Ogbonna Ngwu

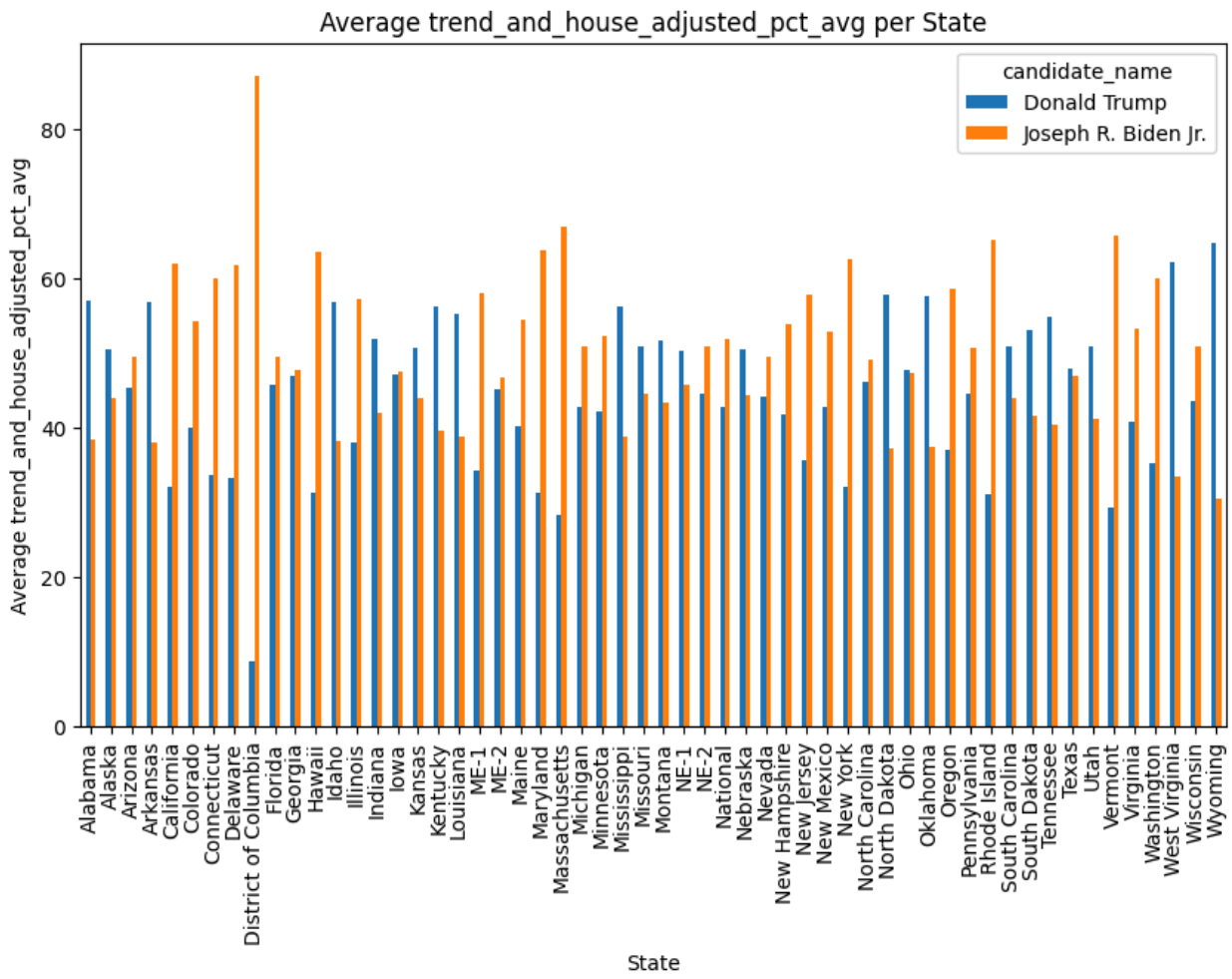
ngwuogbonnaprince@gmail.com | 08165533706

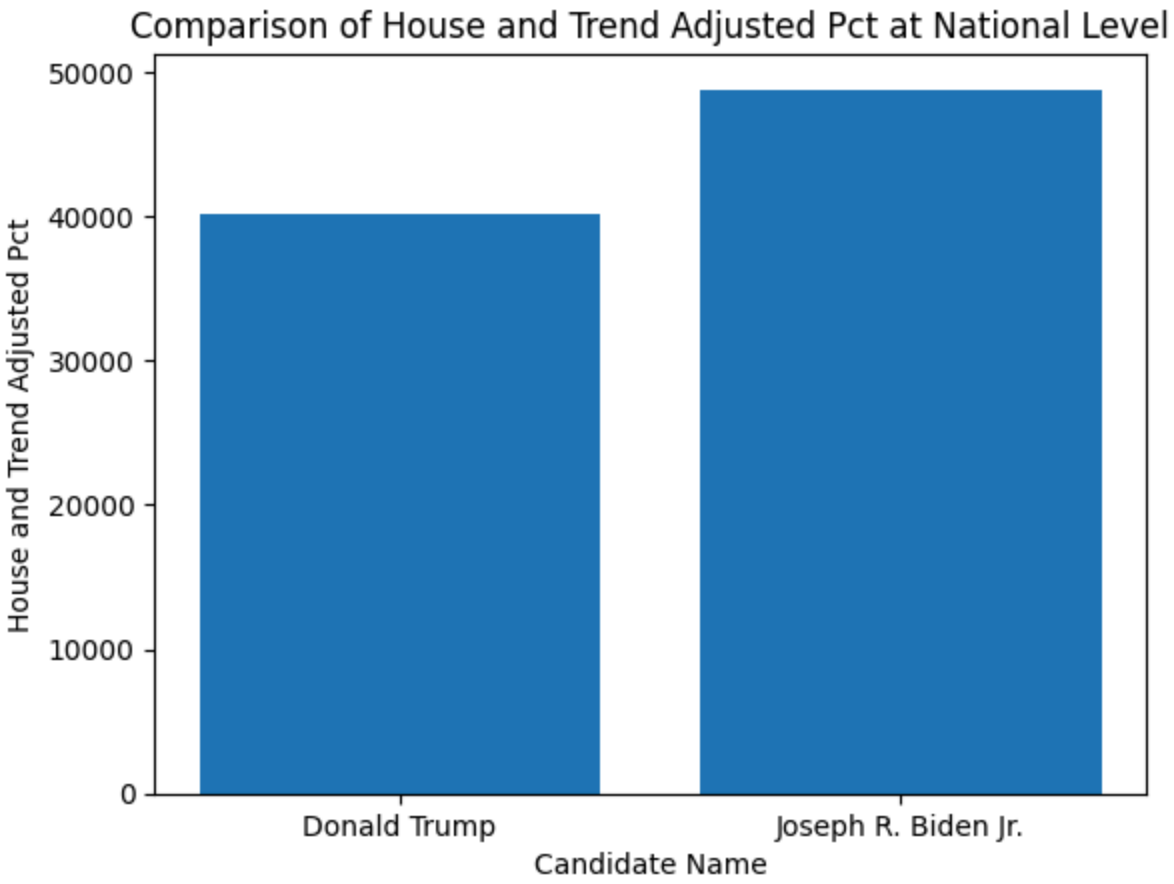
Report: Predicting Future U.S. Elections Using Pre-2020 Survey Data

The steps taken and findings from the analysis are summarized below.

Data Preprocessing and Feature Engineering

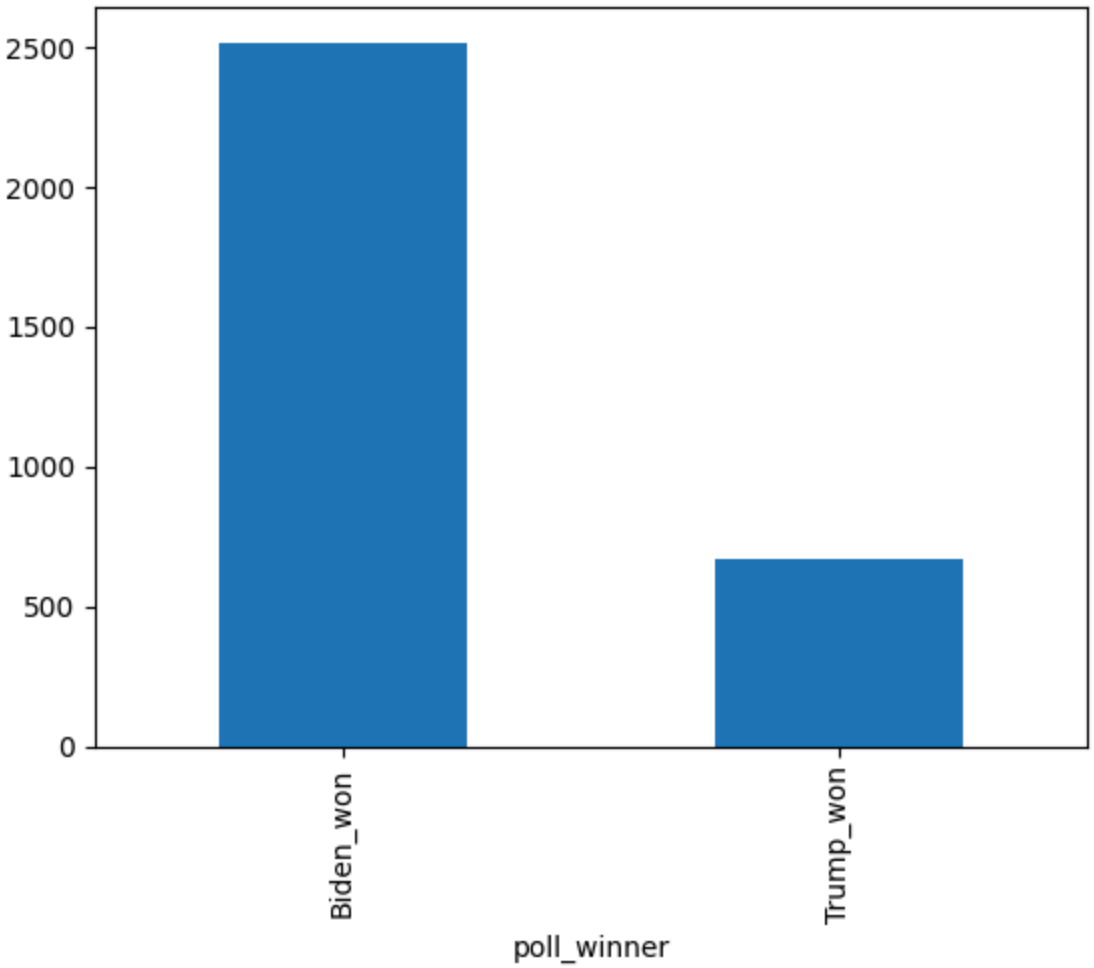
1. **Data Cleaning:**
   - Checked for null values in the dataset.
   - Filled null values in the 'tracking' column with 'F' (assuming binary nature) and the 'samplesize' column with the mean value.
   - Converted date columns ('modeldate', 'enddate', 'startdate') to datetime format.
2. **Descriptive Analysis:**
   - Used `.describe()` to summarize numeric feature distributions.
   - Checked value counts in the 'state' column, noting 'National' as the most frequent followed by 'Wisconsin'.
   - Examined candidate counts per state and overall, in the survey.
3. **Data Exploration:**
   - Visualized average state-level winners and national level winners.


Average trend_and_house_adjusted_pct_avg per State
candidate_name
Donald Trump
Joseph R. Biden Jr.
Average trend_and_house_adjusted_pct_avg
0
20
40
60
80
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
ME-1
ME-2
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
NE-1
NE-2
National
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming
State



- Calculated means for 'trend_and_house_adjusted', 'pct', and 'house_adjusted_pct' to master the data distribution.

4. **Feature Engineering:**
   - Derived new features:
     - `adjustment_impact`: Difference between 'trend_and_house_adjusted_pct' and 'pct'. This captures the overall impact of all possible adjustments to ‘pct’ feature by both trend and house effects.
     - `house_effect`: Difference between 'house_adjusted_pct' and 'pct'. Captures the house bias on the ‘pct’ of each candidate.
     - `trend_effect`: Difference between 'house_adjusted_pct' and 'trend_and_house_adjusted_pct'. Captures the effect of the candidate's public/social standing on the ‘pct’ feature.
     - `poll_to_poll_change`: Difference in subsequent polls for each candidate to capture voter bias and behavior changes.
   - One-hot encoded the 'population' column.
   - Created the target variable `poll_winner` based on 'trend_and_house_adjusted_pct' as this feature captured the ‘pct’ after trend and house effects have been factored in.

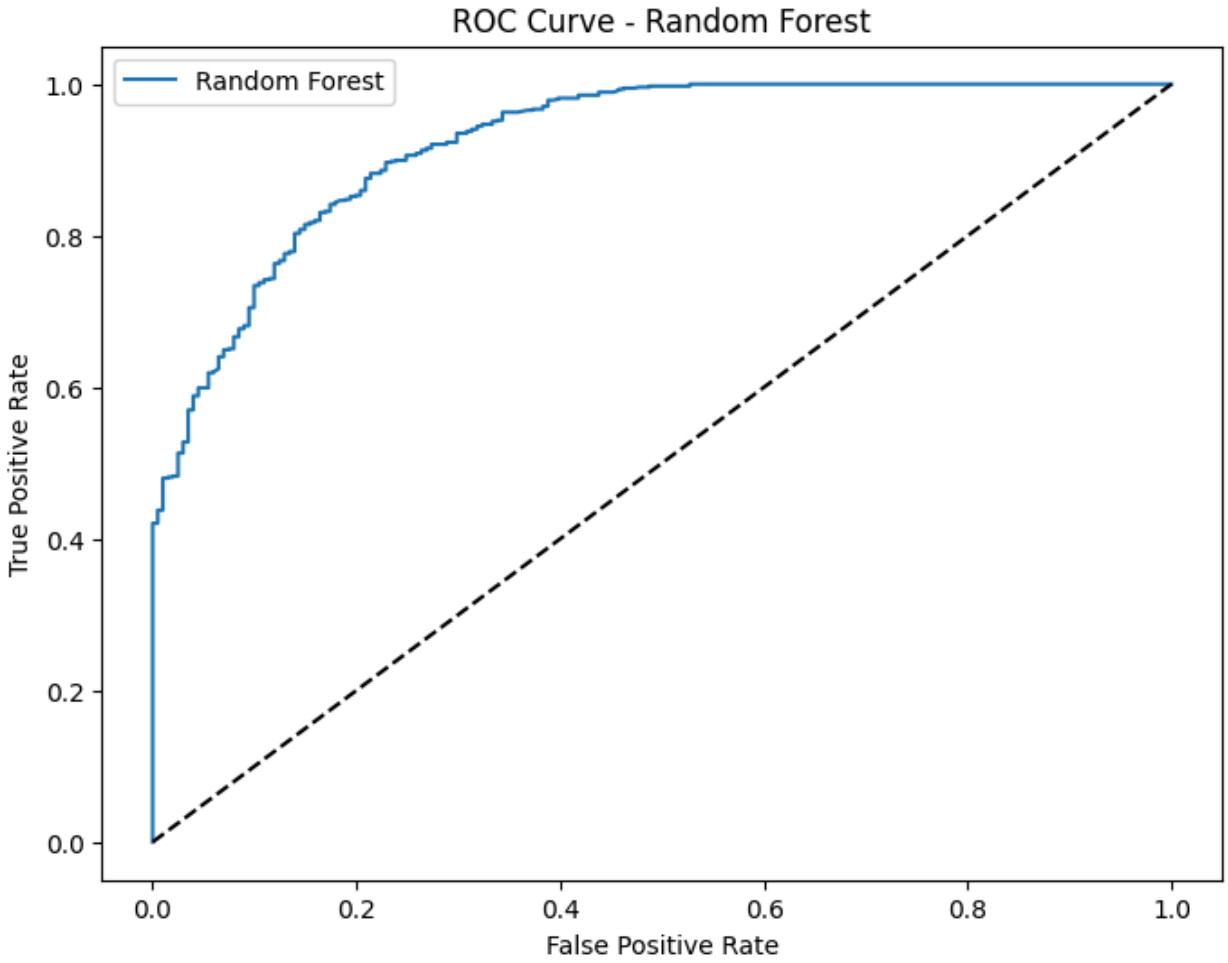


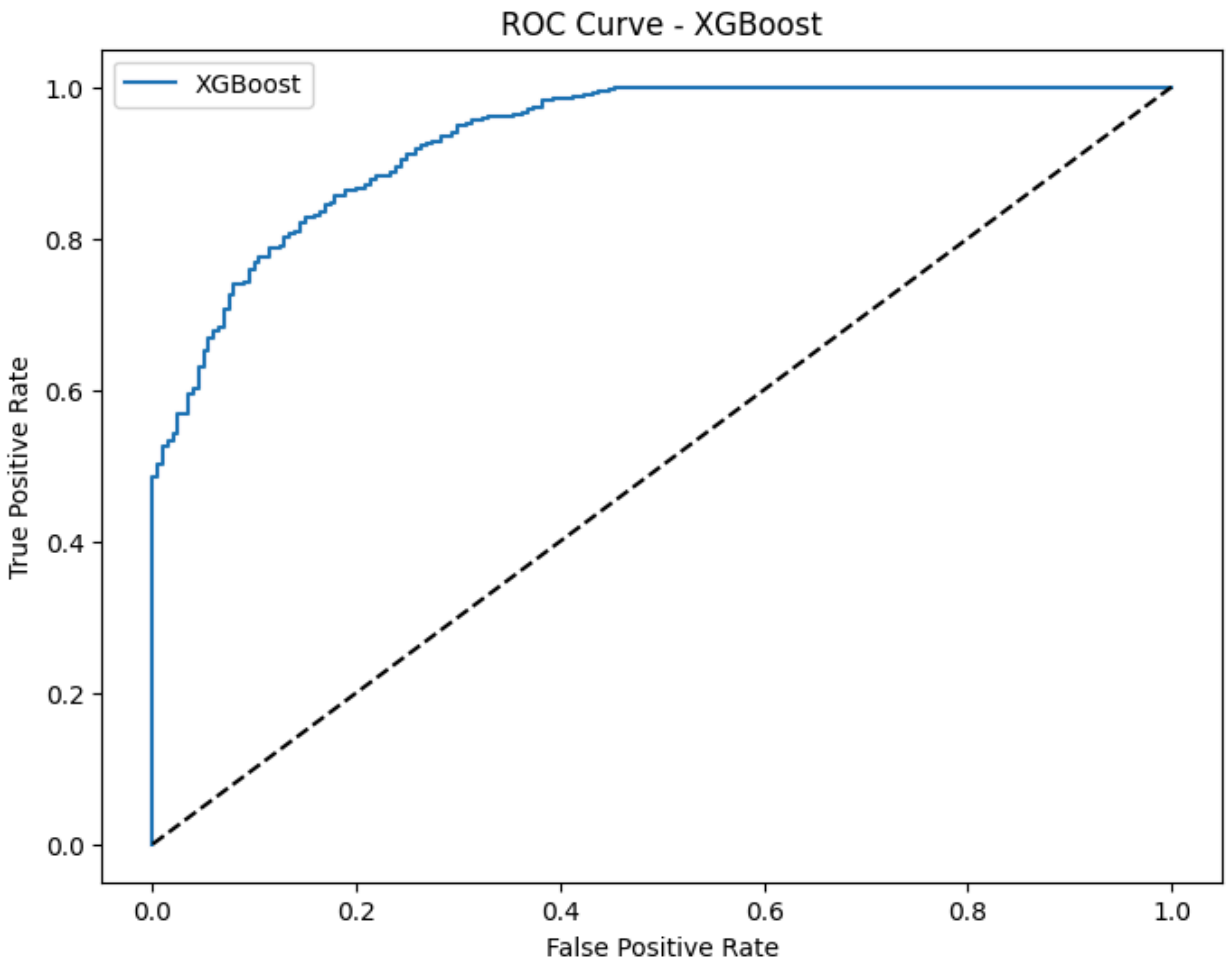
- Converted binary 'tracking' column ('T' to 1, 'F' to 0).
- Engineered a 'high_electoral_votes' feature indicating states with high electoral votes. This feature is important as it is a pointer at a candidate's possible outcome on the crucial Electoral College votes that determines the winner of the election.

5. **Data Preparation:**
   - Split data into separate datasets for each candidate and merged them while dropping common columns.
   - Renamed columns for clarity on candidate specific attributes.
   - Checked and dropped independent features with high correlation to avoid multicollinearity bias for my model.
6. **Model Development:**
   - **Data Splitting**:
     - Used stratified fold method to maintain candidate distribution, split 70 for training set, 30 for test set.
   - **Model Training:**
     - Trained two classifiers: RandomForestClassifier and XGBoostClassifier.
     - Applied cross-validation (cv=4).
     - Used L1 regularization to prevent overfitting.

- Employed GridSearchCV for parameter optimization.

- **Model Evaluation:**
  - Evaluated models using F1 score, Recall, Precision, and Accuracy.
  - Used confusion matrix and ROC curve for further evaluation.





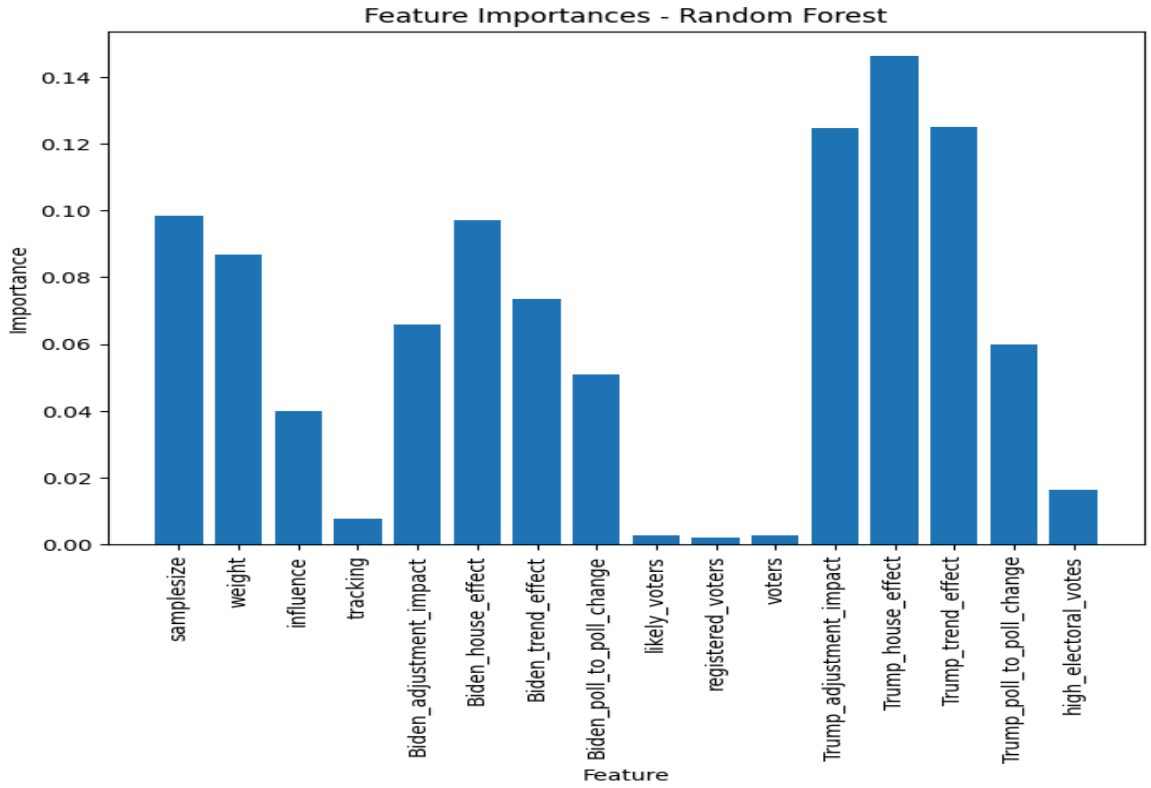
7. Findings.
   - General Observations:
     - The dataset was relatively clean but not perfectly structured.
     - Majority of polls were conducted at the national level.
     - Joe Biden led in approximately 70% of the states and the majority of high electoral vote states.
     - Biden won 2518 out of 3188 polls, whereas Trump won 201 polls.
   - Model Performance:
     - RandomForestClassifier:
       - F1 Score: 0.939
       - Recall: 0.985
       - Precision: 0.897
       - Accuracy: 0.899
       - Class 0 (Trump wins):
         - Correctly classified: 116
         - Misclassified as Biden: 85
       - Class 1 (Biden wins):
         - Correctly classified: 746

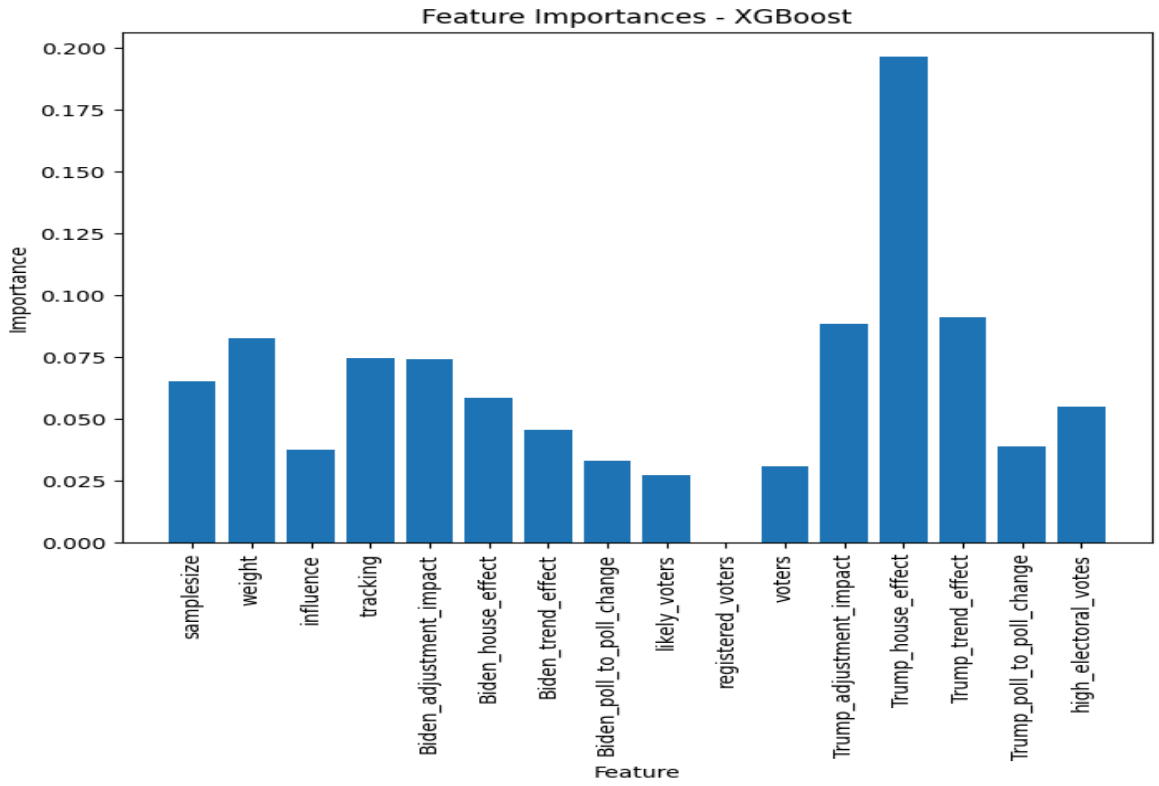
- Misclassified as Trump: 11

0 [116 85]
1 [11 746]

- XGBoostClassifier:
  - F1 Score: 0.94
  - Recall: 0.987
  - Precision: 0.897
  - Accuracy: 0.90
  - Class 0 (Trump wins):
    - Correctly classified: 116
    - Misclassified as Biden: 85
  - Class 1 (Biden wins):
    - Correctly classified: 746
    - Misclassified as Trump: 10

0 [116 85]
1 [10 746]





8. Model metrics implications:

Both models perform similarly in terms of F1 Score, Recall, and Precision. However, XGBoostClassifier has a slight edge in Recall and Accuracy:

- Recall: Indicates that the models are very good at identifying Biden wins correctly (high true positive rate).
- Precision: Shows the models' reliability in predicting Biden wins, with a small trade-off in misclassifying some Trump wins.
- F1 Score: Balances the trade-off between precision and recall, indicating both models handle the imbalanced data reasonably well.
- Accuracy: While high, it's less informative in this context due to the imbalance in classes.

9. Model Insights:

1. The both Algorithms gave much weight F to house effects of both candidates as the prime factors in determining whether a candidate wins or lose an election.
2. The regularization of both models gave higher weight to Trump related effect so as to easily recognize Trump wins, as this is the lesser class of the two outcomes of the election.

3. In the USA election, what really matters most are the public perception of the candidate and the decision of the House.
4. XGBoost generalized more on the test set, giving better performance metrics scores than Random Forest.
5. Given more iteration time, these two models would do way better than its current performance.
6. High electoral vote states played a role in the model in prediction of election winner.
7. Joe Biden won the majority of the states, including at National level, hence the winner of the presidential election.
8. Important features include **trend effects** on each candidate, **house effects**, and **adjustment effects** that were derived through feature engineering. with no impact on the model from feature scaling.

10. Analysis Challenges:

1. The data wasn't enough to train a robust machine learning model to generalize well on unseen data.
2. Data description and meta data provided wasn't enough, took a lot of research and questions to understand. It was almost all about educated guess and ambiguity. That's where we thrive tho, question asking and testing hypotheses and asking a lot of questions.
3. Time to model development. Even after submitting before the deadline, I'm still working on the models.