Cyrus Luong

Data Bootcamp

Professor Koehler

Final Project Write-up

May 12, 2025

**Project Introduction**

From 2016-2020, two contentious presidential elections and the COVID-19 pandemic highlighted various divisions between different demographics in the US. These demographic divisions manifested themselves economically, socially, and, especially amidst the backdrop of the contentious 2020 presidential election, politically. Analyzing and understanding the correlations between various demographic factors and voting patterns allows for the best comprehension and interpretation of future electoral outcomes and political dynamics amongst different demographic groups.

This final analysis seeks to build the best predictive regression model to predict a county’s political vote swing from 2016-2020 based on its demographic data. By analyzing which demographic factors are the best predictors of a county’s vote swing, this project will be able to extract the most salient features for better political analysis. Additionally, by utilizing three different regression models in the final vote swing analysis, the best analysis can be extracted from the three, as well as determine which model is best for potential future analysis.

**2020 Demographics Dataset Description**

To source the dataset, publicly available data on economic, political, and diversity status were collected and compiled. The dataset is primarily composed of numeric features, aside from a county’s state and name. For political data, the dataset contains voting data by county, tracking raw vote totals as well as individual vote percentage for the Democratic and Republican candidates in the 2016 and 2020 elections. The percentage data for each election was used to calculate each county’s vote swing from 2016 to 2020, which is the target variable for the predictive analysis.

For diversity data, the dataset contains the racial and gender breakdown of each county by percentages, tracking data for the percentage of men, women, Hispanic, White, Black, Native American, Asian, and Pacific Islander in each county. Additionally, the dataset also tracks the percentage of citizens that are of voting age by county.

Economic data is also well documented for each county, allowing for data on average household income, unemployment and poverty levels, as well as the details surrounding proportions of types of work, commute types and lengths, and sectors of work. For the data on income, the dataset tracks income and income error, income and income error per capita, poverty and child poverty levels, and the employment and unemployment rates. For occupation data, the dataset tracks percentages in professional occupations, the service industry, office jobs, construction, production, as well as work in the public and private sectors, or family work. It also tracks modes of commute, whether by driving, carpooling, public transportation, walking, other forms of transportation, or remote work, as well as the mean commute time.

Finally, the dataset covers COVID-19 data by county, tracking each county’s total population, COVID cases, and COVID deaths. The data on deaths, cases, and population were also used to calculate COVID deaths and COVID cases per capita, as well as the mortality rate in each county.

**Regression Models and Methods**

This analysis will utilize three different regression models to calculate the predicted vote swings from 2016-2020 by county, using the demographic data highlighted in the above section. The three regression models utilized will be linear regression, K-nearest neighbors (KNN), and neural networks.

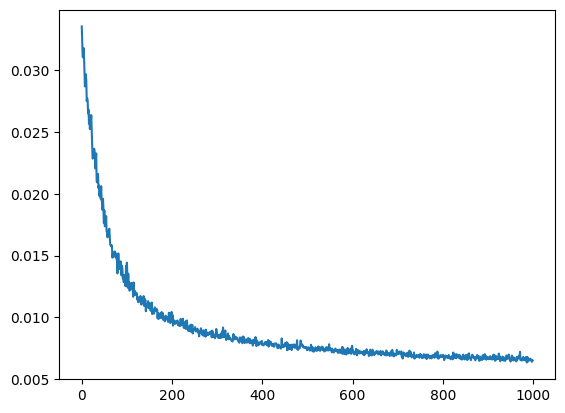
With the dataset, there is inconsistency with the reporting of percentages by feature, with some features like political vote share reported in decimal form, while features like racial diversity are reported in percentage form. To remedy this issue, before inputting the data into each model, the numeric data is scaled using sklearn’s StandardScaler function, normalized around a set mean and standard deviation.

To select the features that would be analyzed as the X values in the regression model analysis, a correlation heatmap was utilized from the exploratory data analysis. Features that had an absolute value of correlation higher than 0.15 were chosen to be the analyzed features, meaning highest correlation to a swing towards either Joe Biden or Donald Trump in 2020. There are 18 features in total that fit this criteria, which are the following: professional career, income per capita, income level, Joe Biden’s 2020 vote percentage, Donald Trump’s 2016 and 2020 vote totals and 2020 vote percentage, Asian, office jobs, latitude, production and construction jobs, poverty and child poverty levels, Hispanic, income error, and COVID deaths per capita. After choosing the features to be inputted into training the model, a train test split is made, with a test size of 50% to prevent overfitting the model on the training data.

To measure the accuracy of each model, each is assessed on its mean squared error from the actual vote swings by county. To create a baseline mean squared error score, a “zeroes list” that guesses no vote swing for each county is used for baseline MSE calculations. This “zeroes list” baseline returns a mean squared error of 0.00366, meaning a mean percentage error of about 6% through just guessing zero percent for every vote swing.

In the linear regression model, each of the 18 features were fed into a pipeline containing the linear regression model, the standard scaler, and sklearn’s PolynomialFeatures, tested on various degrees. This permits additional features to be fed into the model consisting of all of the possible combinations of multiplying each feature by each other, allowing for further complexity and achieving a lower mean squared error for the model.

In the k-nearest neighbors model, the main parameter to be changed is the number of neighbors to calculate the average from. To select the highest performing number of neighbors, different parameters were tested to find the number of neighbors that yielded the lowest MSE from its predictions.

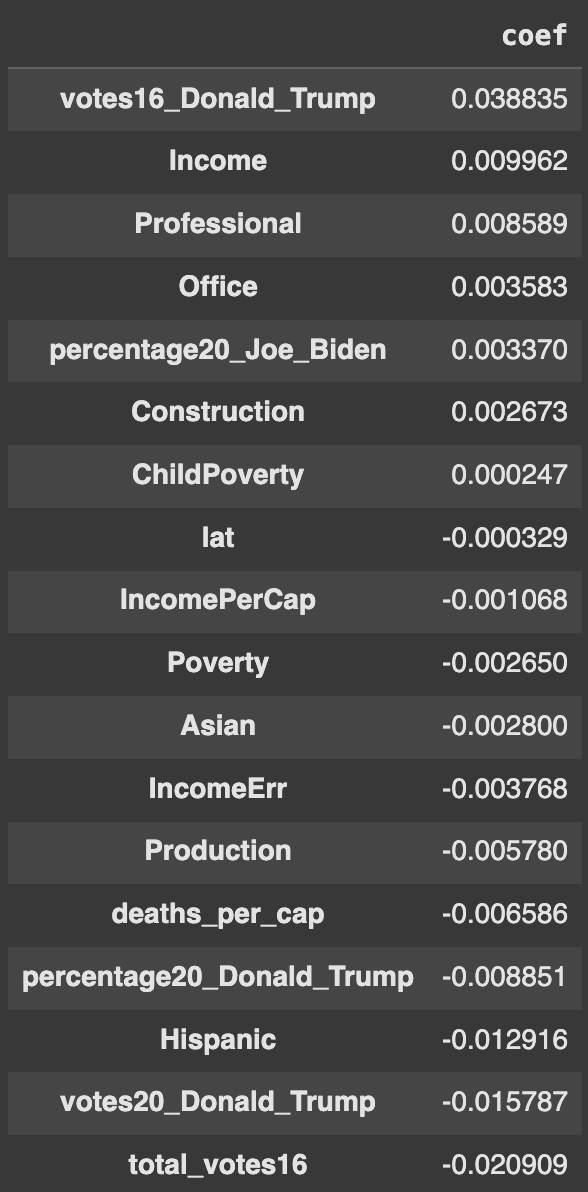
For the neural network model, multiple layers of neurons and models are used to try and use features to scale predictions to be accurate to the true results. First, the original 18 features are inputted into the first hidden layer, with an output of 300 features. There are ReLU activations in between, as well as a DropOut function that prevents overfitting the model by randomly zeroing out neurons while training the neural network. This process repeats with another hidden layer turning the 300 features into 150, and finally another layer turning those 150 features into one final output. This neural network is then trained over 1000 epochs, where every epoch represents a slight decrease in the “loss” of the model, which is measured using mean squared error. Although there are some slight up-and-down abnormalities when it comes to the losses over the 1000 epochs due to the randomness of the DropOut function, there is an overall downward trend in the losses over the training. This trend can be seen in the figure below, measuring the overall mean squared error over the 1000 epochs.

**Regression Model Results and Interpretation**

After comparing the mean squared errors of each model, the most accurate one that emerged was the k-nearest neighbors model, tuned to use an n\_neighbors variable of 3. This provided the lowest mean squared error when testing the model on the test data, with an MSE of 0.00188, or an error of about 4%, about 2% better than the baseline prediction.

In comparison, the linear regression model, using just one polynomial variable, saw the best performance out of the linear regression models, with a test mean squared error of 0.00281, or about a 5% error, 1% better than the baseline. The neural network model, after 1000 epochs, reached a mean squared error of 0.00626, about an 8%, or 2% higher than the baseline.

The reason why this particular dataset may have been more suited towards an analysis using a k-nearest neighbors model rather than a linear regression or a neural network model is due to its more clustered and nonlinear nature. These specific features, which measure economic and racial groups, tend to vote in clusters, and so a k-nearest neighbors model that can average these clusters would perform better than other models. Additionally, a k-nearest neighbors model only requires an average of surrounding data points in each cluster. This means that the model is agnostic of the shape of the dataset, allowing for the analysis of non-linear data sets like the 2020 demographics.

This nonlinearity in the dataset is also why the linear regression and neural network models may not have performed as well. The linear regression model, like its name suggests, operates under the assumption that all of the data analyzed follows a linear relationship. With a nonlinear dataset like the one analyzed, the linear regression model, as well as the neural network model that utilized linear hidden layers, was unable to capture the complexities of the dataset’s shape. Additionally, outliers on either side of the vote swing dataset most likely skewed the model, as some counties, particularly those with a high concentration of certain demographics, swung far above the average to the Democrats or the Republicans.

Moving beyond analyzing each regression model’s accuracy, we can also analyze the correlations of each of the features to a vote swing. Using the linear regression model set to just one polynomial feature allows for the interpretation of the coefficients of each individual feature input into the model, a list of which can be seen in the figure on the right. By extracting these coefficients, we can observe which features have the highest effect on a swing towards Joe Biden (positive coefficient) or a swing towards Donald Trump (negative coefficient).

Disregarding the features that are related to vote share and vote count, some of the highest coefficients towards a positive swing to Joe Biden included a higher income level, as well as professional and office jobs. On the other hand, some of the highest coefficients towards a negative swing to Donald Trump included Hispanics, Asians, and jobs in production. These line up with the general trends seen in the 2020 election as well as each candidate’s political appeal and overall campaign tone.

Joe Biden, with his 2020 campaign, presented a calm, steady status quo message, which was appealing to those seeking more grounded leadership after four years of the first Trump administration. This group of people tended to skew higher income and white-collar, reflected in the aforementioned positive coefficients indicating a vote swing to Joe Biden.

Donald Trump, with his 2020 campaign, tapped into racial and economic divides with a populist message similar to his 2016 campaign. The COVID-19 pandemic had worsened and highlighted these divides, and those facing economic hardships were attracted towards his populist messaging. This included many minorities, like Asians and Hispanics, or those in blue-collar jobs like production jobs, seen in the negative coefficients of features that indicated a swing towards Trump. By analyzing the coefficients of each of these features, the demographics that saw the highest swings towards either party could be targeted by political advertising or groups for future electoral success.

**Conclusion and Next Steps**

Using the most effective regression model, the k-nearest neighbors model, as well as the supplemental models, key takeaways are able to be interpreted surrounding the connections between certain demographic variables and a county’s political swing from the 2016 to the 2020 election. The first key takeaway from the regression analysis is the most effective type of model for examining this kind of data in the future, the k-nearest neighbors model, which allows for analysis of data that is non-linear and in clusters, much like the characteristics of a demographic dataset. Another key takeaway beyond the type of model is how to isolate variables that have the greatest impact on political dynamics and electoral outcomes.

With these two takeaways, when demographic data is available surrounding the 2024 presidential election, further analysis can be conducted to analyze which features affected vote share and vote swing the most in that election. This could be particularly useful to the Democratic party after their loss in the 2024 election, allowing them to pinpoint the demographics or regions of the country with which they have slipped the most. By pinpointing these weaknesses in their political support, they can categorically target these groups using advertising or targeted messaging to achieve a political comeback by the next election.