CS - 418 Introduction to Data Science

Final Report

San Francisco Crime Classification

Deexith Mysore Nagaraj - dmysor3@uic.edu
Pavan Holenarasipura - pholen2@uic.edu

Contents:

AIM

To categorize the different types of crime in San Francisco using the San Francisco Crime Dataset.
We are trying to apply classification and clustering techniques over San Francisco's crime dataset.

Data Set Description

The dataset has the following features recorded.
Dates, Category, Descript, DayOfWeek, PdDistrict, Resolution, Address, X, Y.

- **Date**: The timestamp of the crime recorded.
- **Category**: The category of the of the crime recorded.
- **Description**: 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

The values are very detailed and doesn't 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 other dataset.

The weather of the day was obtained from the a weather data set. The weather data set had a lot of null values, hence weather for each day is considered rather than every hour and each day weather value was merged to the main dataset based on the location and the date, time information.

The Geohash library was used to obtain the zipcodes of each location based on the latitude and longitude information.

The Final Dataset has the following features.

- **Date**: The timestamp of the crime recorded.
- **Category**: The category of the of the crime recorded.
- **Description**: 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.
- **Zip-code**: Zip-code of the area where the crime was reported.
- **Weather**: Weather information of each day.

Feature Extraction

- **Date**: From the given timestamp in the crime dataset, the year, month, date and the time of the crime are extracted.

- **Time:** From the given timestamp in the crime dataset, the time of occurrence of the crime is extracted.

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

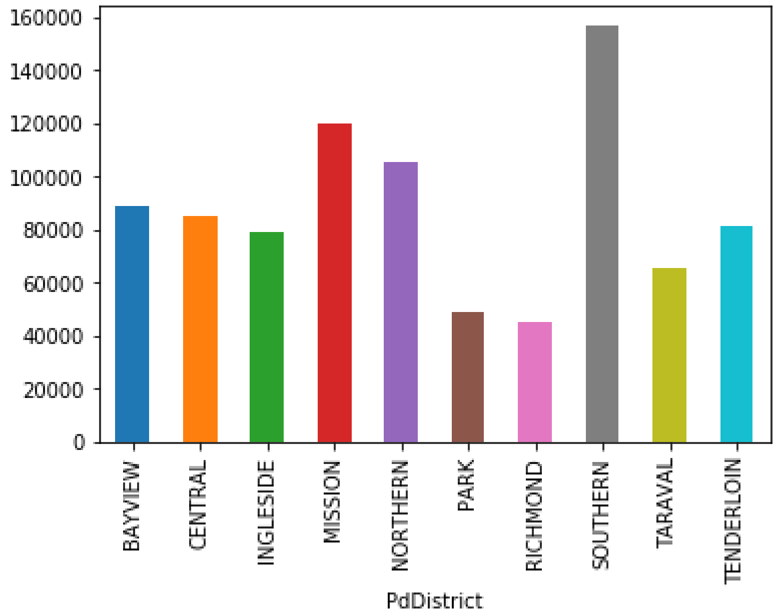
- **Zip-code**: Zip-code of the region where the crime was reported.

- **Weather:** The weather of the day during which the crime occured.

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

Exploratory Analysis

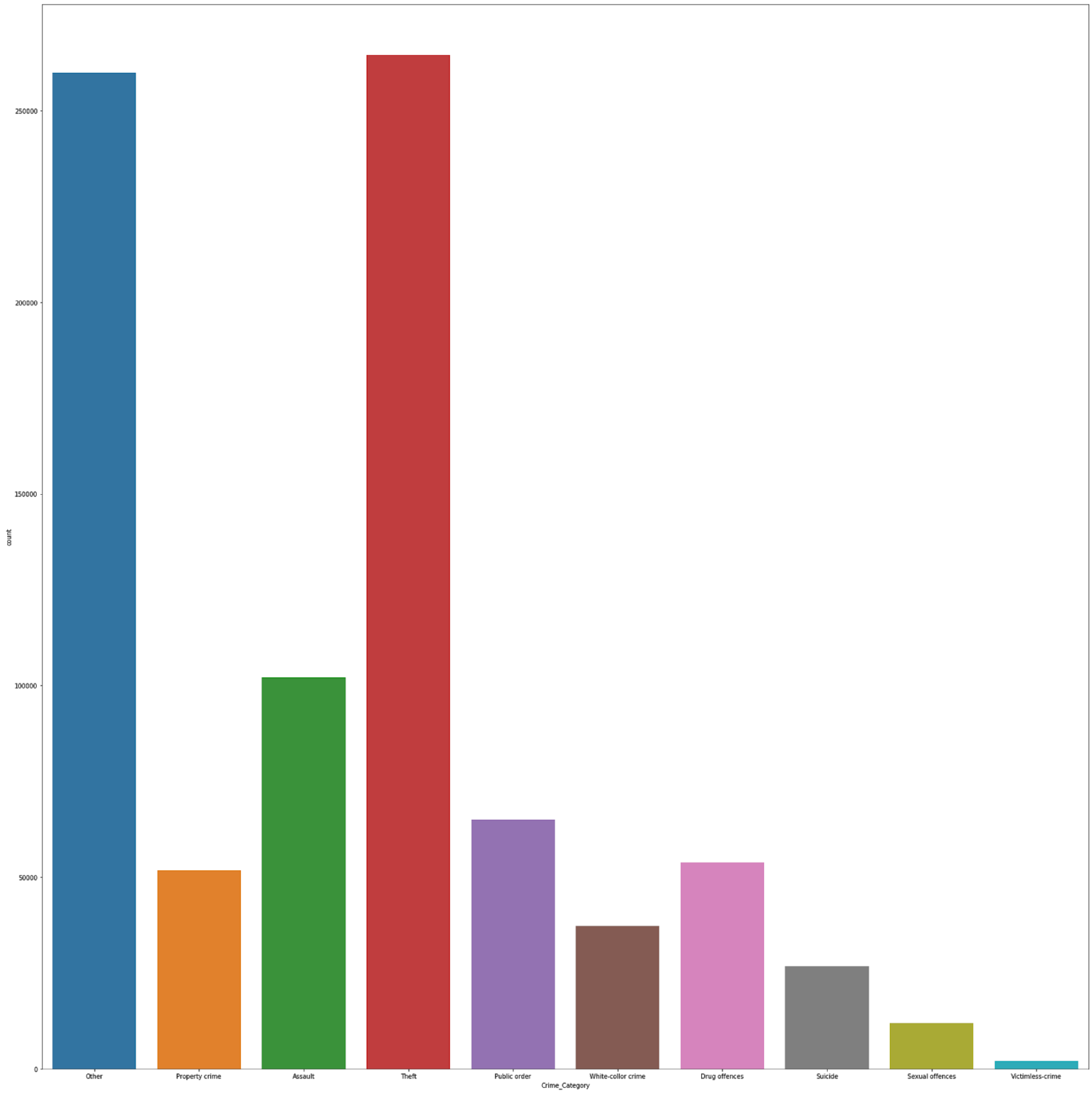
The exploratory analysis was performed on all the features to summarize the main characteristics of the dataset.



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.

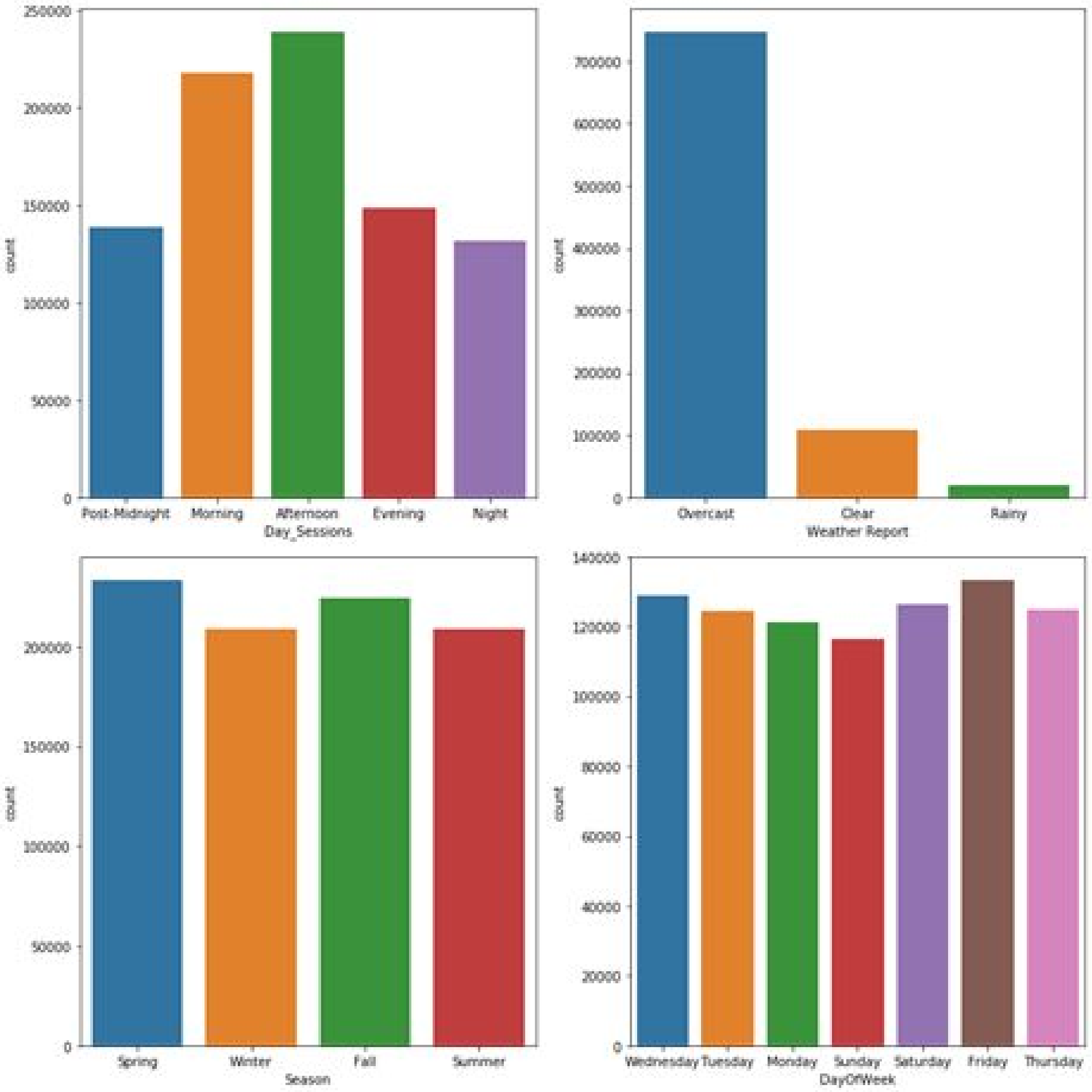
1. **Crime:**The models built for classifying all the 39 classes perform very poorly. Hence, the crimes are grouped based on the categories.
    - **Theft** : LARCENY/THEFT , VEHICLE THEFT, 'BURGLARY.
    - **Sexual offenses:** SEX OFFENSES NON FORCIBLE, SEX OFFENSES FORCIBLE, PORNOGRAPHY/OBSCENE MAT, 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 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, 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.

3. **Sessions**: The whole day is divided into 5 sessions as the morning, afternoon, evening, night and post-midnight for better generalization of the crimes recorded in a day.

4. **Weather**: The weather feature obtained from the weather dataset is grouped into 3 categories - Overcast, Clear and Rainy.



Building the model

Classification:

After Performing the Exploratory analysis, the following features were used for classification:
PdDistrict, Weather, DayofWeek, Session.

Results

The Summary of Results obtained are:

Classifier	***Accuracy, Precision, Recall, F1 Score***
KNN	0.297531771632702, 0.25624441256286706, 0.31593356599523625, 0.2749205150139882.
Decision Tree	0.3506780868088218, 0.318143608476642, 0.48890564354700833, 0.3666771009191756.
Random Forest	0.3506171523019914, 0.31642295565768586, 0.4888206903509098, 0.36647755221217937.
Decision Tree using Boosting	0.3506171523019914, 0.31642295565768586, 0.4888206903509098, 0.36647755221217937

Clustering:

After Performing the Exploratory analysis, the following features were used for Clustering.: ‘Day’, ‘Year’, ‘Day_Sessions’.
The model is trained to cluster all the 39 classes.

Results

The Summary of Results obtained are:

Clustering	*Adjusted Random Score,* *Normalized Mutual Score.*
K Means	0.6717632262971134, 0.8339434441098231
K++	0.5180286688056641, 0.7779241403337509

Conclusion

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/