**Improving Voter Turnout: A Data Science Approach to Examining Personal and Environmental Factors and their Relationship to Voter Participation**

Team Number (or name): Charlotte1

Team Members: Sthanu Ramakrishnan, Jared Leach, Divya Garg, Marc Rizzo, Paul Blase

Course: DSBA/ MBAD 6211- Advanced Business Analytics

Instructor: Dr. Mousavi

Date Submitted: 04/30/2019

# Executive Summary

## **Business Problem Statement**

Voter participation is a key aspect of any nation’s democracy to function properly. Lower voter turnout is a significant problem that affects states/regions/countries around the world, which is undesirable. Individual states/regions/counties must have some tools available to increase voter participation and expand opportunities of eligible voters. With the techniques applied in data science and analytics, we might identify and predict the personal and environmental factors that aid in predicting voter participation in Mecklenburg County. Better predictive models for voter turnout will allow the community to better assist those votes that become unheard due to individual or multiple factors.

## **Business Goal**

The main goal of our research was to create a classifier to predict the likelihood of voter participation in upcoming elections better. By creating a classifier that accurately captures voter engagement, we hope to enable Mecklenburg county to focus more directly on areas where election participation is low. This classifier will, in turn, help these elections to depict the true sentiments of the county better.

## **Data Profile**

We obtained the dataset provided by Quality of Life, a website providing insight into different neighborhood profiles within the county. In this data, Mecklenburg County is broken down into 474 regions. After significant data manipulation and imputation, the cleaned data set resulted in 406 observations with 13 variables examined. The target variable for this study was voter participation in 2016 which examines the percentage of voter participation in each neighborhood profile. Out of 498 features in the dataset, we focused on the following attributes:

* Voter participation in previous years: percentage of voter participation in neighborhood profile
* White\_Population: Provides the percentage of area’s Caucasian population
* Age\_of\_Residents: Median age of residents in the area
* Household\_Income: Median household income
* Employment\_Rate: Percentage of adults employed in a given area
* Proficiency\_Elementary\_Schools: Percentage of students grades 3-5 passing end of year testing
* Bachelors\_Degree/High\_School\_Diploma: Percentage of population with given degrees
* Pubilc\_Nutrition\_Assistance: Percentage of residents on aide such as food stamps
* Prenatal\_Care: Percentage of births where the mother received adequate prenatal care
* Home\_Ownership: Percentage of homes that are occupied by the owner

## **Results**

Through the use of various methods such as clustering, random forest, and ordinal logistic regression, we were able to classify the dependent variable into four main groups of estimated voter participation: Low, LowMid, HighMid, and High. Our final models were able to accurately classify around 75% of the time which is a decent performance. For the Ordinal Logistic Regression, the High class was more often correctly predicted than the others, with a precision of 87.5%; Low, LowMid, and HighMid only had precisions of 78.6%, 71.4%, and 73.5% respectively. The recall for the classes also showed decent variation, with Low having 88%, LowMid having 73.5%, HighMid having 69.4%, and High having 80.8%.

# Project Report

## **Introduction**

According to Wikipedia**,** voter turnout is the percentage of eligible voters who cast a ballot in an election. All citizens of a country must be encouraged to participate in the political process for a political candidate to be fully supported (Unknown Wikipedia Users 2019). As cited in Americanprogess.org, almost 92 million eligible Americans did not vote in the 2016 presidential elections. In the 2014 midterm elections, an estimated 143 million eligible Americans failed to vote, marking the lowest voter participation in 72 years. While the United States of America has a lower level of voter turnout compared to most other developed nations, some of these top other nations still have voter turnout of less than 80% (DeSilver 2018). While the focus of this analysis is on a local district American election turnout. Voter participation is the key aspect of any nation’s democracy to function properly, and all the eligible citizens must have the opportunity to vote (Root and Kennedy 2018).

Eligible voters are clear about the barriers that keep them from voting. For example, a 2017 Pew study examined why registered voters refrained from voting in the 2016 elections and found the most common reason among respondents—25 percent—to be that they **“**Didn’t like candidates or campaign issues,” followed by “Not interested, felt vote wouldn’t make a difference,” at 15 percent; “Too busy or conflicting schedule,” 14 percent; “Illness or disability,” 12 percent; “Out of town or away from home,” 8 percent; “Registration problems,” 4 percent; “Forgot to vote,” 3 percent; “Transportation problems,” 3 percent; and “inconvenient hours or polling place,” 2 percent.” (Lopez and Flores 2017). This example illustrates that people have a diversity of reasons for abstaining from an election. Also, these responses show the importance of identifying a model that best captures the ability to predict voter participation and thus allow the community to assist those at risk of not voting.

Individual states/regions/counties must have some tools available to increase voter participation and expand voting opportunities for eligible voters. These tools may include resources to directly help the citizen vote such as streamlining voter registration with automatic voter registration, same-day voter registration (SDR), preregistration of 16- and 17-year-olds, and online voter registration. These tools may also include data science and analytics to predict individuals at-risk not to vote ahead of time. Both kinds of implements can help ensure a result more representative of the population.

Elected bodies of the government are more representative, and the laws are fairer when all eligible Americans can have their voices heard and to participate in elections. For voters who are disengaged and disenchanted with the political process, robust civics education programs and integrated voter engagement initiatives can drive participation by re-energizing voters and providing them with reasons and opportunities to cast ballots on the issues that matter most to them and their communities. Furthermore, states must have in place affirmative voter registration and voting policies to ensure that eligible voters who want to vote can and not become blocked by unnecessary and overly burdensome obstacles. By adopting these policies, America can find its 92 million missing voters and improve the voting experience for all eligible voters (Root and Kennedy 2018).This work involves identifying and predicting the personal and environmental factors that aid in predicting voter participation in Mecklenburg County. Better predictive models for voter turnout will allow the community to assist better those votes that become unheard due to individual or multiple factors.

**Background**

There has been ample research conducted in previous studies involving voter participation and voter turnout in the election process. One area of research and analysis is the role that voter turnout and participation has with demographic characteristics of voters. Age is a variable that is often examined or collected in research samples involving people. As cited in Costa, Schaffner, and Prevost, 2018, current population surveys indicate that in both past elections and the public’s most recent elections we see at least 15% differences between 18 to 25-year-olds and 25-years and older when it comes to voter participation (Costa et al. 2018). This observation would be a substantial difference if given a large sample size such as the country and Mecklenburg County demographics. Also, other research has concluded that there a statistical relationship between age and voting participation (Stockemer and Rocher 2017). Another variable of interest when it comes to voter demographics is the race of the potential voter. Recent research, while examining a variety of other factors like neighborhood/population demographics of voters and candidate race, confirmed that minority voting is mostly lower in number (Fraga 2016). Finally, evidence has shown that there indeed an education effect on U.S. voter participation where populations with lower education are less likely to vote (Gallego 2010). Some studies have even begun to try to examine and combine multiple demographic variables and found that there are some interactive effects of education, race, and age (Xu 2005). These collective findings suggest that when attempting to identify variables that play a role in determining whether someone will cast a vote or not, we should include some form of consideration for individual demographics such as age, race, and education.

In addition to intrapersonal factors that may have a relationship with voting participation, there is also room to consider the role of contextual factors for individuals and their families. First, reports describe that there is a trend between voter turnout and income of voter where lower-income individuals are less likely to vote and is one of the stronger independent factors on voting (Parlapiano and Pearce 2016). However, recent studies have found mixed results. One article found no statistically significant independent relationship between household income and voting participation (Cebula 2017). Other studies have stated that lower-income voters do not participate in the political process as much, but that this may be part of a larger mechanism affected by societal income inequality (Fenzl 2018). Like income, studies have also shown that the unemployment rate has a positive relationship with voter participation where higher unemployment rate meant higher voter participation (Cebula 2017). Furthermore, recent research shows that in U.S. local elections home ownership is related to higher voter turnout compared to those renting (Jiang 2018). These studies show the importance of considering variables that relate to the voter’s immediate environment like income, employment, and access to adequate housing when it comes to determining whether someone votes in an election or not.

Research shows that those with chronic illnesses, mental health concerns, disabilities, the flu, or some other impairment are typically less likely to vote than those who are in good health. This outcome suggests a link between population health and politics. Healthcare, and access to it, then becomes an important factor in determining voter turnout in a given area (Urbatsch 2017). Research has also found that people in lower-income areas, as well as those with no insurance, disabilities, and poor health, were all significantly less likely to vote than those in the general population (Gollust and Rahn 2015). Access to healthcare also affects those who are within walking distance of a precinct or who recently gave birth and their likelihood to vote (Bhatti et al. 2019; Burden et al. 2017).

There also are several community factors that appear to play a role when it comes to if someone will vote in an election or not. Regarding crime rates, research has shown that higher levels of crime have a negative relationship with voter turnout and participation (Akhmetkarimov 2008). This finding means that the higher the crime rate in that area the lower the voter turnout in that same area. Another factor in determining the likelihood of voter participation is the population density of a given area. Many people in the US view cities and other developed urban areas as storehouses of education, money, and culture, assume that cities would be more knowledgeable in politics, and therefore more likely to vote in political elections (Preuss 1981). This view is flawed, as some figures show counties with low population density have higher rates of voter participation than urban counties do. For example, in its last census, Mecklenburg County ranked 22nd out of the top 50 metro areas in total population but was not even in the top 50 in population-weighted density only averaging 785 people per square mile (Mcmap 2010). This decreased density could lead to lowered voter turnout across Mecklenburg County.

Other immediate, and more direct, environmental factors may be relevant in predicting voter turnout. As cited in an article by Budds in 2018, "A study in Baltimore found that neighborhoods with a higher density of tree canopy also have higher levels of social capital—meaning neighbors are more close-knit and more likely to trust each other" (Budds 2018). Additionally, a sense of community and interactiveness often leads to increases in voter turnout within urban areas (Vemuri et al. 2011). People that may not have necessarily planned on voting may change their mind due to interactions and relationships with people in their local community with those who are planning to vote or who have already voted in the past. Therefore, within Mecklenburg County (primarily urban county), voter turnout could be expected to increase with environmental resources such as more available greenspace and environmental resources that promote social capital such as access to public assistance programs and early child care centers.

Distance to a precinct, or the easy access to transportation to reach a precinct, is another factor that can affect voter turnout. People may look at the cost of being able to go vote or the time wasted voting. Within suburban areas, research has found that distance has its largest impact on turnout (for distances as small as 2-5 miles) (Gimpel and Schuknecht 2003). In urban areas, there are many obstacles, such as lights and frequent construction. These various obstacles can make voting less appealing to these possible voters. This finding is not necessarily true of more rural areas. In these rural precincts, turnout rates are higher, even though the distance traveled to get there is typically 6 to 10 miles which may explain why travel routes in rural areas are typically more direct routes or at the very least unimpeded (Gimple and Schuknecht 2003). Given these findings, we see that access to efficient transportation is important to consider when attempting to predict whether someone may vote or not.

There are a variety of internal and external factors that play a role in voter participation. Looking at these variables in isolation may not be that helpful overall. However, examining them together may require more robust analyses. Data science has tried to answer these questions regarding voter participation and what personal and ecological factors impact voter turnout, but the level of that research is small and has significant room for development. Some studies use more common and basic statistical analysis to explore factors related to voter participation such as regression analysis (e.g., Akhmetkarimov 2008; Cebula 2017). Other studies have tried to use more rigorous analyses such as PROBIT, logistic regression, and even artificial neural networks (e.g., Gallego 2010; Weber et al. 2018; Xu 2005). These models explore only a few variables rather than a more comprehensive or complete list of both personal and environmental factors. This analysis will attempt to not only replicate findings or add new information to the academic conversation, but also create a more robust and realistic model to predict voter outcome with more advanced machine learning and data science.

## **Data**

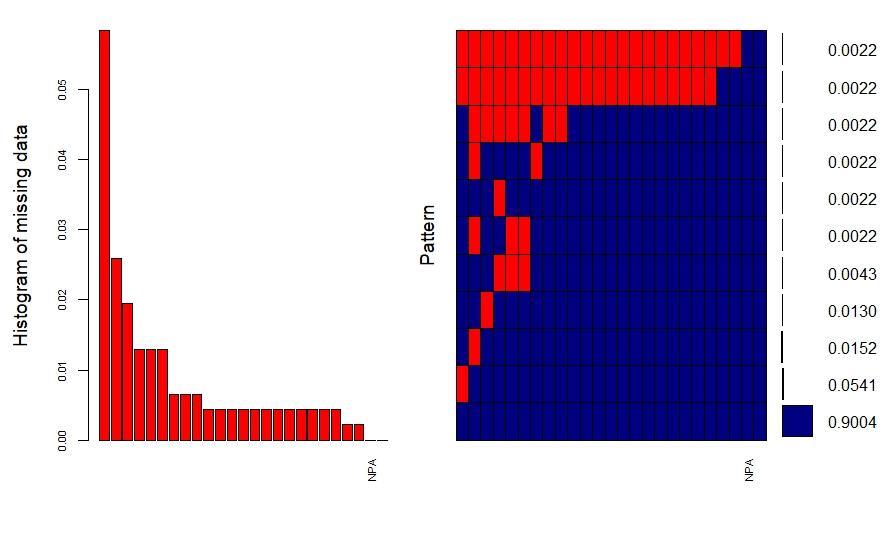
In evaluating voter activity in Mecklenburg County, we have used a regional dataset documenting information in the county. Provided by Quality of Life, a website providing insight on different areas within the county, Mecklenburg is broken down into 474 regions. Within these regions, a wide array of subjects is provided, from health care to economic standing, about the residents and local facilities.

Focusing on all the provided variables will cause issues with degrees of freedom and appropriately comparing observations. Based on the research conducted by the team, the focus shifted to the following variables:

* Population\_Density: Number of people per acre in a community area
* White\_Population; Black\_Population: Provides the percentage of area’s population from a specified race
* Age\_of\_Residents: Median age of residents in the area
* Older\_Adult\_Population\_2016: Percentage of adults over the age of 65
* Household\_Income: Median household income
* Employment\_Rate: Percentage of adults employed in a given area
* Proficiency\_Elementary\_Schools: Percentage of students grades 3-5 passing end of year testing
* Bachelors\_Degree/High\_School\_Diploma: Percentage of population with given degrees
* Public\_Nutrition\_Assistance: Percentage of residents on aide such as food stamps
* Public\_Health\_Insurance: Percentage of residents with public health insurance
* Early\_Care\_Proximity: Percent of housing units with 1/2 mile of a licensed early care and education program
* Prenatal\_Care: Percentage of births where the mother received adequate prenatal care
* Births\_to\_Adolescents: Percentage of births to females under 19
* Violent\_Crime\_Rate: Violent offenses per 1000 residents
* Transit\_proximity: Percentage of housing units within a half mile of a train stop
* Commuters\_Driving\_Alone: Percentage of workers that drive to work alone
* Home\_Ownership: Percentage of homes that are occupied by the owner

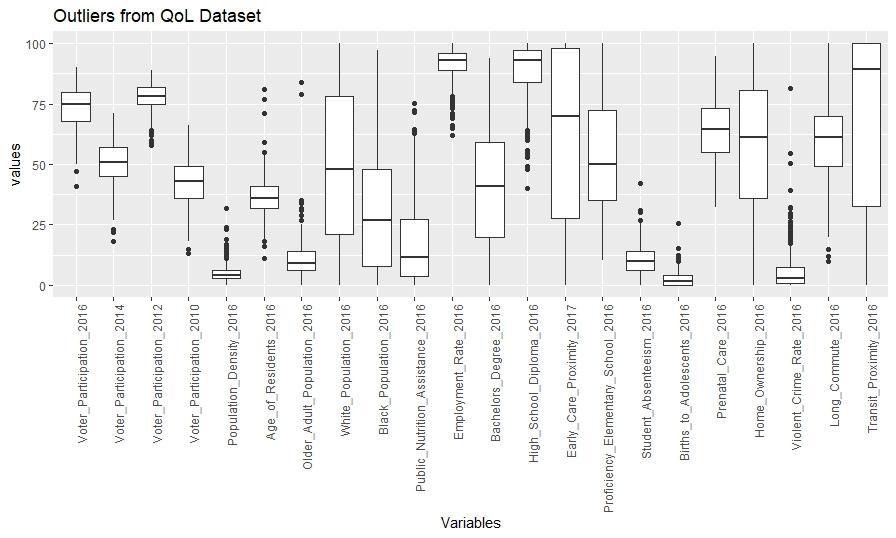
The target variable for this study is a voter participation field, provided by the QoL dataset. This percentage reflects voting activity in the past two presidential elections, as well as the congressional elections in 2010 and 2014. Evaluating the remaining 498 features from the dataset, along with national polling data, we hope to capture voter engagement and how voters participate dependent on the ballot, external sentiment to elections, and circumstances.

From this data, there are a few missing values, specifically within the violent crime rate. In order to ensure the best estimation of all values in the data, the team utilized data imputation. Here is a visual describing the missing values:



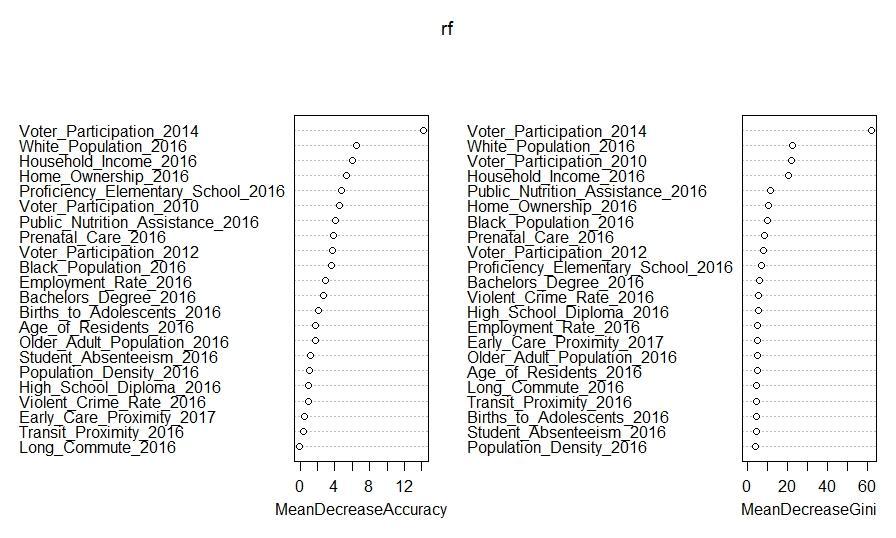
We see a few observations (from areas 122 & 285), are missing all values but ID and population density. Additionally, area 62 had seven missing values, so we move to remove observations with more than five missing values. This action removes the three observations, leaving 459 observations to assess. With this, our highest values of missing values are violent crime rate (25), public health insurance (9), and household income (6). While comparing variable correlation, we find that public health insurance and public nutrition assistance are highly correlated, with an R^2 = .97. With this, we can utilize public nutrition assistance instead, with only three missing values, to represent the use of public aid when modeling. The remaining values were imputed using the mice package, predicting the value based on the remaining features in our dataset. These 41 imputed values lied between seven separate features of the data, which the team documented for later assessment.

Following imputation of values, we moved on to evaluate the distribution of these values to determine if there were outliers or effects on that could affect the modeling process.



While several variables contain outliers, we have some multiplicity in the variables we have kept to this point. For example, retaining median age and an indicator variable for old age in the same modeling set would introduce multicollinearity. In running a simple logistic regression model of all retained variables, we calculated VIF values. In calculating the square root of these values, we see all but one has a value greater than two. This finding indicates strong collinearity of basically all variables at that point.

The final step of data manipulation required assessing the variables tested in a model format including all variables listed above to reaffirm the team’s logic in removing features. Using random forest, we looked for insights into whether or not variables were key in indicating a classification of voter participation(Low participation = <68%, Low-Mid = 68-75, High-Mid = 75-80, High = >80). Similarly, clustering was initialized on the set and identified three groupings, resembling a high, medium, and low participation. Based on variable distinction between groups from clustering, as well as variable importance testing in a random forest model, the team identified specific features as more valuable to predicting voter participation than others.



Feature groups initially deemed valuable, such as population density and access to transportation, were found to lack the value to modeling the data. Keeping in mind collinearity, but trying to balance the importance provided by the test, final decisions were made to remove 14 of the 26 variables initially selected.

With the completion of data imputation and manipulation, this is the information on the 406 remaining observations:



With this data, we evaluate the likelihood of voter turnout and contributing factors using several modeling techniques. First, we intend to use clustering to evaluate similar regions within Mecklenburg County. Assuming demographics and environment is an indicator of the voting interest of citizens, classifying areas within the county together could provide a valuable outlook on how the region is broken down by the factors provided to us. Also, we hope to effectively evaluate the probability of turnout by utilizing more sophisticated logistic regression along with random forest, looking at the values provided by the database to assess the outcome likelihood of voting within the different sections we are testing.

## **Method**

**Cluster Analyses**

For our analysis of voter prediction models, the researchers will be testing three different machine learning/analytical models. To start, we will be performing a preliminary analysis of the final QOL data set by utilizing cluster analysis. This type of analysis will yield an initial look into the grouping of the data and make descriptive statements about differences, or lack thereof, between distinct group within our data set. After testing for clusters within the data, this paper will test two forms of classification systems: Random Forest and Ordinal Logistic Regression as more robust models that can better assess an individual’s voter participation in an election.

To start, we will first be performing a hierarchical cluster analysis using Ward’s method. Because the team cleaned the data set due to outliers and variables that were highly correlated, we are starting with Ward’s method hierarchical clustering. Its formula seeks to optimize R² by minimizing the error sum of squares as one merges each cluster through the agglomerative process. This sequence makes the results less susceptible to noise and outliers. Also, Ward’s hierarchical clustering serves as a great way to initialize K-means clustering which is another clustering method that we will apply to the data.

After initializing our cluster analysis with Ward’s method hierarchical clustering, we will be analyzing the data with K-means clustering analysis to verify previous results with a different clustering method. K-means clustering is useful because it is a basic algorithm that can be applied very easily. The process of K-means clustering involves grouping data based on the distance of data points from each centroid of a cluster. A weakness of K-means clustering is its susceptibility to noise and outliers. However, in this data set outliers were controlled for/eliminated before clustering. In K-means clustering, the researcher determines the number of clusters to use when grouping said data. To validate our selected k-value (i.e., number of clusters) the team will be using both internal and external measures to enhance result validity. Internal measures here will include generating within-group error sum of square plots which will quantify distances between data points in the same cluster. This analysis will also utilize the Silhouette Coefficient. Put another way; this coefficient is a mathematical function that calculates on a -1 to 1 scale the cohesion within each cluster and separation across clusters simultaneously. Taking the average of these coefficients, analysts desire to be as close to a value of 1 as possible. For an external measure of validity, we will be using the Hopkins Measure to assess the clustering tendency. The Hopkins Measure uses random and actual data and measures their respective distances to the nearest neighbor in the original data. This measure generates a value between 0 and 1 where values close to 0 or 1 mean that there are valid clusters to be found in this data. A value close to .5 indicates these data points are not different from random points and that no valid clusters exist. By using multiple clustering methods as well as internal and external measurements of validity, cluster analysis is an excellent way to gain preliminary information on groupings. This analysis sets a foundation for more elaborate classification models regarding voter behavior.

**Random Forest**

The second voter classification model that we integrated into the research is the random forest algorithm. The random forest algorithm stems from the decision tree classification method. Decision trees randomly assess observations and split these values using features from the dataset, fitting the tree to the training data in the hopes of creating a tree that can accurately fit the test data. Unfortunately, decision trees are often prone to overfitting, which can occur when a decision tree too accurately predicts the training data and does not assess the set on the data’s characteristics. This overfitting makes it hard to generalize the tree to the test data. While pruning the decision tree deters the model from overfitting, removing too many leaves can create too broad an interpretation of the data.

Random forest algorithms help to resolve this trade-off. The random forest algorithm is an ensemble method that can be used to average multiple decision trees in order to determine the optimal number of nodes and which trees to use for splitting the data. Random forest begins with random record selection. First, multiple trees are trained on 63.2% of the data. The data for each tree is selected randomly, and then each tree is split on a predetermined m or mtry (which is a list of utilized variables). Once split, the model tests each tree on the remaining 36.8% of the data, and the classifier calculates the OOB or out-of-bag classification error rate. The 'votes' of each tree, or that tree's prediction, are counted, and the model chooses the classification that has the most aggregated votes over all the trees.

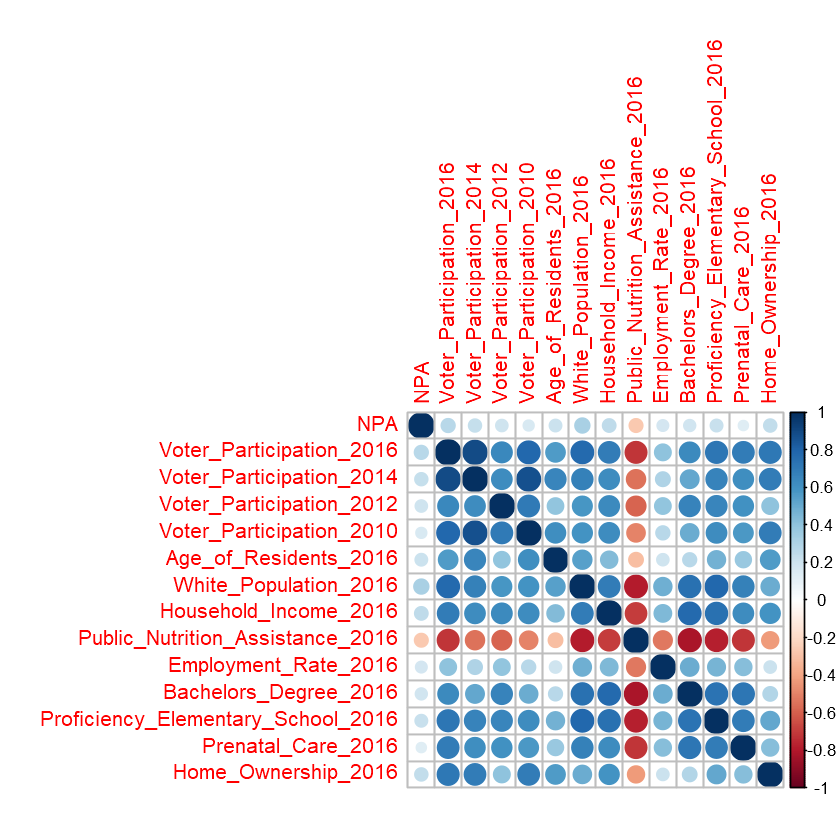
While a good method for classification, random forest is not the best for regression and also struggles with a bias towards multiple-level categorical variables. However, this method can be fine-tuned by finding both the optimal ntree and optimal mtry. Keeping the default mtry (the square root of all predictors) we build multiple trees with different ntree values until we find the ntree value that both stabilizes and minimizes the OOB. The model keeps this ntree value, and we determine the optimal mtry using tunerf function. Once we have both the optimal ntree and mtry, we build the random forest again, using these values to get the optimal forest.

By using the random forest method, rather than just the decision tree method, one can reduce overfitting and eliminate the problem of outliers. This approach, as well as using clustering and more advanced methods such as Ordinal Logistic Regression, helps to identify better classification and more accurately predict voter behavior.

**Ordinal Logistic Regression**

The third model that we built for voter classification is the Ordinal Logistic Regression model. We performed Logistic Regression with the derived variable “Turnout” as the dependent variable. The “Turnout” variable comes from the variable “Voter\_Participation\_2016” variable. The variable “Turnout” is classified into four categories viz. "Low", "LowMid", "HighMid", "High". The team separated these observations into these categories based on the quartile values of “Voter\_Participation\_2016”.

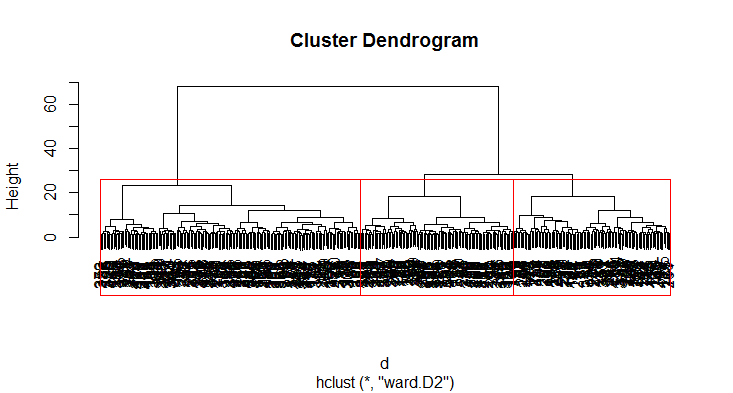
The model was trained using the Cumulative link model (clm) because the clm function takes care of ordinal classification and allows structured thresholds which makes it a powerful model. Also, the regression analysis is performed in depth using the clm function in R. Variable “Turnout” was regressed over 12 independent variables to predict the most likely Voter Turnout category. To check the statistical relationship or dependence among the independent variables, the group created a correlation matrix.

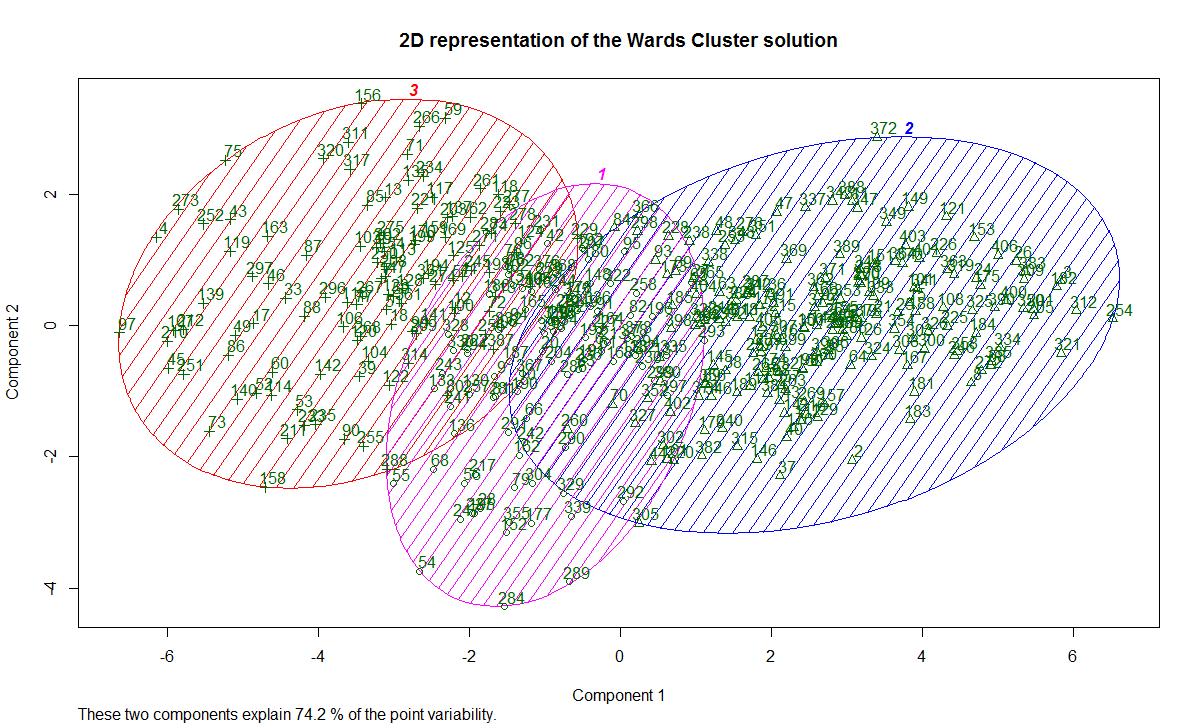


**Results & Discussions**

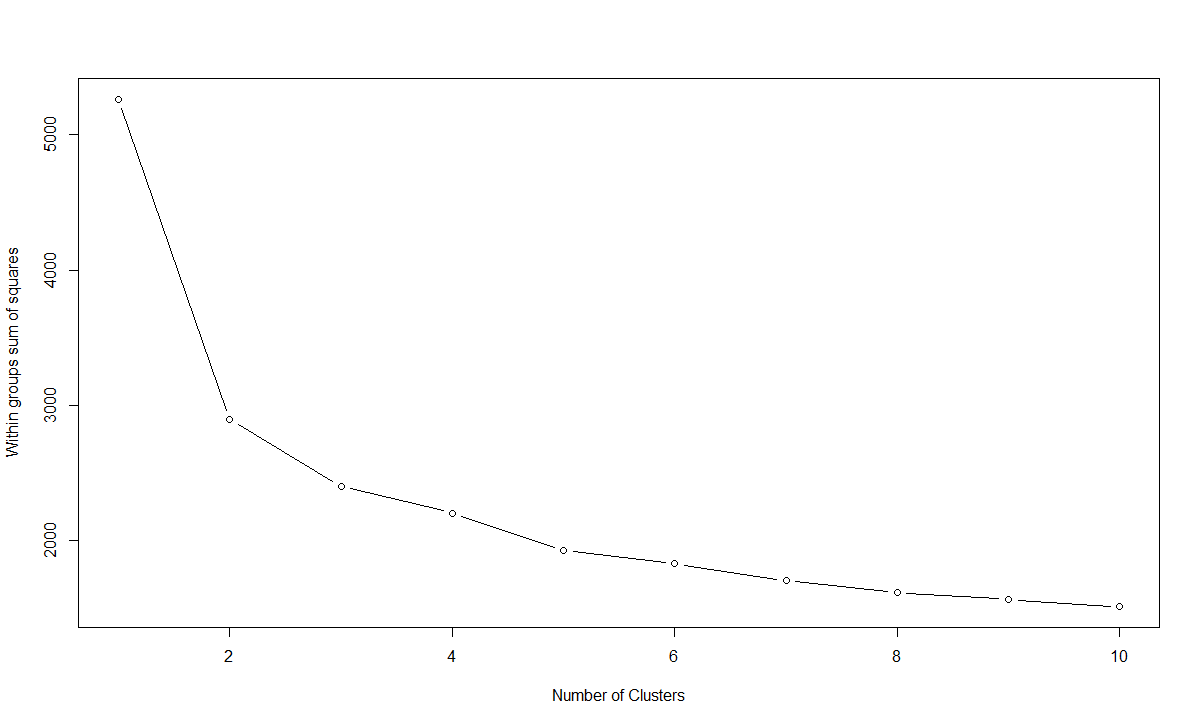
**Cluster Analyses**

After preparing our data to perform cluster analyses by standardizing all the variables to a similar numeric scale, we started by testing whether or not our data shows real/valid clustering by performing the Hopkin’s Measure. Our results showed a Hopkins Value of 0.258. This value indicates that there may be some suggestion of valid clusters of voters within our data set which may warrant further exploration of clustering results. The results, however, cannot clearly say definitively that these clusters are real or no better than random points. Starting with Ward’s method for hierarchical clustering, there is both a dendrogram and a 2D representation of our cluster analysis below to illustrate the results. Upon visual inspection of the dendrogram results, we see there are three cluster groups worth highlighting. These groups/clusters are outlined in red. The 2D representation below shows these data points plotted and organized by cluster. This 2D representation alone explains almost 75% of each data point’s variability.

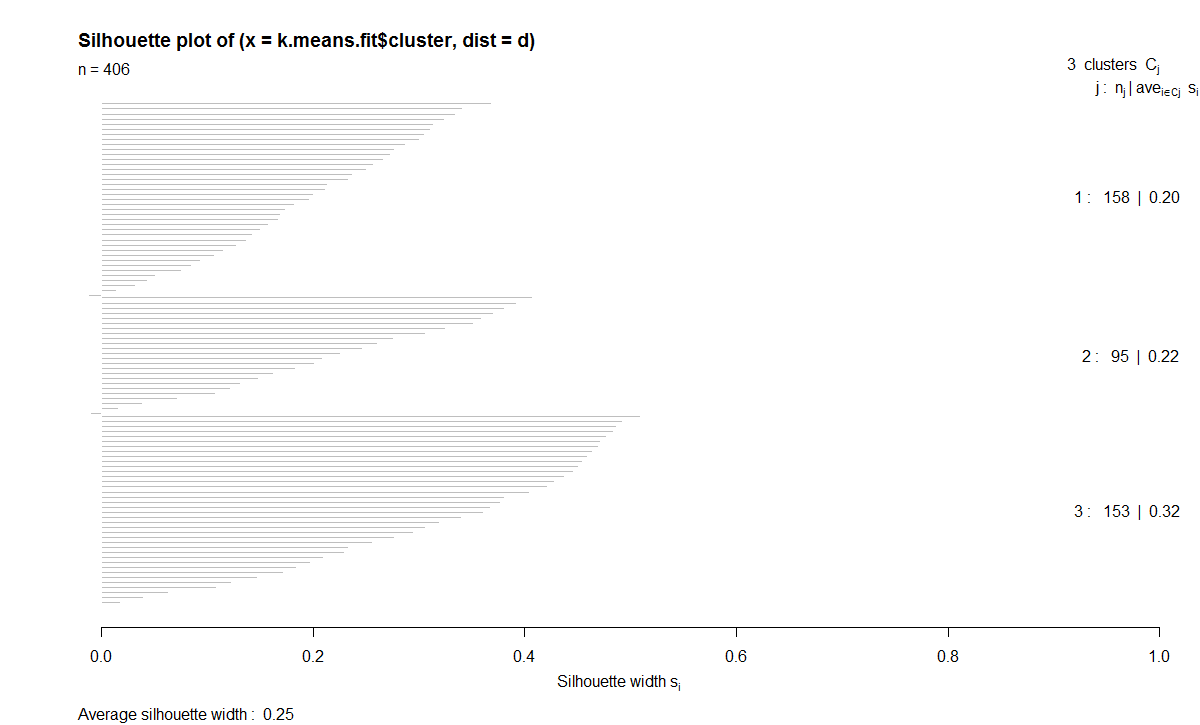


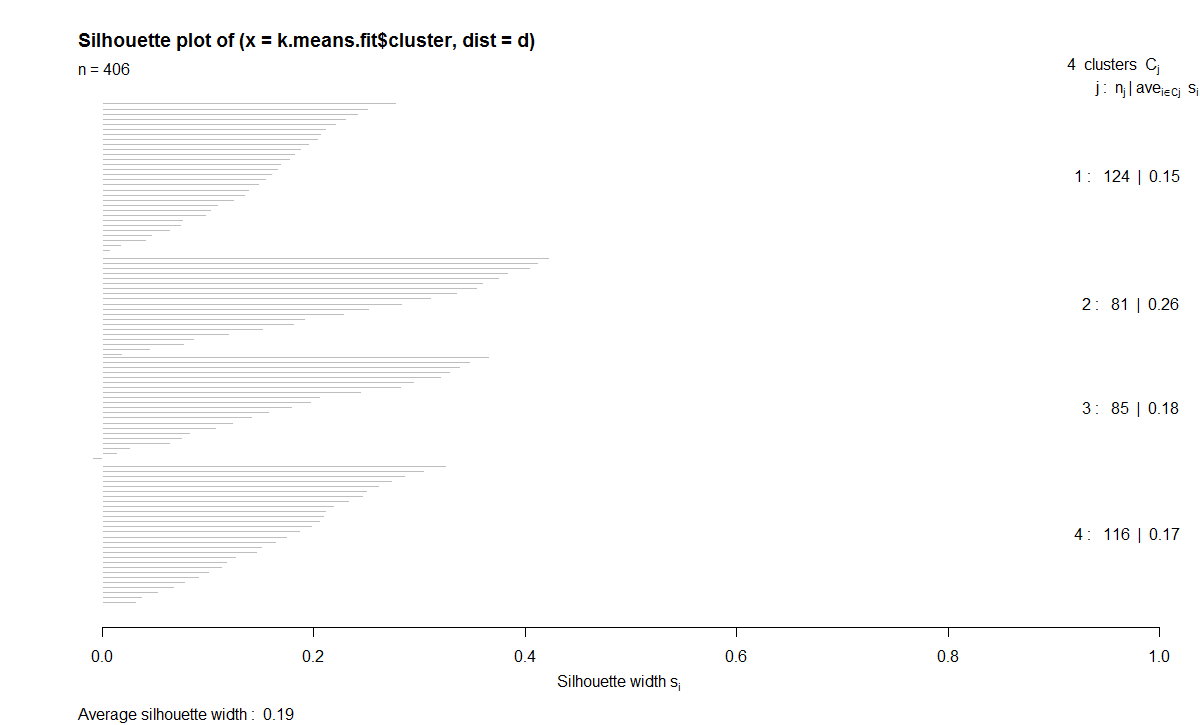


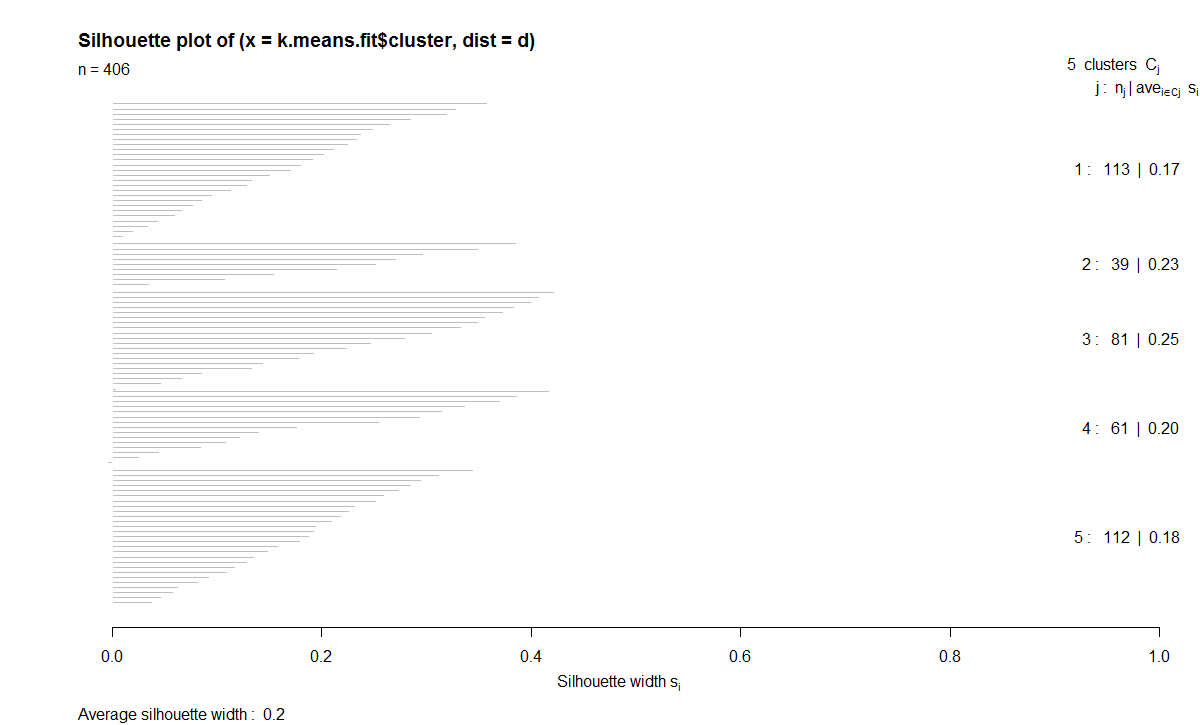
Next, we move to our K-means cluster analysis. First, to internally assess our cluster analysis we generated a within-cluster error sum of squares plot based on the number of clusters used with our data and found the following results:

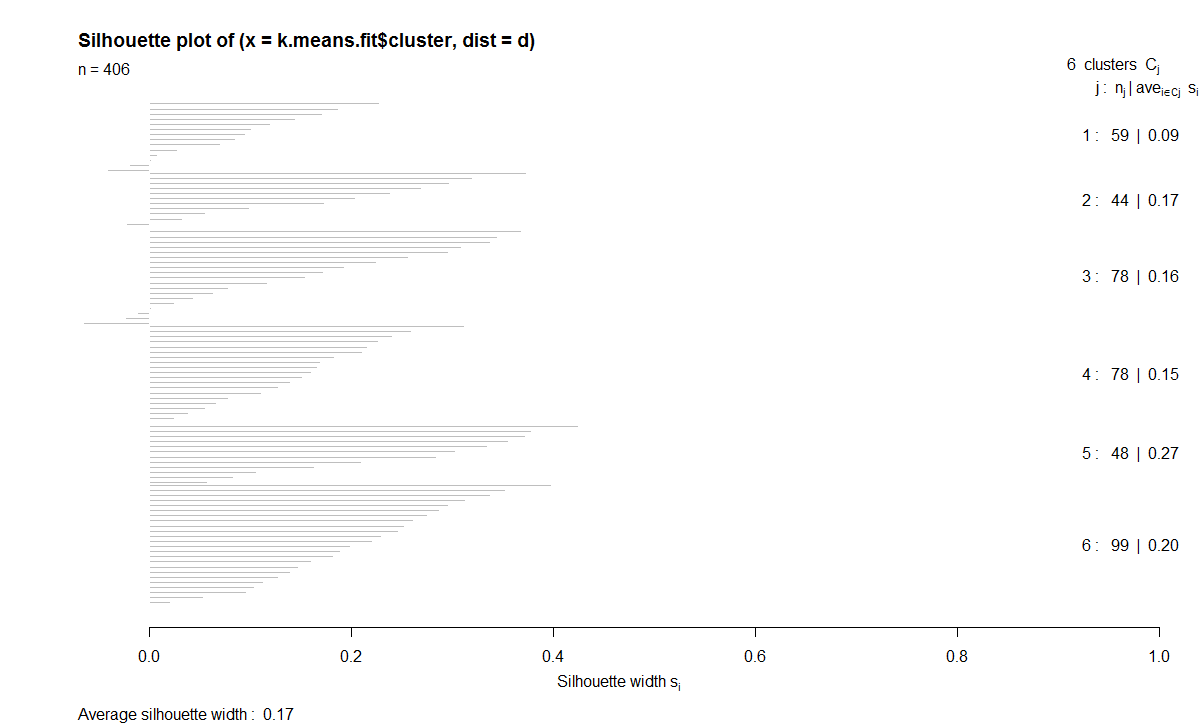


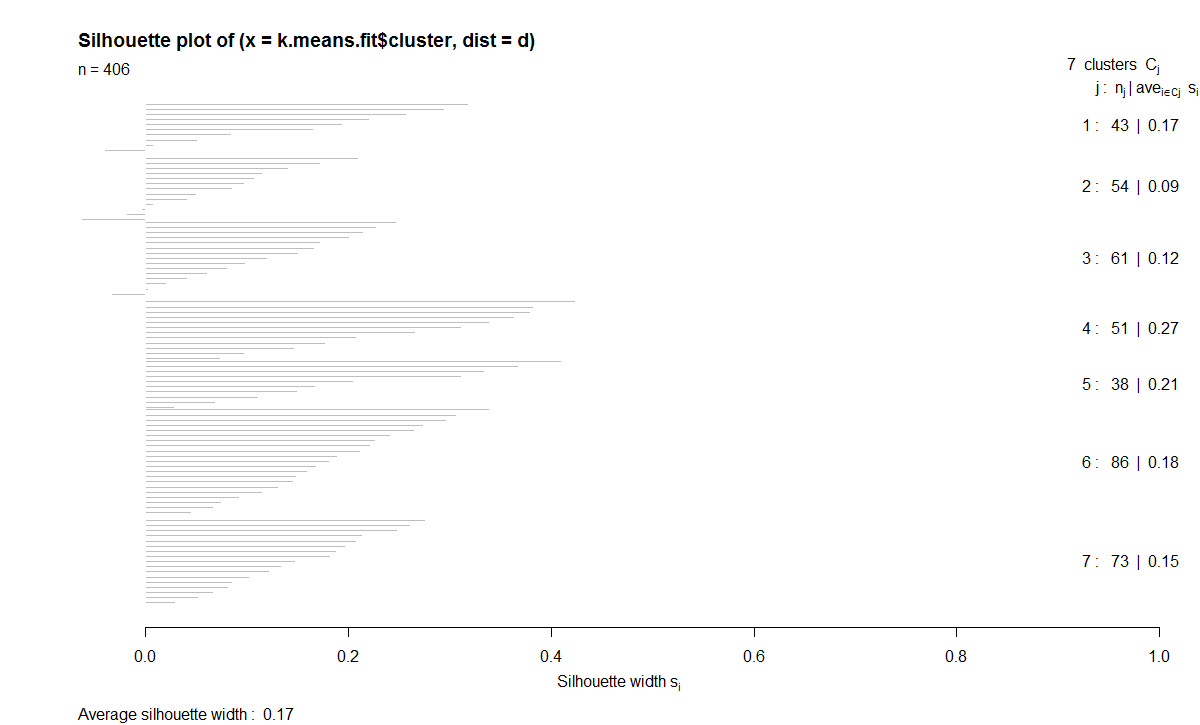
This plot shows that a valid cluster number for this data may be anywhere from 3 to 8 clusters. Using the Silhouette Coefficient (another internal measure assessment tool for clusters), we can try each K-means cluster number within that hypothesized range and determine the best K-value based on which K-value has the highest average Silhouette Coefficient. Below are the results for each Silhouette Coefficient:

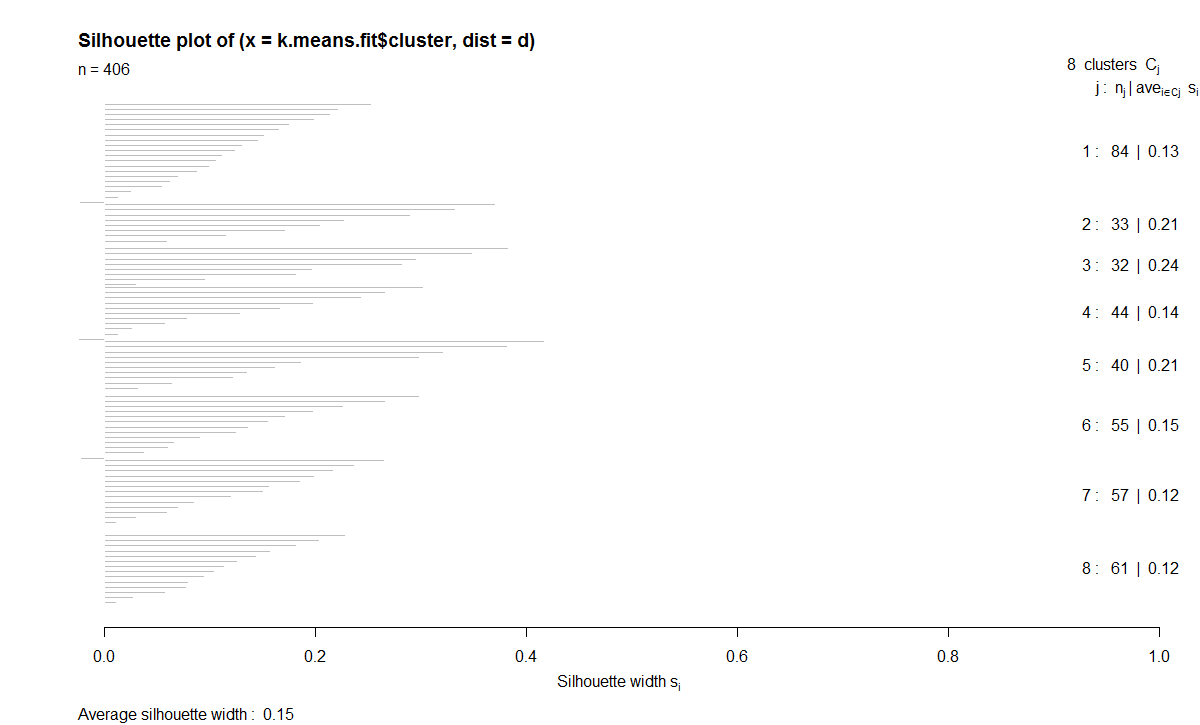




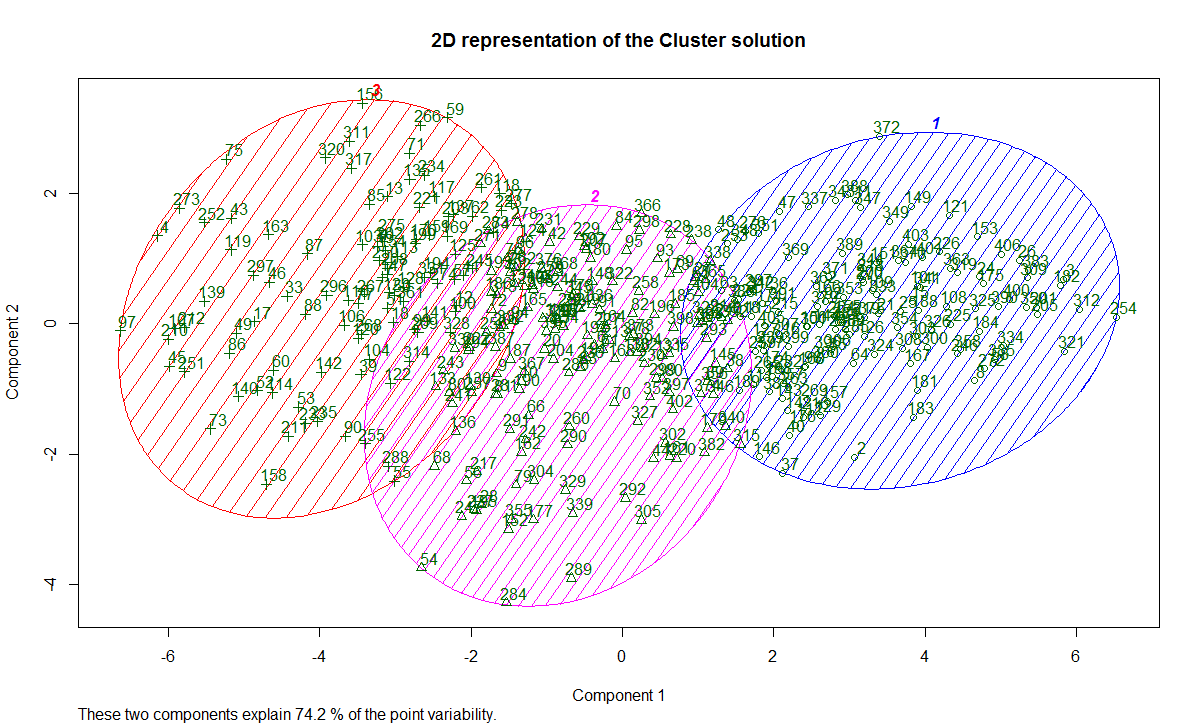






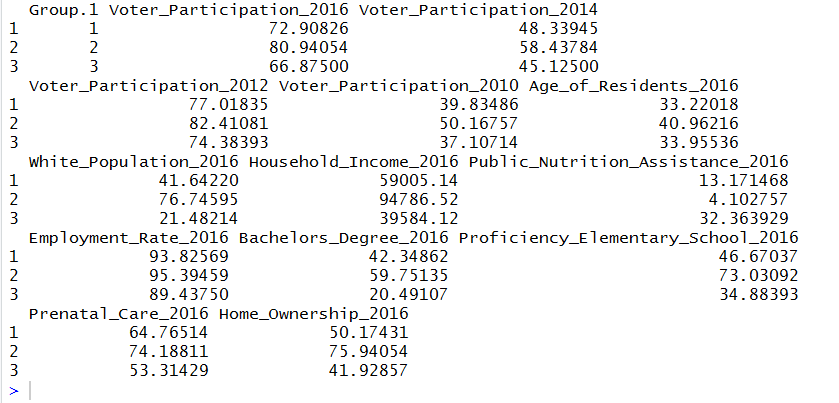


These outcomes suggest that our best K-value for our K-means clustering is three clusters because when using three clusters our average silhouette width was the closest to a value one within our acceptable range of k-values from the within-cluster error sum of squares plot (3 to 8 clusters). Using a K-value = 3, we see that cluster 1 has a size of 142 neighborhoods in it, cluster 2 has 162 neighborhoods, and cluster 3 contains 102 neighborhoods. Below is a 2D representation of the K-means cluster analyses. On this 2D representation, the reader again can observe nearly 75% of the variance is explained by these two dimensions.

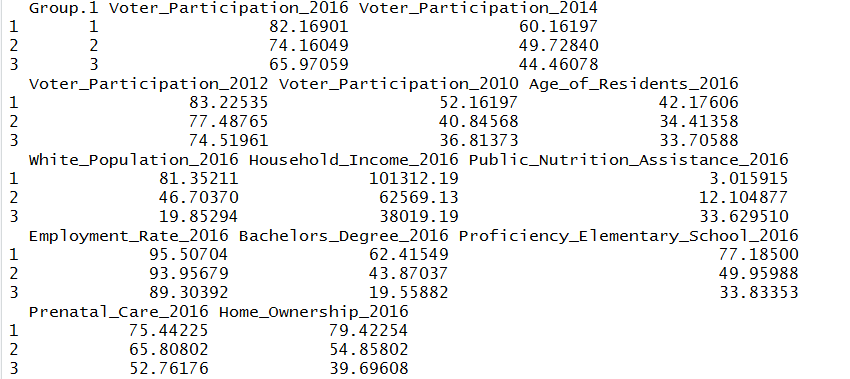


In addition to finding and visualizing 3 clusters in both types of cluster analyses, we also found average values of the neighborhoods grouped within each cluster for all variables considered. These averages within the clusters show some interesting trends between both K-Means and Ward’s Hierarchical Cluster Analyses that are worth noting. These average values for each cluster are posted below:

*Ward’s Hierarchical Clustering:*



*K-Means Clustering:*

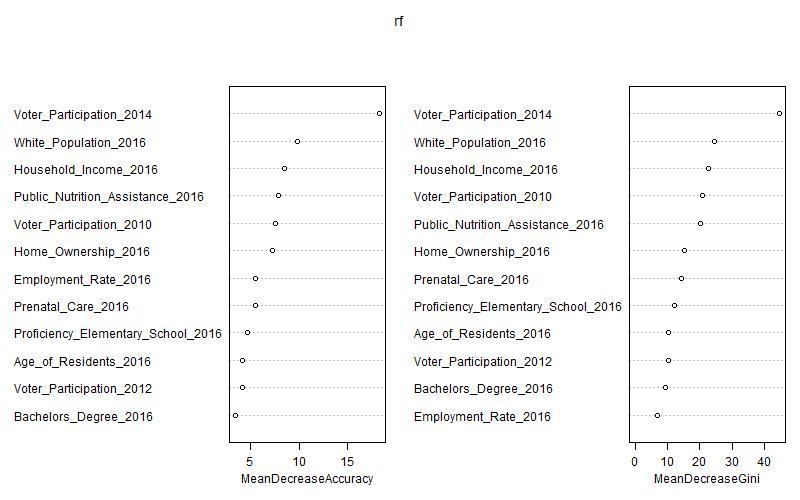


Examining the results from both types of cluster analyses, we see that between the three clusters there is a range of voting participation from 82% to 65% on average. The cluster with the highest voter turnout for the most recent major election also had the highest level of voter participation in the last three major elections, had the highest age, highest proportion of white residents, highest household income, lowest proportion of public nutrition assistance, highest employment rate, highest educated, highest access to prenatal care, and highest levels of homeownership. These trends were observed both in Ward’s hierarchical clustering and in the k-means clustering analyses. The exact opposite trends were true for neighborhoods in the cluster with the lowest average voter turnout in the 2016 major election, except the average age of residents in the hierarchical cluster results only. These results, as a whole, may indicate that there may be some valid groupings of neighborhoods in Mecklenburg County based on a variety of individual and environmental variables such as neighborhood employment rate, homeownership rate, and voting participation.

Furthermore, between these clusters, we also see some trends that emerge in the variable averages of the clusters. However, these conclusions cannot be extrapolated too far since these cluster analyses are only a preliminary analysis of the data that utilizes a fair amount of human judgment. Determining whether or not a certain variable(s) predicts/classifies higher voter turnout requires more elaborate machine learning models such as random forest and logistic regression.

**Random Forest**

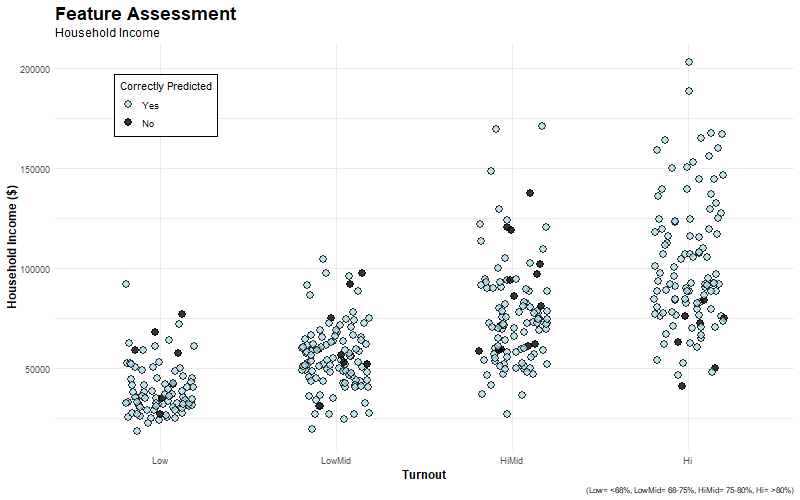
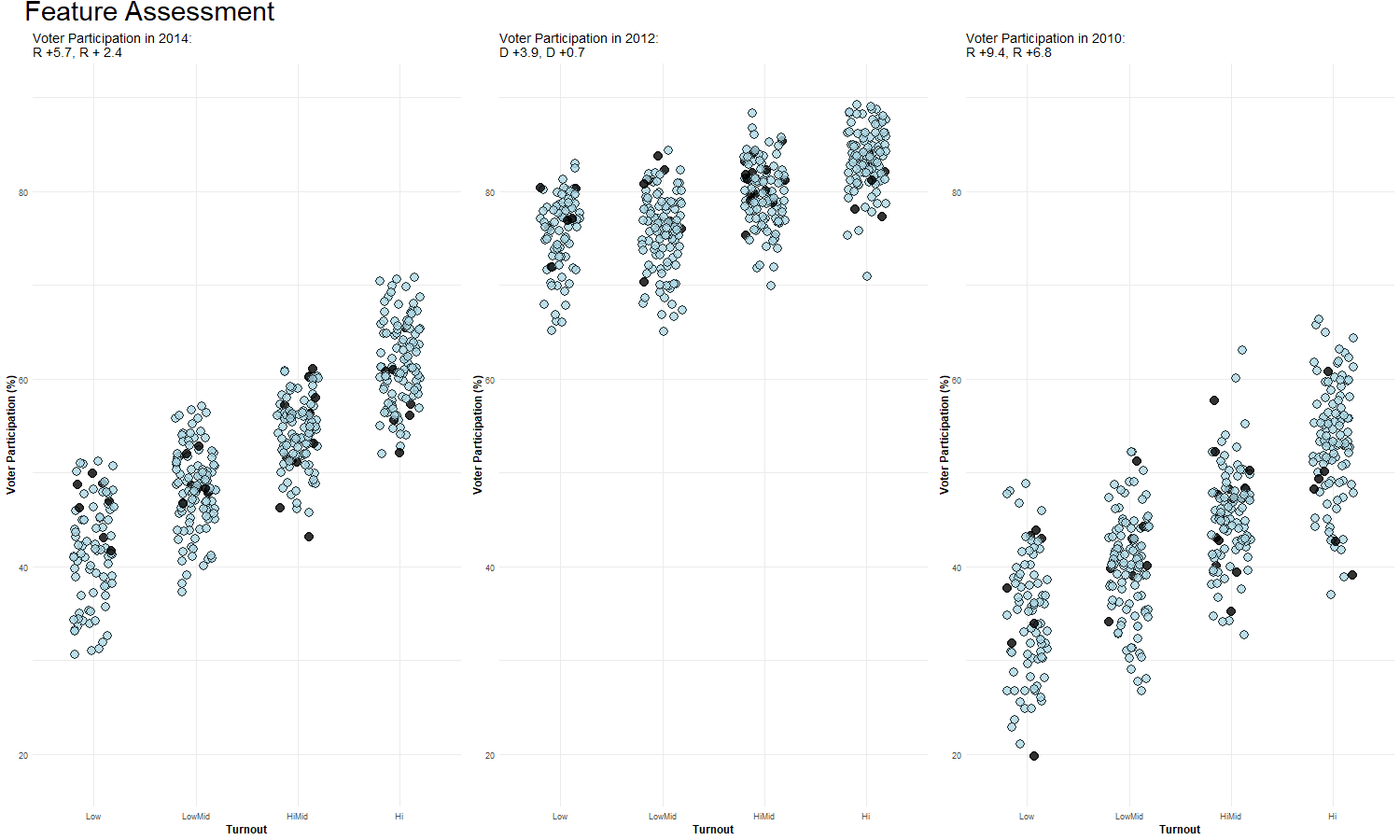
Based on the tests conducted by the team, this was one interpretation of the dataset provided:



While these values contribute to the models produced by multiple iterations of random forests conducted, seven are consistently valuable in assessing the data: voter participation from 2014, 2012, and 2010, homeownership, public nutrition assistance, Caucasian population, and household income. On the other hand, some features that would suggest economic wellness (employment rate, percent with bachelor’s degree, elementary school proficiency) did not provide as valuable insight into voter activity. This result could be because they are indicators, but are more transitive, while the features that were more utilized directly pointed to income differences and variation in means that seems to direct the target.

Though to some level collinear, there is a clear association between disposable income and voting. In part, this could be because the current policy in North Carolina requires a driver’s license in order to vote. However, regardless of background, the public nutrition assistance and the correlated public healthcare assistance, seem to indicate that income has some hand in participation as well.

Although other features provide valuable information, past voting tendencies are the best predictor for future voting, as indicated by the model. As shown below, there is not as much distinction between classes in the 2012 election, but enough variation in 2014 and 2010. The latter two were strong Republican performances on a national level, while 2012 was successful for Democrats, re-electing President Obama.

****

There were two possible conclusions to draw from the voting groups:

1. Republican years draw more voters to the polls, better reflecting the commitment to voting consistently by the citizens of Mecklenburg County. The white population, more likely to vote Republican, of an area is significant in turnout. Polling and projections that look positive could cause cascading that leads these voters to participate in mass.
2. Presidential elections create a ticket that draws all voters. With candidates covered thoroughly on a national scale, it is easier to bring out voters from all classifications. These elections also offer less insight. However, 2014 and 2010 being years for congressional elections, candidates in these elections are less visible, creating distinctions between less interested groups and those that are more committed to participating consistently.

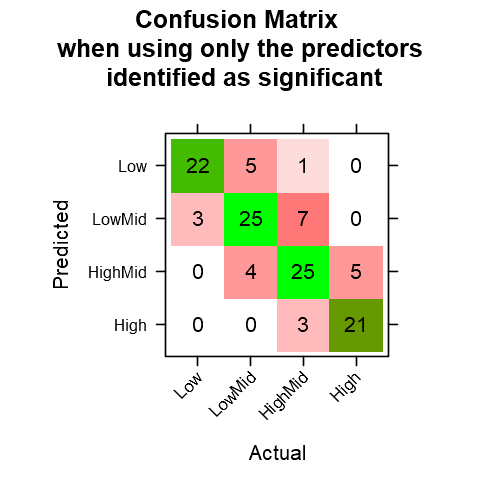
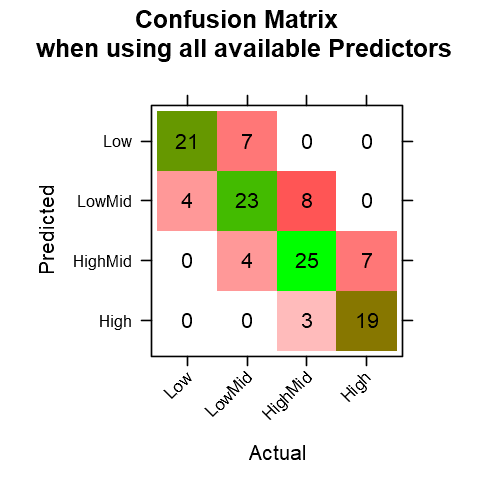
The dataset cannot confirm the true reasoning behind this difference. However, we will likely say the second is true based on the high turnout again during the 2016 presidential election. In any case, running a further-refined random forest indicates a clear value in a few aspects of the new dataset. Variables pointing to income differences and past participation in elections separate the classifications of the target best. Also, variables with limited range, such as the data from presidential elections, can only offer so much insight. Variance in a feature offers the ability to capture differences, and this aided the model significantly.

**Ordinal Logistic Regression**

We ran the ‘clm’ (Cumulative Link Model) model on our cleaned dataset of 406 records with Turnout as the response variable and the remaining 12 variables of our cleaned dataset as our predictors. The summary output of the clm model gives the weight of the predictors, their standard errors, and the significance of the variable in predicting the response class which is measured by the p-value. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis. Thus, we can reject the null hypothesis which is ‘the selected predictor does not influence the Turnout category. The following variables were found to be significant with p-values less than the Significance Level of 0.05.

* Voter\_Participation\_2014
* White\_Population\_2016
* Public\_Nutrition\_Assistance\_2016
* Home\_Ownership\_2016

We used the clm model again to test how the model performs on new data. The prediction was first made using all the available predictors from our cleaned data set and then with only the four significant predictors variables we identified above. The entire data of 406 was split into training and test data sets at a ratio of 70:30. The model was trained using the training data set. The logistic regression was then tested by predicting the ‘Turnout’ category of the test data set which was then compared to the Actual ‘Turnout’ category of the test data record. Using all the available predictor variables yielded an accuracy of 0.727 in the test data set, meaning that the model correctly predicted the ‘Turnout’ category for 72.7% of test data records. This translates to 88 observations out of our 121 observations present in our test data. The diagonal indicated by green in the first Confusion matrix below represents the number of correctly predicted observations in each Turnout category. The other cells in the confusion matrix represent the number incorrectly predicted in each Turnout category. It adds up to 33 which is the remainder of our 121 record test data. The model developed by using only the four significant predictors yielded an accuracy of 76.9%. There is also a confusion matrix produced below representing the four significant predictor variable model representing both our correctly and incorrectly predicted observations.



Below are the statistics results of the prediction model on Test data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Statistics by Class for 4 Predictor Model** | | | | |
|  | **Low** | **LowMid** | **HighMid** | **High** |
| Sensitivity (Recall) | 0.88 | 0.735 | 0.694 | 0.808 |
| Specificity | 0.938 | 0.885 | 0.894 | 0.968 |
| Pos Pred Value (Precision) | 0.786 | 0.714 | 0.735 | 0.875 |
| Neg Pred Value | 0.968 | 0.895 | 0.874 | 0.948 |
| Prevalence | 0.207 | 0.281 | 0.298 | 0.215 |
| Detection Rate | 0.182 | 0.207 | 0.207 | 0.174 |
| Detection Prevalence | 0.231 | 0.289 | 0.281 | 0.198 |
| Balanced Accuracy | 0.909 | 0.81 | 0.794 | 0.888 |

The above table shows that the High Turnout is predicted correctly more often than the other Turnout categories with a Precision of 87.5%. Below is another way of representing the Confusion Matrix as a jitter cluster. Our future work will be to see how the observations move away from the actual category when each predictor is added to our model.



An ideal jitter cluster looks something like the plot below.

To summarize, the Ordinal Logistic Regression is a good method to predict a four-category Target variable. The prediction accuracy of 78% shows that the performance of the model is not bad. We were able to identify the key predictors from our Ordinal Logit Model output.

**Conclusion**

Voter turnout is a significant mechanism for the proper function of a nation's democracy. Although, despite voter turnout importance, research has indicated that both the United States and other established nations have difficulty with lower voter turnout. Reasons for these issues with voter turnout vary greatly from internal to external circumstances. Empirical research has pointed to a multitude of factors that may have a relationship with voter turnout. This research includes reasons such as simple demographics of the voter, to income difference and access to social support resources. Regardless of the reasons, this social occurrence illustrates the need for identifying a model that can best predict voter participation and assist those at risk of not voting.

To create this model, the team used a regional dataset provided by Quality of Life that broke Mecklenburg County down into 474 neighborhood profiles and collected data on each neighborhood profile from voter participation, economic standing, employment rate, and resident characteristics to name a few. After examining descriptive statistics on our variables of interest, we had 13 variables remaining with 406 observations to study. The group used a variety of methods to build our models and attempt to answer our problem statement. Initially, the team utilized cluster analysis to assess and identify potential groupings within our data. Next, the team used two more elaborate analytical models to classify voter participation in our data. These two models are random forest and ordinal logistic regression.

Starting with the cluster analysis, we found some suggestion that the dataset had valid clustering groups after performing a Hopkins Measure. While the results are based primarily on human judgment, the results of our clusters indicate, overall, that cluster(s) with higher voter turnout in Mecklenburg County have higher proportions of variables such as age, income, employment rate, education and lower proportion of non-white residents, and public nutrition assistance to name a few. Overall, the clustering analysis shows some evidence for valid groupings of voter participation in Mecklenburg County based on a variety of life variables. Our random forest analysis revealed that factors such as previous voter participation, home ownership, nutrition assistance, being Caucasian, and income are more valuable in classifying the level of voter participation. However, other predictor variables used were more transient. These results point to the conclusion that there is a distinction between disposable income and voting as well as past voting behavior and voting which is confirmed by a further refined random forest model. Finally, in running the Cumulative Link Model, we conducted an Ordinal Logistic Regression to analyze voter turnout in our neighborhood profiles further. Using a p-value cut-off of .05, the team found voter participation in the 2014 election, White population in 2016, level of public nutrition assistance in 2016, and proportion of homeownership in 2016 to be significant predictors of voter turnout in 2016. In assessing the model with all our original predictor variables, our accuracy was 72.7%, but when only including the four significant predictors our accuracy improved to 76.9%. Results for this analysis conclude that ordinal logistic regression can be a good categorical method to predict low voter turnout from high voter turnout and that key factors that predict this degree of voter turnout include previous voting, race, public resources, and home ownership.

To summarize, the team’s findings indicate that in Mecklenburg County we do observe grouping in voter behavior based on a variety of variables. Furthermore, upon further examination using multiple types of analytical models, we find a consensus that previous voting behavior, race, and income related factors such as home-ownership and access to public assistance programs were significant indicators in voter participation. Limitations of this study include the fact that our data was limited in scope to one U.S. county and only collected data at the neighborhood level rather than at an individual voter level. Furthermore, our data needed significant imputation and manipulation when attempting to run our analytical models. Finally, our dataset was only able to reliably collect complete data in one period making it difficult to examine these variables across time. Further directions for this continued work include, first, finding more robust and complete datasets to replicate the findings of this research. This includes possibly using a dataset that collects these variables of interest at an individual level. Next, these models should be examined using panel data to explore the role of time on our predictors for voter participation. Finally, these, and future, findings could be helpful in identifying the role that certain voting laws and regulatory procedures have on voter participation as a whole.

**References**

(2019). Ordinal Logistic Regression | R Data Analysis Examples. [online] [Stats.idre.ucla.edu](http://stats.idre.ucla.edu/). Available at: <https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/> [Accessed 27 Mar. 2019].

(2019). R help - Package 'MASS' (polr): Error in svd(X) : infinite or missing values in 'x'. [online] Available at: <http://r.789695.n4.nabble.com/Package-MASS-polr-Error-in-svd-X-infinite-or-missing-values-in-x-td4635829.html> [Accessed 27 Mar. 2019].

(2019). What is the difference between categorical, ordinal and interval variables?. [online] [Stats.idre.ucla.edu](http://stats.idre.ucla.edu/). Available at: <https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/> [Accessed 27 Mar. 2019].

Akhmetkarimov, B. 2008. "Revisiting the Profile of the American Voter in the Context of Declining Turnout," *Alternatives: Turkish Journal of International Relations* (7:2/3), pp. 111-133.

Bhatti, Y., Hansen, K. M., Naurin, E., Stolle, D., and Wass, H. 2019. "Can You Deliver a Baby and Vote? The Effect of the First Stages of Parenthood on Voter Turnout," *Journal of Elections, Public Opinion & Parties* (29:1), pp. 61-81.

Budds, D. 2018. "20 Things Every City Can Do to Boost the Quality of Public Life." from https://www.curbed.com/2018/6/21/17484062/center-for-active-design-civic-guidelines

Burden, B. C., Fletcher, J. M., Herd, P., Moynihan, D. P., and Jones, B. M. 2017. "How Different Forms of Health Matter to Political Participation," *Journal of Politics* (79:1), pp. 166-178.

Cebula, R. J. 2017. "Unemployment and Voter Turnout Revisited: A Brief Note," *Electoral Studies* (48), pp. 149-152.

Costa, M., Schaffner, B. F., and Prevost, A. 2018. "Walking the Walk? Experiments on the Effect of Pledging to Vote on Youth Turnout," *PLoS ONE* (13:5), pp. 1-16.

Cumulative Link Models for Ordinal Regression with the R Package ordinal (2019). [online] Available at: <https://cran.rproject.org/web/packages/ordinal/vignettes/clm_article.pdf> [Accessed 27 Mar. 2019].

DeSilver, D. 2018. "U.S. Trails Most Developed Countries in Voter Turnout." from https://www.pewresearch.org/fact-tank/2018/05/21/u-s-voter-turnout-trails-most-developed-countries/

Fenzl, M. 2018. "Income Inequality and Party (De)Polarisation," *West European Politics* (41:6), pp. 1262-1281.

Fraga, B. L. 2016. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout," *American Journal of Political Science* (60:1), pp. 97-122.

Gallego, A. 2010. "Understanding Unequal Turnout: Education and Voting in Comparative Perspective," *Electoral Studies* (29:2), pp. 239-248.

Gimpel, J. G., and Schuknecht, J. E. 2003. "Political Participation and the Accessibility of the Ballot Box," *Political Geography* (22:5), p. 471.

Gollust, S. E., and Rahn, W. M. 2015. "The Bodies Politic: Chronic Health Conditions and Voter Turnout in the 2008 Election," *Journal of Health Politics, Policy & Law* (40:6), pp. 1115-1153.

<https://mcmap.org/qol/#47/>

[Huahuahuahua.github.io](http://huahuahuahua.github.io/). (2019). machine\_learning\_course\_project. [online] Available at: <https://huahuahuahua.github.io/Machine_Learning/machine_learning_course_project.html> [Accessed 27 Mar. 2019].

Jiang, B. 2018. "Homeownership and Voter Turnout in U.S. Local Elections," *Journal of Housing Economics* (41), pp. 168-183.

Lopez, G., and Flores, A. 2017. "Dislike of Candidates or Campaign Issues Was Most Common Reason for Not Voting in 2016." from <http://www.pewresearch.org/fact-tank/2017/06/01/dislike-of-candidates-or-campaign-issues-was-most-common-reason-for-not-voting-in-2016/>

Ordinal Logistic Regression | R Data Analysis Examples. [online] [Stats.idre.ucla.edu](http://stats.idre.ucla.edu/). Available at: <https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/> [Accessed 27 Mar. 2019].

Parlapiano, A., and Pearce, A. 2016. "For Every 10 U.S. Adults, Six Vote and Four Don’t. What Separates Them?," *New York Times* (165:57355), pp. A16-A16.

Preuss, G. G. 1981. "The Effects of Density and Urban Residence on Voter Turnout," *Population and Environment* (4:4), pp. 246-265.

Root, D., and Kennedy, L. 2018. "Increasing Voter Participation in America." from <https://www.americanprogress.org/issues/democracy/reports/2018/07/11/453319/increasing-voter-participation-america/>

Stockemer, D., and Rocher, F. 2017. "Age, Political Knowledge and Electoral Turnout: A Case Study of Canada," *Commonwealth & Comparative Politics* (55:1), pp. 41-62.

Unknown Wikipedia Users. 2019. "Voter Turnout." from <https://en.wikipedia.org/wiki/Voter_turnout#Reasons_for_not_voting>

Urbatsch, R. 2017. "Influenza and Voter Turnout," *Scandinavian Political Studies* (40:1), pp. 107-119.

Vemuri, A. W., Grove, J. M., Wilson, M. A., and Burch Jr, W. R. 2011. "A Tale of Two Scales: Evaluating the Relationship among Life Satisfaction, Social Capital, Income, and the Natural Environment at Individual and Neighborhood Levels in Metropolitan Baltimore," *Environment & Behavior* (43:1), pp. 3-25.

Weber, P., Weber, N., Goesele, M., and Kabst, R. 2018. "Prospect for Knowledge in Survey Data," *Social Science Computer Review* (36:5), pp. 575-590.

Xu, J. 2005. "Why Do Minorities Participate Less? The Effects of Immigration, Education, and Electoral Process on Asian American Voter Registration and Turnout," *Social Science Research* (34:4), pp. 682-702.

# Appendix

library(ggplot2)

library(dplyr)

library(VIM)

library(readxl)

library(mice)

library(car)

library(ROCR)

library(rgdal)

library(tmap)

#Adding additional information

meck\_data <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 2)

meck\_data1 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 3)

meck\_data2 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 4)

meck\_data3 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 5)

meck\_data4 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 6)

meck\_data5 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 7)

meck\_data6 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 8)

meck\_data7 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 9)

meck\_data8 <- read\_excel("qol-data/QOL\_Data\_October\_2018.xls", skip = 1, sheet = 10)

#Voter Participation melt

voter\_part <- meck\_data3[, c(1,10,12,14,16,18)]

#Correlation of Voting Participation

voter\_part\_comp <- voter\_part[complete.cases(voter\_part), ]

cor(voter\_part\_comp)

#Similar election participation (Outcomes and prior polling listed next to per RCP)

cor(voter\_part\_comp[, -c(1, 2)], voter\_part\_comp$Voter\_Participation\_2016) #D +2.1, D +3.3

cor(voter\_part\_comp[, -c(1, 3)], voter\_part\_comp$Voter\_Participation\_2015) #Gallup Prez Ap: 49%

cor(voter\_part\_comp[, -c(1, 4)], voter\_part\_comp$Voter\_Participation\_2014) #R +5.7, R + 2.4

cor(voter\_part\_comp[, -c(1, 5)], voter\_part\_comp$Voter\_Participation\_2012) #D +3.9, D +0.7

cor(voter\_part\_comp[, -c(1, 6)], voter\_part\_comp$Voter\_Participation\_2010) #R +9.4, R +6.8

#CORRELATION TABLES

#SHEET 2

meck\_data\_cor <- merge(voter\_part[, -3], meck\_data, by = "NPA")

colSums(is.na(meck\_data\_cor))

meck\_data\_cor <- meck\_data\_cor[colSums(is.na(meck\_data\_cor)) < 10]

meck\_data\_cor <- meck\_data\_cor[complete.cases(meck\_data\_cor), -1]

a <- cor(meck\_data\_cor[, -c(1:4)], meck\_data\_cor$Voter\_Participation\_2016)

b <- cor(meck\_data\_cor[, -c(1:4)], meck\_data\_cor$Voter\_Participation\_2014)

c <- cor(meck\_data\_cor[, -c(1:4)], meck\_data\_cor$Voter\_Participation\_2012)

d <- cor(meck\_data\_cor[, -c(1:4)], meck\_data\_cor$Voter\_Participation\_2010)

cor(meck\_data\_cor[, -c(1)], meck\_data\_cor$Voter\_Participation\_2016)

#SHEET 3

meck\_data1\_cor <- merge(voter\_part[, -3], meck\_data1, by = "NPA")

colSums(is.na(meck\_data1\_cor))

meck\_data1\_cor <- meck\_data1\_cor[colSums(is.na(meck\_data1\_cor)) < 10]

meck\_data1\_cor <- meck\_data1\_cor[complete.cases(meck\_data1\_cor), -1]

a <- rbind(a, cor(meck\_data1\_cor[, -c(1:4)], meck\_data1\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data1\_cor[, -c(1:4)], meck\_data1\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data1\_cor[, -c(1:4)], meck\_data1\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data1\_cor[, -c(1:4)], meck\_data1\_cor$Voter\_Participation\_2011))

cor(meck\_data1\_cor[, -c(1)], meck\_data1\_cor$Voter\_Participation\_2016)

#SHEET 4

meck\_data2\_cor <- merge(voter\_part[, -3], meck\_data2, by = "NPA")

colSums(is.na(meck\_data2\_cor))

meck\_data2\_cor <- meck\_data2\_cor[colSums(is.na(meck\_data2\_cor)) < 10]

meck\_data2\_cor <- meck\_data2\_cor[complete.cases(meck\_data2\_cor), -1]

a <- rbind(a, cor(meck\_data2\_cor[, -c(1:4)], meck\_data2\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data2\_cor[, -c(1:4)], meck\_data2\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data2\_cor[, -c(1:4)], meck\_data2\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data2\_cor[, -c(1:4)], meck\_data2\_cor$Voter\_Participation\_2011))

cor(meck\_data2\_cor[, -c(1)], meck\_data2\_cor$Voter\_Participation\_2016)

#SHEET 5

meck\_data3\_cor <- merge(voter\_part[, -3], meck\_data3, by = "NPA")

colSums(is.na(meck\_data3\_cor))

meck\_data3\_cor <- meck\_data3\_cor[colSums(is.na(meck\_data3\_cor)) < 10]

meck\_data3\_cor <- meck\_data3\_cor[complete.cases(meck\_data3\_cor), -1]

a <- rbind(a, cor(meck\_data3\_cor[, -c(1:4)], meck\_data3\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data3\_cor[, -c(1:4)], meck\_data3\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data3\_cor[, -c(1:4)], meck\_data3\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data3\_cor[, -c(1:4)], meck\_data3\_cor$Voter\_Participation\_2011))

#SHEET 6

meck\_data4\_cor <- merge(voter\_part[, -3], meck\_data4, by = "NPA")

colSums(is.na(meck\_data4\_cor))

meck\_data4\_cor <- meck\_data4\_cor[colSums(is.na(meck\_data4\_cor)) < 10]

meck\_data4\_cor <- meck\_data4\_cor[complete.cases(meck\_data4\_cor), -c(1, 20)]

a <- rbind(a, cor(meck\_data4\_cor[, -c(1:4)], meck\_data4\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data4\_cor[, -c(1:4)], meck\_data4\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data4\_cor[, -c(1:4)], meck\_data4\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data4\_cor[, -c(1:4)], meck\_data4\_cor$Voter\_Participation\_2011))

cor(meck\_data4\_cor[, -c(1)], meck\_data4\_cor$Voter\_Participation\_2016)

#SHEET 7

meck\_data5\_cor <- merge(voter\_part[, -3], meck\_data5, by = "NPA")

#colSums(is.na(meck\_data5\_cor))

#meck\_data5\_cor <- meck\_data5\_cor[colSums(is.na(meck\_data5\_cor)) < 15]

meck\_data5\_cor <- meck\_data5\_cor[complete.cases(meck\_data5\_cor), -1]

a <- rbind(a, cor(meck\_data5\_cor[, -c(1:4)], meck\_data5\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data5\_cor[, -c(1:4)], meck\_data5\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data5\_cor[, -c(1:4)], meck\_data5\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data5\_cor[, -c(1:4)], meck\_data5\_cor$Voter\_Participation\_2011))

cor(meck\_data5\_cor[, -c(1)], meck\_data5\_cor$Voter\_Participation\_2016)

#SHEET 8

meck\_data6\_cor <- merge(voter\_part[, -3], meck\_data6, by = "NPA")

colSums(is.na(meck\_data6\_cor))

meck\_data6\_cor <- meck\_data6\_cor[colSums(is.na(meck\_data6\_cor)) < 30]

meck\_data6\_cor <- meck\_data6\_cor[complete.cases(meck\_data6\_cor), -1]

a <- rbind(a, cor(meck\_data6\_cor[, -c(1:4)], meck\_data6\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data6\_cor[, -c(1:4)], meck\_data6\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data6\_cor[, -c(1:4)], meck\_data6\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data6\_cor[, -c(1:4)], meck\_data6\_cor$Voter\_Participation\_2011))

#SHEET 9

meck\_data7\_cor <- merge(voter\_part[, -3], meck\_data7, by = "NPA")

colSums(is.na(meck\_data7\_cor))

meck\_data7\_cor <- meck\_data7\_cor[colSums(is.na(meck\_data7\_cor)) < 50]

meck\_data7\_cor <- meck\_data7\_cor[complete.cases(meck\_data7\_cor), -1]

a <- rbind(a, cor(meck\_data7\_cor[, -c(1:4)], meck\_data7\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data7\_cor[, -c(1:4)], meck\_data7\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data7\_cor[, -c(1:4)], meck\_data7\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data7\_cor[, -c(1:4)], meck\_data7\_cor$Voter\_Participation\_2011))

#SHEET 10

meck\_data8\_cor <- merge(voter\_part[, -3], meck\_data8, by = "NPA")

colSums(is.na(meck\_data8\_cor))

meck\_data8\_cor <- meck\_data8\_cor[colSums(is.na(meck\_data8\_cor)) < 50]

meck\_data8\_cor <- meck\_data8\_cor[complete.cases(meck\_data8\_cor), -1]

a <- rbind(a, cor(meck\_data8\_cor[, -c(1:4)], meck\_data8\_cor$Voter\_Participation\_2016))

b <- rbind(a, cor(meck\_data8\_cor[, -c(1:4)], meck\_data8\_cor$Voter\_Participation\_2014))

c <- rbind(a, cor(meck\_data8\_cor[, -c(1:4)], meck\_data8\_cor$Voter\_Participation\_2012))

d <- rbind(a, cor(meck\_data8\_cor[, -c(1:4)], meck\_data8\_cor$Voter\_Participation\_2011))

rm(meck\_data\_cor, meck\_data1\_cor, meck\_data2\_cor, meck\_data3\_cor, meck\_data4\_cor, meck\_data5\_cor,

  meck\_data6\_cor, meck\_data7\_cor, meck\_data8\_cor)

#Evaluating Correlation Values

#2016 Election

a <- data.frame(cbind(rownames(a), a), stringsAsFactors = FALSE)

colnames(a) <- c("Feature", "Cor")

a$Cor <- as.numeric(a$Cor)

a <- a[order(a$Cor), ]

head(a, 10)

tail(a, 10)

a\_2016 <- a[grepl("^.+(2016)$", a$Feature),]

head(a\_2016, 10)

tail(a\_2016, 10)

#2014 Election

b <- data.frame(cbind(rownames(b), b), stringsAsFactors = FALSE)

colnames(b) <- c("Feature", "Cor")

b$Cor <- as.numeric(b$Cor)

b <- b[order(b$Cor), ]

head(b, 10)

tail(b, 10)

#2012 Election

c <- data.frame(cbind(rownames(c), c), stringsAsFactors = FALSE)

colnames(c) <- c("Feature", "Cor")

c$Cor <- as.numeric(c$Cor)

c <- c[order(c$Cor), ]

head(c, 10)

tail(c, 10)

#2010 Election

d <- data.frame(cbind(rownames(d), d), stringsAsFactors = FALSE)

colnames(d) <- c("Feature", "Cor")

d$Cor <- as.numeric(d$Cor)

d <- d[order(d$Cor), ]

head(d, 10)

tail(d, 10)

#CREATING DATAFRAME FOR TESTING

data <- cbind(voter\_part[,-3], meck\_data[, c(6, 16, 20, 22, 26)], meck\_data1[, c(2, 4, 8)],

          meck\_data2[, c(2, 4, 6, 26, 39)], meck\_data5[, c(34, 41, 48)],

          meck\_data6[, 114], meck\_data7[, 4],

          meck\_data8[, c(2, 19)])

rm(meck\_data, meck\_data1, meck\_data2, meck\_data3, meck\_data4, meck\_data5,

  meck\_data6, meck\_data7, meck\_data8, voter\_part)

#ASSESSING NA's and OUTLIERS

#Missing values

colSums(is.na(data))

aggr\_plot <- aggr(data, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

#Missing observations

data$NPA[is.na(data$Early\_Care\_Proximity\_2017)]

data$NPA[is.na(data$Long\_Commute\_2016)]

data$NPA[is.na(data$Home\_Ownership\_2016)]

data$NPA[is.na(data$Household\_Income\_2016)]

#Number of NA's by row

sum(is.na(data[data$NPA == 62,]))

sum(is.na(data[data$NPA == 122,]))

sum(is.na(data[data$NPA == 243,]))

sum(is.na(data[data$NPA == 285,]))

sum(is.na(data[data$NPA == 456,]))

sum(is.na(data[data$NPA == 468,]))

#Removing incomplete observations

data <- data[rowSums(is.na(data)) < 5, ]

aggr\_plot <- aggr(data, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

colSums(is.na(data))

#Correlation Heat Map (complete cases)

cor(data[complete.cases(data), -1])

#IMPUTING MISSING VALUES

#Margin plot for Violent Crime Rate

data\_cor <- data[complete.cases(data), ]

cor(data\_cor[, -c(1, 23)], data\_cor$Violent\_Crime\_Rate\_2016)

cor(data\_cor[, -c(1, 21)], data\_cor$`Public\_Health\_Insurance \_2017`)

#Violent crime vs Public Health Insurance

marginplot(data[c(21,23)])

#Violent crime vs Public Nutrition Assistance

marginplot(data[c(12,23)])

#Public Health Insurance vs Public Nutrition Assistance

marginplot(data[c(12,21)])

#Public Nutrition and Public Health are at R^2 = .97

sum(is.na(data[, -c(1, 21)]))

x <- list("Household\_Income\_2016", "Public\_Nutrition\_Assistance\_2016", "Proficiency\_Elementary\_School\_2016",

      "Births\_to\_Adolescents\_2016", "Prenatal\_Care\_2016", "Public\_Health\_Insurance \_2017",

      "Violent\_Crime\_Rate\_2016")

#Imputed value table

na\_table <- data[,colnames(data) %in% x]

na\_table <- data.frame(ifelse(is.na(na\_table), 1, 0))

na\_table <- cbind(data$NPA, na\_table)

#Impute values

data\_imp <- data[, -c(1, 21)]

tempData <- mice(data\_imp, m=5, maxit=50, meth='pmm', seed=500)

summary(tempData)

tempData$imp$Violent\_Crime\_Rate\_2016

data\_imp <- complete(tempData, 1)

data\_imp <- cbind(data$NPA, data\_imp)

colnames(data\_imp)[1] <- "NPA"

#Final VIM plot

aggr\_plot <- aggr(data\_imp, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

colSums(is.na(data\_imp))

#OUTLIERS

#Ploting features

boxplot(data\_imp$Household\_Income\_2016)

boxplot(data\_imp$Population\_Density\_2016)

ggplot(stack(data\_imp[, -c(1, 11)]), aes(x = ind, y = values)) +

 geom\_boxplot() +

 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

 ggtitle("Outliers from QoL Dataset") +

 xlab("Variables")

remove\_outlier = function(df, column) {

 #find oyputs from the boxplot's stats output

 Q1 <- summary(df[,column])[2]

 Q3 <- summary(df[,column])[5]

 IQR <- (Q3 - Q1)

 lowerlimit <- (Q1 - 1.5\*(IQR))

 upperlimit <- (Q3 + 3\*(IQR))

 cat('Column:', column, '\n')

 cat('Q1:', Q1, '\n')

 cat('Q3:', Q3, '\n')

 cat('IQR:', IQR, '\n')

 cat('lowerlimit:', lowerlimit, '\n')

 cat('upperlimit:', upperlimit, '\n')

 #remove the outliers

 rec1 <- dim(df)[1]

 df <- df[df[,column]>lowerlimit & df[,column]<upperlimit,]

 rec2 <- dim(df)[1]

 cat('Number of outliers removed from ', column, ' :', (rec1-rec2), '\n')

 df

}

outliers\_table <-

 remove\_ouliers\_from\_all\_columns <- function(testdata) {

for (col in colnames(testdata)){

  if (col == "NPA"){

    next

  }

  testdata <- remove\_outlier(testdata, col)

}

testdata

 }

str(data\_imp)

testdata <- remove\_ouliers\_from\_all\_columns(data\_imp[, -c(6, 7, 12, 13, 15, 23, 25)])

dim(testdata)

colnames(testdata)

#Binning high voter turnout

#All Data POints

data\_imp$High\_Turnout <- as.factor(ifelse(data\_imp$Voter\_Participation\_2016 >= 75, 1, 0))

table(data\_imp$High\_Turnout == "1")

summary(data\_imp$Voter\_Participation\_2016)

ggplot(data\_imp, aes(Voter\_Participation\_2016)) +

 geom\_density(fill = "lightblue", alpha = 0.6) +

 geom\_vline(xintercept = 75, linetype = "dashed", color = "red", size = 1)

data\_test <- data\_imp[complete.cases(data\_imp), -(1:2)]

#Run Logistic Regression

smp\_size <- floor(0.55 \* nrow(data\_test))

set.seed(25)

train\_ind <- sample(seq\_len(nrow(data\_test)), size = smp\_size)

train <- data\_test[train\_ind, ]

test <- data\_test[-train\_ind, ]

logmodel <- glm(High\_Turnout ~ ., data = train, family = "binomial")

train$pred <- predict(logmodel, train, type = "response")

train$pred\_val <- ifelse(train$pred >= .5, 1, 0)

table(train$High\_Turnout, train$pred\_val)

ggplot(train, aes(x = Violent\_Crime\_Rate\_2016, y = pred)) +

 geom\_point(aes(color = High\_Turnout))

#Predicting test values

test$pred <- predict(logmodel, test)

test$pred\_val <- ifelse(test$pred >= .5, 1, 0)

table(test$High\_Turnout, test$pred\_val)

ggplot(test, aes(x = Violent\_Crime\_Rate\_2016, y = pred)) +

 geom\_point(aes(color = High\_Turnout))

#ROC

predicted\_values <- predict(logmodel, test, type= "response")

pred <- prediction(predicted\_values, test$High\_Turnout)

perf <- performance(pred, measure = "tpr", x.measure = "fpr")

auc <- performance(pred, measure = "auc")

auc <- auc@y.values[[1]]

roc.data <- data.frame(fpr=unlist(perf@x.values),

                   tpr=unlist(perf@y.values),

                   model="ANN")

ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +

 geom\_ribbon(alpha=0.2) +

 geom\_line(aes(y=tpr)) +

 ggtitle(paste0("ROC Curve w/ AUC=", auc))

#Collinearity

vif(logmodel)

sqrt(vif(logmodel)) > 2

#W/O OUTLIERS

testdata$High\_Turnout <- as.factor(ifelse(testdata$Voter\_Participation\_2016 >= 75, 1, 0))

table(testdata$High\_Turnout == "1")

summary(testdata$Voter\_Participation\_2016)

ggplot(testdata, aes(Voter\_Participation\_2016)) +

 geom\_density(fill = "lightblue", alpha = 0.6) +

 geom\_vline(xintercept = 75, linetype = "dashed", color = "red", size = 1)

test\_data <- testdata[complete.cases(testdata), -2]

#Run Logistic Regression

smp\_size <- floor(0.6 \* nrow(testdata))

set.seed(505)

train\_ind <- sample(seq\_len(nrow(testdata)), size = smp\_size)

train <- test\_data[train\_ind, ]

test <- test\_data[-train\_ind, ]

train\_1 <- train[,-1]

logmodel <- glm(High\_Turnout ~ ., data = train\_1, family = "binomial")

train$pred <- predict(logmodel, train\_1, type = "response")

train$pred\_val <- ifelse(train$pred >= .5, 1, 0)

table(train$High\_Turnout, train$pred\_val)

ggplot(train, aes(x = Violent\_Crime\_Rate\_2016, y = pred)) +

 geom\_point(aes(color = High\_Turnout))

#Predicting test values

test\_1 <- test[,-1]

test$pred <- predict(logmodel, test\_1)

test$pred\_val <- ifelse(test$pred >= .5, 1, 0)

table(test$High\_Turnout, test$pred\_val)

ggplot(test, aes(x = Violent\_Crime\_Rate\_2016, y = pred)) +

 geom\_point(aes(color = High\_Turnout))

# Influential Observations

# added variable plots

av.Plots(logmodel)

# Cook's D plot

# identify D values > 4/(n-k-1)

cutoff <- 4/((nrow(data\_imp)-length(logmodel$coefficients)-2))

plot(logmodel, which=4, cook.levels=cutoff)

# Influence Plot

influencePlot(logmodel, id.method="identify", main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )

#ROC

predicted\_values <- predict(logmodel, test, type= "response")

pred <- prediction(predicted\_values, test$High\_Turnout)

perf <- performance(pred, measure = "tpr", x.measure = "fpr")

auc <- performance(pred, measure = "auc")

auc <- auc@y.values[[1]]

roc.data <- data.frame(fpr=unlist(perf@x.values),

                   tpr=unlist(perf@y.values),

                   model="ANN")

ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +

 geom\_ribbon(alpha=0.2) +

 geom\_line(aes(y=tpr)) +

 ggtitle(paste0("ROC Curve w/ AUC=", auc))

#Collinearity

sqrt(vif(logmodel)) > 2

#Added Variable Plotting

avPlots(logmodel, id.n=2, id.cex=0.7)

outlierTest(logmodel)

#Viz on outcomes

meck\_county <- readOGR(dsn = "qol-data", layer = "NPA\_2014\_meck")

summary(meck\_county)

plot(meck\_county)

shp <- merge(meck\_county, data\_imp, by = "NPA")

tm\_shape(shp) +

 tm\_fill("High\_Turnout") +

 tm\_borders()

test\_data <- rbind(train, test)

test\_data$pred\_val <- as.factor(test\_data$pred\_val)

test\_data$correct <- as.factor(ifelse(test\_data$High\_Turnout == test\_data$pred\_val, 1, 0))

shp1 <- merge(meck\_county, test\_data, by = "NPA")

tm\_shape(shp1) +

 tm\_fill("pred\_val") +

 tm\_borders()

tm\_shape(shp1) +

 tm\_fill("correct") +

 tm\_borders()

#Saving dataframes

write.csv(data\_imp, "full\_test\_set.csv")

ggplot(stack(testdata[, -c(1, 9)]), aes(x = ind, y = values)) +

 geom\_boxplot() +

 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

 ggtitle("Outliers from QoL Dataset") +

 xlab("Variables")

write.csv(testdata, "cleaned\_test\_set.csv")

#Create Table for Variables in Test Data

library(gridExtra)

data\_info <- matrix(ncol = 7, nrow = ncol(testdata) - 2)

dim(testdata)

round(mean(testdata[, 18]), 3)

for(i in 1:(ncol(testdata) - 2)){

 data\_info[i, 1] <- colnames(testdata[i + 1])

 data\_info[i, 2] <- quantile(testdata[, i + 1])[[1]]

 data\_info[i, 3] <- quantile(testdata[, i + 1])[[2]]

 data\_info[i, 4] <- quantile(testdata[, i + 1])[[3]]

 data\_info[i, 5] <- round(mean(testdata[, i + 1]), 3)

 data\_info[i, 6] <- quantile(testdata[, i + 1])[[4]]

 data\_info[i, 7] <- quantile(testdata[, i + 1])[[5]]

}

colnames(data\_info) <- c("Variable", "Min", "1st Q.", "Median", "Mean", "3rd Q.", "Max" )

grid.table(data\_info)

png(filename="data\_info.png")

plot(grid.table(data\_info))

dev.off()

**Cluster Analysis Code:**

#Activating packages for cluster analysis#

library(cluster)

library(clustertend)

library(dbscan)

#reading the data file for cluster analysis#

project\_data <- read.csv("final\_cleaned\_test\_set.csv")

head(project\_data)

colnames(project\_data)

#standardization of data for cluster analysis#

df <- scale(project\_data[,2:14])

head(df)

#Ensuring that we have valid clusters#

hopkins(df, n = nrow(df)-1)

#Ward's method Hierachial clustering#

d <- dist(df, method = "euclidean")

H.ward <- hclust(d, method ="ward.D2")

plot(H.ward)

groups <- cutree(H.ward, k=3)

rect.hclust(H.ward, k=3, border="red")

clusplot(df, groups, main='2D representation of the Wards Cluster solution', color=TRUE, shade=TRUE, labels=2, lines=0)

#determining ideal cluster number for K means#

withinssplot <- function(project\_data, nc=15){

 wss <- (nrow(project\_data)-1)\*sum(apply(project\_data,2,var))

 for (i in 2:nc){

   wss[i] <- sum(kmeans(project\_data, centers=i)$withinss)}

 plot(1:nc, wss, type="b", xlab="Number of Clusters",

      ylab="Within groups sum of squares")}

withinssplot(df, nc=10)

d <- dist(df, method = "euclidean")

k.means.fit <- kmeans(df, 3)

plot(silhouette(k.means.fit$cluster, d))

k.means.fit <- kmeans(df, 4)

plot(silhouette(k.means.fit$cluster, d))

k.means.fit <- kmeans(df, 5)

plot(silhouette(k.means.fit$cluster, d))

k.means.fit <- kmeans(df, 6)

plot(silhouette(k.means.fit$cluster, d))

k.means.fit <- kmeans(df, 7)

plot(silhouette(k.means.fit$cluster, d))

k.means.fit <- kmeans(df, 8)

plot(silhouette(k.means.fit$cluster, d))

#final k value for k means cluster#

k.means.fit <- kmeans(df, 3)

attributes(k.means.fit)

k.means.fit$centers

k.means.fit$size

#2D representation of cluster analysis#

clusplot(df, k.means.fit$cluster, main= '2D representation of the Cluster solution', color=TRUE, shade=TRUE, labels=2, lines=0)

#Observation of results within clusters#

df$kmeans <- k.means.fit$cluster

df$hclust <- groups

df$NPA <- project\_data$NPA

final\_data <- merge(x=df, y=project\_data, key="NPA")

aggregate(final\_data[,c("Voter\_Participation\_2016", "Voter\_Participation\_2014", "Voter\_Participation\_2012", "Voter\_Participation\_2010", "Age\_of\_Residents\_2016", "White\_Population\_2016", "Household\_Income\_2016", "Public\_Nutrition\_Assistance\_2016", "Employment\_Rate\_2016", "Bachelors\_Degree\_2016", "Proficiency\_Elementary\_School\_2016", "Prenatal\_Care\_2016", "Home\_Ownership\_2016")], list(final\_data$kmeans), mean)

aggregate(final\_data[,c("Voter\_Participation\_2016", "Voter\_Participation\_2014", "Voter\_Participation\_2012", "Voter\_Participation\_2010", "Age\_of\_Residents\_2016", "White\_Population\_2016", "Household\_Income\_2016", "Public\_Nutrition\_Assistance\_2016", "Employment\_Rate\_2016", "Bachelors\_Degree\_2016", "Proficiency\_Elementary\_School\_2016", "Prenatal\_Care\_2016", "Home\_Ownership\_2016")], list(final\_data$hclust), mean)

**Random Forest Code:**

library(ggplot2)

library(dplyr)

library(ROCR)

library(pROC)

library(rgdal)

library(tmap)

library(randomForest)

library(caret)

library(grid)

library(gridExtra)

set.seed(42)

data <- read.csv("cleaned\_test\_set.csv")

#Visualizing spread of target

ggplot(data, aes(Voter\_Participation\_2016)) +

geom\_density(fill = "lightblue", alpha = 0.6) +

geom\_vline(xintercept = 75, linetype = "dashed", color = "red", size = 1) +

ggtitle("Initial Group Break")

#Creating new splits

summary(data$Voter\_Participation\_2016)

mean(data$Voter\_Participation\_2016)

ggplot(data, aes(Voter\_Participation\_2016)) +

geom\_density(fill = "lightblue", alpha = 0.6) +

geom\_vline(xintercept = 68, linetype = "dashed", color = "red", size = 1) +

geom\_vline(xintercept = 75, linetype = "dashed", color = "red", size = 1) +

geom\_vline(xintercept = 80, linetype = "dashed", color = "red", size = 1) +

ggtitle("New Group Breaks")

#New voter categories

data$Turnout <- cut(data$Voter\_Participation\_2016, breaks = c(-Inf, 68, 75, 80, Inf),

labels = c("Low", "LowMid", "HiMid", "Hi"))

table(data$Turnout)

#Visualizing past participation

ggplot(data, aes(x = Voter\_Participation\_2012, y = Voter\_Participation\_2014, col = Turnout)) +

geom\_point()

#Min/Max of prior turnout by group

data %>%

group\_by(Turnout) %>%

summarise(min\_val = ifelse(n() > 1, min(Voter\_Participation\_2014), 0),

max\_val = ifelse(n() > 1, max(Voter\_Participation\_2014), 0))

data %>%

group\_by(Turnout) %>%

summarise(min\_val = ifelse(n() > 1, min(Voter\_Participation\_2012), 0),

max\_val = ifelse(n() > 1, max(Voter\_Participation\_2012), 0))

data %>%

group\_by(Turnout) %>%

summarise(min\_val = ifelse(n() > 1, min(Voter\_Participation\_2010), 0),

max\_val = ifelse(n() > 1, max(Voter\_Participation\_2010), 0))

#Mapping out groupings

meck\_county <- readOGR(dsn = "qol-data", layer = "NPA\_2014\_meck")

shp <- merge(meck\_county, data, by = "NPA")

tm\_shape(shp) +

tm\_borders() +

tm\_fill("Turnout")

#Running Random Forest

#Train/test

smp\_size <- floor(0.7 \* nrow(data))

train\_ind <- sample(seq\_len(nrow(data)), size = smp\_size)

train <- data[train\_ind, ]

test <- data[-train\_ind, ]

#Mtry

mtry <- tuneRF(train[, -c(1, 2, 15)], train$Turnout, ntreeTry=100,

stepFactor=1.5, improve=0.01, trace=TRUE, plot=TRUE)

best.m <- mtry[mtry[, 2] == min(mtry[, 2]), 1]

print(mtry)

print(best.m)

#Running Model

rf <-randomForest(Turnout~., data=train[, -c(1, 2)], mtry=best.m, importance=TRUE, ntree=100)

print(rf)

#Variable Importance

importance(rf)

varImpPlot(rf)

jpeg("varImpv1.jpeg", 800, 500)

varImpPlot(rf)

dev.off()

#Predicing test values

#Probability matrix for train

pred1 = predict(rf,type = "prob")

#Probability matrix for test

predicted\_values = predict(rf, type = "response", test)

final\_data <- cbind(test, predicted\_values)

#Confusion Matrix

confusionMatrix(final\_data$predicted\_values, final\_data$Turnout)

confusionMatrix(final\_data$predicted\_values, final\_data$Turnout)$table %>%

plotCM()

#ROC

predicted\_values <- predict(rf, test, type = "prob")

pred <- prediction(predicted\_values, test$Turnout)

perf <- performance(pred, measure = "tpr", x.measure = "fpr")

auc <- performance(pred, measure = "auc")

auc <- auc@y.values[[1]]

roc.data <- data.frame(fpr=unlist(perf@x.values),

tpr=unlist(perf@y.values),

model="ANN")

ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +

geom\_ribbon(alpha=0.2) +

geom\_line(aes(y=tpr)) +

ggtitle(paste("ROC Curve w/ AUC=", auc))

#ROC (STACKOVERFLOW)

votes <- rf$votes[, 2]

rf.roc <- roc(test$Turnout, rf$votes[,2])

plot(rf.roc)

auc(rf.roc)

#Visualization: Assessing Variables

data$pred <- predict(rf, data[, -c(1:2)], type = "response")

data$pred\_correct <- as.factor(ifelse(data$Turnout == data$pred, 1, 0))

data$pred\_correct <- factor(data$pred\_correct, levels = rev(levels(data$pred\_correct)))

#Base layer

p <- ggplot(data, aes(x = Turnout, fill = pred\_correct)) +

scale\_fill\_manual("Correctly Predicted", values = c("1" = "LightBLue", "0" = "Black"),

lab = c("Yes", "No")) +

theme\_minimal() +

xlab("Turnout")

#Feature Assessment

p + geom\_jitter(aes(y = Public\_Nutrition\_Assistance\_2016 ),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("Public Nutrition Assistance (%)") +

labs(title = "Feature Assessment", subtitle = "Public Nutrition Assistance",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = c(0.85, 0.85))

png(filename="public\_aid.png", 800, 500)

p + geom\_jitter(aes(y = Public\_Nutrition\_Assistance\_2016 ),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("Public Nutrition Assistance (%)") +

labs(title = "Feature Assessment", subtitle = "Public Nutrition Assistance",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = c(0.85, 0.85))

dev.off()

p + geom\_jitter(aes(y = Household\_Income\_2016),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("Household Income ($)") +

labs(title = "Feature Assessment", subtitle = "Household Income",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = c(0.15, 0.85))

png(filename="income\_viz.png", 800, 500)

p + geom\_jitter(aes(y = Household\_Income\_2016),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("Household Income ($)") +

labs(title = "Feature Assessment", subtitle = "Household Income",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = c(0.15, 0.85))

dev.off()

p + geom\_jitter(aes(y = White\_Population\_2016),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("White Population (%)") +

labs(title = "Feature Assessment", subtitle = "White Population",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = "none")

p + geom\_jitter(aes(y = Home\_Ownership\_2016),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("Home Ownership (%)") +

labs(title = "Feature Assessment", subtitle = "Home Ownership",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.position = "none")

png(filename="ownership\_viz.png", 800, 500)

p + geom\_jitter(aes(y = Home\_Ownership\_2016),

shape = 21, size = 3.5, color = "black", width = .2, alpha = 0.8) +

ylab("Home Ownership (%)") +

labs(title = "Feature Assessment", subtitle = "Home Ownership",

caption = "(Low= <68%, LowMid= 68-75%, HiMid= 75-80%, Hi= >80%)") +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 12),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.position = "none")

dev.off()

#Voter Participation

vp2014 <- p + geom\_jitter(aes(y = Voter\_Participation\_2014),

shape = 21, size = 4.5, color = "black", width = .2, alpha = 0.8) +

ylab("Voter Participation (%)") +

labs(subtitle = "Voter Participation in 2014: \nR +5.7, R + 2.4") +

ylim(18, 90) +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 14),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = "none")

vp2012 <- p + geom\_jitter(aes(y = Voter\_Participation\_2012),

shape = 21, size = 4.5, color = "black", width = .2, alpha = 0.8) +

ylab("Voter Participation (%)") +

labs(subtitle = "Voter Participation in 2012: \nD +3.9, D +0.7") +

ylim(18, 90) +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 14),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = "none")

vp2010 <- p + geom\_jitter(aes(y = Voter\_Participation\_2010),

shape = 21, size = 4.5, color = "black", width = .2, alpha = 0.8) +

ylab("Voter Participation (%)") +

labs(subtitle = "Voter Participation in 2010: \nR +9.4, R +6.8") +

ylim(18, 90) +

theme(plot.title = element\_text(size = 18, face = "bold"),

plot.subtitle = element\_text(size = 14),

plot.caption = element\_text(size = 8),

axis.title = element\_text(size = 12, face = "bold"),

legend.background = element\_rect(fill="white",

size=0.5, linetype="solid"),

legend.position = "none")

#Outcome, Prior Polling

grid.arrange(vp2014, vp2012, vp2010, ncol=3,

top = textGrob(expression("Feature Assessment"),

gp=gpar(fontsize=20, fontface="bold"),vjust=.3,hjust=3.75))

png(filename="voter\_part\_comp1.png", 1500, 900)

grid.arrange(vp2014, vp2012, vp2010, ncol=3,

top = textGrob(expression("Feature Assessment"),

gp=gpar(fontsize=29, fontface="bold"),vjust=.3,hjust=2.7))

dev.off()

**Ordinal Logistic Regression Code:**

library(foreign)

library(ggplot2)

library(MASS) # for using polr function

library(Hmisc)

library(reshape2)

library(readxl)

library(ordinal) # for using clm function

library(dplyr) # for using nrows and filter functions

library(corrplot) # for correlations plot

library(caret) # for confusion matrix

library(e1071) # for confusion matrix

library(lattice) # for confusion matrix plot using levelplot

getwd()

setwd("E:/+++UNCC/Academics/2019\_Spring-DSBA 6211\_Advanced Business Analytics/Group Project/")

getwd()

data <- read.csv("cleaned\_test\_set.csv")

dim(data)

head(data)

colSums(is.na(data))

summary(data$Voter\_Participation\_2016)

correlations <- cor(data)

corrplot(correlations, method="circle")

data$Turnout <- cut(data$Voter\_Participation\_2016,

                   breaks = c(-Inf, 68, 75, 80, Inf),

                       labels = c("Low", "LowMid", "HighMid", "High"))

summary(data$Turnout)

# Code Reference : https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

# We are using polr command from the MASS package to estimate an ordered logistic regression model

## fit ordered logit model and store results 'm'

m <- polr(Turnout ~ Voter\_Participation\_2014 + Voter\_Participation\_2012 + Voter\_Participation\_2010 + Age\_of\_Residents\_2016 + White\_Population\_2016 + Household\_Income\_2016 + Public\_Nutrition\_Assistance\_2016 + Employment\_Rate\_2016 + Bachelors\_Degree\_2016 + Proficiency\_Elementary\_School\_2016 + Prenatal\_Care\_2016 + Home\_Ownership\_2016, data = data, Hess=TRUE)

# Hess=TRUE to have the model return the observed information matrix from optimization (called the Hessian) which is used to get standard errors.

## view a summary of the model

options("scipen"=100, "digits"=4) #Force R not to use exponential notation

print(m)

# The intercepts or cutpoints indicate where the latent variable is cut to make the 4 Turnout categories in our data

summary(m)

m1 <- clm(Turnout ~ Voter\_Participation\_2014 + Voter\_Participation\_2012 + Voter\_Participation\_2010 + Age\_of\_Residents\_2016 + White\_Population\_2016 + Household\_Income\_2016 + Public\_Nutrition\_Assistance\_2016 + Employment\_Rate\_2016 + Bachelors\_Degree\_2016 + Proficiency\_Elementary\_School\_2016 + Prenatal\_Care\_2016 + Home\_Ownership\_2016, data=data)

summary(m1)

coeff <- coef(summary(m1))

cis <- confint.default(m1) # CIs for parameter estimates assuming normality

# If the 95% CI does not cross 0, the parameter estimate is statistically significant.

coeff\_matrix <- cbind(coeff,cis)

coeff\_matrix

odd\_ratio <- exp(cbind(OR = tail(coeff\_matrix$Estimate, -3), cis))

odd\_ratio

colnames(coeff\_matrix)[colnames(coeff\_matrix)=="Pr(>|z|)"] <- "p\_value"

coeff\_matrix <- as.data.frame(coeff\_matrix)

coeff\_matrix

significant\_coeff\_matrix <- coeff\_matrix[coeff\_matrix$p\_value <= 0.05,]

significant\_coeff\_matrix

#Getting Indexes for test dataset, Size = 30% of the data.

# Since we require 70% of the data for training, keeping only 0.3

set.seed(32)

indexes = sample(1:nrow(data), size=0.3\*nrow(data), replace=FALSE)

# Split data

train\_data = data[-indexes,]

cat('Train Data Dimesnsions : ', dim(train\_data), '\n')

test\_data = data[indexes,]

cat('Test Data Dimesnsions : ', dim(test\_data), '\n')

m2 <- clm(Turnout ~ Voter\_Participation\_2014 + Voter\_Participation\_2012 + Voter\_Participation\_2010 + Age\_of\_Residents\_2016 + White\_Population\_2016 + Household\_Income\_2016 + Public\_Nutrition\_Assistance\_2016 + Employment\_Rate\_2016 + Bachelors\_Degree\_2016 + Proficiency\_Elementary\_School\_2016 + Prenatal\_Care\_2016 + Home\_Ownership\_2016, data=train\_data, link = "logit")

summary(m2)

predicted <- predict(m2,test\_data,type = "class")

predicted

table(test\_data$Turnout,predicted$fit)

cm <- confusionMatrix(predicted$fit, test\_data$Turnout)

cm

cm$table

png(filename = "plot1.png", width = 480, height = 480, units = "px")

plot(cm$table)

dev.off()

# https://huahuahuahua.github.io/Machine\_Learning/machine\_learning\_course\_project.html

qplot(test\_data$Turnout

     , predicted$fit

     , colour=test\_data$Turnout

     , geom = c("jitter")

     , main = "Predicted vs. Actual Test data", xlab = "Actual", ylab = "Predicted")

qplot(train\_data$Turnout

     , train\_data$Turnout

     , colour=train\_data$Turnout

     , geom = c("jitter")

     , main = "Predicted vs. Actual Training Data (both are same)", xlab = "Actual", ylab = "Predicted")

A <- matrix(cm$table[1:16], nrow = 4, ncol = 4, byrow = FALSE)

colnames(A) <- c("Low", "LowMid", "HighMid", "High")

rownames(A) <- c("Low", "LowMid", "HighMid", "High")

A <- as.data.frame(A)

A

options(repr.plot.width=4, repr.plot.height=4)

levelplot(t(A[c(nrow(A):1) , ]),

         col.regions = colorRampPalette(c("white","red", "green"))(1e3),

         scales=list(x=list(rot=45)),

         panel=function(...) {

          arg <- list(...)

          panel.levelplot(...)

          panel.text(arg$x, arg$y, round(arg$z,1))},

         main="Confusion Matrix \n when using all available Predictors",

         xlab='Actual',

         ylab='Predicted',

         colorkey=FALSE)

m3 <- clm(Turnout ~ Voter\_Participation\_2014 + White\_Population\_2016 + Public\_Nutrition\_Assistance\_2016 + Home\_Ownership\_2016

         , data=train\_data

         , link = "logit")

summary(m3)

predicted3 <- predict(m3,test\_data,type = "class")

predicted3

cm3 <- confusionMatrix(predicted3$fit, test\_data$Turnout)

cm3

B <- matrix(cm3$table[1:16], nrow = 4, ncol = 4, byrow = FALSE)

colnames(B) <- c("Low", "LowMid", "HighMid", "High")

rownames(B) <- c("Low", "LowMid", "HighMid", "High")

B <- as.data.frame(B)

B

# options(repr.plot.width=4, repr.plot.height=4)

levelplot(t(B[c(nrow(B):1) , ]),

         col.regions = colorRampPalette(c("white","red", "green"))(1e3),

         scales=list(x=list(rot=45)),

         panel=function(...) {

          arg <- list(...)

          panel.levelplot(...)

          panel.text(arg$x, arg$y, round(arg$z,1))},

         main="Confusion Matrix \n when using only the predictors \n identified as significant",

         xlab='Actual',

         ylab='Predicted',

         colorkey=FALSE)