Predicting County-Level Election Outcomes

Snigdha Majeti

January 6, 2025

Integrating Diverse Data Sources to Forecast Election Results at the County Level

Introduction

The report presents an analysis of predictive models developed to forecast the outcomes of recent U.S. presidential elections at the county level. Various data sources, including polling data, betting odds, satellite imagery, and other relevant information, were utilized to construct the models. The report outlines the data sources employed, detailing the steps of data cleaning and processing undertaken to prepare the data for analysis. It further discusses the modeling approaches used, the challenges encountered, and potential improvements to enhance the accuracy and reliability of the predictions. The conclusion summarizes the findings and reflects on the effectiveness of the model in predicting election outcomes.

Data Sources and Justification

To build accurate predictive models for U.S. presidential election outcomes at the county level, a comprehensive view of various influencing factors is essential. These factors include historical election results, real-time public opinion, socio-economic conditions, and geographic features. Incorporating diverse datasets allows for a deeper understanding of how these elements influence voting behavior and election outcomes. Below is a brief overview of the datasets employed and the rationale behind their selection to inform the development of these models.

Election Results

Dataset Link: US County-Level Election Results
Overview: This dataset provides detailed historical election results at the county level across multiple U.S. presidential elections.
Reason for Selection: Historical election results are key for identifying voting patterns and trends across counties, which are crucial for building predictive models.

Betting Odds

Dataset Links: 2024 President Betting Odds, 2020 President Betting Odds, 2016 President Betting Odds
Overview: These datasets provide betting odds from various sources for U.S. presidential elections. They represent market-driven predictions about election outcomes.
Reason for Selection: Betting odds reflect collective sentiment and expectations, offering valuable insights into the likely outcomes of elections.

Polls

Dataset Link: FiveThirtyEight Polling Data
Overview: This dataset aggregates polling data from multiple sources, offering a snapshot of public opinion and voter sentiment.
Reason for Selection: Polling data provides real-time insights into voter intentions and sentiments, which are essential for understanding the political landscape leading up to the election.

Average Party Vote Share (Previous 2 Cycles)

Dataset Link: MIT Election Lab
Overview: This dataset includes historical election data and the votes received by major candidates over the past years.
Reason for Selection: This dataset was used to calculate the average vote share by party from the past two cycles, allowing for the adjustment of polling and betting odds from state-level predictions to more accurate county-level estimates.

Cartographic Data

Dataset Links: 2023 Cartographic Boundary File, 2020 Cartographic Boundary File, 2016 Cartographic Boundary File
Overview: This dataset provides geographic boundary files defining the counties and equivalent areas across the U.S.
Reason for Selection: Geospatial data is used to extract satellite images of counties and analyze land and water areas, helping with spatial analysis of election outcomes.

Unemployment Rates

Dataset Link: Kaggle Unemployment Rates
Overview: This dataset includes unemployment rates at the county level from 1990 to 2023.
Reason for Selection: Economic factors, like unemployment rates, often influence voter behavior, and this dataset helps account for economic conditions affecting election outcomes.

Census Data

Dataset Link: U.S. Census Data
Overview: This dataset provides demographic data, including population size, age, race, income, and other socio-economic information.
Reason for Selection: Demographic data is crucial for understanding voting behavior, as factors such as age, race, and population changes can significantly impact election outcomes.

Voter Turnout Data

Dataset Link: Ballotpedia Voter Turnout Analysis
Overview: This dataset provides insights into voter turnout in the 2024 U.S. general election.
Reason for Selection: Voter turnout is a crucial factor in election outcomes, and this dataset refines predictions by analyzing turnout variations across counties in previous elections.

Data Cleaning and Processing

The data cleaning and preprocessing involved merging election results, unemployment rates, census data, satellite images, voter turnout, party vote percentages, polls, and betting odds from 2016, 2020, and 2024. Missing values were addressed, and adjustments were made to allow for detailed county-level analysis.

Election Results (2016, 2020, 2024)

- Combined election results from 2016, 2020, and 2024 into a single dataset.
- Standardized column names and data types to ensure consistency across all years.
- Converted state abbreviations in the 2016 data to full state names for uniformity.
- Ensured there were no null values in the dataset.

Unemployment Rates (2016, 2020, 2023)

- Calculated the average unemployment rate for each county in the years 2016, 2020, and 2023.
- Replaced the year in the 2023 data with 2024 and merged it with data from other years into a single dataset.
- Ensured no null values in the dataset.

Census Data (2016, 2020, 2023)

- Merged census data on population estimates and changes across age groups and races for 2016, 2020, and 2023.

- Replaced the year in the 2023 data with 2024 and merged it with data from other years into a single dataset.
- Ensured no null values in dataset.

Satellite Images (2016, 2020, 2023)

- Extracted satellite images for 2016, 2020, and 2023, filtering out images with over 10% no-data or black pixels.
- Imputed missing pixel values by replacing them with the mean of non-black pixels.
- Used a shapefile to associate image data with counties, keeping only relevant columns (fips, ALAND, AWATER, image_name).
- Replace the year in 2023 data to 2024 and merged all satellite images into one dataset, ensuring no null values.

Voter Turnout Adjustment (2016, 2020, 2024)

- Combined state-level voter turnout data from 2016, 2020, and 2024 into a single dataset, ensuring no null values.
- Merged state-level voter turnout data with census and election data by year and state.
- Filled missing voter turnout values with national averages by year.
- Created adjusted county-level voter turnout percentages using:
  - **Eligible Population:** Multiplied state-level turnout by the ratio of eligible voters in the county to the state.
  - **Total Votes:** Multiplied state-level turnout by the ratio of total votes in the county to the state.
  - Clipped adjusted turnout values to 100% to prevent excess.

Party Vote Percentages (2008-2023)

- Analyzed historical election data (1990-2023) to calculate party vote percentages for 2008-2023.
- Filled missing vote percentages by averaging across parties and years.
- Estimated the average party vote share for the 2016, 2020, and 2024 elections using data from the previous two cycles, ensuring no null values.

Polls Data (2016, 2020, 2024)

- Combined state-level poll data for the Democratic and Republican parties from 2016, 2020, and 2024 into one dataset, ensuring no null values.
- Adjusted state-level polls to the county level by:

- Merging state-level polls with county-level data.
- Filling missing polling values by averaging polls for each party per year.
- Calculating county-level polls by multiplying state-level polls by a ratio of difference between the historical party vote share and the state-level polling percentages.
- Clipped adjusted polls to 100% to prevent exceeding.

Betting Odds Data (2016, 2020, 2024)

- Collected national-level betting odds data and calculated a weighted average, giving more weight to recent data.
- Combined national betting odds for 2016, 2020, and 2024 into one dataset with no null values.
- Adjusted national betting odds to the county level by multiplying by the ratio of the county's average party percentage to the national average.
- Clipped adjusted betting odds percentage to 100% to prevent exceeding.

Final Dataset Preparation

- Merged all county-level datasets (election results, unemployment, census, satellite images, voter turnout, party votes, polls, and betting odds) based on fips, state, and county name.
- Removed duplicates and ensured the final dataset had no null values.

EDA and Feature Selection

The final dataset consists of 1,427 columns, including 1,420 features and 7 target variables. Census data, originally divided into 18 age groups, is consolidated into three broader categories: 0 to 24, 25 to 64, and 65+. Numerical features of integer type with more than 60% zero values are removed, along with those having a limited number of unique values. Highly correlated features are discarded, and for correlated feature pairs across age groups, the working-age group (25-64 years) is retained. Following feature reduction, the dataset contains 30 features and 2 target variables.

After conducting exploratory data analysis (EDA), additional features are removed, leaving only one target variable: `per_gop`, which represents the percentage of votes for the Republican Party (Grand Old Party). The `per_dem` variable is excluded, as it is complementary to `per_gop`. A total of 26 features are retained. For the feature `adjusted_gop_pct_estimate` (which represents the GOP's polling percentage), noise is added to mitigate the risk of data leakage, as it is highly correlated with the target variable.

Modeling Approach

Data is split into training and testing sets, with 20% of the data allocated for testing. For the `year` and `fips` features, label encoding is applied. Numerical features are normalized and scaled using MinMaxScaler. Image data is preprocessed using the `preprocess_input` function, tailored for the MobileNetV2 model.

Ensemble Model: An ensemble approach is used, consisting of a Gradient Boosting model trained on tabular data (excluding images) and a pre-trained MobileNetV2 model trained on satellite images. The predictions from both models are combined through a weighted average, with the weights determined by the R^2 scores from both the training and testing datasets. Hyperparameter optimization is performed using `RandomizedSearchCV`, exploring a grid of parameters.

Multimodal Neural Network: A multimodal neural network is developed, integrating both image and tabular data as input, with MobileNetV2 serving as the pretrained backbone for image features. Various hyperparameters, such as regularizer constants and dropout rates, are tested to mitigate overfitting and improve model generalization.

Feature Fusion Model: In the Feature Fusion approach, features are extracted from the image data using pre-trained MobileNetV2. Principal Component Analysis (PCA) is then applied to reduce the dimensionality of the image features while retaining variance. These image features are combined with the normalized tabular features and fed into a Gradient Boosting model. Hyperparameters are tuned through a parameter grid search to identify the best-performing model.

Metrics	Ensemble Model	Multimodal Neural Network	Feature Fusion Model
MSE_Training	0.0031	0.0005	0.0041
R^2_Training	0.8780	0.9796	0.8362
MSE_Testing	0.0062	0.0064	0.0066
R^2_Testing	0.7546	0.7470	0.7390

Table 1: Performance Metrics of Models

Based on the performance of the three models presented in Table 1, the Feature Fusion Model is selected due to the small gap of less than 10% between R^2_train and R^2_test, indicating minimal overfitting and strong generalization from training to testing data. Therefore, the Feature Fusion model is selected as the final model.

Potential Improvements

The model performance can be improved further with several strategies. Expanding the dataset by including election years or incorporating more diverse features such as candidate characteristics, political party alignment, and relevant images could provide richer context. Refining feature selection and introducing new variables that capture regional and political dynamics and applying regularization techniques would further enhance accuracy.

Conclusion

In conclusion, the feature fusion model delivers the best performance, achieving 92% accuracy, 93% precision, 98% recall, and 95% F1 Score on the 2024 data. These results demonstrate the model's strong ability to predict election outcomes per county accurately, capturing the key patterns and dynamics involved. The high precision and recall indicate that the model performs well in both identifying true positives and minimizing false negatives, while the impressive F1 score reflects a balanced and reliable performance. Given its effectiveness in making accurate predictions, the feature fusion model stands out as the most robust and reliable choice for this task.