PROJECT 02 REPORT

Sai Bharath Kumar Bandi, Sachin Dattatraya Hedge, Martin Bautista, Mohmed Hira

CS 418 – Introduction to Data Science

INTRODUCTION:

The objective of this project is to do Regression, Classification, and Clustering on the results for the 2018 United States Senate elections and demographic information for United States counties. We will use the given datasets (demographics_test.csv and merged_train.csv). We have also used Python 3 for the given tasks and used Jupyter Notebook for implementation.

TASKS AND ANSWERS:

1. (5 pts.) 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?

Code: x_train, x_valid, y_train, y_validate = train_test_split(merged.iloc[:,0:-1], merged['Party'], test_size = 0.25, random_state = 0)

A. We used the Holdout method.
   Partitioned the merged dataset into training set and validation set with the ratio 75:25 and considered the parameter random_state = 0

2. (5 pts.) Standardize the training set and the validation set.

```
# Standardize the training set and the validation set

std_train=x_train.select_dtypes(include=[np.int64,np.float64])
std_train=std_train.iloc[:,1:-2]
train_columns=std_train.columns
std_valid=x_valid.select_dtypes(include=[np.int64,np.float64])
std_valid=std_valid.iloc[:,1:-2]
scaler = StandardScaler()
scaler.fit(std_train)
x_train_scaled = scaler.transform(std_train)
x_validate_scaled = scaler.transform(std_valid)
x_train_scaled_df=pd.DataFrame(x_train_scaled,index=std_train.index,columns=train_columns)
x_validate_scaled_df=pd.DataFrame(x_validate_scaled,index=std_valid.index,columns=std_valid.columns)
```

A. The variables [‘State’, ‘County’, ‘FIPS’. ‘Democratic’, ‘Republican’, ‘Party’] are removed
   The x_valid and x_train is standardized.

3. (25 pts.) 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.

A. The regression models we considered are the Multi-linear regression model, LASSO regression model, Ridge regression model and the Elastic Net.

The best regression model for predicting number of votes for **Democratic** is:

LASSO regression with the variables - **'Total Population','Percent Black, not Hispanic or Latino','Percent Less than Bachelor's Degree'**
The performance of the model is:
R square - 0.9505081178945095
Adjusted R square - 0.9500048106188604
RMSE - 12456.252078729762

The best regression model for predicting number of votes for **Republican** is:

Multi-linear regression with the variables - **'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 65 and Older', 'Percent Unemployed', 'Median Household Income', 'Percent Rural'**
The performance of the model is:
R square - 0.7302080671530998
Adjusted R square - 0.7227655310745646
RMSE - 15749.245925443487

We have selected different combinations of variables:

1. All variables
2. ['Percent White, not Hispanic or Latino','Percent Black, not Hispanic or Latino','Percent Hispanic or Latino','Percent Less than Bachelor's Degree','Percent Age 29 and Under','Percent Age 65 and Older']
3. ['Total Population','Percent Black, not Hispanic or Latino','Percent Less than Bachelor's Degree','Percent Age 29 and Under','Percent Age 65 and Older']
4. ['Total Population','Percent Black, not Hispanic or Latino','Percent Less than Bachelor's Degree']

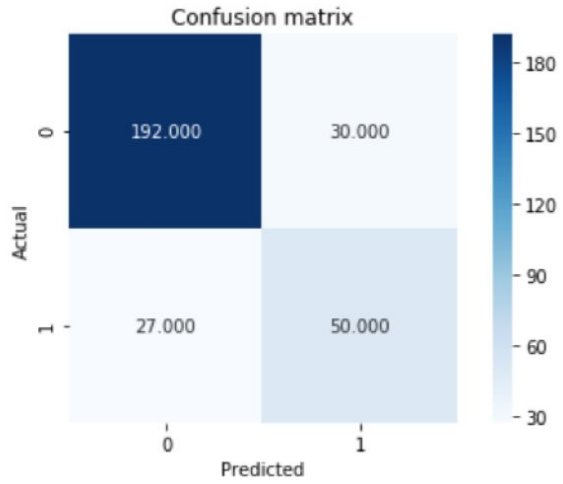
These variable combinations are chosen based on adjusted r-square value

4. (25 pts.) 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?

A. The two classification techniques we used were Decision Tree and Naive Bayes. The best of the two classification models proved to be Decision Trees as we got better values for the metrics. The performance of the model was (0.870, 0.636). The parameters we chose for said

model included The minimum number of samples required to be a leaf node at 2 because that improved the results for the f1-score. The variables that we chose were [‘Percent Hispanic or Latino’,’Percent Black, not Hispanic or Latino’,’Percent White, not Hispanic or Latino’,’Percent Less than Bachelor’s Degree’] and we chose these because we received a better f1-score with them.

Using the variables 'Percent Hispanic or Latino', 'Percent Black, not Hispanic or Latino', 'Percent White, not Hispanic or Latino', 'Percent Less than Bachelor's Degree'



This model (BEST) accurately classifies 242 observations of the for the party. The variable were chosen based on comparing the F1 score.

Different combinations of variables and parameters are chosen and mentioned below

Decision Tree:

1) Variables: All variables
    a) F1 score of: 0.836, and 0.544
    b) Parameters: Default
2) Variables: Percent white, Less than bachelors degree
    a) F1 score of: 0.855, and 0.589.
    b) Parameter: gini, min_sample_leaf = 2. (Rest Default)
3) Variables: Median Household income, Percent age >=29, Percent age 65=<, Percent White, Percent less than bachelors.
    a) F1 score: 0.862, 0.601.
    b) Parameters: gini, min_sample_leaf = 2, min_sample_split = 3. (Rest Default)
4) Variables: Percent White, Percent less than bachelors.
    a) F1 Score. 0.855, 0.589.
    b) Parameters: gini, min_sample_leaf = 2. (Rest Default)

Naive Bayes:

5) Variables: Percent age >=29, Percent age 65=<, Percent White, Percent less than bachelors
    a) F1 score of: 0.864, and 0.563

    b) Parameters: Default
6) Variables: Median Household income, Percent white, Less than bachelors degree
    a) F1 score of: 0.853, and 0.524.
    b) Parameter: Default
7) Variables:Total Population, Percent white, Less than bachelors degree.
    a) F1 score: 0.877, 0.571.
    b) Parameters: 2e-09, (Rest Default)
8) Variables: Percent White, Percent less than bachelors.
    a) F1 Score. 0.857, 0.539.
    b) Parameters: 3e-09 (Rest Default)

5. (25 pts.) 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?

A. The clustering models we considered are - K-means clustering and the DBSCAN

The best clustering model is the K-means clustering with the below parameters - (n_clusters = 2, init = 'k-means++', n_init = 10, random_state = 0)

The variables we considered are - ['Percent Black, not Hispanic or Latino', 'Percent Age 29 and Under', 'Percent Less than Bachelor's Degree', 'Percent Unemployed', 'Percent Female']

The performance of the model:
Adjusted Rand Index - 0.301132997830686
Silhouette Coefficient - 0.2596874751140196
This model is clustering 898 observations as one cluster of which 759 observations are clustered correctly and 297 observations as one cluster of which 186 observations are clustered correctly.

The different combinations of the parameters are:
K-means:
(n_clusters = 2, init = 'k-means++', n_init = 10, random_state = 0)
(n_clusters = 2, init = 'k-means++', n_init = 3, random_state = 0)
(n_clusters = 2, init = 'random', n_init = 10, random_state = 0)
(n_clusters = 2, init = 'random', n_init = 10, random_state = 0)

DBSCAN:
(eps = 2, min_samples = 5, metric = "euclidean")

(eps = 2, min_samples = 7, metric = "euclidean")
(eps = 1.8, min_samples = 5, metric = "euclidean")
(eps = 1.8, min_samples = 7, metric = "euclidean")

We have chosen different combinations of the variables for clustering:

1. All variables
2. 'Total Population', 'Percent Black, not Hispanic or Latino', 'Percent Age 65 and Older', 'Percent Less than Bachelor's Degree', 'Percent Unemployed', 'Median Household Income', 'Percent Rural'
3. 'Total Population', 'Percent White, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 29 and Under', 'Percent Age 65 and Older', 'Percent Less than Bachelor's Degree', 'Percent Unemployed', 'Median Household Income', 'Percent Rural'
4. 'Percent Black, not Hispanic or Latino', 'Percent Age 29 and Under', 'Percent Less than Bachelor's Degree', 'Percent Unemployed', 'Percent Female'

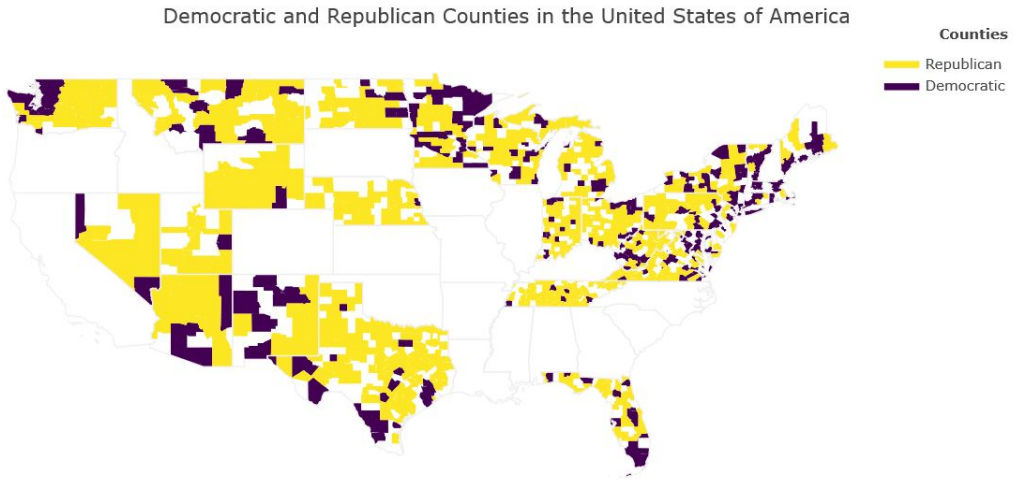
These variables combinations are chosen by comparing the adjusted rand index value

6. (10 pts.) 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?

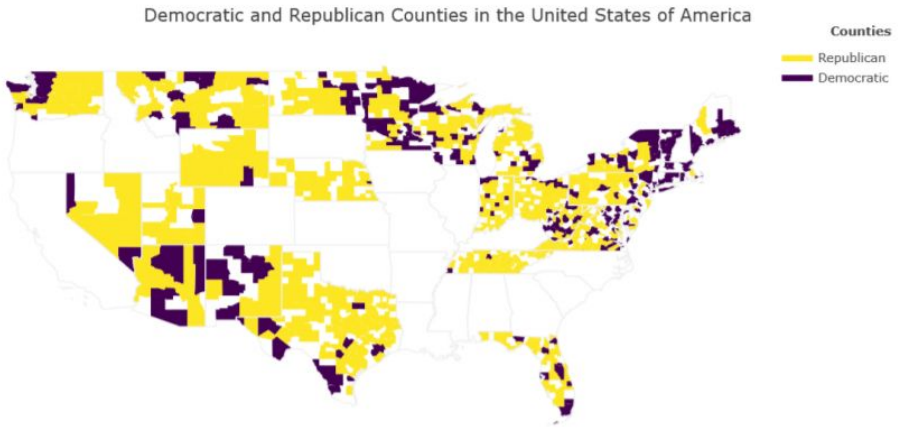
A. The best classification model is the Decision Tree with parameters - (criterion="entropy",random_state = 0,splitter='best',min_samples_leaf=2)

Fit the model with the standardized train dataset with the variables - ['Percent Hispanic or Latino','Percent Black, not Hispanic or Latino','Percent White, not Hispanic or Latino','Percent Less than Bachelor's Degree']

Classify the standardized validation dataset and standardized train dataset using the above classifier. Now the concat the two datasets and plot the map using the plotly library.



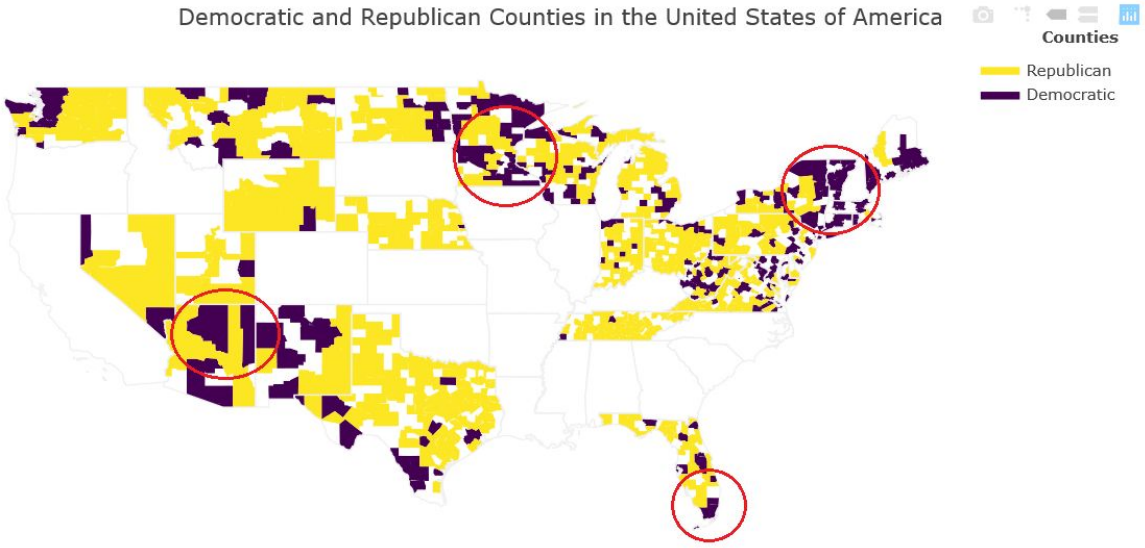
When we compare the resulting map with the project 01 map, which is:



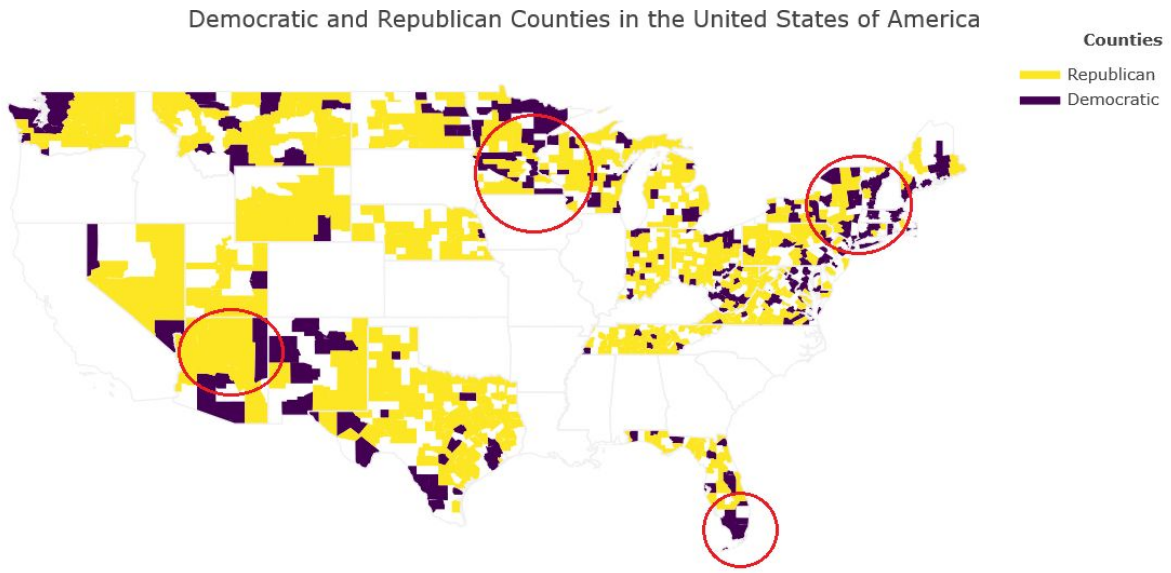
Conclusion:

- In both the maps, we have more Republican counties than Democratic counties
- Some counties in the highlighted area are classified differently:

Project 02 map:



Project 01 map:



7. (5 pts.) 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.

A. The best regression model is the LASSO regression for Democratic and multi-linear regression for Republic with the below variables:
Democratic: ['Total Population','Percent Black, not Hispanic or Latino','Percent Less than Bachelor's Degree']
Republican: ['Total Population', 'Percent White, not Hispanic or Latino', 'Percent Hispanic or Latino', 'Percent Foreign Born', 'Percent Age 65 and Older', 'Percent Unemployed', 'Median Household Income', 'Percent Rural']

The best classification model is Decision Tree with the parameters - (criterion="entropy",random_state = 0,splitter='best',min_samples_leaf=2) and the variables are - ['Percent Hispanic or Latino','Percent Black, not Hispanic or Latino','Percent White, not Hispanic or Latino','Percent Less than Bachelor's Degree']

- Load the demographic_test data and standardize the data.
- Predict and classify the data using the above models
- Export the data into **Task07_result.csv**