Predicting Voting Trends in the United States Based on Demographics with Machine Learning

Ian Jeffries, @00535868
University of Salford, Manchester, United Kingdom

Abstract

This paper is academic in nature and explores the use of two data mining techniques: classification and association rules mining. The same dataset is used for both classification and association rules mining, and thus provides the over-arching theme for this paper. The first section uses classification to predict presidential winners by county in the United States based on voter demographics. The second section attempts to mine association rules to discover underlying trends in voting habits. Following the steps of CRISP-DM methodology, each section will discuss methods for applying correct data mining models and comparisons between R and SAS Enterprise Miner.

CRISP-DM

The CRISP-DM model is a commonly used methodology for data mining experts. It creates a standard blueprint for any data mining project, and is the methodology used in this paper. There will be five parts within each of the three sections that follow the tasks outlined in the CRISP-DM model [1]:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation

The reference model can be seen in Figure 1. [2]

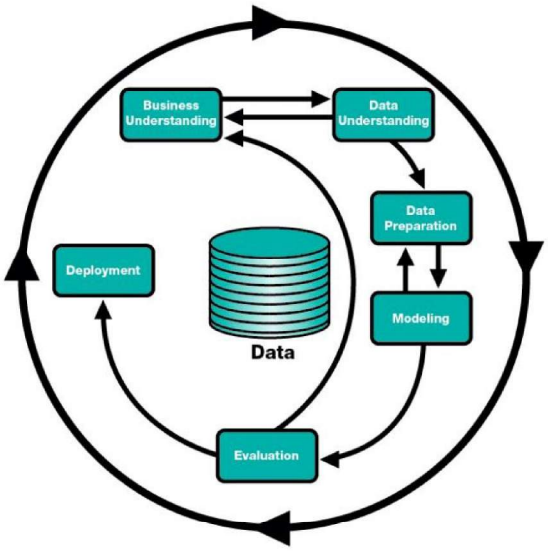

Figure 1: CRISP-DM Reference Model

Section 1: Classification

1. Business Understanding

1.1 Background

Every four years the American people vote on which individual will lead one of the world's greatest superpowers. This transition of power has many implications on world politics, and as a result political parties spend millions of dollars every election cycle.

The question must be asked, where does this money go? The breakdown of spending in the 2016 US election from April 1st to June 30th is seen in Figure 2. [3]

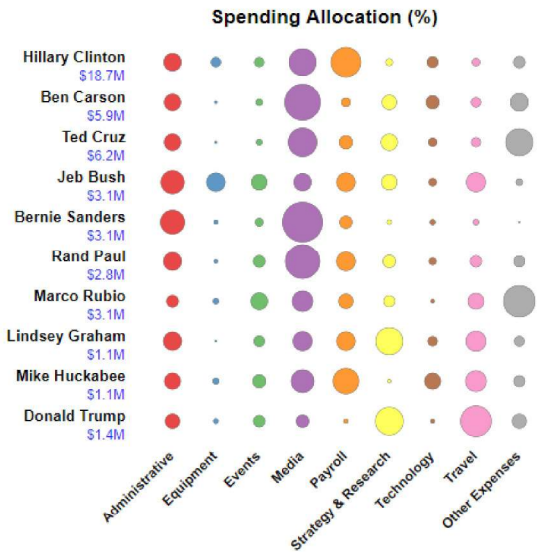

Figure 2: Campaign Spending Allocations from April 1st – June 30th

As seen above, one of the biggest expenditures is in media costs. There is also a substantial amount spent on strategy and research. This implies that campaigns spend a significant part of their budget trying to influence potential voters.

What if Data Science could help narrow their scope even more? How much money could be saved by identifying which way a county will vote based on demographics? If a classification model could be developed, campaigns would know where they have strong support despite changing demographics. This will enable them to focus their funds on potential swing counties and to ignore counties in which they have very little chance of winning.

1.2 Business Objectives

The following study attempts to identify demographics that are strong indicators in voting preference, and to use 2016 election results to train a classification model. This model will then be used to predict election outcomes by county. Three models will be compared, and the most accurate model will be used for deployment. Success will be determined on whether election results can be reasonably predicted by county, with an accuracy over 70%.

1.3 Related Works

Given the importance of American elections and their impact on the global community, many studies have been done to predict election outcomes. A study by Christine Doig-Cardet and Diego Garcia-Olano [4] applied clustering and classification to predict election results by county using 2012 election results. Their study utilizes data from Measure of America [5], rather than the US Census Data used for this study.

Another study by Mohammad Zolghadr, Seyed Niaki, and S. T. A. Niaki [6] apply artificial neural networks (ANN) and support vector regression (SVR) models on election results from 1952 to 2012, using the last three elections to validate the models. Rather than prediction based on demographics, this study focuses on general variables such as unemployment rate, approval rating, and consecutive terms the incumbent party has been in office.

There don't appear to be any well-known studies that use the dataset from this study to predict the 2016 election results. Studies were conducted to predict primary results on Kaggle, [7] but it doesn't seem to have been applied to the 2016 general election.

2. Data Understanding

2.1 Search Strategy

Many websites were researched to try and find an appropriate dataset, including Data.gov, UCI Machine Learning Repository, Google Dataset Search and Kaggle. These sites were searched using terms such as "United States County Demographics" and "2016 Election Results". Ultimately, two separate datasets were selected from Kaggle.

2.2 Data Description

Ben Hamner, who collected the data from a multitude of sources, uploaded the first dataset used in analysis to Kaggle. [8]

Contained within the dataset are the demographics of 3,143 counties in the United States. There was a total of 51 different attributes, with demographics such as age, race, education level, and gender. (The full list of attributes can be found in Appendix A) Economic measures are also listed including median household income and retail sales per capita. All attributes were organized by county and state.

Steve Palley uploaded the second dataset used for this study to Kaggle. [9] This dataset includes the number of votes for each candidate in the 2016 Election by county, along with the total percent of the vote gained and the final winner in each county.

To create a final working dataset, the two sources listed above had to be combined. The county demographics would serve as independent variables, and the election results would theoretically be dependent upon those demographics.

2.3 Data Exploration

The data was explored in preparation for the data cleaning and transformation phase. A quick look at the demographics dataset revealed that state summaries were mixed in with the county level statistics. There didn't seem to be any missing values, but some of the statistics were displayed as a percent and some were integers. Due to this, the data would need to be normalized.

To select only relevant attributes, a scatterplot was created to explore if there was a correlation between certain demographics and the percentage of the county that voted democratic. (This was done after normalization, which is discussed in the next section) Each demographic was plotted on the same scale and fit to a regression line to discover relevant demographics. (This was achieved using the R package ggplot2) [10] The result can be seen in Figure 3 on the next page.

The dataset containing election results had "NA" values in the county column and included duplicate rows for each county as it listed the number of votes for each candidate on a separate line. The data would need to be pivoted to enable the merging of the two datasets based on county.

3. Data Preparation

3.1 Data Selection

Attributes that did not seem to have a correlation to voting preference were removed from the demographics dataset, as seen in Figure 3. This reduced the number of attributes from 51 down to 28.

The relevant attributes from the election results dataset were the "candidate names" and "percent of vote attained" columns. The county and state attributes were used to join to the demographics dataset.

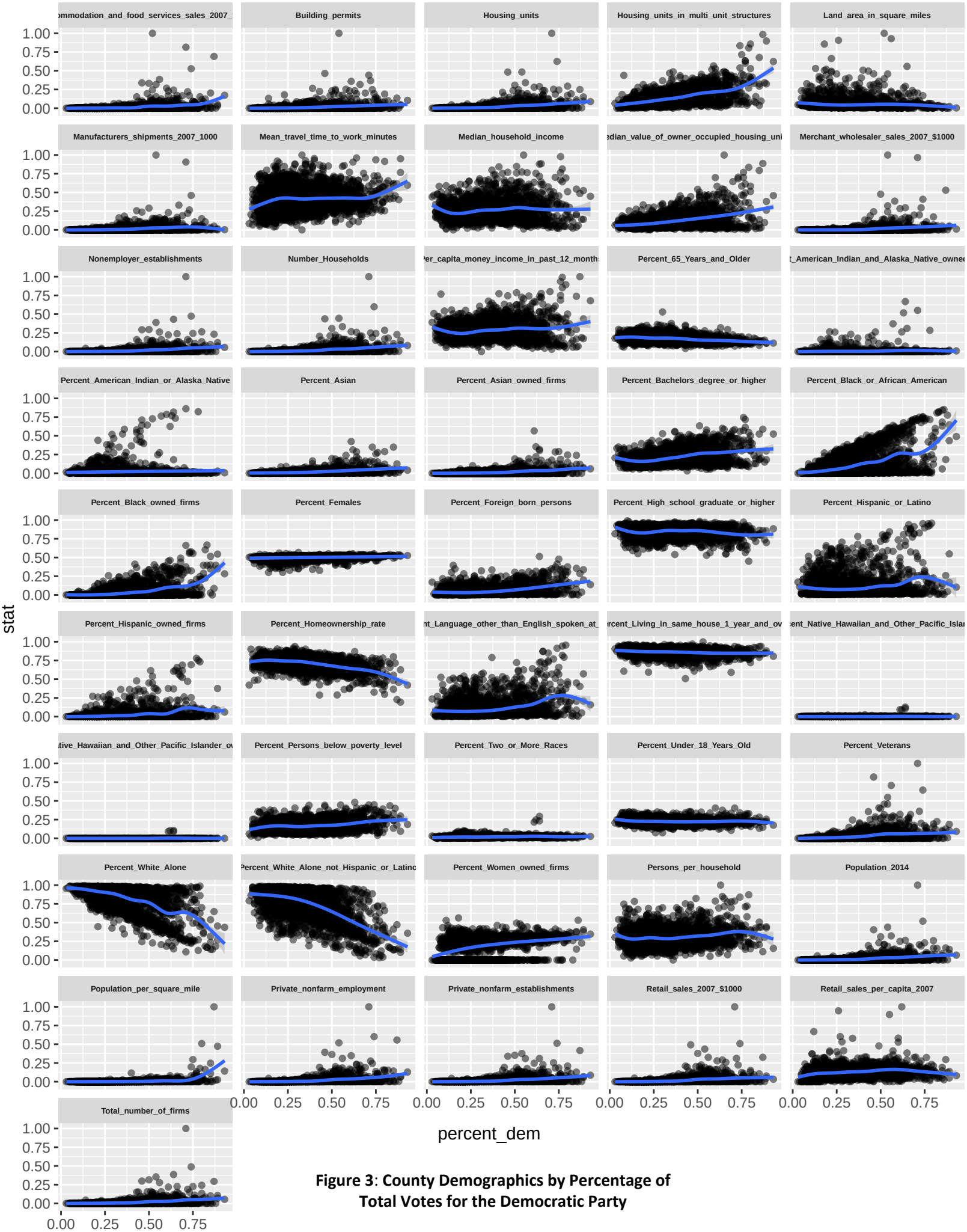

Figure 3: County Demographics by Percentage of Total Votes for the Democratic Party

3.2 Data Cleaning

Rows containing state summary information were isolated in the demographics dataset using R. The scope of this study focuses on county-level results; thus, the state-level demographics were irrelevant and removed from the data. (The full R code can be found in Appendix C)

The election results dataset contained several "NA" values in the "county" column, which were removed from the dataset. Candidates other than Donald Trump or Hillary Clinton were also removed, as there were no third-party winners in any of the counties.

3.3 Data Integration

To merge these two datasets, the "candidates" column in the election results dataset had to be pivoted into separate columns for each candidate, to ensure there was only one row per county. This was achieved using the tidyr package in R. [11]

Once complete, an inner join was performed connecting the county and state attributes in both datasets using the R package "dplyr". [12] All attributes remaining in the demographics dataset were included, taking only "candidate name" and "vote percentage" from the election results dataset.

3.4 Data Formatting

Most demographics were reported as a percentage of the population, while other demographics (median income, persons per household) were reported as integers. To prepare the data for modelling in addition to looking for trends in voting preference, all attributes were normalized on a scale of 0 to 1. The attributes representing a percentage of the population are already on a scale of 0 to 1 and are therefore excluded from the normalization process. Figure 4 contains the formula used to normalize the remaining attributes [13]:

$$A' = \left(\frac{A - \min value\ of\ A}{\max value\ of\ A - \min value\ of\ A}\right) * (D - C) + C$$

Where,

A' contains Min-Max Normalized data one
If pre defined boundary is [C, D]
If A is the range of original data

Figure 4: Formula for data normalization

The column header names were not intuitive for graphing analysis and were replaced by a data dictionary created within excel on the back end.

The final step was to change the candidate names to symmetric binary variables, with "0" representing a vote for the Republican Party and "1" representing a vote for the Democratic Party.

4. Modelling

4.1 Model Selection

To find the most accurate election result predictor, three models were selected for comparison: K-Nearest Neighbour [14], Decision Trees [15], and Artificial Neural Networks. [16] Each of the three models were compared on their overall accuracy, along with their precision percentages. The precision metric was chosen because a false positive could lead to more lost counties, as they were believed to have been secure. The accuracy and precision formulas can be seen in Figure 5.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Figure 5: Formulas for Precision and Accuracy

Models were created and compared using both R and SAS Enterprise Miner.

4.2 Model Building in R

The K-Nearest Neighbour model was built using the "KNN" function from the "class" package. [17] This package uses Euclidean

Distance for classification based on "K" nearest neighbours. Euclidean Distance is calculated as follows:

$$d = \sqrt{\sum_{i=1}^{p} (v_{1i} - v_{2i})^2}$$

Where "d" is equal to the distance between two variables and v1 and v2 are equal to variables one and two.

To find an optimal "K", values 1 to 15 were tested on sample data, using an 80/20 split for training data to test data. The accuracy percentages for each K-value can be seen in Figure 6.

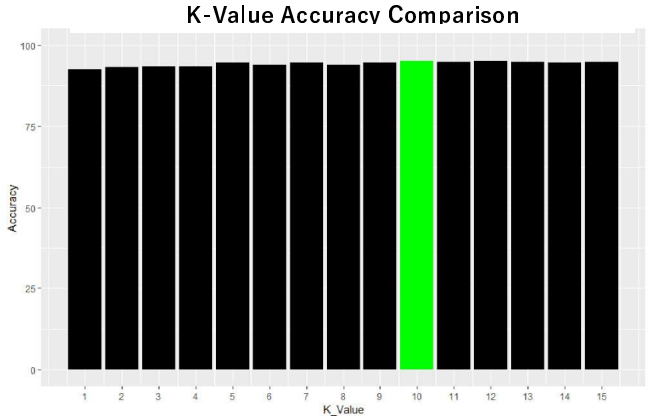

Figure 6: K-Value Accuracy % Compared

As seen above, a "K" value of 10 yielded the highest accuracy at 95.12%. Precision was also compared to see if the results varied dramatically. (See Figure 7) A "K" value of 12 yielded a precision percentage that was a full point above a "K" value of 10, (91.78% compared to 90.79%) while only decreasing accuracy by .15%. As a result, a "K" value of 12 will be used for this study.

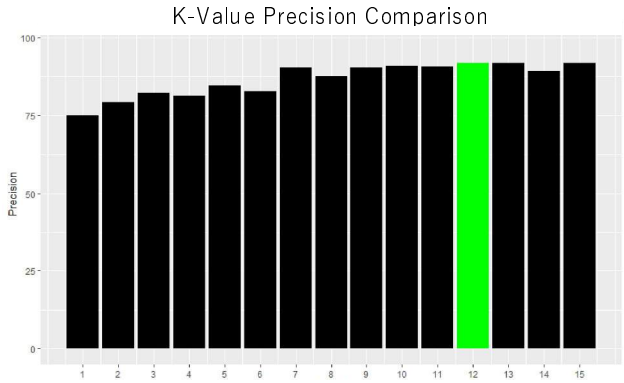

Figure 7: K-Value Precision % Compared

The Decision Tree model was built using the "rpart" [18] and "rpart.plot" [19] packages. An 80/20 split for training data to test data was used to train and then validate the decision tree. The full tree can be seen in Figure 8, with a full-size image available in Appendix B.

As seen in Figure 8, R generated an optimal tree that resulted in 11 leaf nodes. The demographics deemed most important (variables with the highest entropy near the top of the tree) were "Percent White Alone not Hispanic or Latino", "Percent Black or African American", and "Percent Bachelor's Degree or Higher."

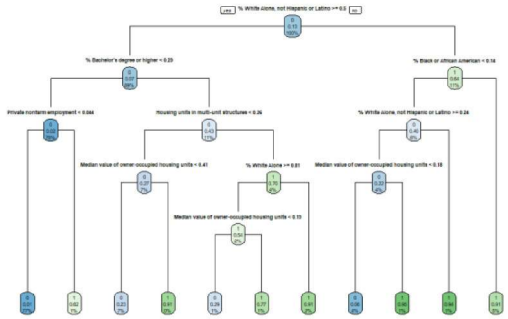

Figure 8: Decision Tree for Voting Preference

The decision tree rules for this model are as follows:

Winner is 0.01 when
- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher < 0.29
- Private nonfarm employment < 0.044

Winner is 0.06 when
- % White Alone, not Hispanic or Latino is 0.24 to 0.50
- Median value of owner-occupied housing units < 0.18
- % Black or African American < 0.14

Winner is 0.23 when
- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher >= 0.29
- Median value of owner-occupied housing units < 0.41
- Housing units in multi-unit structures < 0.26

Winner is 0.29 when
- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher >= 0.29
- Median value of owner-occupied housing units < 0.19
- Housing units in multi-unit structures >= 0.26
- % White Alone >= 0.81

Winner is 0.62 when
- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher < 0.29

- Private nonfarm employment >= 0.044

Winner is 0.77 when

- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher >= 0.29
- Median value of owner-occupied housing units >= 0.19
- Housing units in multi-unit structures >= 0.26
- % White Alone >= 0.81

Winner is 0.91 when

- % White Alone, not Hispanic or Latino < 0.50
- % Black or African American >= 0.14

Winner is 0.91 when

- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher >= 0.29
- Median value of owner-occupied housing units >= 0.41
- Housing units in multi-unit structures < 0.26

Winner is 0.91 when

- % White Alone, not Hispanic or Latino >= 0.50
- % Bachelor's degree or higher >= 0.29
- Housing units in multi-unit structures >= 0.26
- % White Alone < 0.81

Winner is 0.94 when

- % White Alone, not Hispanic or Latino < 0.24
- % Black or African American < 0.14

Winner is 0.95 when

- % White Alone, not Hispanic or Latino is 0.24 to 0.50
- Median value of owner-occupied housing units >= 0.18
- % Black or African American < 0.14

The winner can be determined from the rules above by rounding: 0 = Republican and 1 = Democrat.

The final model, artificial neural networks, was built using the "neuralnet" package, [20] which uses backpropagation to refine its model. (A method of comparing model prediction results to the training data results and refining the model until the difference is minimized) To ensure an accurate model as possible, 13 possible hidden node values were compared, as seen in Figure 9. (A default threshold of .1 was used in all modelling)

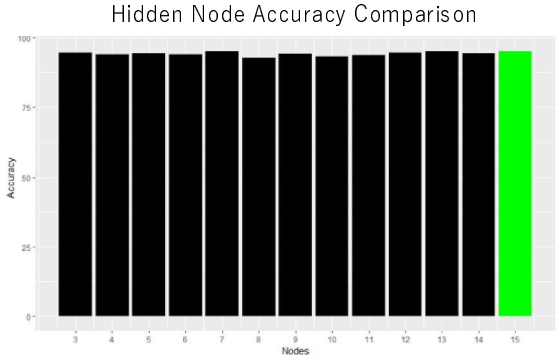

Figure 9: Hidden Node Accuracy % Compared

15 hidden nodes yielded the highest accuracy percentage, but a comparison of precision must be considered as well. (Figure 10) In this case, 3 hidden nodes returned a higher precision than 15 nodes. (88.31% compared to 84.44%) Accuracy only declined by .45%, resulting in a hidden node value of 3 utilized for this study.

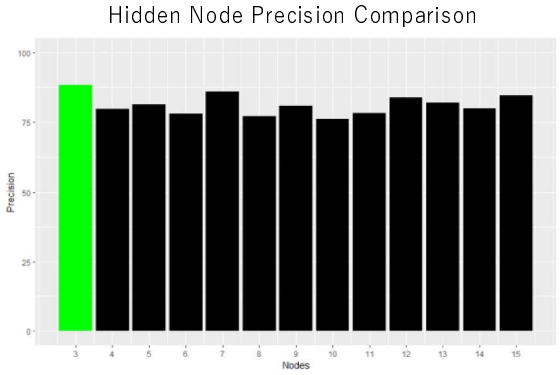

Figure 10: Hidden Node Precision % Compared

4.3 Model Building in SAS Enterprise Miner

Using the cleaned dataset from R, the same classification models were built in SAS Enterprise Miner.

A decision tree was built using the default dataset allocation of 40-30-30 respectively for training, validation, and test datasets.

Variable worth was calculated to obtain an indication of the most relevant attributes. Variables with the highest worth were "Percent White", "Housing Units in Multi-Unit Structures", and "Percent Black or African American". Figure 11 includes the entire breakdown.

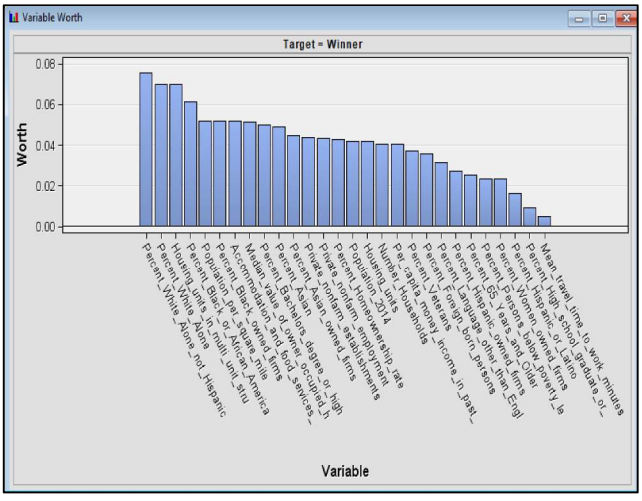

Figure 11: Variable Worth in SAS Miner

A decision tree was created, with a leaf size maximum of 8 and the number of rules set to 5. The final tree is pictured below:

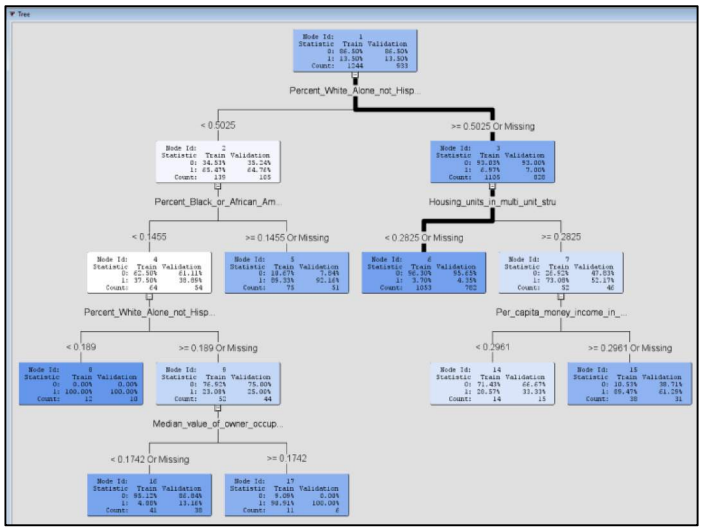

Figure 12: SAS Miner Decision Tree

The rules for this model are as follows:

Node = 5

- if Percent_White_Alone_not_Hispanic < 0.5025
- AND Percent_Black_or_African_America >= 0.1455 or MISSING

  then
  - Tree Node Identifier = 5
  - Number of Observations = 75
  - Predicted: Winner=1 = 0.89
  - Predicted: Winner=0 = 0.11

Node = 6

- if Percent_White_Alone_not_Hispanic >= 0.5025 or MISSING
- AND Housing_units_in_multi_unit_stru < 0.2825 or MISSING

  then
  - Tree Node Identifier = 6
  - Number of Observations = 1053
  - Predicted: Winner=1 = 0.04
  - Predicted: Winner=0 = 0.96

Node = 8

- if Percent_White_Alone_not_Hispanic < 0.189
- AND Percent_Black_or_African_America < 0.1455

  then
  - Tree Node Identifier = 8
  - Number of Observations = 12
  - Predicted: Winner=1 = 1.00
  - Predicted: Winner=0 = 0.00

Node = 14

- if Percent_White_Alone_not_Hispanic >= 0.5025 or MISSING
- AND Per_capita_money_income_in_past_ < 0.29612
- AND Housing_units_in_multi_unit_stru >= 0.2825

  then
  - Tree Node Identifier = 14
  - Number of Observations = 14
  - Predicted: Winner=1 = 0.29
  - Predicted: Winner=0 = 0.71

Node = 15

- if Percent_White_Alone_not_Hispanic >= 0.5025 or MISSING
- AND Per_capita_money_income_in_past_ >= 0.29612 or MISSING
- AND Housing_units_in_multi_unit_stru >= 0.2825

  then
  - Tree Node Identifier = 15
  - Number of Observations = 38
  - Predicted: Winner=1 = 0.89
  - Predicted: Winner=0 = 0.11

Node = 16

- if Percent_White_Alone_not_Hispanic < 0.5025 AND Percent_White_Alone_not_Hispanic >= 0.189 or MISSING
- AND Percent_Black_or_African_America < 0.1455
- AND Median_value_of_owner_occupied_h < 0.17419 or MISSING

  then
  - Tree Node Identifier = 16
  - Number of Observations = 41
  - Predicted: Winner=1 = 0.05
  - Predicted: Winner=0 = 0.95

Node = 17

- if Percent_White_Alone_not_Hispanic < 0.5025 AND Percent_White_Alone_not_Hispanic >= 0.189 or MISSING
- AND Percent_Black_or_African_America < 0.1455
- AND Median_value_of_owner_occupied_h >= 0.17419

  then
  - Tree Node Identifier = 17
  - Number of Observations = 11
  - Predicted: Winner=1 = 0.91
  - Predicted: Winner=0 = 0.09

The KNN model was created using the SAS model "MBR", which classifies based on "K" nearest neighbours. For simplicity, a "K" value of 12 was used to match the model built in R.

The final model, ANN, was created using the SAS model "AutoNeural". This model allows SAS to find optimal configurations for a neural network. The only manual adjustment added was a hidden layer limit of 3, to match the model built in R. The final process flow can be seen in Figure 13.

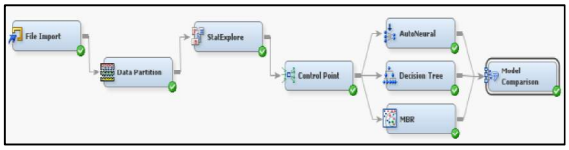

Figure 13: SAS Miner Process Flow

4.4 Model Assessment

A comparison of all three models in R confirms that the best model in terms of accuracy and precision is the K-Nearest Neighbour model, as seen in Figure 14.

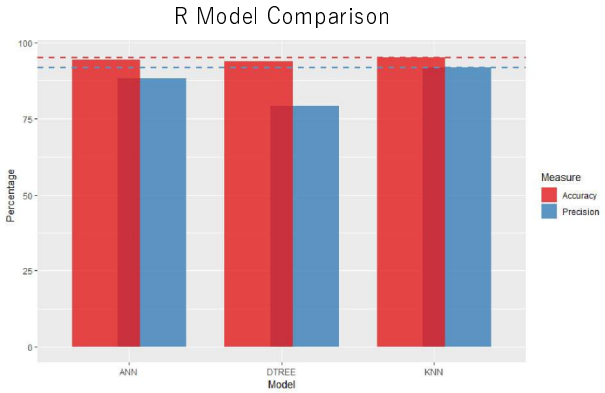

Figure 14: R Model Comparison

If R is being used in further studies, it is recommended that the KNN model be used. Not only does it yield the highest accuracy, but the precision percentage is a full 3 points above the ANN model and almost 12 points higher than the Decision Tree Model.

	KNN	DTREE	ANN
Accuracy	94.97%	93.79%	94.23%
Precision	91.78%	79.12%	88.31%

A comparison of all three models in SAS Miner yielded similar results to R. An initial review of the cumulative lift for each model implies that either the MBR or AutoNeural models produced the highest lift for the longest period. (Cumulative lift represents the benefit of using a predictive model compared to using no model) [21]

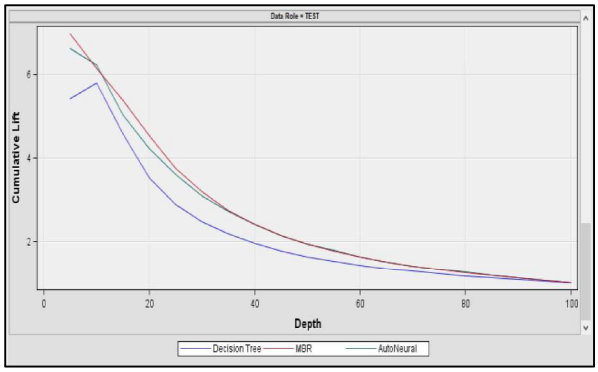

Figure 15: Model Lift Comparison

A comparison of 3 key metrics provides greater clarity on which was the most accurate model:

Data Role	Target Variable	Statistics Label	Tree	MBR	AutoNeural
Test	Winner	Test: Average Squared Error	0.068086	0.047108	0.061753
Test	Winner	Test: Cumulative Lift	5.803023	6.144377	6.223151
Test	Winner	Test: Misclassification Rate	0.075027	0.073955	0.080386

Figure 16: SAS Model Comparison

As seen in Figure 16, MBR had the lowest average squared error in addition to the lowest misclassification rate. Although the AutoNeural model slightly outperformed MBR in terms of cumulative lift, the MBR model was selected for this study due to its greater overall accuracy compared to the other two models.

5. Evaluation

5.1 Evaluate Results

Data mining techniques and CRISP-DM methodology have demonstrated that there is indeed a strong correlation between demographics and voting preference in American elections. The K-Nearest Neighbour model yielded the highest accuracy in terms of classification, which strongly indicates that counties sharing similar demographics vote along similar party lines.

Both SAS Miner and R reveal that race is a strong indicator in voting preference, specifically the percentage of the population that are White or African American. According to the variable worth presented in Figure 11, 3 of the top 4 variables are demographics relating to race. As seen in the decision tree in Figure 12, a county with a population greater than 50% White has an 86.50% chance of voting Republican, based on the training data. The opposite is true of African American populations, indicating that if a county is greater than 14.6% African American there is a roughly 89% chance of voting Democratic.

Several other factors were key indicators in voting preference, such as “Number of

Housing Units in Multi-Unit Structures", which is an indicator of densely populated urban areas. According to the decision tree, counties above a normalized value of .28 Multi-Unit Housing Structures has a 73% chance of voting Democratic. As shown in Figure 17, highly populated counties tended to vote more strongly Democratic.

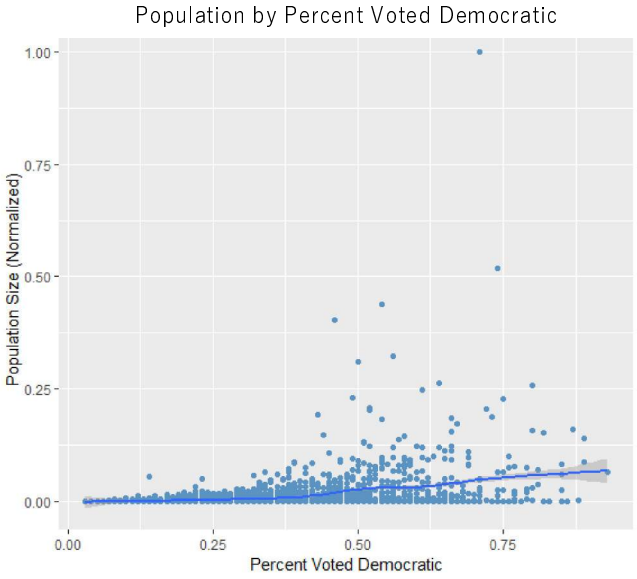

Figure 17: County Population by Percentage of Total Votes for the Democratic Party

The final two attributes discussed relate to economic indicators. "Per Capita Money Income" and "Median Value of Owner Occupied Home" were both high-worth variables, with less wealthy counties voting predominantly Republican. (Figure 12) The voting trends for the key demographic indicators can be seen in Figure 18. (Created using a regression line - shaded portion indicates standard error)

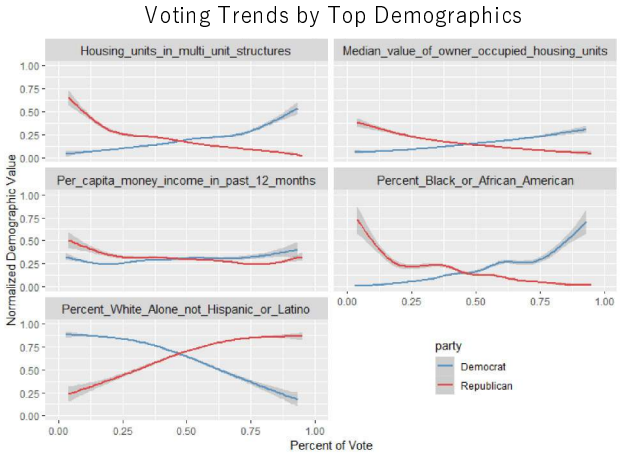

Figure 18: Voting Trends by Key Demographics

The conclusion is that demographics can be used to create an accurate prediction model for election results. The most accurate model yielded roughly 95% accuracy in R and 93% in SAS Miner, well above the 70% threshold. In fact, none of the three models in either software tool was under 90% accuracy. In addition, the top model yielded a precision rate of 91.78%.

Considering the findings of this study, it has been concluded that there is an advantage to employing supervised machine learning to aid in campaign strategy. The next step to fully employ the classification algorithm would be to assign a percent certainty to each county classification, and then map the results to form a comprehensive campaign strategy.

6. Ethical Implications

There are many ethical implications that come with classifying counties by demographics, specifically with how the learning algorithm can be used.

A study done in 2017 showed that strict identification laws have a differentially negative impact on the turnout of racial and ethnic minorities in elections. [22] It is possible that the classification algorithm from this study could be used to identify counties with a high percent of ethnic minorities and impose harsher identification laws to skew the vote.

The political party in power could also use the algorithm to re-draw electoral districts to include demographics that more heavily favour their party, also known as gerrymandering. [23] This tactic can have far-reaching impacts on American politics, shaping the political landscape for years to come.

Section 2: Association Rules Mining

1. Business Understanding

1.1 Background

Now that a classification model has been built using demographics data and election results, the next logical step is to mine the data for interesting association rules. The background will remain the same as Section 1: given the millions of dollars spent on advertising for campaigns each year, machine learning is needed to help narrow the focus on counties that naturally lean to one political party based on demographics.

1.2 Business Objectives

In this study, unsupervised machine learning will be leveraged to find interesting association rules between demographic data and voting preference. Success will be determined on whether interesting rules can be mined with a high degree of support and confidence. The threshold will be determined through testing, but a minimum confidence of 50% is required for a rule to be interesting.

2. Data Understanding

2.1 Data Description

The datasets used in this study are identical to the datasets from Section 1, and all descriptions and search strategies can be found there. This section will focus on data exploration relevant to association rules mining.

2.3 Data Exploration

The datasets were explored and cleaned thoroughly in section 1. The datasets used in this study have already been normalized, "NA" values were removed, and attributes selected only if they had some correlation to voting preference.

Exploring the data further revealed some manipulation would be required to enable association rules mining. All attributes, apart from "county", "state", and "winner", are ratio-scaled numeric attributes. These attributes need to be converted to ordinal attributes to effectively mine for rules.

3. Data Preparation

3.1 Data Selection

All relevant demographic attributes were kept, along with the winner in each county. The only attributes removed from the original classification dataset were the "county" and "state" attributes, as they would not be needed in the mining process.

3.2 Data Formatting

To convert the ratio-scaled numeric attributes to ordinal attributes, a binning method was used. Each of the attributes were sorted into 5 bins, to mine rules that included a range of values. (For example, bin 1 for the "Population_2014" attribute would be from 0_to_0.199) To accomplish this, the "bin_data" function from the "mltools" package in R was used. [24]

There are several options for binning data using this tool. Method 1 is to bin data into an equal number of observations within each group. Method 2 is to bin data into equally spaced groups from min to max, which does not consider how many observations fall into each group. To hedge against the possibility that a large amount of values will fall into one group and thus skew the results, both binning methods were used.

The "bin_data" function automatically grouped the values, but the output was not in a viewer-friendly format. As an example, the value of one of the bins was as follows:

Percent_65_Years_and_Older
[0.000894409486015041, 0.00187242397880161)

The square brackets are inclusive whereas the round brackets are exclusive. To cut down on the number of digits, in addition to enhancing readability, all bins were relabelled to the following format:

Percent_65_Years_and_Older
0.0009_to_0.0018

For simplicity, all decimals were rounded to the 4th digit and the exclusive side of the bin was .0001 less than the next occurring bin. This ensured no overlap and an easily readable format for rule output. (All labels were reformatted using Excel and implemented using the "revalue" function from the "plyr" package in R. [25])

4. Modelling

4.1 Model Selection

The Apriori Algorithm [26] will be used to mine interesting rules from the dataset by pruning small subsets under a certain support measurement. Appropriate confidence and support measures will be set after significant testing on the data. (See Figure 19)

$$\text{Supp}\,(A \Rightarrow B) = \frac{\text{support containing both A and B}}{\text{total of tuples}}$$

$$\text{Conf}\,(A \Rightarrow B) = \frac{\text{support containing both A and B}}{\text{tuples containing A}}$$

Figure 19: Support and Confidence Formulas

4.2 Model Building in R

The first step for model creation in R was to determine which binning method should be used for modelling. There would be a disproportionally high support for any bin that has a large N, so it would make sense to use the method that places an equal amount of values in each bin. Two charts were created for visual confirmation.

Method 1 was meant to bin values into equal sized bins. As seen in Figure 1, most max bin sizes for any demographic hovers around 20% of the observations. This would make sense as there are 5 equally sized bins – the max of any given bin should be 20%. There are 5 attributes that are above a 25% threshold, which have been highlighted in red. (Figure 20) Closer inspection revealed that three of these attributes, "Percent Black owned firms", "Percent Asian owned firms", and "Percent Hispanic owned firms", had a high number of 0 values, exceeding 50% of total instances. This resulted in only 2 bins being created for each of these attributes. As a result, they were be removed from the dataset. Only 2 bins would result in a high support, but not very useful information. The attributes "Percent Women owned firms" and "Percent Asian" also resulted in fewer bins than 5, and consequently were removed.

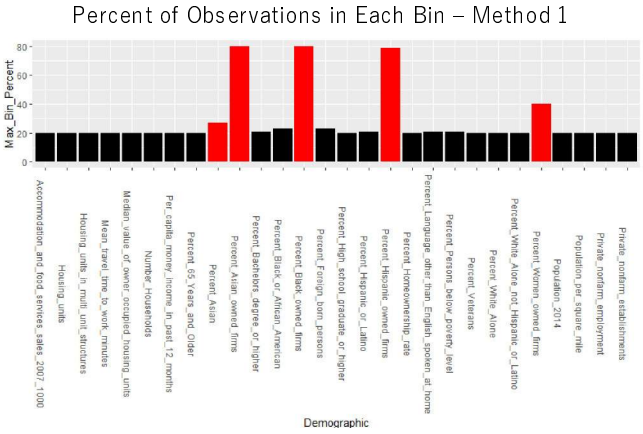

Figure 20: Max Bin Comparison for Method 1

Method 2 binned values on equally spaced bin lengths, not taking into account how many values fell into each bin. The distribution of the max bin percentage can be seen in Figure 21.

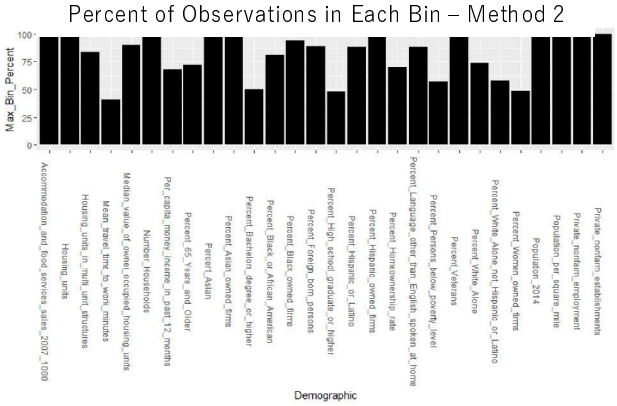

Figure 21: Max Bin Comparison for Method 2

As seen above, most attributes have as many as 100% of the total observations in one bin. It is apparent at a glance that this binning method will not work to mine useful association rules, due to the high concentration of values in only one bin. Given these findings, Method 1 will be used for model creation.

Initial testing also confirmed that separate rules would need to be created for Democratic counties and Republican counties. Only 13.5% of all counties voted Democratic, and as a result a higher support only returns rules with "Winner=Republican" on the right-hand side.

To find the optimal parameters for confidence and support, "ruleExplorer" was used from the "arulesViz" package. [27] This allows for interactive rule adjustment within the Shiny Server. For all rules, a minimum number of items within each rule was set to 2, with a maximum of 5. (A minimum of 2 was selected based on certain demographics having great impact on voting preference)

In the interest of readability, an "N" number of rules equal to 15 would be ideal. There were over 100 rules with a minimum confidence of 90% and support of 15% for "Winner=Republican". Changing these metrics slightly to 18.9% support and 92.3% confidence resulted in 15 rules.

The support threshold was lowered to find rules relating to democratic counties, due to the low number of total incidents. After thorough testing, a support metric of 3.1% and a confidence metric of 81.7% yielded 16 rules. The rules generated in R can be seen in Figure 24 on the next page.

4.3 Model Building in SAS Miner

To build the model in SAS Miner, the data had to be rearranged into a format that SAS can interpret. Every demographic column was combined into two key-value pairs, with each county labelled as an integer ranging from 1 to 3110. This essentially created a "market basket" format, as seen in Figure 22.

County	Demographic
1	Population_2014: 0.0036_to_0.0091
1	Percent_65_Years_and_Older: 0.041_to_0.1417
1	Percent_White_Alone: 0.759_to_0.8915
1	Percent_Black_or_African_American: 0.152_to_0.851
1	Percent_Hispanic_or_Latino: 0.018_to_0.0289
1	Number_Households: 0.0044_to_0.0106
1	Winner: Republican

Figure 22: Example of New Data Format for SAS Miner

Once the data was imported into SAS Miner, two separate tracks for each political party were created to mine the data using the "Association Node". The initial confidence and support parameters were set to equal the parameters set in R, which returned 40 rules for the Republican Party and only 1 rule for the Democratic Party. Some trial and error resulted in a Confidence threshold of 94% and Support threshold of 20% returning 16 rules for the Republican Party. A Confidence of 71% and Support of 2% resulted in 14 rules for the Democratic Party. The rules generated along with the process flow can be seen in Figures 23, 25 and 26.

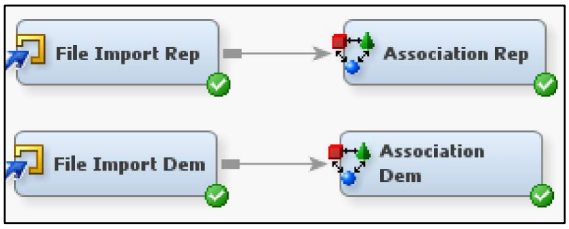

Figure 23: SAS Miner Process Flow

Association Rules Created in R

Rule #	LHS	RHS	confidence	lift	support	count
1	{Percent_White_Alone=0.108_to_0.7589,Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.82	6.10	0.04	112
2	{Percent_White_Alone=0.108_to_0.7589,Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Percent_Veterans=0.0221_to_1}	{Winner=Democrat}	0.82	6.10	0.03	98
3	{Percent_White_Alone=0.108_to_0.7589,Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985}	{Winner=Democrat}	0.82	6.06	0.04	113
4	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.83	6.17	0.04	115
5	{Percent_White_Alone=0.108_to_0.7589,Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Percent_Homeownership_rate=0.194_to_0.6679,Housing_units_in_multi_unit_structures=0.178_to_0.985}	{Winner=Democrat}	0.82	6.08	0.03	101
6	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Percent_Homeownership_rate=0.194_to_0.6679,Housing_units_in_multi_unit_structures=0.178_to_0.985,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.84	6.22	0.03	100
7	{Percent_Foreign_born_persons=0.066_to_0.513,Percent_Homeownership_rate=0.194_to_0.6679,Median_value_of_owner_occupied_housing_units=0.1445_to_1,Accommodation_and_food_services_sales_2007_1000=0.0057_to_1}	{Winner=Democrat}	0.82	6.09	0.03	97
8	{Percent_Bachelors_degree_or_higher=0.254_to_0.744,Percent_Homeownership_rate=0.194_to_0.6679,Median_value_of_owner_occupied_housing_units=0.1445_to_1,Accommodation_and_food_services_sales_2007_1000=0.0057_to_1}	{Winner=Democrat}	0.82	6.10	0.03	98
9	{Percent_White_Alone=0.108_to_0.7589,Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.88	6.50	0.03	100
10	{Population_2014=0.0092_to_1,Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.82	6.08	0.03	101
11	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Percent_Veterans=0.0221_to_1,Housing_units_in_multi_unit_structures=0.178_to_0.985,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.83	6.17	0.03	100
12	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units=0.0116_to_1,Housing_units_in_multi_unit_structures=0.178_to_0.985,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.83	6.14	0.03	102
13	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Number_Households=0.0107_to_1,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.83	6.14	0.03	102
14	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Private_nonfarm_establishments=0.0079_to_1,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.82	6.09	0.03	102
15	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Private_nonfarm_employment=0.0076_to_1,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.82	6.05	0.03	103
16	{Percent_White_Alone_not_Hispanic_or_Latino=0.031_to_0.6147,Housing_units_in_multi_unit_structures=0.178_to_0.985,Accommodation_and_food_services_sales_2007_1000=0.0057_to_1,Population_per_square_mile=0.0023_to_1}	{Winner=Democrat}	0.82	6.09	0.03	102
1	{Housing_units_in_multi_unit_structures=0.055_to_0.0819}	{Winner=Republican}	0.95	1.09	0.19	588
2	{Percent_White_Alone_not_Hispanic_or_Latino=0.888_to_0.9419}	{Winner=Republican}	0.96	1.11	0.19	598
3	{Per_capita_money_income_in_past_12_months=0.1928_to_0.2387}	{Winner=Republican}	0.95	1.10	0.19	590
4	{Per_capita_money_income_in_past_12_months=0.2388_to_0.2825}	{Winner=Republican}	0.95	1.10	0.19	594
5	{Percent_White_Alone=0.9652_to_0.993}	{Winner=Republican}	0.98	1.13	0.20	608
6	{Percent_White_Alone_not_Hispanic_or_Latino=0.942_to_0.986}	{Winner=Republican}	0.98	1.14	0.20	614
7	{Median_value_of_owner_occupied_housing_units=0.0692_to_0.0975}	{Winner=Republican}	0.96	1.10	0.19	597
8	{Percent_Homeownership_rate=0.752_to_0.7839}	{Winner=Republican}	0.96	1.11	0.19	598
9	{Percent_65_Years_and_Older=0.183_to_0.2079}	{Winner=Republican}	0.95	1.10	0.19	598
10	{Percent_65_Years_and_Older=0.208_to_0.529}	{Winner=Republican}	0.96	1.11	0.19	601
11	{Percent_Homeownership_rate=0.784_to_0.938}	{Winner=Republican}	0.96	1.11	0.19	602
12	{Percent_White_Alone=0.942_to_0.9651}	{Winner=Republican}	0.95	1.10	0.19	596
13	{Percent_Homeownership_rate=0.718_to_0.7519}	{Winner=Republican}	0.94	1.08	0.19	589
14	{Percent_Bachelors_degree_or_higher=0.16_to_0.1939}	{Winner=Republican}	0.95	1.09	0.19	606
15	{Percent_Foreign_born_persons=0.01_to_0.0189}	{Winner=Republican}	0.95	1.10	0.21	664

Figure 24: Association Rules Created in R

Rule Index	Left Hand of Rule	Right Hand of Rule	Confidence (%)	Lift	Support (%)	Transaction Count
1	Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147 & Percent_Bachelors_degree_or_higher: 0.254_to_0.744	Winner: Democrat	83.48	6.18	3.09	96.00
2	Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147 & Median_value_of_owner_occupied_housing_units: 0.1445_to_1	Winner: Democrat	81.42	6.03	2.96	92.00
3	Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147 & Per_capita_money_income_in_past_12_months: 0.3375_to_1	Winner: Democrat	78.31	5.80	2.09	65.00
4	Population_per_square_mile: 0.0023_to_1 & Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147	Winner: Democrat	77.11	5.71	4.12	128.00
5	Mean_travel_time_to_work_minutes: 0.5361_to_1 & Housing_units_in_multi_unit_structures: 0.178_to_0.985	Winner: Democrat	76.32	5.65	1.86	58.00
6	Percent_White_Alone: 0.108_to_0.7589 & Percent_Bachelors_degree_or_higher: 0.254_to_0.744	Winner: Democrat	74.22	5.50	3.05	95.00
7	Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147 & Housing_units_in_multi_unit_structures: 0.178_to_0.985	Winner: Democrat	74.18	5.49	4.34	135.00
8	Percent_White_Alone: 0.108_to_0.7589 & Housing_units_in_multi_unit_structures: 0.178_to_0.985	Winner: Democrat	73.18	5.42	4.21	131.00
9	Percent_White_Alone: 0.108_to_0.7589 & Per_capita_money_income_in_past_12_months: 0.3375_to_1	Winner: Democrat	73.03	5.41	2.09	65.00
10	Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147 & Accommodation_and_food_services_sales_2007_1000: 0.0057_to_1	Winner: Democrat	71.51	5.30	3.95	123.00
11	Private_nonfarm_establishments: 0.0079_to_1 & Percent_White_Alone_not_Hispanic_or_Latino: 0.031_to_0.6147	Winner: Democrat	71.43	5.29	3.86	120.00
12	Percent_White_Alone: 0.108_to_0.7589 & Percent_Foreign_born_persons: 0.066_to_0.513	Winner: Democrat	71.43	5.29	3.05	95.00
13	Percent_White_Alone: 0.108_to_0.7589 & Median_value_of_owner_occupied_housing_units: 0.1445_to_1	Winner: Democrat	71.19	5.27	2.70	84.00
14	Percent_Homeownership_rate: 0.194_to_0.6679 & Per_capita_money_income_in_past_12_months: 0.3375_to_1	Winner: Democrat	71.11	5.27	3.09	96.00

Figure 25: SAS Miner Democrat Association Rules

Rule Index	Left Hand of Rule	Right Hand of Rule	Confidence (%)	Lift	Support(%)	Transaction Count
73	Percent_White_Alone_not_Hispanic_or_Latino: 0.942_to_0.986	Winner: Republican	98.40	1.14	19.74	614.00
74	Percent_White_Alone: 0.9652_to_0.993	Winner: Republican	97.75	1.13	19.55	608.00
75	Percent_White_Alone_not_Hispanic_or_Latino: 0.888_to_0.9419	Winner: Republican	96.30	1.11	19.23	598.00
76	Percent_Homeownership_rate: 0.784_to_0.938	Winner: Republican	96.01	1.11	19.36	602.00
77	Percent_65_Years_and_Older: 0.208_to_0.529	Winner: Republican	95.85	1.11	19.32	601.00
78	Percent_Homeownership_rate: 0.752_to_0.7839	Winner: Republican	95.68	1.11	19.23	598.00
79	Median_value_of_owner_occupied_housing_units: 0.0692_to_0.0975	Winner: Republican	95.52	1.10	19.20	597.00
80	Per_capita_money_income_in_past_12_months: 0.2388_to_0.2825	Winner: Republican	95.50	1.10	19.10	594.00
81	Percent_65_Years_and_Older: 0.183_to_0.2079	Winner: Republican	95.37	1.10	19.23	598.00
82	Percent_Bachelors_degree_or_higher: 0.129_to_0.1599	Winner: Republican	95.11	1.10	18.75	583.00
83	Percent_White_Alone: 0.942_to_0.9651	Winner: Republican	95.06	1.10	19.16	596.00
84	Housing_units_in_multi_unit_structures: 0_to_0.0549	Winner: Republican	94.88	1.10	18.49	575.00
85	Percent_Foreign_born_persons: 0.01_to_0.0189	Winner: Republican	94.86	1.10	21.35	664.00
86	Per_capita_money_income_in_past_12_months: 0.1928_to_0.2387	Winner: Republican	94.86	1.10	18.97	590.00
87	Percent_Bachelors_degree_or_higher: 0.16_to_0.1939	Winner: Republican	94.69	1.09	19.49	606.00
88	Housing_units_in_multi_unit_structures: 0.055_to_0.0819	Winner: Republican	94.69	1.09	18.91	588.00
89	Percent_Language_other_than_English_spoken_at_home: 0_to_0.0249	Winner: Republican	94.27	1.09	18.52	576.00

Figure 26: SAS Miner Republican Association Rules

4.3 Model Assessment

Testing many different Confidence and Support thresholds was required in both R and SAS Miner to get as close to 15 rules as possible. As a result, R created 15 rules for the Republican Party and 16 rules for the Democratic Party, while SAS Miner created 16 rules for the Republican Party and 14 rules for the Democratic Party. As seen in Figures 24 and 26, many of the rules for the Republican Party are similar between the two platforms. The rules for the Democratic Party vary more significantly, due to the "maximum items" in a rule being set to 3 in SAS Miner and 5 in R. There did not appear to be a way to limit to one item on the right-hand side of the rule in SAS Miner, resulting in other demographics combining with voting preference on the right-hand side. To eliminate this, the item threshold had to be set to 3.

Visualizing the rules in R using the "arulesViz" package [27] enables the identification of the most important rules, as seen in Figure 27.

There are three rules that could be considered the most interesting within the Republican Association Rules, due to higher than average Confidence and Support measures. (Rules 3, 6 and 15, which can be referenced using the table on page 14)

The Democratic Rules Scatterplot produced 4 extremely interesting rules. (Rules 1, 3, 4 and 9)

Plotting the results in 3 dimensions using SAS Miner enables the addition of Lift as a parameter. (Lift = confidence of rule / expected confidence of rule) This allows an even deeper look at the association rules, identifying three potentially valuable outlying rules for the Republican Party. (Figure 28 on the next page)

A 3D plot for the Democratic Party doesn't reveal any clear outliers; but two rules were selected for their high Confidence metrics and a cluster of 4 rules were selected for their high Support metrics. (Rule indexes can be found from the SAS Miner rules table in Figures 25 and 26 on page 15)

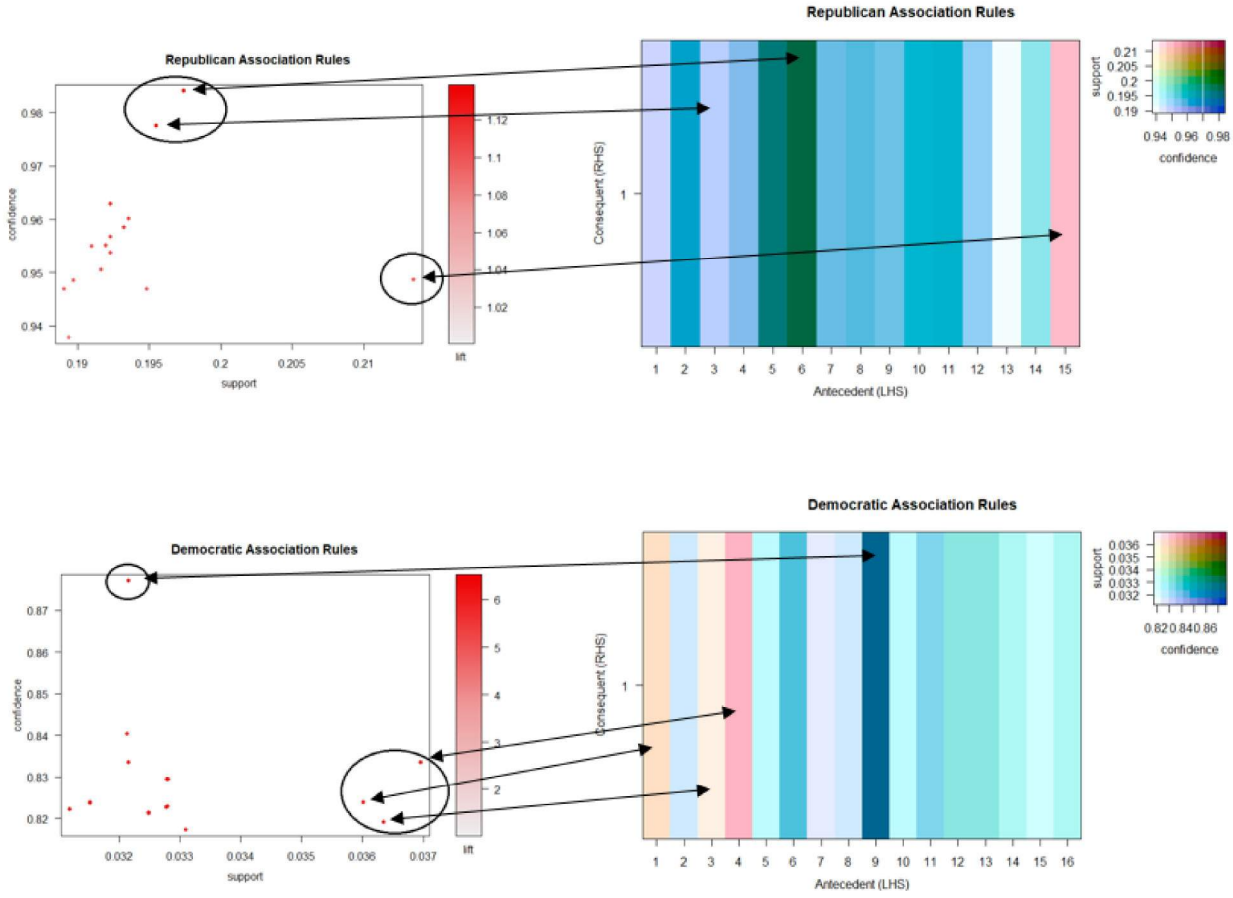

Figure 27: Outlying Rules by Rule Number

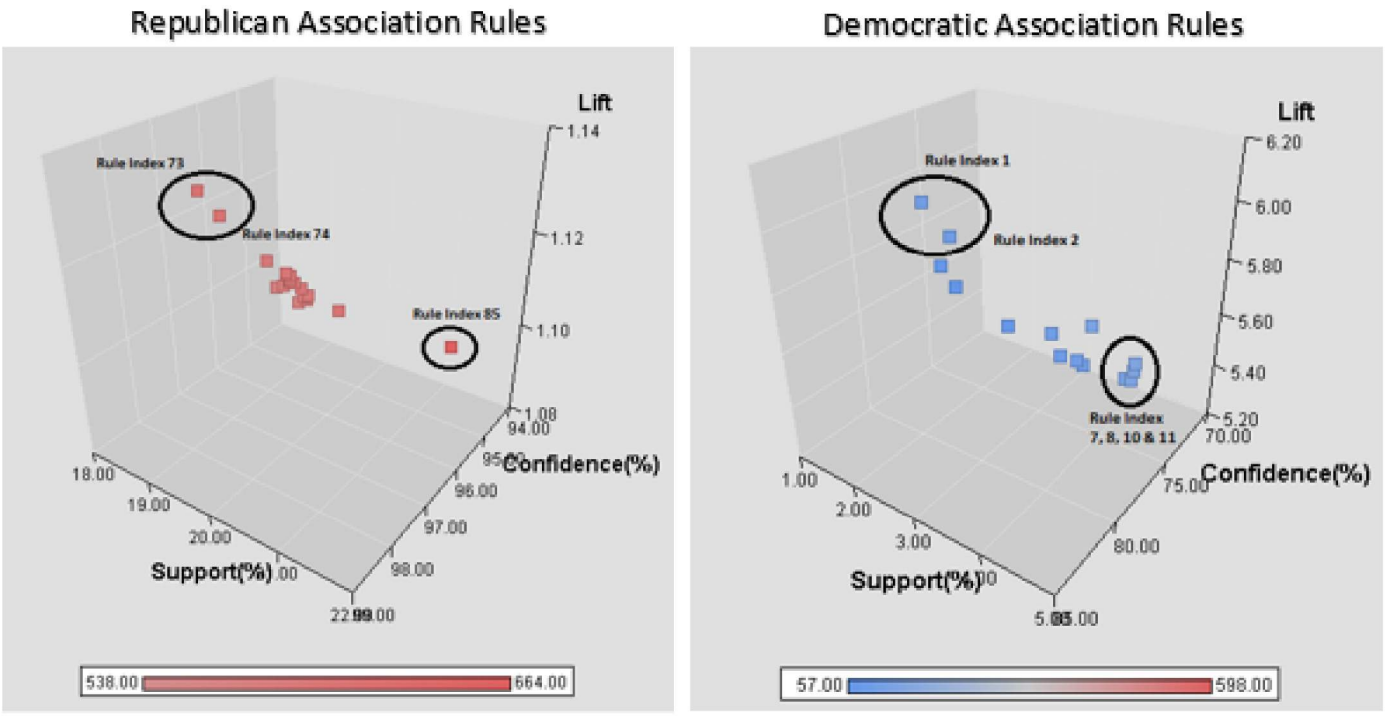

Figure 28: 3D Plot of Rules in SAS Miner

5. Evaluation

5.1 Evaluate Results

Employing unsupervised machine learning to mine for Association Rules revealed once again that race and population size are key indicators in voting preference.

As seen in Figure 24 on page 14, 14 out of 16 of the Democratic Rules contained the attribute "Percent White", indicating that a county must typically be below 61.4% white to vote Democratic. Many of these rules combined with high values for "Housing in Multi Unit Structures", "Population per Square Mile", and "Private Non-Farm Establishments", all of which are indications of densely populated urban areas. In fact, all four of the outlying Democratic rules contained some combination of these demographics.

Two attributes showed up in the Democratic Rules that were not deemed important from the classification study: "Percent Veterans" and "Percent Bachelor's Degree or Higher". Based on rules 2 and 11 in R, counties with a White Population under 61% and a veteran population from 2% - 100% vote Democratic, with an almost 83% Confidence. According to rule 8 in R, higher educated counties tend to vote more strongly Democratic.

SAS Miner revealed similar rules for the Democratic Party, with a lower confidence. This is due to the complexities of the rule length in R, which was not replicated in SAS Miner. One attribute unique to SAS was "Mean Travel Time to Work: .54_to_1" (Rule Index 5), which is yet another indicator of densely populated counties.

The rules generated for the Republican Party are much more straightforward, which implies that just a few key demographics have a very strong influence on Republican voting preference, compared to a complex combination of demographics that influence Democratic voting preference.

According to the rule table generated in R, counties with a high white population (ranging from 88% to 99%) have a very high confidence of voting Republican, with some rules as high as 98% confident. Two of the

most interesting rules (6 and 15) relate to race, with high Confidence in high White populations and low Foreign-Born populations. The third most interesting rule (3) shows that poorer counties (.19 to .24 normalized) tend to vote Republican with a 95% confidence. An interesting rule reveals that poorly educated counties (Bachelor's Degree or Higher = 16% to 19%) also vote Republican, with a confidence of 95%. SAS Miner returned rule almost identical to those created in R.

In conclusion, it has been shown that interesting rules with a high degree of confidence can be mined using demographics data. The Democratic rules have a Confidence well above the 50% threshold, with a high Lift metric. The support metrics are low for these rules, due to the low percentage of counties that voted Democratic overall.

The Republican Rules have extremely high Confidence metrics (all above 94%) and a Support above 19%. These rules have a much lower Lift than their Democratic counterparts but are still above 1.

Unsupervised machine learning combined with classification has revealed voting patterns based on demographics with a high degree of confidence. This can be utilized within a campaign to focus on counties at a much more granular level.

References

[1] Shearer, C. (2000), 'The CRISP-DM Model: The New Blueprint for Data Mining', *Journal of Data Warehousing, Volume 5, Number,* pp. 13-22

[2] Crisp-dm.eu. (2018). *CRISP-DM by Smart Vision Europe » Data Mining Phases*. [online] Available at: http://crisp-dm.eu/reference-model/ [Accessed 15 Nov. 2018].

[3] *Where Does All That Campaign Money Go? [Interactive Graphic]*. (2016, March 08). Retrieved November 15, 2018, from http://metrocosm.com/where-does-all-that-campaign-money-go-interactive-graphic/

[4] DOIG-CARDET, C., & GARCIA-OLANO, D. (n.d.). Http://diegoolano.com/reports/US-2012election-blueislands.pdf. Retrieved December 5, 2018, from http://diegoolano.com/reports/US-2012election-blueislands.pdf

[5] *Measure of America: A Program of the Social Science Research Council.* (n.d.). Retrieved December 5, 2018, from http://www.measureofamerica.org/

[6] Zolghadr, M., Niaki, S. A., & Niaki, S. (2017). Modeling and forecasting US presidential election using learning algorithms. *Journal of Industrial Engineering International*, 14(3), 491-500. doi:10.1007/s40092-017-0238-2

[7] Hamner, B. (2016, July 01). *2016 US Election*. Retrieved December 5, 2018, from https://www.kaggle.com/benhamner/2016-us-election/kernels

[8] Hamner, B. (2016, July 01). *2016 US Election*. Retrieved November 15, 2018, from https://www.kaggle.com/benhamner/2016-us-election/home

[9] Palley, S. (2016, November 19). *2016 US Presidential Election Vote By County*. Retrieved November 15, 2018, from https://www.kaggle.com/stevepalley/2016uspresidentialvotebycounty/home

[10] Wickham, H. (n.d.). Ggplot2 v3.1.0. Retrieved November 15, 2018, from https://www.rdocumentation.org/packages/ggplot2/versions/3.1.0

[11] Wickham, H. (n.d.). Tidyr v0.8.2. Retrieved November 15, 2018, from https://www.rdocumentation.org/packages/tidyr/versions/0.8.2

[12] Wickham, H. (n.d.). Dplyr v0.5.0. Retrieved November 18, 2018, from https://www.rdocumentation.org/packages/dplyr/versions/0.5.0

[13] Patro, S. K., & Sahu, K. K. (2015). *Normalization: A Preprocessing Stage*. Iarjset, 20-22. doi:10.17148/iarjset.2015.2305

[14] Guo G., Wang H., Bell D., Bi Y., Greer K. (2003) KNN Model-Based Approach in Classification. In: Meersman R., Tari Z., Schmidt D.C. (eds) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. *Lecture Notes in Computer Science*, vol 2888. Springer, Berlin, Heidelberg

[15] Quinlan, J.R. (1986) *Induction of Decision Trees, Machine Learning Volume 1*, Issue 1, pp 81 – 106

[16] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117. doi:10.1016/j.neunet.2014.09.003

[17] Ripley, B. (n.d.). Class v7.3-14. Retrieved November 17, 2018, from https://www.rdocumentation.org/packages/class/versions/7.3-14

[18] Atkinson, B. (n.d.). Rpart v4.1-13. Retrieved November 17, 2018, from https://www.rdocumentation.org/packages/rpart/versions/4.1-13

[19] Milborrow, S. (n.d.). Rpart.plot v3.0.5. Retrieved November 19, 2018, from

https://www.rdocumentation.org/packages/rpart.plot/versions/3.0.5

[20] Guenther, F. (n.d.). Neuralnet. Retrieved November 19, 2018, from https://www.rdocumentation.org/packages/neuralnet/versions/1.33/topics/neuralnet-package

[21] Naseri, M. B., & Elliott, G. (2010). A Comparative Analysis of Artificial Neural Networks and Logistic Regression. *Journal of Decision Systems*, 19(3), 291-312. doi:10.3166/jds.19.291-312

[22] Hajnal, Z., Lajevardi, N., & Nielson, L. (2017). Voter identification laws and the suppression of minority votes. *The Journal of Politics*, 79(2), 363. doi:http://dx.doi.org/10.1086/688343

[23] Engstrom, E. J. (2016). *Partisan gerrymandering and the construction of American democracy*. Ann Arbor: University of Michigan Press. doi:file:///C:/Users/SAP420/Downloads/649968.pdf

[24] Gorman, B. (n.d.). Mltools v0.3.5. Retrieved November 27, 2018, from https://www.rdocumentation.org/packages/mltools/versions/0.3.5

[25] Wickham, H. (n.d.). Plyr v1.8.4. Retrieved November 27, 2018, from https://www.rdocumentation.org/packages/plyr/versions/1.8.4

[26] Bhargava, M., & Selwal, A. (2013). Association Rule mining using Apriori Algorithm: A review. *International Journal of Advanced Research in Computer Science*, 4(2), 327-329. Retrieved November 27, 2018, from www.ijarcs.info.

[27] Hahsler, M. (n.d.). ArulesViz v1.3-1. Retrieved November 29, 2018, from https://www.rdocumentation.org/packages/arulesViz/versions/1.3-1

Appendix A: Full Attribute List

Attribute Name	Description
AFN120207	Accommodation and food services sales, 2007 ($1,000)
AGE135214	Persons under 5 years, percent, 2014
AGE295214	Persons under 18 years, percent, 2014
AGE775214	Persons 65 years and over, percent, 2014
area_name	County Name
BPS030214	Building permits, 2014
BZA010213	Private nonfarm establishments, 2013
BZA110213	Private nonfarm employment, 2013
BZA115213	Private nonfarm employment, percent change, 2012-2013
EDU635213	High school graduate or higher, percent of persons age 25+, 2009-2013
EDU685213	Bachelor's degree or higher, percent of persons age 25+, 2009-2013
HSD310213	Persons per household, 2009-2013
HSD410213	Households, 2009-2013
HSG010214	Housing units, 2014
HSG096213	Housing units in multi-unit structures, percent, 2009-2013
HSG445213	Homeownership rate, 2009-2013
HSG495213	Median value of owner-occupied housing units, 2009-2013
INC110213	Median household income, 2009-2013
INC910213	Per capita money income in past 12 months (2013 dollars), 2009-2013
LFE305213	Mean travel time to work (minutes), workers age 16+, 2009-2013
LND110210	Land area in square miles, 2010
MAN450207	Manufacturers shipments, 2007 ($1,000)
NES010213	Nonemployer establishments, 2013
Percent_Dem	Percent Democratic
Percent_Rep	Percent Republican
POP010210	Population, 2010
POP060210	Population per square mile, 2010
POP645213	Foreign born persons, percent, 2009-2013
POP715213	Living in same house 1 year & over, percent, 2009-2013
POP815213	Language other than English spoken at home, pct age 5+, 2009-2013
PST040210	Population, 2010 (April 1) estimates base
PST045214	Population, 2014 estimate
PST120214	Population, percent change - April 1, 2010 to July 1, 2014
PVY020213	Persons below poverty level, percent, 2009-2013
RHI125214	White alone, percent, 2014
RHI225214	Black or African American alone, percent, 2014
RHI325214	American Indian and Alaska Native alone, percent, 2014
RHI425214	Asian alone, percent, 2014
RHI525214	Native Hawaiian and Other Pacific Islander alone, percent, 2014
RHI625214	Two or More Races, percent, 2014
RHI725214	Hispanic or Latino, percent, 2014
RHI825214	White alone, not Hispanic or Latino, percent, 2014
RTN130207	Retail sales, 2007 ($1,000)
RTN131207	Retail sales per capita, 2007
SBO001207	Total number of firms, 2007
SBO015207	Women-owned firms, percent, 2007
SBO115207	American Indian- and Alaska Native-owned firms, percent, 2007
SBO215207	Asian-owned firms, percent, 2007
SBO315207	Black-owned firms, percent, 2007
SBO415207	Hispanic-owned firms, percent, 2007
SBO515207	Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007
SEX255214	Female persons, percent, 2014
state_abbreviation	State
VET605213	Veterans, 2009-2013
WTN220207	Merchant wholesaler sales, 2007 ($1,000)

Appendix B: R Decision Tree