**New York Police Department Arrest Data**

**Classification on Level of Crime**

*Final Report*

**Group members:**  
Natasha Arokium  
Jiffar Abakoyas

Bethold Owusu

Thierno Diallo

**Introduction & Motivation**

New York City ranks #25 in the United States list of cities with most violent crime rates[[1]](#footnote-1). Since we all reside in New York City and safety is a common concern, we thought it would be useful to understand which factors about crimes in NYC allow us to make predictions. In order to do this, we analyze the New York Police Department's citywide crime statistics for 2019. The NYPD dataset contains arrest information that contains information such as the type of offence, the age and race of the offenders, the borough where the crime occurred, and other relevant information for analysis.

Our main machine learning task was to predict the level of crime committed from other descriptions of the crime. We think this is important to help us know what factors are important in knowing that a felony will occur. If the level of crime can’t be predicted from other features of the crime, then that will also be important information to note.

**Preliminary Exploratory Analysis**

To do better machine learning, first did the following EDA to understand the data:

**The Data**

There were 214,000 rows in our dataset, and the distribution of the level of crime was as follows. The original dataset consisted of 17 features and 1 target column. Out of the 17 features, we used the following 8 columns for our classification model:

A close up of text on a black background

Description automatically generated

**Table 1: Column names used in our classification task. (Level of Offense is our target variable)**

**The Target**

The target class had five classes: Felony, Misdemeanor, Infraction, Violation and Unknown. Since over 95% of the data belonged to the felony or misdemeanor class, we dropped the other three classes before training our classifiers. The two classes were split 60/40 between misdemeanor and felony respectively.

**A screenshot of a cell phone

Description automatically generated**

**Figure 1: Number of arrests by level of crime**

We did visualization on some of the feature columns in order to determine how useful they would be towards our classification model. Among them is borough,

**Borough**

**A picture containing drawing

Description automatically generated**

**Figure 2: Number of arrests by Borough**

Above it shows that crime levels by type of crime defer by borough. We expect then that the borough will be a useful feature in predicting the level of crime.

**Age**

A screenshot of a cell phone

Description automatically generated

**Figure 3: Age and level of crime**

Age showed significant variability in terms of the number of crimes and the types of crime. This was another feature of importance in distinguishing between types of crimes.

**Machine Learning**

We created the following classifier models and trained them on our data and tested them using a separate testing set. The following are our results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | Precision | Recall | F1-Score | Comment |
| Dummy | 0.52 | 0.52 | 0.52 | Approximates coin toss |
| Logistic Regression (LR) | 0.61 | 0.62 | 0.57 | Slightly improved |
| Logistic Regression (PCA) | 0.61 | 0.62 | 0.56 | Dimensions reduced by 70%. Accuracy same. |
| SVM | 0.61 | 0.62 | 0.57 | Same as LR |
| Decision Tree | 0.59 | 0.60 | 0.59 | Same as LR |
| Random Forest | 0.61 | 0.62 | 0.57 | Same as LR |
| Random Forest, PCA | 0.61 | 0.62 | 0.56 | Same as LR |

**Table 2: Classifier models and prediction results on test set**

**PCA**

We used PCA in order to see both whether it is able to significantly reduce our features, as well as improve our accuracy.

A screenshot of a cell phone

Description automatically generated

**Figure 5: Principal Component Analysis (50 components)**

PCA was able to allow us to reduce the number of columns from 150 to 50, and still maintain up to 90% of the variability of the data. This was a welcome computational and storage advantage. Nevertheless, it was not useful at increasing the accuracy of our classifiers. We performed PCA based classification on both Random Forest and Logistic Regression (Results on Table-2 above).

Since our classifiers performed similarly, we selected logistic regression for further analysis using hyperparameter tuning in an attempt to reach higher levels of accuracy. Nevertheless, the prediction was not improved by any significant margin. Using 5-fold cross validation on gridsearchCV of varies learning rates and regularization values, the best result was 0.63 accuracy (std 0.003).

**Random Forest for Feature Importance**

We also used Random Forest for Feature Importance. We were able to look at the top features for the classifier.

A screenshot of a cell phone

Description automatically generated

**Figure 3: Top 20 features from Random Forest Feature Importance**

We see above that age and month provide some indication to the classifier, but their importance values are less than 5 percent (age-group being in 25 to 44 being the feature of highest importance).

**Result**

After employing different classifier models on the dataset, it is safe to conclude that the dataset does not have rich features that allow for higher accuracy at the prediction of the level of crime. We are led to believe that it is not possible to predict the level of crime committed by a perpetrator from looking at the characteristics in our features.

**Reference**

MacDonald, John; *The Effects of Local Police Surges on Crime and Arrests in New York City;* PloS One; June 2016

NYPD Crime and Enforcement Activity Reports; <https://www1.nyc.gov/site/nypd/stats/reports-analysis/crime-enf.page>

Patternizr; <https://statescoop.com/how-the-nypd-is-using-machine-learning-to-spot-crime-patterns/>

# [Southall](https://www.nytimes.com/by/ashley-southall), Ashley; *Crime in New York City Plunges to a Level Not Seen Since the 1950s*; Dec 2017; <https://www.nytimes.com/2017/12/27/nyregion/new-york-city-crime-2017.html>

1. According to the USA today article "*Dangerous States: Which States have the highest rates of violent Crime and most murder*", Jan 2020. [↑](#footnote-ref-1)