent predictive accuracy. In the United States, forecasters predicted 100% of state races in the 2012 federal election, though with poorer performance in 2016 (Jackman, 2014; Kennedy et al., 2017; Linzer, 2013); (Wezerek & Gus, 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, 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 and 52% accurate for the elections from 1987 to 2015 (Murr et al., 2021). Synthetic models in the United Kingdom also had a highly variable constituency-level accuracy, between 57 and 100%, for the elections from

1987 to 2017 (Curtice & Firth, 2008; Fisher, 2016; Murr et al., 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 of Neunhoeffer et al. (2020) demonstrates the most consistently excellent results, highlighting the potential benefits of utilizing ML techniques over swing models.

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 of 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 will 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 et al., 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 (e.g., 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. Reliance on historical data adds new sources of potential error (Lewis-Beck, 2005). Therefore, for an ML model to be preferred over the pendulum, its enhanced performance should justify its greater complexity.

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 2.1), the ML and pendulum models, and the process for fitting, evaluating, and comparing these models (Section 2.2).

2.1. Data

ML models are trained on data from past elections to predict future ones. This section details the predictors and the dependent variable in this study. For a list of all predictors see Appendix Table 1.

2.1.1. Dependent variable: Seat winner party

Australia has a multiparty system, but the two major parties, the Australian Labor Party (ALP) and the Liberal–National Coalition Party (LNP), have won the vast majority of federal seats since 1949 (see Fig. 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 provide detailed data about election preferences before 2007 conveniently (Australian Electoral Commission, 2001, 2004, 2007, 2010, 2013, 2016, 2019a, 2022). So, because at least one election is required for training an ML model, I examine the five elections between 2010 and 2022.

2.1.2. Polling

Polling averages are inputs for ML models and pendulums (see Section 2.2.2). 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.

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 2 lists all data sources. Details such as nonresponse or undecided rates are often unavailable in these aggregate sources.

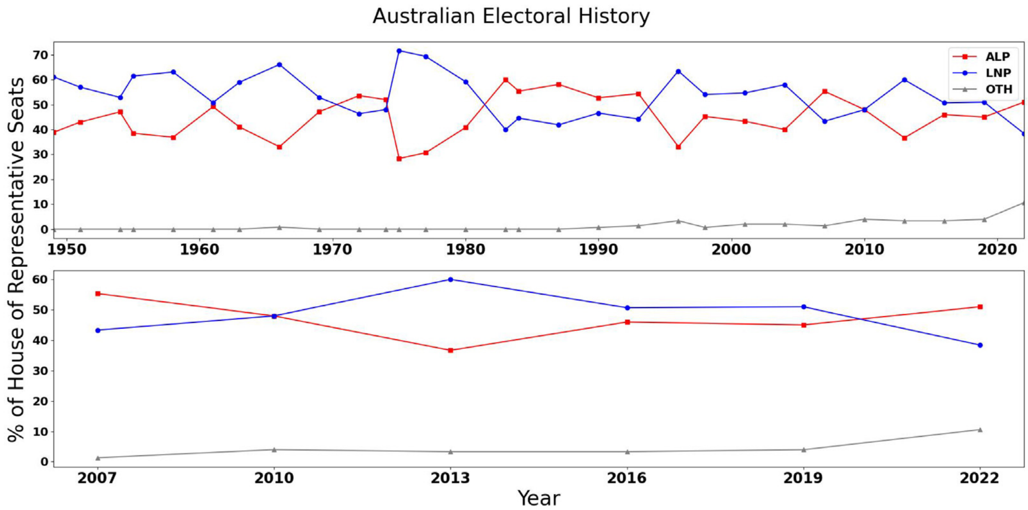

Fig. 1. Timeline of election outcomes. The top figure shows Australian federal elections since 1949. The bottom figure shows only the elections tested in this study.

In 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 postcodes (and thus their electoral seats), sample sizes are usually around 1000, randomly drawn from the entire population, resulting in small sample sizes for individual seats (Bowe, 2023); (PhantomTrend, 2016). Furthermore, aggregated poll results are consistently available, but not individual respondent data, making it impossible to disaggregate polls by seat (Bowe, 2023). Pollsters conduct swing-seat-specific surveys, but their results are not consistently available publicly (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, and 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 three and six months from election day as well as two 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 2.2). 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 1). Fig. 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 2.2 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, 2019, 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 pendula deviating from forecasts conducted by other analysts for the same election.

2.1.3. 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 2.2.2).

2.1.4. State effects

These predictors indicate the 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.

2.1.5. Pendulum predictions

In some feature sets (see Section 2.2.3), I use seat-level predictions from a TPP proportional swing model (see Section 2.2.2) as inputs for the ML models. The ML model alters the pendulum's predictions in these trials by accounting for other variables, such as socioeconomic or state effects.

2.1.6. Fundamentals

Fundamentals are predictors based on the proposition that voters punish incumbents for poor economic performance and are reluctant to change the government otherwise (for an overview, see Hummel & Rothschild, 2014). For this reason, I use macroeconomic variables and

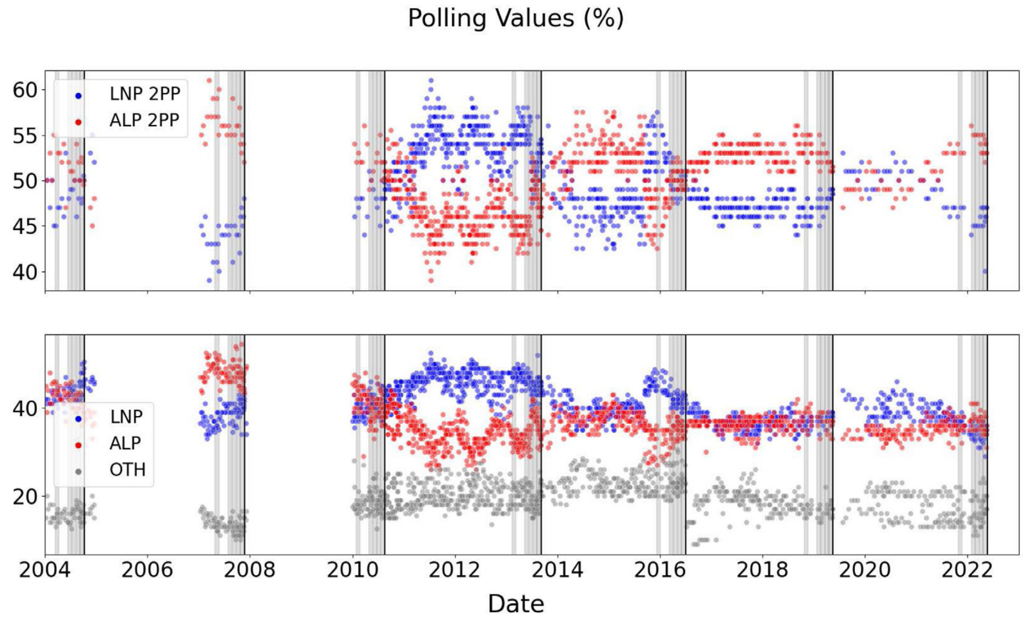

Fig. 2. Polling values and the four-week time windows (gray boxes) in which polling averages are calculated. The black lines indicate election days.

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 & Rothschild, 2014; Jennings et al., 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 (Australian Bureau of 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 2.1. I include median income, rent, mortgage payment, 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 retrospective inference time) to estimate how voters feel about the macroeconomy in the run-up to the election.

2.1.7. 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 2.2.).

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 et al. (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 2.2.1), 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 2.2.1 and 2.2.2). Next, I present the sets of features that I will vary to test the value of different types of predictors (Section 2.2.3). Then I 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 (Section 2.2.4). Additionally, I describe how I measure uncertainty in each model (Section 2.2.5). Finally, I explain how I will use permutation feature importance (PFI; Altmann et al., 2010) to explain how the most successful ML models make their predictions (Section 2.2.6).

2.2.1. 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 et al. (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 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 (Appendix Section 3.2.). 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 3.1. I discuss hyperparameters in Section 2.2.4.

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 2.2.3. 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).

2.2.2. Pendula

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 3.3. Table 1 provides an overview of each variation.

2.2.3. 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 et al., 2009). Instead, I test different sets of features for each ML model that reflect various forecasting strategies, as detailed in Table 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 in order 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 3.1). Appendix Table 1 provides the complete list of predictors and their respective feature sets.

2.2.4. Temporal cross-validation and hyperparameters

To ensure that my forecasts reflect out-of-sample estimates as they could have been made at the time, I use temporal cross-validation (TCV; Bergmeir & 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., 2023). 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 et al., 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 & 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 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 that in 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.

2.2.5. Uncertainty calibration

This paper will present the uncertainty around each estimated election outcome. Pendula, classifiers, and regressors handle forecast probabilities differently. For each model, I will provide the most likely forecast and the

Table 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 incumbents. 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 predict seats won by OTH candidates better. It calculates swings for major and minor parties using multiparty polls instead of TPP. If the swing against an incumbent party is enough for it to lose the seat, that seat is predicted to go to the party that received the second highest votes in the previous election.

Table 2
Summary of feature sets.

Feature set name	Description
Fundamentals	These are the socioeconomic variables described in Section 2.1.6, previous election vote shares, and state effects.
Polls and fundamentals	The same as the fundamentals set but includes polling averages described in Section 2.1.2.
Pendulum, polls, and fundamentals	In this feature set, I also include seat-level predictions from a TPP proportional pendulum (see Section 2.2.2). 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.

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 that 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.

2.2.6. Model interpretation

I aim to select the best ML model based on the evaluation criteria (Section 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 (e.g., TPP polls for ALP and LNP), withholding one does not lower the model's accuracy

Table 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.

Lead (Months)	Year	Model	All seats			Marginal seats		
			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

Notes: 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%. See Section 2.2.5 for details; refer to Appendix Section 4 for the complete table.

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 2.1. 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.

3. Results

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 4.1 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.

Fig. 3 shows each model's classification accuracy score (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 2.2.5). Table 3 presents a tabular summary of Fig. 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 2.2.5).

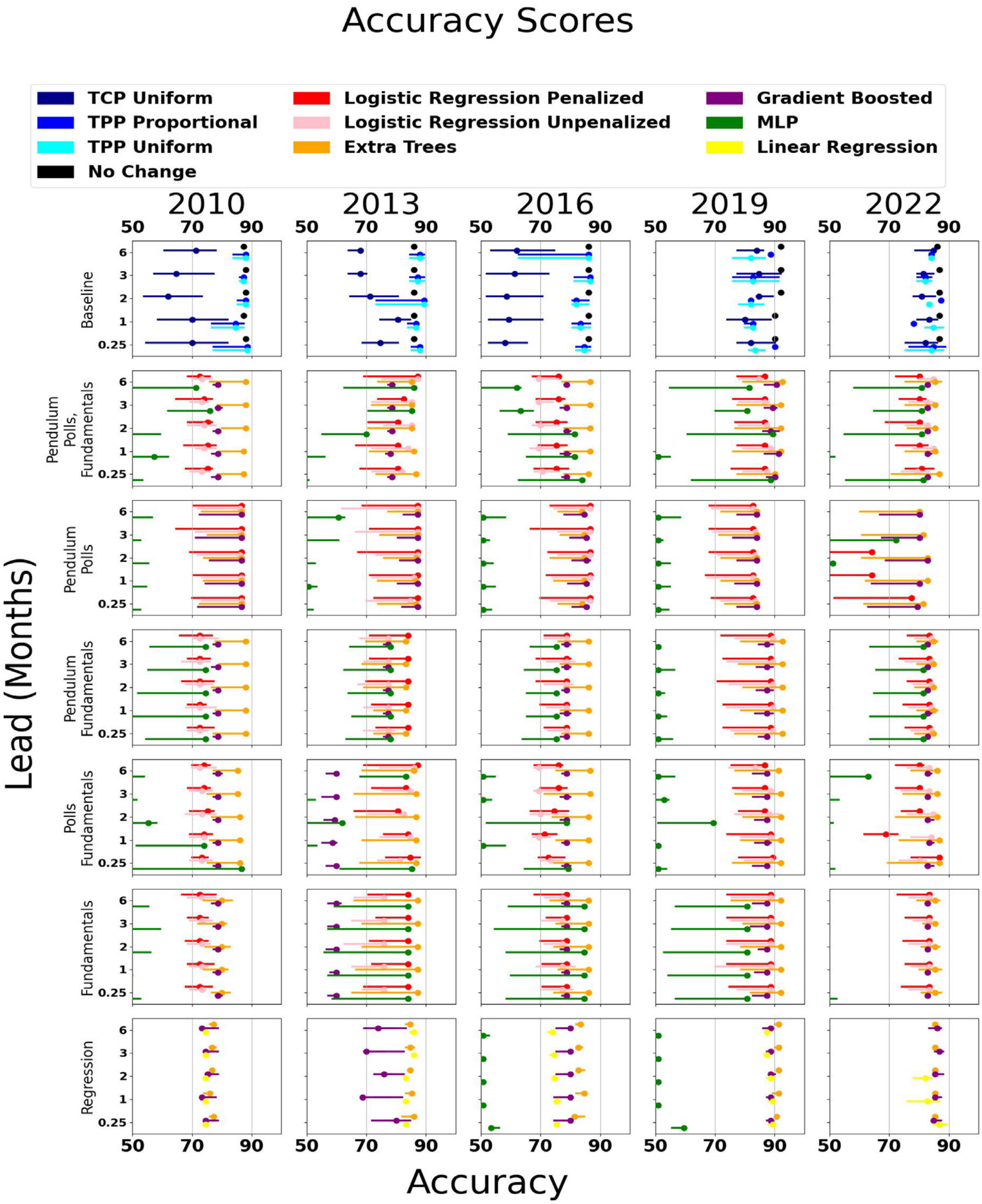

Fig. 3. Classification accuracy scores of each ML model and pendulum across every election and lead time. Note that any variation in the "no change" models is caused by newly redistricted seats.

In contrast, the most likely outcome, according to each of the pendulum models, is often roughly in the middle of its error bar. 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 2.2.5). For someone using a pendulum to predict an election, this would be like sensitivity-analyzing how the pendulum's forecasts changed as they tried 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. Also, as discussed, predicting cross-bench 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 3.2. Additionally, Appendix Section 4.3 presents the seat classification accuracies for marginal and crossbench seats. Table 3 presents the marginal seat classification accuracies for the best performing ML and pendulum models (discussed below).

Before discussing the ML models, I first determine which of the baseline models was the best in order to compare ML models against that technique. In the 2019 and 2022 election 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 Fig. 3). Tree-based models incorporating fundamental features outperform the pendula in 2019, achieving over 90% accuracy on average; however, they perform much worse in elections 2010–2016. Regression models perform better than the pendula 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 4.5.

For seat classification accuracy, I evaluate that ET with pendulum, polls, and fundamental features performed best overall. This model is consistently roughly as good as, or slightly better than the best pendulum at any election at one month before election day. This model is also as good as or slightly better than the best pendulum in any election at both six and three months out, except for 2013 (see Table 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.

Fig. 4 presents the forecasts of seat shares for the ALP, with equivalent figures for LNP and OTH in Appendix Section 4.4. 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 in every election from three months lead time. Table 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 pendula 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 a minority government, which is much closer to the actual outcome (a slim two-seat majority for LNP) than the proportional pendulum (which predicted 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 does 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.

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 (Fig. 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 Fig. 3). I chose the TPP proportional model for its higher accuracy than the other pendulum models at this lead time (Appendix Section 4.7 presents this analysis using the uniform pendulum instead).

Fig. 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 4.6. 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 changes in 2019, when the pendulum

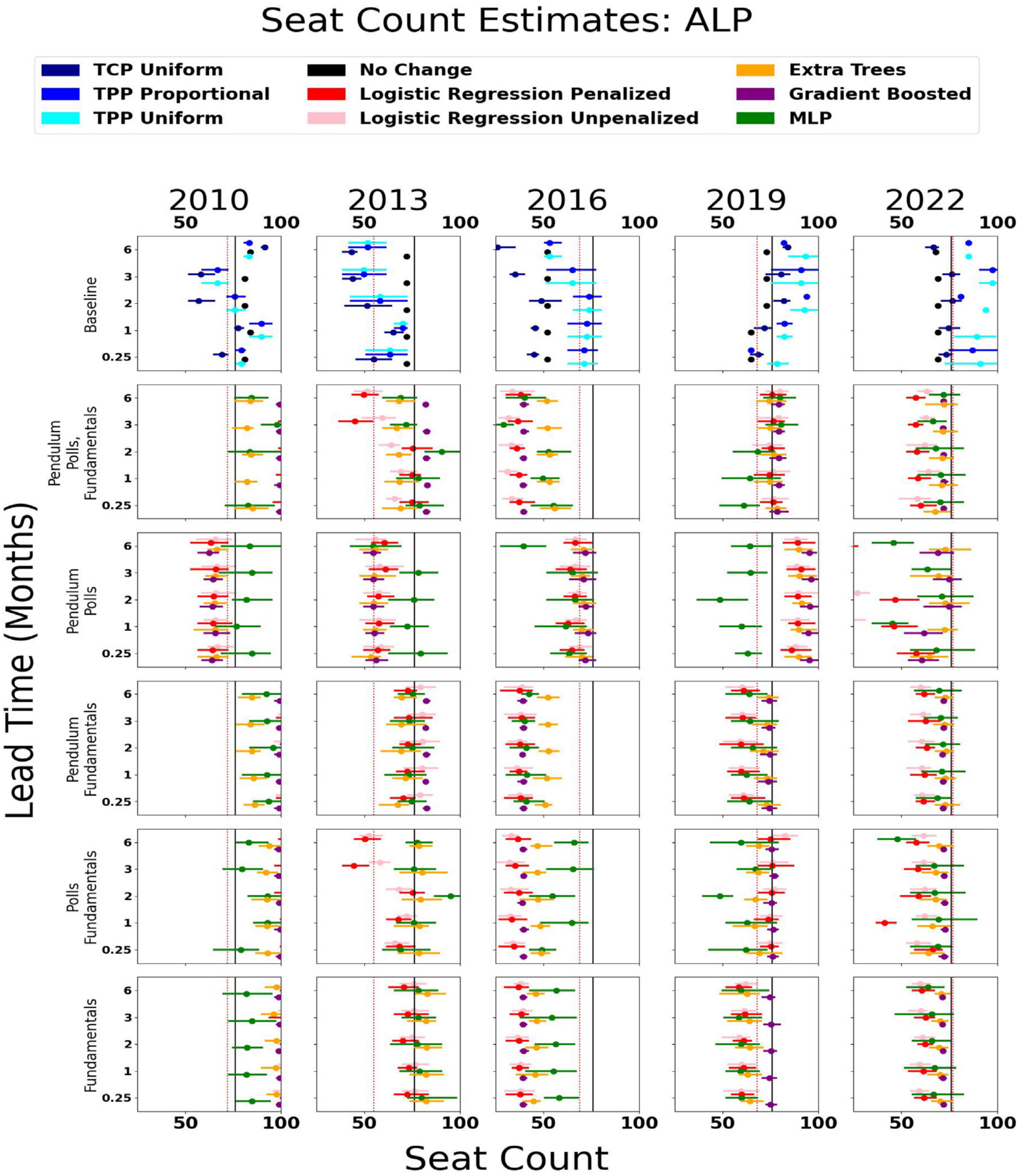

Fig. 4. The count of seats for the ALP as predicted by each model. The dotted line indicates the actual seat count for that election, and the solid line represents the 76 seats required to form a majority government.

misclassifies a substantial 26 seats, whereas the ET model misclassifies only 11. The ET continues to outperform the pendulum in 2022, when it miscalls 22 seats, compared with the pendulum's 27. Looking at the composition of misclassifications, it seems the ET model favors the incumbent more than the pendulum. Hence, ET gives 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 the 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 election day 2022 (Fig. 4). In tight elections such as 2010, 2016, 2019,

Table 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.

Lead (Months)	Year	Feature set	Model	Mean error	Lower bound (95% CI)	Upper bound (95% CI)
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

Notes: 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. See Section 2.2.5 for details; refer to Appendix Section 4 for the complete table.

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.

Fig. 6 shows the permutation feature importance (PFI) for each feature block in the best-performing ET model. The model relies primarily 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, Fig. 3 shows that including polls improves predictions slightly in some cases (e.g., six months before the 2013 election), indicating some value in incorporating polls.

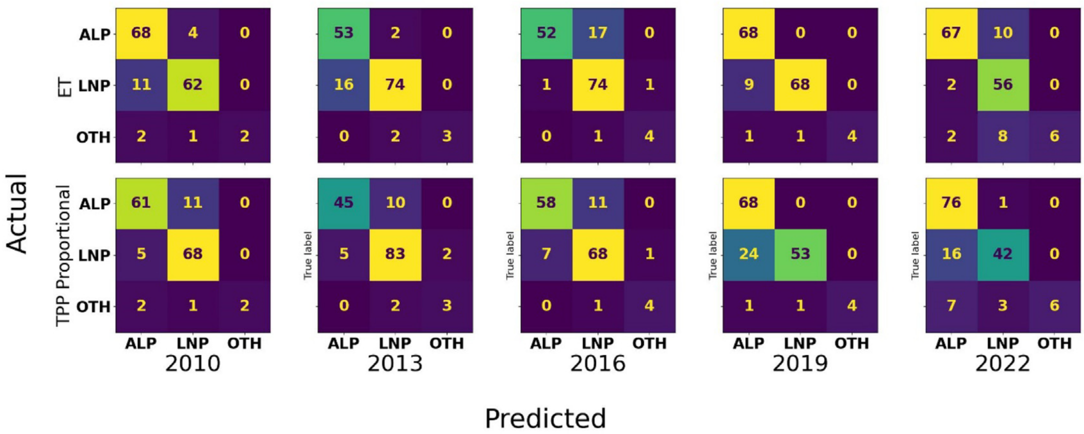

Fig. 5. Confusion matrices for the best model in this study (ET with all predictors) versus the best pendulum (TPP proportional). Each model's most likely predictions were made three months before election day. For example, the top leftmost number (68) indicates that in 2010, the ET model predicted that 68 seats would go to the ALP, and they 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.

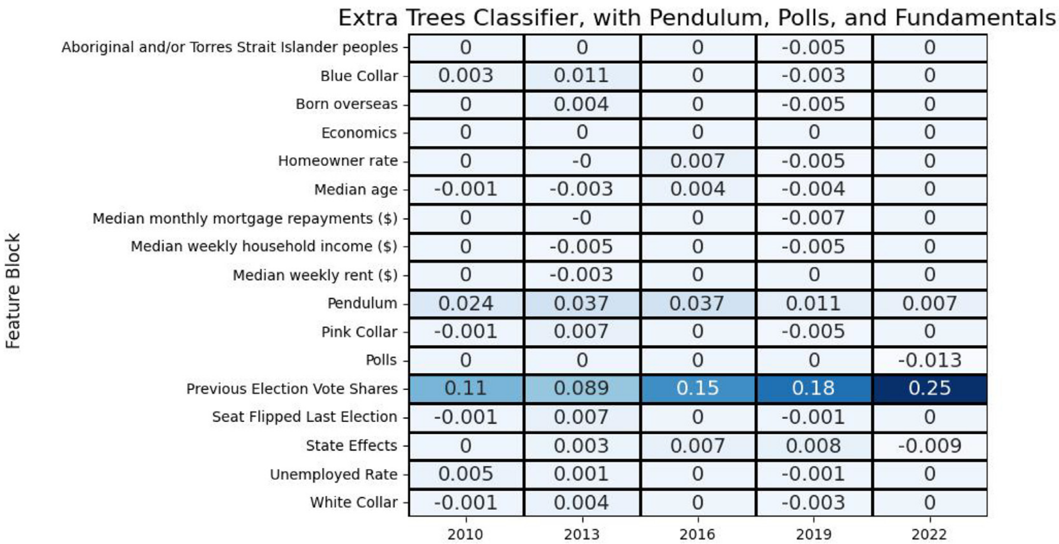

Fig. 6. The PFI of each block of predictors for the best-performing model (ET with pendulum, polls, and fundamental predictors). Higher positive values indicate that including the feature block improves accuracy; negative values indicate that withholding it improves accuracy.

3.3. Evaluation

Table 5 compares the ET model with 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 (from 2019 to 2022) with a three-month lead time. However, seat count error was often higher than for 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.

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. 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 application in other democracies.

Table 5
Comparing the best discovered ML model with the evaluation criteria.

Criteria	Extra trees with pendulum, polls, and fundamentals
Seat count accuracy (see Fig. 4, Table 4, and Appendix Section 4.4).	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 Fig. 3, Table 3, and Appendix Section 4.3).	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 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 (Fig. 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 (Fig. 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 (Figs. 3 and 4).

This paper trialed a wide range of algorithms and types of predictors and revealed two notable models. First, 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 5 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. Second, a gradient-boosted classifier with pendulum and polling predictors forecast 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 those of seat contests. Even an improvement of a few seats could help better anticipate the outcomes of tight elections such as 2010, 2016, 2019, and 2022.

Further improvements can be made in 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 excelled at both tasks. This might be achieved by ensembling the two techniques. Finally, tree-based models are inherently more complex 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 simulate closely how the model would have forecast 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 these data are rarely made available across multiple elections and lead times. Novel data sources could be explored, especially if they helped predict swing seats, which are typically the hardest to forecast. Marginal-seat-specific polls often fail to predict swing seats accurately (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.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used 'Grammarly' in order to perform some basic copy editing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: John Collins reports financial support was provided by University of Mannheim. John Collins reports a relationship with University of Mannheim that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment of funding sources

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany, research project "Prediction-based Adaptive Designs for Panel Surveys" (KE 2185/3-1).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2025.02.002.

Data and code availability

The data and code are available as a reproducibility package available at https://osf.io/rvc2f/?view_only=a6f7c8360a3d4312ab6681dea0e1500e.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. http://dx.doi.org/10.1093/bioinformatics/btq134.

Argandoña-Mamani, A., Ormeño-Alarcón, T., Iparraguirre-Villanueva, O., Paulino-Moreno, C., & Cabanillas-Carbonell, M. (2024). In X.-S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of eighth international congress on information and communication technology*, *Predicting election results with machine learning—a review* (pp. 989–1001). Springer Nature, http://dx.doi.org/10.1007/978-981-99-3043-2_82.

Arnold, C., Biedebach, L., Küpfer, A., & Neunhoeffer, M. (2023). *The role of hyperparameters in machine learning models and how to tune them.*

Australian Bureau of Statistics (2021). *TableBuilder — Australian bureau of statistics*. https://www.abs.gov.au/statistics/microdata-tablebuilder/tablebuilder.

Australian Electoral Commission (2001). *AEC: when: past electoral events*. https://results.aec.gov.au/10822/Website/index.html.

Australian Electoral Commission (2004). *Election 2004: the official election report and results*. https://results.aec.gov.au/12246/default.htm.

Australian Electoral Commission (2007). *2007 federal election results*. https://results.aec.gov.au/13745/.

Australian Electoral Commission (2010). *Australian electoral commission virtual tally room*. https://results.aec.gov.au/15508/Website/Default.htm.

Australian Electoral Commission (2011). *Compulsory voting in Australia. Australian electoral commission*. https://www.aec.gov.au/About_AEC/publications/voting/.

Australian Electoral Commission (2013). *Australian electoral commission virtual tally room*. https://results.aec.gov.au/17496/Website/Default.htm.

Australian Electoral Commission (2016). *2016 federal election. Australian electoral commission*. https://results.aec.gov.au/20499/Website/HouseDefault-20499.htm.

Australian Electoral Commission (2019a). *2019 federal election. Australian electoral commission*. https://results.aec.gov.au/24310/Website/HouseDefault-24310.htm.

Australian Electoral Commission (2019b). *AEC fact-sheet. Australian electoral commission*. https://www.aec.gov.au/Elections/federal_elections/2022/files/downloads/fact-sheet-national-seat-status-2022-federal-election.pdf.

Australian Electoral Commission (2022). *2022 federal election. Australian electoral commission*. https://results.aec.gov.au/27966/Website/HouseDefault-27966.htm.

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. http://dx.doi.org/10.1016/j.ins.2011.12.028.

Bowe, W. (2021a). *YouGov MRP poll (part one) and more – the poll bludger*. https://www.pollbludger.net/2022/05/11/yougov-mrp-poll-part-one-and-more/.

Bowe, W. (2021b). *YouGov MRP poll (part two): labor 80 seats, coalition 63, others 8 – the poll bludger*. https://www.pollbludger.net/2022/05/12/yougov-mrp-poll-part-two-labor-80-seats-coalition-63-others-8/.

Bowe, W. (2023). *The poll bludger – analysis and discussion of elections and opinion polls in Australia*. https://www.pollbludger.net/.

Browne, B. (2022). *Between sense and nonesense: the predictive power of the electoral pendulum*. https://apo.org.au/node/317924.

Curtice, J., & Firth, D. (2008). Exit polling in a cold climate: The BBC–ITV experience in Britain in 2005. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(3), 509–539. http://dx.doi.org/10.1111/j.1467-985X.2007.00536.x.

Evershed, N., & Nicholas, J. (2022). How we used machine learning to cover the Australian election. *The Guardian*, https://www.theguardian.com/info/2022/jul/15/how-we-used-machine-learning-to-cover-the-australian-election.

Fisher, S. D. (2016). Piecing it all together and forecasting who governs: The 2015 British general election. *Electoral Studies*, *41*, 234–238. http://dx.doi.org/10.1016/j.electstud.2015.11.014.

Gelman, A., Hullman, J., Wlezien, C., & Elliott Morris, G. (2020). Information, incentives, and goals in election forecasts. *Judgment and Decision Making*, *15*(5), 863–880. http://dx.doi.org/10.1017/S1930297500007981.

Goot, M. (2021). How good are the polls? Australian election predictions, 1993–2019. *Australian Journal of Political Science*, *56*(1), Article 1. http://dx.doi.org/10.1080/10361146.2020.1825616.

Goot, M. (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*, 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.

Goot, M. (2023). Seat-by-seat polling versus the pendulum. In A. Gauja, M. Sawer, & J. Sheppard (Eds.), *Watershed* (1st ed.). (pp. 383–412). ANU Press, https://www.jstor.org/stable/jj.12639070.24.

Graefe, A. (2019). Accuracy of German federal election forecasts, 2013 & 2017. *International Journal of Forecasting*, *35*(3), 868–877. http://dx.doi.org/10.1016/j.ijforecast.2019.01.004.

Green, A. (2016). *Pendulum*. https://abc.net.au/news/elections/federal/2016/guide/pendulum.

Green, A. (2019). *Pendulum*. https://abc.net.au/news/elections/federal/2019/guide/pendulum.

Green, A. (2022). *Antony Green's election calculator—federal election 2022*. https://abc.net.au/news/elections/federal/2022/guide/calculator.

Greenop-Roberts, H. (2022). Forecasting federal elections: New data from 2010–2019 and a discussion of alternative and emerging methods. *Australian Economic Review*, *55*(1), 1. http://dx.doi.org/10.1111/1467-8462.12450.

Gschwend, T. (2017). *Zweitstimme.org — die prognose zur bundestagswahl 2017*. http://2017.zweitstimme.org/.

Gschwend, T., Müller, K., Munzert, S., Neunhoeffer, M., & Stoetzer, L. F. (2022). The zweitstimme model: A dynamic forecast of the 2021 German federal election. *PS: Political Science & Politics*, *55*(1), 85–90. http://dx.doi.org/10.1017/S1049096521000913.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer, http://dx.doi.org/10.1007/978-0-387-84858-7.

Hummel, P., & Rothschild, D. (2014). Fundamental models for forecasting elections at the state level. *Electoral Studies*, *35*, 123–139. http://dx.doi.org/10.1016/j.electstud.2014.05.002.

Jackman, S. (2005). Pooling the polls over an election campaign. *Australian Journal of Political Science*, *40*(4), 499–517. http://dx.doi.org/10.1080/10361140500302472.

Jackman, S. (2014). The predictive power of uniform swing. *PS: Political Science & Politics*, *47*(2), 317–321. http://dx.doi.org/10.1017/S1049096514000109.

Jackman, S., & Marks, G. N. (1994). Forecasting Australian elections: 1993, and all that. *Australian Journal of Political Science*, *29*(2), 2. http://dx.doi.org/10.1080/00323269408402294.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (Vol. 103)*. Springer, http://dx.doi.org/10.1007/978-1-4614-7138-7.

Jennings, W., Lewis-Beck, M., & Wlezien, C. (2020). Election forecasting: Too far out? *International Journal of Forecasting*, *36*(3), 949–962. http://dx.doi.org/10.1016/j.ijforecast.2019.12.002.

Kang, S., & Oh, H.-S. (2023). Forecasting South Korea's presidential election via multiparty dynamic Bayesian modeling. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2023.01.004.

Kefford, G. (2021). *Political parties and campaigning in Australia: data, digital and field*. Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-68234-7.

Kennedy, R., Wojcik, S., & Lazer, D. (2017). Improving election prediction internationally. *Science*, *355*(6324), 515–520. http://dx.doi.org/10.1126/science.aal2887.

Leigh, A., & Wolfers, J. (2006). *Competing approaches to forecasting elections: economic models, opinion polling and prediction markets: IZA Discussion Papers, 1972.*

Lewis-Beck, M. S. (2005). Election forecasting: Principles and practice. *The British Journal of Politics and International Relations*, *7*(2), 145–164. http://dx.doi.org/10.1111/j.1467-856X.2005.00178.x.

Lewis-Beck, M. S., & Dassonneville, R. (2015). Forecasting elections in Europe: Synthetic models. *Research & Politics*, *2*(1), Article 2053168014565128. http://dx.doi.org/10.1177/2053168014565128.

Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108*(501), 124–134. http://dx.doi.org/10.1080/01621459.2012.737735.

Mackerras, M. (1976). Uniform swing: Analysis of the 1975 election. *Politics*, *11*(1), 41–46. http://dx.doi.org/10.1080/00323267608401542.

Magalhães, P. C., Aguiar-Conraria, L., & Lewis-Beck, M. S. (2012). Forecasting Spanish elections. *International Journal of Forecasting*, *28*(4), 769–776. http://dx.doi.org/10.1016/j.ijforecast.2012.04.007.

Montalvo, J. G., Papaspiliopoulos, O., & Stumpf-Fétizon, T. (2019). Bayesian forecasting of electoral outcomes with new parties' competition. *European Journal of Political Economy*, *59*, 52–70. http://dx.doi.org/10.1016/j.ejpoleco.2019.01.006.

Munzert, S. (2017). Forecasting elections at the constituency level: A correction–combination procedure. *International Journal of Forecasting*, *33*(2), 467–481. http://dx.doi.org/10.1016/j.ijforecast.2016.12.001.

Murr, A. E., Stegmaier, M., & Lewis-Beck, M. S. (2021). Vote expectations versus vote intentions: Rival forecasting strategies. *British Journal of Political Science*, *51*(1), 60–67. http://dx.doi.org/10.1017/S0007123419000061.

Nadeau, R., & Lewis-Beck, M. S. (2020). Election forecasts: Cracking the Danish case. *International Journal of Forecasting*, *36*(3), 892–898. http://dx.doi.org/10.1016/j.ijforecast.2019.09.007.

Neunhoeffer, M., Gschwend, T., Munzert, S., & Stoetzer, L. F. (2020). Ein ansatz zur vorhersage der erststimmenanteile bei bundestagswahlen. *Politische Vierteljahresschrift*, *61*(1), 111–130. http://dx.doi.org/10.1007/s11615-019-00216-3.

Pack, M. (2023). How did MRP do in Australia? [Substack newsletter]. *The Week in Polls*, https://theweekinpolls.substack.com/p/how-did-mrp-do-in-australia.

PhantomTrend (2016). *PhantomTrend/ptcode [R]*. (Original work published 2014). https://github.com/PhantomTrend/ptcode/blob/32db0c38d7d622364b8e385eb90bc390460ae6ec/PollingData/NationalData.csv.

Reed, J. (2022). *The reed pendulum*. https://resolvestrategic.com/__static/943e394a485a5a962f38548cbc2edacc/the-reed-pendulum-post-2022-updated-with-byelections.pdf?dl=1.

Rothschild, D. (2015). Combining forecasts for elections: Accurate, relevant, and timely. *International Journal of Forecasting*, *31*(3), 952–964. http://dx.doi.org/10.1016/j.ijforecast.2014.08.006.

Rustika, K. (2021). *Successes and lessons from the YouGov MRP — YouGov*. https://au.yougov.com/politics/articles/42933-successes-and-lessons-from-yougov-mrp.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development* (pp. 1310–1315). https://ieeexplore.ieee.org/abstract/document/7724478.

Thai, H.-T., Mentré, F., Holford, N. H. G., Veyrat-Follet, C., & Comets, E. (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical Statistics*, *12*(3), 129–140. http://dx.doi.org/10.1002/pst.1561.

Theis, S. E., Hense, A., & Damrath, U. (2005). Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteorological Applications*, *12*(3), 257–268. http://dx.doi.org/10.1017/S1350482705001763.

Turgeon, M., & Rennó, L. (2012). Forecasting Brazilian presidential elections: Solving the N problem. *International Journal of Forecasting*, *28*(4), 804–812. http://dx.doi.org/10.1016/j.ijforecast.2012.04.003.

Umeda, M. (2023). Aggregating qualitative district-level campaign assessments to forecast election results: Evidence from Japan. *International Journal of Forecasting*, *39*(2), 956–966. http://dx.doi.org/10.1016/j.ijforecast.2022.03.006.

Wezerek, J. B., & Gus (2019). *How good are FiveThirtyEight forecasts?* FiveThirtyEight. https://projects.fivethirtyeight.com/checking-our-work/.

Wolfers, J., & Leigh, A. (2002). Three tools for forecasting federal elections: Lessons from 2001. *Australian Journal of Political Science*, *37*(2), 2. http://dx.doi.org/10.1080/10361140220148115.

Xie, Y., Li, C., Li, M., Liu, F., & Taukenova, M. (2023). An overview of deterministic and probabilistic forecasting methods of wind energy. *IScience*, *26*(1), Article 105804. http://dx.doi.org/10.1016/j.isci.2022.105804.

YouGov (2021). *YouGov's first Australian MRP poll with the Australian conservation foundation — YouGov*. https://au.yougov.com/politics/articles/37812-yougov-first-australian-mrp-poll.