SDS 323 Project: Predicting County-Level Voter Turnout

Aaron Grubbs Kaushik Koirala Khue Tran Matthew Tran

I. Abstract

This report attempts to determine the relevant factors that contribute to voter turnout using data obtained from the 2016 election with four predictive models: KNN, linear regression, random forest, and PCA. Overall KNN produced the best model with an RMSE of 0.09, but the other models performed fairly with RMSEs roughly between 8 and 10. All the models conclude that the most relevant factors in predicting voter turnout to be ones involved in socioeconomic status; specifically negative ones such as lower educational attainment. From these results, political candidates and the government can allocate their resources more effectively to fulfill their voter turnout objectives.

II. Introduction

In the United States, elections for public office occur at several levels. However, elections for president receive the most attention, utilize the most resources, and persuade the most of the public to turnout and vote.[1] In the 2016 election cycle, at the federal level, over 36% of all election spending was on the presidential election alone, while the remaining nearly 64% percent were spent across several hundred congressional elections.[2]. These resources are often dispensed to target particular demographics, and campaigns often rely on the groups of high-turnout voting blocs.[3] In this context, this report aims to determine the best statistical model for predicting voter turnout as a percentage of the approximate voting age population in a presidential election at the county level, using information such as the racial composition of a county, the educational demographics of a county, county poverty rates, and several other county level statistics. This report will also explore which county-level factors are most crucial in determining a county's voter turnout. This report will use county-level data from the 2016 presidential election as the training data to build models that will predict county level voter turnouts. In the future, thhese predictive models can be leveraged, from the campaigns' perspective, to predict which counties will have the highest turnout and more efficiently allocate electoral resources. From a governmental perspective, these models can also be used to predict the counties most vulnerable to low voter participation due to shifting demographics or population metrics, and deploy government resources to instead encourage greater participation.

III. Methods

Data

The data for this report was collected mainly from two different internet sources. One was used to collect demographic information at the county level.[4] This demographic information included descriptive statistics for a county such as the poverty rate, or education levels. The second source specifically provided county level voting statistics, such as the number of Democrat votes, the number of Republican votes, and most importantly the voter turnout.[5] It is important to note that this voter turnout is based on an estimate of the Voting Age Population (VAP) for the county. Because the estimate can be an underestimate, it is possible for a county to report more than 100% turnout, however this has only once happened. The demographic

[1] https://www.americanprogress.org/issues/democracy/reports/2018/07/11/453319/increasing-voter-participation-america/

[2] https://www.opensecrets.org/overview/cost.php?display=T&infl=Y

[3] https://www.aarp.org/politics-society/government-elections/info-2018/power-role-older-voters.html

[4] https://public.opendatasoft.com/explore/dataset/usa-2016-presidential-election-by-county/table/?disjunctive.state

[5] http://proximityone.com/elections2016.htm

information in the first dataset was combined with the turnout information provided in the second dataset. Many feature columns in the demographics dataset, such as weather and location data, were removed to prevent sparse, redundant data from affecting model complexity, interpretability and computational efficiency. Additionally, it must be noted that county level data for counties in Alaska could not be obtained. The simplified data used for analysis can be found **here**. Many of the techniques in this report may use subsets of the features in the linked dataset to build the most optimal model.

Regression Models

The following 4 statistical models will be used for regression to predict the voter turnout level for a county:

- Linear Model
- PCA-based Regression
- KNN
- RandomForests

Metrics such as the RMSE will be compared and discussed for each of the models. The methodology behind the implementation of each of these models will be discussed in this section.

Linear Regression

This method of model production uses a stepwise/sequential approach. Each variable in consideration was evaluated by their significance, missingness, and proportion of variance accounted for on the variable "Turnout" and ordered from most impactful to least impactful. Starting with the most impactful variable "Poverty Rate" additional variables were added to the model in sequence. With each additional variable a model dataset was created by removing instances of missingness and outliers determined by using two times cook's distance as the cutoff. Using the new model data, the new model and the previous model were compared using ANOVA to evaluate whether there was a significant improvement between the models. This process was repeated until the final model was produced. Of note some variables were removed from consideration as the process went on, without testing the variance inflation factor, due to inclusions of previous related variables. For example, the Democrat percentage is closely related to the Republican percentage so only one would be appropriate to include.

PCA-based Regression

The working dataset contains 50 variables, which can lead to overfitting in high dimensional space when used with regression. So we will implement PCA as a dimension reducing method to extract important variables which will then be used in a regression model. The coefficients from the component regression model will not be directly interpretable, but will provide insights into which variables are significant as predictors of voter turnout.

KNN

In this method, we want to perform k nearest neighbors (KNN) to see if any of our features impact the voter turnout. In order to measure this, we select a set of features such that our root-mean squared error (RMSE) is minimized. This is done by analyzing the all of thes features and checking if any combination of them have an impact on the RMSE on whether it decreases or increases based on the number of features used in the model.

For features that have an "N/A" in the data, we are going to assume that these "no shows" are zero - we apply the is.na function to fill in these cells. After that, we eliminate variables that have too many missing values - these include: Poor_Physical_Health, Poor_Mental_Health, Low_Birthweight, Teen_Births, Adult_Smoking, Adult_Obesity, Diabetes, STDs, and HIV_Rate. We also eliminate County and State, since they are categorical features that have no affect on our numerical features. Finally, we scale all of our features since KNN must use scaled features to run properly.

RandomForests

In order to use the RandomForests bagging model, some features present in the data will not be used. The county name is merely a descriptor variable for which we are trying to determine voter turnout and will not be used. The state that a county is in is a categorical variable present in the dataset, but creating one hot features for the 49 states in the dataset will create a sparse dataset. Additionally, some county health statistics features such as HIV rate, physical and mental health, and adult smoking rates will be dropped due to their sparsity.[6] Lastly, the raw vote tally for people who turned out to vote in a particular county will also not be considered as predicting turnout as a percentage when raw vote numbers have already been obtained is futile.

The data will be regressed on Turnout. Because RandomForest is a bagging technique that finds the true signal by counterintuitively amplifying the noise, there aren't concerns of overfitting or a lack cross-validation. The RMSE will be computed on the test set, but the RMSE is expected to be relatively similar across several random train-test splits. Then, with the trained random forest model a variable Importance Plot will be generated that will show each variable's impact on the MSE of the overall model, implying the importance of that particular variable in predicting turnout.

IV. Results

Linear Regression

The following table shows the semi-partial correlations from the constructed linear regression model.

```
##         Poverty_Rate       LT_High_School    Children_1_Parent
##          0.693111169          0.946836549          0.089618455
## Poor_Physical_Health          Teen_Births           Median_Age
##          0.178266868          0.223846382          2.309000006
##              Service             Democrat     White_Not_Latino
##          0.114585363          0.901302366          0.119742616
##                Green    School_Enrollment        Adult_Obesity
##          0.432052927          0.003191164          0.159887079
##      Low_Birthweight     Total_Population          Agriculture
##          0.115273480          0.777422742          0.067962360
##          Libertarian
##          0.407341581
```

The final model uses poverty rate, less than high school education percentage, children living with 1 parent percentage, number of poor physical health days, teen births, median age, service occupation percentage, Democrat percentage, White (Not Latino) population percentage, Green party percentage, school enrollment, adult obesity, low birthweight, total population, agriculture occupation percentage, and Libertarian party percentage to predict the turnout rate of a county during an election. The RMSE of this model is 9.5604484.This model was shown to be overall significant with a p-value <0.05 and can account for 11% of the variance in turnout rate. Using the semi-partial correlation median age is shown to be the most impactful variable accounting for 2.3% of the variance while holding all other variables constant. Based on the significant coefficients the top three variables that increase voter turnout is median age, Democrats percentage, and poverty rate while the top three variables that decrease voter turnout are adult obesity, Libertarian percentage, and the number of poor physical health days.

PCA-based Regression

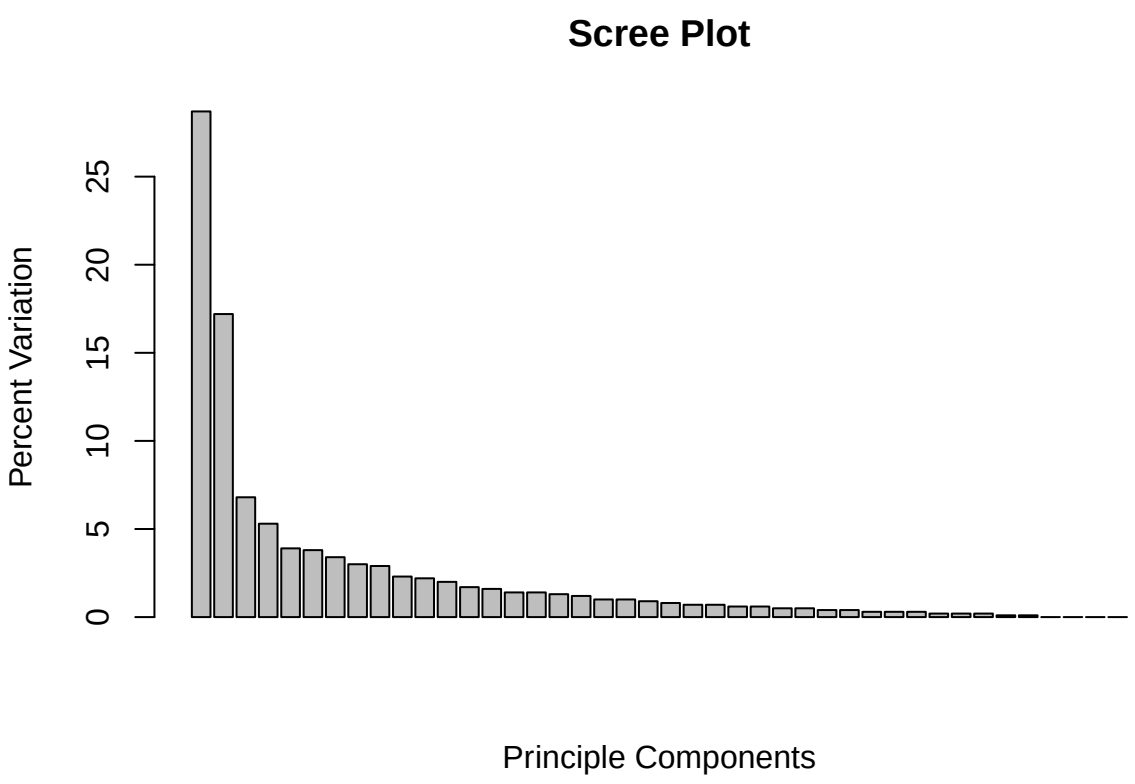
We first load in the dataset and remove additional feature columns with sparse data that would remove a significant amount of observations, namely "Adult_Smoking", "Poor_Physical_Health", "Poor_Mental_Health", and "HIV_Rate". Other variables omitted at this step were categorical labels "County" and "State" that would not be included in the principal components. Also, the variable "Votes" was removed since "Turnout"

[6] While not true for every model, sparse features have to be dropped so an excessive amount of rows and data are not omitted due to missing data.

was calculated from the number of votes. Finally, we temporarily separate the "Turnout" variable as this is the target variable.

The PCA is performed on the remaining 42 variables after scaling. The results are summarized in the table and scree plot below. From the summary, we see that the first PC only accounts for 28.68% of variance in the data. Using the first three components, we can account for 52.68% of variance.

```
## Importance of components:
##                            PC1    PC2     PC3     PC4    PC5    PC6     PC7
## Standard deviation     3.4704 2.6902 1.68720 1.49019 1.2766 1.2600 1.19225
## Proportion of Variance 0.2868 0.1723 0.06778 0.05287 0.0388 0.0378 0.03384
## Cumulative Proportion  0.2868 0.4591 0.52684 0.57971 0.6185 0.6563 0.69016
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.11806 1.10520 0.98559 0.95169 0.92054 0.85174 0.82385
## Proportion of Variance 0.02976 0.02908 0.02313 0.02156 0.02018 0.01727 0.01616
## Cumulative Proportion  0.71992 0.74901 0.77214 0.79370 0.81388 0.83115 0.84731
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.77247 0.76002 0.74232 0.69597 0.66272 0.65379 0.62749
## Proportion of Variance 0.01421 0.01375 0.01312 0.01153 0.01046 0.01018 0.00937
## Cumulative Proportion  0.86152 0.87527 0.88839 0.89992 0.91038 0.92056 0.92993
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.59219 0.55160 0.54078 0.51669 0.49599 0.47948 0.45738
## Proportion of Variance 0.00835 0.00724 0.00696 0.00636 0.00586 0.00547 0.00498
## Cumulative Proportion  0.93828 0.94553 0.95249 0.95884 0.96470 0.97018 0.97516
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.43358 0.41818 0.36881 0.35260 0.34366 0.28583 0.27078
## Proportion of Variance 0.00448 0.00416 0.00324 0.00296 0.00281 0.00195 0.00175
## Cumulative Proportion  0.97963 0.98380 0.98703 0.98999 0.99281 0.99475 0.99650
##                           PC36    PC37   PC38    PC39    PC40     PC41     PC42
## Standard deviation     0.26012 0.18718 0.1713 0.08876 0.08445 0.004896 0.001758
## Proportion of Variance 0.00161 0.00083 0.0007 0.00019 0.00017 0.000000 0.000000
## Cumulative Proportion  0.99811 0.99894 0.9996 0.99983 1.00000 1.000000 1.000000
```



We now look at the loadings for PC1. From ranking the top 10 variables in decreasing absolute values, we see that the strongest associated variables are negative, namely "Child_Family_Poverty", "LT_High_School", followed by "Poverty_Rate", "Teen_Births", "LT6_Poverty", "Elderly_Poverty", and "Diabetes". Conversely, the most positively correlated variables are "GT_High_School", "GT_Bachelors_Degree", and "Management".

```
## Child_Family_Poverty       LT_High_School       GT_High_School
##           -0.2513700           -0.2470270            0.2462103
##         Poverty_Rate          Teen_Births          LT6_Poverty
##           -0.2409496           -0.2372629           -0.2330763
##       Eldery_Poverty  GT_Bachelors_Degree             Diabetes
##           -0.2181859            0.2131081           -0.2068900
##           Management
##            0.2024462
```

```
##
## Call:
## lm(formula = Turnout ~ PC1 + PC2 + PC3, data = test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.868  -6.298  -0.089   6.400  43.872
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57.77828    0.18739 308.334  < 2e-16 ***
## PC1          0.35422    0.05401   6.559 6.49e-11 ***
```

```
## PC2          0.18937    0.06967   2.718  0.00661 **
## PC3         -0.93071    0.11109  -8.378  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.666 on 2657 degrees of freedom
## Multiple R-squared:  0.04342,    Adjusted R-squared:  0.04234
## F-statistic:  40.2 on 3 and 2657 DF,  p-value: < 2.2e-16
```

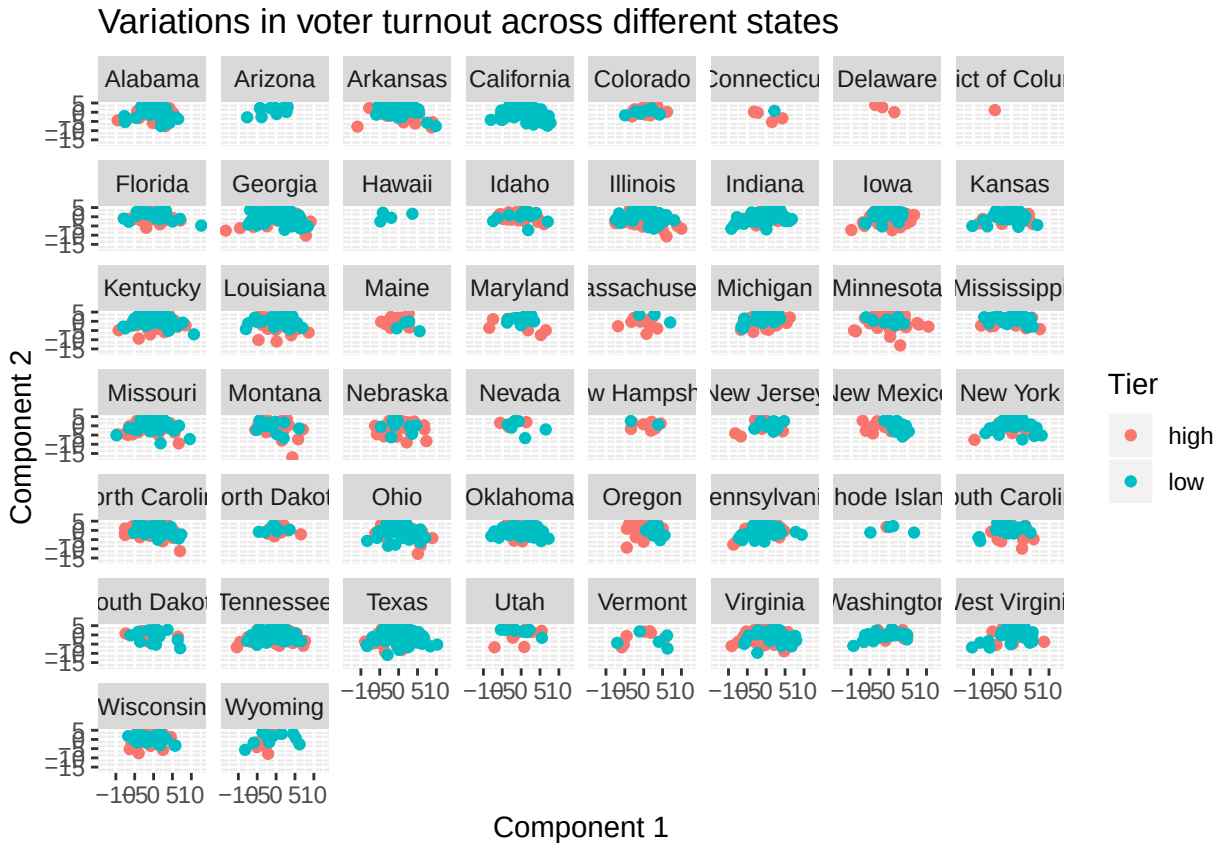
In comparison to the linear model above, the feature variables selected in common using PCA are "Poverty_Rate", "LT_High_School", and "Teen_Births", thus verifying that these are important predictors for voter turnout. Now we build a regression model using the principle components. The first model with PC1, PC2, and PC3 as predictors performed with an RMSE of 9.6591565. Overall, this was a significant model with p<2.2e-16 as summarized.

```
##
## Call:
## lm(formula = Turnout ~ PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 +
##     PC8, data = test2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.669  -6.242  -0.024   6.243  43.830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57.77828    0.18323 315.338  < 2e-16 ***
## PC1          0.35422    0.05281   6.708 2.41e-11 ***
## PC2          0.18937    0.06812   2.780  0.00548 **
## PC3         -0.93071    0.10862  -8.569  < 2e-16 ***
## PC4          0.98133    0.12298   7.980 2.16e-15 ***
## PC5         -0.11288    0.14355  -0.786  0.43173
## PC6          0.38939    0.14544   2.677  0.00747 **
## PC7          1.05136    0.15371   6.840 9.80e-12 ***
## PC8          0.48720    0.16391   2.972  0.00298 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.452 on 2652 degrees of freedom
## Multiple R-squared:  0.08716,    Adjusted R-squared:  0.08441
## F-statistic: 31.65 on 8 and 2652 DF,  p-value: < 2.2e-16
```

While the coefficients of a principle component regression model cannot be easily interpreted, we recognize the results as verification that the strongly correlated variables of children in families in poverty, percentage with high school diploma, and poverty rate are the best predictors that we have. To confirm this, we perform another regression model including up to PC8, which in total accounts for 71.99% variance in the data. This model was also significant (p<2.2e-16) with RMSE 9.4357269, which is not a significant improvement from the previous model using only three PCs.

From the results of PCA, we plot the data along the first two principal components. To better visualize the data since the model does not account for the majority of variations, we split the dataset into two tiers: high turnout (above mean) and low turnout (below mean). The plot between PC1 and PC2 shows no clear distinction between the two tiers, as is the case for PC2 and PC3, and PC1 and PC3 (shown in appendix). This trend is fairly consistent across all states as shown in the faceted plot below with variations in each state.

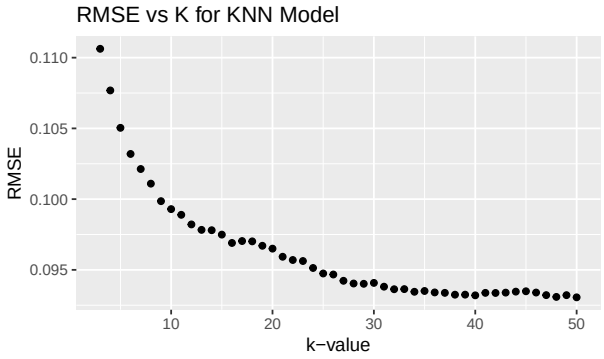
```
## Joining, by = c("State", "Precincts", "Republican", "Democrat", "Green", "Libertarian", "LT_High_Sch
```



KNN

The following is the average (across 50 train-test split partitions) RMSE vs k-value plot for the voter turnout. Here we are using all of the remaining features, and checking if these features will result is a low RMSE:

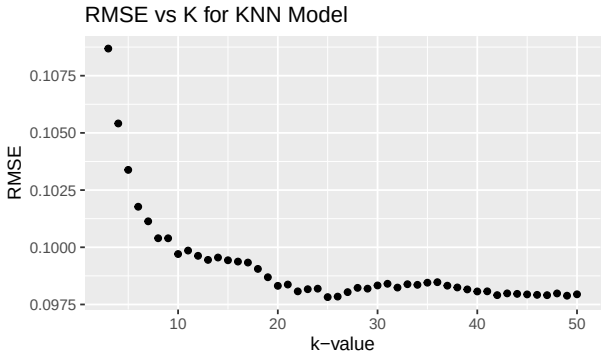
```
## Using parallel package.
##   * Set seed with set.rseed().
##   * Disable this message with options(`mosaic:parallelMessage` = FALSE)
```



The best k-value using knn is 50 and the RMSE at that k-value is 0.0930583.

Now in the second run of the KNN algorithm, we want to remove some more features. In this case we only include the following features for our train/test split: Precincts, Republican, Democrat, Green, Libertarian, LT_High_School, GT_Bachelors_Degree, GT_High_School, Graduate_Degree, School_Enrollment, Median_Earn_2010.

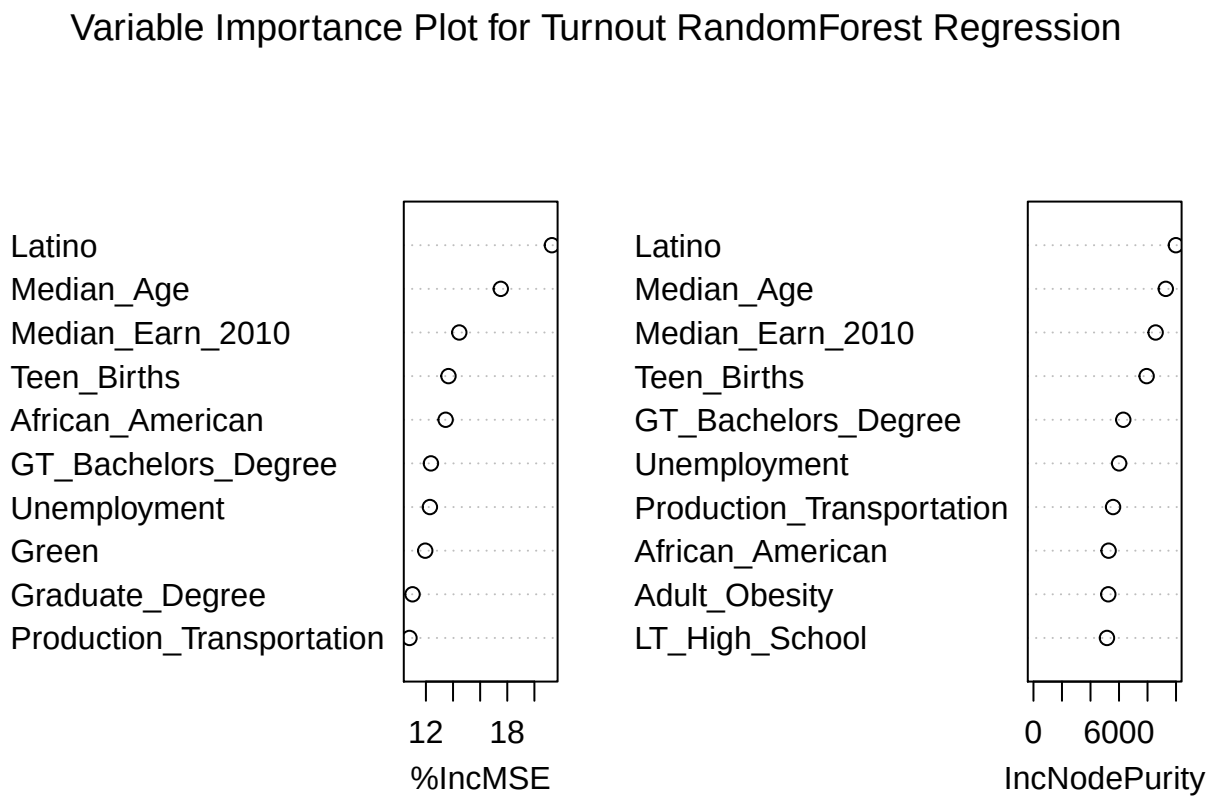
```
## Using parallel package.
##   * Set seed with set.rseed().
##   * Disable this message with options(`mosaic:parallelMessage` = FALSE)
```



The best k-value using knn is 25 and the RMSE at that k-value is 0.097822.

Random Forests

The RMSE for the RandomForests model is 8.9429714. The plot below shows the Variable Importance Plot for the RandomForests model, with the 10 most important variables. This plot shows if the variable is ignored and not allowed to split in the randomforest, how much worse the MSE gets percent wise.



Some of the most important variables, for example, that are contributing to how this model would predict turnout in a county seem to be the Latino population percentage in a county, the median age in a county, and the African-American population percentage in a county.

V. Conclusion

Based on the learning models' RMSEs, KNN seemed to perform the best, which may be due to the inherent flexibility in the model compared to the other methods. It performed at two orders of magnitude better than the other models, but it failed in indicating specific factors that contributed significantly to voter turnout. The other three models (linear regression, PCA and regression, random forest) resulted in lower RMSEs. This is resonable given the range of variables in the dataset and baseline variations in the voting population. The linear and random forest models provided better indications as to which factors could strongly predict turnout. A large limitation in the linear regression model is the removal of 470 counties due to lack of data and 185 counties removed for being outliers. These removed counties could have accounted for other trends that were not included. Also, this method is computationally intensive since the variables are added and removed manually and are subject to human error. Similarly, the random forest approach suffered greatly from unavailable data and due to supporting a greater number of variables. Overall, PCA produced results that were difficult to interpret, but served to confirm the important variables used in linear regression.

The low socioeconomic indicators in these models, such as teen birth rates, unemplotment rates, and poor physical health, negatively correlated with voter turnout for all of the models. This implies that these counties may not be the most efficient places to concentrate on during an election. Another factor that may be important is the relative make-up of the county's political affiliation. Based on some aspects of the linear model, an increase in independent political affiliations tends to decrease voter turnout overall which can be seen as places where one could attempt to polarize the voter base. It is important to note however that there are many variations and factors unaccounted for such as culture and religion in the dataset that may have large impacts on voter tendencies that can be used to improve these models further.