DATA 606 Final Delivery

**Chicago Crime Analysis**

**Final Summary Report**

**606 Capstone Project**

**By**

**Sai Vinay Pabbisetty**

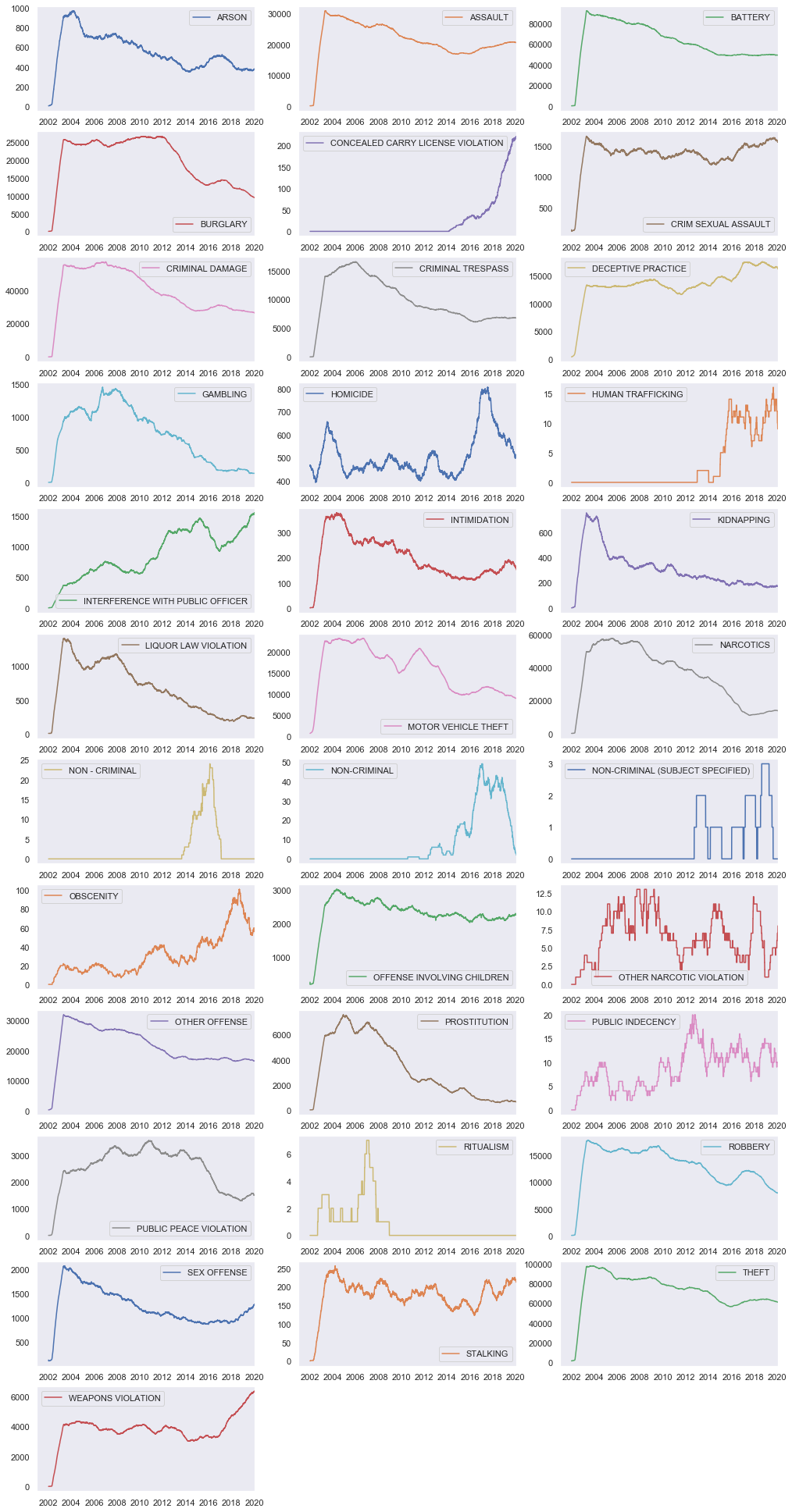
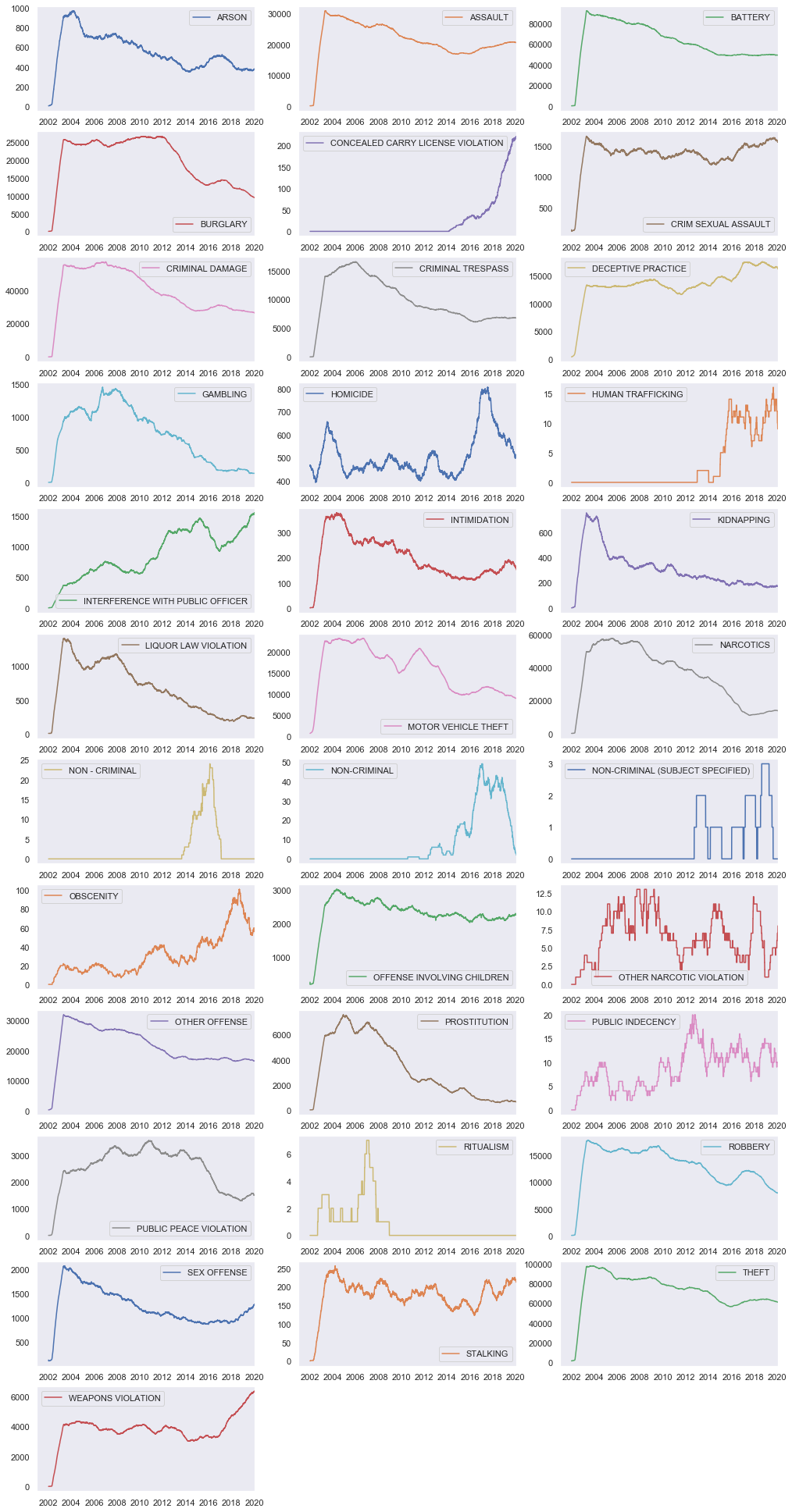
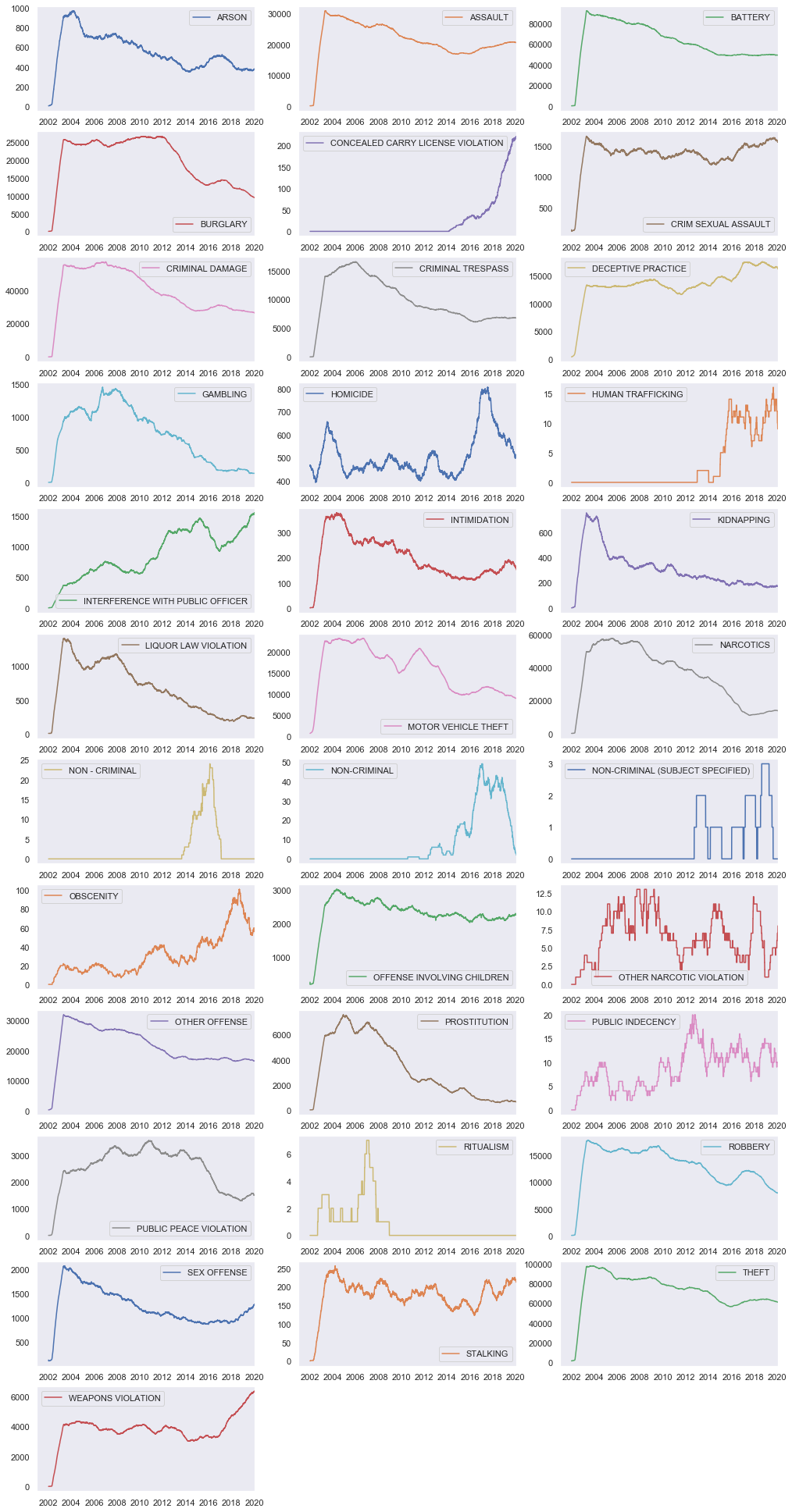
**Instructor: Ergun Simsek**

**Project Goal**: My Project goal was to predict the crime type from other features like when and where the crime has taken place. And to explore the data by getting meaningful and useful trends to answer some questions related to crime.

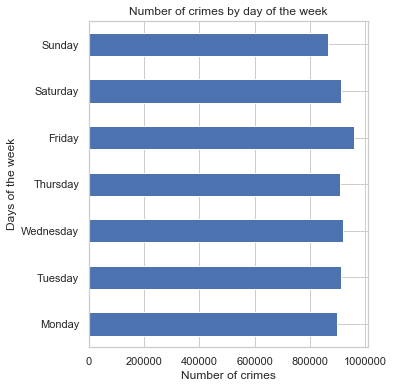
**Motivation**: How the crimes have evolved in Chicago? Which Areas of Chicago are safer and which are not in terms of crime? Can we predict the type of crim based on other features?

**Exploratory Data Analysis:**

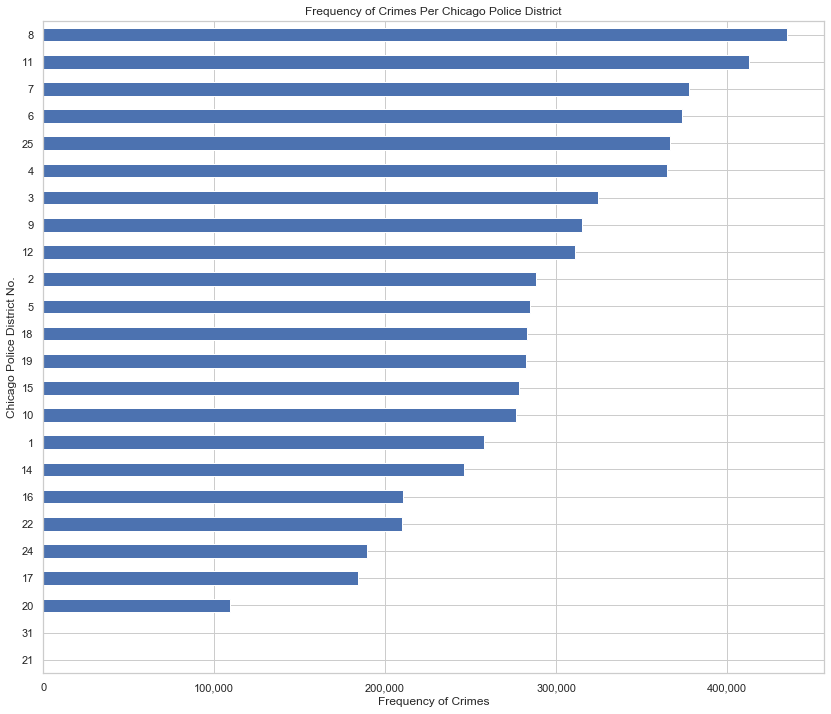
I gained some good insights by doing exploratory data analysis for Chicago Crime Data. I am including five interesting visualizations my EDA has produced, Firstly I plotted a graph to see how each crime has evolved over the years from 2001 to 2020. Out of all the sub graphs that I observed in that graph I noticed that Weapons violation, Concealed carry license violation and Interference with public officer are the crimes that got increased over the years linearly and reached there maximum in recent years. You can observe that in the below picture.



I continued my analysis by plotting a graph to answer this question: On which day of the week are more most no. of crimes committed in Chicago?. And the results were as follows:



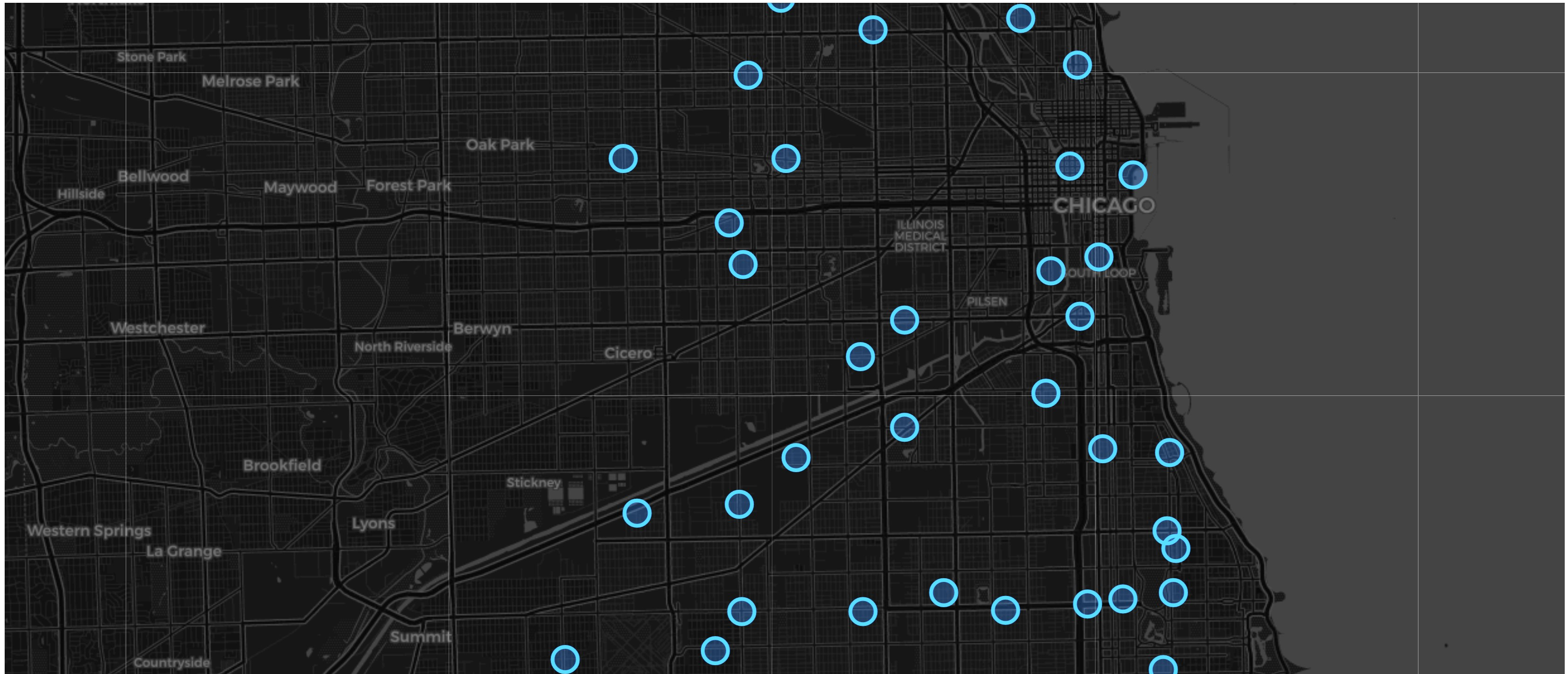
As per the graph we can say that number of crimes committed are distributed evenly throughout the week but On Friday more crimes are committed when compared to other days in a week. I also plotted a graph to see how the crimes are distributed across different districts in Chicago and it came out like this:



From the above graph we can infer that District no.8 has more crime frequency than all other districts in Chicago which makes it the dangerous place to live in terms of crime in Chicago. Districts- 20,21,31 are the safer places to live in Chicago according to the graph as they are having lower crime frequency.

**Crime Mapping:**

Using the longitudinal and latitudinal location data I have implemented crime mapping using a package called folium. It gives us a interactive window as output in which we can zoom in and zoom out to see the crime spots that are plotted on the map. The image looks like following:



The blue circles in the map image above show us crime spots which are plotted on the Chicago map using the longitudinal and latitudinal data.

**Data Cleaning, Feature Engineering and Data Wrangling:**

Firstly, I made sure that the data is not having any inconsistencies, missing values, and null values before starting the Feature Engineering. I then decided to focus on the required features, so dropped the columns which are not required for modelling like Updated On, ID, Case number and separate Longitude and Latitude columns.

After dropping the columns, I have factorized required columns which are in categorical form as I must fit the data to models. Then I started preprocessing the target variable by plotting graphs to see which classes are least occurred to combine them into a single class called OTHERS. OTHERS is a combination of crime types like Homicide, Kidnapping, Stalking, Obscenity, Ritualism etc. which are very less frequent in the given data set.

Target variable column is now ready for factorizing to convert it from categorical form to numerical form which the models can parse and understand.

Now every categorical and numerical feature is ready for training a model, Then I have plotted a correlation plot with respect to the target to find out which features are important for making predictions. I have included the features with at least a correlation of 0.2 and discarded the rest of features which I felt are not useful which has very less correlation (< 0.2).

**Splitting Data:**

Now the data is completely clean and ready for training the models. I had split the data into train(80%) and test(20%) with a random state of 3

**Feature Set Used : ['IUCR', 'Description', 'FBI Code', 'Location Description', 'Domestic']**

**Target Class : Primary Type**

**Training Set Size : (80000, 23)**

**Test Set Size : (20000, 23)**

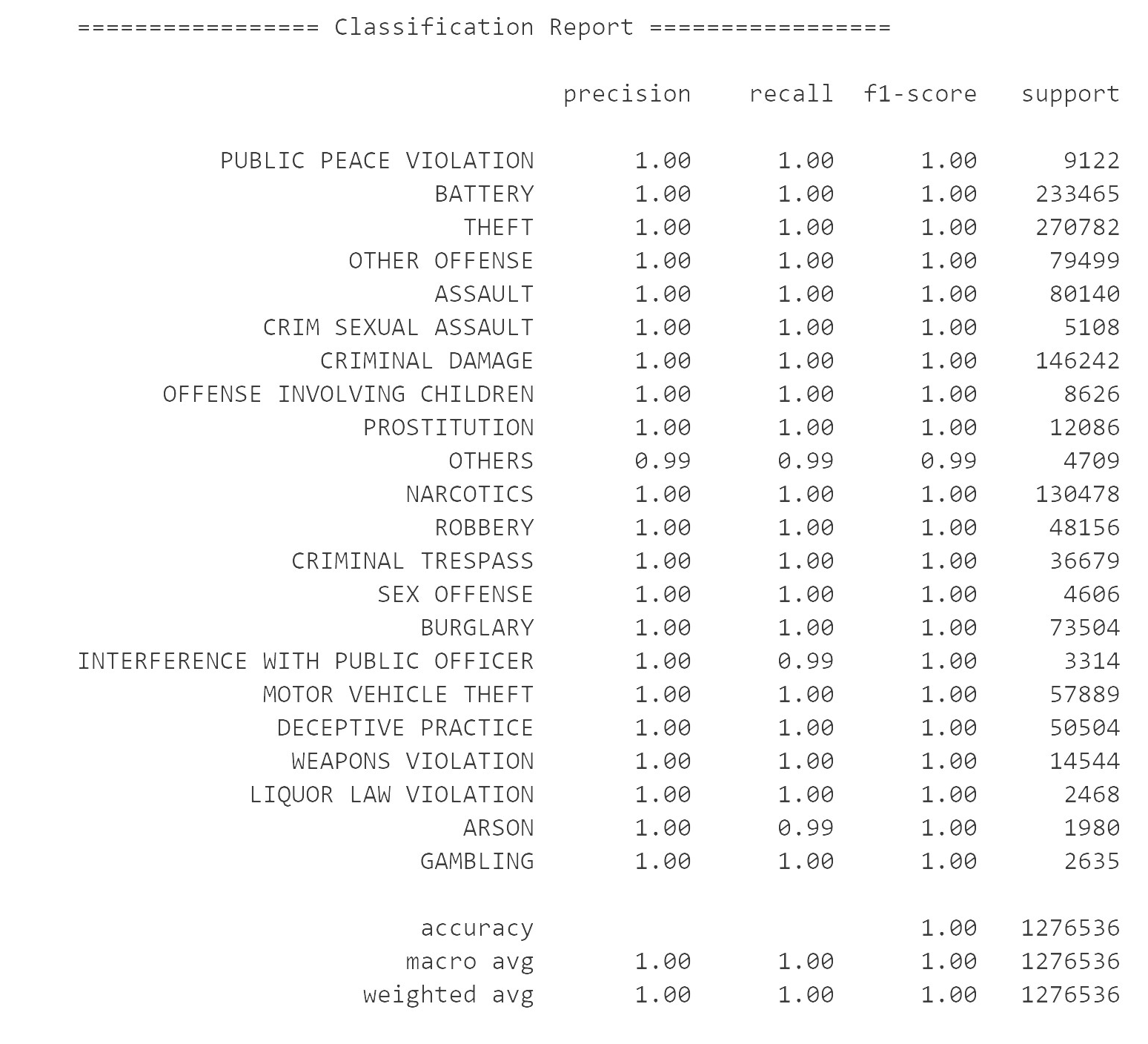
Training Set and Testing set might look very small because I have sampled the data to 100000 rows, As it is a very big data set it takes so long to run the code if you don’t sample the data. But the results were almost equal if we train the whole 6-7 million rows of data.

**Training Models & Obtained Results:**

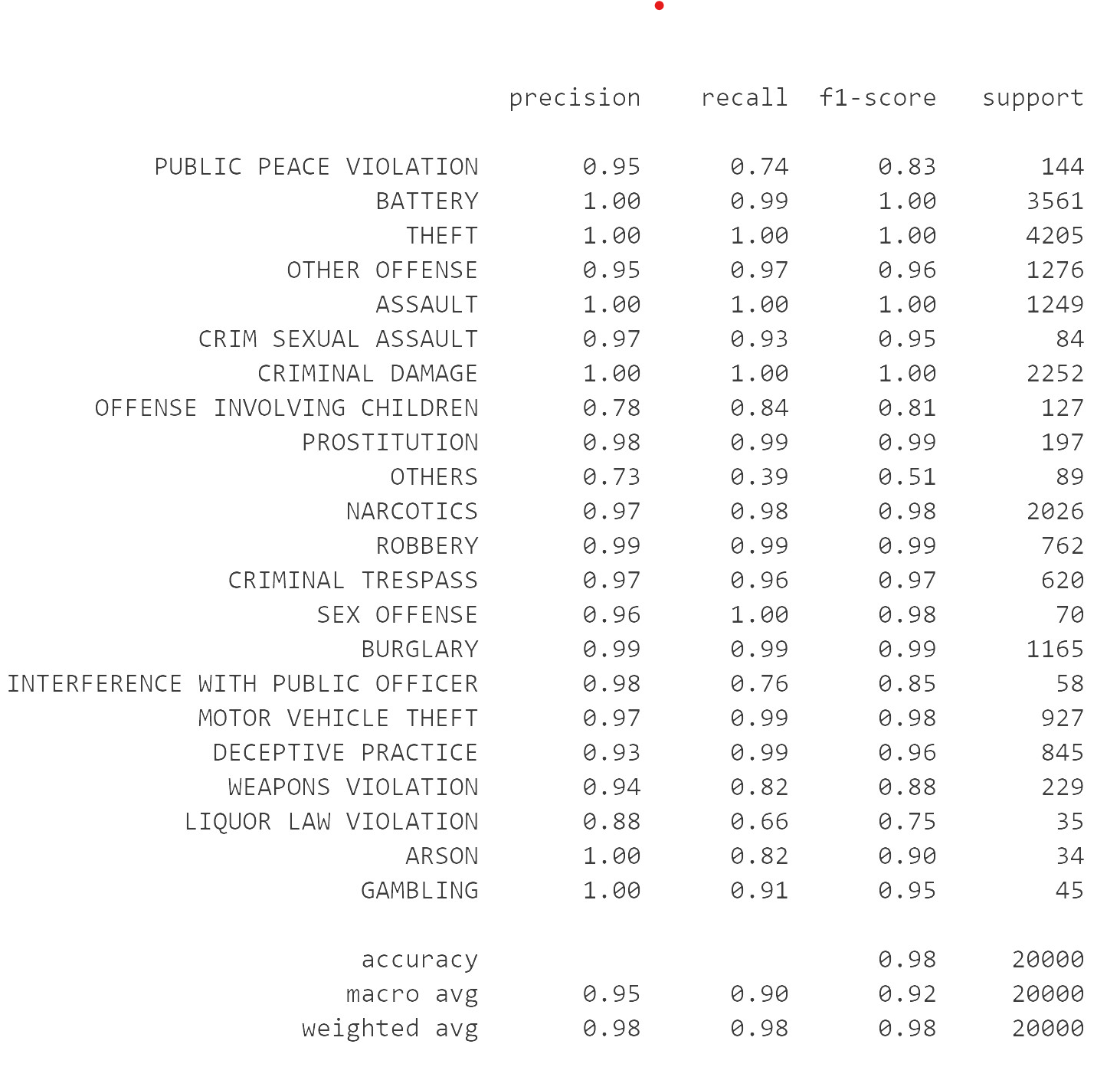
I have used three different models to predict my target “Primary Type”, which is the type of crime committed. The three models that I have used are as follows:

1. Random forest Classifier,
2. Multilayer Perceptron Neural Network
3. K Nearest Neighbor Classifier

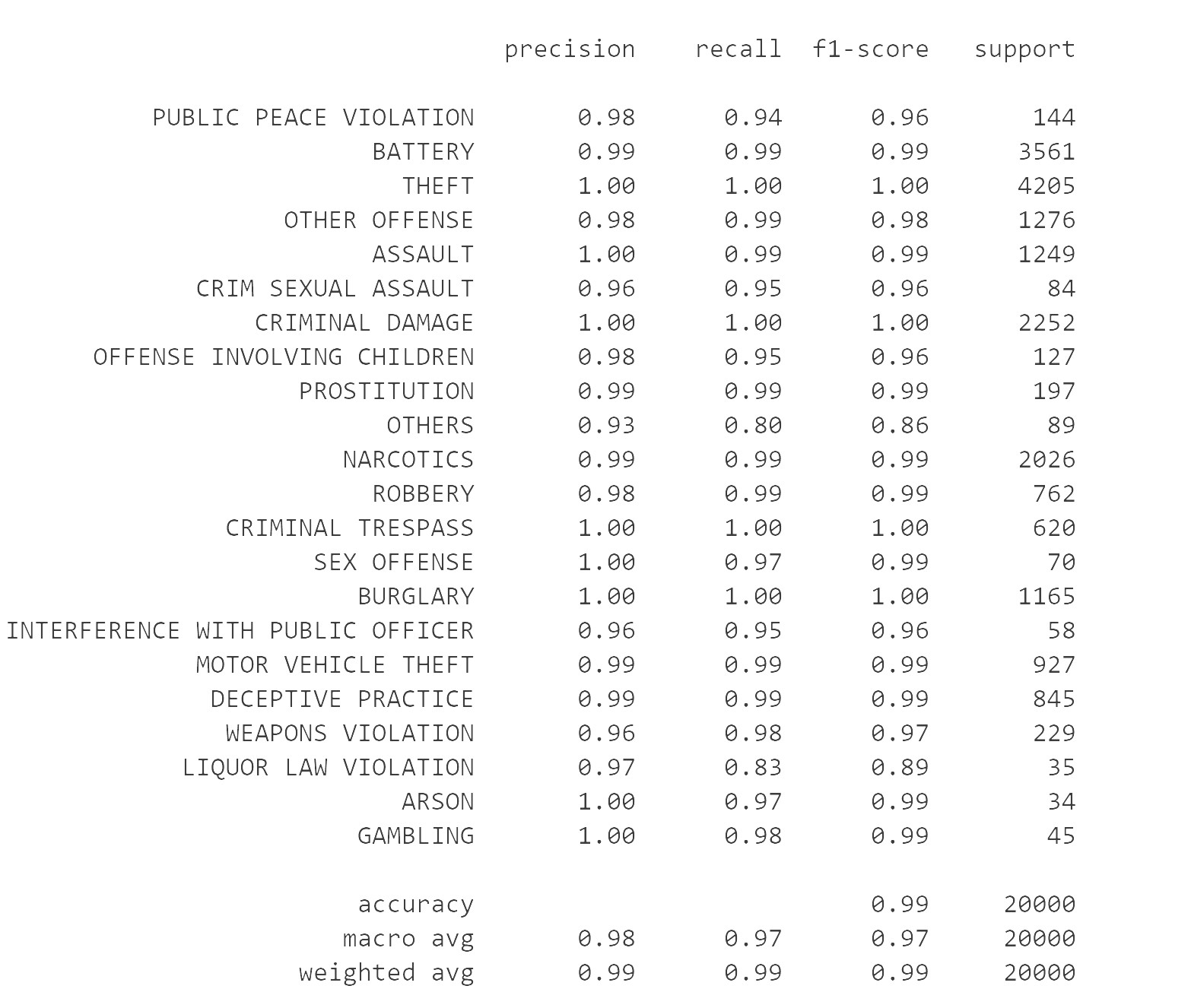
Random forest Classifier performed well by predicting almost all the classes of target correctly. The Classification generated for random forest classifier model predictions looks as follows:



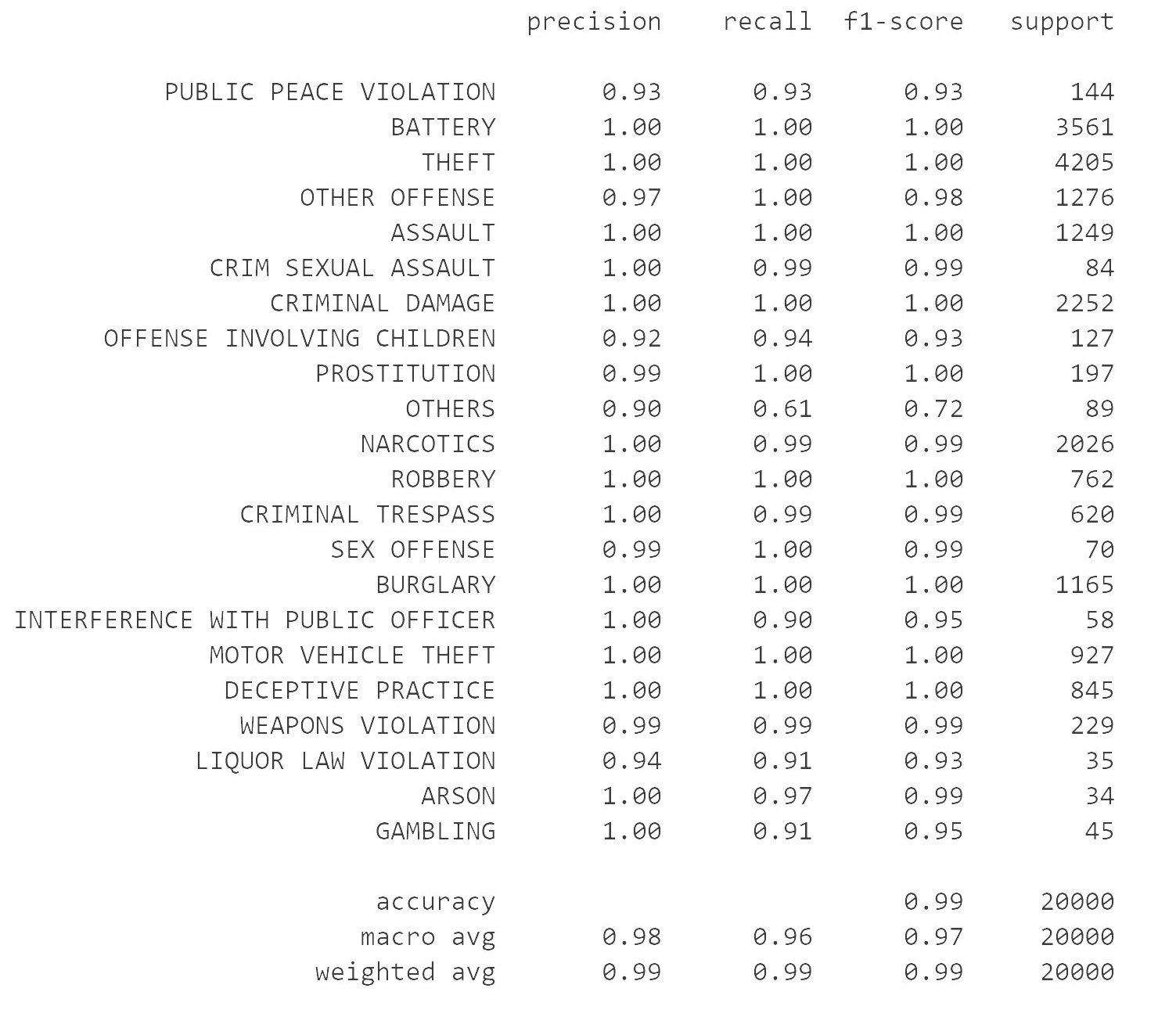
Multilayer perceptron Neural Network model also did a good job in making the predictions. But it got confused with some classes like Public Peace Violation crime, Liquor Law Violation crime and OTHERS class which is a combined class of less frequently committed crimes. The Classification Report for this model is as follows:



Even K nearest neighbors Classifier faced some problem in predicting some classes like Liquor Law violation and OTHERS class but it is comparatively better than Multilayer Perceptron Neural network classifier. The classification report for KNN Classifier is as follows:



After training all these models I used ensemble voting classifier by combining all three models to make better predictions but even voting classifier got less scores when compared to Random Forest Classifier model alone. I think this difference in scores is created because of the OTHERS class which is a combination of all less frequently committed crimes. By excluding the less frequently committed crime type classes I expect ensemble voting classifier to perform better than other models as it combines all three model predictions to make better predictions. The Classification Report for ensemble Voting Classifier with all 3 models looks as follows:



**References:**

\* X. Zhang, Z. Hu, R. Li and Z. Zheng, "Detecting and mapping crime hot spots based on improved attribute oriented induce clustering," 2010 18th International Conference on Geoinformatics, Beijing, 2010

\* https://blog.dominodatalab.com/creating-interactive-crime-maps-with-folium/

\* McClendon, Lawrence & Meghanathan, Natarajan. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: An International Journal. 2. 1-12. 10.5121/mlaij.2015.2101.