**2024 Presidential Election Prediction Project**

**IE 490 ML Final Report**

**Problem Statement:**

Predicting the voting tendencies of states in US presidential elections is a complex task with significant implications for political strategy and governance. The president of the United States is tasked with a very challenging job of leading not just country, but in many ways the world as well. Their policies, leadership and executive actions play a major role in the economic, political and military policies. Based on preliminary research, we can also see that companies are also paying more attention to the election outcomes in recent years. From the US Chamber of Commerce, we can see an increase the mentioning of policy risk from 10-K filings, a required document that every company must file with the SEC. As more companies are becoming sensitive to shifts in policy and how that would affect their daily operations, it becomes even more crucial to be able to accurately predict presidential elections, and hence the corresponding policies associated with each party.

A graph with a line and red line

Description automatically generated

*Chart 1: Biplot of the Factors in the 2020 US County Description Dataset.*

The hypothesis posits that by analyzing historical voting patterns and demographic factors, it is possible to forecast whether a state will vote red (Republican) or blue (Democratic) in upcoming elections. This hypothesis is based on the assumption that past voting behaviors, coupled with demographic characteristics, contain predictive signals for future election outcomes. The objective is to develop a predictive model that accurately classifies states' political affiliations. Addressing this issue would yield benefits to decision making in both governmental and private endeavors by addressing public policy uncertainty associated with presidential elections, as well as promoting stability and enhancing discourse around elections . In order to address this challenge, four machine learning models were trained and assessed through misclassification error rates. Misclassification was selected as the primary metric for success due to the nature of the problem as a binary classification case. The remainder of this report is organized as follows: the Data Sources section outlines key characteristics of the training and validation datasets for the machine learning models, as well as preprocessing; the Goals and Methods section outlines the sub-objectives in achieving the overall goal of county political classification, as well as methods used in generating predictions; the results section covers the model selection process, county predictions, and key insights from the training process; and lastly, the Future Work section covers extensions to the proposed model and avenues for future research.

**Data Sources:**

For this project, four datasets of 47 features were created corresponding to the 2012, 2016, 2020, and 2024 election cycles. Each dataset consists of 3113 observations corresponding to the counties in the US. Data was sourced from the US Census Bureau’s Five-Year American Community Survey (ACS) and Economic Census (EC) leveraging the “tidycensus” and “censusapi” APIs within R Studio. A summary of the features for the datasets is shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| County name | Mean travel time to work | Percentage of homeowner-occupied units | Asian alone | American Indian and Alaskan Native-owned firms |
| State name | Median household income | Median value of owner-occupied units | Native Hawaiian or Pacific Islander alone | Asian-owned firms |
| Total number of households | Per capita income | Total population | Two or more races | Native Hawaiian or Pacific Islander owned firms |
| Average household size | Percentage of individuals below the poverty level | Female persons, percentage | Hispanic or Latino | Hispanic-owned firms |
| High school education or more attained, percentage | Percentage of housing units | Persons under 5 years, percentage | White alone, not hispanic or latino | Women-owned firms |
| Bachelors degree education or higher attained, percentage | Percentage of two-unit housing structures | Persons over 18 years, percentage | Private nonfarm establishments | Retail sales |
| Civilian Veterans, percentage | Percentage of three or four unit housing structures | Persons over 65, percentage | Private nonfarm employment | Accommodation and Food sales |
| Homeowners with desidence in the same house 1 year and over | Percentage of 5-9 unit housing structures | White alone | Nonemployer establishments | Population density |
| Foreign born persons, percentage | Percentage of 10-19 unit housing structures | Black or African American alone | Total number of firms | Election Outcome |
| Individuals with a language other than english spoken at home, percentage | Percentage of 20+ unit housing structures | Alaskan native or American Indian alone | Number of Black-owned firms |  |

*Table 1: Summary of predictors for election outcomes.*

The data-scraping process revealed two core areas of concern that would need to be addressed through preprocessing. Firstly, there were several instances of missing values in the datasets. These missing values arose for two main reasons. The first was related to sampling limitations of the Census Bureau's surveys. Since all the surveys are conducted on subsets of the US population in each year, there were instances where data was not available for rural counties during all of the election cycle years. In these cases, data was substituted from the earliest available preceding year. The second source of missing values stemmed from sampling procedures for the US Economic Census. Economic Census data from after 2017 is organized into core-based statistical areas (CBSAs). CBSAs are clusters of economically linked counties. Due to the structure of the dataset for years after 2017, data needed to be disaggregated into county-level observations. In these cases, the number of observations for the CBSAs were divided among their constituent counties following a population-weighted scheme. For counties outside CBSAs, data was substituted from the earliest available, preceding year; furthermore based on observations from the cleaning process, it was determined that data from counties in Alaska would be excluded from the analysis and final prediction model due to a lack of recorded information from that state.

The second area of concern was that Census Bureau datasets are limited in the availability of data for certain years. Data on population demographics for the ACS is only supported until 2022, and the EC is only available between 2017 and 2022. In order to address this issue, the 2016 dataset, and the 2020 and 2024 datasets were generated using data closest to the associated elections. Another area of note is that the EC API calls occasionally return “NA” values for the number of businesses owned by minority groups in certain counties. It is unclear whether these missing values are due to a lack of information in the surveys or whether they are meant as zero values. As such, they are assumed to be zeros for the purposes of the modeling.

All in all, each dataset includes one categorical feature in the form of the state of the county, and 46 numeric features in the population parameter. The datasets also include the results of the associated presidential elections for the recorded years. County-level election results were sourced from the [2008-2016 Election Results](https://github.com/tonmcg/US_County_Level_Election_Results_08-20/blob/master/2020_US_County_Level_Presidential_Results.csv) dataset, which despite the repository’s naming also includes results from the 2020 election cycle. Counties were classified as Republican or Democrat according to the recorded observations for the percentages of GOP and Democrat votes in each county to produce a categorical output variable. The 2012-2016 datasets will be leveraged for training purposes, and the 2020 set will be reserved for validation.

**Goals and Methods**

Our main goal for this project will be to predict which party a state will win in 2024 based on the predictors in the dataset. Three sub-objectives will be used to guide progression towards the overall goal:

1. Distributions will be characterized for each predictor variable. These distributions will be used to inform model design strategy and provide insights into the underlying dynamics of county classification. Data validation will also be conducted to identify correlations between predictors as well as evaluate outliers & high leverage points in the dataset.
2. Based on the results from sub-objective 1, multiple models will be trained to classify observations as Red or Blue. The number and type of models used will be dependent on whether the associated model assumptions for each classification model can be justified. Models will then be assessed based on precision, specificity, overall classification accuracy, and other relevant metrics.
3. Variable importance analysis will be conducted to identify core features for county classification with the various models. This importance analysis will be conducted using multiple methods for validation, and confidence thresholds of 95% will be used for p-value based approaches.

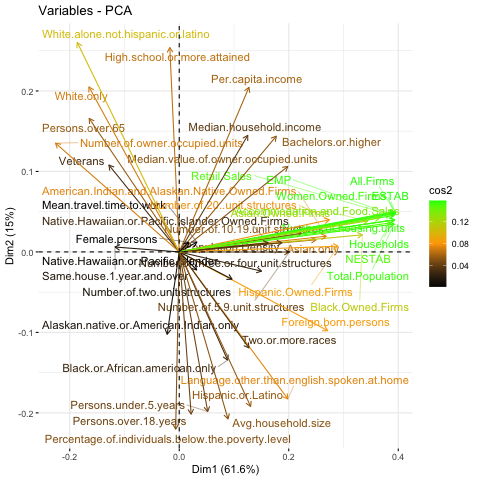
To ensure the reliability of our analysis, the initial phase of our methodology involves rigorous data cleaning. We will handle missing values through removal, verify consistency in county representation by checking for duplicates and standardizing names, address discrepancies through outlier detection and consistency checks, and ensure data completeness by verifying all relevant variables are present and addressing missing data. These steps will guarantee that our dataset is robust and ready for subsequent analysis, laying the foundation for accurate modeling and interpretation of results.

As part of our predictive modeling approach, we will first employ Principal Component Analysis (PCA) to reduce the dimensionality of our dataset and extract the most important information from our predictor variables. By transforming the original set of predictor variables into a new set of orthogonal variables called principal components, PCA will enable us to capture the underlying patterns and relationships within our data while minimizing information loss. We will then look at feature distributions and look for any multicollinearity present in our large list of predictors. The results of the latter defined our selected models which were: Ridge Regression, Lasso Regression, Polynomial SVM, and Radial SVM. Selected for their ability to handle multicollinearity of our parameters that proved to have an effect on classification.

We trained all our models with 2016 data, using cross-validation to select the best-fitting parameters for each.

**Results**

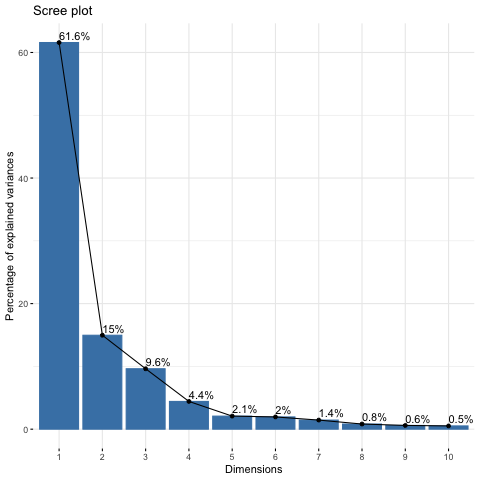
Given the large number of numerical features in the dataset, it was determined that some form of dimensionality reduction would be required before beginning the analysis. Principal component analysis was conducted on the numeric features to assess their significance. PCA was conducted on the predictors in the 2020 County feature dataset and all the various descriptors in it because missing value analysis and cleaning is still in progress for the 2012 and 2016 datasets. In Figure 1, we can see a Biplot of all the factors that are in the County description dataset.



*Figure 1: Biplot of the Factors in the 2020 US County Description Dataset.*

As can be seen from the biplot, most of the features have large magnitudes when compared to the first two principal components, implying that the selected features all contribute significant information. There are only two features which do not appear to contribute much variability to the dataset, and those are the percentage of female persons in each county and the mean travel time to work. Special attention will be payed to these features during the preliminary stages of model training to determine if these low-variance factors play a significant role in prediction accuracy, or whether they should be removed to reduce model complexity.

After plotting the first two principal components, the proportion of variance explained by the principle components was plotted to determine the appropriate number of components to use in training. In Figure 2, we can see the Proportion of Explained Variance for each dimension being utilized in the function. By using the elbow rule, it can be concluded that after five dimensions, the PEV slowly flattens out, implying that this is how many principal components should be used to minimize model complexity while preserving the variance of the original dataset.

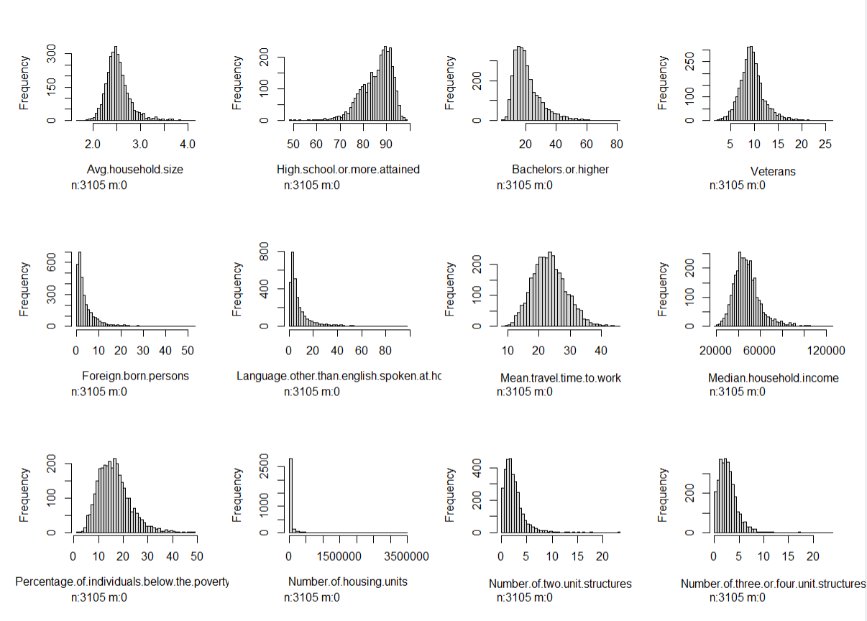


*Figure 2: Proportion of Explained Variance vs Dimensions Graph*

The next thing we worked on was looking at the distribution of each parameter/predictor in our dataset to visualize their characterization. As shown in Figure 3, we can see most of the predictors tend to follow normal or exponential distributions, which means they are valid for being considered in the models we would be testing. Up next we created a covariance matrix to compare each variable’s relationship with each other, as shown in Figure 4. We saw a good number of predictors with strong direct relationships with each other, especially data regarding economic factors. These predictors (such as number of black-owned firms, number of women-owned firms, etc.) tended to have a stronger correlation with each other, as shown in the dark blue circles on the bottom right of the covariance plot. There were a handful that also had inverse relationships, such as number of occupied units and number of structures or number of homeowners, which made sense from the mechanistic and societal point of view. All in all, there was evidence of multicollinearity from the covariance matrix. Thus, we decided on using the following four models in our analysis:

* Lasso Regression
* Ridge Regression
* Polynomial SVM
* Radial SVM

These models were selected due to their effectiveness in addressing multicollinearity, feature reduction, and high-dimensional datasets. At the beginning of the project, it was hypothesized that the ridge regression model would struggle the most due to difficulty in fully eliminating features; as such, it was selected as the baseline for comparison.



*Figure 3: Snapshot of predictor distributions*

The four models were trained on the dataset corresponding to the 2016 election cycle and assessed with the 2020 election as the validation set. To test and compare the performances, we conducted a plot comparing the true positive rate to the false positive rate on the validation set. This plot is shown below in Figure 5. The Lasso and Ridge Regression models both exhibited nearly identical, strong performance. Polynomial SVM also performed well, but Radial SVM had the poorest results from the initial testing.

A screen shot of a grid

Description automatically generated

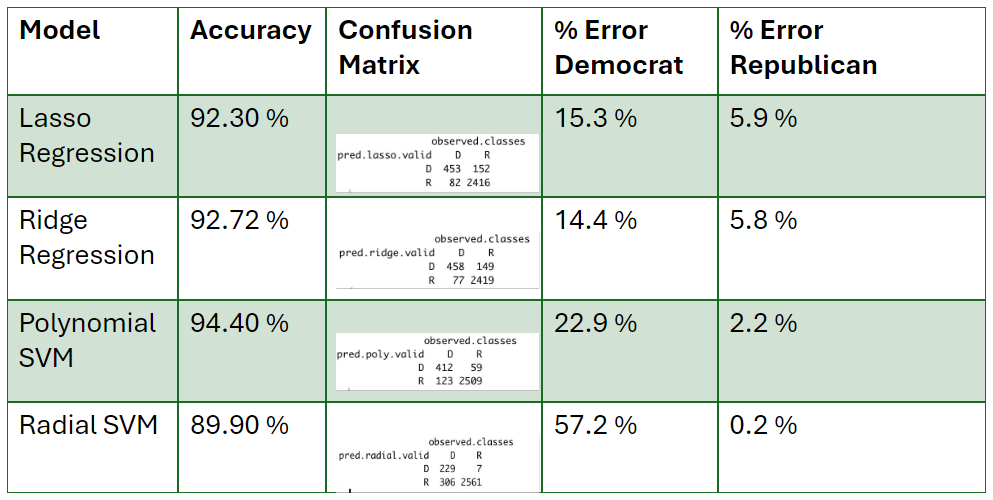
*Figure 4: Covariance plot of the Predictors in the Dataset*

A graph of a performance of tested models

Description automatically generated

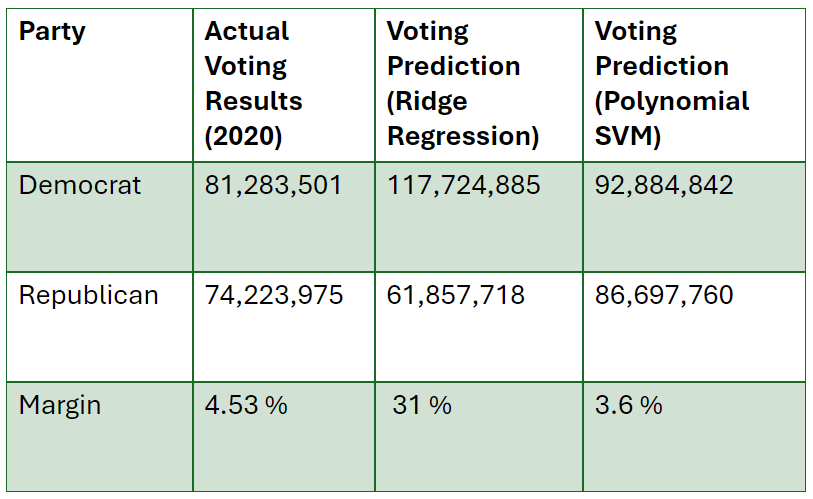
*Figure 5: True Positive Rate vs False Positive Rate for Each Model Tested*

Next, we created confusion matrices to compare the sensitivity, specificity, and overall accuracy of the models on predicting counties as Republican or Democrat. Results of these assessments are shown below in Figure 6:



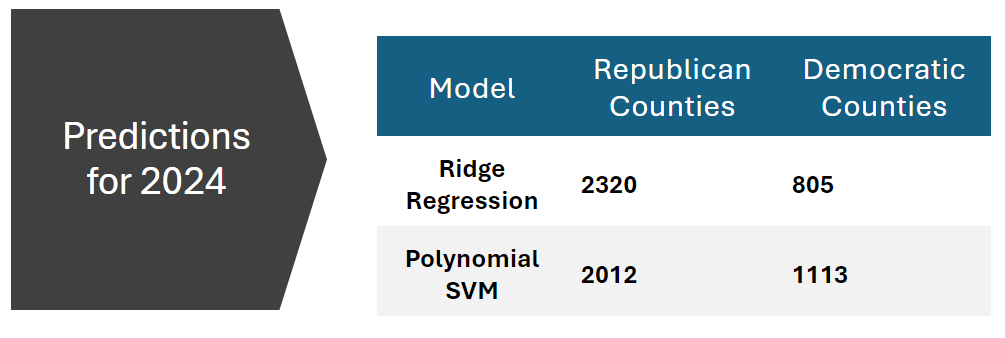
*Figure 6: Accuracy and Misclassification Comparison of Each Model*

The accuracy was the highest for polynomial SVM, followed by the ridge regression base model and lasso regression, while radial SVM had the lowest accuracy rate; however, assessing the accuracy in predicting Republican or Democrat can paint a different picture of the model performance. Comparing our two best-performing models, Polynomial SVM and Ridge Regression, we can see that the Polynomial SVM had a lower percent error for Republican counties and a higher percent error for Democratic counties. Comparing overall accuracy and in-class accuracy would serve as an important distinction during the next stage of the project: translating county political leanings into predicted votes. The best accuracy model, Polynomial SVM, was compared to the baseline Ridge Regression model to produce predictions of voter turnouts for the 2020 validation set. Both predictions were then compared to actual voter turnouts for validation. During this stage of the analysis it was revealed that the misclassified counties from both models tended to have higher average populations. As such, misclassification as Democratic would lead to significant skewing in the analysis on total votes per party. The results of the voter predictions for the 2020 election cycle are shown below in Figure 7:



*Figure 7: Validation Analysis on 2020 Dataset using Ridge Regression and Polynomial SVM*

As can be seen, the Polynomial SVM predictions produced a very similar percent difference in votes between the two parties to the actual value from the 2020 election, implying the model is able to produce accurate results. The actual 2020 margin was 4.53% in the Democratic candidates favor while the polynomial SVM had a margin of 3.6% in the Democratic candidate while ridge regression had a whopping difference of 31%.



*Figure 8: 2024 Prediction Using the Ridge Regression and Polynomial SVM*

Ultimately we made our predictions for 2024 using ridge regression and polynomial SVM, as they were the best performing models we had generated. The results, as shown in Figure 8, show 2320 Republican counties and 805 Democratic counties for the ridge regression model and 2012 Republican counties and 1113 Democratic counties for polynomial SVM.

**Challenges and Lessons Learned**

While working on the first phase of our project, we encountered several challenges that needed careful consideration and adaptation of our methodology. The largest challenge in the project was securing an appropriate dataset for training the models. Originally, it was planned to leverage a dataset from Kaggle for training of the machine learning models; however, upon further inspection it was revealed that the information present was too outdated to be effective in fully supporting all the goals of the project. As such, extensive data-scraping was conducted in order to provide a solid foundation for the project. This data scraping and the following cleaning took up significantly more time than expected; however, the results proved worthwhile as shown by the accuracy of the model when compared to the actual 2020 election results. As such, a large lesson learned was the importance of taking time to properly set up the project to ensure success later on.

Another major challenge that arose during the progression of the project was when we received our results. As shown earlier, the misclassification rates for Democratic counties and Republican counties are vastly different, displaying an a level of difficulty when identifying Democratic counties as the error rates were over 10% whereas for Republican counties they were less than 5%. It is important to look deeper and understand nuances in the dataset as this would help us understand any deficiencies in the model. Based on further research, we found out that Democratic counties tend to be larger (urban and suburban areas), which can lead to more diversity in voting patterns whereas rural areas very strongly vote for Republican candidates, emphasizing how the size of the county can lead to differences in the predictions. Despite these obstacles, we implemented strategies to mitigate the impact of discrepancies and incompleteness, such as variable renaming, enabling us to proceed with our analysis effectively.

**Future Work**

As mentioned earlier, the error rate for Democratic counties and Republican counties were drastically different. A major player for this is because the nature of each county is different. Democratic counties tend to be larger, mostly urban or suburban areas with more mix of votes and lower margins. Republican counties tend to have strongholds in rural areas, which are very lopsided with high margins favoring Republicans. This difference in the nature of each county leads to the different error rates when each model tries to predict counties and hence leading to a higher error rate for Democratic counties as the margins are not as dramatic as those for rural or Republican counties. Hence, a sensitivity analysis on the counties involved, their population and voter turnout would drastically help in more accurately profiling each county and how they would vote.

Another important thing we would want to consider is the time periods we would want to use. The whole goal of this project is to better understand trends in elections which have an impact on public policy, which affects everyone in the country and the world. Hence, using datasets from earlier years to see changes in trends would significantly help us decipher any trends over the years. This can significantly aid in understanding not just national and federal election outcomes, but demographic shifts and economic changes in local regions, states and districts to see their patterns and policies. Additionally, if we can expand this model to understand and analyze other elections such as congressional, gubernatorial and local elections, it can help us understand which party and policies would be in place in certain areas, states and regions.