CS418 - INTRODUCTION TO DATA SCIENCE

PROJECT 2 - REPORT

Task 1. Partition the merged dataset into a training set and a validation set using the holdout method or the cross-validation method. How did you partition the dataset?

The data was partition using the holdout method. The data was split using 75% as the training set and 25% for the validation set. Since we were predicting votes and party the partition used democrat, republican and party variables and the other part used all the other variables that can be applied.

Task 2. Standardize the training set and the validation set.

To standardize the training set and validation set, the training set was used to do the scaling. First we pass it to the fit function to get the means and standard deviation for the columns. Then the scaling was applied to the training set and validation set which each returned a 2d array.

Task 3. Build a linear regression model to predict the number of votes cast for the Democratic party in each county. Consider multiple combinations of predictor variables. Compute evaluation metrics for the validation set and report your results. What is the best performing linear regression model? What is the performance of the model? How did you select the variables of the model? • Repeat this task for the number of votes cast for the Republican party in each county.

Best performing regression model was LASSO regression model with all predictors in achieved an R squared value of 0.90 when predicting democratic votes and it achieved a value of 0.83 when predicting republican votes. The LASSO regression was compared to a simple linear regression model with percent foreign born and median household income as predictors which scored 0.2157 for predicting democratic votes and 0.1234 when predicting republican votes. The variables were selected based on what was thought to be the variables that correlated the best with which party had the most votes.

Task 4. Build a classification model to classify each county as Democratic or Republican. Consider at least two different classification techniques with multiple combinations of parameters and multiple combinations of variables. Compute evaluation metrics for the validation set and report your results. What is the best performing classification model? What is the performance of the model? How did you select the parameters of the model? How did you select the variables of the model?

Four classification models are used to test the dataset and get to know the best classification model. The models used were k-nearest Neighbors, Support Vector Machine(SVM), Decision trees and Naive Bayes classifiers. The model is built upon a few variables(predictors) that was selected based upon the higher impact they have on the output of the result. The predictors chosen were ‘Percent White, not Hispanic or Latino’, ‘Percent Black, not Hispanic or Latino’, ;Percent Hispanic or Latino’, ‘Percent Age 29 or Under’, ‘Percent Age 65 or Older’, ‘Percent Less than High School Degree’, ‘Percent Less than Bachelor’s Degree’ and ‘Percent Rural’.

The best performing model was SVM with RBF kernel. The performance of this model was determined based on the value of accuracy which was found to be 0.8394, which is better than all the other models. All the classification models were built using different parameters in each classifier, but the default parameters were found to be the best performing ones among all. Those iterations with different parameters are not shown in the code.

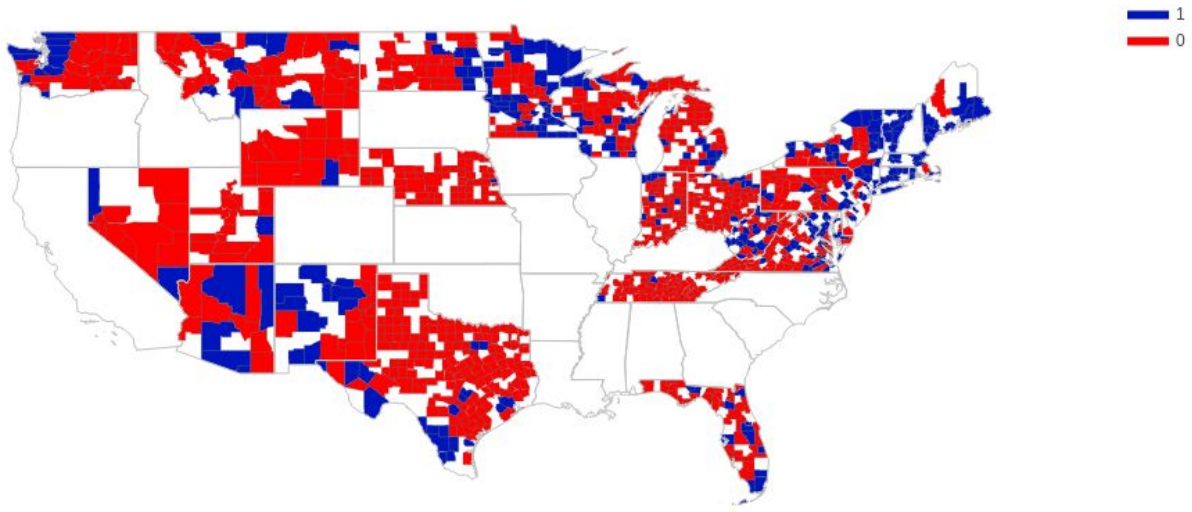
The classifier classifies each state based on the selected predictors into either 0 or 1 relating to Republican or Democratic, respectively.

Task 5. Build a clustering model to cluster the counties. Consider at least two different clustering techniques with multiple combinations of parameters and multiple combinations of variables. Compute unsupervised and supervised evaluation metrics for the validation set with the party of the counties (Democratic or Republican) as the true cluster and report your results. What is the best performing clustering model? What is the performance of the model? How did you select the parameters of model? How did you select the variables of the model?

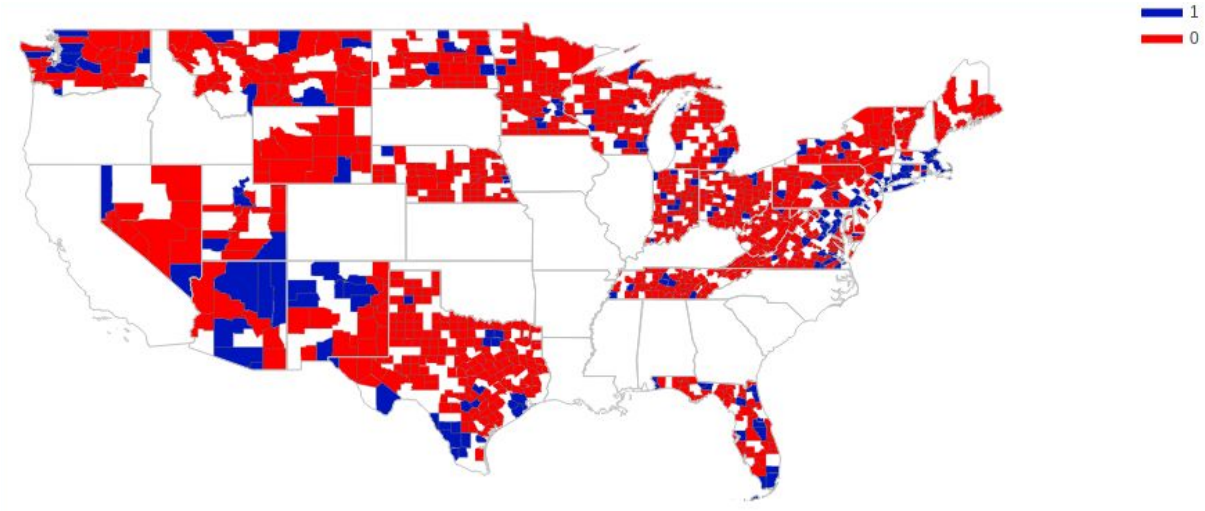
The best performing model was k means n_clusters=2, random_state=0 and random method of initialization. The performance of this model in terms of supervised metrics was .1975, that is the adjusted rand index score, unsupervised is .307 this is the mean Silhouette Coefficient. The highest overall supervised score was .1989 while the highest mean Silhouette Coefficient was .7233. Given that we have the necessary data to use supervised evaluation metrics the former score was seen as more important to evaluating the real world performance of the model though it must be stated that the performance of clustering was not as good as the classification methods like SVM. As far as selecting the parameters, the only sensible number of clusters for this purpose was 2 modification of other parameters such as the method of initialization did provide higher performance metrics but the contingency matrix was incorrect. As far as the variables including all demographic data provided the best model, considering unsupervised performance metrics however this would not be the case, the best contingency matrix and supervised performance metrics were achieved with all demographic data.

Task 6. Create a map of Democratic counties and Republican counties using the counties' FIPS codes and Python's Plotly library (plot.ly/python/county-choropleth/). Compare with the map of Democratic counties and Republican counties created in Project 01. What conclusions do you make from the plots?

Project 1



Project 2



The most notable difference is in the northeastern US (new england) the model predicted that the vast majority of those counties should be should be republican, when in reality they were democratic, as race seemed to be an effective predictor of voting patterns in other parts of the US it seems this is the cause of misclassification in this area, it is possible if we removed the variables that deal with race/ethnicity this region would have been predicted accurately although other regions in the US would be misclassified as a result as evident from the accuracy in testing different combinations of variables, the final conclusion to draw from this is that no model is perfect, as the misprediction of the northeast is mirrored in the northern

midwest in the same way and there does not seem to be a way with the data provided to correct this issue without reducing the overall performance of the model.

Task 7. Use your best performing regression and classification models to predict the number of votes cast for the Democratic party in each county, the number of votes cast for the Republican party in each county, and the party (Democratic or Republican) of each county for the test dataset (demographics_test.csv). Save the output in a single CSV file. For the expected format of the output, see sample_output.csv.

Step 1: To do the regression the test data was scaled according to the training set

Step 2: LASSO regression was used since that was the best performing one from the tests. The first model predicts the democratic votes using the training set on the scaled test set. The second model is the same as before but it predicts republican votes instead.

Step 3: After doing all the regressions upon the Democratic and Republican columns, the next step is to predict which party won in each state using the best classifier from the ones implemented. Since, SVM has the best metrics among the different classifiers, the SVM model built using the training set is taken and the demographics_test.csv file is fed into the model to test by selecting the predictors that was initially used in the model building process. It classifies each data row as either 0 or 1 relating to Republican or Democratic, respectively.

Step 4: All the found results are stored in a csv file named ‘output.csv’. This csv is created using the numpy.savetxt function, by providing all the resulting arrays.