Cook County Sentencing Data Analysis

Executive Summary

By Srikanth Nanduri

Background & Purpose:

The objective of this project two-fold. The first goal is to build a predictive model which using the features calculates/predicts the length of the sentencing for the individual. The second goal is to create a classifying algorithm which will appropriately classify the race of the individual based on the given set of features. The year 2020 came with a variety of challenges from the Presidential Election to the widespread Pandemic, the year was filled with difficult problems that required society to band together (socially distant, of course) and solve them. One of the more prevalent social issues that emerged in the past 10 months is the Black Lives Matter movement. The BLM movement questioned and shined light on the injustices faced by many Black individuals specifically in the justice system. In my research to investigate the problem and find data driven insights, I came across the Cook County dataset which contains the detail sentencing data between 2008 – 2015. The data contains 241,659 rows, 41 columns.

Results and Analysis:

For the first objective, the goal was to predict the 'COMMITMENT_TERM' value. after using Linear Regression which included Ordinary Least Squares, Lasso, Ridge and Stochastic Gradient Descent Regressor models, the predictive ability of the model was approximately 18 percent and a root mean squared error that ranged between 10.80 to 11.15. The features in the dataset did not provide enough predictive power to accurately predict the sentencing time.

For the second objective, the goal was to accurately classify the 'RACE' attribute using the give set of features. The analysis began with Clustering and Principal Component Analysis to validate if unsupervised learning would better classify the data than supervised learning. However, the analysis for both judged through the Completeness and Homogeneity Scores was 0.005 and 0.01 which means members of one class belonged to more than one cluster and clusters did not contain a homogenous group of classes.

To take the classification problem one step further, the analysis began with a Decision Tree Classifier which proved to be the best model at classifying the data accurately with a 69 percent accuracy rate. Then, using that model, the Random Forest algorithm was implemented to get an aggregated group of results for multiple decision trees which came out to approximately 68 percent accuracy. The last step was to find the most optimal Random Forest algorithm parameters to see if there was any additional improvement. The results remained at a 69 percent accuracy.

Conclusion & Recommendations:

Overall, there are a few takeaways from the data analysis on this dataset. Based on the calculations, we can see that the Ordinary Least Squares and Linear Regression Models performed the best. Meanwhile, the SGD and Lasso regression methods produced a higher RMSE value. We can clearly say that based on the models created, the Linear Regression was the simplest and easiest model to replicate. However, we need to keep in mind a few things.

The first is that the model's overall predictive ability measured through R-squared coefficient is approximately 18% at best. This basically means the model can only explain 18% of the variance in the data which is a weak value. The second thing that we can also see the weakness in predictive power when we look at the Root Mean Squared Error. Using the RMSE on the training data, the model is

generally off by approximately 8 years in predicting the length of the prison sentence. Similarly, the model is even worse when predicting the testing data where it is off by 10+ years.

Even so, we can see some clear insights from the analysis namely the impact of certain features and the number of features required to have the best model. The most informative variables are as follows:

1. CHARGE_COUNT
2. OFFENSE_CATEGORY_Homicide
3. OFFENSE_CATEGORY_Narcotics
4. OFFENSE_CATEGORY_Sex Crimes
5. CHARGE_DISPOSITION_Not Guilty

Another element to keep in mind is the need for domain knowledge and understanding. While we can use the basic data available to predict to a certain degree the length of the prison sentence, the fact of the matter is that there are quite a few subjective elements that determine how long someone's sentence is. For example, the severity of the crime committed (1st degree versus 3rd degree) the actual circumstances involving the crime committed and other case specifics could not be considered as data elements.

Moreover, clustering does not enable us to sort out the RACE of an individual being sentenced. The completeness score, which calculates whether all members of a give class are assigned to the same cluster, shows that the clustering method did not accurately classify the Race. Next, the homogeneity score, which measures each cluster contains only members of a single class, shows that the levels within RACE were scattered across multiple cluster. This makes sense because when we saw the initial distribution of the label data, a few of the categories only represented less than 0.1% of the population, however, the clustering algorithm calculated significant values for the smaller labels.

Lastly, based on the analysis, the Random Forest Classifier with the most optimized parameters generates a 69% accuracy in correctly classifying the data. This is not significantly different from the regular decision tree classifier. Moreover, the most important feature in correctly classifying the data are 'OFFENSE_CATEGORY_Narcotics', 'OFFENSE_CATEGORY_DUI', and 'COMMITMENT_TERM'. Lastly, the verdict (guilty versus not guilty) of the case did not provide any discernable value in classifying the data according.

Overall, in order to improve the models and derive further insights, there are a few things that are required. Specifically, a stronger understanding of the legal systems and the underlying details on how the sentences are carried out. Next, there needs to be greater clarity on how the data is aggregated and maintained within the Cook County database and potentially identify and communicate with Cook County legislators on whether additional features could be made available to improve data insights. And finally, the location data could be utilized to map out where the criminal activities are carried out and how that location affects the ‘COMMITMENT_TERM’ and if it is limited to a specific ‘RACE.’