Thip Rattanavilay DSC530 Term Final

March 4, 2021

This dataset describes crimes in the City of Los Angeles dating 2010 to present. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. The dataset refers to the last update of 2010-2021 and contains 26 columns and 1751418 rows, each of which corresponds to a crime incident.

Our work started with a satisfying Data Cleaning process, during which I changed original dataset formats with more standard ones, which are also easier to handle. I also added several useful columns to the original dataset (such as Year Occurred, Month Occurred, Timeslot Occurred, etc), which have been used for the analysis computation during section (4), namely Explorative Analysis section. In this section I have reported several analysis using plots such as bar plots, box plots and histograms. During the analysis computation is also emerged a strange anomaly regarding Victim Ages, which prompted us to carry out a further iteration of data cleaning with the objective of remove as more as possible this anomaly. Doing this, I have successfully cleaned in a better way our dataset removing, even if partially, the anomaly. Furthermore, this section has figure out other significant analysis about our dataset, regarding for example the most frequent crimes and lapons used in the city of LA. I show also interesting statistics about crimes distribution over years, months, timeslots and Weekdays, and some of them Ire a bit surprising.

As result I discovered interesting (and in some cases unexpected) correlations betIen some of these variables. The most unexpected of these indicates that Victim Sex is highly correlated with the type of structure, vehicle, or location where the crime took place (Premise Code), but also with the type of the Crime (Crime Code). I reported these results on a Correlation Matrix. As done for Unsupervised Learning, also in Supervised Learning I worried about using meaningful. I chose these as target variables since they represent information which could be unknown. Therefore, trying to predict this information could be interesting and useful. So as result of my finding I can be satisfied, since I provided simple tools potentially usable in the real world, yet I don’t think I can support authorities (such as FBI and police) in their work just yet...

Lastly For my regression analysis test, I decided to do a multiple regression test to anslr the question of can I predict when a crime is going to occur. I build baseline model, which divides test set into five groups based on its average of 4 values (number of crimes of each factor: location, month, date, hour) in descending order. If given features belong to first group, the model predicts. As the p-value of density is 0 (small), the changes in crime rate have got close relation with changes in density. R-squared value is found to be 0.525 with only density as predictor variable. This means that 52.5% variability of crime rate is explained by density feature. Co-effient estimate of 0.0086 indicates one value increase of density would cause 0.0086 value increase in crime rate. Standard Errors assume that the covariance matrix of the errors is correctly specified. The condition number is large, 3.28e+06. This might indicate that there are strong multicollinearity or other numerical problems.

In conclusion, this project led to interesting results, analysis and statistics, but also provided useful tools both for authorities and population, which that allows a better understanding of crimes in LA.