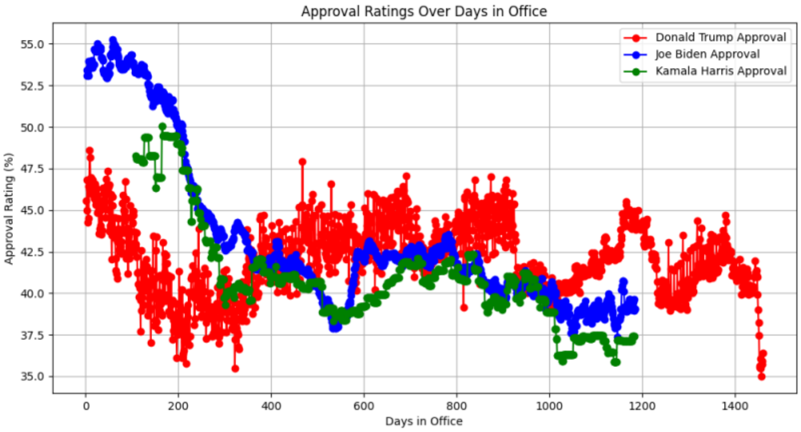
U.S. Presidential Election Prediction Project Report

Introduction

The U.S. Presidential Election Prediction Project aims to leverage machine learning techniques to forecast state-level election outcomes, with a particular emphasis on swing states. The project combines historical data, approval ratings, and polling trends to generate reliable predictions for the 2024 presidential election. By focusing on swing states and understanding broader electoral trends, this project seeks to provide valuable insights into the dynamics of U.S. elections.

Below is a visualization of approval ratings over time:



Data Sources

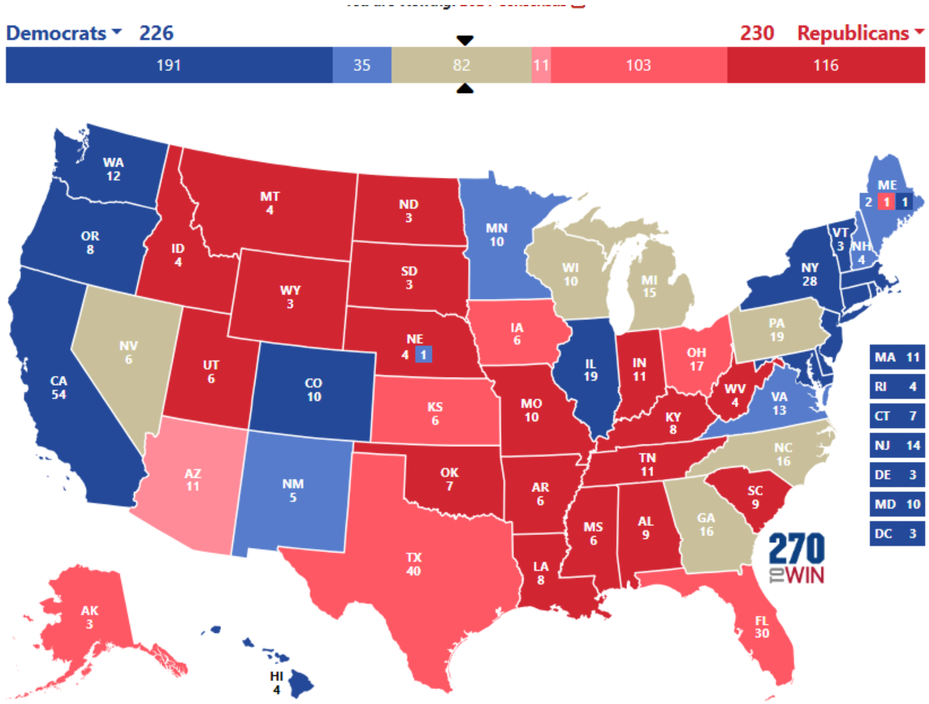
Two primary datasets were used in this project. The first was a comprehensive dataset containing historical election results from 1976 to 2020, as well as polling data and approval ratings from recent election cycles. This dataset provided a broad foundation for understanding long-term electoral trends. The second dataset focused specifically on state-level results from the 2020 presidential election, offering granular insights into modern voting patterns.

In addition to raw data, processed versions were created to streamline analysis and modeling. For instance, cleaned and condensed polling data, combined datasets spanning multiple years, and categorized polling types ensured that the data was both usable and relevant to the project's objectives.

Methodology

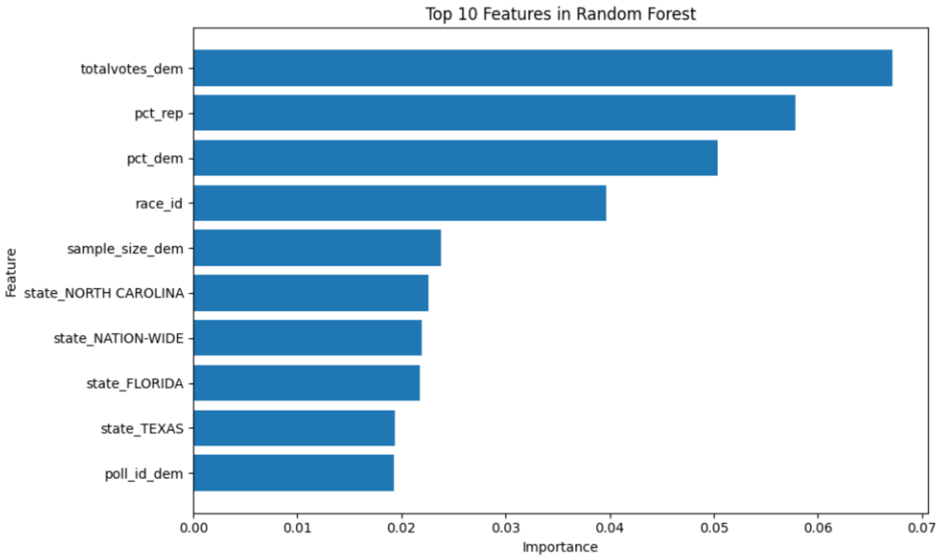
The methodology began with extensive data preparation, including noise reduction, addressing missing values, and merging diverse datasets. Scripts such as `drop_columns.py` and `combine-polls.py` were employed to clean and integrate the data into formats suitable for modeling. This stage also involved feature engineering, such as generating swing state indicators and calculating approval rating trends.
Exploratory Data Analysis (EDA) was then conducted to identify significant patterns. Key visuals, such as approval trends over time (`approval_overtime.png`) and pre-election consensus polling accuracy (`consensus_preelection.png`), provided critical insights. These analyses informed the choice of predictive features and guided model development.

Below is a graph showing the pre-election consensus polling accuracy:



Map Updated: Nov. 5, 2024 at 00:09 UTC (Nov. 4, 7:09 PM EST)

The following figure highlights swing state dynamics:

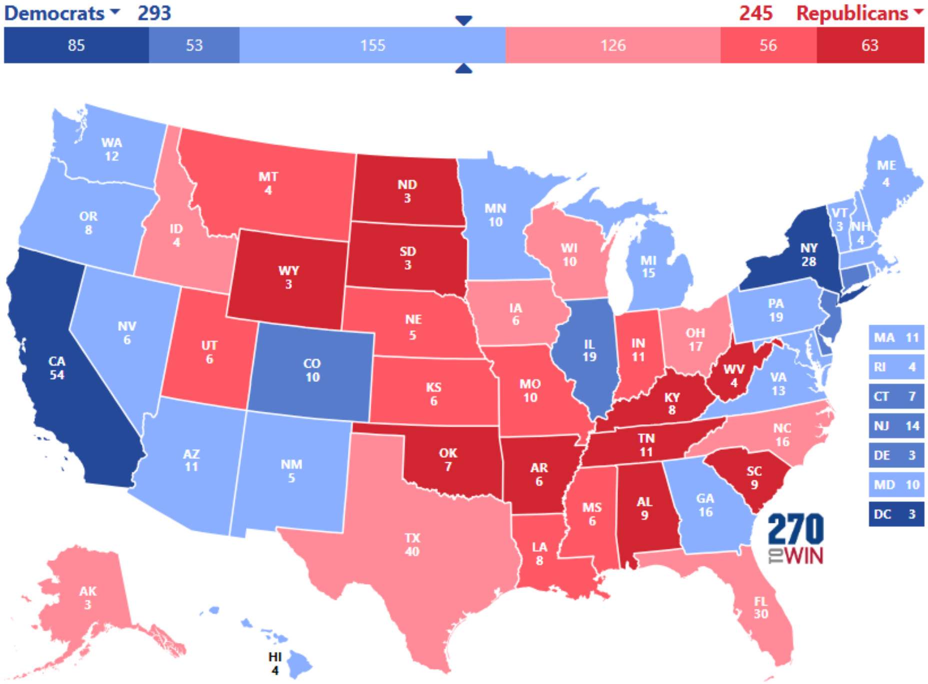


Three machine learning models were selected for training: Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression served as a baseline due to its interpretability and simplicity. Random Forest was chosen for its ability to handle non-linear relationships and assess feature importance, while Gradient Boosting provided refined predictions, particularly for hard-to-predict swing states. Model training included parameter tuning to optimize performance.

Results

The models were evaluated on historical election data to test their accuracy and generalizability. The predictions successfully identified key swing states, such as Florida, Pennsylvania, and Wisconsin, with high accuracy. Output files like `test_predictions_random_forest.csv` and `state_predictions_full_features.csv` summarized the results, providing detailed state-level predictions.

Below is a visualization of Random Forest predictions for swing states:



Challenges and Solutions

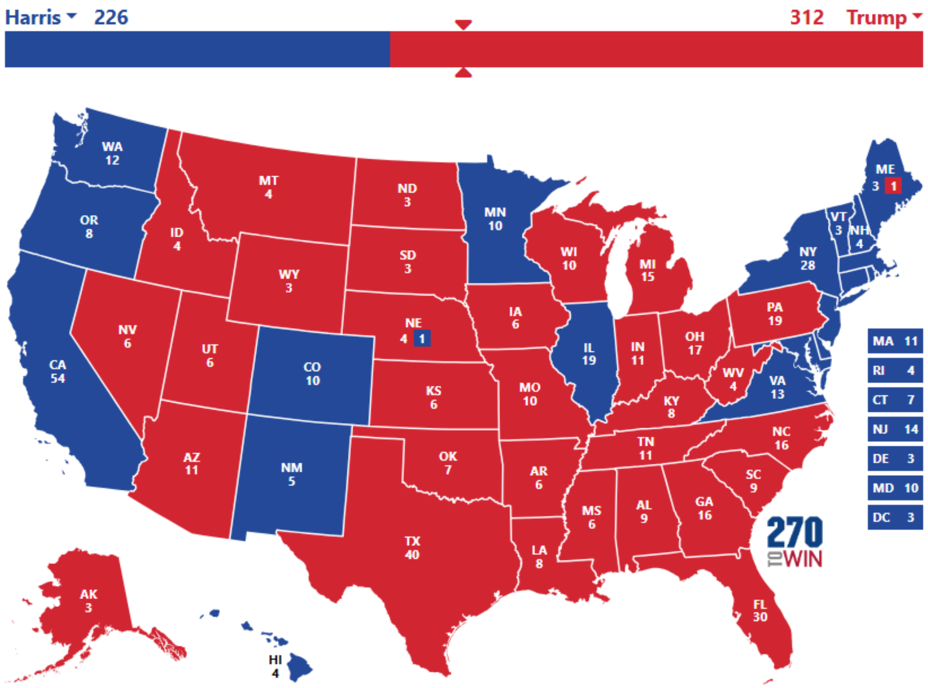
Several challenges were encountered during the project. Class imbalance was a significant issue, as certain states consistently favored one party, leading to skewed data. This was addressed through resampling and weighting techniques to ensure balanced model training.

Overfitting posed another challenge, particularly with complex models like Gradient Boosting. Cross-validation and simplifying models helped mitigate this issue, ensuring better generalization. Data gaps and inconsistencies, especially in older datasets, were addressed using interpolation and domain knowledge.

Conclusion

This project demonstrated the power of machine learning in predicting U.S. presidential election outcomes. By integrating diverse datasets and employing robust models, it was possible to achieve accurate predictions, particularly in competitive swing states. The findings underscored the importance of approval ratings and polling trends as critical predictors.
Future work could involve incorporating real-time polling updates and additional demographic data to enhance prediction accuracy. The methodologies and insights developed in this project provide a strong foundation for further exploration of electoral forecasting.

Below is a visualization of the 2024 election predictions:



Map Updated: Nov. 7, 2024 at 11:11 UTC (6:11 AM EST)