Shayne Erickson

Prof. Lam

ECON 4980-001

9 March 2021

Predictive Model for Determining Primary Chicago Crime Type

*Overview*

This project was in part designed to allow us to show our knowledge of python, yet the primary goal was for us to apply our in-class knowledge of machine learning techniques. My machine learning model took in 6 categories of input values, returning whether or not a crime was a theft. Using my method for this one crime type, the learning and prediction can certainly be extended, one-by-one, to each other category of crime.

*Data Usefulness*

The data we were given to analyze was the Chicago police log of crime from 2001 to the present. It came with a few issues regarding the “cleanliness,” so the first step was to obtain something reliable. The data is obviously extremely large, so the data set that I use in my program is a 10% sampling of the data. This command can be removed in order to use the entire dataset for predictions with a larger basis. However, we are looking to (i) execute this program with an acceptable run-time and (ii) make a conjecture about this crime type overall given only a sample of the data set. After this, I removed incomplete rows of data (null values in at least one column). I did not select a particular year, due to the nature of my “seasons” input. From here, I was able to work with a complete, reliable, representative data set.

*Input and Target Data*

The inputs for my dataset were: (1) the season slot that the crime occurred within, (2) the day of the week on this crime occurred, (3) the time slot throughout the day in which it occurred, (4) the shortest distance from the crime’s coordinate location to one of the twenty-two Chicago police precincts, (5) whether or not an arrest was made, and (6) whether or not this was a domestic incident. The output is a 1 or 0 – 1 indicating a theft was the primary type of the crime that occurred given these parameters, and 0 indicating that a theft was not the primary crime type. My rationale is as follows: crime is seasonal (winter brings more Christmas-related crime, for example); crime is cyclical, since weekends may see more of one type than others due to scheduling and the occurance of opportunity crimes; crime occurs at different times – night times are much more prone to activity due to offenders’ limited visibility; crime has the possibility of occurring less nearest police stations; and the arrest/domestic categories can indicate the severity of the crime, a hint as to the crime type that might escalate to such levels.

*Chicago Crime Analysis*

My program was able to analyze the crime dataset presented, returning whether or not a crime is of primary type “THEFT” with an accuracy score of 80% (+/- about 1%). This is not extremely strong, as there could be a number of vital factors not included or unnecessary factors included in the input variables list. In addition, it is a relatively limited predictor. However, by switching what the ‘target’ variable is set to – i.e. changing the below line of code:

theft\_type = numpy.where(dataset['Primary Type']=='THEFT', 1, 0)

so that the word “THEFT” is replaced with the new crime type of choice:

crime\_type = numpy.where(dataset['Primary Type']== “CRIME\_TYPE”, 1, 0)

then we can run the regression for each type of crime.

There are two alternatives for this multiple-output: (i) run a multinomial regression or (ii) calculate a probability for each crime, rather than a 1 or 0, and return the crime type with the highest probability of occurring. People’s behaviors, especially in the moment of a crime, have a substantial degree of unpredictability, so an array of most-likely crime types (given these particular inputs) may give more insight into the range of crime types that might occur with similar circumstances, rather than whether or not a particular crime type did occur given circumstances.