ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
Department of Statistics

Classification-US Election

US Primary Elections Analysis

Author: Chalikias Konstantinos

2025

Contents

1 Introduction

The goal of this project is to develop a classification model using data from the 2016 U.S. primary elections. Specifically, a model that can predict whether Donald Trump received the majority of Republican votes in a given county.

Two datasets are used for this task. The first contains county-level voting results, including the total votes for each candidate in both the Republican and Democratic primaries. The second provides socioeconomic statistics for each county, such as population, racial composition, age distribution, income levels, and other relevant factors.

2 Exploratory analysis

Before building the classification model, it is important to examine the data in detail to better understand its structure, distribution, and potential patterns. There are a total of 2712 counties eligible for analysis after the cleaning process with 51 socioeconomic variables for each and the total votes for every candidate.

Out of the 35 million total votes in the 2016 primaries, 14.23 million were for the Democrat party and 20.76 million for the Republican.

Figure 1 presents a map with the most voted party per state. Blue for Democrat and red for Republican party

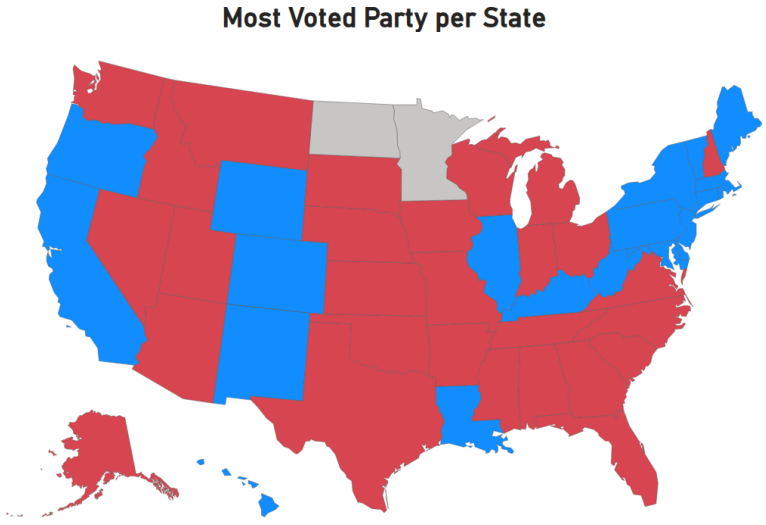

Figure 1: Most supported party in each state. (Democrat = blue, Republican = red)

Figure 2 presents the total votes that each candidate received.

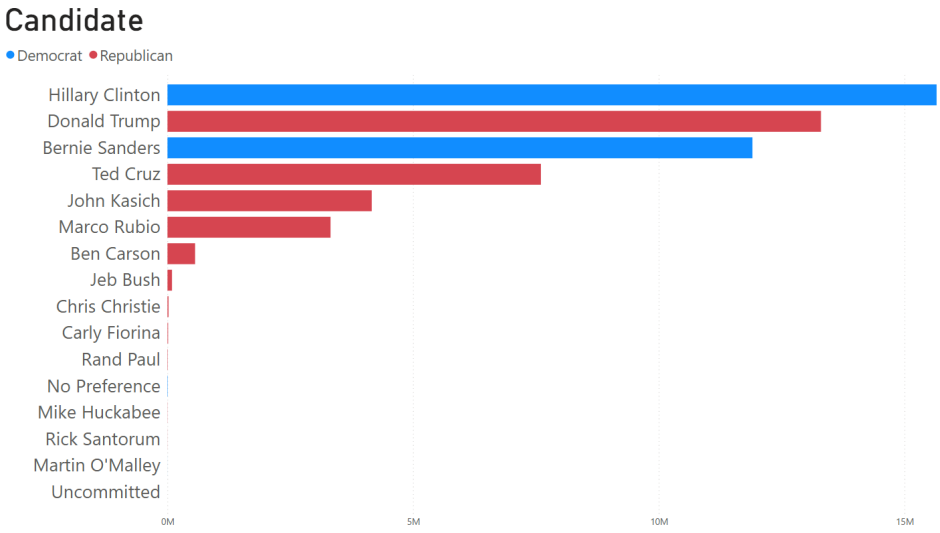

Figure 2: Total votes received by candidates in the Democratic and Republican primaries

Donald Trump and Hillary Clinton dominated the Republican and Democratic primaries, respectively. Other prominent candidates included Ted Cruz, John Kasich, and Marco Rubio in the Republican race, and Bernie Sanders in the Democratic race.

Socioeconomic Figures

The following figures present socioeconomic statistics.

Figure 3 shows the total number of votes cast in each state by income per capita and most voted party. States further to the right have higher per capita income, while states higher on the plot recorded more votes. The most voted party is distinguished by color. The figure suggests that wealthier states tend to have higher support for the Democratic Party.

Figure 4 shows the total number of votes cast in each state by education and most voted party. Education is measured as the percentage of the population over age 25 with a bachelor's degree or higher. States further to the right a higher share of educated residents, while states higher on the plot recorded more votes. The most voted party is distinguished by color. The figure indicates that states with higher education levels tend to have higher support for the Democratic Party.

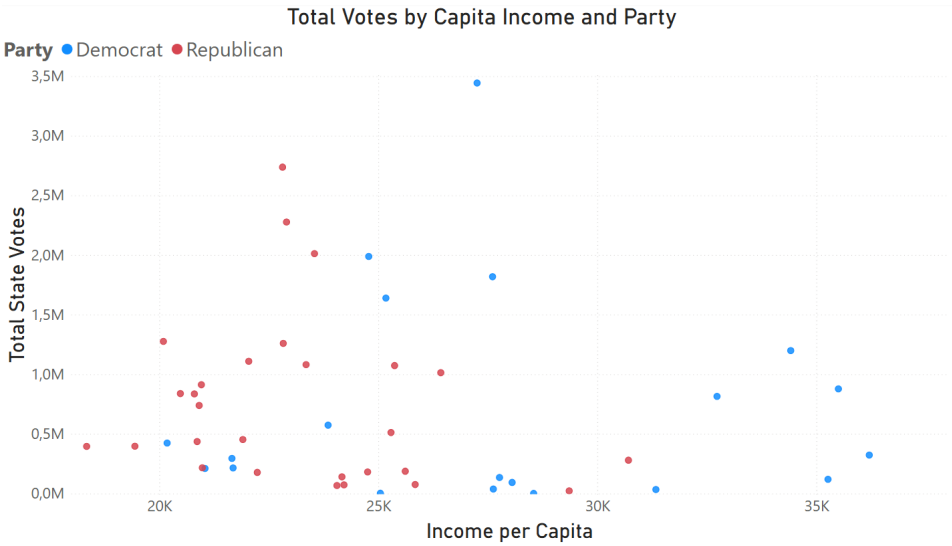

Figure 3: Votes in each state by income per capita and most supported party

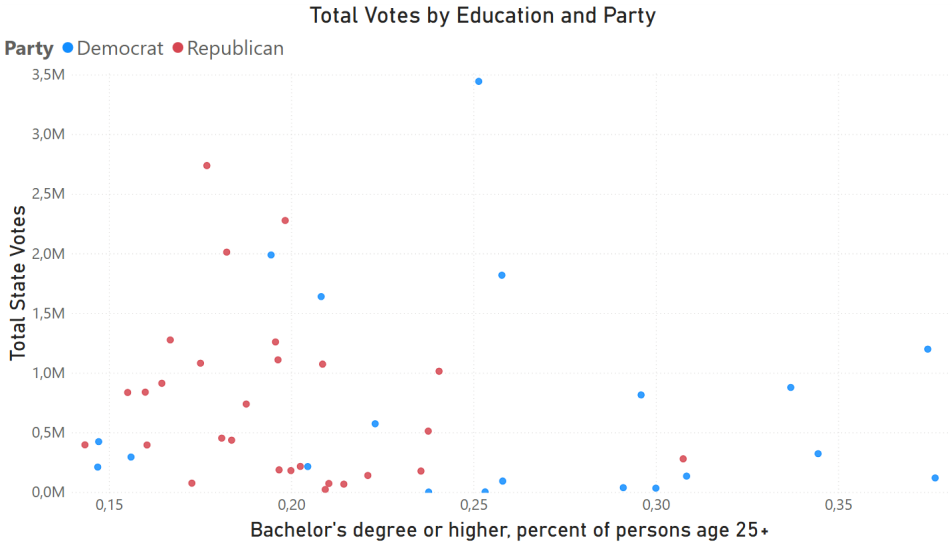

Figure 4: Votes in each state by education level and most supported party

Trump Figures

Since the goal of this analysis is to build a classification model to predict Trump's chances of winning a county, the following plots focus specifically on his performance.

Figure 5 presents a histogram of the fraction of votes Donald Trump received in each county. The distribution shows that, in most counties, Trump secured between 25% and 50% of the votes. However, there are also a number of counties where his support was substantially higher, ranging from 60% to 80%.

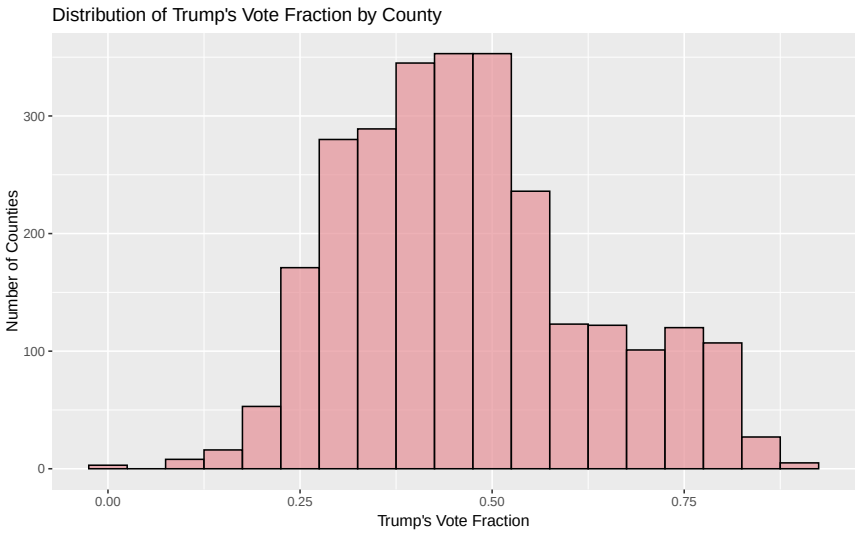

Figure 5: Histogram of Fraction of votes Donald Trump received in each county

Trump won 1351 out of the total 2712 counties, that is about 50%. The following Figure 6 presents the distribution of counties won by state.

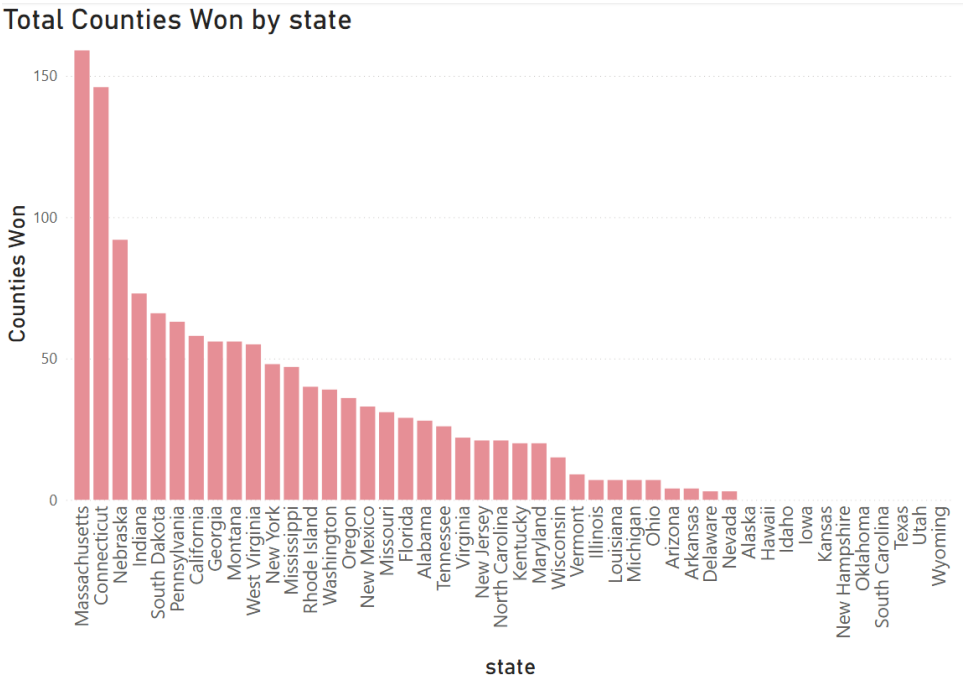

Figure 6: Number of counties won by Trump in each state

3 Classification models

The model is designed to predict whether Donald Trump received the majority of Republican votes in a given county. To achieve this, a new dummy variable will be created: it will take the value 1 if Trump received more than 50% of the votes in a county, and 0 otherwise. These two outcomes form the target classes for the classification model.

3.1 Linear Discriminant Analysis

The first method, Linear Discriminant Analysis (LDA), aims to find a linear combination of features that best separates the two classes. The objective is to maximize the variance between classes while minimizing the variance within each class. In other words, LDA seeks to place the two classes as far apart as possible, while keeping counties within the same class as close together as possible.

The end result of LDA is a linear classifier that assigns each county to one of the two classes. The decision boundary is centered at zero: counties with positive scores

are classified into the first class (Trump majority) and those with negative scores into the second class (no Trump majority).

Figure 7 presents a density plot of the linear classifier scores across all counties, with the actual outcomes distinguished by color. In an ideal scenario, the two distributions would not overlap at all. In practice, however, some overlap is observed in the plot, which indicates that the classes cannot be perfectly separated.

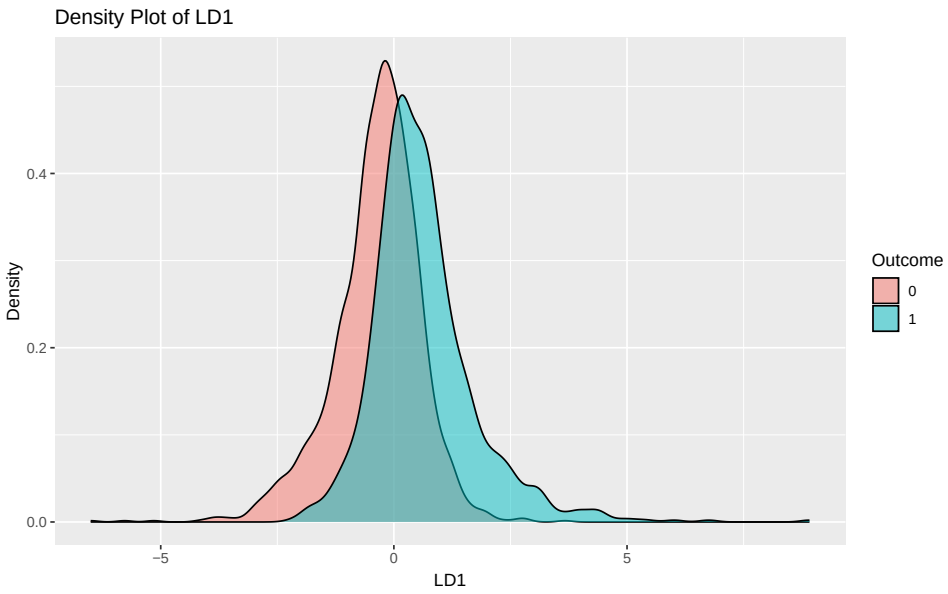

Figure 7: Density plot of LDA classifier scores by outcome class

An experiment was conducted to evaluate and compare the performance of the models. The full dataset was split into training and testing subsets, with 80% of the data used for training and 20% reserved for testing. Model performance was measured by accuracy, defined as the percentage of correct predictions on the test set. This will be repeated 1000 times and the result is the average accuracy for all models.

In this case, the model achieved an accuracy of approximately 71%. Since a random guess would be expected to yield around 50% accuracy, this result represents a clear improvement.

3.2 Generalized Linear Model

The next model is a regression-based approach that assumes a binomial distribution for the target variable. It estimates the probability that a county belongs to each class by

incorporating its socioeconomic and demographic characteristics as predictors. Each county is then assigned to the class with the highest predicted probability.

Using the same evaluation procedure as with the LDA model, the Generalized Linear Model achieved an accuracy of about 71.1%. This represents a slight numerical improvement, but not a substantial difference in performance.

3.2.1 Generalized Linear Model Improvement

In an effort to improve the previous model, instead of using all 51 socioeconomic variables available, some will be removed. Variables were iteratively removed using the Akaike Information Criterion (AIC) as a guide. AIC, an estimator of prediction error, was minimized by sequentially eliminating variables until no further improvement could be achieved.

The final model retained 33 variables with the AIC improved from 2980 to 2956. The model's accuracy increased slightly to 71.6%, which represents only a marginal improvement and not a meaningful difference in performance.

3.3 K-nn Nearest Neighbors

The idea behind this algorithm is that counties with similar characteristics are likely to belong to the same class. Each county is therefore assigned to the class most common among its k nearest neighbors—the counties that are most similar to it.

A critical step in applying this method is choosing the value of k. If k is too small, the model becomes highly sensitive to noise, as a single outlier neighbor can strongly influence the classification. If k is too large, the model may become too generalized, averaging over many neighbors and potentially ignoring local structure in the data. The optimal value of k is determined by testing different values through cross-validation and selecting the one that yields the best accuracy.

Figure 8 compares model accuracy for values of k ranging from 1 to 50. Accuracy is relatively low for very small k, but increases steadily as k grows. The model reaches its peak performance for values between k=10 and k=18, where accuracy stabilizes around 72.5%. Beyond k=18, accuracy gradually declines. This suggests that any k within the 10–18 range would be a reasonable choice.

Final model, using k = 15 achieved an accuracy of approximately 72.5%. Again, a slight numerical improvement but not a meaningful difference in performance.

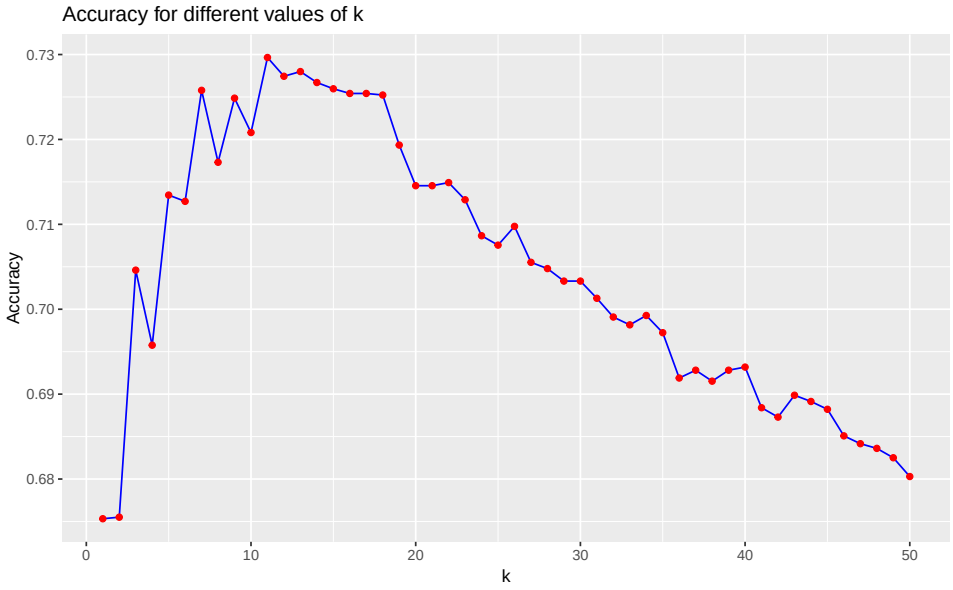

Figure 8: Model accuracy for different values of k in the Knn model

3.4 Final Model Comparison

The results for all models are summarized in Table 1. The differences in performance across models are minimal and not statistically significant. Any of the models can be considered acceptable, as all perform substantially better than random guessing.

Table 1: Model Accuracy Results

Model	**Accuracy**
LDA	0.7093
GLM (full)	0.7114
GLM (Improved)	0.7142
K-NN	0.7252

4 Conclusion

This project set out to develop a classification model to predict whether Donald Trump received the majority of Republican primary votes in U.S. counties using voting results and socioeconomic data. Several classification models were applied, including Linear Discriminant Analysis (LDA), Generalized Linear Models (GLM), and k-Nearest

Neighbors (KNN).

The results show that all models achieved comparable accuracy, around 71–72%, which is a clear improvement over random guessing. While KNN provided a slight numerical advantage, the differences across models were not statistically significant. This suggests that the underlying patterns in the data are being captured consistently, regardless of the specific modeling approach.

An important takeaway is that socioeconomic and demographic characteristics are indeed informative predictors of county-level voting behavior in the Republican primaries. However, the relatively modest accuracy also highlights the complexity of electoral outcomes and the influence of factors not included in the dataset, such as campaign dynamics, media coverage, or local political issues.