Capstone Project 1: In Depth Analysis

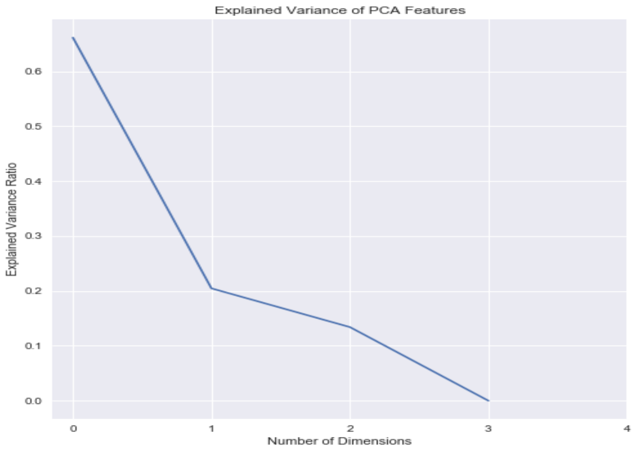
I. Introduction

The final goal of this project is to build a model to predict the unemployment rate for a given region of the United States. The data has been cleaned and formatted in a way that is easy to analyze. Some exploratory analysis was done to identify trends in the data, and to determine the best approach for building the model. Some of the variables were eliminated, since it was determined that they were not useful for building the model. Null or missing values were either eliminated, or replaced using the most appropriate method for each case. The categorical variables were encoded by numerical values so that they can be used in the algorithms for building a model.

The data frame contains almost 35,000 rows and 11 columns. Before we use machine learning algorithms to build the model, a little more work is required. We need to determine which features are most useful for predicting unemployment rates, and we also need to deal with features that are correlated with each other.

II. Analysis and Feature Selection

In the data set, there were four different variables related to education: ‘Percent with Less Than High School’, ‘Percent with a High School Education’, ‘Percent with Some College’, and ‘Percent with a Bachelors Degree’. In the previous analysis, it was obvious that these features are strongly correlated with each other. Because of this, we probably do not need all four of these variables in the model. I used Principal Component Analysis (PCA) to analyze these features. This is an algorithm that can be used to reduce the dimensions of the data. So rather than having four different variables related to education, we can reduce it to 2 or 3 variables. By plotting the explained variance of the features (shown below), I was able to determine that most of the variance in the data was explained by three of the features. I will include these 3 features in the model for the education data.

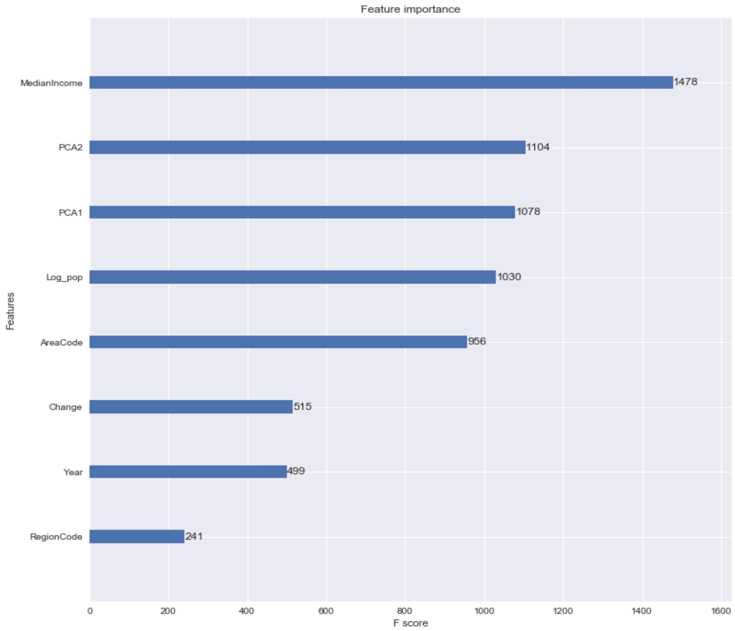


One of the variables in my data set is the population estimate for each area. Some of these are for an entire state, and the rest are for individual counties. Because of this, there is a lot of variation in the values of this variable (state populations are much higher than county populations). I took the log of all the values in this column, and created a new column with these new values. I also scaled all the data, which will help us to build a better model.

III. Fitting the Model

There are many different types of models that can be used for prediction. The next step is to determine which model will work best for this particular problem. The first attempt for a model was a linear model. This will give us a baseline to compare to as we try other types of models. To fit the linear model, I dropped some of the columns that were not needed, then split the data into training and testing sets. I set aside 30% of the data as the test set, and used the other 70% as the training data. The training set is used to fit the model, then the testing data is used to test the model, to determine how well it works on new data. The initial model had an accuracy score of about 25%, while the mean absolute error was about 0.07.

Next, I used a popular machine learning algorithm called XGBoost (Extreme Gradient Boosting) to fit a model. This uses decision trees to fit a model using specified parameters. Using this algorithm, I got a mean absolute error of 0.027. XGBoost also allows us to look at feature importance, to determine which features have the greatest impact in the model. I constructed a bar plot showing the feature importance for this model. The feature importance is determined by how many times that particular feature is split on across all trees in the model. The plot is shown below.



I decided to find other data to try to improve the model. I found minimum wage data for each state in each year, and also the CPI for each year. After adding this data, I fit the models again and compared the mean absolute errors. Using XGBoost, the error decreased slightly. By plotting the feature importance again, I was able to determine that minimum wage was more important than CPI.

The XGBoost algorithm has several hyperparameters that can affect the accuracy and efficiency of the model. I used cross-validation with a randomized search to find the optimal values for these parameters. I fit the model again, using these values for the hyperparameters, and got a mean absolute error of about 0.026. Without tuning the parameters, I got a very small value for the mean absolute error, and it did not improve much after tuning the parameters. This algorithm seems to be doing a very good job of building a predictive model to predict unemployment rates.

IV. Conclusion

There are other variables that may also affect unemployment rates, and it might be possible to improve the model slightly by adding this additional data. However, this model does a good job of prediction without being too complex. This model could be used to predict the unemployment rate for a given region if one or more of the predictor variables changes. For example, if the population of a region is expected to increase and the median income is also expected to increase, how will this affect the unemployment rate? By answering questions like this, agencies will be able to know how to plan for future growth of a city, and implement strategies to minimize unemployment in these regions.