DATA 698: Masters Research Project - Proposal

Gabriel Campos, Gabriella Martinez

Last edited September 15, 2024

Thesis Proposal: Predicting Election Outcomes Based on Historical and Census Data

Introduction

Data science has become integral to modern business practices, with organizations across various sectors implementing continuous monitoring and predictive analytics to enhance performance and meet key performance indicators (KPIs). Recent advancements have advanced statistical and data-driven methodologies, allowing for the integration of both quantitative and categorical variables in outcome forecasting. However, certain sectors present ongoing complexities in weighting the numerous influencing factors for accurate forecasting. The prediction of presidential election outcomes is one such area that continues to pose significant challenges despite these methodological advancements. Addressing these challenges is crucial for a comprehensive understanding of democratic processes. Employing data science techniques in election forecasting not only represents a valuable learning opportunity for students but also contributes to the broader field of political analysis.

Research Problem

The challenge of anticipating election outcomes, lies in the diverse dynamics and complexity that drive voter behavior. Historically, election forecasting has relied heavily on polling data and surveys. More recently, social media has been utilized as a tool to gauge candidate favorability, offering an alternative approach to prediction. However, both methods are prone to significant bias and can be skewed by unrepresentative demographic samples.

This thesis will develop a data-driven model to predict election outcomes by utilizing historical voter data. The Random Forest algorithm, as taught in the CUNY School of Professional Studies curriculum, will be employed for the predictive analysis. The limitations and shortcomings of traditional forecasting methods which often overlook granular demographic and socioeconomic data will be compared to the proposed data-driven approach. Furthermore, historical election patterns (such as the long-standing two-party system with Republican and Democratic controlling presidential outcomes) will be noted alongside its potential influence on forecasting. The research will aim to confront these shortcomings by utilizing county-level census data, voter turnout records, and additional categorical data such as income, education and party affiliation. By integrating historical trends and socioeconomic data, this model aims to enhance prediction accuracy and identify major factors influencing it. Significant political and social factors that may influence the election but are not incorporated into the model will be discussed in the final analysis.

Objectives

The primary objectives of this thesis are as follows:

- To use historical voter turnout, census data, and demographic details to predict future election results.
- To identify and highlight key variables, such as income and education, that may influence voter behavior but are not directly included in the prediction model.
- To employ predictive models taught in CUNY SPS Data Science curriculum for forecasting,
- To compare the model's predictions with historical election trends, focusing on major political parties (Republican and Democratic) and other key variables specific to each election cycle.

Methodology

This thesis will employ a data-driven methodology to develop a predictive model for election outcomes.

Data Collection:

Historical voter turnout and election results will be sourced from the County Presidential Election Returns dataset (2000-2020) available on the Harvard Dataverse platform (dataverse.harvard.edu). This dataset provides comprehensive county-level election data, which will serve as the foundation for the analysis.

Demographic and socioeconomic data, including information on age, race, gender, income, and education level, will be obtained from the U.S. Census Bureau's official website (www.census.gov). Specifically, datasets titled "Citizen Voting Age by Race and Ethnicity" for relevant years will be utilized. For example, to analyze the 2016 election, the dataset covering "Citizen Voting Age by Race and Ethnicity 2013-2017" will be employed, reflecting the period from the last inauguration up to the subsequent inauguration.

Preprocessing:

Data will be cleaned, processed, and normalized to ensure compatibility between different datasets (e.g., aligning census data with county-level voter turnout data, using FIPS numbers as geolocation/Unique ID for merging, etc.).

The election voting metrics will be combined with socioeconomic and demographic data to create a comprehensive dataset. This integration will enable a detailed analysis by aligning voting outcomes with relevant socioeconomic and demographic factors. Missing values will be handled through imputation, and variables will be encoded for use in the model.

Model Development:

Classifier will be developed to predict election outcomes, with features including historical voter turnout, demographic composition, and socioeconomic factors at the county level.

The model will be trained on past election data and validated both with cross-validation methods and the actual election result itself.

Model Evaluation:

Performance metrics such as accuracy, precision, recall, and F1-score will be used as well as a comparison of the election result.

Feature importance within the model will be analyzed to identify the most influential predictors.

Notable Variables:

The historical dominance of the Republican and Democratic parties in past elections will be examined and may be incorporated into the model, with alternative parties classified as "Other" for analytical purposes. This analysis will provide context for the model's predictions and assist in explaining any deviations from established patterns.

Significance

The significance of this research lies in its ability to improve the accuracy of election outcome predictions using a data-driven approach. By incorporating granular demographic and socioeconomic data at the county level, this model will provide more detailed insights into the factors driving voter behavior. Policymakers, campaign strategists, and political analysts can use these insights to better understand voting patterns and prepare for future elections. Additionally, this research can help identify key variables that could influence election results but are often overlooked in traditional forecasting models.

Conclusion and Timeline

This project will begin with data collection and preprocessing, which will take one month. Model development will follow over the next month, including training, validation, and tuning of the predictive model. The analysis phase, where model results are compared with historical trends and live results, will include notable variables, and will take an additional two weeks. Finally, the remaining time will be allocated for revisions, writing, and thesis submission.

The goal of this research is to offer a more refined approach to predicting election outcomes, using historical data and demographic insights, to contribute to the broader understanding of electoral behavior in the U.S.

Related Links

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap/2017-2021-CVAP.html

https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap/2013-2017-CVAP.html

https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap/2013-cvap.html

https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap/2009-cvap.html