Project 02: Regression, Classification, and Clustering

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?

First, we separate the dataset to two parts. Y that contains dependent variables 'Republican', 'Democratic', 'Party', 'Total Population'' and 'FIPS' that is for the mapping task. data_x that contains remaining independent variables. Then we use holdout method to partition the Y and data_x to 20% test set and 80% training set randomly.

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

Use StandardScaler trained by x_train on x_train, x_test and data_x

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?

result:

R_squared: 0.9384, model built by "Total Population"

R_squared: 0.9378, model built by "Total Population", "Percent White", "Percent Female"

R_squared: 0.9353, model built by "Total Population", "Median Household Income", "Percent Unemployed"

R_squared: 0.9510, model built by LASSO (alpha = 3000) based on "Total Population", "Percent Foreign Born", "Percent Less than bachelor's degree.

LASSO(alpha = 3000) has the best performance, 95.1% of the variability in the number of votes cast for the Democratic party can be explained by "Total Population", "Percent Foreign Born", "Percent Less than bachelor's degree"

• Repeat this task for the number of votes cast for the Republican party in each county.

result:

adjusted_R_squared: 0.6295, model built by "Total Population"

adjusted_R_squared: 0.6386, model built by "Total Population", "Percent White", "Percent Female"

adjusted_R_squared: 0.6697, model built by LASSO (alpha = 1000) based on 9 features

adjusted_R_squared: 0.6960, model built by LASSO (alpha = 100) based on all features

adjusted_R_squared:0.6455, model built by "Total Population", "Median Household Income", "Percent Unemployed"

LASSO (alpha = 100) has the best performance, 69.6% of the variability in the number of votes cast for the Republican party can be explained by all features.

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?

result

accuracy: 0.6652, F1_score: [0.771 0.375] use decision tree with "Total Population" to predict Party

accuracy: 0.6861, F1_score: [0.791 0.369] use decision tree with "Total Population", "Median Household Income" to predict Party

accuracy: 0.7071, F1_score: [0.804 0.416] use decision tree with "Total Population", "Median Household Income", "Percent White" to predict Party

accuracy: 0.7698, F1_score: [0.844 0.560] use decision tree with all features to predict Party

accuracy: 0.7907, F1_score: [0.869 0.468] use SVC with "Total Population" to predict Party

accuracy: 0.7656, F1_score: [0.858 0.317] use SVC with "Total Population", "Median Household Income" to predict Party

accuracy: 0.8117, F1_score: [0.883 0.516] use SVC with "Total Population", "Median Household Income","Percent White" to predict Party

accuracy: 0.8493, F1_score: [0.904 0.640] use SVC with all features to predict Party

accuracy: 0.8075, F1_score: [0.877 0.557] use KNeighbors k=3 with all features to predict Party

accuracy: 0.8284, F1_score: [0.893 0.559] use KNeighbors k=6 with all features to predict Party

model built by SVC with all features has the best performance, use radial basis function kernel as parameter.

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?

result

adjusted_rand_index and silhouette_coefficient: [0.0634, 0.6589] single linkage, one variables

adjusted_rand_index and silhouette_coefficient: [0.1276, 0.6688] complete linkage, one variables

adjusted_rand_index and silhouette_coefficient: [0.0028, 0.6910] single linkag, e two variables

adjusted_rand_index and silhouette_coefficient: [0.0835, 0.5250] complete linkage, two variables

adjusted_rand_index and silhouette_coefficient: [0.0028, 0.6408] single linkag, e three variables

adjusted_rand_index and silhouette_coefficient: [0.0137, 0.6121] complete linkage, three variables

adjusted_rand_index and silhouette_coefficient: [0.0453, 0.5521] kmeans method iteration = 10, three variables

adjusted_rand_index and silhouette_coefficient: [0.0453, 0.5521] kmeans method iteration = 100, three variables

adjusted_rand_index and silhouette_coefficient: [0.1087, 0.7053] kmeans method iteration = 10, all variables

adjusted_rand_index and silhouette_coefficient: [0.1295, 0.5525] DBSCAN, three features,eps=0.7, min_samples=10

adjusted_rand_index and silhouette_coefficient: [0.0574, 0.4809] DBSCAN, all features, eps=2.9, min_samples=10

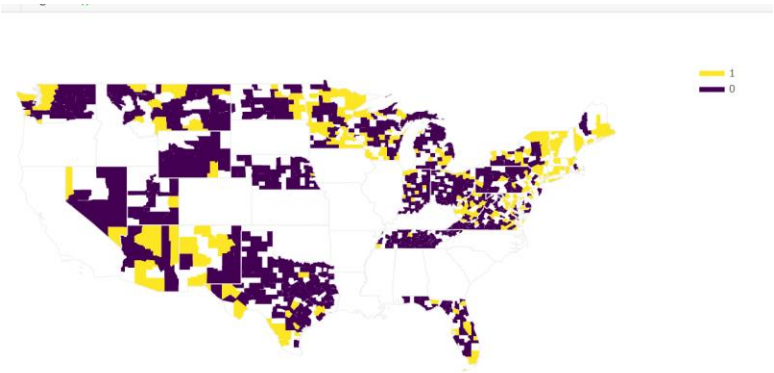
Use DBSCAN(eps=0.7, min_samples=10) with "Total Population", "Median Household Income", "percent of white" features to cluster party has the best performance.

The adjusted_rand_index is about 0.13,which means that model is not that good for clustering 'party' based on these three features, also means if use this model, different party cannot be clearly separated based on these three features. silhouette coefficient is about 0.5525, different clusters' centroids separate in a distance, from this aspect, the clustering is good.

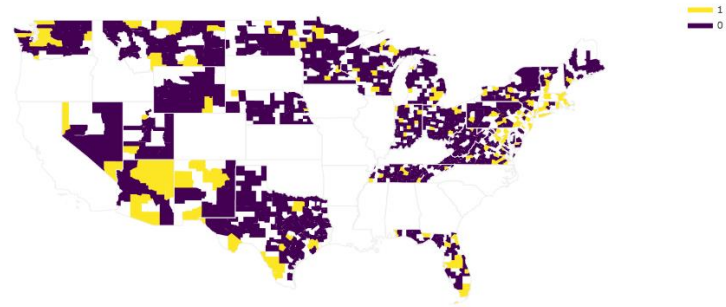
The true cluster's silhouette coefficient is 0.2189 based on these three features, the true cluster's silhouette coefficient is 0.1506 based on all features

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?

Project 1



Project 2



This map is similar to the previous one, the reason why some counties are classified with different color in this project we analyze are these they are belong to test sample that were not be trained by model so there exist some error, or these counties are located around the classification boundary, the model in order to avoid overfit, just mislabel these counties.

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.