Playing the Trump Card: An Analysis of Voting Data from the 2016 Presidential Elections at the County Level

Jeff Austin (V00856801), Iain Emslie (V00825434), Jason Hu (V00841056)

In partial fulfillment of the academic requirements of SENG474
Department of Computer Science, University of Victoria
December 5th, 2018

Abstract

The 2016 presidential election in the United States was one of the most controversial in recent history. Many pundits and voters were surprised by the result and the election of Donald Trump. Polls which predicted a high probability of the victory of Hillary Clinton were proven wrong. The Cambridge Analytica Scandal in which a small data mining company used Facebook user data to influence voters was an important moment in the history of online privacy and the public's awareness of the possibilities of data mining. The election has been speculated to reveal deep divisions in American society. The purpose of this project was to investigate the results of the election to understand some of the differences in voting patterns between groups of people.

Contents

1 Introduction

The objectives of this project were to discover interesting correlations in the 2016 US presidential election datasets and to determine if we could accurately classify which presidential candidate counties were likely to have the majority vote based on the county's demographics.

To accomplish these objectives, we first completed a data preparation and preprocessing step to aggregate several datasets and ensure they were in useable condition. The primary set of county features of interest consisted of the demographics and economic performance. We were interested in determining if we could prove the stereotypes of the typical Trump or Clinton voters: moreover, we were even more interested in seeing if these features could reveal unexpected correlations that would *disprove* the stereotypes. The second phase of the project aimed to achieve exactly that.

During the second phase, we compared and visualized a number of features pairs and triplets that were anecdotally thought to indicate which direction a subset of the population was likely to vote for. These comparisons included with the assumed outcome based on cultural stereotypes:

- Homeownership. Homeownership is greater in rural regions than it is in urban regions. It is believed that the subset of the population lived in dense urban regions were more likely to vote for Clinton.
- Population density. The land area of a county vs its population is also an indicator of how rural a region is and it is believed that voters in less densely populated rural regions are more likely to vote for Trump.
- Veterans. It is believed that those who have served in the military are more likely to vote for Trump.
- Affluence. It is believed that voters who are closer to the poverty line were more likely to vote for Trump.
- Female voter race. It is believed that Trump was only interested in appeasing caucasian voters and ostracized minority voters: especially female minorities. It was also believed that females, in general, were more likely to vote for Clinton.

The final phase of the project involved determining if the voting behavior of a county could be classified based on its properties: rather than voting predominantly based on regional culture or values. For this phase, we developed and evaluated six classifiers to determine which could best classify counties based on multiple evaluation metrics. Lastly, we treated Vancouver Island as if it were a county and attempted to determine if our own political alignment was congruent with that of other counties in the United States.

2 Data Preparation and Preprocessing

For our project, we used three comma-separated values (CSV) datasets: the 2016 US Presidential election county votes [1], 2014 US census county demographics [2], and 2016 US county unemployment [3]. In their initial state, these datasets were unsuitable for data mining; therefore, we first needed to perform some data preparation on them.

Data preparation is often defined as involving four distinct tasks: data cleaning, data integration, data transformation, and data reduction [4]. This process is so important that, in the *Handbook of Data Mining*, Pyle asserts that: "Data preparation consumes 60 to 90% of the time needed to mine data—and contributes 75 to 90% to the mining project's success" [5]. Due to the importance of data preparation, we expended significant effort in the early stages of our project to ensure our data was in a useful state that would require little to no modification to data mine.

Our data preparation involved first converting the CSV files into database tables. As databases are designed to process large quantities of data, we felt this decision would provide us with the best tools to preprocess our data. Once the data was in a database, we proceeded to perform the four tasks of data preparation. These tasks are described in detail in the following subsections. In order to make the preprocessed data accessible, we developed a basic web server that would query the database on startup and then export and serve a preprocessed CSV file. Our data was now prepared and available to be read in as a remote resource using *Pandas*.

2.1 Data Cleaning

The task of data cleaning involves eliminating inconsistencies or incoherent data. For our data, the data cleaning consisted primarily of making the county naming format consistent. In some datasets, the counties followed a "Name County" convention while in others simply a "Name". The secondary task was to ensure that all datasets contained state abbreviations as well as a Federal Information Processing Standards (FIPS) codes. FIPS codes are used to uniquely identify counties and county equivalents in the United States and were used as a means to combine the data during the data integration task.

2.2 Data Integration

The task of data integration involves integrating databases (or database tables) from various sources into a single dataset. This task was critical for our project as we were using three datasets. Due to the care that was taken in the data cleaning task to ensure that all datasets were at the county granulation and all contained FIPS codes, data integration became little more than a series of joins on the FIPS codes.

2.3 Data Reduction

The task of data reduction involves reducing the number of features in a dataset with the goal of reducing data mining processing time. For our project, joining together all features during data integration results in ~90 features: many of which are duplicates or irrelevant. After performing data reduction, we were left with 59 features.

3 Data Mining

We used *Python* and *scikit-learn* to analyze the data gathered in our database. The primary goals of this stage of the project were to find interesting results and correlations between the data and to create visualizations of these results.

We created Python methods in our Jupyter Notebook so that we could easily analyze individual attributes from the full CSV file. This allowed us to correlate subsets of the attributes to create 2D and 3D visualizations of attributes in the dataset.

One example of an interesting result that we found is demonstrated in the image below.

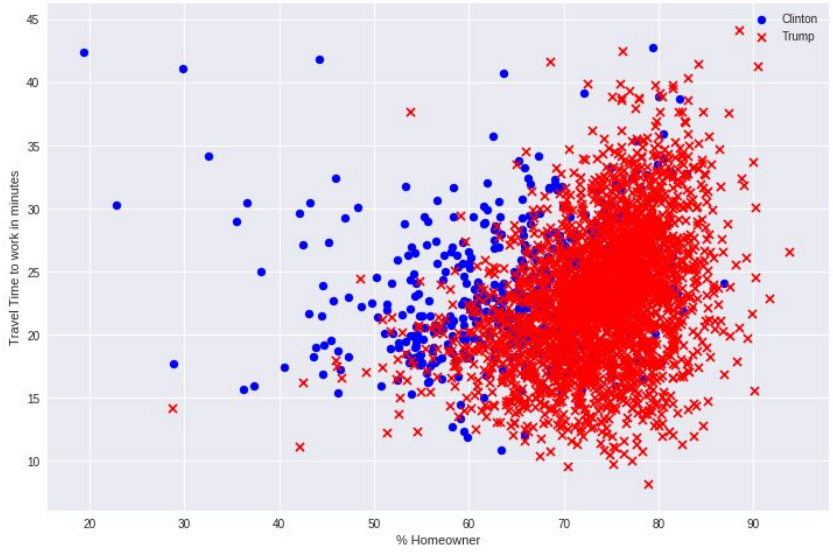


Figure 1: Relationship between Mean Travel Time to Work and Homeownership Rate

The plot indicates that counties which voted for Trump generally have a higher rate of home ownership than those that voted for Clinton. Our hypothesis for this is that voters in urban areas are both more likely to vote for Clinton and more likely to rent. Whereas Trump supporters are more likely to live in rural areas or smaller cities and to own their own homes. The travel time doesn't seem to have much of an effect on how people voted.

The following plot illustrates the relationship between population per square mile and the size of the county.

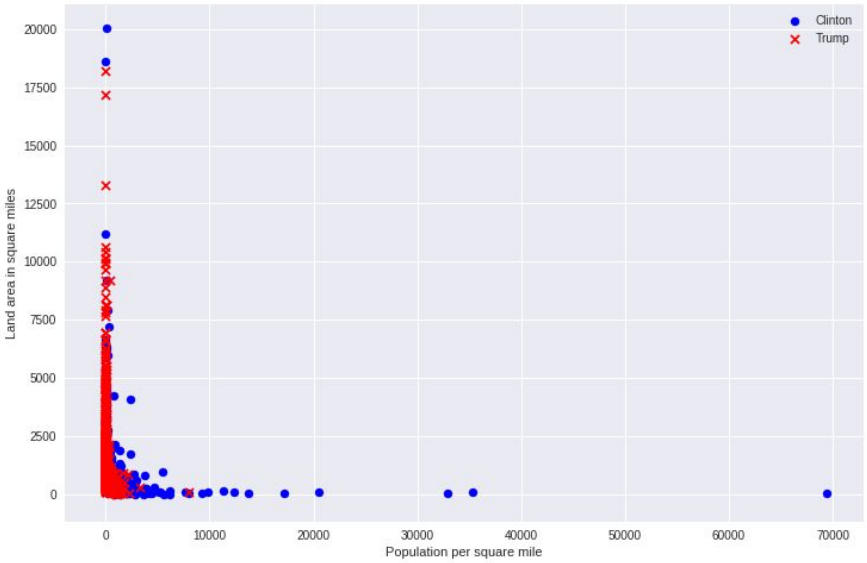


Figure 2: Relationship between Land Area in Square Miles and Population per Square Mile

There is an indication that the most populated counties voted for Clinton over Trump. The size of the county and voting preference is less clear. A large number of counties that voted for both Clinton and Trump are clustered closely together with outliers for both candidates.

The following chart is a three-dimensional plot of the relationship between the number of veterans, the poverty rate and the per capita income at the county level.

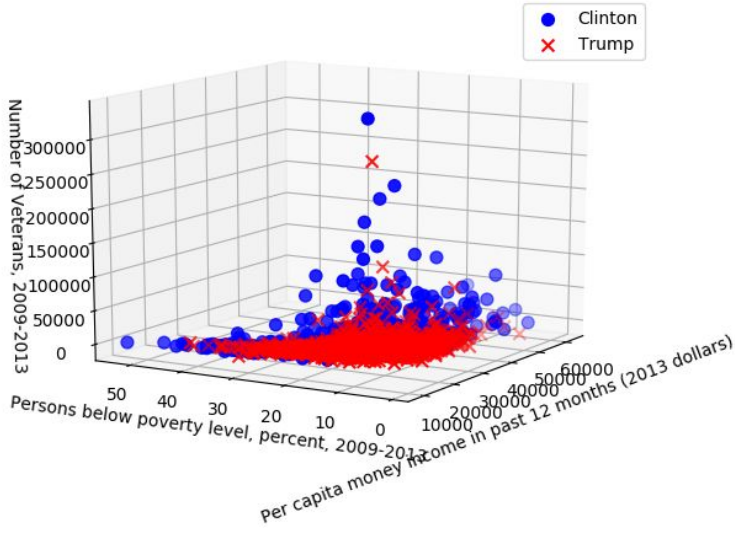


Figure 2: Relationship between Number of Veterans, Percent of People Below the Poverty Line and Per Capita Income For 2013

Figure 2 contains a surprising result for us. The counties with the highest number of veterans were more likely to vote for Clinton over Trump. This was the opposite of what we expected. There does not seem to be an obvious distinction between Trump and Clinton voters in terms of income or poverty level.

However, there are more Clinton voters on the extreme ends of the poverty and income ranges whereas the counties which voted for Trump were more focused in the middle range.

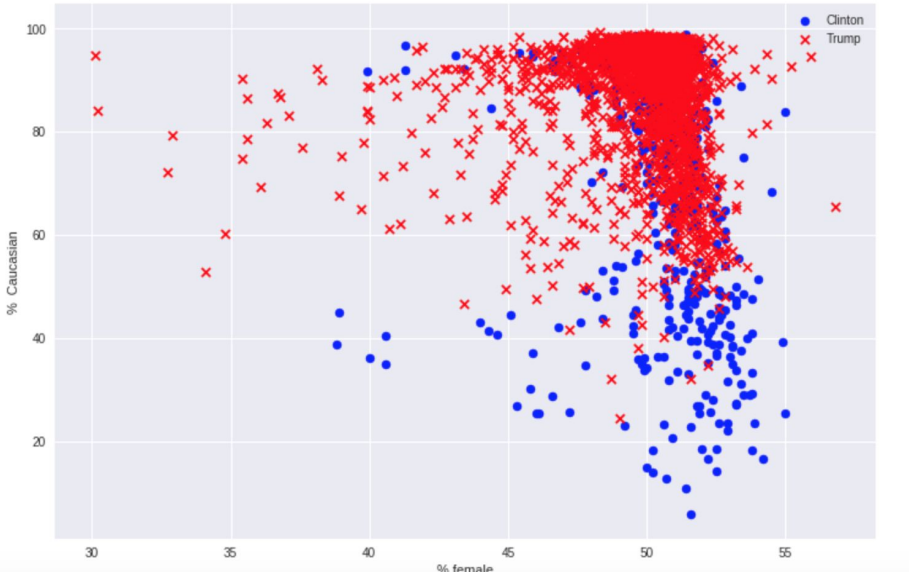


Figure 3: Relationship between Percent Caucasian and Percent Female

Figure 3 demonstrates that the counties with a higher percentage of Caucasian voters were more likely to vote for Donald Trump. Those with a lower percentage were more likely to vote for Hillary Clinton. In addition, those counties with a lower percentage of men were more likely to vote for Trump.

Figure 4: Relationship between Percent African American and Percent Female

Figure 4 shows that the counties with the highest percentage of African-American voters were likely to vote for Clinton. Those with the lowest percentage were more likely to vote for Donald Trump.

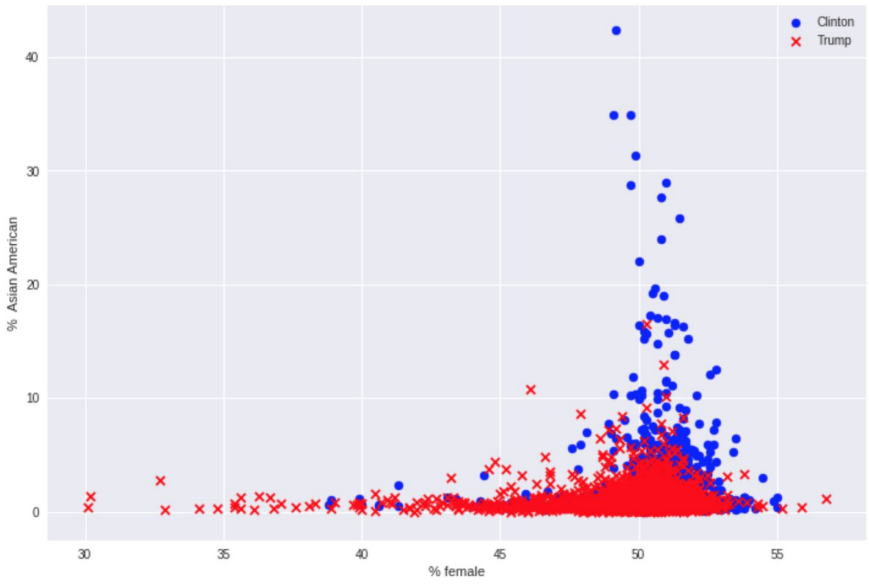


Figure 5: Relationship between Percent Asian American and Percent Female

Figure 5 shows that the counties with the highest number of Asian-Americans were more likely to vote for Clinton. This seems to be less pronounced than for the previous two figures.

4 Classification

The primary goals of this stage of the project were to test different classification methods and evaluate the performance of these classification methods. We developed a set of classifiers using *scikit-learn* and *Python* that are capable of accurately predicting which candidate a county is most likely to vote for. To find the best classifier, we compared six classifiers: Logistic Regression, Gaussian Naive Bayes, Perceptron, Support Vector Machine, K-Nearest Neighbours, and Decision Tree.

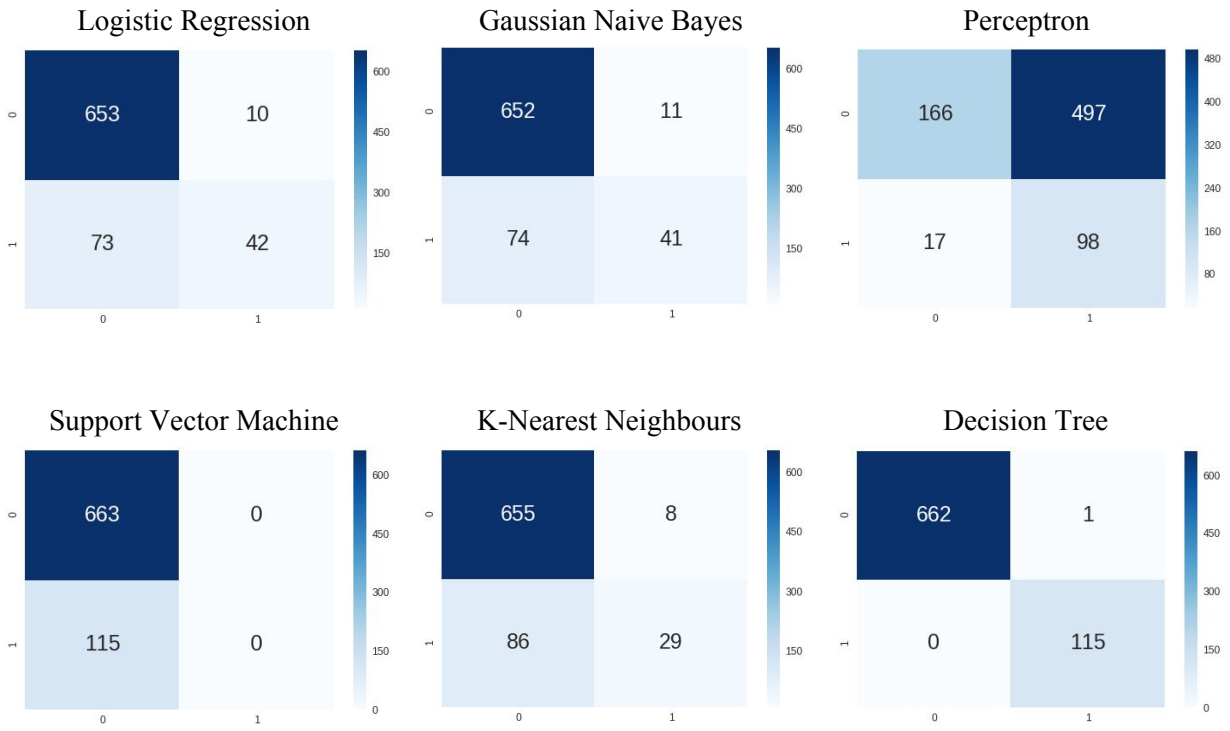
Prior to training the classifiers, an additional data pre-processing step was required to remove any irrelevant features: including the state name, county name, and party votes. Of the 59 features in our original dataset, we retained 55 relevant features to use for training. We used the *scikit-learn train_test_split* method to split the 3111 county data into random train and test subsets and used the default parameters which produced a $75-25\%$ split of the input data. With the data split into subsets, we were prepared to train and evaluate each classifier.

4.1 Classifier Evaluation

To evaluate each classifier performance, we first examined their prediction accuracy and then confirmed that it is a valid measure by checking it the ROC space and F-Measure were congruent. Below in figure 6 is a graph displaying the prediction accuracies of each of the six classifiers:

Figure 6: Classifier prediction accuracies

Accuracy is often considered a poor metric for measuring performance if the data is heavily skewed or if error costs are unequal [7]; however, for our dataset, the data is almost evenly distributed - 46.1% Trump, 48.2% Clinton - and the error cost are exactly equal. To confirm this hypothesis, we then generated the confusion matrices for each classifier.



For a binary classifier in *scikit-learn*, the count of true negatives is $C_{0,0}$, false negatives is $C_{1,0}$, true positives is $C_{1,1}$ and false positives is $C_{0,1}$ [6]. Using a confusion matrix, we can compute many useful evaluation metrics: primarily true-positive rate (TPR), false-positive rate (FPR), and F-Measure which considers both the precision and recall when computing a score. These new metrics can be found below in table 1.

Classifier	TPR	FPR	Precision	Recall	F-Measure
Logistic Regression	0.365	0.015	0.792	0.365	0.268
Gaussian Naive Bayes	0.357	0.017	0.774	0.357	0.259
Perceptron	0.852	0.75	0.164	0.852	0.139
Support Vector Machine	0	0	0	0	0
K-Nearest Neighbours	0.252	0.012	0.763	0.252	0.191
Decision Tree	1	0.002	0.983	1	0.659

Table 1: Classifier evaluation metrics

From the F-Measure score, we can see that accuracy continues to be a valid measure of the classifier performance. An interesting result is that the Support Vector Machine appears to follow the strategy of never predicting a positive classification; meaning that it never produces a false positive but also produces no true positives.

As our classification problem is a discrete classifier is one that outputs only a class label, for each classifier we produce a false-positive rate (FPR) and true-positive rate (TPR) pair which corresponds to a single point in the receiver operating characteristics (ROC) space. The ROC space is useful for organizing classifiers and visualizing their performance [8]. An ROC space containing plotting all six classifiers can be seen below in figure 7.

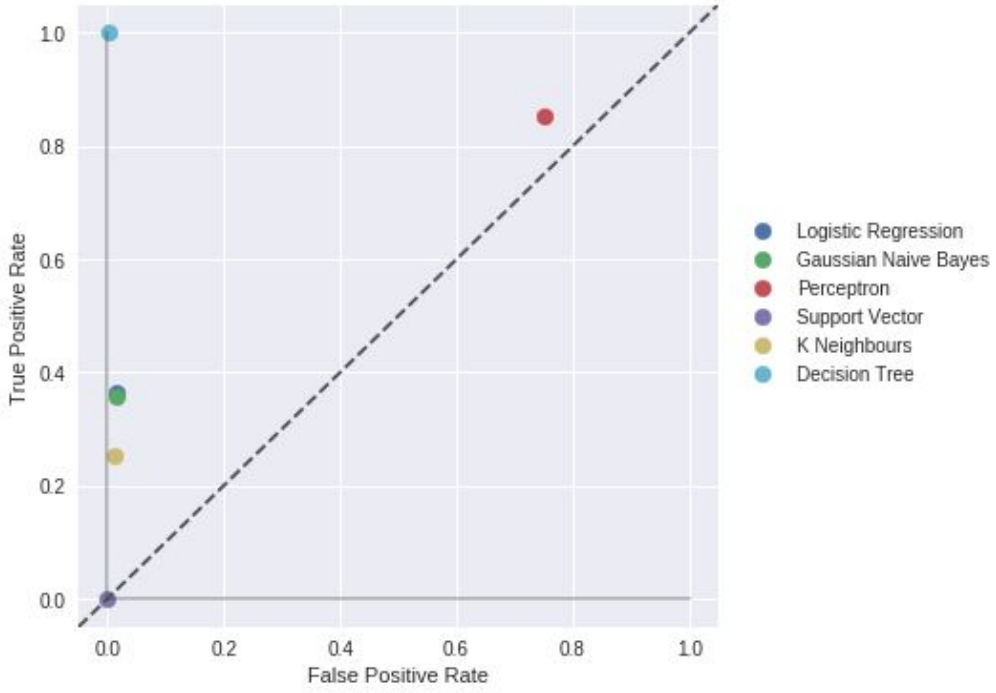


Figure 7: Classifier comparison in ROC space

As a classifier in ROC space is considered better than another if it is closer to the northwest, we once again find the Decision Tree classifier to perform best followed by Logistic Regression and Gaussian Naive Bayes. We, therefore, find that the prediction accuracy, F-measure, and ROC space are congruent and that the Decision Tree classifier performs the best for our dataset.

4.2 Vancouver Island

Here on Vancouver Island, at least anecdotally, there is a strong anti-trump opinion. To determine if this is due to our demographics/population/economy or if it is due to our beliefs/culture/values, we gathered all the relevant features for Vancouver Island to create an instance that we could use our classifiers to predict.

Nearly all the required information was found using the 2016 Canadian Census information provided by Stats Canada [6]; however, some information was unavailable. The list of unavailable features can be found in Appendix A; furthermore, any financial features were converted into US dollars at an exchange rate of $1\ CAD = 0.753882\ USD$.

When the new "Vancouver Island County" is predicted using the six classifiers we evaluated in the subsequent section, we get an even split of Trump vs Clinton. We find that *Logistic Regression, Gaussian Naive Bayes*, and *Perceptron* vote for Clinton while *Support Vector Machine, K-Nearest Neighbours*, and *Decision Tree* vote for Trump; although, it is worth noting that *Support Vector Machine* always classifies

instances as Trump.

As the Decision Tree classifier outperformed all other classifiers by all metrics, we could conclude that if Vancouver Island was a county in the United States, that we likely would have voted for Trump; however, this result was relatively inconclusive as both the second and third best classifiers determined we would vote for Clinton.

5 Conclusion

The objectives of this project were to discover interesting correlations in the 2016 US presidential election datasets and to determine if we could accurately classify which presidential candidate counties were likely to have the majority vote based on the county's demographics. In doing so, we challenged a number of stereotypes and found that most were confirmed to be true.

These stereotypes included that Caucasian females were more likely to vote for Trump while all other races were more likely to vote for Clinton. We also confirmed that voters from rural regions - where homeownership is greater and population density is lower - were more likely to vote for Trump while those from urban or densely populated regions were likely to vote for Clinton. Perhaps the most interesting result that from rejecting the stereotype that veterans were likely to vote for Trump; instead, we found that there was a strong indication that veterans would vote for Clinton. Another surprising result was that voters from most extremes of the income spectrum were more likely to vote for Clinton while those in the center would vote for Trump: the opposite of what was expected. Ultimately, it appears that many of the stereotypes do exist for a reason but that they should not be trusted as in some cases the truth is actually the inverse.

In order to determine if cultural beliefs and values were homogeneously distributed throughout all counties, we developed six classifiers. The rationale was that if we would accurately classify counties, then it is likely that a county's demographics and economic performance were sufficient indicators of political alignment. We discovered that a *Decision Tree* classifier performed best by all evaluation measures and could predict the likely candidate with 99% accuracy; furthermore, *Logistic Regression* and *Gaussian Naive Bayes* were also capable of predicting with an 89% accuracy. This level of performance is a powerful confirmation that demographics and economic performance were sufficient to predict the 2016 election results. Lastly, we examined how these indicators would classify Vancouver Island and if it was consistent with the anecdotal belief of our political alignment. Unfortunately, this result was relatively inconclusive with half of the classifiers - including the *Decision Tree* - indicating we would vote for Trump while the other half - including *Logistic Regression* and *Gaussian Naive Bayes* - indicating we would vote for Clinton.

From the results we discovered throughout this project, we have determined that national stereotypes and county demographics and economic performance can very often be used as strong indicators of political alignment; however, they should be examined carefully as they can occasionally be misleading or outright false.

References

[1] Hamner, Ben. 2016 US Election Version 8. Available: https://www.kaggle.com/benhamner/2016-us-election. [Accessed: 2-Dec-2018].

[2] MuonNeutrino. US Census Demographic Data, Version 1. Available: https://www.kaggle.com/muonneutrino/us-census-demographic-data. [Accessed: 2-Dec-2018].

[3] Jay Ravaliya. US Unemployment Rate by Country, 1990-2006, Version 2. Available: https://www.kaggle.com/jayrav13/unemployment-by-county-us. [Accessed: 2-Dec-2018].

[4] Kochanski, Andrzej., 2010, Data preparation. COMPUTER METHODS IN MATERIALS SCIENCE. Vol. 10. 25-29. Available: https://www.researchgate.net/publication/299350639_Data_preparation

[5] Pyle, D., 2003, Data Collection, Preparation, Quality, and Visualization, The Handbook of Data Mining ed. Nong Ye, Lawrence Erlbaum Associates, New Jersey. Available: http://read.pudn.com/downloads159/ebook/710349/5GreatMatlabBooks/HandbookOfDataMining.pdf

[6] "sklearn.metrics.confusion_matrix," 1.4. Support Vector Machines - scikit-learn 0.19.2 documentation. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html. [Accessed: 29-Nov-2018].

[7] F. Provost, T.Fawcett, "Analysis and Visualization of Classier Performance Comparison under Imprecise Class and Cost Distributions". Available: https://www.nssl.noaa.gov/users/brooks/public_html/feda/papers/ProvostandFawcettKDD-97.pdf

[8] T. Fawcett, "An introduction to ROC analysis". Available: https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X

[9] "Census Profile, 2016 Census Vancouver Island and Coast [Economic region], British Columbia and British Columbia [Province]," Census Profile, 2016 Census, 24-Apr-2018. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=ER&Code1=5910&Geo2=PR&Code2=59&Data=Count&SearchText=Vancouver&SearchType=Begins&SearchPR=01&B1=All&TABID=1. [Accessed: 30-Nov-2018].

Appendix A - Missing Features

The information for all features below was unavailable from Stats Canada. For all missing features, we used the mean average across all counties as the value for Vancouver Island with the exception of "VET605213" which was approximated using BC veteran information retrieved from Veterans Affairs Canada.

POP715213 - Living in same house 1 year & over, percent, 2009-2013
POP645213 - Foreign born persons, percent, 2009-2013
VET605213 - Veterans, 2009-2013
BZA010213 - Private nonfarm establishments, 2013
BZA110213 - Private nonfarm employment, 2013
BZA115213 - Private nonfarm employment, percent change, 2012-2013
NES010213 - Nonemployer establishments, 2013
SBO001207 - Total number of firms, 2007
SBO315207 - Black-owned firms, percent, 2007
SBO115207 - American Indian- and Alaska Native-owned firms, percent, 2007
SBO215207 - Asian-owned firms, percent, 2007
SBO515207 - Native Hawaiian- and Other Pacific Islander-owned firms, percent, 2007
SBO415207 - Hispanic-owned firms, percent, 2007
SBO015207 - Women-owned firms, percent, 2007
MAN450207 - Manufacturers shipments, 2007 ($1,000)
WTN220207 - Merchant wholesaler sales, 2007 ($1,000)
RTN130207 - Retail sales, 2007 ($1,000)
RTN131207 - Retail sales per capita, 2007
AFN120207 - Accommodation and food services sales, 2007 ($1,000)
BPS030214 - Building permits, 2014