Sneha Dubey

Dr. Chen

CSCI 184

17 May 2024

Project Proposal

Project Title: Voter Demographics & Classification Models

Students: Sneha Dubey

Dataset: The dataset I will be using for this project is a subset of the Cooperative Election Study (CES) 2022 Study Common Content. The original, full dataset can be found here.

This study's dataset includes various demographic information pertaining to a nationally-representative sample of around 60,000 American voters; the researchers collected information such as each voter's home zip code, birthday, gender, education level, race, religion, past votes, and political affiliations.

Because this study was so comprehensive in its gathering of about 706 feature variables, I will be using a subset of the data to conduct this project. Specifically, I will be focussing upon the features that discuss physical/ideological characteristics of each of the 60,000 voters, with the target variable being TS_partyreg (which political party each of the voters are registered with). This means that I will be removing all features I consider to be distractions/noise, such as those pertaining to how the voters have voted in the past, from consideration (e.g. 2016/2020 presidential elections, senate elections, congressional elections, governors, etc.).

Even though I must heavily preprocess this data before I can begin working with it, I've chosen this dataset because it is important for me to truly evaluate each model's performance as it relates to *real* data. In addition, using data from an organisation such as the

CES allows for me to apply any and all of my findings to my understanding of the political climate around me; the real-world applicability of this project is therefore enhanced.

Project Idea: My project aims to explore the relationship between an individual's demographic information and their political party affiliation, and to successfully train an optimal classification model to further predict a voter's party allegiance based on their characteristics. The steps to be followed in this project are as follows:

1. Preprocess the CES 2022 Study data, dropping all unnecessary features and encoding data to be model-friendly; split into training and testing data
2. Train the following classification models on the same training data, minimising the differences in their finetuning

    Decision Tree
    Random Forest
    Support Vector Machine
    K-Nearest Neighbours
    Logistic Regression
    Naive Bayes
    XGBoost
    AdaBoost

3. Test each of the models using the testing data, ranking each of them in terms of their success (based on their performance metrics); determine the most-successful model
4. Display the relationships not only between the features and target, but between each of the features themselves for real-world analysis
5. Create a UI for the user to input their characteristics, or random characteristics of their choosing, to see how the most-successful model would classify them
6. Summarise all of my findings and apply them to the real world

Software Needed: This project will be solely completed within Jupyter Notebook files. The requirements to engage with the code include Python 3 and jupyter notebook to be downloaded and installed, as well as several Python libraries (exactly which ones are to be determined) to be installed into the working directory.

Papers to Read:

- Evaluation of Classification Models in Machine Learning
- Modelling Voting Behaviour During a General Election Campaign Using Dynamic Bayesian Networks
- Predicting Propensity to Vote with Machine Learning