Kalyan Kumar Paladugula- Michael Ybarra - Zoheb Mohammed CS418 Data Science

Project 2 - Fall 2019

- Q1, How did you partition the dataset?
    - The data set was partition by using the Hold-out method. For training and testing the data, we used 80% of the data to train and the remaining 20% of the data to test.

- Q2,

```
# TASK 2
x_train.head()
```

	Total Population	Percent White, not Hispanic or Latino	Percent Black, not Hispanic or Latino	Percent Hispanic or Latino	Percent Foreign Born	Percent Female	Percent Age 29 and Under	Percent Age 65 and Older	Median Household Income	Percent Unemployed	Percent Less than High School Degree	Percent Less than Bachelor's Degree	Percent Rural
49	0.083065	0.727571	0.191791	0.122607	0.147765	0.935078	0.321157	0.835212	0.205124	0.530498	0.292139	0.806858	0.310283
314	0.002336	0.967382	0.000151	0.022349	0.023163	0.875524	0.226376	0.618222	0.340126	0.185821	0.102249	0.649536	1.000000
331	0.004520	0.952289	0.000893	0.025489	0.045743	0.875024	0.221253	0.707377	0.153223	0.609891	0.248889	0.798457	0.797938
1022	0.008226	0.928527	0.082687	0.013381	0.024674	0.881316	0.328605	0.532165	0.303621	0.346107	0.350446	0.809291	0.469030
808	0.001632	0.588984	0.081922	0.368115	0.040630	0.866553	0.400685	0.515187	0.356295	0.515541	0.232292	0.841751	1.000000

- Q3, What is the best performing linear regression model? What is the performance of the model? How did you select the variables of the model?

    - The Democratic model with 7 predictor variables : "Total Population", "Percent Black, not Hispanic or Latino", "Percent Less than Bachelor's Degree", "Percent Foreign Born", "Median Household Income", "Percent Unemployed", "Percent Less than High School Degree" is the best performing model.
      The Adjusted R2 value of this model is 0.870 and root mean square error for validation set is 502922217.

    - The Republican model with 10 predictor variables :
      "Total Population", "Percent White, not Hispanic or Latino", "Percent Less than Bachelor's Degree", "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 Rural" is the best performing model.
The Adjusted R2 value of this model is 0.855 and root mean squared error is 470853100.

- We picked the variables by reviewing project 1, then we checked the remaining variables. We started off with a single variable that gave us the highest F1 score and then continue to add more variables until we were able to get the highest F1 score. Some combinations gave us lower scores and were discarded but not deleted from the jupyter lab file.

- Q4, 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?
    - Currently the best performing data classification model we used for this project was the K-nearest neighbors classifier algorithm with k = 4 with all variables based on the following results:
        A. The Decision tree with parameters:
        "Total Population", "Percent Less than Bachelor's Degree", "Percent Black, not Hispanic or Latino", and "Percent Foreign Born","Percent Rural" is the best classifier in case of decision tree classifiers in terms of F1 Score [0.60162602 0.86197183]

        B. In case of K Nearest Neighbour classifier with k=3, the classifier with four parameters: "Total Population", "Percent White, not Hispanic or Latino", "Percent Black, not Hispanic or Latino", "Percent Hispanic or Latino" has the best performance in terms of F1 Scores [0.625 0.8852459]

        C. In case of k nearest neighbor classifier with k=4, the model with all parameters: has the best performance. [0.67213115 0.88764045]

        D. In case of Naive Bayes Classifier, the classifier with the variables: "Total Population", "Percent White, not Hispanic or Latino", "Percent Rural" has the best performance [0.57943925 0.8787062]

        E. In case of DBSCAN with all variables, we got only one cluster.

- K-nearest neighbors gave us the best performance in terms of an F1 scores [0.625 0.8852459]

- We used the same method as in finding the best linear regression model. Start by picking one parameter, one variable and stacking up additional variables and parameters until we were satisfied with the highest F1 that we could obtain with the various combinations.

- Q5, 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 clustering model that we tested was the K-Means clustering method based on the following results:

  A) K-means
  i) The K-means clustering model with 1 variable: "Total Population" has the best performance in terms of silhouette coefficient (better cluster quality i.e. high cohesion and separation) [0.903] (with adjusted_rand_index value of [0.120])

  ii) The K-means clustering model with 3 variables: "Total Population", "Percent Female", "Percent Age 29 and Under", "Percent Less than Bachelor's Degree" has the best performance in terms of adjusted_rand_index (true clusters) [0.289] (with silhouette coefficient (cluster quality i.e. cohesion and separation) value of [0.432])

  iii) The K-means clustering model with 2 variables: "Total Population", "Percent Female" has the best performance in terms of both adjusted_rand_index (true clusters) and silhouette_coefficient (cluster quality i.e. cohesion and separation) [0.122, 0.752]

  B) Single Linkage clustering
  i) the clustering model with 1 variable: "Total Population" has the best performance in terms of silhouette coefficient (better cluster quality i.e. high cohesion and separation) [0.953] (with adjusted_rand_index value of [0.0056])

  ii) the clustering model with 1 variable: "Percent White, not Hispanic or Latino" has the best performance [0.06341778206546299, 0.6589687284593577] in terms

of adjusted_rand_index (true clusters) [0.063] (with silhouette coefficient (cluster quality i.e. cohesion and separation) value of [0.659])

iii) the clustering model with 1 variable: "Percent White, not Hispanic or Latino" also has the best performance in terms of both adjusted_rand_index (true clusters) and silhouette_coefficient (cluster quality i.e. cohesion and separation) [0.063, 0.659]

C) Complete Linkage clustering

i) the clustering model with 1 variable: "Total Population" has the best performance in terms of silhouette coefficient (better cluster quality i.e. high cohesion and separation) [0.953] (with adjusted_rand_index value of [0.0056])

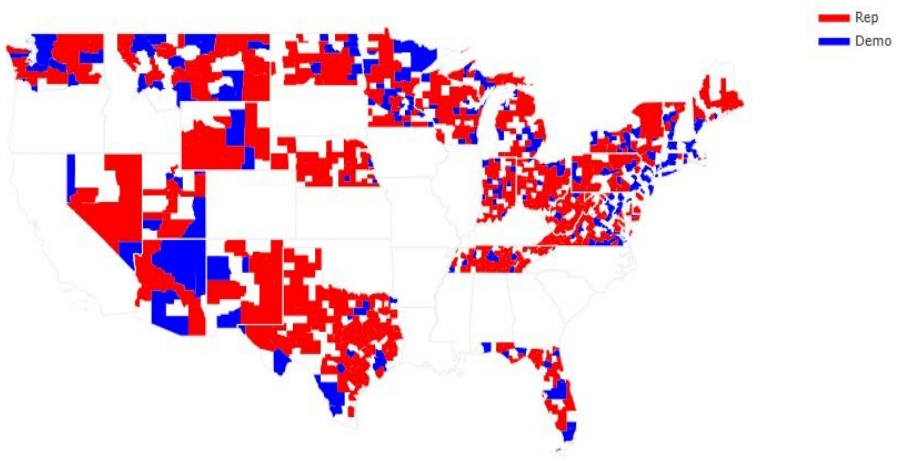
ii) the clustering model with 2 variables: "Percent White, not Hispanic or Latino", "Percent Less than High School Degree" has the best performance in terms of adjusted_rand_index (true clusters) [0.144] (with silhouette coefficient (cluster quality i.e. cohesion and separation) value of [0.588])

iii) the clustering model with 1 variable: "Percent White, not Hispanic or Latino" has the best performance in terms of both adjusted_rand_index (true clusters) and silhouette_coefficient (cluster quality i.e. cohesion and separation) [0.128, 0.669]

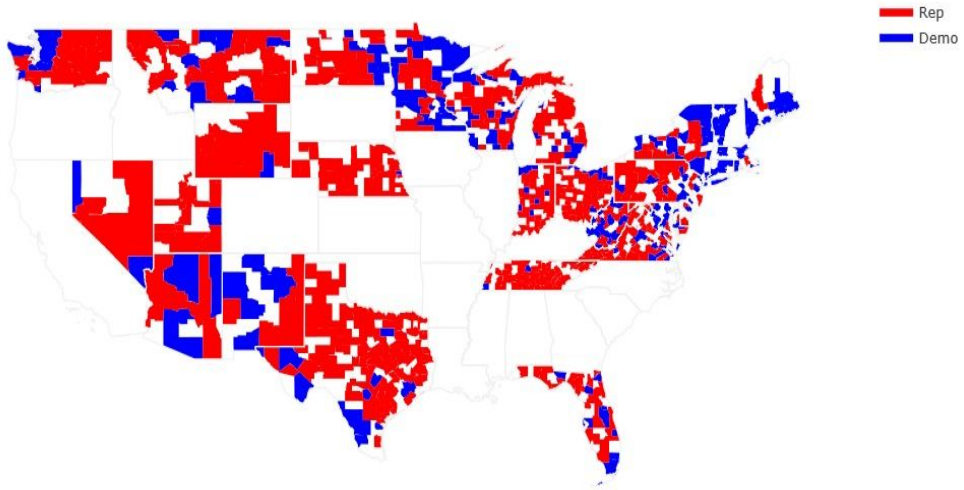
- We continued to use the same technique as we use in linear regression model by picking one to start with then increasing the amount of variables until we had a high enough number that would satisfy our model.

- Q6, 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



- From the above two plots, the best classifier we used to predict the counties had wrongly predicted some counties as Republican and some as Democratic.
- The accuracy obtained is 85.7 percent and F1 Score is [0.72900158 0.90278567]

- Q7,

```
out.head()
```

	State	County	Democratic	Republican	Party
0	NV	eureka	0.0	8888.0	0.0
1	TX	zavala	0.0	4464.0	0.0
2	VA	king george	20148.0	23612.0	0.0
3	OH	hamilton	453771.0	276454.0	1.0
4	TX	austin	11205.0	9091.0	1.0