Predicting 2024 U.S. Election Outcomes Using Machine Learning

Final Report

Team: Zubair Atha, Eesha Barua, Sibi Marappan, Justin Mathew, Pranitha Natarajen

Introduction

Forecasting U.S. presidential elections is a complex task that combines historical insights with modern data analysis. The Baseline Bayesian Model was the foundation of this project, initially used to explore trends and test data assumptions. Building on this, we developed a machine-learning pipeline to predict the 2024 election results using advanced models.

Datasets Used

To ensure comprehensive coverage, the following datasets were utilized:

- **1968–2016 Polling Data:** Historical trends sourced from the MIT Election Lab.
- **2020 and 2024 Polling Averages:** Real-time state-level polling data obtained from FiveThirtyEight.
- **State-Level Election Results:** Historical outcomes from public records for analysis and validation.

These datasets provided both depth and breadth, enabling a robust foundation for predictive modeling and validating state-level results.

Models Evaluated

Five models were implemented and tested to assess their predictive capabilities:

- **Baseline Bayesian Model:** Establishes probabilistic benchmarks based on historical polling data.
- **Linear Regression:** A straightforward approach to capturing linear trends in vote shares.
- **Random Forest:** An ensemble method leveraging decision trees for robust predictions.

- **Decision Tree:** A simpler tree-based model for state-level predictions.
- **XGBoost:** An advanced boosting technique for improving prediction accuracy.

Each model's strengths and limitations were explored, with the goal of understanding which approach best captures electoral dynamics.

1 Methodology

1.1 Baseline Model (Bayesian)

The Baseline Bayesian Model was the starting point of this project, providing a foundational benchmark. It leveraged historical polling trends to estimate the probabilities of Democratic and Republican wins at the state level. This model was instrumental in identifying patterns and informed the design of more advanced machine-learning models, though its simplicity resulted in higher error rates.

1.2 Training Process

Building on the insights gained from the Baseline Model, advanced machine learning models were trained to predict vote shares for the Democratic and Republican parties. The training process included:

- Splitting the dataset into **training (80%)** and **testing (20%)** subsets to validate performance.
- Using historical data from 2004 to 2020 for training, ensuring models captured relevant trends.
- Incorporating engineered features like state partisan lean, days until the election, and incumbency indicators to enhance accuracy.
- Optimizing hyperparameters for Random Forest and XGBoost using grid search.

The models were trained to predict the vote shares for both parties, with the target variable being the actual vote share. These predictions were aggregated to calculate state-level electoral votes, ultimately determining the winner for each state and the overall election.

1.3 Evaluation Metrics

The models were evaluated on:

- **Mean Absolute Error (MAE):** Measures prediction accuracy.

- **State Accuracy:** Proportion of states with correct predictions.
- **Prediction Correctness:** Whether the predicted winner matches the real-world outcome.

The combination of these metrics ensured both state-level and overall predictive accuracy were rigorously evaluated.

1.4 Electoral Vote Comparison

Table 1 summarizes the predicted electoral votes for Harris (Democrat) and Trump (Republican) across all models, alongside the real-world result for 2024:

Model	Harris (DEM)	Trump (REP)	Winner	Correct Prediction
Baseline (Bayesian)	289	249	Harris	✗
Linear Regression	247	289	Trump	✓
Random Forest	267	269	Trump	✓
Decision Tree	267	269	Trump	✓
XGBoost	267	269	Trump	✓
Real-Life Result	226	312	Trump	✓

Table 1: Electoral Vote Comparison

2 Conclusion and Final Results

2.1 Results Overview

Model	MAE (Harris)	MAE (Trump)	State Accuracy
Baseline (Bayesian)	63.00	63.00	0.00
Linear Regression	21.00	23.00	1.00
Random Forest	41.00	43.00	1.00
Decision Tree	41.00	43.00	1.00
XGBoost	41.00	43.00	1.00

Table 2: Model Metrics Comparison

This table highlights the superior performance of Linear Regression in terms of both MAE and state prediction accuracy, while ensemble methods like Random Forest and XGBoost also performed well.

2.2 Insights and Future Scope

The Linear Regression model demonstrated superior performance due to its simplicity and better alignment with feature-engineered data. Future work could explore voter demographics, campaign data, and real-time sentiment analysis for improved predictions.