MATHS 1006: Data Taming & Prediction

Final Report Instructions

Semester 2 2024

Due date: 11:59 pm, Friday November 1st (week 13)

Scenario

You are employed as a data scientist at Policy and Opinion, an organisation that works to better understand the public opinions of the public in various countries. The organisation has been increasingly interested in the recent attempts to predict how people will vote in US elections. One way this is achieved is by using large surveys or opinion polls where respondents are asked before the election whether they intend to vote and if they were going to vote, who they would vote for. After the election, some people are followed up to see who they decided to vote for.

You have been provided with a very large dataset based on the 2016 Cooperative Electoral Survey (Ansolabehere & Schaffner, 2017). We have chosen the 2016 year of this survey because it was an interesting year for prediction (many polls predicted a different outcome than what occurred). It might seem funny to attempt to fit a prediction model for an event that has already occurred, but this can be helpful to assess the quality of prediction models. After the election, some people are followed up to see who they decided to vote for, which adds an extra layer for predictive analysis.

Predictive models are used in real world polling to make predictions for a particular area will vote, but this is a bit too advanced for this class, so in this report we are just going to focus on predictions for individuals like those that are in our survey. If you're interested in this bigger process, Lauren Kennedy will be giving a talk in November for the Adelaide Data Science Centre.

This dataset has been modified for the purposes from the original survey responses for the purposes of assessment, and the models you fit are the *first step* of working with population estimation and election prediction.

To complete this project you should make sure you do the following:

Step 1: Cleaning and preparing

- Clean the data. Remember to justify any decisions you make, and if you are not sure about a particular decision, you can discuss with the data stakeholder (Tayla, or your tutor) to consider the implications.
- Set the seed as your student id number (without the "a")
- Take a sample of 3000 observations from the bigger dataset; 1500 democrat voters and 1500 republican voters

Step 2: Exploratory data analysis

Before making predictions, the company are interested in exploring some of the patterns in the data. Specifically, they would like to visualise:

- How do 2016 votes vary for those with a student loan and those without?
- What is the relationship between 2016 vote and education attainment, and is this influenced by student loan status?

- For each income category, what proportion of 2016 voters cast the same vote as they did in 2012?

You might find other visualisations and tables useful in both cleaning and understanding the data, but you do not need to include every plot you made in your report!

Step 3: Fit models

Split your dataset into training and test sets. Preprocess both datasets using tidymodels.

Your main aim is to build a model to predict vote outcome (democrat/republican) for an individual based on appropriate variables in the provided dataset. This may be all variables in the dataset or a subset of variables if not all are useful. After consultation with an expert statistician, the company have decided they would like you to compare the following three models:

- A logistic regression;
- A K-nearest neighbours model with a range of 1 to 100 and 20 levels; and
- A random forest with 100 trees and 5 levels.

Step 4: Compare models

You are required to determine and explain which of the above is the best model for predicting vote outcome. You may want to consider metrics such as accuracy, AUC, sensitivity, and specificity. It is important to provide an explanation to the founders about how you compared the models and why you made the choices you did, to fully justify your decision. For your chosen model, identify the most important variables.

What to submit:

After completing the above steps, you will need to prepare a reproducible report. A succinct but reproducible statistical appendix (like you saw in assignment 4) should be included that contains your full analyses (steps 1-4). You should also write a reproducible report that contains the following sections. You may include figures and tables in your report as well as your appendix, but you should also ensure that those you include are chosen carefully and selectively. We advise that you complete steps 1-4 and write the appendix before writing the report. As a guide, your main report should be approximately 6 pages and your statistical appendix (which includes code, output and tables) should be approximately 10-15 pages.

Formatting

Provide a report in the following format:

- An executive summary, with key results outlined in plain English;
- A methods section, outlining what data was analysed, steps that were taken, and stating the software used for your analysis;
- A results section, including an exploratory data analysis, and interpreting and evaluating your models in language the founders will understand;
- A discussion section, in which you discuss your models' outcomes and predictions, with specific relevance to the founders' objectives;
- A conclusion, in which you summarise your findings and recommendations to the founders, written in plain English; and
- A statistical aappendix, including all R code you used to perform your analysis, output where required to show how the processes you implement are working as well as any technical material beyond what you might expect the readers to understand.

Some more rules about your report:

- **You must complete this assignment using R Markdown**;

- Your report must be submitted as **pdf only** on MyUni;
- You must include **units** when providing results;
- Include any working when providing solutions;
- Provide all numerical answers to **3 decimal places**;
- Make sure you include both your code and R output / plots in your appendix;
- Make sure any tables or plots included have informative captions;
- You can submit more than once if you find errors and your latest submission will be marked;
- Make sure you only upload one document for your final submission. If you submit multiple pages (i.e. one per question) you will be deducted 10% per page submitted;
- No marks deducted for projects submitted within 24 hours of the deadline. After 24 hours, the project is not marked and you get zero; and
- Finally, make sure you check your submitted assignment is the correct one, as we cannot accept other submissions after the due date.

Explanation of the variables

Variable	Description
`ballot_ID`	Unique ballot identifying number
`state`	State of residence
`county`	County of residence
`voted_2016`	Who did you vote for in the 2016 presidential election? (democrat/republican)
`student_loan`	Do you have a student education loan? (yes/no)
`child_u18`	Do you have a child under the age of 18? (yes/no)
`followed_event_fb`	Have you followed a political event on FB? (yes/no)
`education`	Highest education attainment (not high school/high school/ undergraduate/postgraduate)
`employment`	Current employment status (employed/unemployed/retired/student/other)
`social_media`	In the past 24 hours, have you used social media? (Yes/No)
`marital_status`	Survey participant's marital status (Divorced/Domestic Partnership/Married/Separated/Single/Widowed)
`home_owner_status`	Survey participant's home owner status (own/rent/other)
`voted_2012`	Who did you vote for in the 2012 election? (democrat/republican)
`birth_year`	What is your year of birth?
`vote_primary`	Did you vote in the 2016 primary election (yes/no)
`vote_method`	How did you vote? (in person/by mail)
`health_insurance`	Do you hold insurance through a government program such as medicare or medicaid? (yes/no)
`faminc`	Family income (see description below)

faminc variable description

level	description
1	Less than $10,000
2	$10,000 - $19,999
3	$20,000 - $29,999
4	$30,000 - $39,999
5	$40,000 - $49,999
6	$50,000 - $59,999
7	$60,000 - $69,999
8	$70,000 - $79,999
9	$80,000 - $99,999
10	$100,000 - $119,999
11	$120,000 - $149,999
12	$150,000 or more
31	Prefer not to say
97	$150,000 - $199,999
13	$200,000 - $249,999
14	$250,000 - $349,999
15	$350,000 - $499,999
16	$500,000 or more
98	Skipped
99	Not asked
32	$250,000 or more

Reference

Ansolabehere, S. and Schaffner, B.F., 2017. CCES Common Content, 2016. Harvard Dataverse. Version V4. DOI: 10.7910/DVN/GDF6Z0. Available at: https://doi.org/10.7910/DVN/GDF6Z0