Project 02: Regression, Classification, and Clustering Report

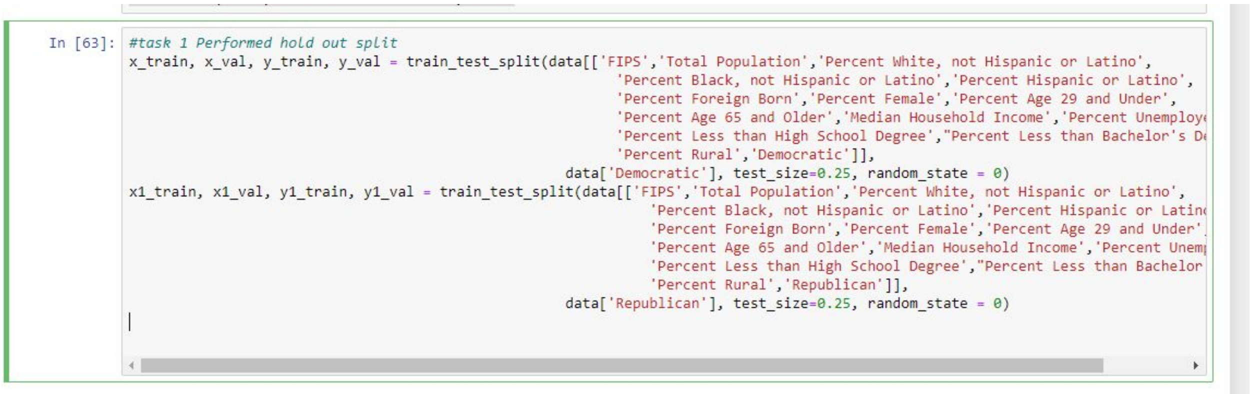
Prepared By: Kevin Patel, Prithvi Bhatt and Preethi Elango

CS 418 at the University of illinois at chicago, Fall 2019.

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?

We used holdout method to partition our dataset into a training set and a validation set.

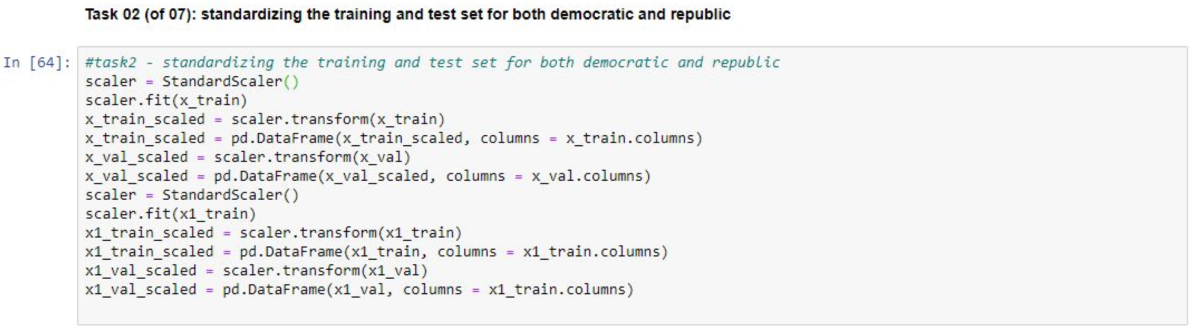
Training set contains 75% of data and validation set contains 25% of the data



```
In [63]: #task 1 Performed hold out split
         x_train, x_val, y_train, y_val = train_test_split(data[['FIPS','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 Unemploy
                                                                 'Percent Less than High School Degree',"Percent Less than Bachelor's De
                                                                 'Percent Rural','Democratic']],
                                                          data['Democratic'], test_size=0.25, random_state = 0)
         x1_train, x1_val, y1_train, y1_val = train_test_split(data[['FIPS','Total Population','Percent White, not Hispanic or Latino',
                                                                     'Percent Black, not Hispanic or Latino','Percent Hispanic or Latin
                                                                     'Percent Foreign Born','Percent Female','Percent Age 29 and Under'
                                                                     'Percent Age 65 and Older','Median Household Income','Percent Unem
                                                                     'Percent Less than High School Degree',"Percent Less than Bachelor
                                                                     'Percent Rural','Republican']],
                                                          data['Republican'], test_size=0.25, random_state = 0)
```

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

We have used standard scalar function for standardizing the training and test set for both democratic and republic.



Task 02 (of 07): standardizing the training and test set for both democratic and republic

In [64]:
```
#task2 - standardizing the training and test set for both democratic and republic
scaler = StandardScaler()
scaler.fit(x_train)
x_train_scaled = scaler.transform(x_train)
x_train_scaled = pd.DataFrame(x_train_scaled, columns = x_train.columns)
x_val_scaled = scaler.transform(x_val)
x_val_scaled = pd.DataFrame(x_val_scaled, columns = x_val.columns)
scaler = StandardScaler()
scaler.fit(x1_train)
x1_train_scaled = scaler.transform(x1_train)
x1_train_scaled = pd.DataFrame(x1_train, columns = x1_train.columns)
x1_val_scaled = scaler.transform(x1_val)
x1_val_scaled = pd.DataFrame(x1_val, columns = x1_train.columns)
```

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.

We built 2 regression models for each party one with single parameter and multiple parameters. We calculated both lasso and ridge regression for both the parameter combinations.

Model 1 variables: Total Population

Model 2 variables: 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,'Median Household Income

Variables are selected based on the positive and negative correlation between them.

For democratic, we got below mentioned values.

Model 1 using one variable:

One variable r2=0.9436415220931658 adj. r2=0.9713158723670933

lasso-one variable r2=0.9436415220931673 adj.r2= 0.971315872367094

Ridge one variable r2=0.9436415220931669 adj.r2= 0.9713158723670938

Model 2 using Multiple variable:

multiple variable = 0.9278780606334682 adj.r2= 0.9622508741653365

lasso multiple variable = 0.9278901279994416 adj.r2= 0.9622573107263362

Ridge multiple variable =0.9277408329226396 adj.r2= 0.9621776759159704

Out of these values, Ridge regression model(using multiple variable) which has higher adjusted R-squared value.Thus, we take it as the best model to predict Total Votes for democratic party.

For predicting Total votes for Republic , we got below mentioned values.

Model 1 using one variable:

One variable r2=0.9436415220931658 adj. r2=0.9713158723670933

lasso-one variable r2=0.9436415220931962 adj.r2= 0.951315872367109

ridge-one variable r2=0.9436415220931962 adj.r2= 0.951315872367109

Model 2 using Multiple variable:

multiple variable = 0.9278780606334682 adj.r2= 0.9622508741653365

lasso multiple variable = 0.9278901279994416 adj.r2= 0.9622573107263362

Ridge multiple variable =0.9277408329226396 adj.r2= 0.9621776759159704

comparing all the models and their r2 and adj r2 values, lasso is best for multiple variable

Out of these values, Lasso regression model(using multiple variable) which has higher adjusted R-squared value. Thus, we take it as the best model to predict Total Votes for republic party.

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?

We have built a 3 different model using 2 combinations of variables.

Variable combination 1: 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

Model 1: KNeighbors = [0.87103594, 0.512   ] **Parameter used:** n_neighbors=5, So that it would give robust results

Model 2: GaussianNB=[0.86474501, 0.58503401]

Model 3: SVC=[0.90947368, 0.6504065] **Parameter used:** kernel =rbf, we use this function when the data is not linearly separable

Variable combination 2: Total Population,Percent White, not Hispanic or Latino,Percent Unemployed,Percent Less than High School Degree,Percent Less than Bachelor's Degree,Percent Rural

Model 1: KNeighbors = [0.85953878, 0.44628099]**Parameter used:** n_neighbors=5, So that it would give robust results

Model 2: GaussianNB= [0.87741935, 0.57142857]

Model 3: SVC= [0.88842105, 0.56910569] **Parameter used:** kernel -=rbf'

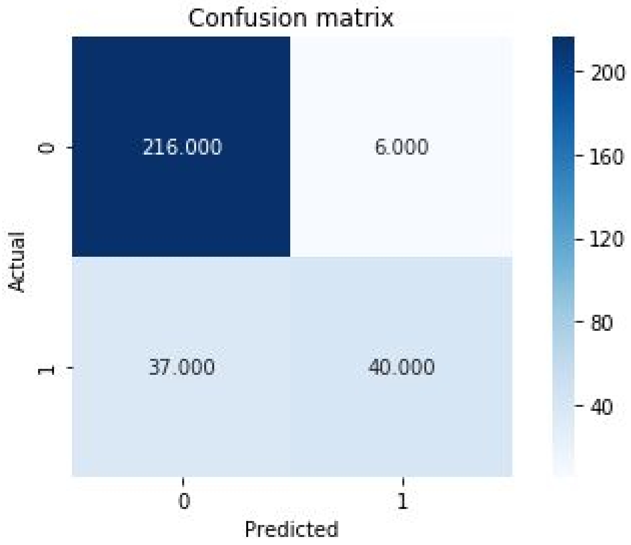
Comparing all the model's F1 score, Best performing model is SVC with variable combination 1 and has F1 score of SVC=[0.90947368, 0.6504065] .

Performance of model

Accuracy_score= 0.8561872909698997

Precision_score = ([0.85375494, 0.86956522])

recall_score= [0.97297297, 0.51948052])



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?

We have built 3 models using 3 combinations of variables.

Variable 1: Total Population

Model 1: single linkage method:

rand_index: 0.005608925119335567 silhouette_coeff = 0.9531008389502824

Model 2: complete linkage method

rand_index: 0.005608925119335567 silhouette_coeff = 0.9531008389502824

Model 3: KMeans

Rand_index: 0.11979747814620154   silhouette_coeff = 0.902954408093495

Variable combination 2: Total Population,Percent White, not Hispanic or Latino, Median Household Income ,Percent Unemployed, Percent Less than High School Degree, Percent Less than Bachelor's Degree,Percent Rural

Model 1: single linkage method

Rand_index: 0.005608925119335567   silhouette_coeff = 0.9523736778861548

Model 2: complete linkage method

Rand_index: 0.005608925119335567   silhouette_coeff = 0.9523736778861548

Model 3: KMeans

Rand_index: 0.11979747814620154 silhouette_coeff = 0.902954408093495

Variable combination 3: 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

Model 1: single linkage method

Rand_index: 0.005608925119335567   silhouette_coeff = 0.9523736738921391

Model 2: complete linkage method

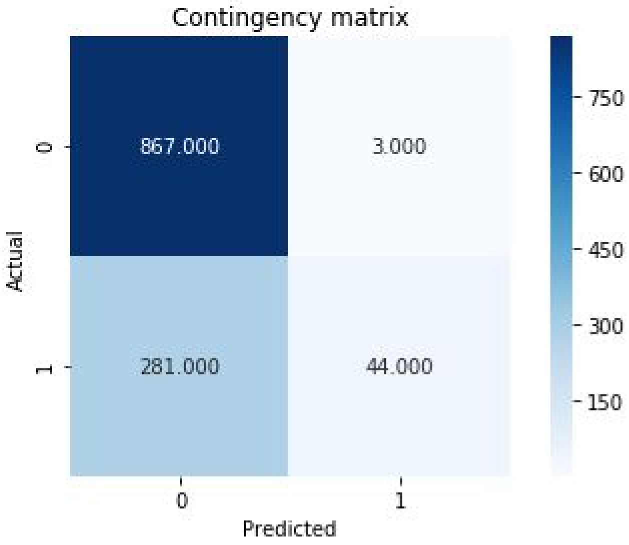
Rand_index: 0.005608925119335567   silhouette_coeff = 0.9523736738921391

Model 3: KMeans

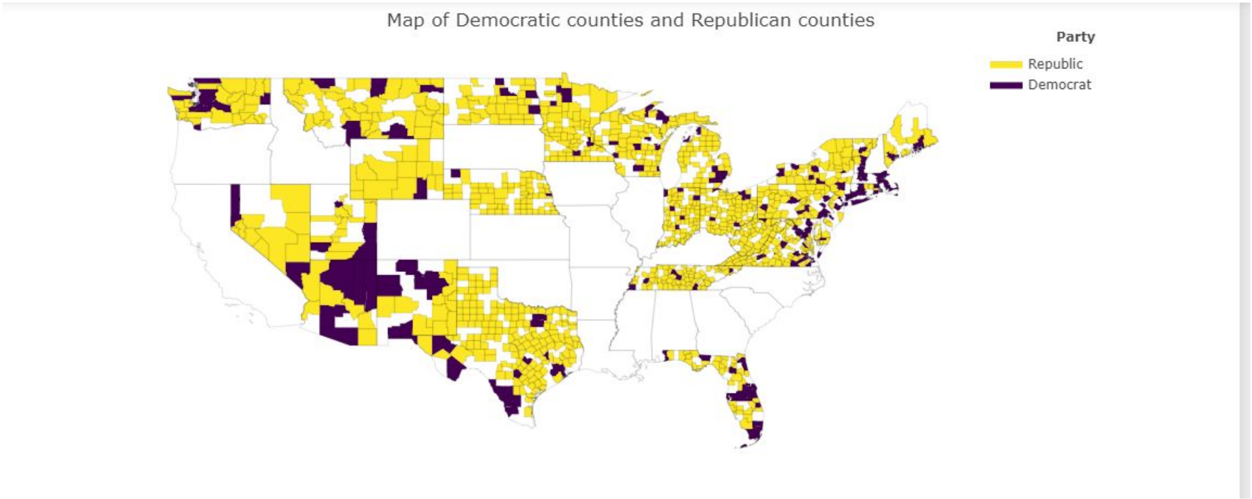
Rand_index: 0.11979747814620154   silhouette_coeff =0.902954408093495

Comparing all the models, we could see kmeans hs higher rand_index and better silhouette_coeff.

Based on the contingency matrix we could see that it classifies democratic data perfectly around 93% of the time.



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/). Compared with the map of Democratic counties and Republican counties created in Project 01. What conclusions do you make from the plots?



We could see minor changes in the classification of counties. Few Democratic counties got changed to Republic and vice versa.

This shows that we see some miss classification compared to Project 1 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.

We have used ridge model to predict Total votes for Democratic party and lasso model to predict Total votes for Republican party

We used SVC model to predict the party classifiers.