Election Outcome Prediction in the United States

…

Venkata Bharath Chakravarthi Kudumula

Introduction

Two-Party Dominance

- Democratic and Republican parties are the two major contenders.

Electoral College System

- 538 total electors
- A candidate needs 270+ votes to win the presidency.

Winner-Takes-All Mechanism

- Most states allocate all their electors to the candidate with the majority vote in that state.

Swing States Matter

- Some states (e.g., Pennsylvania, Wisconsin) can go either way — these determine close elections.

Problem Statement:

The purpose of this project is to investigate prediction of U.S. presidential elections at the county level by implementing a variety of socio-economic, demographic and historical election data. While traditional election forecasting relies mainly on number polls and opinions of experts, the approach taken in this study is data-driven using machine learning models to evaluate patterns in voter behaviour.

This project will combine data from the U.S. Census (American Housing Survey) and the certified election results, for the purpose of creating strong classification and regression models to:

- Accurately classify if a county will likely vote Democratic or Republican, and
- Predict the percentage of votes the Democratic candidate is expected to achieve.

Additionally, a general aim is to be able to understand the predominant factors driving electoral outcomes, assess how accurate publicly available data can be, and analyse how patterns in race, income, employment and commuting habits relate and translate to voting behaviours across units.

Data Overview

Datasets Used

- **U.S. Census Data (ACS 2015 & 2017)**
  - Demographics, income, employment, commuting patterns
- **U.S. Election Data (2020)**
  - County/state-level certified vote counts (President, Senate, Governors)

Key Feature Categories

- **Demographics:** Total Population, Race, Gender
- **Economics:** Income, Poverty, Employment
- **Transportation:** Commute modes, Work-at-home
- **Election Results:** Vote counts and % for REP/DEM across races

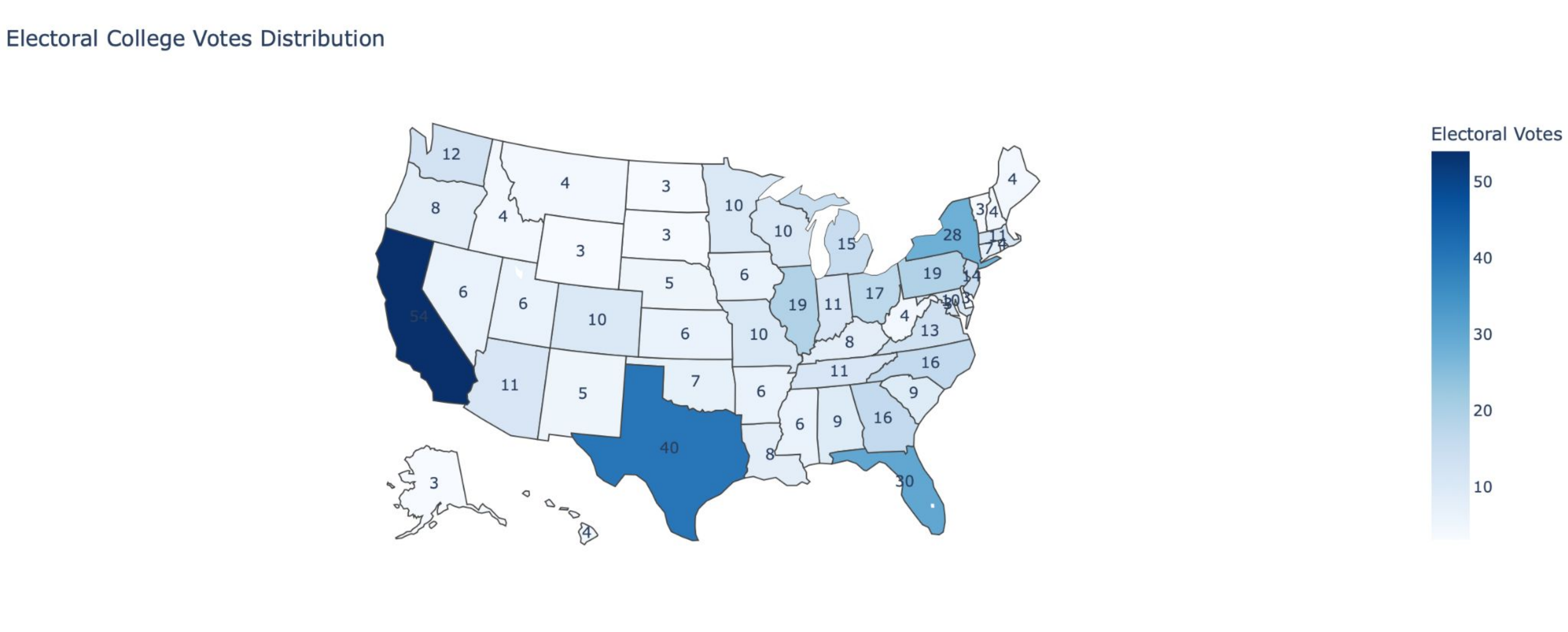
Data Sources

- Kaggle (Compiled from U.S. Census Bureau and State Election Offices)
- Public domain, licensed under Creative Commons

Summary Statistics

Electoral College Votes Distribution :

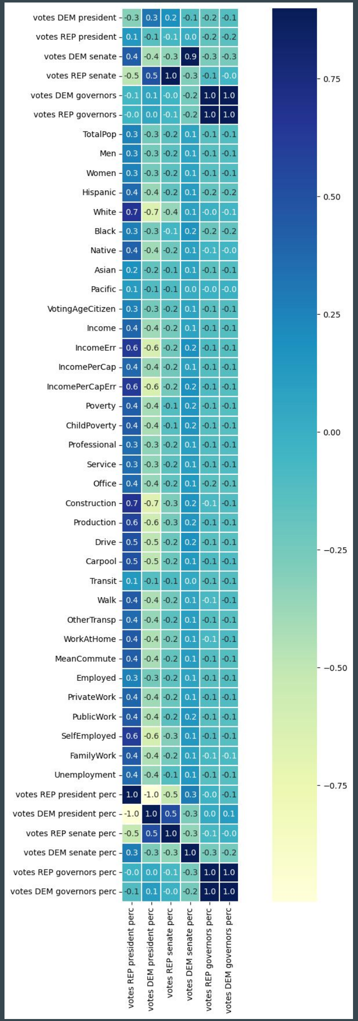
- California is the state with most electoral votes followed by Texas, Florida and Newyork.
- States with the Fewest Electoral Votes (3 votes each): Alaska, Delaware, North Dakota, South Dakota, Vermont, Wyoming, and the District of Columbia.



Spearman Correlation Heatmap

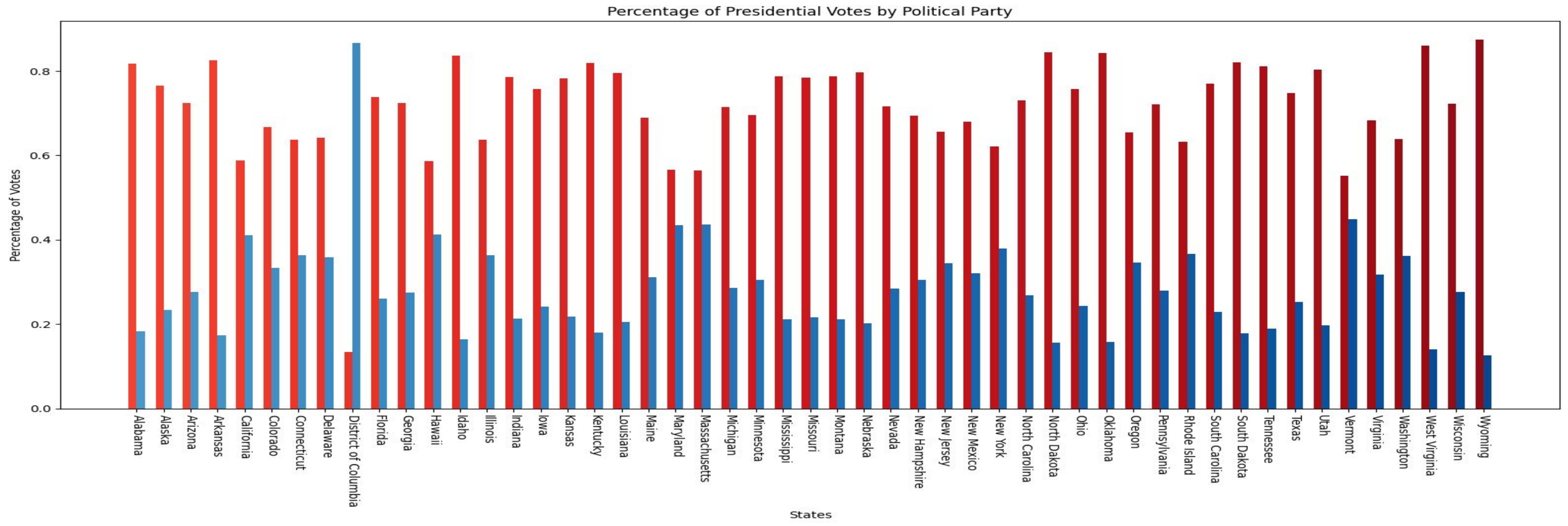
- Republican president's vote are really correlated with race. He receives votes mostly from white people. Moreover his votes come from people with low income, non self-employed

- Democratic president has majority of correlation perfectly opposite



No. of Votes by Political Party

- Partisan Strongholds: States like Wyoming, Alabama, and Oklahoma are heavily Republican, while California, New York, and Massachusetts lean strongly Democratic.
- Swing State Competition: States like Pennsylvania, Wisconsin, and Michigan show close margins, highlighting their battleground status.
- D.C. as an Outlier: District of Columbia overwhelmingly favors Democrats, as seen in its dominant blue bar.

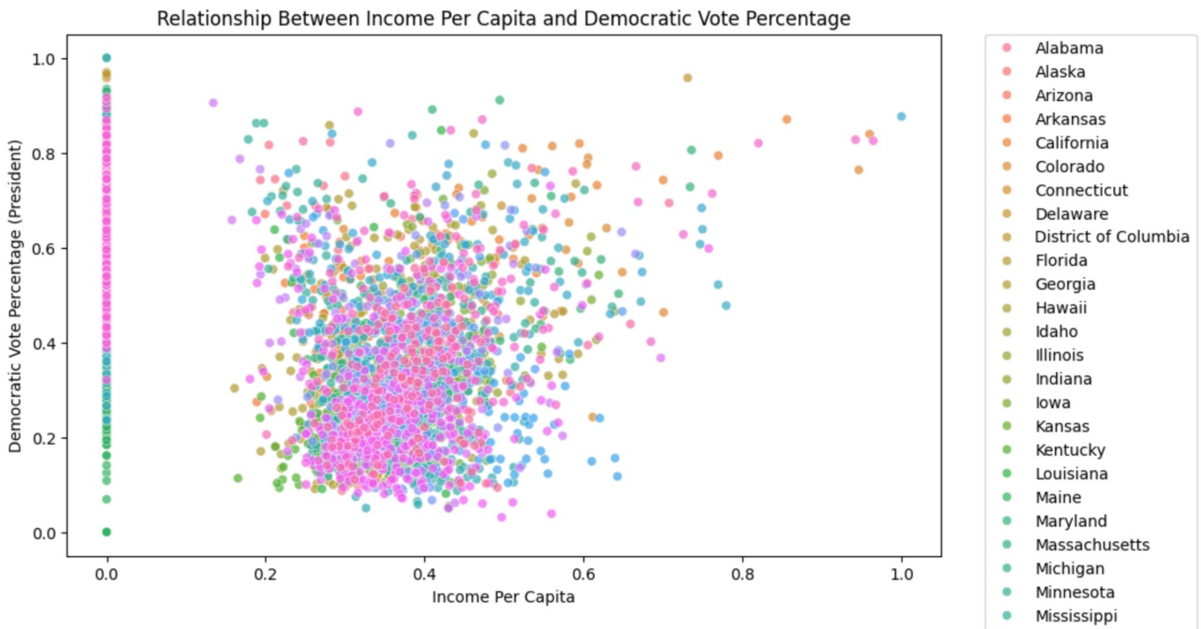


Income PerCapita Vs Democratic Presidential Vote Share

- Each dot represents a county's income per capita vs. Democratic presidential vote share.
- No strong linear relationship is observed—income alone does not predict voting behavior.
- Low-income counties show mixed support, including strong Democratic leaning in some areas.

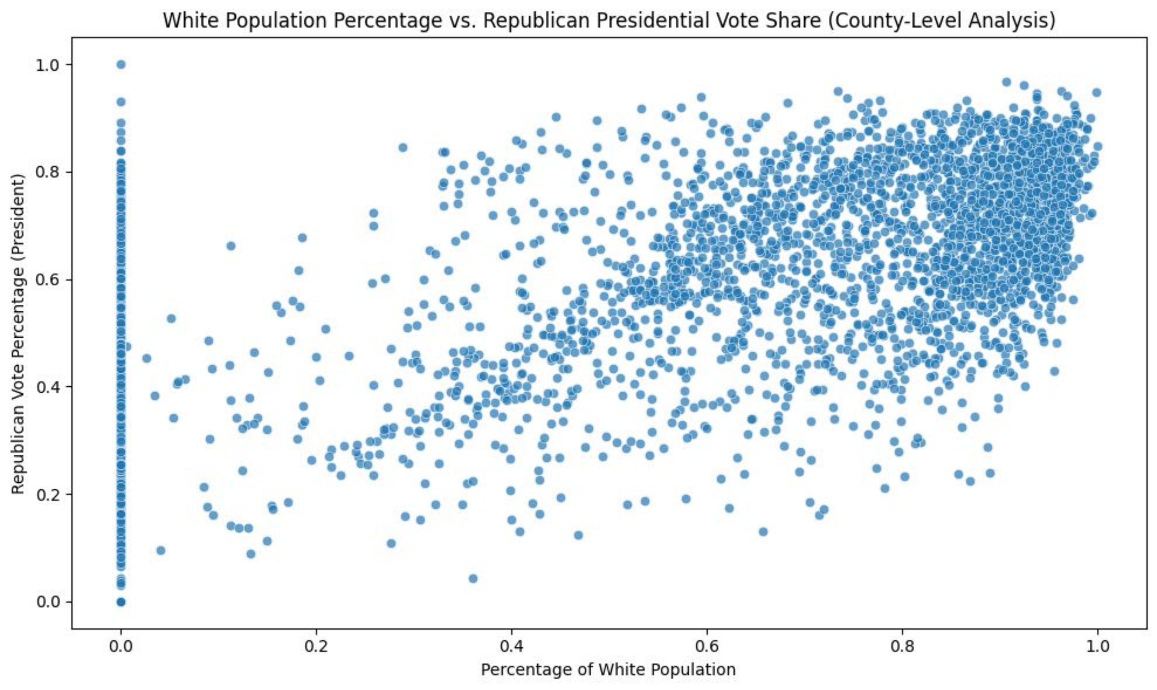
- High-income counties generally lean moderately Democratic, but the trend is not universal.

- Regional differences (state-wise coloring) indicate context-specific voting patterns.



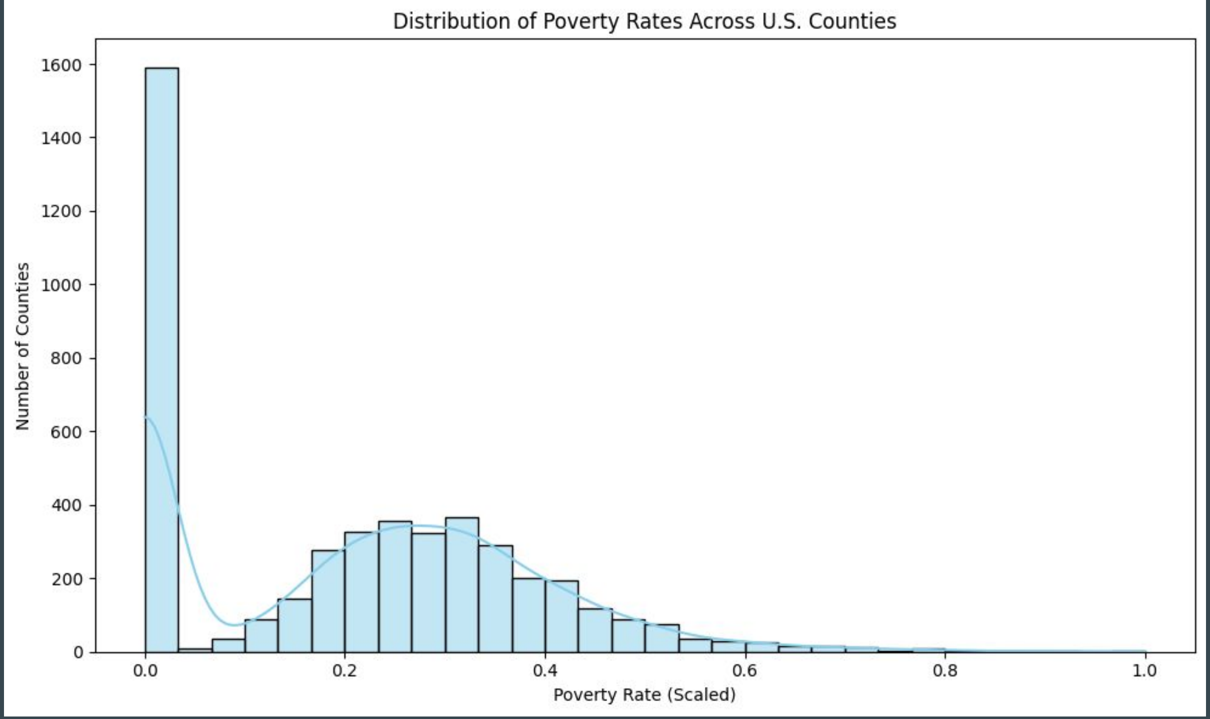
White Population Vs Republican Presidential Vote Share

- Counties with higher White population percentages tend to have higher Republican vote shares.
- There is a visible upward trend, especially as the White population exceeds 60%.
- Counties with lower White populations (below 20%) show more diverse voting behaviors.
- This pattern aligns with demographic voting theories, where Republican support tends to increase in predominantly White regions.

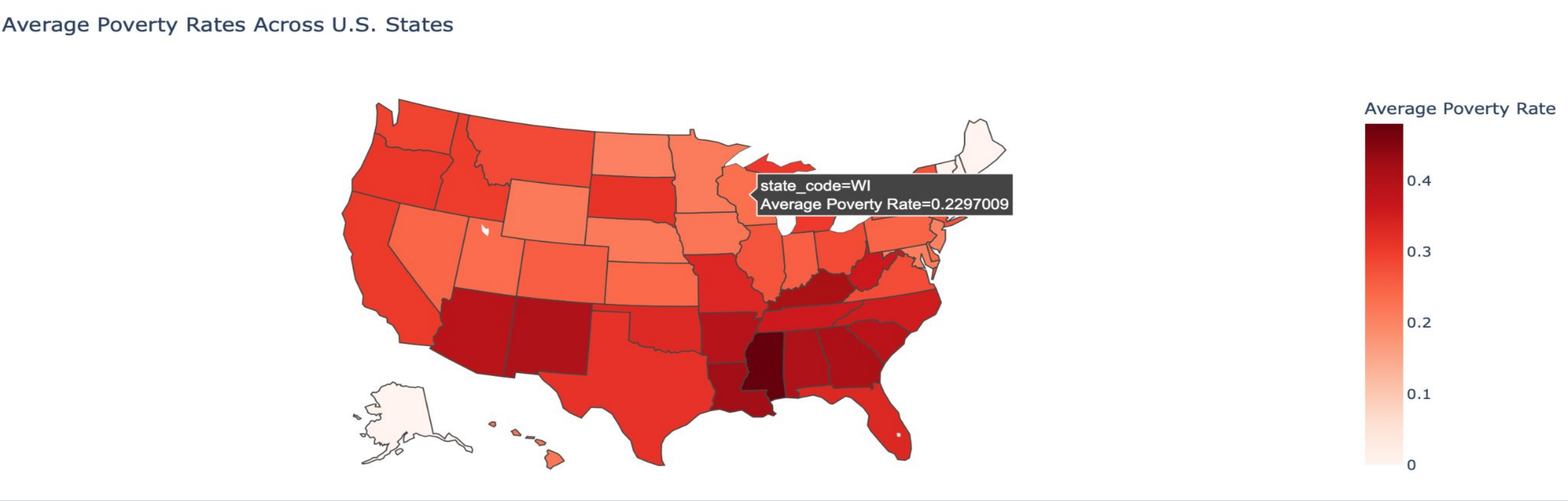


Distribution of Poverty Rates across US

- The distribution is right-skewed, meaning most counties have low to moderate poverty rates.
- A large number of counties (over 1500) report very low poverty rates - close to 0 on the scaled axis.
- Fewer counties experience high poverty rates, but those that do may represent economically vulnerable regions.
- The spread shows diversity in economic conditions across U.S. counties, which could impact voter priorities and behavior.



Geographic Distribution of Average Poverty rates across US


Average Poverty Rates Across U.S. States
state_code=WI
Average Poverty Rate=0.2297009
Average Poverty Rate
0.4
0.3
0.2
0.1
0

- Southern states like Mississippi, Louisiana, and Arkansas show the highest average poverty rates, highlighted by darker shades on the map.
- Northeastern states like Massachusetts, Connecticut, and New Jersey show lower average poverty rates, indicated by lighter shades.
- The map reveals regional economic disparities, which could influence political priorities and voter behavior across states.
- This spatial pattern supports the importance of considering geography when analyzing election outcomes.

Classification - Logistic Regression

Target:
Predict whether the **Democratic party wins (1)** or not (0)

Feature Selection:

- Dropped all vote-related columns
- Only socio-economic & demographic features used

Train-Test Split:

- 80-20 stratified split to preserve class balance

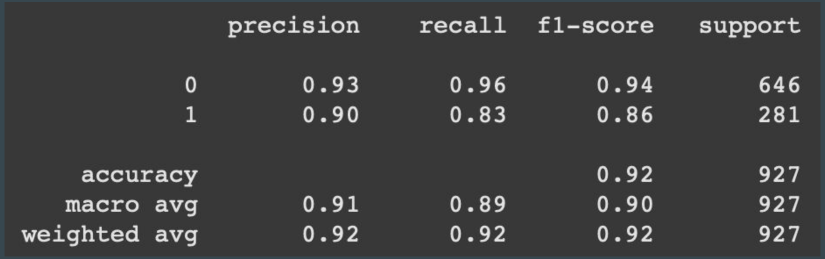
Model Used:

- **Logistic Regression** with max_iter=1000

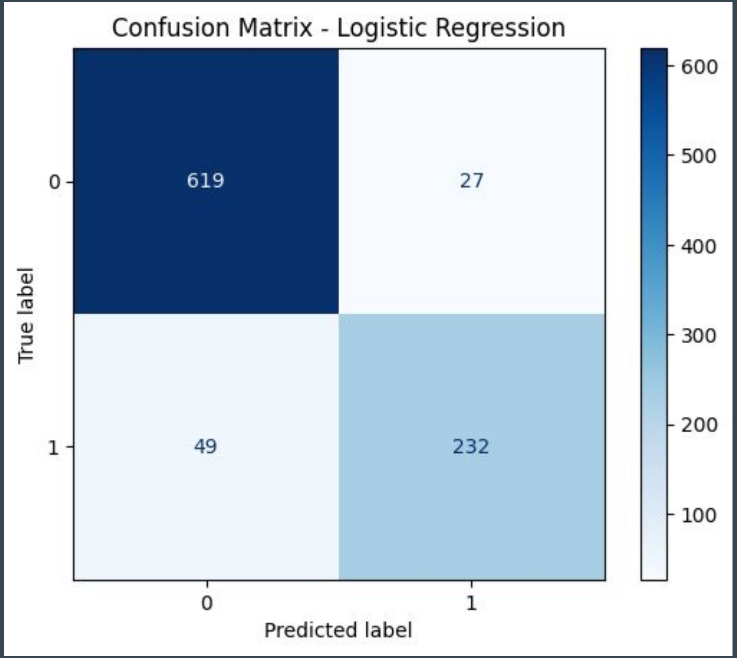
Model Evaluation

Classification Report & Confusion Matrix :

- 619 Republican counties, and 232 Democrat counties were classified properly by the model.
- The model incorrectly classified 27 counties (Districts) as Democrat, and it inaccurately classified 49 Democrat counties as Republican.
- Overall, the model achieved 92% accuracy; Democrat precision was measured at 90%, with recall returned at 83%.
- Overall these results indicate good performance, especially in clearly partisan areas; with a small amount of misclassifications in loosely contested areas.

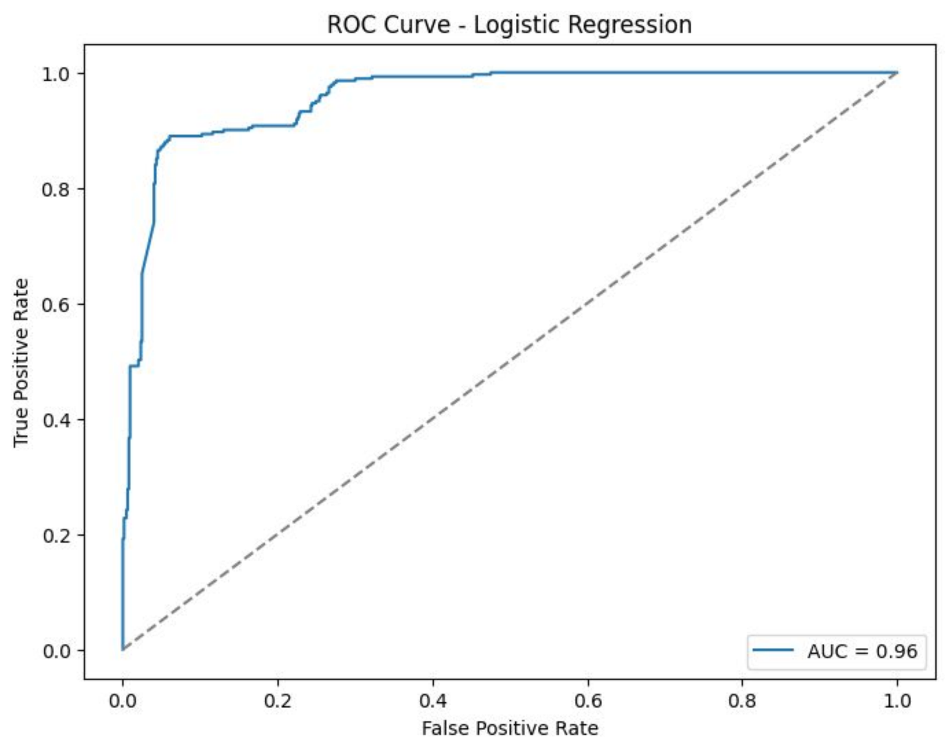


```
              precision    recall  f1-score   support

           0       0.93      0.96      0.94       646
           1       0.90      0.83      0.86       281

    accuracy                           0.92       927
   macro avg       0.91      0.89      0.90       927
weighted avg       0.92      0.92      0.92       927
```



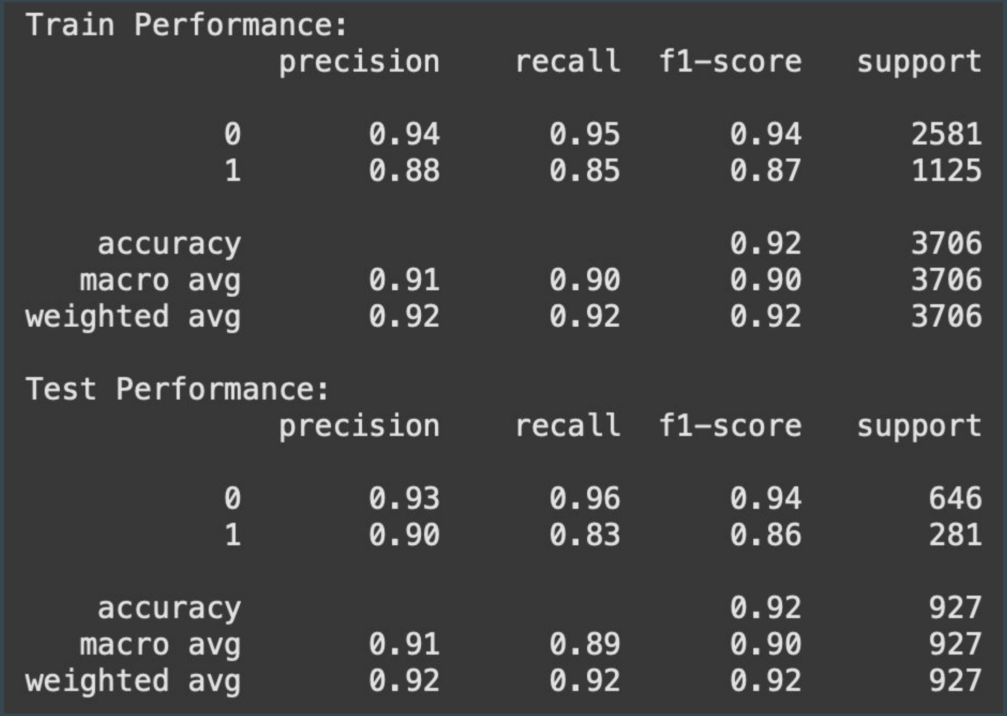
AUC-ROC Curve

- The AUC score is 0.96, which shows the model separates Democrat and Republican wins very effectively.
- The curve stays close to the top-left corner, indicating high sensitivity and specificity.
- This confirms the model is not making random guesses and performs reliably across different thresholds.



Overfitting Check

After looking at how the model performed in both the training and test sets, I notice that the accuracy and F1-scores are almost the same without a large dip in performance on unseen data. This indicates that the model is not at risk of overfitting and generalizes well. Because the model already attained high and stable performance in this task without hyperparameter tuning, I conclude this sufficient logistic regression model serves our purpose of our classification task in hand.



```
Train Performance:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94      2581
           1       0.88      0.85      0.87      1125

    accuracy                           0.92      3706
   macro avg       0.91      0.90      0.90      3706
weighted avg       0.92      0.92      0.92      3706

Test Performance:
              precision    recall  f1-score   support

           0       0.93      0.96      0.94       646
           1       0.90      0.83      0.86       281

    accuracy                           0.92       927
   macro avg       0.91      0.89      0.90       927
weighted avg       0.92      0.92      0.92       927
```

Feature Importance using Coefficients

Interpretation Based on Coefficients:

- Positive coefficients → Increase likelihood of **Democrat win**
- Negative coefficients → Increase likelihood of **Republican win**

Top Positive Influencers (Democrat):

- **Black population** , **Asian population** , **Professional jobs** , **Higher income per capita**

Top Negative Influencers (Republican):

- **White population** , **Construction jobs** , **DEM/REP senate %** (note: high vote % already indicates lean)

Key Insight:

- **Race** , **occupation** , and **income levels** are among the most predictive features of party preference.

Regression Task - Random Forest

Goal:
Predict the **percentage of votes** received by the **Democratic presidential candidate**
Initial Attempt – Linear Regression:

- Used as a baseline
- Underperformed, especially in swing or high-variance counties

Final Model – Random Forest Regressor:

- Handled non-linear patterns and feature interactions better
- Significantly improved accuracy over linear model

Target Variable:

- votes DEM president perc

Train-Test Split:

- 80-20 split using socio-economic and demographic features only

Model Evaluation (Before Tuning)

Model Used:

- Random Forest Regressor (default settings)

Evaluation Metrics:

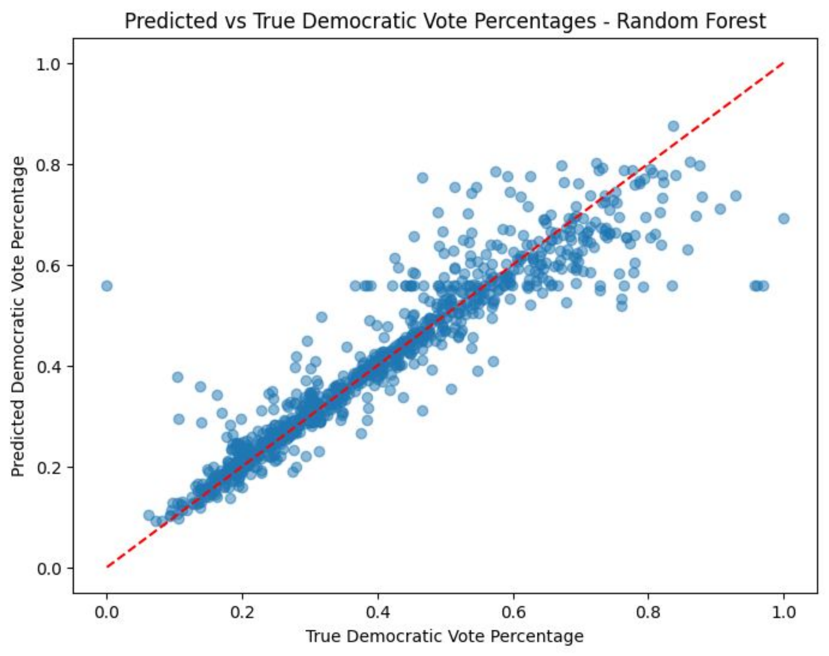
- **Mean Squared Error (MSE):** 0.0044
- **R² Score:** 0.8744 → Model explains ~87% of the variance in vote percentages

Performance Insight:

- Strong correlation between **true vs predicted** vote shares
- Slight underperformance in very high or low Democratic vote share counties

Baseline Summary:

- Solid fit without tuning
- Good foundation to improve further via hyperparameter optimization



Hyperparameter Tuning

Tuning Method:

- Used **RandomizedSearchCV** with 3-fold cross-validation
- Optimized for **R² score**

Parameters Tuned:

- n_estimators: 100, 200, 300
- max_depth: 10, 20, 30, None
- min_samples_split: 2, 5, 10
- max_features: 'sqrt', 'log2', None

Best Combination Found:

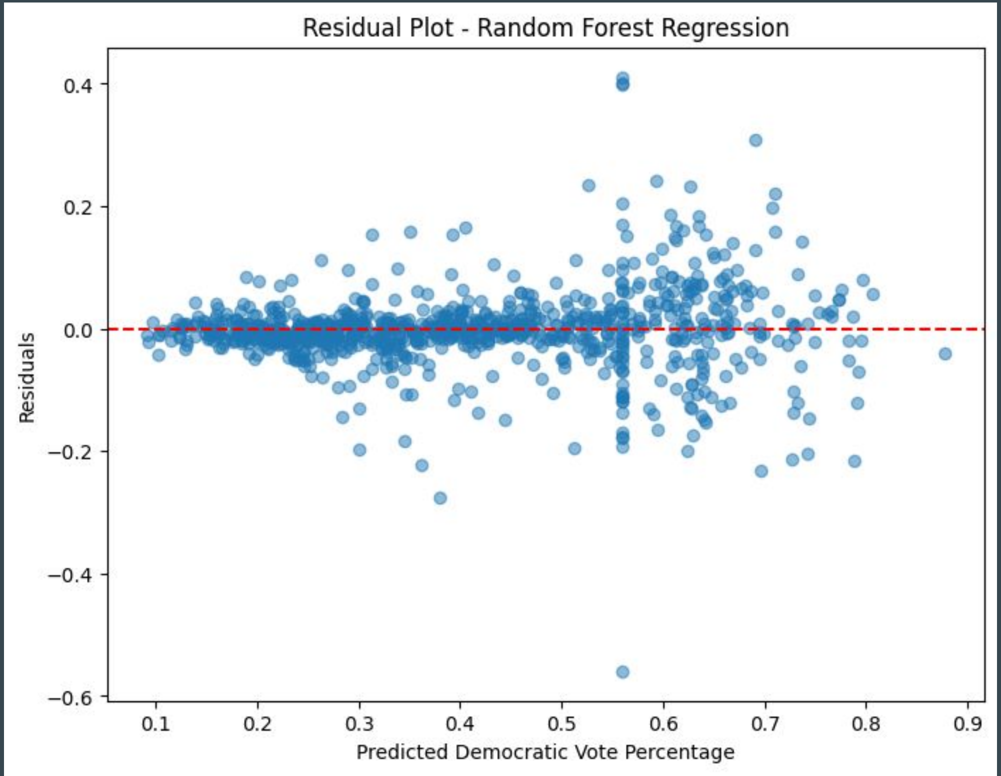
- n_estimators=100, max_depth=20, min_samples_split=2, max_features=None

Improved Performance:

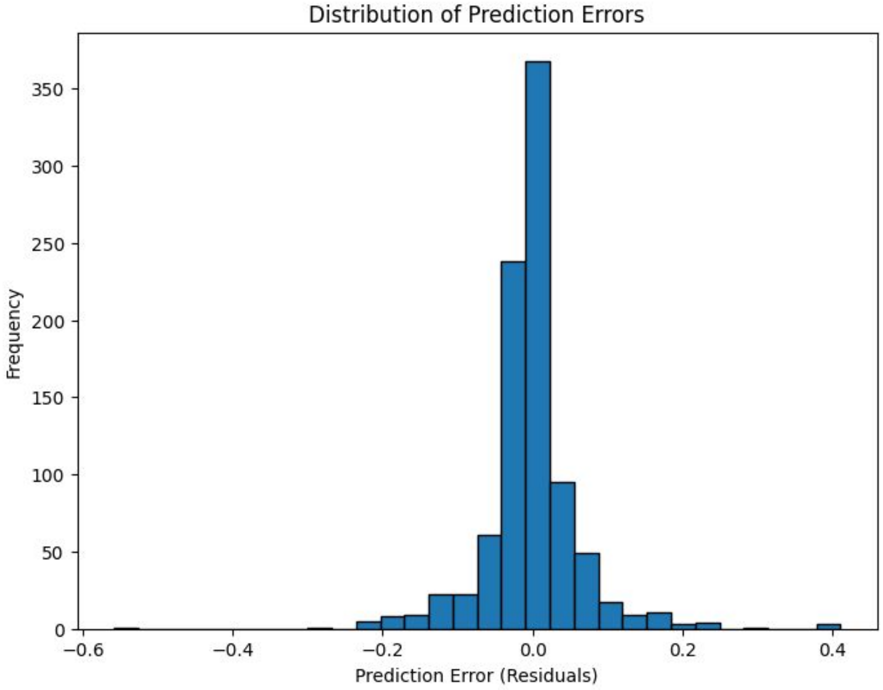
- **Mean Squared Error:** 0.0042
- **R² Score:** 0.8822 → Slight improvement over untuned model

Residual Analysis

I plotted the residuals to evaluate if there were any systematic errors in the model. I can see that the points are fairly randomly distributed around the zero line which is a good sign, albeit there was slightly more spread and variability when the predicted percentages were between 0.5 and 0.7, so this is the range the model struggles. Finally, I don't see any major patterns of overfitting or bias from the plot, but it indicates the models less confidence in the higher prediction ranges.

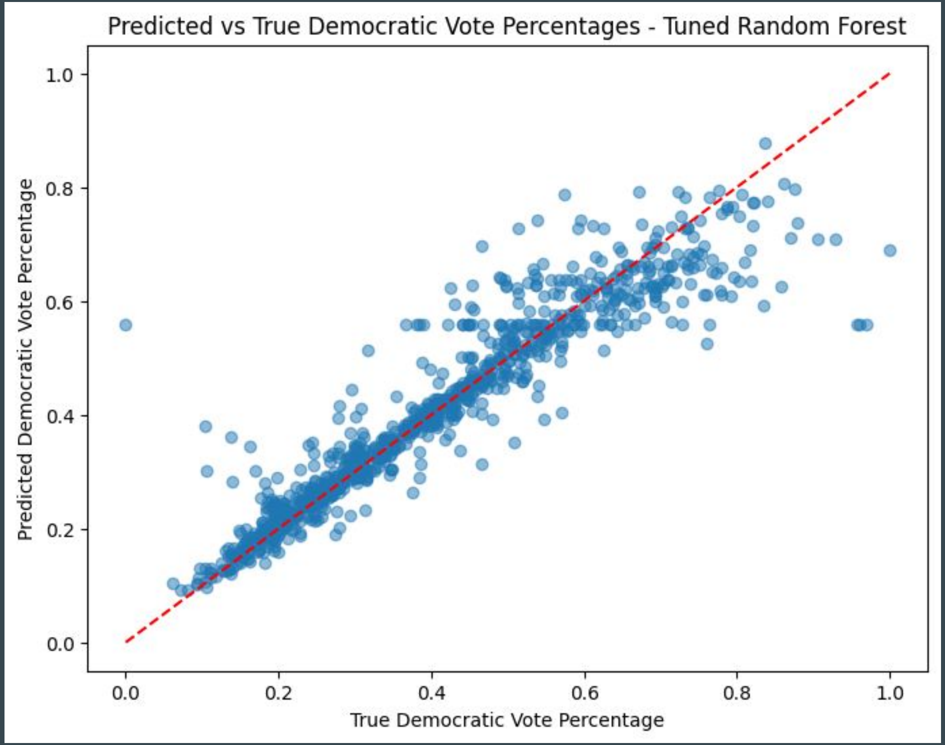


I graph the distribution of prediction errors to look for evidence that the model might be biased towards under- or over-predicting. The plot looks symmetrical and centered around zero which is a good sign. It tells me that the model's errors are balanced and not skewed in one direction or the other. Furthermore, the majority of the errors are small and fall within a narrow range which adds evidence to the case that the model is reliable.



Scatterplot for Tuned Model

I plotted out the predicted vs. true vote percentages before and after tuning, and the tuned model has a tighter spread (less spread) and increased density around the ideal diagonal line, particularly in the middle and especially the higher vote percentage ranges. This supports that the tuning improved model precision and, therefore, made the predicted values somewhat more consistent with the actual values. Though again, the improvement from tuning was not huge, the visual comparison does aid in supporting the small but constant performance improvement, as noted in the improvement of the R² score.



Feature Importance

Purpose

Identify which features most influence Democratic vote percentage prediction

Top Predictive Features

- Race-related variables (e.g., % White, % Black)
- Income and Employment indicators (e.g., IncomePerCap, Professional jobs)
- Commute patterns (e.g., Public transit, Walk to work)

Insights

- Areas with more diversity and higher income lean Democratic
- Commute types and job sector distributions also reflect urban vs rural patterns

Limitations and Biases

Data Imbalance:

- Republican and Democratic parties dominate most counties
- Minor parties excluded due to near-zero vote share

Missing Real-Time Sentiment:

- Social, political events and public opinion not captured in static demographic data

Model Assumptions:

- Assumes historical patterns repeat
- Ignores campaign dynamics, turnout variability, and third-party impact

Conclusion

Objective Achieved:
Successfully built classification and regression models to predict U.S. election outcomes using socio-economic and demographic data.
Model Highlights:

- **Classification Accuracy:** 92% using Logistic Regression
- **Regression R² Score:** 0.88 with tuned Random Forest Regressor

Key Insights:

- **Race**, **income**, and **employment sectors** are top predictors of voting behavior
- Urban areas and diverse communities show higher Democratic support

Takeaway:

- Data-driven modeling can explain and predict voting patterns with high accuracy
- Useful in identifying key swing regions and understanding electoral trends