ECE 5424 - Advanced Machine Learning

Final Project Report

County Presidential Election Results Prediction Based On Fundamentals: A Comparison

*****

Team Members

Hulya Dogan, hulyad@vt.edu

Peter Chen, chenp3@vt.edu

Minh T. Nguyen, mnguyen0226@vt.edu

Instructor

Professor Creed F. Jones, Ph.D., crjones4@vt.edu

Institution

Virginia Polytechnic Institute and State University

Location

Blacksburg, VA, USA

Date

December 1st, 2022

Abstract

The United States of America is the top influential nation in the world [14] which has dynamic relations and effects on public policy [19], global economy/trades [5], and geopolitics. Thus, the US Presidential Election is one of the most significant events that caught millions of people's and international leaders' attention. In class, we learned multiple unpredictable vital factors could determine the election results, e.g., Truman defeated Dewey in the 1948 US presidential election [18]. Several research attempts have been made to accurately predict the presidential election results, which can bring political and economic advantages to an organization or a nation. *In this study, we aim to predict presidential election results with machine learning techniques for the 2020 election based on analyzing the fundamentals of the four U.S. presidential elections from 2004 to 2016.*

Accomplishment & Acknowledgement

Here are the contributions of our team members:

- Hulya:
    - Finding and preparing the data
    - Writing: Abstract, Sections I, II, III (B), VI.
    - Code: DataPreprocessing.ipynb
    - Data: data_dictionary.csv
    - Data: edit_summary_stats.csv
    - Data: data_combined.csv
- Peter:
    - Writing: Section III(A), IV(A, C), Section V.
    - Code: us_election_prediction_svm_with_feature_selection.ipynb
    - Code: us_election_prediction_mlp_with_feature_selection.ipynb
    - Code: us_election_prediction_logistic_with_feature_selection.ipynb
    - Code: get_important_features.ipynb
    - Code: features_analysis.ipynb
    - Code: countypresPreProcessed.ipynb
- Minh:
    - Writing: Section IV (A, B), Section V (A, B, D), Section VI.
    - Code: data_exploration.ipynb.
    - Code: us_election_prediction_decision_tree_no_feature_selection.ipynb.
    - Code: us_election_prediction_random_forest_no_feature_selection.ipynb.
    - Code: us_election_prediction_svm_no_feature_selection.ipynb.
    - Code: us_election_prediction_mlp_no_feature_selection.ipynb.

All team members have agreed upon the contributions; the contribution can be traced back via our team's Github. Please contact one of our team members if you have any questions or concerns. All team members have reviewed and approved the submitted codes, report, and datasets.

Our team would like to thank Professor Creed F. Jones for providing quality teaching materials, lectures, and guidance for our project.

Foreword

The project report is divided into seven sections. To make the report easy to read and concise, we only provide a description, analysis, and results of our work (Section I - VII). In addition to this report, our team submitted all our fully commented, reproducible codes (in JupyterNotebook or Python, in separate files) and datasets. Our team recommends checking our code in Jupyter Notebook as they are easier to comprehend, and the results of each cell are presented.

Table of Contents

I. Introduction

There is a vast amount of literature in different disciplines, such as economics, political science, and data science, about what factors affect the prediction of election outcomes. Various data are being considered to predict the election results, such as social media posts, survey results, referendum judgments, etc. The prediction models are receiving increasing attention worldwide, especially during the election season. Especially in the US, there is a long tradition of election forecasting which could get accurate results if appropriate methods are used. [4] [5] [7].

To predict the election results, especially in the United States, there are various sources such as fundamental variables, for example, economic data, survey data from pre-election polls, or social media posts about the election. All of these sources bring advantages and disadvantages to the forecasting process. For example, the data for fundamental variables, such as household income, education levels, race, and homeownership, are available much earlier than pre-election polls. Therefore, these fundamental variables could help forecast the presidential election results [1] [3] [2].

In our project, we highlight the potential for predicting the United States presidential election outcomes at the county level based on the fundamental variables acquired from American Community Survey data (ACS). Fundamentals refer to variables independent of the current election rhetoric, the campaign performance of a candidate immediately before an election, or social media posts. Fundamental variables include individuals' annual income, annual total family income, age, gender, marital status, race, citizenship status, language spoken at home, education level, and employment status at the individual level. Using these fundamental variables, we aim to determine whether we can predict election outcomes.

II. Background

Election prediction has a long history in the United States. There are several approaches to forecasting election results, and there are a variety of models based on fundamental variables [1] [8] [6] [13] [15]. One study that predicts the presidential elections based on fundamental variables is the Abramowitz article [1]. In the article, the author aimed to predict the 2012 election based on Barack Obama's performance in office and a wide range of issues ranging from government spending and health care to immigration and gay marriage. The essential Time for Change Model was used based on the results of the 16 presidential elections since World War II. Most of the presidential election predictions had high predictive accuracy.

In Schaffner et al. study [8], the authors examined the data from a national survey conducted during the final week of October 2016. Using unique measures of attitudes on racism and sexism, coupled with a question designed to tap into dissatisfaction with personal economic

conditions, they were able to determine to what extent each of these explanations helped to explain vote choices in 2016 and, ultimately, whether either of these explanations can explain the education gap in vote choice among whites. They found that while economic dissatisfaction was an important part of the story, racism and sexism were much more impactful in predicting support for Trump among white voters. Specifically, sexism and racism explain close to two-thirds of the educational gap among white voters in the 2016 presidential vote.

Another research by Akee et al. [2] studied family income and intergenerational voting behavior. They investigated family income affected voting in US elections across two generations from the same household. The results confirm a strong inter-generational correlation in voting between parents and their children. They also showed that a family's economic circumstances during childhood plays a role in influencing levels of political participation in the United States.

Immigration is another example of the fundamental variables, and it is one of the most divisive political issues in the United States and many Western countries. Edo et al.[6] estimated the impact of immigration on voting for far-left and far-right candidates in France, using panel data on presidential elections from 1988 to 2017. What they predicted is immigration increases support for far-right candidates. They also found that immigration has a weak negative effect on support for far-left candidates, which could be explained by reduced support for redistribution.

As stated above, there is a dearth of literature on predicting presidential election results based on several fundamental variables in the United States. With this project, we hope to contribute to the literature since what we used to predict presidential elections are independent of current election rhetoric and campaign performances, which could immediately impact people's voting decisions.

III. Research Design, Methods, & Datasets

A. Research Design & Methods

Six steps have been outlined for the project. These steps are based on the machine learning model development pipeline we learned during the ECE 5424 lectures. Throughout our experiments, we will follow these guidelines strictly to get the best results.

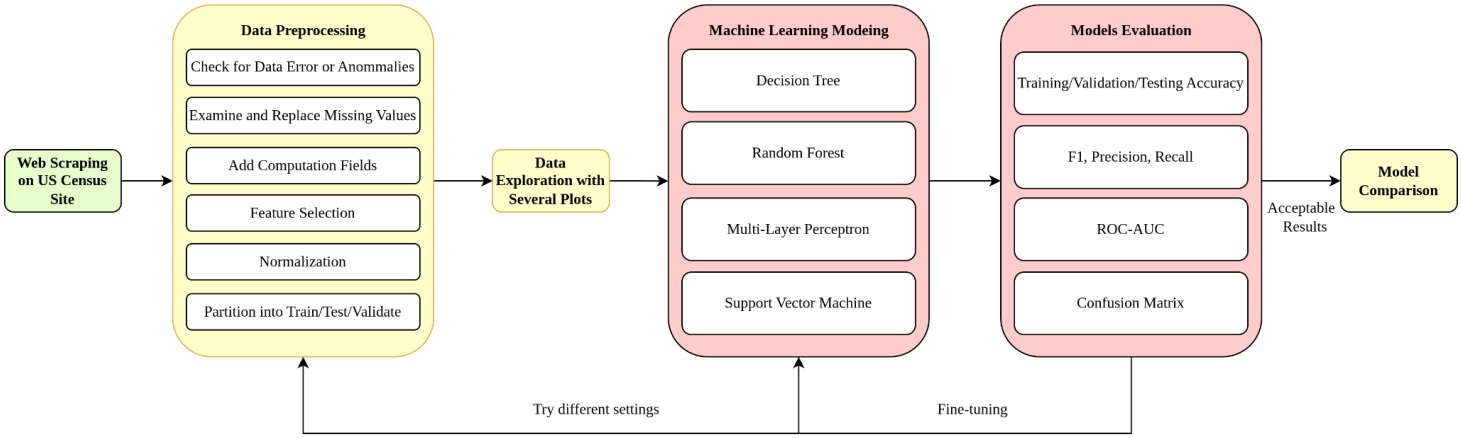

Figure 1. Machine Learning Modeling Life Cycle.

Step 1 - Web Scraping: Datasets and demographic information will be scraped on the US Census website via https://usa.ipums.org/usa/. Here we have the county vote dataset and county census dataset.

Step 2 - Data Preprocessing: As the datasets are raw, several data preprocessing methods will be applied, including the six steps above: a) Check for data error or anomaly, b) Examine and replace missing values, c) Add computational fields, d) Feature selection, e) Normalization, f) Partition into train/test/validate. Lastly, we will merge the two processed datasets (county vote and county census).

Step 3 - Data Exploration: The processed data are represented in high-dimension. Thus, data visualization tools such as histograms or distribution plots will help understand the datasets.

Step 4 - Machine Learning Modeling: Four machine learning models (within the scope of our class) will be experienced for the classification task: Decision Tree, Random Forest, Multi-Layer Perceptron, and Support Vector Machine. Different parameters will be experienced.

Step 5 - Models Evaluations: Trained model will be evaluated with several metrics, including accuracy, F1, Precision, Recall, ROC-AUC, Confusion matrix (TP/TN/FP/FN), Training/Validating/Testing curves (for overfitting and underfitting evaluation).

Step 6 - Model Comparison: Different models' types and settings will be compared, and provided insights into their results.

As the task of determining the winners of the elections, how will we determine the winner? Our team knows there will be two types of votes in real elections - national and electoral college. However, since we have the dataset which maps the census of each county to the vote results (Red or Blue) of that county, we make the winner declaration process easier by counting the total

predicted votes of both parties in each county. Thus, whichever party has the larger counting will win the election.

In data exploration, as we do binary classification, we check whether the dataset is balanced. We count the votes for both parties and plot the bar graph for all four election years. If both parties have a similar number of votes, the dataset is balanced, which is great for unbiased classification.

In terms of feature elections, we used forward and backward methods as they have implementations in Scikit-Learn. Furthermore, Recursive Feature Elimination with Cross-Validation (RFECV), part of the backward method, automatically determines the number of optimal features and the features themselves.

In machine learning modeling, we will train, validate, and test the four models we learned in class (Decision Tree, Random Forest, Multi-Layer Perceptron, and Support Vector Machine). We chose these models as they are in the scope of the class, and they have been proven to perform well (in various tabular datasets in homework). First, The dataset will be split into training, validating, and testing sets with ratios of 80%, 10%, and 10%, respectively. All four machine learning models will be trained and validated on the training and validating sets. Then the model will be tested and evaluated with several metrics. Our team iterates this process until we get acceptable results. Then the model will be compared between feature-selected models vs. non-feature-selected models and pick the best model.

B. Datasets

Our purpose in this study is to see if we can predict election outcomes based on fundamentals in a county. Fundamentals refer to variables independent of the current election rhetoric, the campaign performance of a candidate immediately before an election, or social media posts. Fundamental variables include individuals' annual income, annual total family income, age, gender, marital status, race, citizenship status, language spoken at home, education level, and employment status at the individual level. We intentionally added citizenship status, marital status, and language spoken at home because immigration reform and gender issues have been at the core of every election for the last two decades. Education level is also a fundamental that changes how individuals approach contemporary political and social issues. Income level, both at the individual and family level, and employment status are indicators of how well the economy has been doing and could indicate voting preferences regardless of the political rhetoric.

We compile our data from two main sources: individual-level American Community Survey (ACS) Census Microdata provided by IPUMS[9] and county-level presidential election data provided by MIT election and Data and Science Lab[7]. The fundamental variables came from

annual ACS samples. These variables are then aggregated at the county level using person weights provided by the ACS. We generated county-level average individual and family incomes and average age variables as continuous variables. For the categorical variables, we calculated the percentages in each county. Some variables, such as years of schooling that indicated the education level of individuals, are regrouped to minimize noise and feed the model with more meaningful variables. As such, years of schooling are re-categorized to reflect those with a high school diploma or lower education, some college or bachelor's degree, master's or professional certificate, and doctoral degree. Similarly, citizenship status and language spoken at home are regrouped into fewer yet more indicative categories. Finally, all income variables are adjusted for inflation using the consumer price index released by the Census Bureau.

MIT releases presidential election data results at the county level from 2000 onward. The data contains the number of votes each party receives as well as the winner of the election in that county. Because many county boundaries and names have changed over time, ACS has a consistent county linkage for around 400 counties after 2006. Therefore, when we combined the two data sets, we had 1671 matching county-year election records, including 2008, 2012, 2016, and 2020. Initially, we planned to predict the 2024 election results, however, due to the lack of census data by the time of our project completion, we decided to move forward with predicting the 2020 elections.

IV. Experiments

A. Data Preprocessing

MIT releases presidential election data results at the county level from 2000 onward. The data contains the number of votes each party receives. We first sorted the dataset by year and "county_fips," a unique identifier for each voting county. Rows without county fips codes were dropped as they could not be correlated to the census dataset. Next, we populated the votes for each political party to their column, compared the total votes of each county between the two dominant Democrat and Republican parties, then, based on the result, generated a new column "winner" to indicate the winning party for each county where 0 represents Democrat, 1 represents Republican.

Due to many county boundaries and names having changed over time, ACS has a consistent county linkage for around 400 counties after 2006. Therefore, when we combined the two data sets, we had 1671 matching county-year election records, including 2008, 2012, 2016, and 2020.

After grouping the two datasets, we have a dimension of (7526 rows x 45 columns). However, we will choose the rows within those four years since we only consider four election years (2008, 2012, 2016, 2020). Thus, we will have a dimension of (1858 rows x 45 columns).

Here are the feature names and their corresponding types:

```
year                   int64
county_fips            int64
inctot               float64
mortamt1             float64
avrg_age             float64
ftotinc              float64
foodstmp_1_freq      float64
foodstmp_2_freq      float64
sex_2_freq           float64
sex_1_freq           float64
marst_5_freq         float64
marst_6_freq         float64
marst_1_freq         float64
marst_4_freq         float64
marst_3_freq         float64
marst_2_freq         float64
race_1_freq          float64
race_2_freq          float64
race_7_freq          float64
race_8_freq          float64
race_5_freq          float64
race_6_freq          float64
race_3_freq          float64
race_3_freq          float64
race_4_freq          float64
race_9_freq          float64
ctz_stat_1_freq      float64
ctz_stat_3_freq      float64
ctz_stat_2_freq      float64
lang_1_freq          float64
lang_2_freq          float64
educ_attain_2.0_freq float64
educ_attain_1.0_freq float64
educ_attain_3.0_freq float64
educ_attain_4.0_freq float64
empstat_1.0_freq     float64
empstat_3.0_freq     float64
empstat_2.0_freq     float64
state_po              object
county_name           object
democrat             float64
green                float64
liberitarian         float64
other                float64
republican           float64
winner               float64
dtype: object
```

Figure 2. Feature names and their data types.

Then, as we consider each row as a data point for our machine learning model, we will drop the categorical state names and abbreviations. We will also drop all the number of votes labels except for the "winner" columns (0 represents Democrat, 1 represents Republican). We aren't concerned about other parties as they do not have a history of winning.

- county_fips: as there is no correlation between fips and vote numbers.
- state_po: as there is no correlation between state appreciation and vote numbers.
- county_name: as there is no correlation between county names and vote numbers.
- democrat: as we do classification task.
- green: as we do classification task.
- libertarian: as we do classification task.
- other: as we do classification tasks.
- republican: as we do classification task.

Thus, we will have a dimension of (1858 rows x 37 columns). Next, we use the rows from 2008, 2012, and 2016 for model training; then, we test the model performance on the 2020 dataset (expectation of a Democratic win).

Note: Although we did not show the output of the other preprocess steps, however, in all of our codes, we have followed all six preprocessing steps: a) Check for data error or anomaly, b) Examine and replace missing values, c) Add computational fields, d) Feature selection, e) Normalization, f) Partition into train/test/validate.

B. Data Exploration

When considering grouping all four election years into one dataset, we can see that both labels are balanced.

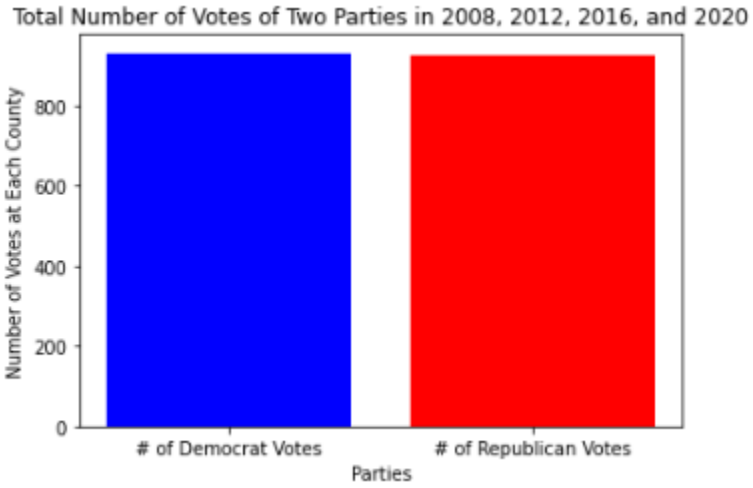

Figure 3. Label counts four election years dataset.

The dataset is also somewhat balanced for both labels when considering separating four election years.

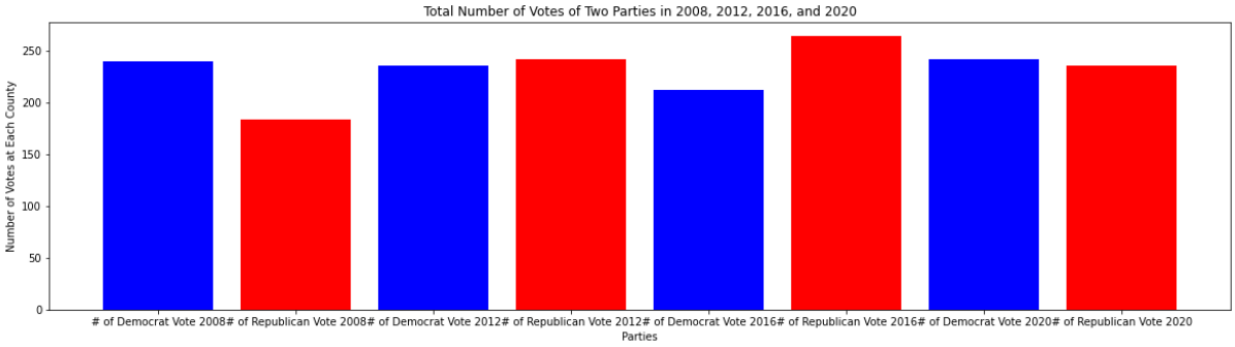

Figure 4. Label counts four election years dataset.

Here, regarding splitting the dataset or grouping them, the dataset is quite balanced for both voting labels. Thus, this is an excellent dataset to avoid unbiased predictions.

C. Feature Selections

Feature selections extract vital variables that determine the outcome of a presidential election while also helping to optimize our prediction models. Using available Scikit-Learn built-in functions, forward and various backward feature selection processes were performed for six models. Logistic Regression, Multilayer Perceptron, and four Support Vector Machine models

using different kernels (linear, poly, RBF, and sigmoid). Feature selections were not made for the Decision Tree and Random Forest. We can see the order of importance by looking at which level of the tree diagram each feature was chosen and the delta in entropy value. Due to computational power and time constraints, feature selection was only completed on one of the three MLP cases (1 hidden layer, 10 nodes). From the observation of results, it was obvious that the backward feature selection process is the better choice as accuracy across all models outperformed forward feature selection. In addition, Recursive Feature Elimination with Cross-Validation (RFECV) is the best choice out of the backward feature selection group, as it can find the optimal number of features without specification from the user or revert to the default value. However, due to some non-linear models from Scikit-Learn lacking the "coef_" attribute, which returns the weight coefficients required to perform RFECV, it was only used on applicable linear models (Logistic Regression & Linear SVM).

Feature selections reduced the number of feature variables to 27 or less compared to the original 36. However, at least one model selected 35 of the 36 features at least once. While this improved the efficiency of individual modeling, we needed to understand which features are essential for determining election results. Therefore, a list of 35 features was ranked by the total times selected, with six being the highest. Those selected four or more times were populated for analysis of voting behaviors discussed in a later section. This list, shown in *Figure 5 has* 27 variables: income, mortgage payment, marital status, race, language, education, and employment status. Population with a Master's Degree was the most crucial variable, used 6 out of 6 times.

	important_features	# of Times Selected (Out of 6)	Description
0	inctot	4	Average Annual Income of Individuals
1	mortamt1	5	Average amount of mortgage payments
3	ftotinc	5	Average Annual Total Family Income
4	marst_5_freq	5	% Widowed
5	marst_6_freq	4	% Never Married
6	marst_1_freq	5	% Married Spouse present
7	marst_4_freq	5	% Divorced
8	marst_3_freq	4	% Seperated
11	race_2_freq	4	% Black/African American
14	race_3_freq	4	% American Indian or Alaska Native
16	ctz_stat_3_freq	5	% Non-Citizen
17	ctz_stat_2_freq	5	% Naturalized Citizen
18	lang_1_freq	5	% English is spoken at home
19	lang_2_freq	4	% Another Language is spoken at home
20	educ_attain_2.0_freq	5	% Some College or Bachelor Degree
22	educ_attain_3.0_freq	6	% Masters or Professional Certificate
23	educ_attain_4.0_freq	5	% Doctoral Degree
24	empstat_1.0_freq	4	% Employed
25	empstat_3.0_freq	5	% Not in the labor force
26	empstat_2.0_freq	5	% Unemployed

Figure 5. Important Features.

D. Machine Learning Modeling

Note: All models are developed in Sklearn. Although in our proposal, we mentioned that we would implement Multi-Layer Perceptron in Tensorflow, after a discussion with Professor Jones, we have decided to implement it in Sklearn. From homework, we learned that for Multi-layer Perceptron, the performance is similar. However, Sklearn is easier to implement.

For the Decision Tree, we have three considered settings. We choose these settings to try possible diverse settings

- Setting 1: criterion='gini', splitter='best'
- Setting 2: criterion='entropy', splitter='best'
- Setting 3: criterion='log_loss', splitter='best'

For Random Forest, we have three considered settings. We choose these settings to try possible diverse settings

- Setting 1: criterion='gini'
- Setting 2: criterion='entropy'
- Setting 3: criterion='log_loss'

For the Support Vector Machine, we have four considered settings. We choose these setting to try possible diverse settings

- Setting 1: kernel='linear'
- Setting 2: kernel='poly''
- Setting 3: kernel='rbf'
- Setting 4: kernel='sigmoid'

For Multi-Layer Perceptron, we have three considered settings. We choose these settings to try possible diverse settings

- Setting 1: 1 hidden layer with 4 neurons
- Setting 2: 1 hidden layer with 10 neurons
- Setting 3: 2 hidden layer with 10 neurons

Note: As later we learned in class, we know that we can use Grid Search to optimize the parameters. However, due to constraints in computational power and time, as well as we did not put this idea into our proposal, our team decided not to implement Grid Search. However, we acknowledge its advantages and will consider this method as a follow-up optimization to our future direction for this project.

V. Results & Analysis

The two tables below show the results of each model setting for the Training/Validation dataset and Testing dataset. Note the **bold** numbers are the best within the same machine learning type settings (such as within the Decision Tree settings), while the highlighted ones are the best for all settings (such as within all Decision Tree, Random Forest, SVM, MLP settings).

In observation of model performance between with and without feature selection, we can see that similar accuracy is achieved but with fewer feature variables. The SVM model using Sigmoid is an exception, where the accuracy of the test set and prediction on the 2020 election dataset increased significantly from roughly 60% to 80%. As we can see, modeling becomes more robust and generalized by condensing the number of features by their importance. Although we sacrificed a bit of accuracy for some models for fewer variables, the value gained by implementing feature selection in simplification more than makes up for it. If more data can be obtained for training, the performance of feature selection would increase and perhaps achieve higher accuracy with lesser features, as we saw in one of the homework assignments.

Table 1. Model's Performance on the Training/Validation (2008, 2012, 2026) Dataset

Models Settings	Training MSE	Validating MSE	Accuracy	F1 Score	Precision Score	Recall Score	ROC-AUC Score
Decision Tree criterion='gini', splitter='best'	**0.000000**	**1.043478**	**0.739130**	**0.727273**	**0.720000**	**0.734694**	**0.738907**
Decision Tree criterion='entropy', splitter='best'	**0.000000**	1.062802	0.734300	0.719388	0.719388	0.719388	0.733547
Decision Tree criterion='log_loss', splitter='best'	**0.000000**	**1.043478**	**0.739130**	**0.727273**	**0.720000**	**0.734694**	**0.738907**
Random Forest criterion='gini'	**0.000000**	**0.666667**	**0.833333**	**0.832117**	**0.795349**	**0.872449**	**0.835307**
Random Forest criterion='entropy'	**0.000000**	0.695652	0.826087	0.825243	0.787037	0.867347	0.828169
Random Forest criterion='log_loss'	**0.000000**	0.734300	0.816425	0.812808	0.785714	0.841837	0.817707
SVM - kernel='linear' (**No** Feature Selection)	0.836439	**0.734300**	**0.816425**	**0.815534**	**0.777778**	0.857143	**0.818480**

SVM - kernel='linear' (Feature Selection) **Total Features = 27**	0.844720	0.772947	0.806763	0.805825	0.768519	0.846939	0.808790
SVM - kernel='poly' (**No** Feature Selection)	0.790890	0.763285	0.809179	0.806846	0.774648	0.841837	0.810827
SVM - kernel='poly' (Feature Selection) **Total Features = 18**	**0.741201**	0.763285	0.809179	0.812352	0.760000	**0.872449**	0.812371
SVM - kernel='rbf' (**No** Feature Selection)	0.840580	0.782609	0.804348	0.806683	0.757848	0.862245	0.807269
SVM - kernel='rbf' (Feature Selection) **Total Features = 18**	0.840580	0.763285	0.809179	0.808717	0.769585	0.852041	0.811342
SVM - kernel='sigmoid' (**No** Feature Selection)	1.370600	1.314010	0.671498	0.674641	0.635135	0.635135	0.673914
SVM - kernel='sigmoid' (Feature Selection) **Total Features = 18**	1.010352	0.956522	0.760870	0.775510	0.697959	**0.872449**	0.766500
MLP - 1 hidden layer with four neurons (**No** Feature Selection)	0.977226	0.946860	0.763285	0.760976	0.728972	0.795918	0.764932
MLP - 1 hidden layer with ten neurons (**No** Feature Selection)	0.939959	**0.801932**	**0.799517**	**0.800000**	0.757991	**0.846939**	**0.801910**
MLP - 1 hidden layer with ten neurons (Feature Selection) **Total Features = 19**	1.006211	0.946860	0.763285	0.762136	0.726852	0.801020	0.765189
MLP - 2 hidden layer with ten neurons (**No** Feature Selection)	**0.886128**	0.821256	0.794686	0.792176	**0.760563**	0.826531	0.796293

Table 2. Model's Performance on the Training/Validation 2020 Dataset

Models Settings	Testing MSE	Accuracy	F1 Score	Precision Score	Recall Score	ROC-AUC Score
Decision Tree criterion='gini', splitter='best'	1.305439	0.673640	0.659389	0.680180	0.639831	0.673221
Decision Tree criterion='entropy', splitter='best'	**0.962343**	**0.759414**	**0.731935**	**0.813472**	**0.665254**	**0.758247**
Decision Tree criterion='log_loss', splitter='best'	1.037657	0.740586	0.701923	0.811111	0.618644	0.739074
Random Forest criterion='gini'	**0.577406**	**0.855649**	**0.850972**	0.867841	**0.834746**	**0.855389**
Random Forest criterion='entropy'	0.686192	0.828452	0.814480	**0.873786**	0.762712	0.827637
Random Forest criterion='log_loss'	0.744770	0.813808	0.794457	0.873096	0.728814	0.812754
SVM - kernel='linear' (**No** Feature Selection)	0.761506	0.809623	0.793651	0.853659	0.741525	0.808779
SVM - kernel='linear' (Feature Selection) **Total Features = 27**	0.786611	0.803347	0.784404	0.855000	0.724576	0.802371
SVM - kernel='poly' (**No** Feature Selection)	0.828452	0.792887	0.792887	0.866310	0.686441	0.686441
SVM - kernel='poly' (Feature Selection) **Total Features = 18**	0.878661	0.780335	0.744526	0.874286	0.648305	0.778698
SVM - kernel='rbf' (**No** Feature Selection)	**0.719665**	0.820084	0.808036	0.853774	0.766949	**0.819425**
SVM - kernel='rbf' (Feature Selection) **Total Features = 18**	0.753138	0.811715	0.788732	**0.884211**	0.711864	0.810478

SVM - kernel='sigmoid' (**No** Feature Selection)	1.171548	**0.707113**	0.699571	0.708696	0.690678	0.706909
SVM - kernel='sigmoid' (Feature Selection) **Total Features = 18**	0.744770	0.813808	**0.809422**	0.818182	**0.800847**	0.813647
MLP - 1 hidden layer with four neurons (**No** Feature Selection)	0.845188	0.788703	0.770975	0.829268	0.720339	0.787855
MLP - 1 hidden layer with ten neurons (**No** Feature Selection)	0.820084	0.794979	0.771028	**0.859375**	0.699153	0.793791
MLP - 1 hidden layer with ten neurons (Feature Selection) **Total Features = 19**	**0.728033**	**0.817992**	**0.815287**	0.817021	**0.813559**	**0.817937**
MLP - 2 hidden layer with ten neurons (**No** Feature Selection)	0.794979	0.801255	0.783599	0.847291	0.728814	0.800357

From observation of results presented, four main machine learning groups were used. Decision Trees, Random Forest, SVM, and MLP. Although their performance varies, the resulting predictions of the 2020 Presidential Election were consistent across. They will each be discussed below:

Decision Tree was the worst performer in terms of accuracy compared to the other four machine learning methods. This is likely due to the high number of total features and a relatively small training data set. For this modeling, the main difference between each model was the criterion used. "Gini," "entropy," and "log_loss." Conceptually "entropy" and "log_loss" are similar as "log_loss" is also known as cross_entropy loss, which measures the disorder of each feature for calculation of information gain. "Gini," on the other hand, measures the rate of misclassification of instances. As shown in table 1, the three models yielded similar performance in accuracy. "Gini" is less complex to compute compared to "entropy" and "log_loss," thus less computationally expensive, and is the default method in Scikit-Learn. Although the resulting performance on the 2020 test dataset was lower using the "Gini" model compared to the other two, this is most likely due to the relatively small training dataset, as the performance between the three is similar. With a larger dataset, "Gini" would be the preferred method due to its simplicity.

Here is the plot of three Decision Tree settings in the 2020 test dataset. The prediction is "correct" to be Democrats winning (although we did not count the vote and winning into complex population votes and electoral-college votes).

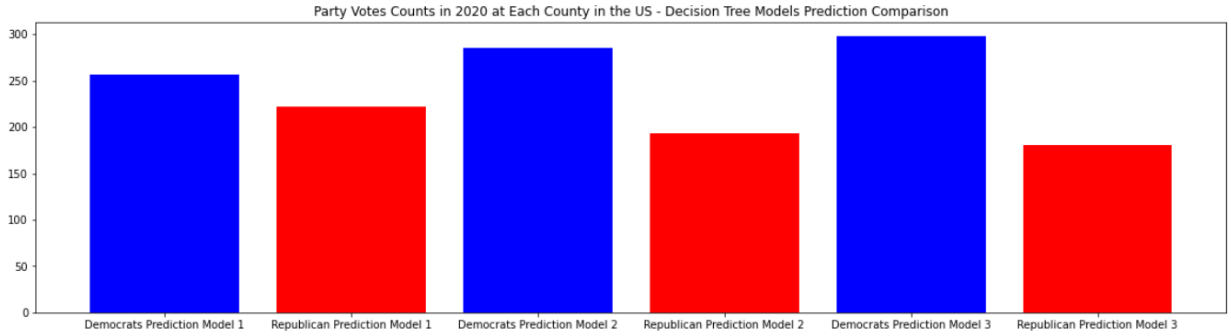

Figure 6. Decision Tree Performance on 2020 Test Dataset.

Random Forest, an extension of Decision Tree, yielded the best model performance out of the four training data. This is likely due to its combined use of K-fold cross-validation, bagging, and subspace sampling, where multiple Decision Tree models are produced based on a different subset of random samples of the dataset with various subsets of features. With all this randomness added to the modeling, the model is more well-trained and generalized, thus more robust for different test sets. Once again, the three criteria yield similar performance, just like the Decision Tree.

Here is the plot of three Random Forest settings in the 2020 test dataset. The prediction is "correct" to be Democrats winning (although we did not count the vote and winning into complex popular votes and electoral-college votes).

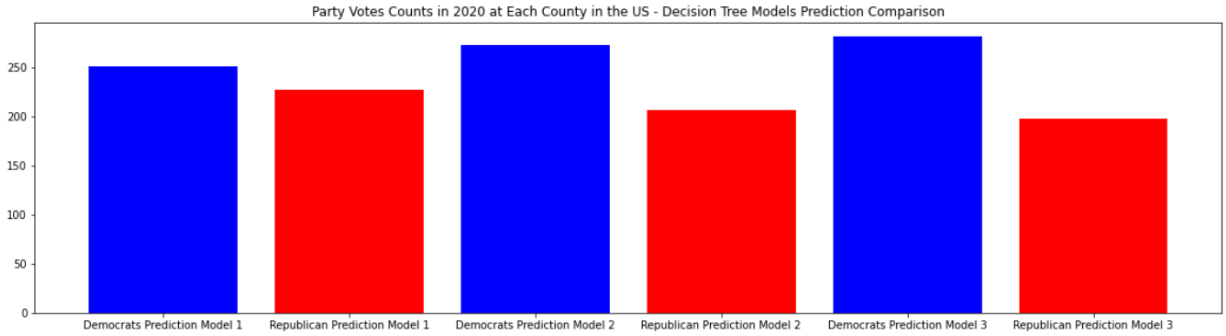

Figure 7. Random Forest Performance on 2020 Test Dataset.

Overall, SVM is the second most effective modeling method for this project. Results across the four SVM models were similar except when the Sigmoid kernel was used. However, that issue was resolved once feature selection was applied, which removed insignificant feature variables. Therefore, that issue was most likely due to noise or outliers introduced by those eliminated features, which prevented the sigmoid function from producing a high-quality classifying

boundary. With its good consistent model performance across the board, relatively simple, and computationally inexpensive to implement, SVM would be an excellent second choice behind Random Forest modeling for this type of project.

Here is the plot of four Support Vector Machine settings in the 2020 test dataset. The prediction is "correct" to be Democrats winning (although we did not count the vote and winning into complex popular votes and electoral-college votes).

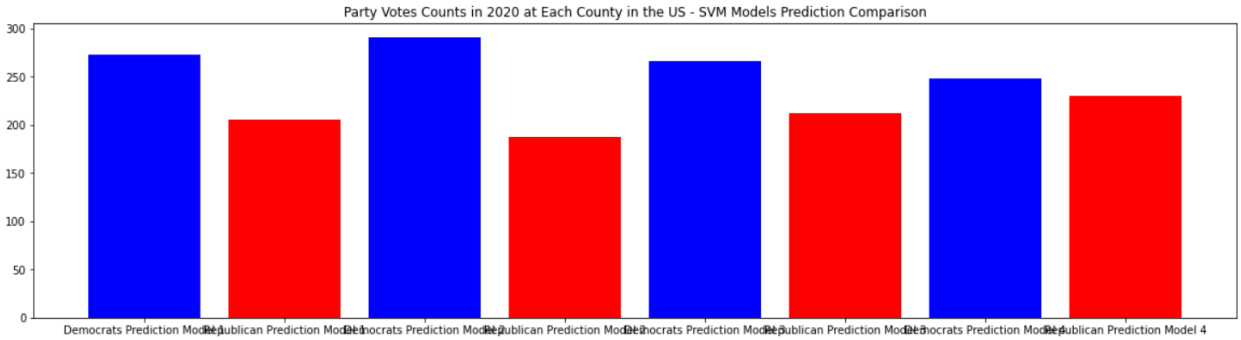

Figure 8. Support Vector Machine Performance on 2020 Test Dataset.

Finally, MLP Classifier's overall performance is lower than SVM and Random Forest but better than the Decision Tree. Its accuracy on test sets hovers around high 70%, with some reaching 80%. However, implementation is more complex than that of SVM and Random Forest. Changing the hidden layer and node parameters could swing model performance more than desired. The "GridSearch" function would be used to find the ideal parameters more effectively, but that will take a significant amount of time to complete, especially on a more extensive data set. MLP Classifier no doubt provides flexibility in modeling and, in theory, can classify any dataset given enough modeling complexity and time. However, it is less efficient in most cases than Random Forest or SVM.

Here is the plot of three Multi-Layer Perceptron settings in the 2020 test dataset. The prediction is "correct" to be Democrats winning (although we did not count the vote and winning into complex popular votes and electoral-college votes).

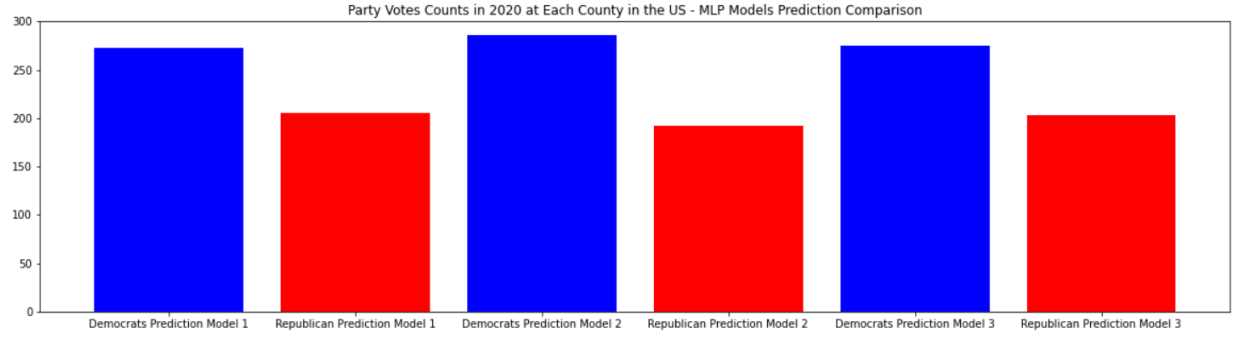

Figure 9. Multi-Layer Perceptron Performance on 2020 Test Dataset.

In analyzing the 27 essential features we extracted in the above section, three spreadsheets were created to examine voters' behavior. In *Figure 10*, we took census data for 2008, 2012, and 2016, compiled two lists of counties based on affiliated political parties, then computed the mean value for every 27 features we selected and graphed them in log base ten scales.

Generally speaking, for these three election years, the population with higher average income, mortgage payment, and education voted Democrat. The single, African American or foreign population also voted Democrat on average. The population of White, Native Americans, married couples, and retirees voted Republican. These findings are mostly consistent with our initial research on the project proposal. However, the notion that higher-income families tend to vote Republican has recently changed.

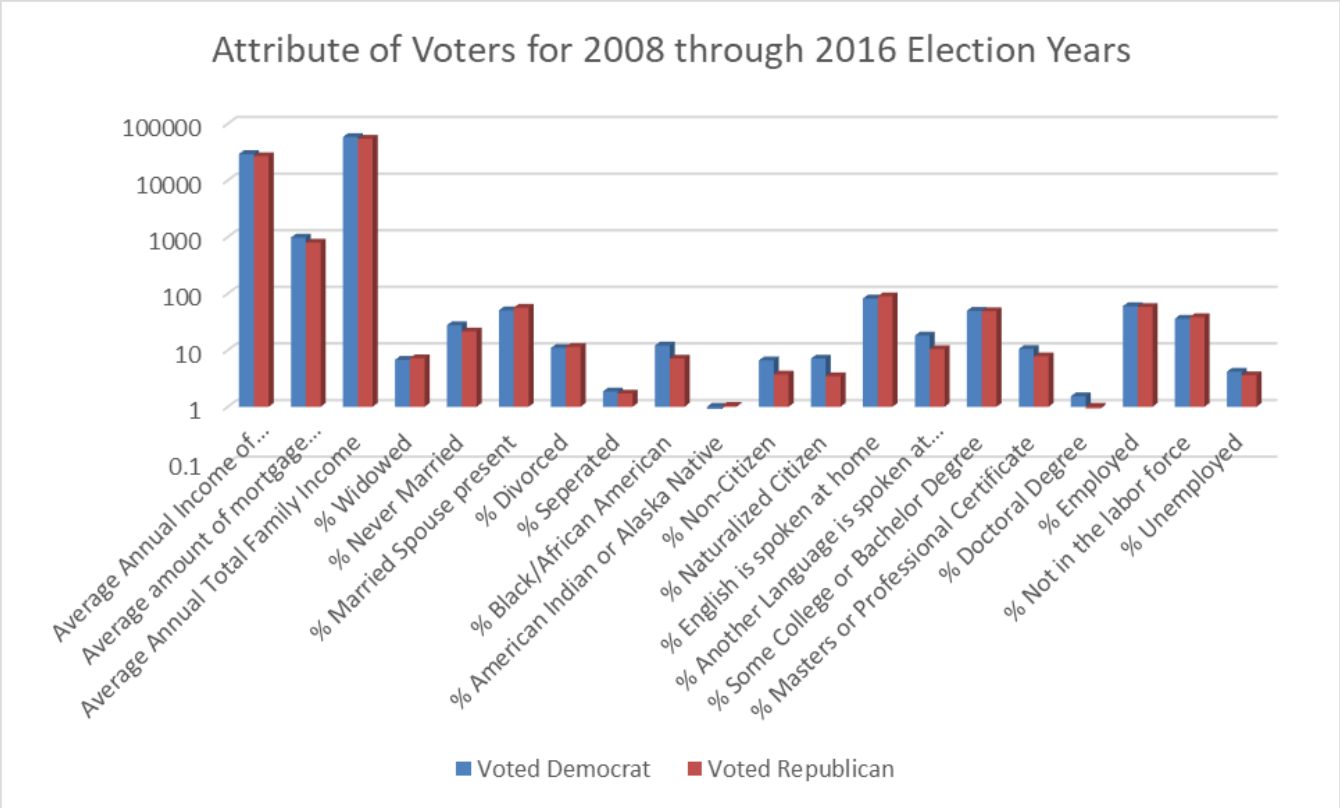

Figure 10. Voters Behavior 2008 through 2016 Election Years.

In *Figure 11*, we looked at the average change in census reports between 2016 and 2020. This gave us some insights into the incumbent party's performance (Republican) during its term, which also affected the 2020 presidential election. From the graph, we can see that there was growth in average income along with mortgage payments which signal GDP growth. However, the unemployment percentage also increased, which does not favor the incumbent. The

population group of single, African American, foreign race, and advanced degree also increased between 2016 and 2020. All these factors favor the Democratic Party in light of recent election results.

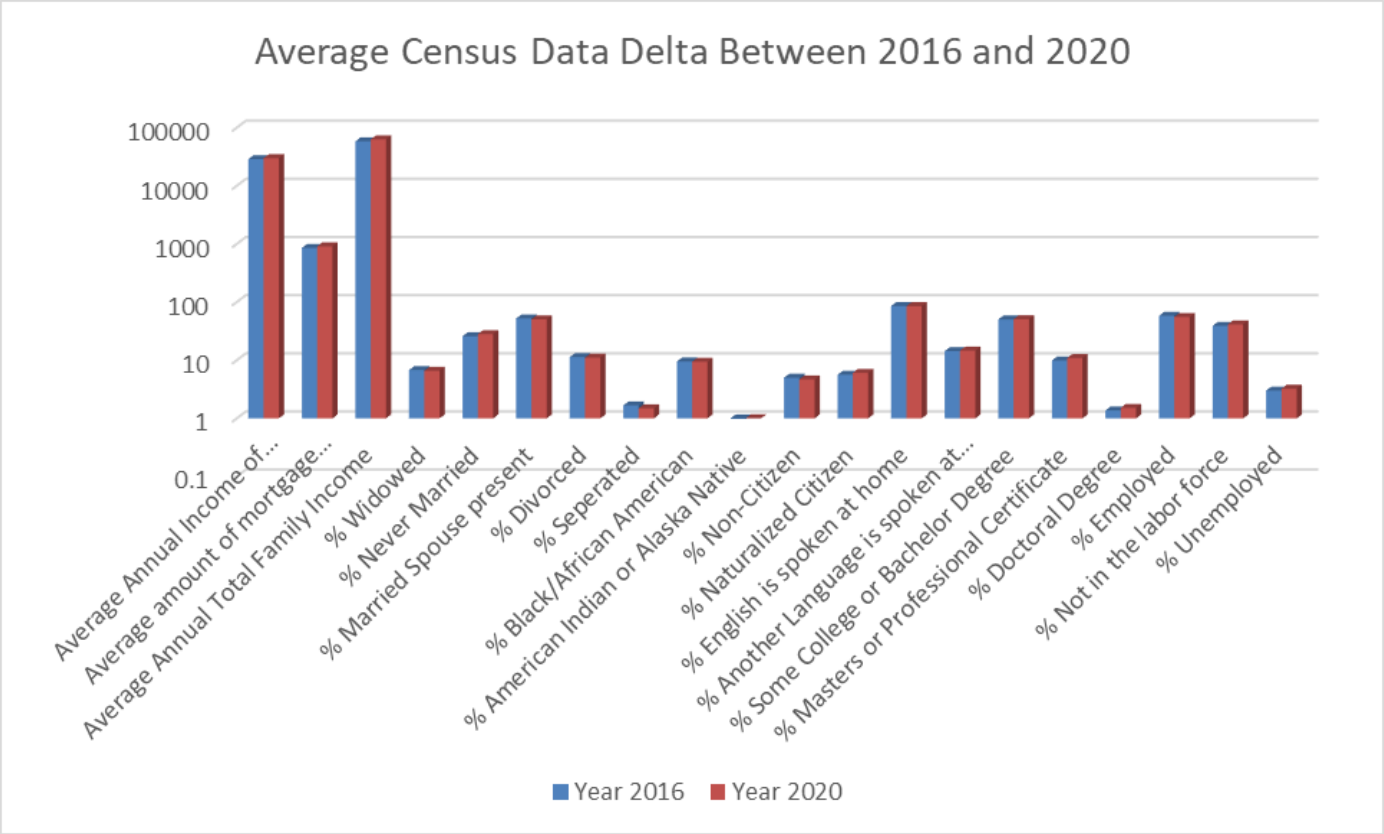

Figure 11. Census Data Delta Between 2016 and 2020.

To our surprise, the Democratic Party won the 2020 Presidential Election, consistent with analysis and prediction models. As shown in *Figure 12,* Voting behaviors remained consistent with the previous three elections. Those single, with higher average annual income, foreign race, and higher education, leaned toward Democrats. And as mentioned above, the population of these groups of people increased from 2016 to 2020, which widened the margin even more. The increase in the unemployment rate hurt the incumbent Republican Party as those who were unemployed voted even more heavily toward Democrats. It should also be noted that the voting behavior of Native Americans in 2020 changed and leaned more toward Democrats. Combining all these factors, it was no surprise that we elected a new president in 2020.

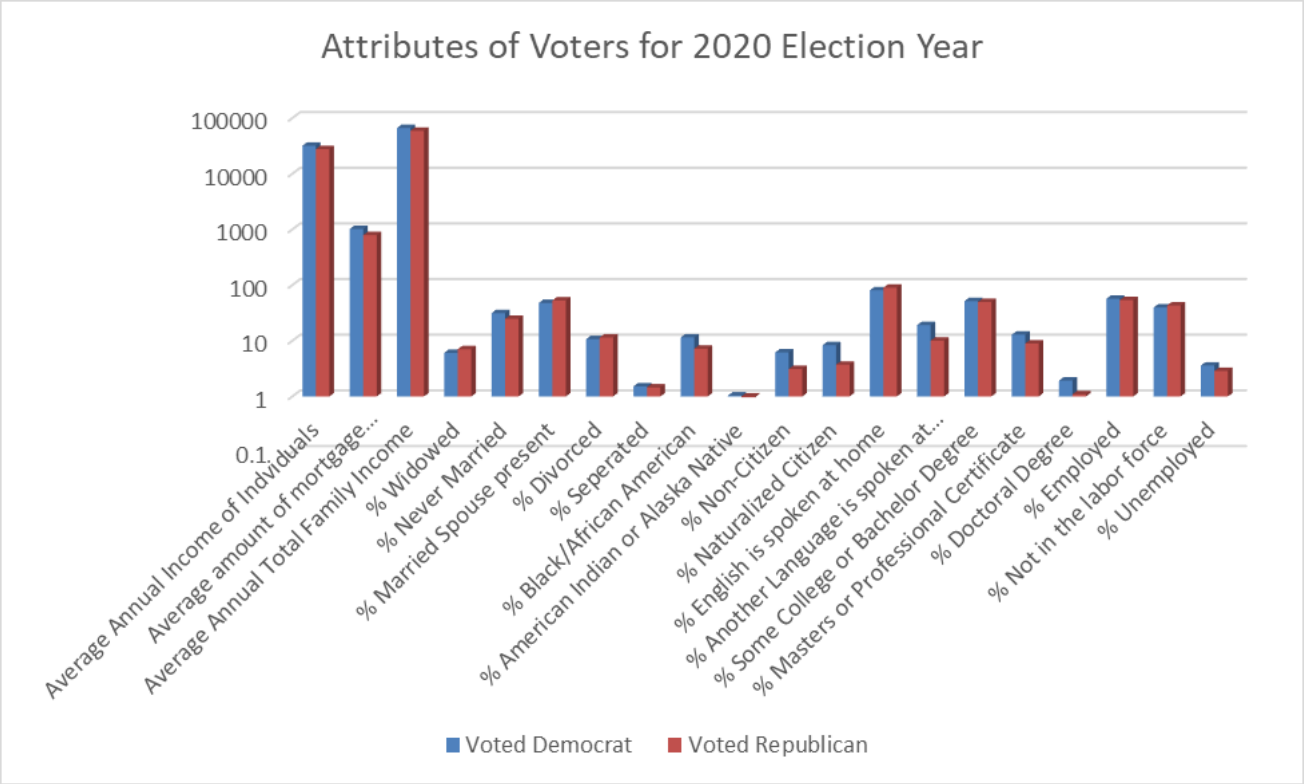

Figure 12. Voters Behavior 2008 through 2016 Election Years.

VI. Conclusion

We investigate if we can predict election outcomes based on fundamentals in a county. We included the variables; annual income of individuals, annual total family income, age, gender, marital status, race, citizenship status, language spoken at home, education level, and employment status at the individual level. We intentionally added some of the variables like citizenship status, marital status, and language spoken at home since they played a significant role in e previous elections. We employed several machine learning algorithms, Decision Trees, Random Forest, SVM, and MLP, to predict the election results and feature selections to extract critical variables that determine the outcome of a presidential election while also helping to optimize our prediction models. Overall based on our analysis, we observed in the last three election years, the population with higher average income, mortgage payment, and education voted Democrat. Likewise, the single, African American, or immigrant population. On the other hand, White Americans, Native Americans, married couples, and retirees voted for Republicans. These findings are mostly consistent with our initial research on the project proposal. However, the notion that higher-income families tend to vote Republican has recently changed due to the increase in the unemployment rate. Overall, we found that the Democratic Party won the 2020 Presidential Election, consistent with analysis and prediction models.

VII. Key References

[1] Abramowitz, Alan. "Forecasting in a polarized era: The time for change model and the 2012 presidential election." PS: Political Science & Politics 45, no. 4 (2012): 618-619.

[2] Akee, Randall, William Copeland, E. Jane Costello, John B. Holbein, and Emilia Simeonova. Family income and the intergenerational transmission of voting behavior: Evidence from an income intervention. No. w24770. National Bureau of Economic Research, 2018.

[3] Amin, Modhurima Dey, Syed Badruddoza, and Jill J. McCluskey. "Predicting access to healthful food retailers with machine learning." Food Policy 99 (2021): 101985.

[4] A. S. R. Manstead, "The psychology of social class: How socioeconomic status impacts thought, feelings, and behavior," *Br. J. Soc. Psychol.*, vol. 57, no. 2, pp. 267–291, 2018.

[5] D. DeSilver, "The politics of American generations: How age affects attitudes and voting behavior," *Pew Research Center*, 09-Jul-2014. [Online]. Available: https://www.pewresearch.org/fact-tank/2014/07/09/the-politics-of-american-generations-how-age-affects-attitudes-and-voting-behavior/. [Accessed: 29-Sep-2022].

[6] Edo, Anthony, Yvonne Giesing, Jonathan Öztunc, and Panu Poutvaara. "Immigration and electoral support for the far-left and the far-right." European Economic Review 115 (2019): 99-143.

[7] MIT Election Data and Science Lab, 2018, "County Presidential Election Returns 2000-2020", https://doi.org/10.7910/DVN/VOQCHQ, Harvard Dataverse, V11, UNF:6:HaZ8GWG8D2abLleXN3uEig== [fileUNF]

[8] Schaffner, Brian F., Matthew MacWilliams, and Tatishe Nteta. "Explaining white polarization in the 2016 vote for president: The sobering role of racism and sexism." In Conference on the US Elections of, pp. 8-9. 2016.

[9] Steven Ruggles, Sarah Flood, Ronald Goeken, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 12.0 [dataset]. Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D010.V12.0

[10] "How presidential elections affect the stock market: U.S. bank," *How Presidential Elections Affect the Stock Market | U.S. Bank*, 13-Jan-2021. [Online]. Available: https://www.usbank.com/investing/financial-perspectives/market-news/how-presidential-elections-affect-the-stock-market.html. [Accessed: 26-Sep-2022].

[11] I. Sabuncu, M. A. Balci, and O. Akguller, "Prediction of USA November 2020 election results using Multifactor Twitter Data Analysis Method," *arXiv.org*, 24-Jan-2021. [Online]. Available: https://arxiv.org/abs/2010.15938v3. [Accessed: 26-Sep-2022].

[12] "The impact of the U.S. election on geopolitics," *Harvard Kennedy School*. [Online]. Available:

https://www.hks.harvard.edu/more/about/leadership-administration/deans-office/deans-remarks/impact-us-election-geopolitics. [Accessed: 26-Sep-2022].

[13] J. Gramlich, "What the 2020 electorate looks like by party, race and ethnicity, age, education and Religion," *Pew Research Center*, 29-May-2021. [Online]. Available: https://www.pewresearch.org/fact-tank/2020/10/26/what-the-2020-electorate-looks-like-by-party-race-and-ethnicity-age-education-and-religion/. [Accessed: 26-Sep-2022].

[14] Jiang, Zhenlong, et al. "Machine learning and simulation-based framework for disaster preparedness prediction." 2021 Winter Simulation Conference (WSC). IEEE, 2021

[15] Mann, J., 2022. *Forecasting the Presidential Election: What can we learn from the models?*. [online] Brookings. Available at: <https://www.brookings.edu/articles/forecasting-the-presidential-election-what-can-we-learn-from-the-models/> [Accessed 29 September 2022].

[16] Mehta, Mihir, et al. "Early stage machine learning–based prediction of US county vulnerability to the COVID-19 pandemic: machine learning approach." JMIR public health and surveillance 6.3 (2020): e19446.

[17] "Most influential countries | U.S. news best countries." [Online]. Available: https://www.usnews.com/news/best-countries/most-influential-countries. [Accessed: 27-Sep-2022].

[18] N. J. Smelser and P. B. Baltes, *International Encyclopedia of the Social & Behavioral Sciences*. Amsterdam etc.: Elsevier, 2001.

[19] Scheinker, David, Areli Valencia, and Fatima Rodriguez. "Identification of factors associated with variation in US county-level obesity prevalence rates using epidemiologic vs machine learning models." JAMA Network open 2.4 (2019): e192884-e192884.

[20] P. Singh, R. S. Sawhney, and K. S. Kahlon, "Forecasting the 2016 US presidential elections using sentiment analysis," *SpringerLink*, 01-Jan-1970. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-68557-1_36. [Accessed: 26-Sep-2022].

[21] "Predicting elections: Experts, polls, and fundamentals," *Upenn.edu*. [Online]. Available: https://www.sas.upenn.edu/~baron/journal/18/18124/jdm18124.html. [Accessed: 29-Sep-2022].

[22] R. D. Endsuy, "Sentiment analysis between Vader and EDA for the US presidential election 2020 on Twitter datasets," *Journal of Applied Data Sciences*. [Online]. Available: http://bright-journal.org/Journal/index.php/JADS/article/view/17. [Accessed: 26-Sep-2022].

[23] "Religious landscape study," *Pew Research Center's Religion & Public Life Project*, 13-Jun-2022. [Online]. Available:

https://www.pewresearch.org/religion/religious-landscape-study/compare/party-affiliation/by/state/. [Accessed: 26-Sep-2022].

[24] S. University, "The 2020 U.S. election, issues and challenges," *Stanford News*, 09-Nov-2020. [Online]. Available: https://news.stanford.edu/2020/09/28/2020-u-s-election-issues-challenges/. [Accessed: 26-Sep-2022].

[25] "Sentiment analysis," *Lexalytics*, 06-Jun-2022. [Online]. Available: https://www.lexalytics.com/technology/sentiment-analysis/. [Accessed: 26-Sep-2022].

[26] "Truman defeats Dewey," *History.com*, 09-Feb-2010. [Online]. Available: https://www.history.com/this-day-in-history/truman-defeats-dewey. [Accessed: 26-Sep-2022].

[27] "What Factors Shape Political Attitudes? [ushistory.org]," *Ushistory.org*. [Online]. Available: https://www.ushistory.org/gov/4b.asp. [Accessed: 29-Sep-2022].