San Francisco Crime Classification

Bharath Janapareddi (HC4765@wayne.edu)
Kavya Nagaraju (HC3344@wayne.edu)
Priyanka Patil (HC6653@wayne.edu)

Abstract

Crime has been prevalent in our society for a very long time and it continues to be so even today. The San Francisco Police Department has continued to register numerous such crime cases daily. Predicting the crime and the crime rate is one of the essential factors in improving the efficiency of the police department and reducing threat for the public. Different machine learning approaches were conceptualized and implemented for predicting future crimes based on a given set of geographical and time-based features. Various classification techniques like Random Forest, KNN, Decision Tree are used. Lastly, our results are experimentally evaluated and compared against previous work. The proposed model finds applications in resource allocation of law enforcement in a Smart City.

***Keywords**:* Crime Classification; Random Forest; KNN; Decision Tree

1. Introduction

San Francisco first boomed in 1849 during the California Gold Rush. The city then expanded both in terms of land area and population. As a result, the crime rate and civil problems also proliferated. Crime continues to be a threat to us and our society and demands serious consideration if we hope to reduce the onset of the repercussions caused by it. However, San Francisco of today is different than what it was at its beginning. Now it is well known for the Silicon Valley and the tech giants than that for its criminal history. With the increase in the crime rate, it is very difficult to predict the crime and prevent it from happening. The city of San Francisco is one amongst the many to have joined this Open Data movement. The data scientists and engineers working alongside the San Francisco Police Department (SFPD) have recorded over 100,000 crime cases in the form of police complaints they have received. With the help of this historical data, many patterns can be uncovered. This would help us predict the crimes that may happen in the future and thereby help the city police better safeguard the population of the city.

1.1 Motivation

The motivation behind taking up this topic for the research is that every aware citizen in today's modern world wants to live in a safe environment and neighborhood. However, it is a known fact that crime in some form exists in our society. Although we cannot control what goes on around us, we can try to take a few steps to aid the government and police authorities in trying to control it. The SFPD has made the Police Complaints data from the year 2003 to 2015 available to the general public with more than 800,000 observations. Hence, taking inspiration from the facts stated above, we decided to process this data provided and analyze it to identify the trends in crime over the years as well as make an attempt to predict the crimes in the future.

1.2 Problem Formulation

Our work can be very well explained in two distinct parts.

1. Performing exploratory data analysis for our dataset.
   We utilize this provided crime dataset to perform exploratory data analysis to observe existing patterns in the crime throughout the city of San Francisco. We also study the crime spread in the city based on the geographical location of each crime, the possible areas of victimization on the streets, seasonal changes in the crime rate and the type, and the hourly variations in crime.

2. Building a prediction model to predict the type of crime.
   Our goal is to build a prediction model using various classification techniques in the crime categories. Therefore, predicting the crime that can occur based on geographical conditions. This will help SFPD plan their patrol and contribute their services to the smart city effectively.

2. Dataset Description

The San Francisco Crime Police Department has made the Police Complaints data from the year 2003 to 2015 available to the general public with more than 800,000 observations and it has the following features.

Variable(s) (9)	Description of the variable
Dates	The timestamp of the crime recorded
Category	The category of crime records
Descript	A short note on the crime
DayOfWeek	The day on which the crime took place
Pddistrict	The police department, under which the crime is reported
Resolution	The status of the crime, resolved or unresolved
Address	The address of the crime scene
X	The latitude of the crime scene
Y	The longitude of the crime scene

2.1 Data Cleaning and Preprocessing

As mentioned, the data consists of 878049 observations of 9 variables. The values are very detailed and do not contain any null values. However, it is hard to determine the relationship between the features and the crime classes. Hence additional information is taken from another dataset namely 'zip code'. An Inner join was performed on the San Francisco dataset and zip code data. The 'Season' and 'Hour' of the day was obtained from the Dates variable in the dataset. The Geohash library was used to obtain the zip codes of each location based on the latitude and longitude information.

The final dataset consists of 230512 observations of 14 variables and it has the following features.

Variable(s)	Description of the variable
Dates	The timestamp of the crime recorded
Category	The category of crime records
Descript	A short note on the crime
DayOfWeek	The day on which the crime took place
Pddistrict	The police department, under which the crime is reported
Resolution	The status of the crime, resolved or unresolved
Address	The address of the crime scene
Lat	The latitude of the crime scene
Lng	The longitude of the crime scene
Season	Seasons information of the month
Geohash	To encode latitude, longitude & grouping nearby points on the globe
Zip	Zip-code of the area where the crime was reported
Population	The total population of area zip code covers
Hour	The hour at which the crime took place

2.2 Feature Extraction

There a lot of features like Address, Time, Date, X, and Y which can be transformed into new features that hold more meaning as compared to the existing ones. The Time feature is in the Timestamp format. It would be interesting to observe patterns in crime by the hour. Hence the Hour field is extracted from

the Time field. The original crime dataset has 39 types of crime recorded. Data reduction has done on the variable category. Since we are trying to predict the future occurrences of crimes, it is essential to have categories about actual criminal activities. But the above labels do not provide any additional information to help us achieve our goal. Thus, these categories are completely filtered out from our dataset. We have notified in the below table about the crime categories. The Geohash codes were extracted from the latitude and longitude of san Francisco crime. Zip-code of the region where the crime was reported. The whole year is divided into 4 groups based on the seasons. This helps in classifying the crimes based on when the crime has occurred the most.

New Crime Category	**Original Categories of the crime**
Theft	Larceny/Theft, Vehicle Theft, Burglary
Sexual Offences	Sex Offenses Forcible, Sex Offenses Non-Forcible, Pornography/Obscene Mat, Prostitution
Public Order	Drunkenness, Suspicious Occ, Bribery, Driving Under The Influence, Recovered Vehicle, Bad Checks, Loitering, Disorderly Conduct, Liquor Laws, Trespass, Weapon Laws
Assault	Robbery, Kidnapping, Assault
Drug Offences	Drug/Narcotic
Property Crime	Trea, Embezzlement, Stolen Property, Vandalism, Arson
White-collar Crime	Fraud, Forgery/Counterfeiting, Secondary Codes
Victimless Crime	Gambling, Runaway
Suicide	Suicide, Family Offenses, Missing Person, Extortion
Other	Warrants, Other Offenses, Non-Criminal

3. Exploratory Data Analysis

The Exploratory Data Analysis was performed on all the features to summarize the main characteristics of the dataset. This is the target label/crime we want to predict. We have 39 crime categories (i.e. classification classes). The distribution of the training sample is very skewed as you can see in Figure-1.

Figure-1: Graph for the Crime Category

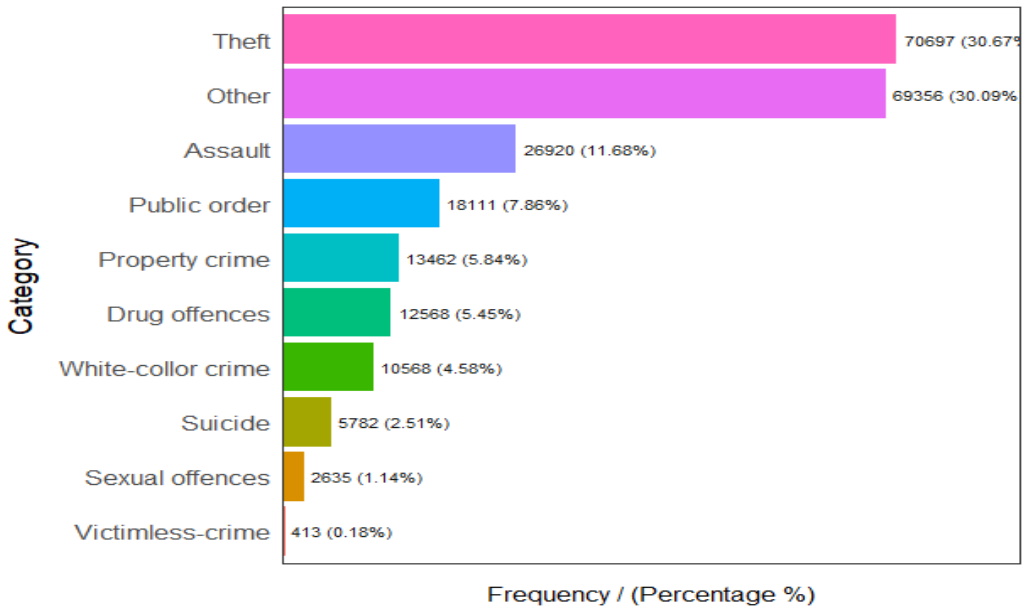
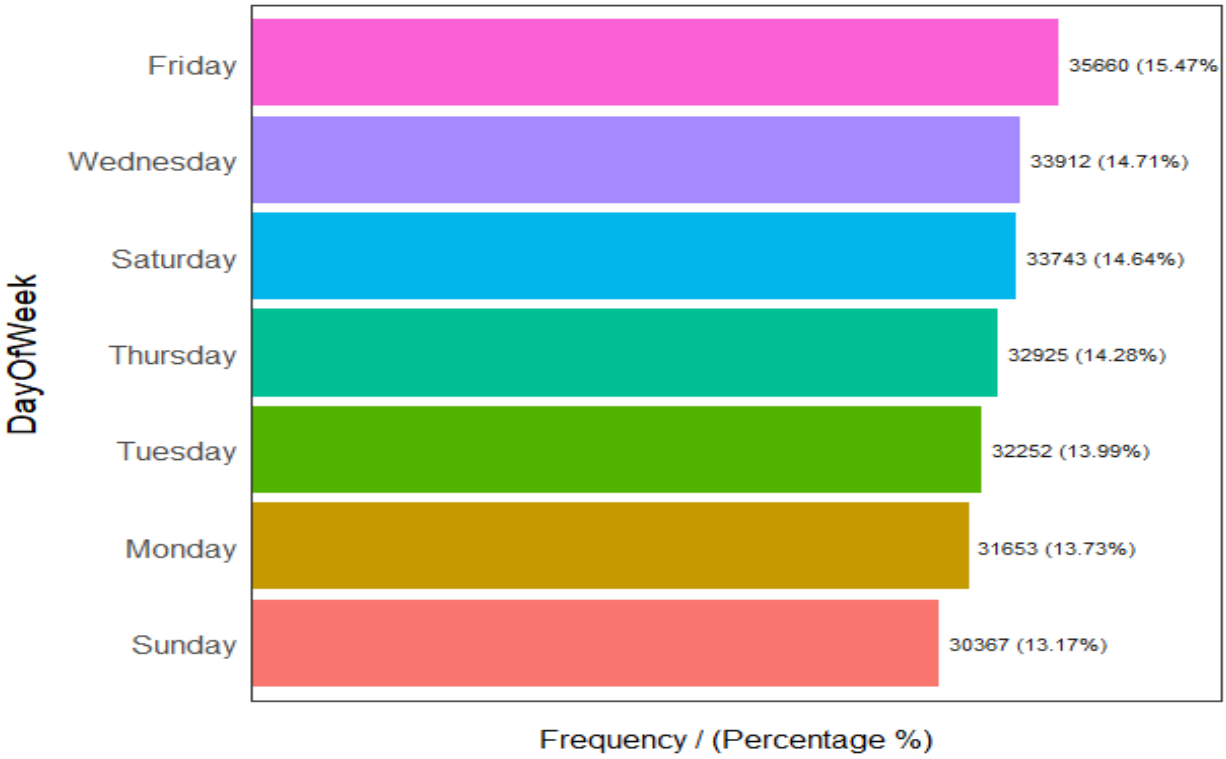


Figure-2: Graph for Day of Week



Crimes seem to be almost evenly distributed across all days of the week. But there seems to be a slight increase on Fridays. Friday night partying culture might have an impact on that spike.

Figure-3: Graph for Rate of crime per District in SF

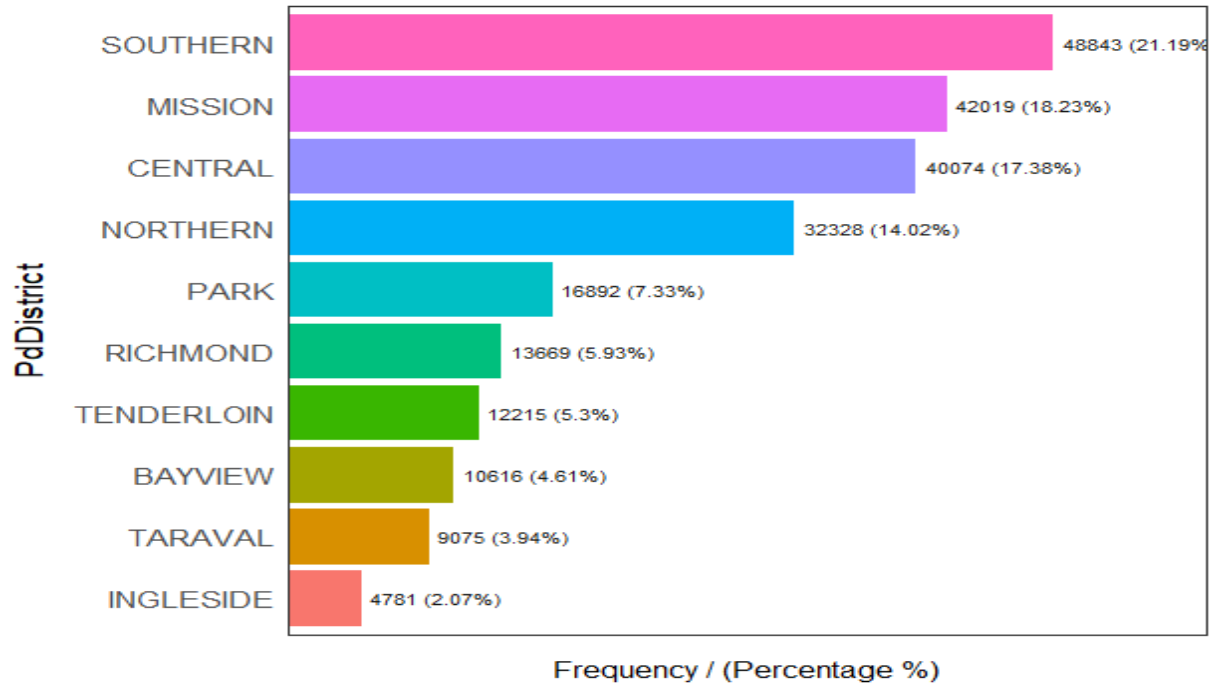


Figure-3 shows the trend of the crime over the years in various districts of San Francisco. These are the Police Districts and each of those includes many other city districts. Here, you can see that the crime in Southern, Mission, Central and Northern districts is on the rise. Whereas, crimes in Taraval and Ingleside have seen a fall in their crime.

3.1 Building Model

In this prediction task, we employed several classification and clustering models to analyze the performance of the model.

Decision Tree

The decision tree classification model forms a tree structure from the dataset. A decision tree is built by dividing a dataset into smaller pieces. At each step in the algorithm, a decision tree node is split into two or more branches until it reaches leaf nodes. Leaf nodes indicate the class labels or results. At each step, the decision tree chooses a feature that best splits the data with the help of two functions: Gini Impurity and Information Gain. Gini Impurity measures the probability of classifying a random sample incorrectly if the label is picked randomly according to the distribution in a branch.

$$IG(p) = \sum_{i=1}^{k} p_i (1 - p_i)$$

Gini Impurity is computed by summing the probability of pi times the probability of mistaking while categorizing an item $(1 - p_i)$. While building the tree, Information Gain helps to decide which feature to split next at each step. Information Gain can be calculated using entropy, which is a function to calculate the expected value at each step. Entropy is defined as:

$$H(T) = IE = -\sum_{i=1}^{k} p_i \log_2 p_i \ldots \text{(ii)}$$

p_i represents the percentage of each feature being present in the child node after a split. The Sum of p_i is always 1. Information Gain can be calculated using the following equation:

$$IG = \text{Entropy(parent)} - \text{Weighted Sum of Entropy(children)} \ldots \text{(iii)}$$

At each step, the decision tree tries to make splits that give the purest child nodes. Our response is category and predictors are Dayofweek, PdDistrict, and hour.

```
> #decision tree
> sanfc.dt <- train(Category ~ DayOfWeek + PdDistrict + hour, data = sanfc.train , method = "rpart")
> sanfc.dt1 <- predict(sanfc.dt, data = sanfc.train)
> table(sanfc.dt1,sanfc.train$Category)

sanfc.dt1            Assault Drug offences Other Property crime Public order Sexual offences Suicide Theft
  Assault                  0             0     0              0            0               0       0     0
  Drug offences            0             0     0              0            0               0       0     0
  Other                11266          5830 28988           4848         7653            1219    2679 20194
  Property crime           0             0     0              0            0               0       0     0
  Public order             0             0     0              0            0               0       0     0
  Sexual offences          0             0     0              0            0               0       0     0
  Suicide                  0             0     0              0            0               0       0     0
  Theft                 9879          4016 25435           5745         6557             877    1850 35433
  Victimless-crime         0             0     0              0            0               0       0     0
  white-collor crime       0             0     0              0            0               0       0     0

sanfc.dt1            Victimless-crime White-collor crime
  Assault                          0                  0
  Drug offences                    0                  0
  Other                          185               4350
  Property crime                   0                  0
  Public order                     0                  0
  Sexual offences                  0                  0
  Suicide                          0                  0
  Theft                          158               3953
  Victimless-crime                 0                  0
  White-collor crime               0                  0
> mean(sanfc.dt1 == sanfc.train$Category)
[1] 0.3556911
>
```

Random Forest:

Random Forests is a very popular assembling learning method that builds many classifiers on the training data and combines all their outputs to make the best predictions on the test data. Thus, the Random Forests algorithm is a variance minimizing algorithm that uses randomness when making a split decision to help avoid overfitting on the training data.

We used the random forest to rank the features based on their importance to predict the labels. In our study, we came across that random forest does not work well with negative values thus we took the absolute of the longitude(X), it didn't make any difference on the dataset but slightly improved the performance of the model.

After performing Exploratory Data Analysis, the following features were used for classification: DayOfWeek, PdDistrict, Hour

```
> test_pred = data.table(knn_test$prob)
> # View testing accuracy.
> print('Testing Accuracy')
[1] "Testing Accuracy"
> print(table(train_model$category_predict == train_pred$V1))

 FALSE   TRUE
159213  71299
> print(prop.table(table(train_model$category_predict == train_pred$V1)))

    FALSE      TRUE
0.6906929 0.3093071
>
```

K-Nearest Neighbors

After performing Exploratory Data Analysis, we have scaled the predictor's Latitude & Longitude.
We further converted factor variables to numeric, namely hour, Population, DayOfWeek, PdDistrict, Season, Geohash to build our model for better classification.

```
> test_pred = data.table(knn_test$prob)
> # View testing accuracy.
> print('Testing Accuracy')
[1] "Testing Accuracy"
> print(table(train_model$category_predict == train_pred$V1))

 FALSE   TRUE
159213  71299
> print(prop.table(table(train_model$category_predict == train_pred$V1)))

    FALSE      TRUE
0.6906929 0.3093071
>
```

CONCLUSIONS

The dataset is highly random, and features were less likely related to the type of crime. However, categorizing the feature and crimes improved the performance of the model slightly. Thus, we can conclude from the above results and plots, that the type of crime is less likely dependant on the external factors such as weather, day, time, and location.

REFERENCES

https://www.kaggle.com/c/sf-crime
https://www.justia.com/criminal/offenses/