UNEMPLOYMENT IN THE UNITED STATES

LOOKING FOR FACTORS THAT AFFECT UNEMPLOYMENT.

Merle Glick

CONTENTS

I. Introduction of Problem

II. The Data

III. Data Wrangling

IV. Analysis

V. Machine Learning

VI. Conclusion

THE QUESTIONS

- How have unemployment rates changed over the past 10 years?
- What factors are correlated with unemployment rates?
- How can we use these factors to predict unemployment rates?

THE DATA

- USDA site
- U.S. Census Bureau
- Unemployment rate for every county of the U.S. from 2007 – 2017
- Population, median income, education level

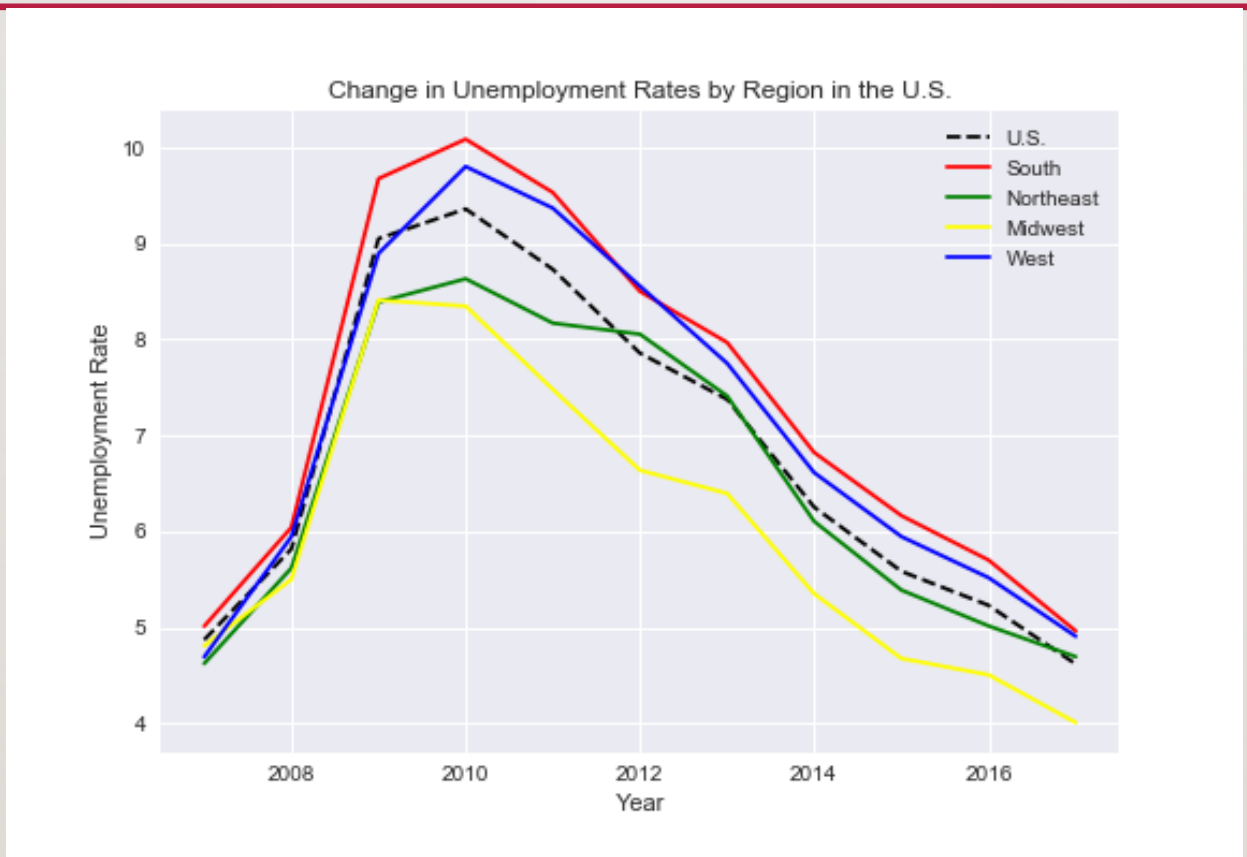
DATA WRANGLING

- Tools: python, pandas, numpy, matplotlib
- Dropped unwanted data – U.S. territories, irrelevant columns
- Converted median income from string to integer
  - Defined a function: ‘remove_non_digits()
  - Removed ’$’ and comma from values

AND MORE DATA WRANGLING

- Merged multiple dataframes
  - Reformatted 'county name' column so that it was the same in all dataframes
  - Used pandas .stack() function to combine population estimates into one column
- Did more cleaning so that columns contained only the necessary information
- Imputed missing values

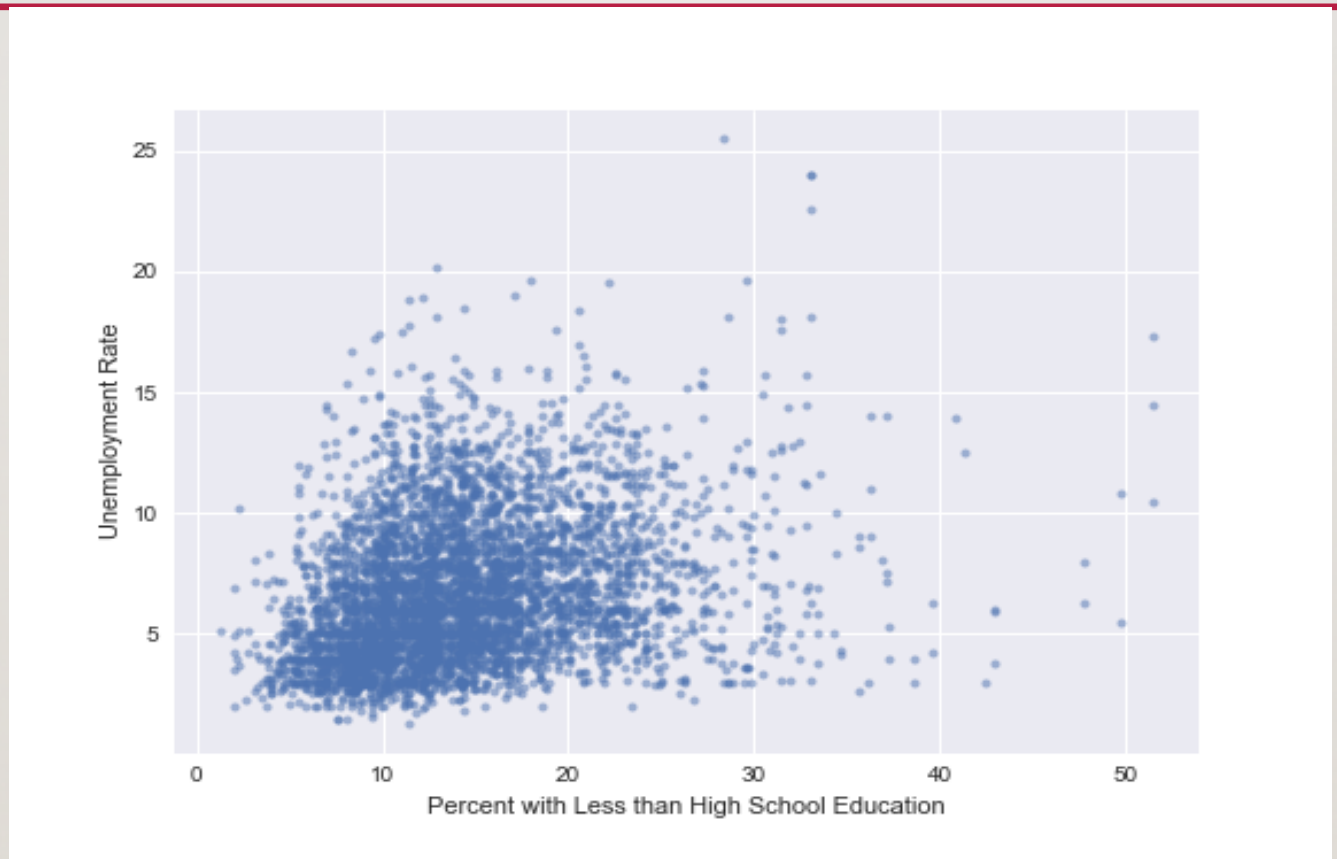
CHANGE IN UNEMPLOYMENT RATE BY REGION



ANALYSIS

- Which variables are strongly correlated with unemployment rates?
- Which variable are correlated with each other?
- Used Seaborn pairplots to visualize correlations
- Used variables that are obviously correlated with unemployment rates

CORRELATION BETWEEN 'LESS THAN HIGH SCHOOL' AND 'UNEMPLOYMENT RATE'



MACHINE LEARNING

- Used Principal Component Analysis to transform ‘Education’ variables
- Reduced 4 Education variables to 3 PCA features
- Scaled the variables with MinMaxScaler, to get better prediction
  - All values are between 0 and 1
- Split data into training and test sets

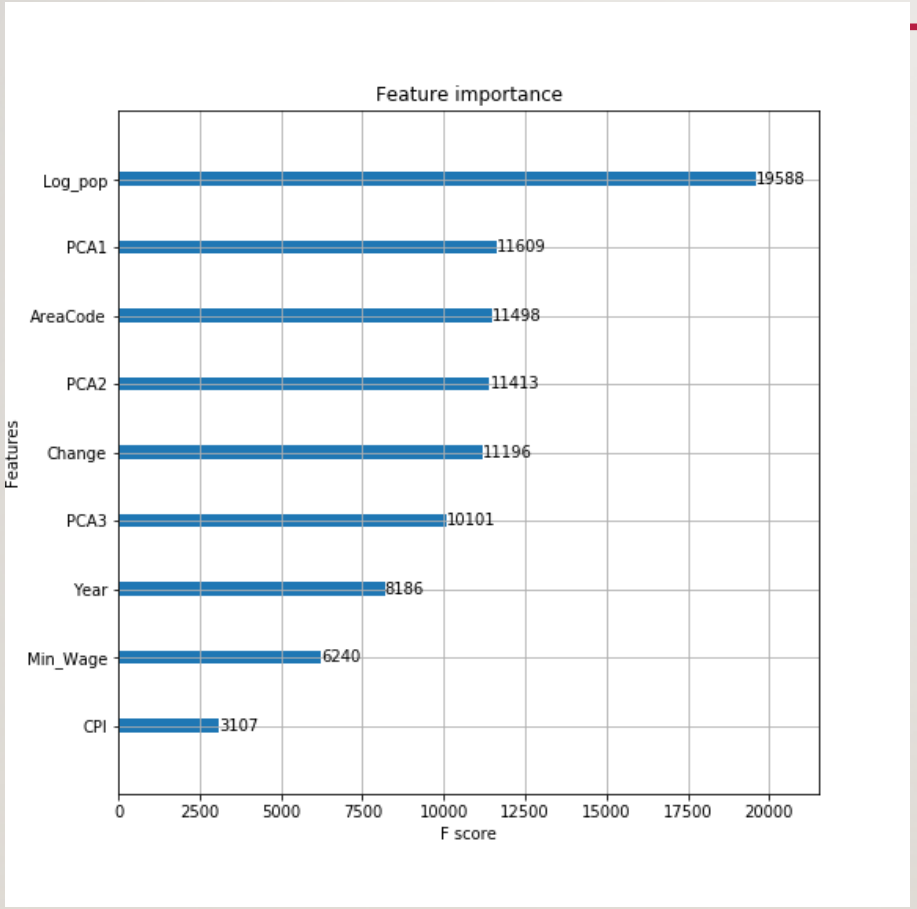
CONSTRUCTING THE MODEL

- Linear Model
- Used 10 features to fit the model
- Used mean absolute error (MAE) as measure of how well the model works
- With linear model, MAE = 0.07

XGBOOST

- A fast, accurate machine learning algorithm that uses gradient boosting
- Uses decision trees to build the model
- Has many hyperparameters that can be tuned for optimal results
- Used sklearn randomized search algorithm to tune parameters
- Resulted in MAE = 0.027, much better than linear model

FEATURE IMPORTANCE WITH XGBOOST



Shows importance of each feature based on how many times it split on that feature while building the model

Population (logarithm of values) is most important feature in this model

PCA features (education data) are all useful in predicting unemployment rate

CATBOOST

- Another machine learning algorithm that uses boosting
- Used randomized search to tune hyperparameters
- Fit model using optimal hyperparameters
- Resulted in MAE = 0.033
- Not quite as good as XGBoost, but still much better than the linear model

CONCLUSION

- Using XGBoost, I was able to construct a model that is good at predicting unemployment rates
- Possible future work: acquire additional data to improve model (might slow down the process)
- Usefulness of model: Agencies could implement strategies to reduce unemployment rates by focusing on factors that have been shown to affect unemployment