Crime Rate NYPD Predictions

Milestone: - Data Collection and Processing

Group 14

Mayur Mahanta

Aniket Sakharkar

857-437-9190 (Mayur Mahanta)

857-230-5126 (Aniket Sakharkar)

mahanta.m@northeastern.edu

sakharkar.a@northeastern.edu

Percentage of Effort Contributed by Mayur: 50%

Percentage of Effort Contributed by Aniket: 50%

Signature of Mayur: *Mayur*

Signature of Aniket: *Aniket*

Submission Date: 27th May 2022

Data Collection:

Here, we are gathering data in a measured and systematic manner to ensure accuracy and facilitate data analysis. The "NYPD_Arrest_Data_Year_to_Date_".csv file was uploaded to Google Collaboratory and read with pd.read_csv() before being stored as a data frame.

Data Processing:

The first set of operations we performed with the uploaded data frame are as follows:

Step 1: Using shape () function on the data frame, we found out the number of rows to be 155507 and the number of columns to be 19.

Step 2: Using info () function, we found out the datatypes of each variable and their corresponding null values

Step 3: In this step, we have assigned a new Model Data Variable to store the same dataset using the function copy () whilst keeping the original data frame intact for visualisation purposes.

Step 4: Here, we are transforming all categorical variables into numeric variables using labelencoder() from the library sklearn. The variables on which we are performing this operation here are, "*PERP_SEX*," "*PERP_RACE*," "*AGE_GROUP*," "*ARREST_BORO*" and "*OFNS_DESC*."

Step 5: After changing the categorical variables to numerical variables, we found out the variables of interest that are "*PERP_SEX*," "*PERP_RACE*," "*AGE_GROUP*," "*ARREST_BORO*" and "*OFNS_DESC*." Secondly, the variables "*PD_CD","KY_CD","LAW_CAT_CD","PD_DESC","ARREST_DATE","New Georeferenced Column","LAW_CODE"* are not of relevance from the perspective of training data but will be used during visualisation.

Step 6: In order to double check null values, we have omitted all non-important object datatypes and selected only the float64 and int64 datatypes. We have executed this using isnull().sum() through which we have eliminated all null values in our Model Data frame.

Step 7: Now, using describe () function, we are retrieving the summary of the statistics pertaining to the Model dataframe.

Step 8: Moving on, we are trying to check for outliers without using visualization at this point. Hence, we are using a predefined statistical function called find_outliers_IQR() which uses the concept of finding outliers using the Interquartile formula. We have called the function on one variable called "*JURISDICTION_CODE*" to find length of the outliers, maximum outlier value and the minimum outlier value.