Predicting election outcomes based on demography

Group 42 - Nicholas Wood (46042210), Michael Yee (46430261), Alex Valenzuela (46042229)

Summary

All Australians of voting age are periodically required to vote in elections to decide which political candidate they want to represent them as a Member of Parliament in the federal House of Representatives. An electorate is the geographic region that an elected Member of Parliament will represent and typically encompasses multiple suburbs or regional areas, and each Australian voter must only vote within the electorate where they reside.

We theorise that it is possible to construct a model based on the various demographics within each electorate to predict the outcome of a federal election in those electorates. We will utilise census data from the Australian Bureau of Statistics (ABS) to build the model and federal election data from the Australian Electoral Commission (AEC) to train and test the model.

Goals

Our primary goal is to develop a model that can accurately predict the outcome of a vote held in a particular electorate based on demographic data for that electorate. Examples of demographic characteristics we will examine include age distribution, sex, income levels, job types, and educational attainment among others.

For our secondary goals, we will aim to identify which demographic attributes are most strongly correlated with political affiliation with the major and minor political parties/independent representatives. We also aim to develop demographic profiles that are characteristic of different types of electorates (safe, leaning, marginal and non-traditional party seats).

Finally we will build an electoral map illustrating the predicted electoral outcome according to our model compared to the actual 2019 Electoral Map. This will be used to assess the overall accuracy of our model in predicting electoral outcomes.

Our Datasets

The first source is from the 2016 Australian census by the Australian Bureau of Statistics (ABS), specifically the datasets are by “place of usual residence” against a range of demographic data specially curated in Census TableBuilder, data is in csv format. The other source is from the 2019 Australian Federal election by the Australian Electoral Commission (AEC), specifically we use datasets based on preferences and two-party preferred by CED/division, data is in csv format.

Initial analysis shows that both electoral data and census data almost perfectly link with each other, and both datasets provide data by Commonwealth Electoral District (CED). The imperfection in the linkage of the datasets is due to some changes to CED’s between 2016-18: 2 new CED’s were introduced, 1 was dissolved & 6 were renamed as well as some CED boundaries being redrawn. Our data cleaning process will rename CED’s accordingly, whilst for the CED’s that are new, dissolved or readjusted we will extract census data by suburb and recombine in line with the correct CED boundary.

Techniques to be used

Regression (Simple Linear & Logistic): Used to predict voting characteristic **(y)** e.g. strong Labor win, marginal Labor win given the main demographic variables **(x's)**: sex, job, religion (Logistic); income & age (Simple Linear).

K-means clustering: Used to determine the main groups of electorates (e.g. strong Labor win, marginal Labor win, plus any hidden groups) by comparing winning vote percentage to age, male to female ratio, etc.

Other available techniques that are yet to be learnt in the course such as k-nearest neighbour & Naive Bayes classifiers will be used if deemed appropriate.

Project Plan

Milestone 1 (Set up) - by week 8

- Finalise and submit project proposal
- Acquire, clean, merge and format all relevant datasets into a single master dataframe
- Set up Github Repository and Jupyter notebook

Milestone 2 (Model Building) - by week 11

- Use initial and subsequent findings/analysis to set up an initial working model
- Assess model's efficacy based on performance against 2019 election data and adjust accordingly

Milestone 3 (Finalisation) - by week 13

- Collate findings for use in the final build of our model
- Use model/analysis to discuss the relationship between certain demographic characteristics and political affiliation for use in presentation.
- Identify drawbacks & limitations of the model for discussion in presentation
- Finalise presentation

Relevant Prior Work

In 2013, Stimson and Shyy examined the demographic determinants of election outcomes for the 2007 federal elections using 2006 census data[1]. They found that socio-economic status, age, multiculturalism, and residence ownership were the primary determinants of voter support for this election. However, they also identified changes over time between the 2001, 2004, and 2007 federal elections in how these determinants correlated with support for each major political party. Therefore, our model may utilise different determinants for predicting the 2019 federal election outcome.

[1] Stimson, RJ & Shyy, T-K 2013, 'And now for something different: modelling socio-political landscapes' The Annals of Regional Science, vol. 50, no. 2, pp. 623–643, doi: 10.1007/s00168-012-0505-5.