Prediction of Crime for NYPD Using Datamining Techniques

Vyshampa Maringanti
University of New Haven
Master's in computer science
vmari2@unh.newhaven.edu

Akhila Gade
University of New Haven
Master's in computer science
agade1@unh.newhaven.edu

Bagyasree Chitikela
University of New Haven
Master's in Computer science
bchit1@unh.newhaven.edu

Abstract

Crime is a common societal issue impacting the quality of living and the economic development of society. In order to provide a stronger solution to this issue, it is very important to consider the trends of crime. By taking crime datasets from the New York Police Department's (NYPD complaint historical information), we examine this trend. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD) from 2006 to the end of last year, 2019. It also comprises of crimes and crime-related incidents categorized by the police department. The location and time of these events are stored in the records. The main aim of this research is to predict the possibility of occurrence of crime. The proposed paper uses multiple algorithms such as Logistic Regression, Random Forest, Decision Trees, KNN(K nearest neighbors) and XG Boost to evaluate the accuracy of the model.

Predicted results cannot be assured of 100% accuracy, but the results show that our model helps to reduce the crime rate to some degree by providing information about the sensitive areas of crime. Getting this kind of information will help people make better decisions in avoid such crime sensitive areas. This model will also be useful for the police forces to improve the level of crime prevention and distribution of police manpower efficiently.

Keywords: Crime, NYPD, Logistic Regression, Random Forest, Decision trees, KNN, XG Boost.

Introduction

Crime is one of the major problems around the world as it affects the quality of life and society's economic growth. As crime increases, departments of law enforcement continue to advocate for new data analysis methods to improve crime analysis and better protect their neighborhoods. Though we cannot predict who will be victims of crime, we can predict the probability of occurrence of crime. So, in order to develop such an effective crime analysis model, we need to study and evaluate the crime data.

This paper uses an NYPD dataset that contains varieties of crimes that occur, based entirely on multiple factors along with locations and offenses from 2006 to 2019. We used various classifiers to determine the occurrence of crime. The purpose of this model is to apply machine learning techniques to determine the probability of occurrence of crime.

Related Work

Different research works related to crimes have been conducted by many researchers, here are some of the related work that we have studied before proceeding to our model.

In this Paper **(Crime Analysis using KMeans Clustering**): Author Initially gathered crime dataset and filtered according to the requirements. Exported all crime dataset file to the rapid miner tool and then performed Normalization operation to perform k means clustering on resultant dataset. Finally, performed plot view and got the clusters to form the crime analysis cluster. To analyze this data author used K-means clustering analysis as it aims to partition n observations into k clusters and each observation belongs to the cluster with nearest mean. K means algorithm complexity is O(tkn), where n is instances, c is clusters, and t is iterations. From these cluster

trends it is concluded that crime murder is decreasing from 1990 to 2011. **Accuracy:** Decreasing trend from 1990-2011 was observed.

In this Paper **(Crime Analysis in India using Data Mining Techniques): –** Proposed approach consists of several steps. In first step, cleaning and transformation – K- means algorithm is used to replace the missing values with the mean/mode value of that attribute. K- means classifier classified crime instances into clusters based on the crime characteristics. Author followed Random Forest and Neural Network algorithms used in this approach to test which gives better results for the given dataset. Random Forest was used to rank on the trees on the basis of priority. Neural Networks used for non-linear modelling and recognizing patterns. In this study, Google map marker clustering (GMAPI) is used, and it is very helpful in reflecting a country's crime-prone regions. GMAPI includes characteristics like the latitude and longitude of a location. As per the GMAPI prediction a police department can easily assign more manpower based on the priority rankings. In locations identified with more crimes patrolling will be increased. **Accuracy:** Achieved good accuracy of 99.93 %.

In this Paper **(Crime Prediction based on crime types and using special and Temporal criminal hotspots):** Apriori Algorithm was used in predicting the type of crime pattern, specific location within a particular time. This algorithm was implemented on location and time features. To obtain more frequent patterns constraint-based mining was used by restricting the extraction process on frequent patterns on three item sets- Location, Day and Time. Naïve Bayesian Classifier was chosen as a classifier as it assumes independent effect between the attributes which matches with the dataset. Dataset division was done randomly into 80% training and 20% for testing.

Author has applied same classifier on training data of different states to obtain different models for crime type prediction. Decision Tree Classifier model and sci-kit learn tool were used to model class label values using simple decision rules from data features. **Accuracy:** Achieved 51% of prediction accuracy in Denver and 54% prediction accuracy in Los Angeles.

In this Paper **(Crime Prediction and Analysis using Machine Learning):** Author used various algorithms to verify the accuracy of the dataset by implementing (KNeighbor Classifier, GaussianNB, MultinomialNB, BernoulliNB, SVC and Decision Tree Classifiers) to predict types of crimes committed over Time (Month/ Hour), No of crimes of all types of crime over the whole city of Chicago, Crimes committed across different locations. After feature selection location and month attribute are used for training. The dataset is divided into pair of xtrain ,ytrain and xtest, y test. The algorithms model is imported form sklearn. Building model was done using model. Fit (xtrain, ytrain). After the model is build using the above process, prediction was done using model Predict(xtest). The accuracy is calculated using accuracy_score imported from metrics - metrics.accuracy_score (ytest, predicted). **Accuracy:** The accuracy of the model is 78 percent.

In this Paper **(Crime Prediction Using Decision Tree(J48) Classification Algorithm):** Author used method for crime prediction is Spiral model methodology (which combines the features of the prototyping model and the waterfall model).This methodology consists of data collection, data-preprocessing, building classification model using the training data and evaluation of the generated models using test data. The approach followed for prediction by using J48 is the block diagrammatic approach for the crime predictive system where they do Data collection from crime historical data and new crime data and then data preprocessing in which data is transformed and

filtered. Then data mining was done by using WEKA tool therefore data was set for visualization in which data was represented in tree. **Accuracy:** Decision Tree(J48) Classification Algorithm predicted the unknown category of crime data to the accuracy of 94.25287% .

In this Paper **(Crime Analysis and Prediction Using Data Mining) :** Author discusses about the crime analysis and prevention for identifying, analyzing patterns and trends in crime. Here the author explained an approach between computer science and criminal justice to develop a data mining procedure which helps in solving crime fast. For Crime Analysis, author used MongoDB for collecting the data since the crime data is unstructured data. Naive Bayes Classifier methodology is used for classification. Apriori Algorithm is used for the Pattern Identification. For Prediction they used the decision tree concept. For representing criminal data, they used a graph database called Neo4j. **Result:** Classification for the data is done based on Bayes theorem which is greater than 90% accuracy.

In this Paper **(Predicting Crime Using Time and Location Data) :** Author used machine-learning techniques to analyze the previous crime datasets and predicts the hotspots for crime based on time and location. This paper uses a different algorithm like Random Forest, Decision tree and different ensemble methods such as Extra Trees, Bagging and AdaBoost to evaluate the accuracy given by each algorithm. The Methods used in this paper to train the dataset are Random Forest, Decision Tree and different assembly methods such as Extra Trees, Bagging and AdaBoost. Bagging gives the accuracy of 99.92%. AdaBoost gives the accuracy of 74.78%.

Proposed Method

Modeling is of various kinds. Predictive Modeling is used to analyze the past data and predict the future outcome. In our project we implemented modeling on NYPD Dataset. NYPD maintains monthly crime data based on police incident reports across New York. The data is saved on the NYPD website with each month corresponding to an Excel file. Different modeling algorithms in Python have been used for this prediction.

Data Preparation:

Initially, we started by filtering the unnecessary duplicate or missing values of records from the dataset of NYPD which is available in Nypd.gov website. Then we exported and uploaded all the records to excel sheet and then to workspace by applying filters on specific columns such as vic_age, vic_sex, suspect_age, suspect_sex, boro_nm etc.

Data Cleaning:

After completing the process of data preparation, we verified all the columns and rows of the dataset objects for null values and duplicates. In this phase, we started off by filling the missing values of offense_description by mapping the offense_code and its corresponding offense_description. We then found that there were still missing values in offense_description. The reason behind this was one of the offense codes does not have any description. Hence, we removed everything related to this offense code. The same procedure was followed with the jurisdiction description. Later, we treated the features by changing their data types from object to integer , category and time based on their characteristics.

EDA:

EDA(Exploratory Data Analysis) is one of the data analysis methodologies used to summarize dataset characteristics with statistical numbers and graphs. In our project we implemented EDA on NYPD Dataset. NYPD maintains monthly crime data based on police incident reports across New York. The data is saved on the NYPD website with each month corresponding to an Excel file. Pre-processing algorithms in python were used to connect directly to excel files in the web source in order to retrieve the required information.

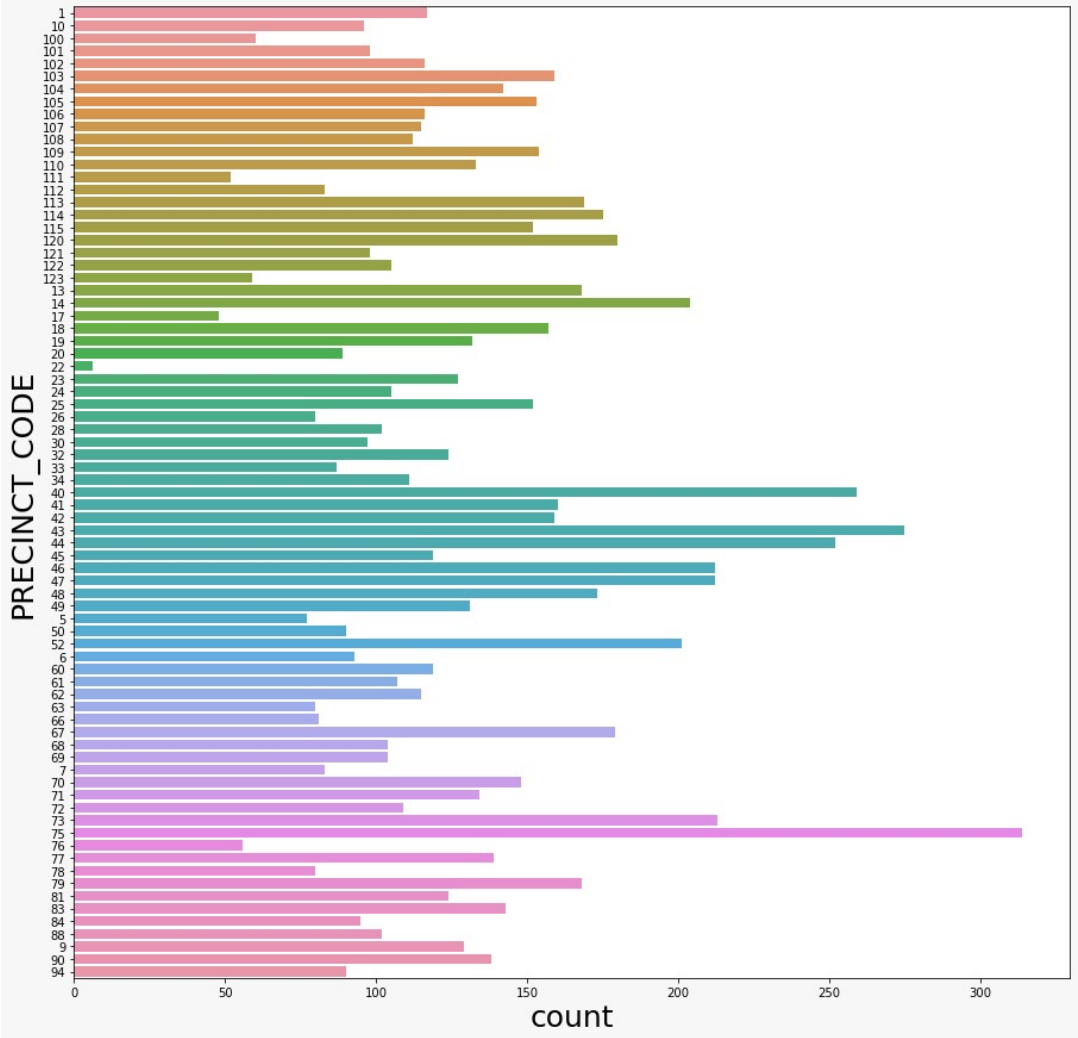
A lot of statistical and machine learning algorithms could be used as an EDA technique. Here, we picked up a few primary techniques to illustrate.

Exploratory Method		**Visualization Technique**
Univariate	Categorical	Count Plot
	Numerical	Histogram
Bivariate	Categorical 'Vs' Numerical	Box Plot, Scatter plot, Pie chart, violin plot, bar plot
	Numerical 'Vs' Numerical	Line Plot or Replot

Data Exploration from Different Perspectives:

1. **Unique Value Count:**

One of the first items that was used during data exploration is to see how many unique values there are in the categorical columns. This gives you an idea of what the data are about. The unique number of categorical columns in the dataset of the precincts is seen here.

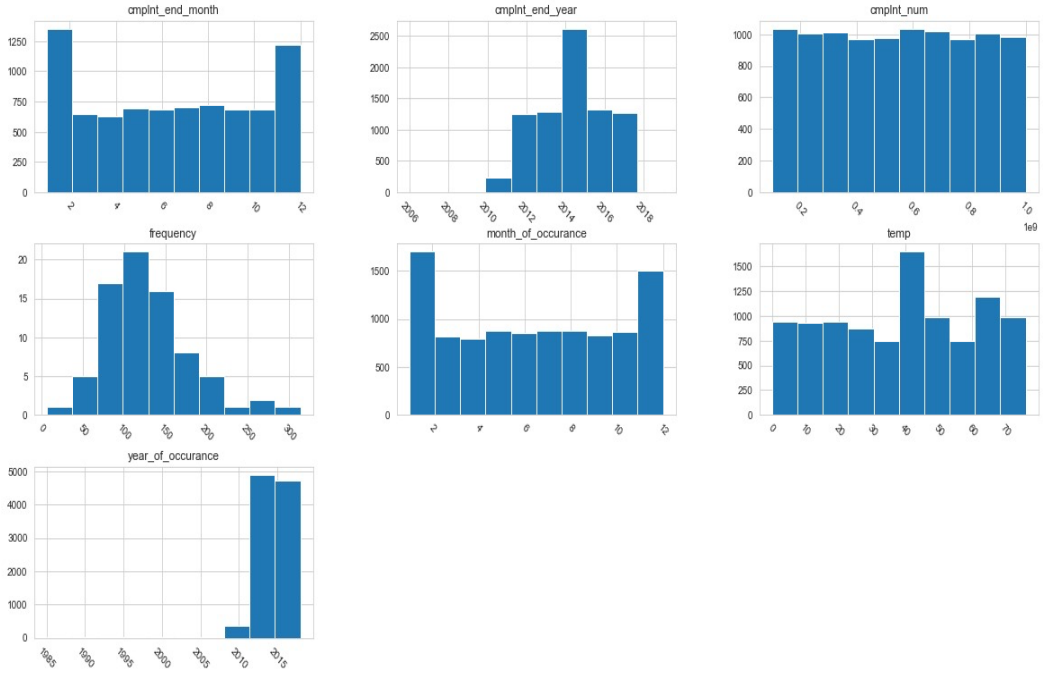


The categorical column with maximum number of unique values is Precinct_code-75 which has received highest number of complaints or unique values that are approximately 350. This column has 77 unique precinct codes among which the highest number of

complaints filed are in precinct-75.This category has minimum value of precinct noted as-precinct-22. If we observe maximum number of unique values is in “precinct-75” column, which means that most of the valuable information required for research is mainly around different complaints in precinct-75.

2. **Histogram:**

It offers details on the spectrum of values in which most values fall. It also provides details as to whether there is a skew in the results. If we make a histogram in the **month_of_occurance** column displays the highest and lowest number of complaints per month. Similarly, **cmplnt_end_month** column describes in which month maximum and minimum number of crimes that have been successfully completed by suspects. The temp is the Label encoded column of **Addr_pct_cd** which is number of complaints at specific precinct. **Frequency (count(temp))** chart shows the highest and lowest frequency values.

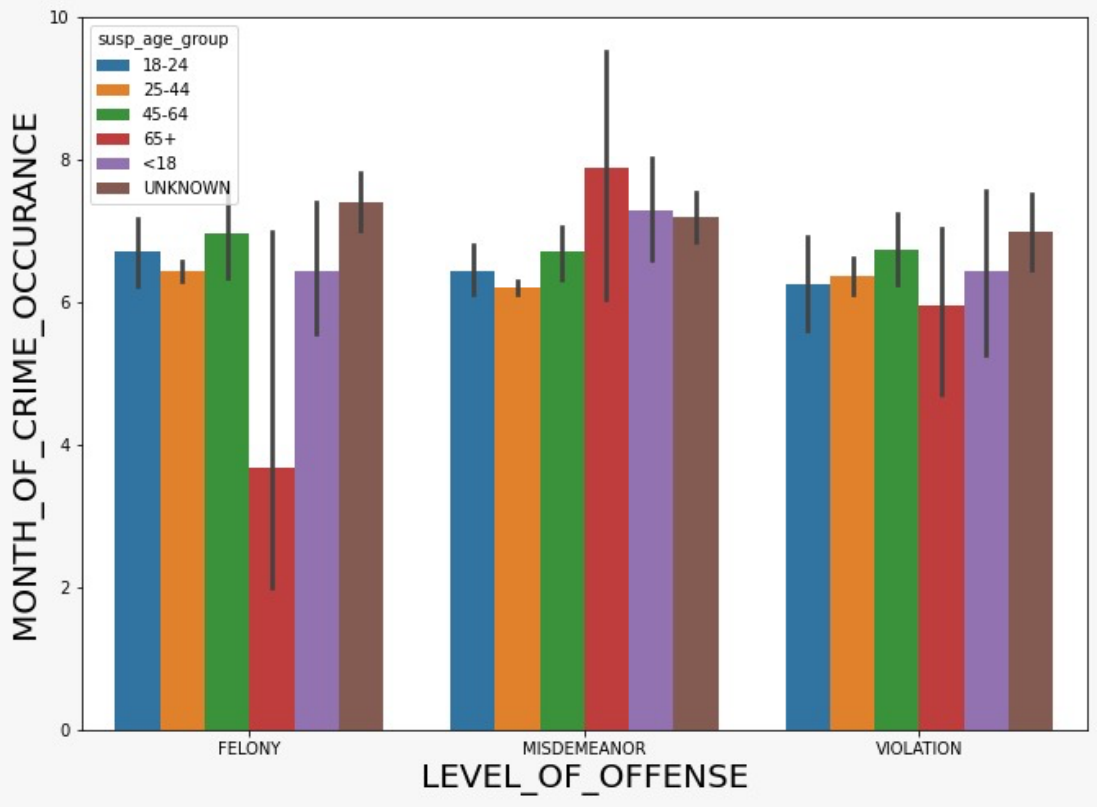


In the above Graphs Frequency graph represents normal distribution curve. Hence, implementation of Linear regression model is possible.

3. Cluster size Analysis:

Grouping data together makes it possible for us to get that high-level perspective. Data groups allow us to look at groups rather than individual data points first. This grouping is also called as clustering or segmentation. As a first step in segmentation, cluster size analysis reveals how data can be separated into various categories.

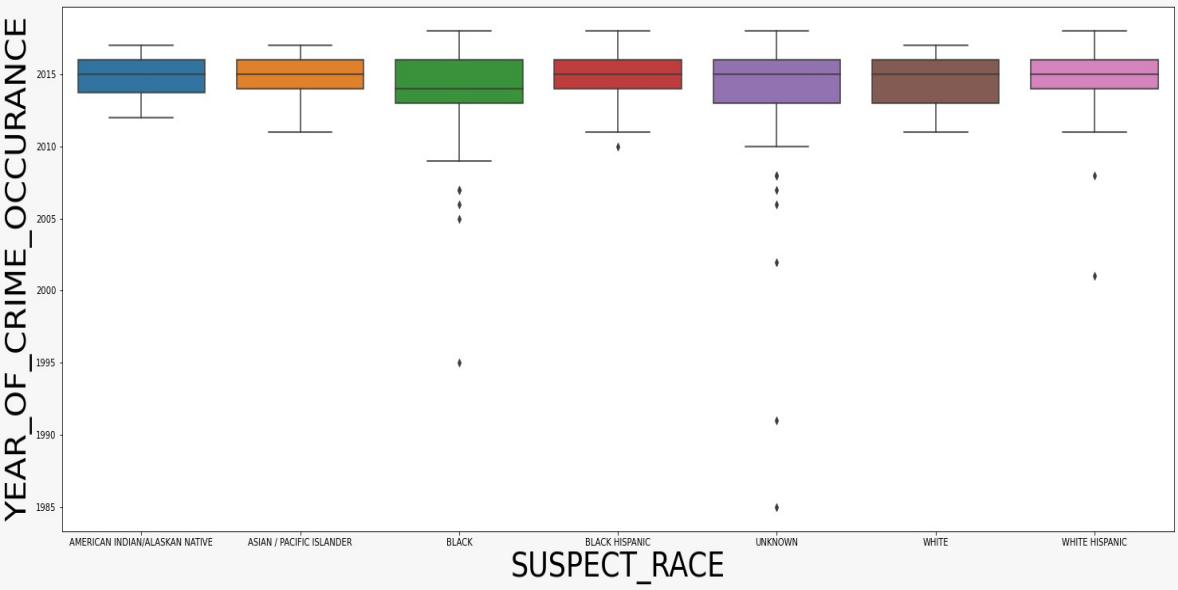
As we may observe, we split all data into three classes, we have clusters that are more or less of the same size. Each group represents different levels of offenses such as FELONY,MISDEMEANOR and VIOLATION. In the below graph we can observe highest level of crime occurred in the month august and the age group of the suspect was 65+ and the level of offense was MISDEMENOR. Lowest offense was also done by the suspects with age group 65+ (offense type= FELONY).



4. Outlier Perspective:

By using outlier detection method, we tried to find out the unusual data which is outlier detection/ Anomaly Detection. Outliers represent unusual, rare, anomaly or exceptional. Outliers analysis also helps to enhance the quality of exploratory data analysis.

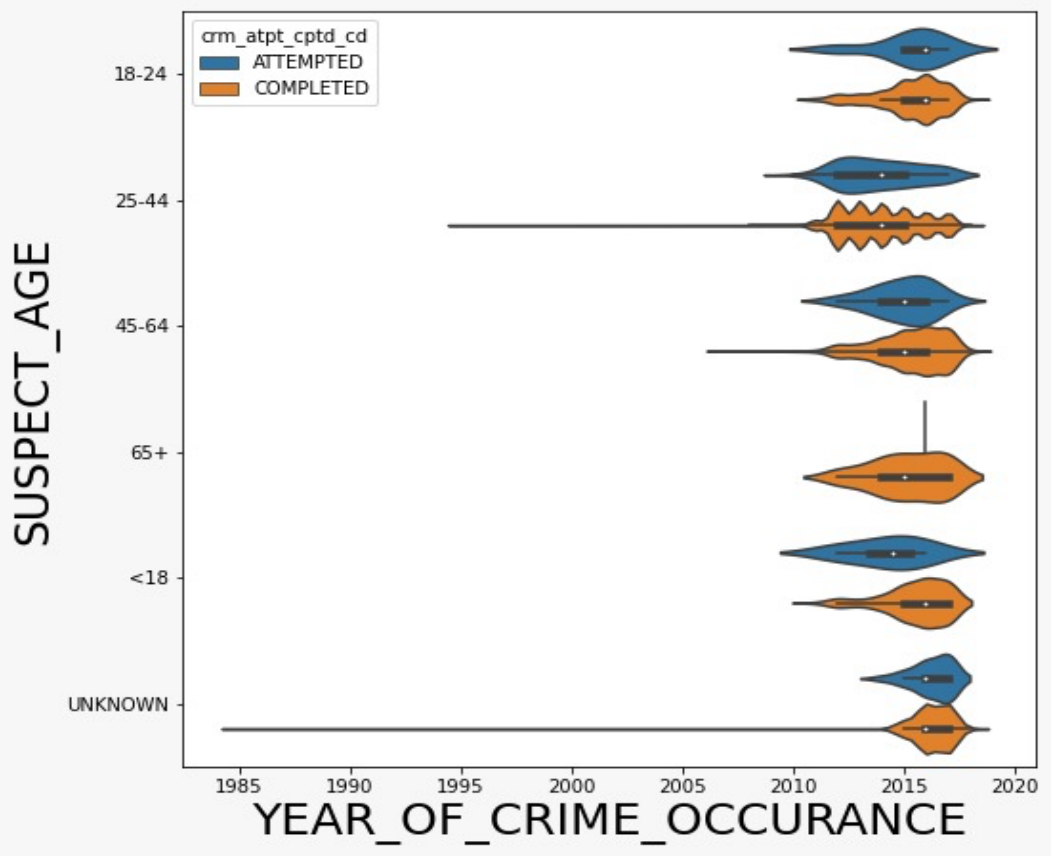
The below graph represents box plot between the columns of **Year_of_crime_occurance** (Numerical)and **Suspect_Race(Categorical).** The below box plot explains the Unknown has maximum outliers as 6. Next highest outliers are in Black race. Least outliers are around white Hispanic race. Also, it can be deduced that most of the crimes were occurred between 2012-2018.



Violin Plots:

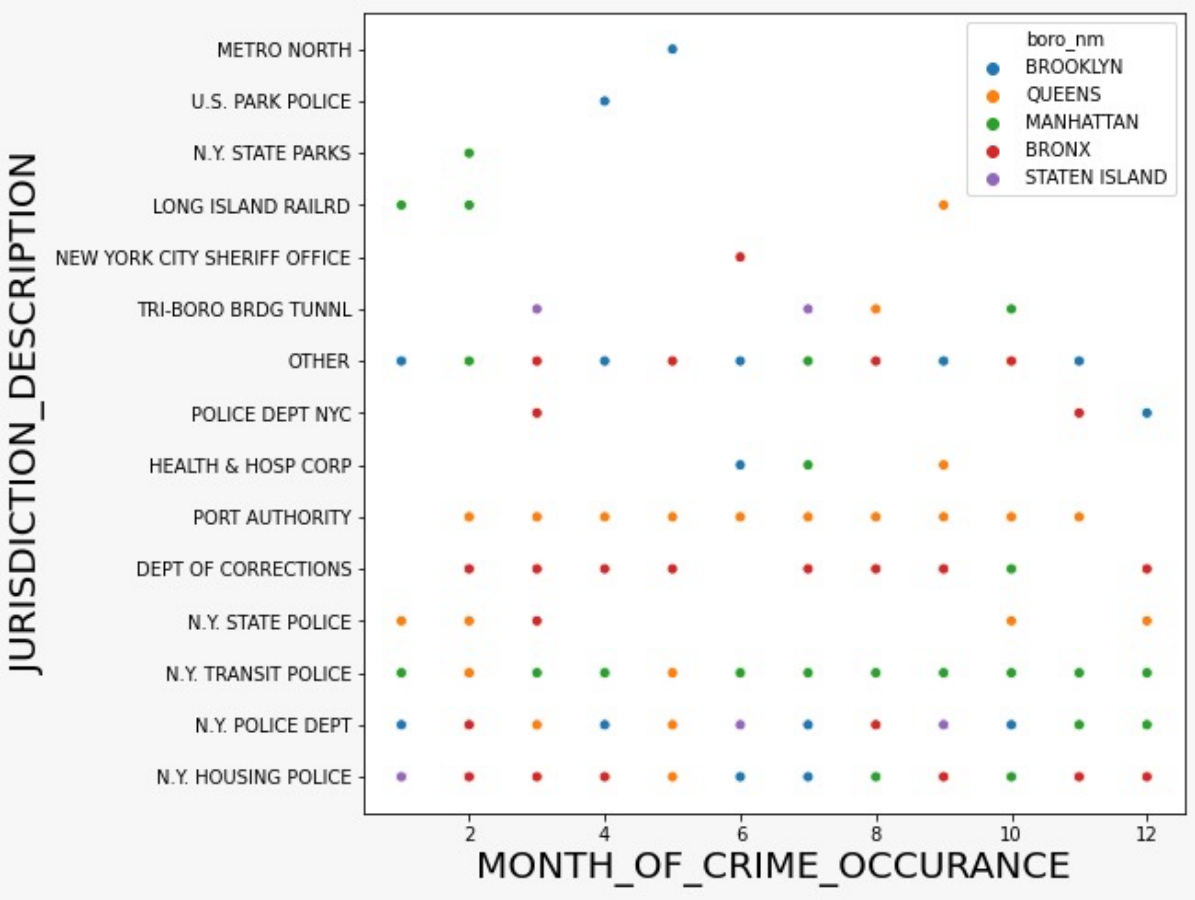
Violin plot is a method for plotting numeric results.Here we used one numerical and one categorical value to plot violin graph. It is very similar to a box plot, with the addition of a rotated kernel density plot on either side. Also, Outliers are defined with a solid black line.In the below violin chart Suspect age is in the Y-Axis and x axis as Year-Of_Crime _Occurance.

Blue label denotes the Number of crimes attempted which is very high between the period of 2012-2017. Yellow label denotes the successfully completed crimes by the suspects with the age group of 65+. Also, we can observe that all these suspects were completed the crimes successfully that they have attempted.This insight might help while doing data modelling.When it comes to outliers most of the outliers were detected during the years 1995 and 2010 at the age group of the suspect is around 44+.



5. **Scatter Plots:**

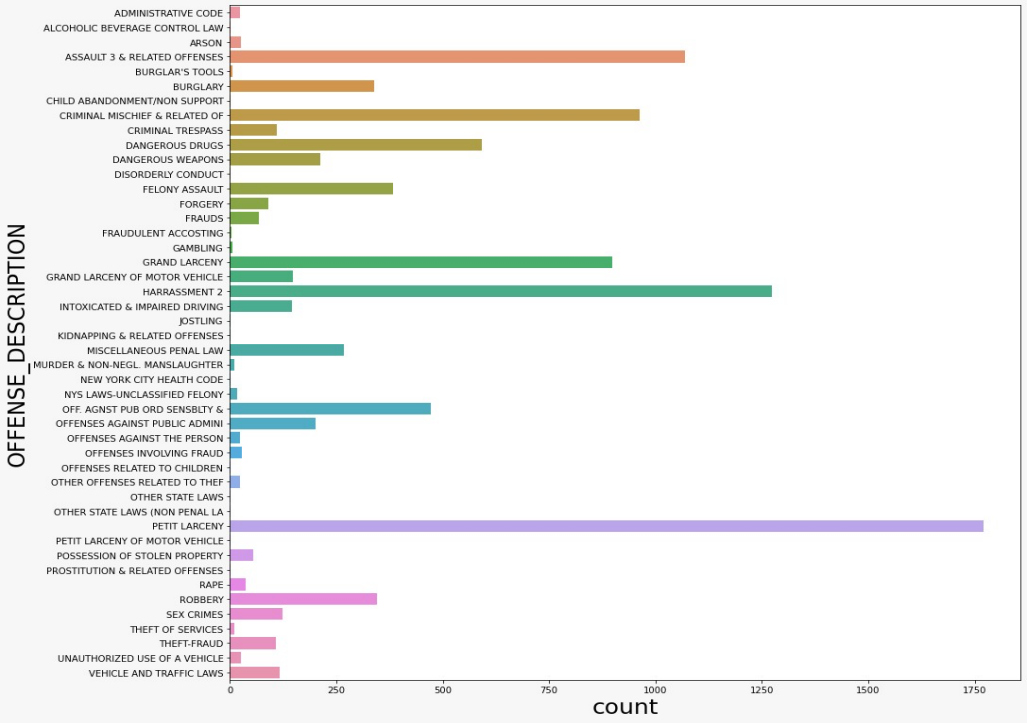
Scatter plot was built between Jurisdiction and Month of crime. If we look at the spread of the crimes in the scatter plot, we can deduce that most of the crimes were occurred in March month in BRONX boro jurisdiction. Even though there are same number of dots in February but there were more outliers in the data. One more observation can be made by looking at the port authority orange dots in the graph which is detection of continuous crime from the months between February and November.



6. Categorical Perspective:

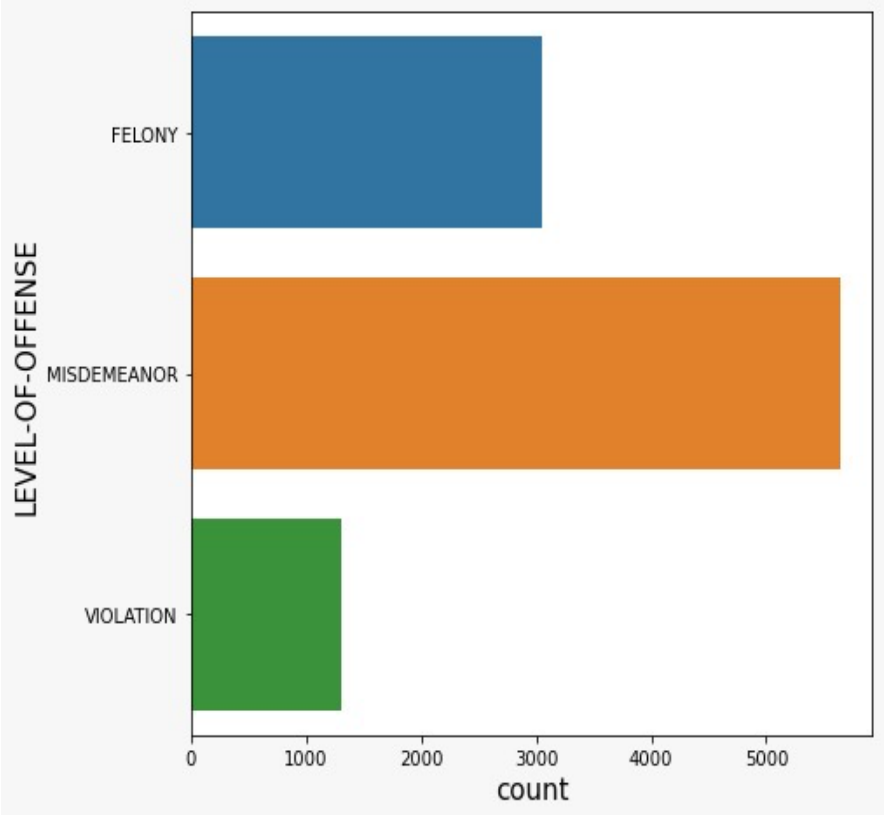
One of the very first things that could be useful during data exploration is to see how many unique values there are in the categorical columns. This gave us an idea of what the data are really about.

a) **Offense Description:** The unique number of categorical columns in the dataset of the unique Offense Descriptions is seen here. X axis represents the count of offenses and Y – axis denotes the Offense description numbered from 1 to 45 labels. Highest number of crimes were for the offense **petit Larceny** which is more than 1750.

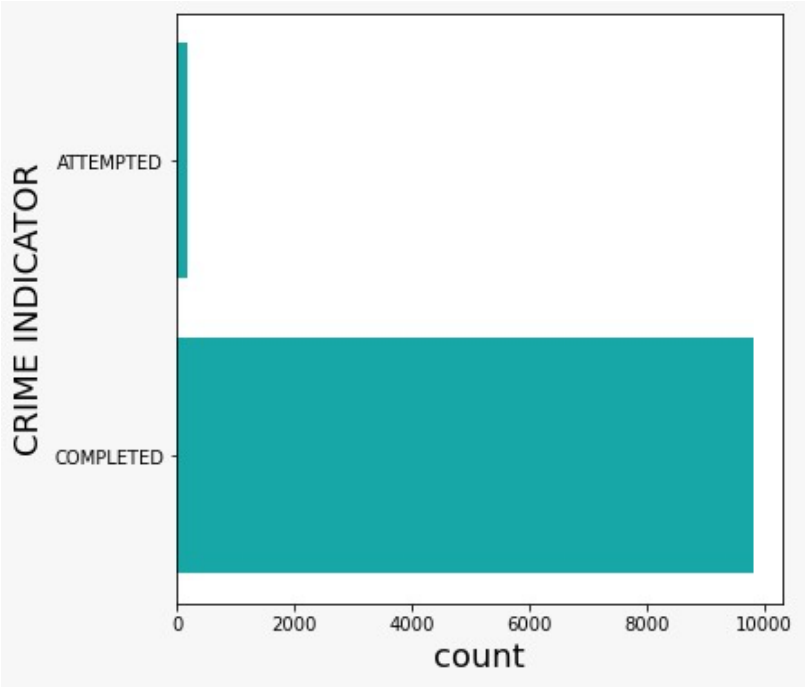


b) **Level_Of_Offense:** The Below graph represents the Level of Offense as a unique categorical value. By looking at the graph we can observe that **Misdemeanor** was the most (more than 5000) committed offense among all these offenses listed. As you can

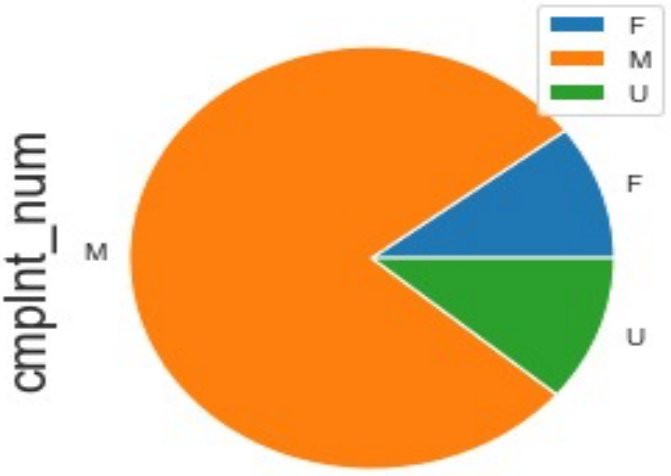
see the green bar in the graph which has Least leavelof offense was recorded as Violation.



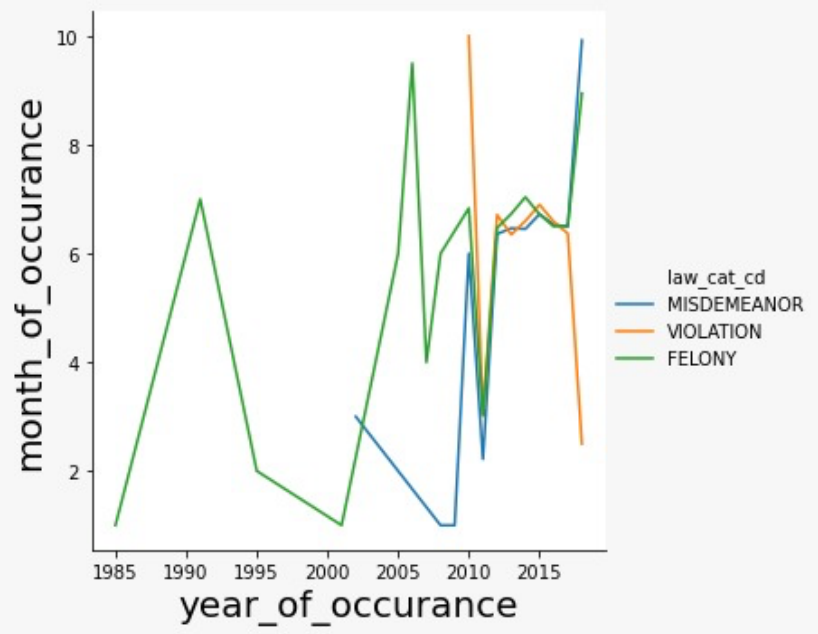
c) **Crime Indicator:** The below graph represents the crimes which have been completed or attempted. Here , the Crime indicator indicates that number of crimes which have been successfully completed or attempted by the suspect. We can observe that a greater number of crimes have been successfully implemented by the suspect with the count of approximately 9800, whereas the crimes that have been interrupted or failed were less than 100.



7. **Categorical Vs Numerical :** The below pie chart is plotted between suspect sex (Categorical)and complaint number (Numerical). Suspect_sex feature has three categories namely, Male(M), Female(F) and Unknown(U). The reason for having Unknown category is because we do not have enough information about suspect gender. We observed from the graph that a greater number of complaints have been filed against Male category.



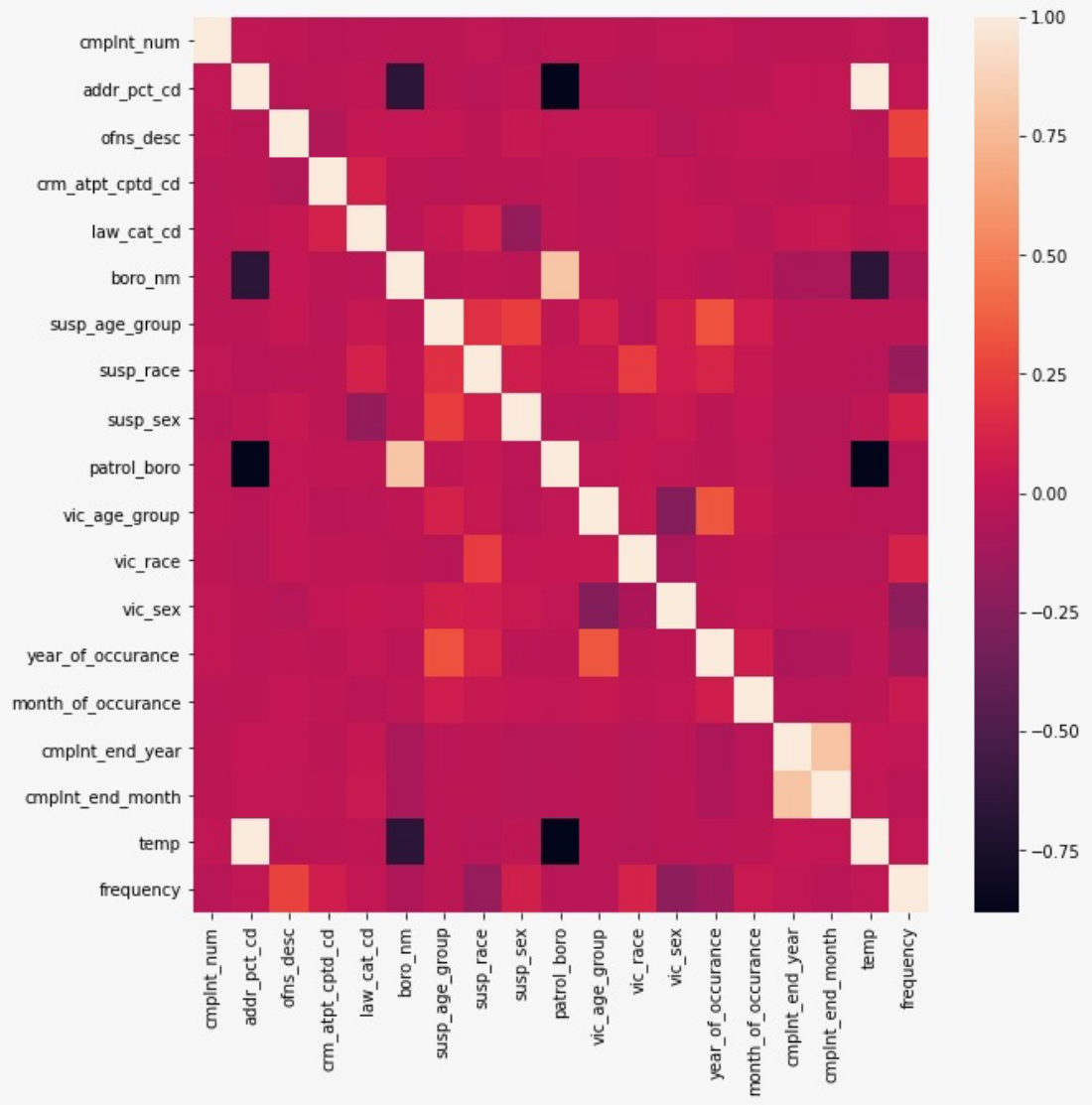
8. **Numerical Vs Numerical:** In this graph X-axis and Y-axis were generated from a single feature called “Cmplnt_fr_dt”. The feature 'cmplnt_fr_dt' has yyyy-mm-dd format, by using DatetimeIndex() function in pandas we divided this into two separate columns as year and month. We also used ‘Law_cat_cd’ (level of offense) attribute as a hue in order to get more insights about the data. The observations from the graph are that FELONY was the most occurring level of offense from 1985 to 2015. Highest value of VIOLATION was observed in October 2010 but decrease in VIOLATION can be observed after 2015.It can also be deduced that MISDEMEANOR trend is increasing after 2011.



9. **Correlation Between Categorical Features:**

Heat map is a very useful data analysis technique for discovering a connection between Categorical variables. There are several categorical columns in our NYPD crime dataset. The below is correlation map by using Heat map. High correlations are shown with light color and low correlations are shown in dark colors like black and dark red. In the below map the top correlations are between **Cmplnt_end_year** and **Cmplnt_end_month**

. The medium correlation is between **Patrol_boro** and **Boro_nm**, **Ofns_desc** and **frequency**, **Year_of_occurance** and **susp_age_group**. Many other categories come under low correlation for example **vic_age_group** and **vic_sex**.

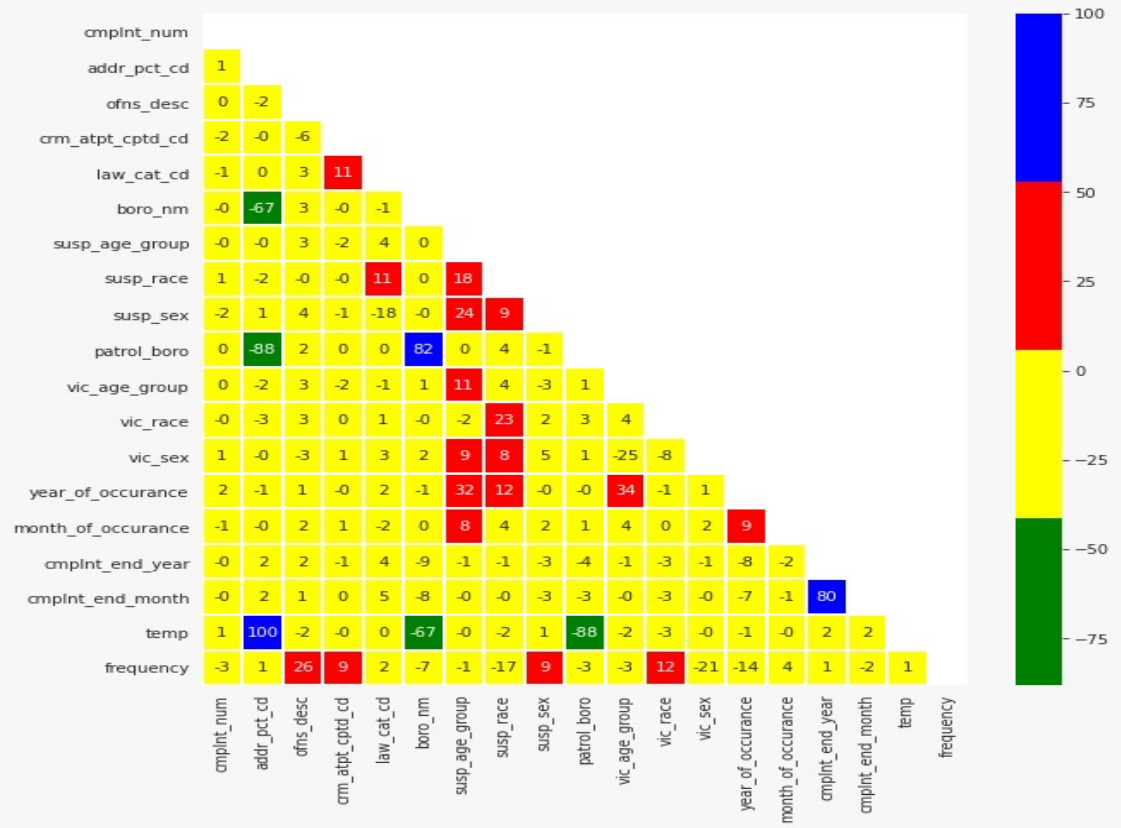


Heat Map Mask: Correlation tests the dependency of two variables. It also tests "how two variables dependent on each other " and "how strongly they connect" means an

increase in one factor and an increase in another. To verify this dependency, we implemented heat map mask by using seaborn library.

Seaborn Heatmap feature has a mask statement that lets us to select the elements from the input data frame. In our example, we want to mask the upper triangular elements to make the lower triangle correlation map.

In the below heat map dark blue and dark red represents Positive correlation and Negative correlation respectively. Hence, Green color represents Negative correlation between the features and blue represents positive correlation between the selected features.



Modelling:

Logistic Regression:

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical).

Logistic Regression has different parameters, but in our model, we used specific parameters such as ‘C’ and ‘penalty’ . Penalty is used to specify the regularization. Usually ‘l2’ is the default penalty, which is used by most of the solvers, if we are not specifying any penalty then no regularization is applied. C parameter is used to control the penalty strength.

XG Boost:

XG Boost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.).

In XG Boost, the parameters used for the model are ‘n_estimators’, ‘max_depth’, ‘reg_lambda’ and ‘learning_rate’. The **n_estimators**: Number of gradient boosted trees. Equivalent to number of boosting rounds, **max_depth**: Maximum tree depth for base learners, **learning_rate**: Boosting learning rate and **reg_lambda** : L2 regularization term on weights.

KNN:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

The number of neighbors (n neighbors) is the most significant KNN hyperparameter. It measures values between 1 and 21. The optimal number of neighbors for our project are 19. We also have some parameters in KNN such as weights and metric.

Decision Trees:

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

The parameters used in Decision Tree model are like Random Forest, but we have an extra parameter namely 'max_depth' . If no max_depth values are specified, then the nodes are expanded until all leaves are either pure or till, they are less than the min_samples_split values.

Random Forest:

It is a bagging-based algorithm with a key difference wherein only a subset of features is selected at random. In other words, every interviewer will only test the interviewee on certain randomly selected qualifications (e.g. a technical interview for testing programming skills and a behavioral interview for evaluating non-technical skills). The parameters used in Random Forest are 'n_estimators', 'min_samples_split', 'min_samples_leaf' . The number of random features to sample at each split point is the most significant parameter. A log scale of 10 to 1,000 might be good value. Until no further change is seen in the model, the number of trees should be increased.

Experimental Results

In this section, we summarize the key results that we obtained from different modelling classifier techniques on the two datasets (training and testing). The dataset used for the modeling includes 60,000 records with 24 features. The dataset is divided into training dataset (70%) and testing dataset (30%). The training dataset was used for training the models and the test dataset was used to test the accuracy of the model.

Accuracy:

We got an accuracy of 99% for all our models for both training and testing data by using all features of the dataset. Below table represents the training and testing data accuracy of different models.

Modeling Technique	Train data accuracy	Test data accuracy
Logistic Regression	99.96	100
Random Forest	100	99.99
XG Boost	99.97	100
K-Nearest Neighbor(KNN)	99.74	99.99
Decision Trees	100	100

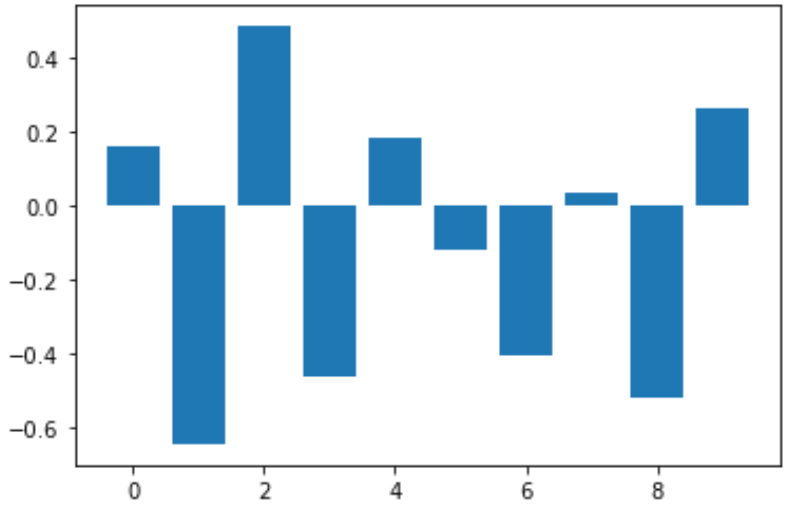
Precision and Recall :

Precision and Recall are the evaluation metrics of a model. Precision helps us to know what proportion of the positive identifications are correct and Recall helps us to know what proportion

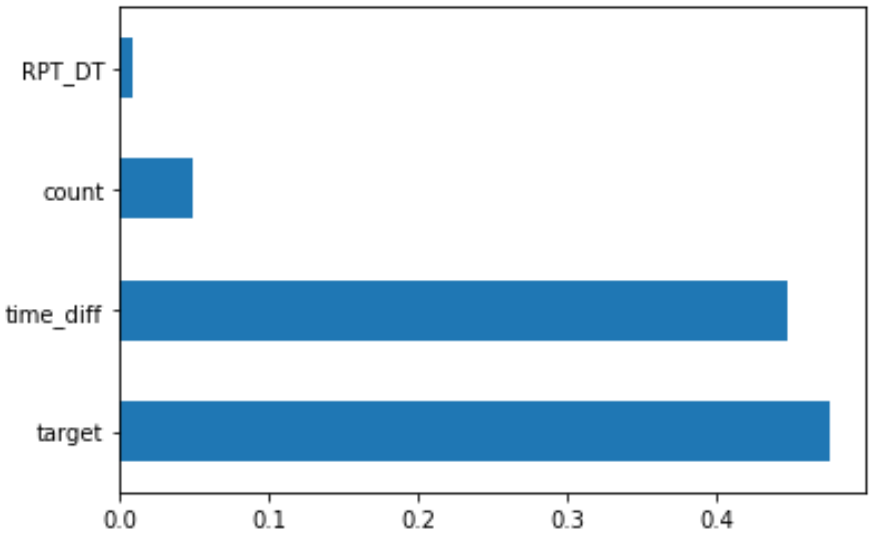
of actual positives are correctly identified. For the dataset and features we used, all the models had a precision and recall of 1.

Feature selection and Feature importance graphs for each model :

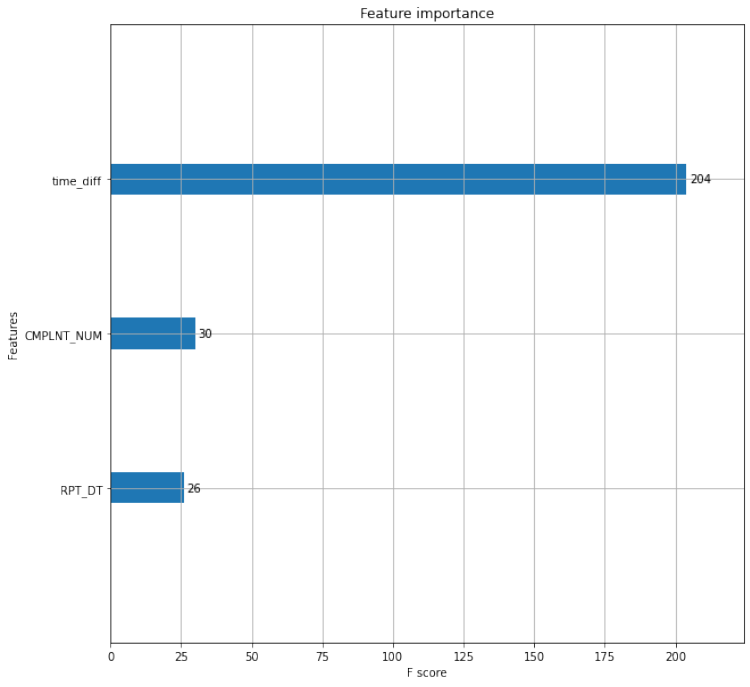
1. **Logistic Regression :** The below are the F-score for each feature used in the model. We can see that the highest score is recorded for Feature 2.



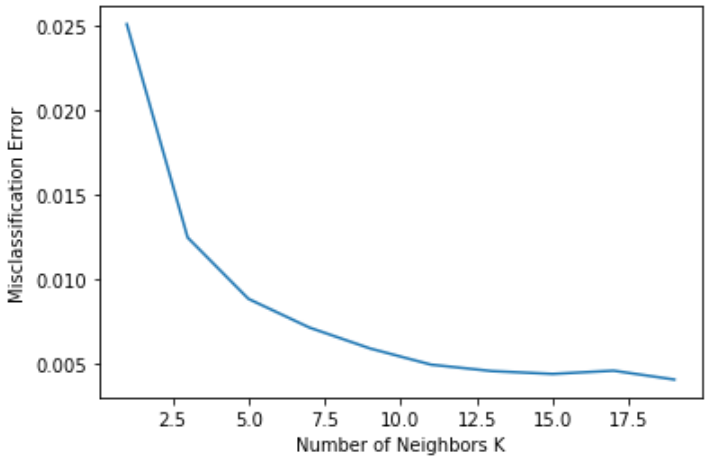
2. **Random Forest :** In the below graph, we can see that the target has the highest score of more than 0.4



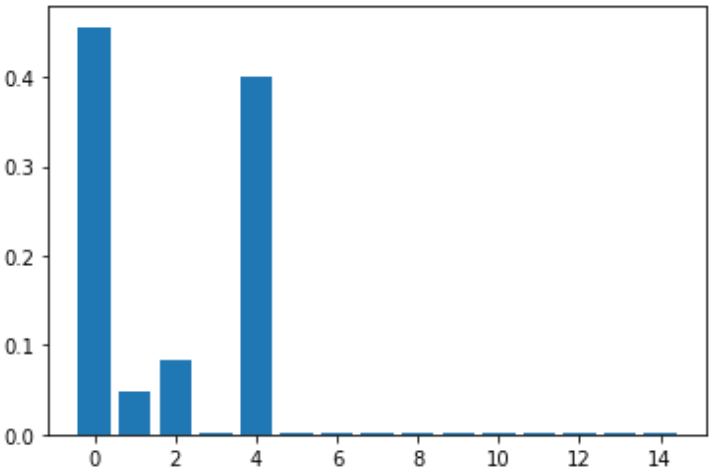
3. **XG Boost:** From the below graph it can be noted that the time_diff has more importance than the other features.



4. **KNN :** In the below graph, we can observe that the misclassification error decreases as the number of neighbors k value increases.



5. **Decision Trees :** From the below values we can see that the highest score is recorded for the Feature 0.



Discussion

Though we researched a lot of papers the below are some papers which are closely related to our work. We took these models as our inspiration to build a predictive model which bridges the gap and achieves a good accuracy. The below are the outcomes we noticed from these papers.

In Paper 1(Jyoti Agarwal, 2013), the author showed the decreasing trend, but they could not predict about the future occurrences of the crime. In Paper 6(Shiju sathyadevan, 2019),have shown the trend and predictability of crime on a specific day using Decision tree, but it would have been more precise if a specific location were chosen. In Paper 3(Tahini Almanie, 2015), Author clearly explained the approach followed for analysis, pattern recognition and other details like dataset divisions. Author also gave the detailed narration of the implementation methods of algorithms on the given dataset. It would be efficient and helpful if this prediction has done with dependent attributes as well. This prediction only works with the Independent attribute set of data. In Paper 5(Emmanual Ahishakiye, 2017),research was conducted on the 1994 instances of data set, which can be considered as weak data analysis. It would have been more precise if the model were trained with massive data which would have been resulted in a better model. The data set used for the research was a numerical dataset. Hence, it does not guarantee the same accuracy for categorical dataset.

In Paper-2(Deepika K.Kk, 2018), The research achieved very good accuracy by using good approach in classifying patterns. But future predictions were not mentioned in this paper and also, in this paper author used K-means clustering algorithm but it is not possible always to specify number of cluster (K) at the beginning as we cannot calculate k value on categorical attributes.

In Paper 4(Bharati, 2018), Though the research was implemented using multiple algorithms the highest accuracy obtained was 78% using K neighbor classifier. The model would have been more accurate if the research were focused on improving the one specific approach. Among all the papers, (Jeisa Yuki, 2019)this paper has acquired highest accuracy with the implementation of Bagging algorithm i.e., 99.92%by using a data set of 6 million instances.

Conclusion and Future work

This paper uses five different types of algorithms to predict occurrence of crime in a given precinct. The predicted results are closer to the actual results as all the algorithms showed an accuracy of 99%. Hence, we can say that the given dataset provides maximum accurate results with high accuracy using different classifiers.

This work can be extended by predicting the exact time and location of the occurrence of offense. By using machine learning models, we can also predict the occurrence of number of crimes per precinct. In an advance level we can also try to predict which gender and what age group may most likely commit a crime.

Appendix

Repository Address

https://github.com/vmari2/Crime-Prediction-of-NYPD.git

References

1. Bharati, A. (2018). Crime Prediction and Analysis using Machine Learning. *International Journal of Engineering and Technology*.
2. Deepika K.Kk, s. v. (2018). Crime Analysis in India using Data Mining Techniques. *International Journal of Engineering and Technology*.
3. Emmanual Ahishakiye, D. T. (2017). Crime Prediction using Deciion Tree Classification . *Internationall journal of computer and Information technology*.
4. Jeisa Yuki, M. M. (2019). Predicting crime using time and location data. *International Journal of Computer application*.
5. Jyoti Agarwal, R. S. (2013). Crime Analysis using KMeans Clustering. *International Journal of Computer Applications*.
6. Shiju sathyadevan, D. s. (2019). Predicting crime using time and Location data. *International Journal of computer application.*
7. Tahini Almanie, R. M. (2015). Crime Prediction based on crime types using spacial and Temporal criminal. *International Journal of Data Mining and Knowledge Management Process*.
8. www.wikipedia.org
9. https://towardsdatascience.com/
10. https://scikit-learn.org/stable/
11. https://www.saedsayad.com/