**San Francisco Crime Classification**

**Kavya** [**Nagaraju(hc3344@wayne.edu**](mailto:Nagaraju(hc3344@wayne.edu)**)**

**Abstract:**

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. We are working on the San Francisco Crime classification Data from Kaggle which was collected from the SF Police Department reporting system.

We analyzed the data from 2003 to 2015 with more than 800,000 observations.

Our initial Hypothesis is to know whether any external factors like weather conditions, seasons, time, and location impact the crimes that occurred.

Using data modeling approaches such as Random Forest, Decision Tree, and KNN we created models that can be used to classify the category of crime given the geographical features.

***Keywords****:* Crime Classification; Random Forest; K-NN; Decision Tree.

**Introduction**

The main purpose of an introduction is to enable the paper to be understood without undue reference to other sources. It should, therefore, have enough background material for this purpose. Generally, highly specialized papers will not need an extensive introduction as interested readers may be expected to be familiar with current literature on the subject. On the other hand, when a paper is likely to interest people working in fields outside the immediate area of the paper, the introduction should contain background material that could otherwise be scattered throughout the literature.

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. 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. Though with the help of data mining and other tools the prediction of crime can be done. This doesn’t mean that the crime will be completely controlled but to some extent, the SFPD can provide helpful information to prevent crime.

**DATA SET DESCRIPTION**

The San Francisco Crime classification data with 800,000 observations has the following features:

* **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.

**DATA CLEANING AND PREPROCESSING**

As mentioned earlier 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 ‘zipcode’. 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 has the following features with 230512 observations of and14 variables :

* **Date s:** 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:
* **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.
* **Lat**: The latitude of the crime scene.
* **Lng:** The longitude of the crime scene.
* **Population:** The total population of area zipcode covers.
* **Hour:** The hour at which the crime took place.

**FEATURE EXTRACTION**

**Hour:** From the given timestamp in the crime dataset, the time of the crime are extracted.

**Crimes**: The original crime dataset, has 39 types of crime recorded.

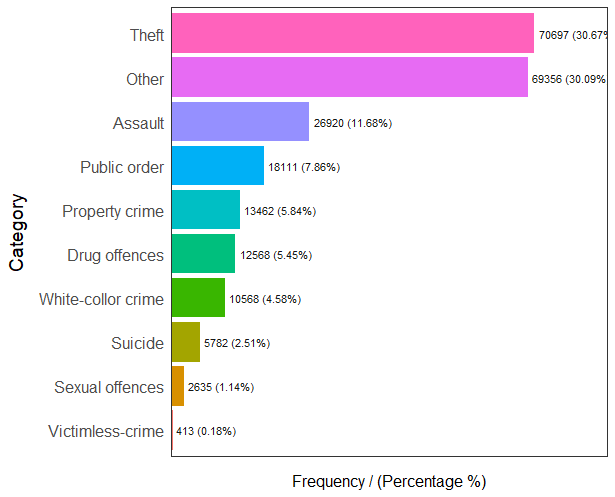
**Zip-code**: Zip-code of the region where the crime was reported.

**Season**: The Season of the year when the crime was reported.

**Geohash**: The Geohash codes were extracted from the latitude and longitude of san Francisco crime

**EXPLORATORY DATA ANALYSIS**

The exploratory data analysis was performed on all the features to summarize the main characteristics of the dataset.



***Fig-1:*** ***Category***

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 the figure.

From the exploratory analysis, we noted the features are very random. Hence, some of the features were categorized/Grouped for better representation of the features and to improve the model accuracy.

The following feature engineering tasks were performed.

1. Crime: The models built for classifying all the 39 classes perform very poorly.

Hence, the crimes are grouped based on the categories.

● Theft: LARCENY/THEFT, VEHICLE THEFT, 'BURGLARY.

● Sexual offenses: SEX OFFENSES NON FORCIBLE, SEX OFFENSES FORCIBLE, PORNOGRAPHY/OBSCENE MAT, PROSTITUTION.

● Public Order: DRUNKENNESS, SUSPICIOUS OCC, DRIVING UNDER THE INFLUENCE, RECOVERED VEHICLE, 'BAD CHECKS', 'LOITERING', 'DISORDERLY CONDUCT', 'LIQUOR LAWS', 'WEAPON LAWS'

● Assault: SEX OFFENSES FORCIBLE, KIDNAPPING, ASSAULT.

● Drug offenses: 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, EXTORTION.

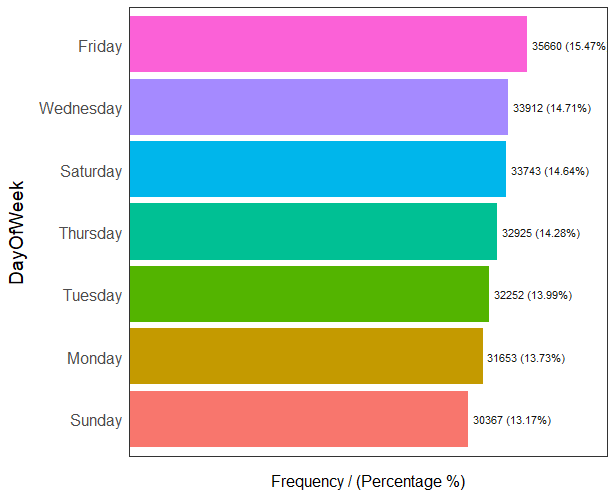
● Other: WARRANTS, OTHER OFFENSES, NON-CRIMINAL,

MISSING PERSON.

2. Seasons: The whole year is divided into 4 groups based on the seasons. This

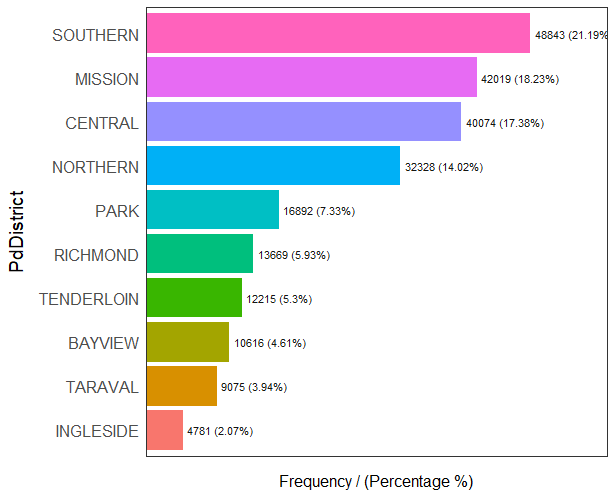
helps in classifying the crimes based on when the crime has occurred the

most.



***Fig-3: Day Of Week***

Crimes seem to be almost evenly distributed across days of the week. They increase on Fridays, though. Friday night parting culture might have an impact on that spike.



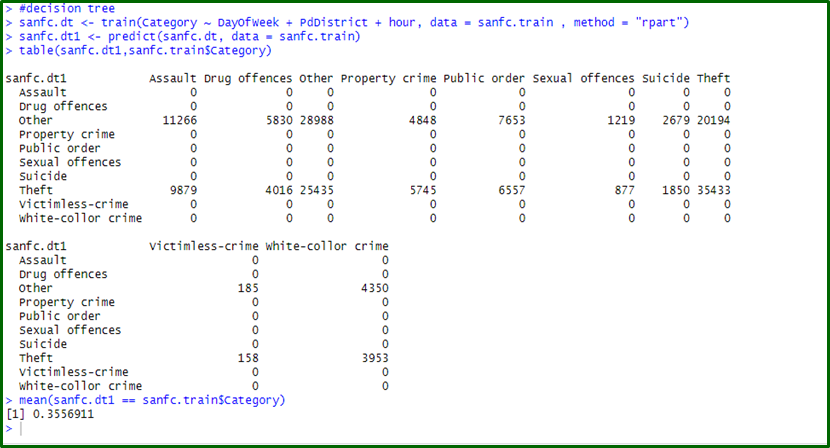
**BUILDING THE MODEL**

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

**Decision Tree:**

The decision tree algorithm works by splitting the dataset recursively, which means that the subsets that arise from a split are further split until a predetermined termination criterion is reached. At each step, the split is made based on the independent variable that results in the largest possible reduction in the heterogeneity of the dependent variable.

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

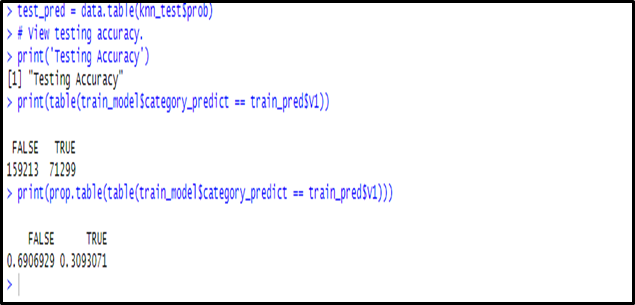


**Random Forest:**

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.

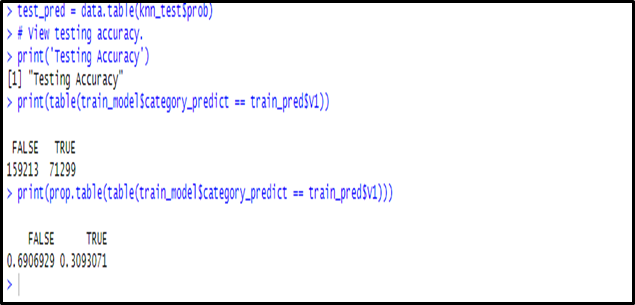
After performing Exploratory Data Analysis, the following features were used for classification: DayOfWeek, PdDistrict, Hour



**K-Nearest Neighbors:**

After performing Exploratory Data Analysis, We have scaled the predictors “Latitude” &“Longitude”.

We further converted factor variables to numeric, namely “hour”,” Population”,”DayOfWeek”,”PdDistrict”,”Season”,”Geohash” to build our model for better classification.



**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 of the external factors such as weather, day, time and location.

**REFERENCES**

● Kaggle - <https://www.kaggle.com/c/sf-crime>

● https://www.justia.com/criminal/offenses/