Project 02: Regression, Classification, and Clustering

By: Shubham Chaudhary, Kirun Haque, Michelle Baginski

Questions:

1. We partitioned the dataset into two disjoint sets, a training set and validation set using the holdout method. The training set was 75% and the test set was 25%.

2. Task 3
   - The best performing linear regression model to predict Democratic votes is the one we found to have the highest R-squared value, which was 0.882. The variables that were in this model are Republican Party, Total Population, Percent Hispanic or Latino, Percent Age 29 and Under, Percent Foreign Born,Percent Less than Bachelor's Degree, Percent Rural. The variables for the models were selected on the basis of significance factor, also known as p-value.

   - The best performing linear regression model to predict Republican votes is the one we found to have the highest R-squared value, which was 0.737. The variables that were in this model are Democratic Party, FIPS, Total Population, Percent Black, not Hispanic or Latino, Percent Hispanic or Latino, 'Percent Foreign Born, Percent Age 65 and Older, Median Household Income, Percent Unemployed, Percent Less than High School Degree, Percent Less than Bachelor's Degree, Percent Rural. The variables for the models were selected on the basis of significance factor, also known as p-value.

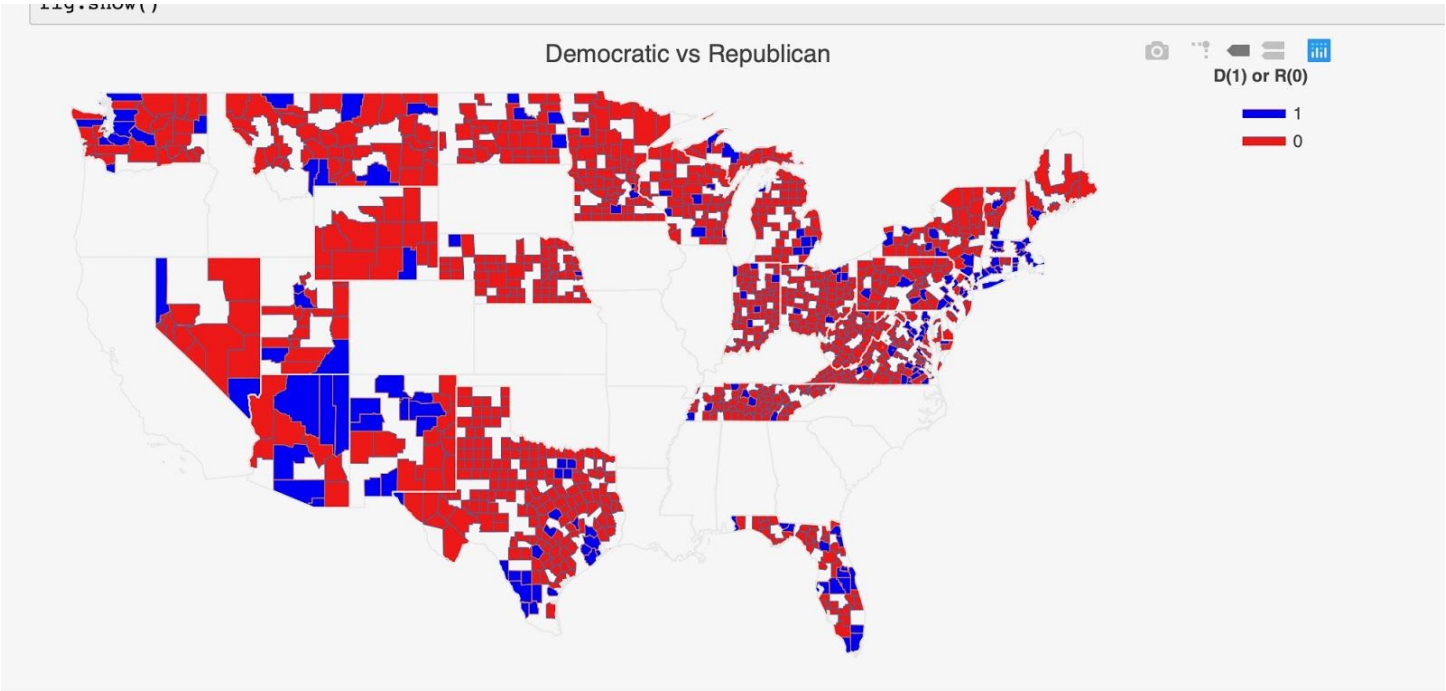
3. Task 4
   The two techniques for classification we have picked are decision trees and SVM. In the decision tree, we have modified the parameters by selecting the criterion to be Gini and the number of splits to be 4. The variables were selected using the correlation coefficient between the independent variables and the variables used in our conclusion from project 1.
   - For decision trees, the best model accuracy is 0.78 and the F1 score is [0.8287037 , 0.55421687] with no modification in the parameters, however we selected a certain amount of variables that is Total Population, Percent Black, not Hispanic or Latino, Percent Hispanic or Latino, Percent Foreign Born, Percent Age 65 and Older, Median Household Income, Percent Unemployed, Percent Less than High School Degree, Percent Less than Bachelor Degree, "Percent Rural.
   - For SVM, the best model accuracy is 0.853 and the F1 score is [0.90756303 0.63934426] with a modification in the selection of the kernel function, which was chosen to be rbf. We selected a certain amount of variables that is Percent White, not Hispanic or Latino, Total Population, Percent Black, not Hispanic or Latino, Percent Hispanic or Latino, Percent Foreign Born, Percent Age 65 and Older, Median Household Income, Percent Unemployed, Percent Less than High School Degree, Percent Less than Bachelor Degree, Percent Rural.

4. Task 5

The two techniques for clustering are K-Means clustering and Agglomerative clustering with Ward's method.

- In K-Means, the unsupervised evaluation metrics are the adjusted rand index, which is 0.225, and the silhouette score which is 0.343. The supervised metrics are 0.788 for the accuracy and the F1 Score is [0.8598338 0.56752137]. The variables we selected are Total Population, Percent Black, not Hispanic or Latino, Percent Hispanic or Latino, Percent Foreign Born, 'Median Household Income, Percent Unemployed, Percent Less than High School Degree, Percent Less than Bachelor's Degree, Percent Rural. We selected the parameters and variables on the basis of the best accuracy and F1 score model.
- In the Agglomerative clustering with Ward's method, the evaluation metrics are the adjusted rand index, which is 0.154 and the silhouette score which is 0.322.
  The supervised metrics are 0.709 for the accuracy and the F1 Score is [0.79233992 0.51738526]. The variables we selected are Total Population, Percent White, not Hispanic or Latino, Percent Hispanic or Latino, Percent Foreign Born, Percent Less than High School Degree, Percent Less than Bachelor's Degree, and Percent Rural. We selected the parameters and variables on the basis of the best accuracy and F1 score model.

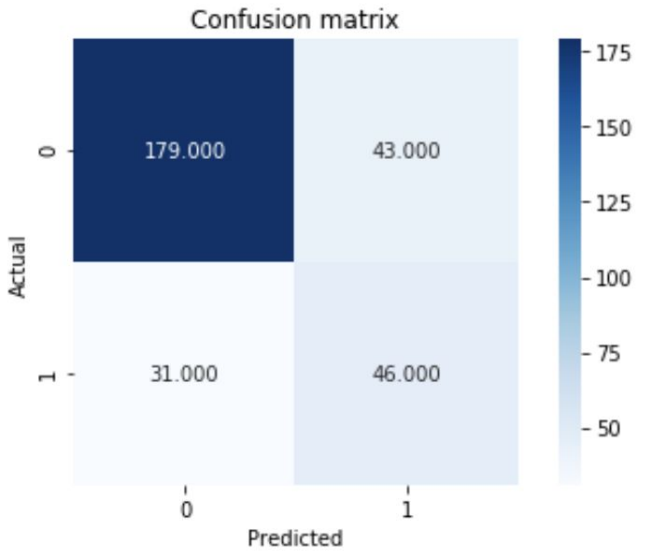
5. Task 6



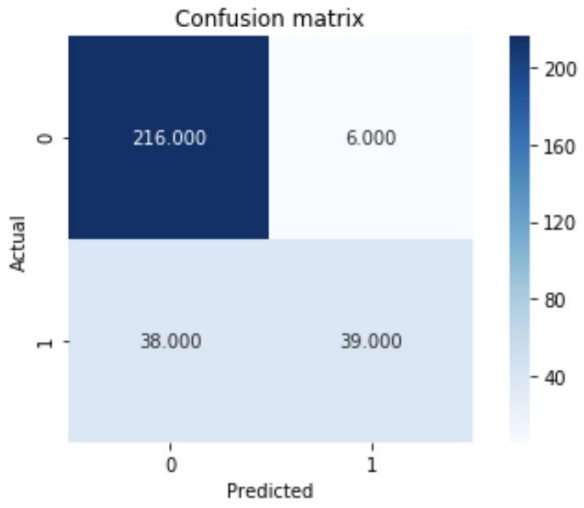
This chart was produced on the basis of the predicted values from the best classification model, which is SVM with the kernel function rbf. By comparing this chart with that of the first project, we concluded that the model was able to predict the Republican counties more accurately than it could predict the Democratic counties.

Best Performance Plots

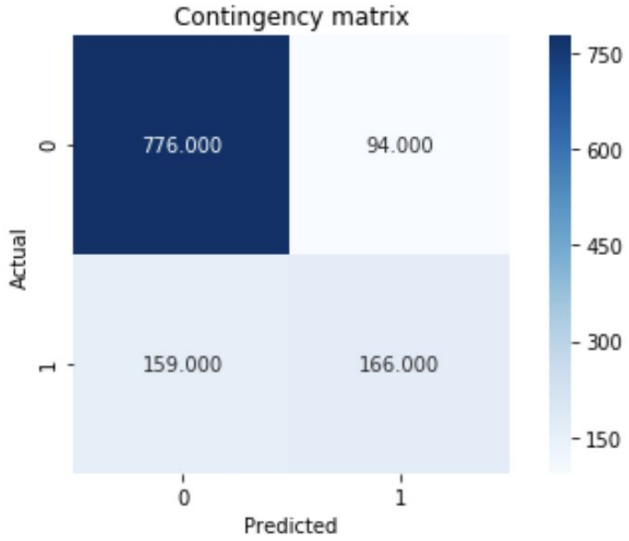
Decision Trees



SVM



K-Means



Agglomerative Clustering with Ward's