IML Hackathon - Challenges 2: Help the Chicago Police Prevent Crime!

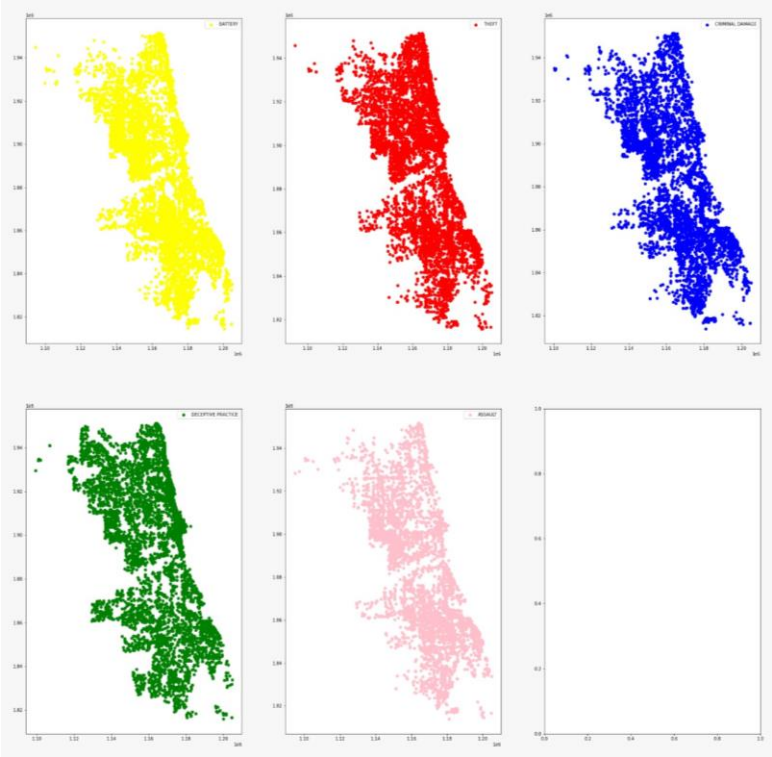
We have given a dataset of 35,000 crimes. The dataset contains facts on crime prevalence that has taken place in Chicago over 2021. There are different attributes of the dataset such as: Location Description, Description, FBI Code, Block, Location, Year, Latitude, Longitude, Month, Day, Hour, Minute, Second and Primary Type.

The implementation of this project id divided into the following steps:

1. Feature selection

   The attributes used for building the model are: 'Arrest', 'Domestic', 'Date'

   We have notice that for the primary mission, most of the location's features are irrelevant to the kind of the crime as we can see in the plot. The crimes distributed the same in Chicago independence to their type.

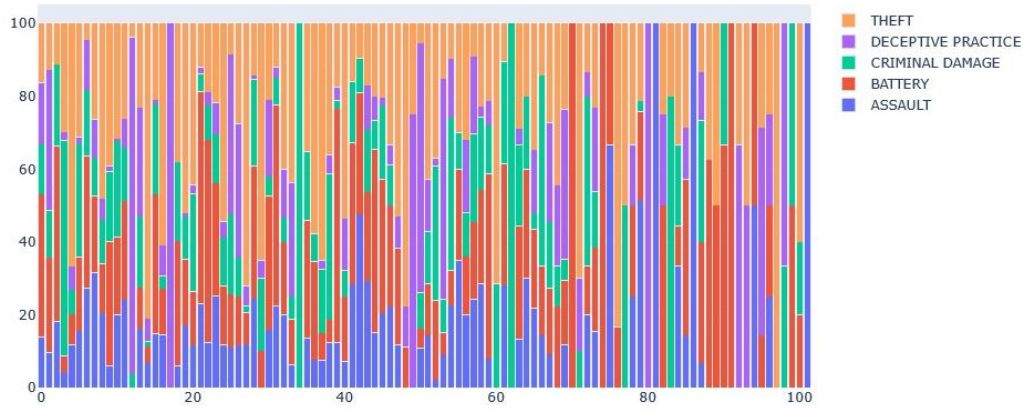


2. Data Preprocessing

   We decided to split the "Date" feature into new attributes like "Day", "weekday" and "hour" which can be used as feature for the model.

   Also, the categorical attributes are converted into dummies values. And the null values are removed.

We saw a connection between the hour and the kind of the crime



3. <u>Building and Training the Model</u>

Our first step in building the model, was to compare some classifier sklearn offer us, For example, RandomForest, SVM, DecisionTree and so on, and letting the sklearn to choose its default hyperparameters. Because RandomForest showed the best accuracy, we decided to choose it, and try to tune it by choosing the best hyperparameters we've found. By tuning the model, we have succeeded to increase the accuracy in ten percent on the train set. Amazing!

In the second mission, we decided to fight crime using different tool set – Unsupervised Learning. What we tried to do was to project our 3D data (x coordinate, y coordinate and time) into the 2D space, and then find the centrums in the data. Then when getting a date as input, return the centrums' coordinates that represent locations with a lot of crime during this time in the day.