DATS 6103 Project Proposal

Authors: Divya Parmar, Kushal Ismael, Bradley Reardon

The problem we selected for this project is voter turnout in American elections. There was a lot of speculation around voter turnout in 2020 with both high-profile candidates and new voting methods due to the pandemic. Furthermore, there is a long-standing interest in why many eligible voters decide not to vote. Turnout has been around 55-60 percent in most modern presidential elections, but rose to 67 percent in the highly salient 2020 presidential election, which was the highest turnout since 1900. However, that leaves around one-third of eligible voters still deciding not to cast a ballot. We want to explore reasons that they may not have voted.

The dataset that we will use is from a poll conducted in September 2020 by Ipsos and FiveThirtyEight on voting behavior for 8,327 respondents. The survey covers a range of questions asking what it means to be a good American, how much the survey respondent agrees or disagrees with systematic racism statements, their trust and faith in the US government, how they are affected by government policy-making, thoughts on voting and past voting actions, and demographic information. We will need to convert some categorical variables into numerical values to conduct a correlation analysis and remove any variables that are highly correlated with one another. Then we will use forward stepwise selection to choose which variables should be included in the models.

We plan to use a random forest model on this dataset to determine what factors may affect voter turnout and predict the likelihood that an individual plans to vote. The model will be standard form. Note that this poll was conducted before the election and prior to most mail-in ballots being sent out, so we are not measuring whether an individual voted or not, simply their intent to do so.

The packages we will use are NumPy, Pandas, Matplotlib, Sklearn, and PyQT5. NumPy and Pandas will be used for data cleaning and preprocessing, Matplotlib will be used for visualizing our findings, Scikit-learn will be used to create, train, and test our models, and PQt5 will be used to create our GUI.

The reference materials we will use to obtain sufficient background on applying the chosen network to our voter analysis are the 6103 Data Mining course materials, the official documentation websites for the packages we will be using, and various websites such as medium.com.

We will judge the performance of our models by checking the accuracy score of our test dataset. The metrics we will calculate are: precision, recall, F1, and MSE. We may use accuracy as a metric, dependent on how balanced the dataset is between voters and non-voters. This will be examined during exploratory data analysis.

Our group plans to meet weekly on Tuesdays to discuss progress on the projects. The table below outlines the progress and deliverables that we plan to achieve at each date.

Proposed Date of Completion	Actions/Deliverables
Tuesday, March 30	• Decide on a dataset • Create GitHub repository
Tuesday, April 6	• Submit group proposal

	• Begin exploratory data analysis
Tuesday, April 13	• Clean and process the data for train-test into random forest and logistic regression models • Apply the models to the data
Tuesday, April 20	• Evaluate the results of the models • Refine models if needed • Begin writing the final report
Tuesday, April 27	• Create GUI • Create presentation
Sunday, May 2	• Recording of presentation • Finalize final report • Individual final reports
Monday, May 3	• Final Project