Prediction of Unemployment Rates

Kalyan Kumar Paladugula
Michael Ybarra
Zoheb Mohammed

Problem Statement

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.

Data Sources

US Census Demographic dataset for the year 2015 from Kaggle.com

US Census Demographic dataset for the year 2017 from Kaggle.com

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

US unemployment dataset for the year 2016 from the American fact finder website

Brief description of solution

Process

1. First, process the datasets to identify any missing values and to remove any attributes that cause multicollinearity.
2. Explore the data and remove any irrelevant or redundant variables.
3. 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. Standardize the dataset
5. Build linear regression models on the 2015 dataset 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. Use that model to predict the unemployment rate of the counties for the year 2017, and calculate the mean squared error and mean absolute error between the actual and predicted unemployment rates.
7. 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. Build a classifier using the 2015 dataset and predict the classes of the counties of the year 2015 and 2017.

Results OverView

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"
F1 scores = [0.7625, 0.86728395, 0.74166667]

Predictions:

Regression:

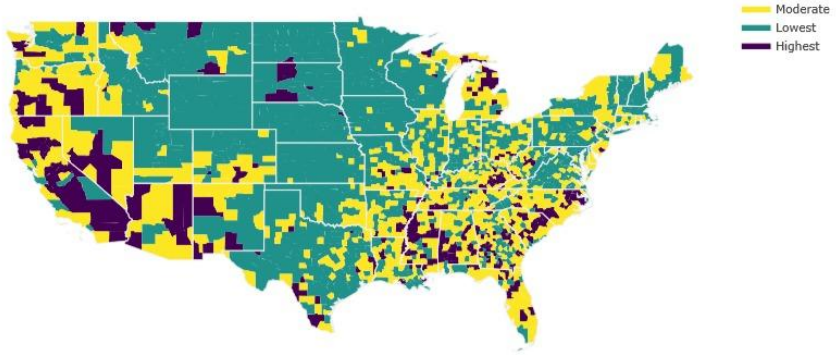
We predicted the unemployment rates of counties for the year 2017 with Root mean square error of 2.26 and mean absolute error of 1.003.

Classification:

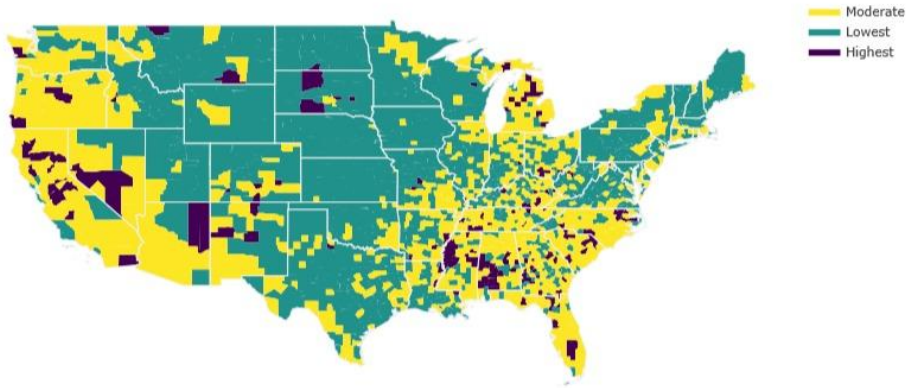
We predicted the classes of counties with the accuracy of 0.81 for the year 2015 and 0.69 for the year 2017.

Results 2015

Actual 2015

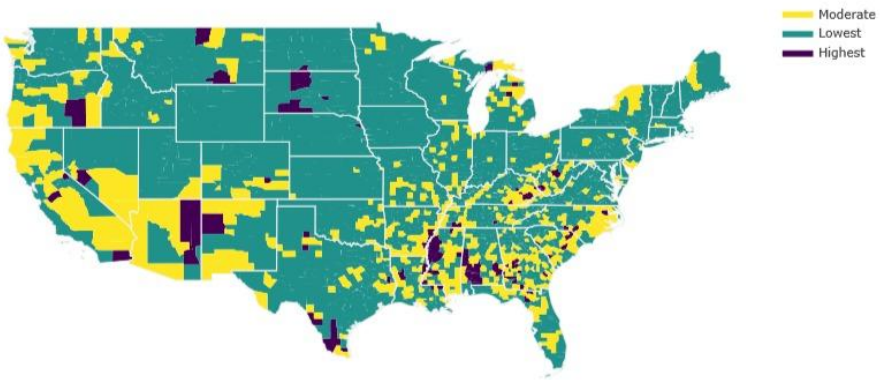


Predicted 2015

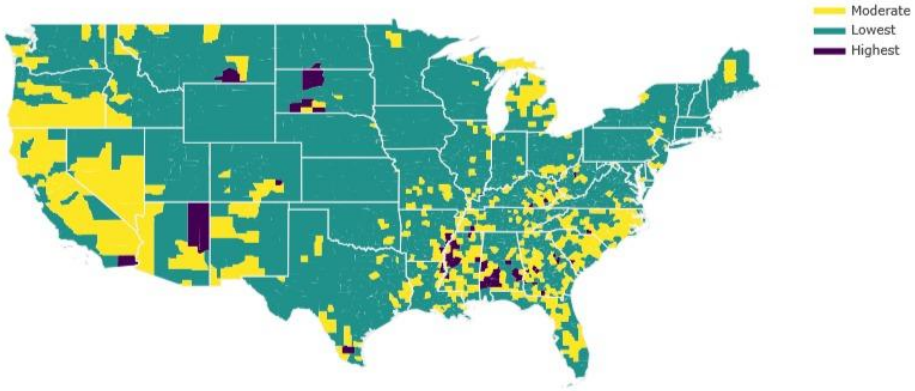


Results 2017

Actual 2017



Predicted 2017



Conclusion

For the year 2015, our classifier predicted some of the classes that were actually highest as moderate and some of the lowest as moderate incorrectly. Our model classifier is biased towards the moderate class.

As the classifier biased towards moderate, it also predicted some of lowest and highest classes as moderate for the year 2017 and it predicted classes of counties with less accuracy than that for the year 2015.

There was a significant decrease in unemployment rates of counties in 2016. There maybe other factor to cause this that we have not discovered.

So, we believe that if we used the 2016 dataset to build models we would have gotten a less root mean square error between actual and predicted unemployment rates using regression, and better accuracy in prediction of classes of counties for the year 2017.

Thank You