Abstract

The objective of this project was to determine whether socioeconomic or demographic variables can be used to predict state-level presidential elections outcomes. To accomplish this, we used two models, a random forest classification model, and a linear regression model. We used data from various government data sources (Census, FRED, LAUS, BLS, and iPUMS), and voting data from the MIT Election Data and Science Lab.

In the random forest classification model, we predicted whether the state voted Democrat or Republican. Our results showed that the three most important predictors were voter participation, unemployment rate, and median income, with an accuracy score of 83.3%.

In comparison, we made two multiple linear regression models. One predicted the percentage of Democrat votes, while the other predicted the percentage of Republican votes. The most important predictors in these models were median income, unemployment rate, and percentage of the population aged 65 and over. The Democrat model had an r-squared value of 0.445, while the Republican model had an r-squared value of 0.403.

Our findings indicate that the strongest predictors of how a state will vote are median income and unemployment rate.

Introduction

Our aim is to create a model to try and predict US Presidential election results based on educational attainment, demographic factors of voting participation rates, age, unemployment, and median income. To help us answer our question and build out our model, we sourced data from the US Census, the Integrated Public Use Microdata Series (iPUMS), the Local Area Unemployment Statistics (LAUS), the Federal Reserve Economic Data (FRED), and The Bureau of Labor Statistics (BLS), along with MIT Election Data and Science Lab as our primary sources. We will be creating this model at the national level, which will cover the timeframe of 2000-2024 (seven US presidential elections). We compiled these data sources together to have one overall dataset to use for our data analysis and machine learning models. Data cleaning consisted of sorting our data by year and by state, filtering our data to include only data from 2000-2024, and also creating proportional datasets where necessary.

Creating predictive models and analyzing factors that impact elections are important because they track trends in education levels and socioeconomic factors in the US, which can have significant implications for the country's health and development. In addition, studying US presidential elections presents us with an opportunity in identifying trends in this data that could help us better understand voting behavior and electoral preferences. For these reasons, we will look to build a model that can utilize this data to help us understand which factors have the most impact on voting in US Presidential elections.

Theoretical Background

This project utilized two models to predict election outcomes: a random tree classification model, and a linear regression model. Both of these models use machine learning algorithms to estimate the relationships between an outcome and a series of explanatory variables.

The explanatory variables considered in our analysis were:

- The percentage of residents who had received a bachelor's degree
- The median household income of the population
- The percentage of the population aged 18-24
- The percentage of the population aged 25-44
- The percentage of the population aged 45-64
- The percentage of the population aged 65+
- The percentage of voter participation
- The median age of the population
- The unemployment rate of the population

Unemployment rate and median income were chosen as measures of a state's economic prosperity and were included to show how well a state is doing economically could affect the outcomes of their votes. The age groups and median age were chosen to get a sense of how different age groups would vote, as current media talks about younger voters leaning more Democrat while older voters tend to lean more Republican instead. For these reasons, we thought this would be a good way to estimate these effects. Voter participation was chosen as it has been reported in recent elections that this has been very low, and people have blamed the lack of motivation to vote as part of the reason for certain recent election outcomes. Examining whether this has changed, and whether this change directly impacts one party's chances or the other would be interesting. And finally, we included the percentage of residents who had received a bachelor's degree, as attending college is often ascribed to a Democratic tilt.

Our random forest model is a classification model that creates a multitude of decision trees through bootstrap aggregating, in which the training algorithm repeatedly selects a random sample with replacement from a training set, and fits tree to those samples. These decision trees are used to sort the data into one of two classifications. In the case of our project, the two classifications are a Democrat win or a Republican win. This random forest model was used for two purposes:

- To serve as a predictive model for future election results.

- To estimate the importance of an explanatory variable on the outcome of the election through its gini score.

The random forest model was chosen for use in this analysis because it does a good job of handling nonlinear relationships and interactions that are common in demographic variables, as well as helping to reduce the influence of multicollinearity on the relative importance of the predictor variables.

Our linear regression model estimates the relationship between the percentage of votes in a state population that were Democrat or Republican and nine different explanatory variables. The relationships are modeled using linear predictor functions whose parameters are estimated from the data. This linear regression model was used for two purposes:

- To serve as a predictive model for future election results, selecting variables which showed a statistically significant effect on predicting the percentage of votes.
- To quantify the strength of the relationships between the percentage of votes and the nine explanatory variables we picked.

The linear regression model was chosen for use in this analysis because it helps interpret the relative effects of the predictor variables, helping us to better understand the influence these variables have on the election outcomes.

Methodology

We took the nine factors mentioned earlier and gathered them into datasets sorted by year and state. The predictor factor datasets (containing the nine demographic and economic factors) were compiled from ACS, BLS, LAUS, FRED, and iPUMS. The voting data was sourced from MIT Election Data and Science Lab. These datasets were then merged with a dataset containing voting data from 2000 to 2024, only including years where national elections took place.

Using this merged dataframe, we created two models to try and predict election results, and to see which of these factors had the most impact on predicting election outcomes.

The first model we considered was a random forest model (supervised), using a dummy variable for Democratic victory. The factors above were used as predictors for the outcome of the state elections. For the national level model, we split our data into a training dataset and a testing dataset, with an 80-20 split. Fitting our model to the training dataset, we then predicted the testing dataset and checked the results of our model with an accuracy score.

After creating a random forest model at the national level, we also wanted to investigate which predictors were the most useful for various swing states, as those are the least predictable. The swing states we chose to look at were Wisconsin, Pennsylvania, Ohio, Michigan, and Florida.

The second model we considered was a linear regression model, predicting either the percentage of Democrat votes, or the percentage of Republican votes. We first created some single predictor linear regression models, predicting percentage of Democrat and Republican votes with median income, unemployment rate, voter participation, and two of the age group percentages (18-24 and 65+).

Next, to get a sense of how these predictors interacted with each other, and to determine which had the greatest effect on predicting the outcomes, we created two multiple linear regression models, one predicting Democrat voter percentage and the other predicting Republican voter percentage using all the factors we listed above. From these models, we removed any statistically insignificant predictors (predictors with p-values less than 0.05), to create reduced models. We also removed the population age percentage predictors for populations 25-44 and populations 45-64, as all the population age percentage predictors were highly correlated together.

We created two reduced multiple linear models. The Democrat model used the predictors unemployment rate, median income, percentage of the population with a

bachelor's or higher, percentage of the population aged 18-24, percentage of the population aged 65+, and the voter participation. The Republican model used the same predictors, except for percentage of the population with a bachelor's or higher.

With these reduced multiple linear models, we scaled the coefficients so we could determine which predictor had the greatest effect on the voter percentages. This was done by scaling each predictor to a mean of 0 and standard deviation of 1 and then using those scaled predictors to fit a multiple linear regression model.

The final comparison we did was to create multiple linear regression models that only used data from the five swing states we mentioned above. After creating the initial models, we created reduced models by removing any variables that were not statistically significant and keeping only the two population age percentage variables we used in the other reduced models.

For both the Democrat and Republican reduced swing state models, the predictors used were unemployment rate, percentage of the population aged 18-24, and percentage of the population aged 65+.

Computational Results

This section presents all computational outputs generated in the analysis pipeline.

Exploratory Data Analysis

Correlation matrices were computed to assess linear relationships between all ten predictor variables and the Democratic and Republican vote percentages. The predictors included unemployment rate, median income, median age, bachelor's attainment, voter participation, voting-age population, and population proportions for the 18–24, 25–44, 45–64, and 65+ age groups. Pairwise scatterplots were generated, illustrating the bivariate relationships between each numerical predictor and both voting-percentage outcomes, with each visualization including a regression line and the univariate distributions of both variables. Boxplots were generated for both proportion-based and continuous predictors to assess distributional characteristics and potential outliers.

Random Forest Classification Models

A national Random Forest classifier was trained using 200 estimators and a maximum depth of 9, with performance calculated on a 72-sample test set.

For the Republican class ("False"), the model achieved a precision of 0.800, a recall of 0.889, and an F1-score of 0.842 (36 cases). For the Democratic class ("True"), the model yielded a precision of 0.875, a recall of 0.778, and an F1-score of 0.824 (36 cases). The model achieved an overall accuracy of 0.833.

State-level models were also developed for classification. In Pennsylvania (N=2), the model produced an accuracy of 0.500. Similarly, the combined five-state model (Wisconsin, Pennsylvania, Ohio, Michigan, and Florida, N=6) also resulted in an overall accuracy of 0.500.

Ordinary Least Squares (OLS) Regression Models

Single Input Regressions

Single-variable OLS regressions were estimated for each predictor against Democratic and Republican vote percentages.

Outcome Variable	Predictor	R2	Coefficient	p−value	RMSE	MAE
Democratic Votes	Median Income	0.172	$3.702\times10-6$	<0.001	0.107	0.077
Republican Votes	Median Income	0.209	$-4.036\times10-6$	<0.001	0.103	0.075
Democratic Votes	Unemployment Rate	0.076	1.710	<0.001	0.113	0.087
Republican Votes	Unemployment Rate	0.029	−1.046	0.001	0.114	0.089
Democratic Votes	Voter Participation	0.103	0.634	<0.001	0.111	0.084
Republican Votes	Voter Participation	0.096	−0.606	<0.001	0.110	0.085
Democratic Votes	Pop 65+ %	0.009	−0.327	0.081	0.117	0.091

Republican Votes	Pop 65+ %	0.000	0.057	0.761	0.116	0.090
Democratic Votes	Pop 18-24 %	0.048	-1.8223	<0.001	0.114	0.083
Republican Votes	Pop 18-24 %	0.059	2.0025	<0.001	0.113	0.083

Multiple National Regression Models

Full national OLS models included all ten predictors (N=306 observations). Reduced national OLS models were developed using variable selection procedures (N=357 observations).

Model	Target	R2	Adj. R-squared	MAE
Full Model	Democratic Votes	0.445	0.426	0.0588
Full Model	Republican Votes	0.403	0.383	0.0618
Reduced Model	Democratic Votes	0.390	0.380	0.0649

Reduced Model	Republican Votes	0.324	0.314	0.0685

Predictors in the final reduced Democratic model included unemployment rate, median income, bachelor's attainment, the 18–24 and 65+ population proportions, and voter participation. The final reduced Republican model retained unemployment rate, median income, the 18–24 and 65+ population proportions, and voter participation.

Standardized coefficients were estimated after scaling all predictors to assess relative impact. In the Democratic reduced model, coefficients ranged from **−0.031 to 0.051**. In the Republican reduced model, coefficients ranged from **−0.047 to 0.029**.

Model	Target	R2	Adj. R-squared	MAE
Scaled Model	Democratic Votes	0.656	0.391	0.015
Scaled Model	Republican Votes	0.691	0.454	0.013

Democrat Predictors	Dem Scaled Model Coefficients
unemployment_rate_normalized	0.0401
median_income_normalized	0.0507
bach_per_normalized	-0.0200
pop_65_per_normalized	-0.0307
voter_participation_normalized	0.0191
pop_18_24_per_normalized	-0.0283

Republican Predictors	Rep Scaled Model Coefficients
Unemployment Rate Normalized	-0.0193
Median Income Normalized	-0.0466
65+ Population percentage Normalized	0.0290

Voter Participation Normalized	-0.0177
18-24 Population Percentage Normalized	0.0290

Swing-State Regression Models

OLS models were estimated using data from five swing states (N=24 for full models; N=28 for reduced models).

Model	Target	R2	Adj. R-squared	MAE
Full Model	Democratic Votes	0.656	0.391	0.015
Full Model	Republican Votes	0.691	0.454	0.013
Reduced Model	Democratic Votes	0.556	0.501	0.028
Reduced Model	Republican Votes	0.238	0.143	0.034

The reduced Democratic swing-state model retained unemployment rate and the 18–24 and 65+ proportions. The reduced Republican swing-state model retained only the 18–24 proportion.

Discussion

This section will cover the discussion and analysis of our results.

Exploratory Data Analysis

In our exploratory analysis, our correlation heatmap and pair plots helped us understand the strength of the relationships between our predictor variables and our dependent variables of Democrat and Republican voting percentages. For Democrat voting percentage, we saw that the strongest correlations were with median income and voter participation with correlations of 0.42 and 0.32, respectively. For Republican voting percentage, the strongest correlations remained the same, with median income and voter participation rates showing correlations at –0.46 and –0.31 respectively. These results very closely reflect each other, and this would make sense since we are measuring the percentage of votes for one party or the other.

Our boxplot analysis also revealed many high and low outliers in all our predictor variables except for bachelor's degree percentage. These outliers in combination with the correlational relationships mentioned earlier were the main drivers for us concluding that our data was suitable for further modelling through random forest classification and linear regression models.

Random Forest Classification Models

At the national level, we saw that our random forest model had an accuracy score of 83.33%, which we believe to be fairly strong given the scope of our analysis. Our model was more precise when predicting a Democrat victory at 87.5% compared to Republican victory at 80%. Overall, this machine learning model was quite effective overall at attempting to predict election results based on the data here.

However, our swing state and individual state level models both had accuracy scores of 50%, indicating that both models were essentially guessing at which factors were most important in predicting election results. We found that focusing specifically on individual states, even in our swing state group, reduced the total amount of data that we could work from, which made our model less effective due to not having as much information to work with. In this case, we did not consider either the swing state or individual random forest models to be effective at predicting election results.

Ordinary Least Squares (OLS) Regression Models

Our OLS regression models focus on single input regression models, along with national and swing state multiple regression models. Our national multiple regression models also include reduced models.

Single Input Regressions

Our single input regression models investigated median income, unemployment rate, voter participation, and the 18-24 voting population percentage and 65+ voting population percentages. All our predictors were statistically significant in these models except for 65+ voting population percentage for Republican voters. The mean absolute errors for these variables were all within a range around 7-9%, with our root mean squared error being in a range of about 10-12%. Of these single variables, the Republican median income predictor showed the lowest RMSE and MAE at 10.3% and 7.5%, respectively. In context, these error ranges are quite large and do not give a strong conclusion in regard to our original question. This result, however, is somewhat expected since we are assessing single-input variables.

Multiple Regression Models

Our multiple linear regression model had an R-squared value of .445, indicating a moderate strength between our model and Democrat voting percentage. The moderate strength can be partly described as this analysis not being a completely comprehensive list of predictors for Democrat voting percentage. Some of these factors that might have allowed us to get a more complete picture but were not included were data on government assistance programs, and possibly Medicare data as well.

Voting age population was the only factor to show up as not statistically significant in our model.

Our multiple linear regression model produced an error rate between approximately +-5-7% from the true voting percentage total for each group on average, while our reduced models fell more within +-6-7% from the true voting percentage total for each group on average.

Our multiple regression analyses also optimized our residual plots in both the full and reduced models.

For our scaled models, we found that for Democrat voting percentage, holding all other variables constant, we would expect to see:

- For unemployment rate, a 4.01% increase in democratic voting percentage.
- For median income, a 5.07% increase in democratic voting percentage. This is the strongest predictor for our Democrat model.
- For the percentage of bachelor degrees attained, a 2.00% decrease in democratic voting percentage.
- For the 65+ voting population, a 3.07% increase in democratic voting percentage.
- For voter participation, a 1.91% increase in democratic voting percentage.
- For the 18-24 voting population, a 2.83% decrease in democratic voting percentage.

And for Republican voting percentage, holding all other variables constant, we would expect to see:

- For unemployment rate, a 1.93% decrease in Republican voting percentage.
- For median income, a 4.66% decrease in Republican voting percentage. This is the strongest predictor for our Republican model.
- For the 65+ voting population, a 2.90% increase in Republican voting percentage.
- For voter participation, a 1.77% decrease in Republican voting percentage.
- For the 18-24 voting population, a 2.90% increase in Republican voting percentage.

Overall, in these models, median income had the greatest impact on both Democrats and Republicans, with Democrats seeing a 5.07% percentage point increase and Republicans seeing a 4.66% percentage point decrease in percentage of votes.

Swing-State Regression Models

In our Swing-State Regression model, the only factor that showed up as statistically significant was unemployment rate for Democratic voting percentage, while unemployment rate and all our population categories were statistically significant for Republican voting percentage. The Republican model having more statistically significant variables also led to it having a higher R-squared value of 0.691 compared to the Democrat model of 0.656.

Our reduced models for both Republican and Democrat voting percentages produced worse R-Squared and mean absolute errors, with the Democrat model's reduced R-Squared dropping by 0.0992 and the Republican model's dropping by 0.4529, reflecting

that the linear regression model in this specific case was not as effective of a model as we would have expected.

Conclusions

This project explored how to create an effective model at predicting US Presidential elections from 2000-2024 based on numerous key socioeconomic variables. The results of our predictive model produced the following conclusions and takeaways:

- Median income was the strongest factor that impacted voting percentage changes for both Democrat and Republican voting percentages in our data.
- Our National random forest model was able to predict election results with an 83% accuracy based on the numerical predictors in our model.
- Unemployment rate, median income, bachelor's degree percentage, all population age ranges percentages (18-24, 25-44, 45-64, 65+), and voter participation all had statistically significant results in our dataset.
- Our multiple linear regression model produced an error rate between approximately +-5-7% from the true voting percentage total for each group on average.
- Our swing state models both showed that unemployment rate was statistically significant, with it being the only statistically significant predictor for the Democrat model, while the Republican model also included the population percentages.

Overall, we believe that this model was quite successful in predicting US Presidential election results and offered key insights into socioeconomic factors that are worth tracking and monitoring elections.

References

Bureau, US Census. “Voting and Registration in the Election of November 2024.” *Census.Gov*, US Census, 24 Apr. 2025, www.census.gov/data/tables/time-series/demo/voting-and-registration/p20-587.html.

Bureau, US Census. “CPS Historical Time Series Tables.” *Census.Gov*, US Census, 25 Aug. 2025, www.census.gov/data/tables/time-series/demo/educational-attainment/cps-historical-time-series.html.

Dataverse, Harvard. “1976-2020-President.TAB - U.S. Presidential Elections.” *Harvard Dataverse*, 2025, dataverse.harvard.edu/file.xhtml?fileId=10244938&version=8.0.

Download center, StatsAmerica. *StatsAmerica Download Center*, 2025, www.statsamerica.org/downloads/default.aspx.

Economic Data, Federal Reserve. “Federal Reserve Economic Data.” *MEDIAN HH INCOME BY STATE 1984 -2023 (Updated Periodically) - FRED\ALFRED - St. Louis Fed*, FRED, 2025, fredaccount.stlouisfed.org/public/datalist/8534/.

MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2024", https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V16, UNF:6:NKTy7eW9uEWX4imXpPxf5g== [fileUNF]

Statista, Statista. “Educational Attainment Distribution in the United States from 1960 to 2022 .” *Statista*, 2025, www.statista.com/statistics/184260/educational-attainment-in-the-us/.