Final Project Report

Group Info: Kalyan Kumar Paladugula (kpalad4, 679025059)
Michael Ybarra (mybarr3, 659036727)
Zoheb Mohammed (zmoham2, 654090066)

1. Problem Selection. Identify a real-world problem (for example, predicting the number of votes cast for the Democratic and the Republican parties in each county during the 2018 United States Senate elections) and propose a data science solution (for example, building linear regression models). Describe your problem and your solution.

   The objective of the project is to predict the unemployment rates of the counties for the year 2017 based on the unemployment rates of counties in the year 2015 using regression, and to predict the classes of counties in 2017 dataset using classifier built on the 2015 dataset.

   Solution Flow Chart:
   1. First, we will process the datasets to identify any missing values and to remove any attributes that cause multicollinearity.
   2. We will explore the data and remove any irrelevant or redundant variables.
   3. We will partition the dataset as training dataset and the test dataset using hold-out method with 80% percent of data as training and 20% as test data.
   4. We will standardize the dataset
   5. We will build linear regression models on the 2015 data set using multiple combinations of variables with unemployment rate as the response variable. Pick the best model based on the adjusted R2 values and the root mean square error on the test dataset.
   6. We use that model to predict the unemployment rate of the counties for the year 2017. And we will calculate the mean squared error from the actual unemployment rates.
   7. Also, we want to classify the counties into 3 classes in terms of unemployment:
      A. 12% and over (Worst)
      B. 8% to 12% (Moderate)
      C. 8% and below (Best)
   8. FInally, we will build a classifier using the 2015 dataset and will predict the classes of the counties in terms of unemployment for the year 2015 and 2017 as well.

2. Data Collection. Identify one or more datasets relevant to your problem. Describe your datasets.

   - We got the demographic datasets for the year 2015 and 2017 from kaggle:
     There are 37 columns and 3220 observations in each dataset:

CensusId	3220 non-null int64
State	3220 non-null object
County	3220 non-null object
TotalPop	3220 non-null int64

Men	3220 non-null	int64
Women	3220 non-null	int64
Hispanic	3220 non-null	float64
White	3220 non-null	float64
Black	3220 non-null	float64
Native	3220 non-null	float64
Asian	3220 non-null	float64
Pacific	3220 non-null	float64
Citizen	3220 non-null	int64
Income	3219 non-null	float64
IncomeErr	3219 non-null	float64
IncomePerCap	3220 non-null	int64
IncomePerCapErr	3220 non-null	int64
Poverty	3220 non-null	float64
ChildPoverty	3219 non-null	float64
Professional	3220 non-null	float64
Service	3220 non-null	float64
Office	3220 non-null	float64
Construction	3220 non-null	float64
Production	3220 non-null	float64
Drive	3220 non-null	float64
Carpool	3220 non-null	float64
Transit	3220 non-null	float64
Walk	3220 non-null	float64
OtherTransp	3220 non-null	float64
WorkAtHome	3220 non-null	float64
MeanCommute	3220 non-null	float64
Employed	3220 non-null	int64
PrivateWork	3220 non-null	float64
PublicWork	3220 non-null	float64
SelfEmployed	3220 non-null	float64
FamilyWork	3220 non-null	float64
Unemployment	3220 non-null	float64

And for the unemployment rates of counties for the year 2013, we used American factfinder website:

US unemployment dataset for the year 2013 from the American fact finder website.

3. Data Preparation. Detect and correct data quality problems (missing data, noise, outliers, etc.) and transform the data into an appropriate format for data analysis. Describe your data preparation process and report the results obtained.

First, we searched for missing values and found few in the Income and ChildPoverty columns. We replaced them with the corresponding median values of their columns.

Then, we identified that the attributes “Women” and “OtherTransp” would cause multicollinearity. So, we removed them using drop method.

And we removed the attributes: “IncomeEr”' and “IncomePerCapErr”, which are not necessary to build any model.

4. Data Exploration. Explore the data using summary statistics and plots and identify the most important variables for data analysis. Describe your data exploration process and report the results obtained.

For Regression:

Using Scatterplots and R2 values, we found some correlated attributes and attributes that are significantly affecting the unemployment rate of counties. We kept one of the correlated attributes that has higher correlation with the response variable: Unemployment Rate.

For Classification:

Using box plots and summary characteristics, we identified attributes that are important to distinguish observations from different classes.

5. Data Modeling. Train and test models using the data. Your data modeling step must include at least two of the following tasks: (1) Regression, (2) Classification, (3) Clustering, and (4) Text analysis. Consider multiple techniques, parameters, and variables. Describe your data modeling process and report the results obtained.

    1. Regression:
    Multiple linear regression:

    First, we built models using the attributes obtained from the data preparation step and then we checked the rest of variables.

    We started building the regression model with the attribute “Poverty” that has the highest correlation with the response variable. Then, we added the second attribute. If the addition of a new variable increases the adjusted R2 and the root mean squared error on the validation set, we included it in the model, else we replaced it with a new attribute.

    We got the best regression model with the predictors:

"Poverty", 'White', "Black", "Service", "WorkAtHome", "Hispanic", 'Mean Commute Time', "Percent Population with age between 16 and 44", 'Asian','Office', 'Construction','Drive', 'Transit', "State Unemployment rate for 2013", "County Unemployment Rate for 2013"
with adj R2 value of 0.839, , R2 value of 0.840, and Root mean square error of 3.04

2. Classification:

We built the classifiers in the same way we built regression models. But here we checked the F1 Scores to pick the best classifier.
We used decision tree, Naive Bayes and K-Nearest neighbour classifiers:
Here are the results:

Decision tree with 1 parameter: "County Unemployment Rate for 2013",
is the best one with F1 Scores:
[0.72093023 0.85846154 0.70815451]

In case of Naive Bayes, the classifier with parameter: "County Unemployment Rate for 2013"
is the best one with F1 Scores:
[0.7625 0.86728395 0.74166667]

In case of K_Nearest Neighbour with k=3, the classifier with parameters: "TotalPop", "Men", "White", "Median Household Income",'Citizen Voting Age Pop', 'County Unemployment Rate for 2013' is the best one with F1 Scores:
[0.75294118 0.85714286 0.70403587]

In case of K_Nearest Neighbour with k=4, the classifier with parameters: "TotalPop","Poverty", "Men", "Black" ,"Citizen Voting Age Pop", "County Unemployment Rate for 2013", 'Percent Population with age between 16 and 44', 'Hispanic', 'Native' is the best one with F1 Scores:
[0.72820513 0.8729927 0.68137255]

Naive Bayes with parameter: "County Unemployment Rate for 2013" has the best F1 scores.

Note: As there is a high correlation between County unemployment rate for 2013 and the unemployment rate, it makes sense why we got the best classifier with the attribute "County Unemployment Rate for 2013" .

6. Results:

Best Models:
Regression:
Predictors:

"Poverty", 'White', "Black", "Service", "WorkAtHome", "Hispanic", 'Mean Commute Time', "Percent Population with age between 16 and 44", 'Asian','Office', 'Construction','Drive', 'Transit', "State Unemployment rate for 2013", "County Unemployment Rate for 2013"
R2 value = 0.840
Adj R2 value = 0.839
Root mean square error = 3.04

Classification:
Naive Bayes with parameter: "County Unemployment Rate for 2013" has the best F1 scores. [0.7625 0.86728395 0.74166667]

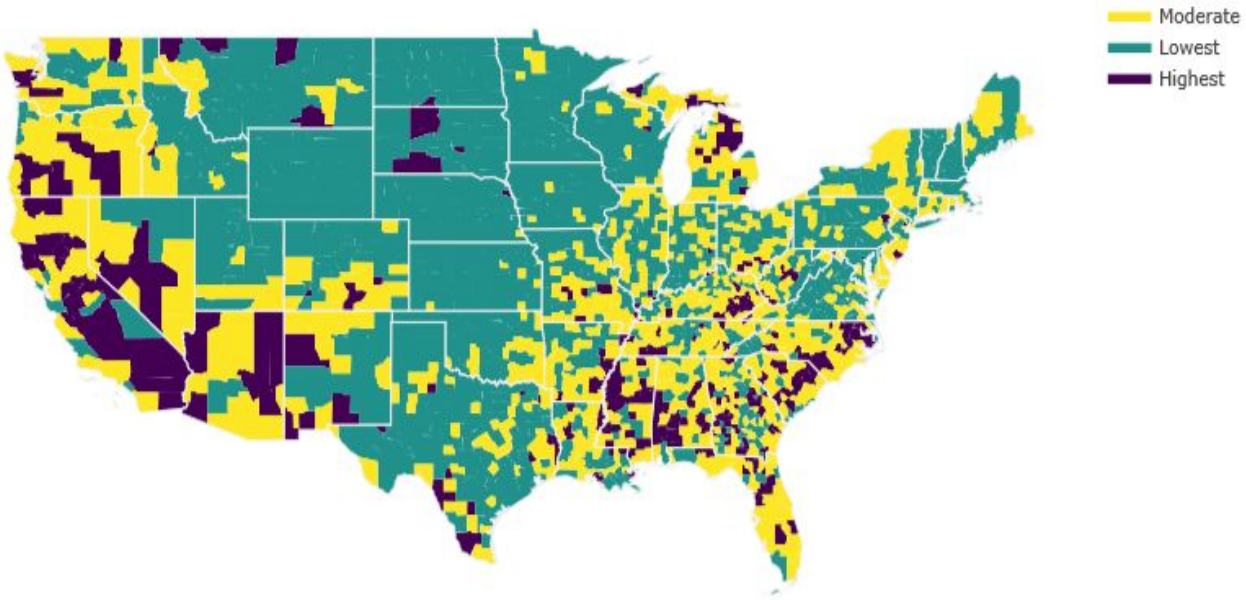
Predictions:
Regression:
We predicted the unemployment rates of counties for the year 2017 with Root mean square error of 2.26 and mean absolute error of 1.003.

Classification:

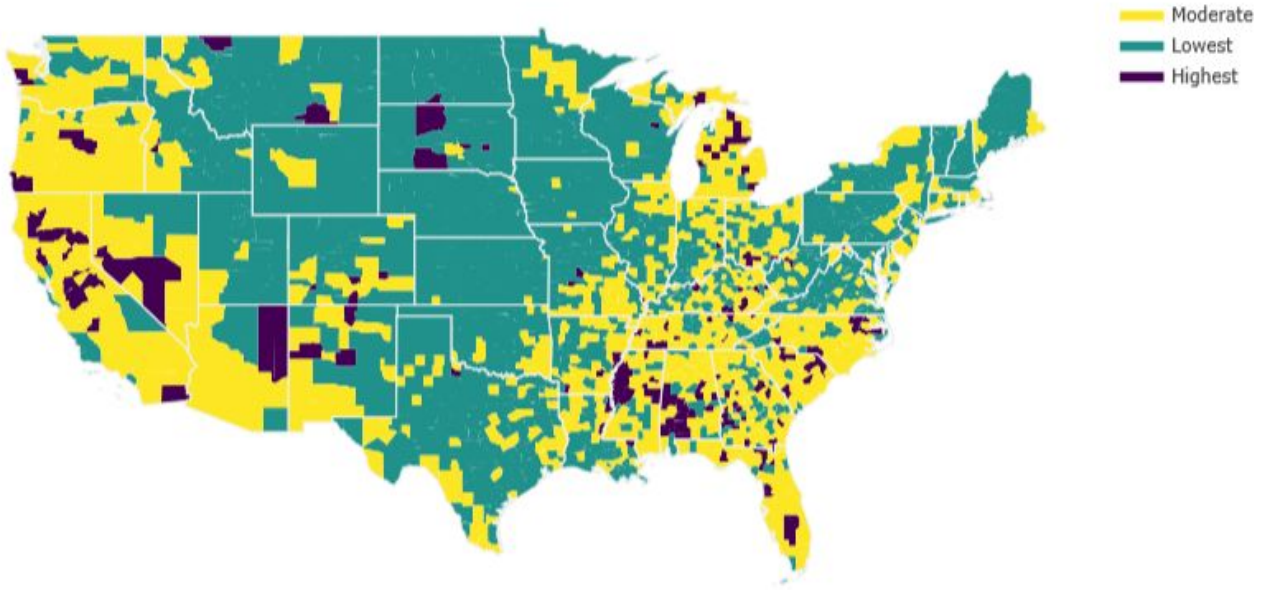
We predicted classes of observations with the accuracy of 0.81 for the year 2015 and 0.69 for the year 2017.

Plots:

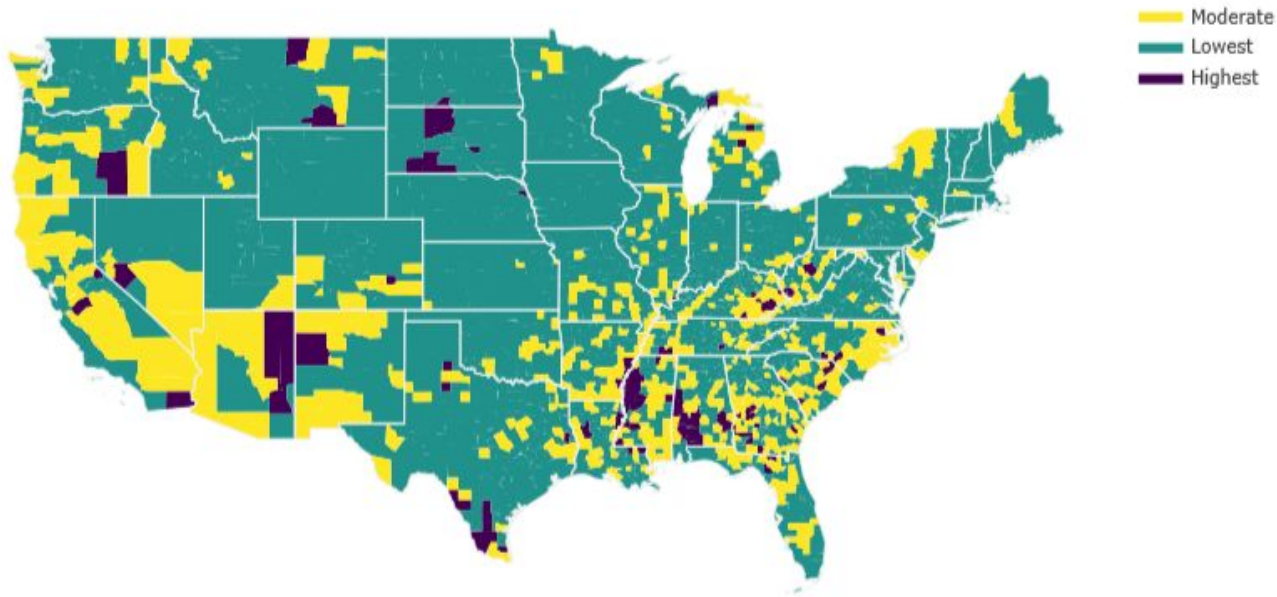
Actual Classes of counties for the year 2015



Predicted Classes of counties for the year 2015



Actual Classes of counties for the year 2017



Predicted Classes of counties for the year 2017