Project Report
San Francisco Crime Classification

Team: The Outliers

Khyati Pawde : 704590958

(Report, Slides. Presentation, Algorithm: Decision Trees, Random Forests)

Mansi Shah : 504591572

(Report, Slides, Visualization, Algorithm: Logistic Regression, K Means)

Pallavi A. Kotkar : 504589593

(Report, Slides, Presentation, Algorithm: Naive Bayes, SVM)

Sravani Kamisetty : 304414410

(Report, Slides, Feature Engineering, Algorithm: KNN)

Description

The project uses the SF crime dataset that provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given a time and location, the project shall predict the category of crime that occurred. This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 01/01/2003 to 05/13/2015. The training set and test set rotate every week, meaning week 1,3,5 belong to test set, week 2,4,6,8 belong to training set.

We will utilize machine learning classification techniques like Logistic Regression, SVM, Decision trees etc. to build a model that predicts the category of the crime. The best performing algorithm will be applied on the test dataset. The predicted labels i.e the category of the crime of the test test model will be evaluated via a Kaggle submission.

Motivation:

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more for its tech scene than its criminal past. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. The city saw more than 20% jumps in both the rate of property crime, such as thefts and burglary, and the rate of violent crime, such as robbery and assault, between 2012 and 2013.

The primary motivation of this project is to make use of the huge amount of data available to help gauge the current crime scenario in San Francisco. We can not only predict the category of crime but also find underlying patterns in the data which can reveal certain criminal psychologies. Such patterns when uncovered can help the public as well as the police department in abatement of such crimes.

Data Fields:

- Dates - timestamp of the crime incident
- Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.
- Descript - detailed description of the crime incident (only in train.csv)
- DayOfWeek - the day of the week
- PdDistrict - name of the Police Department District

- ➔ Resolution - how the crime incident was resolved (only in train.csv)
- ➔ Address - the approximate street address of the crime incident
- ➔ X - Longitude
- ➔ Y - Latitude

Exploratory Analysis

Firstly we plotted the counts of the categories of crimes. The below graph shows that the top 5 crimes are - Larceny, Assault, Drug/Narcotics, Vehicle Theft and Vandalism and they contribute to more than 70% of the total crimes in the dataset.

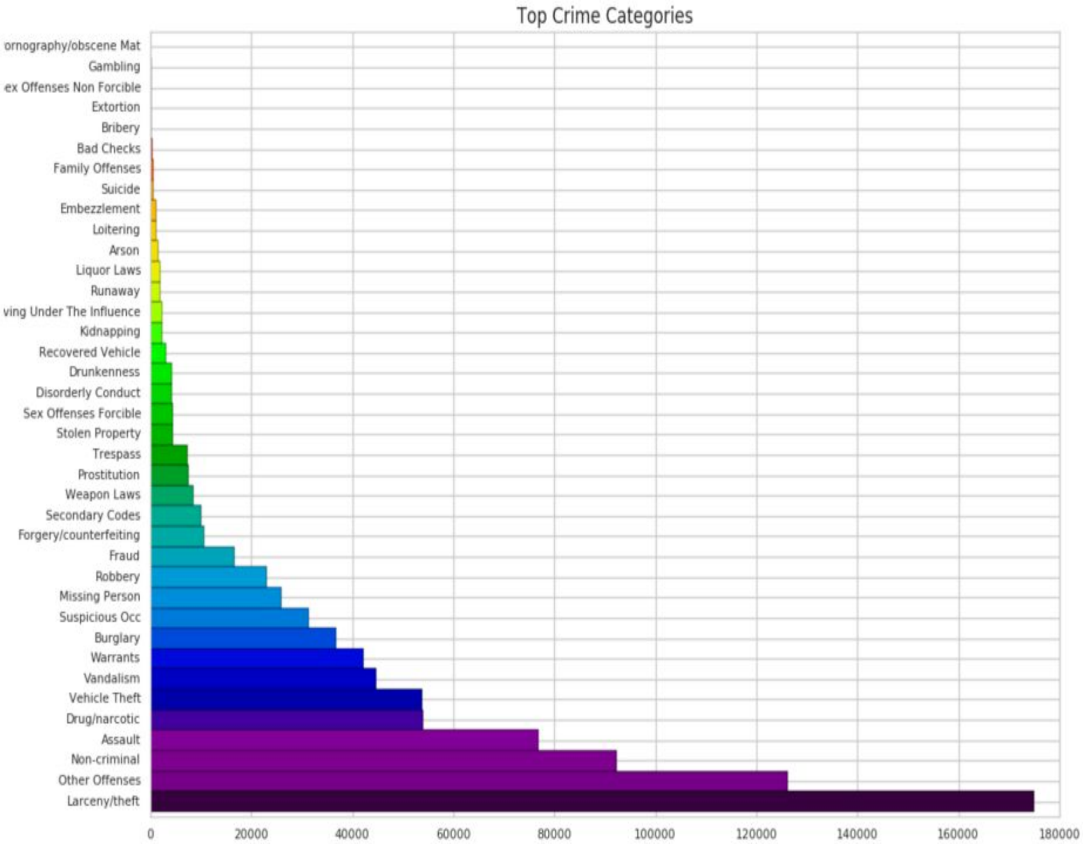


Figure 1: Top crime categories

We took the top 5 crime categories and plotted them on a SF map using the latitude and longitude to get an idea of the location/areas where the crimes occur. From the map below it’s clear that most of the crimes happen in and around Tenderloin

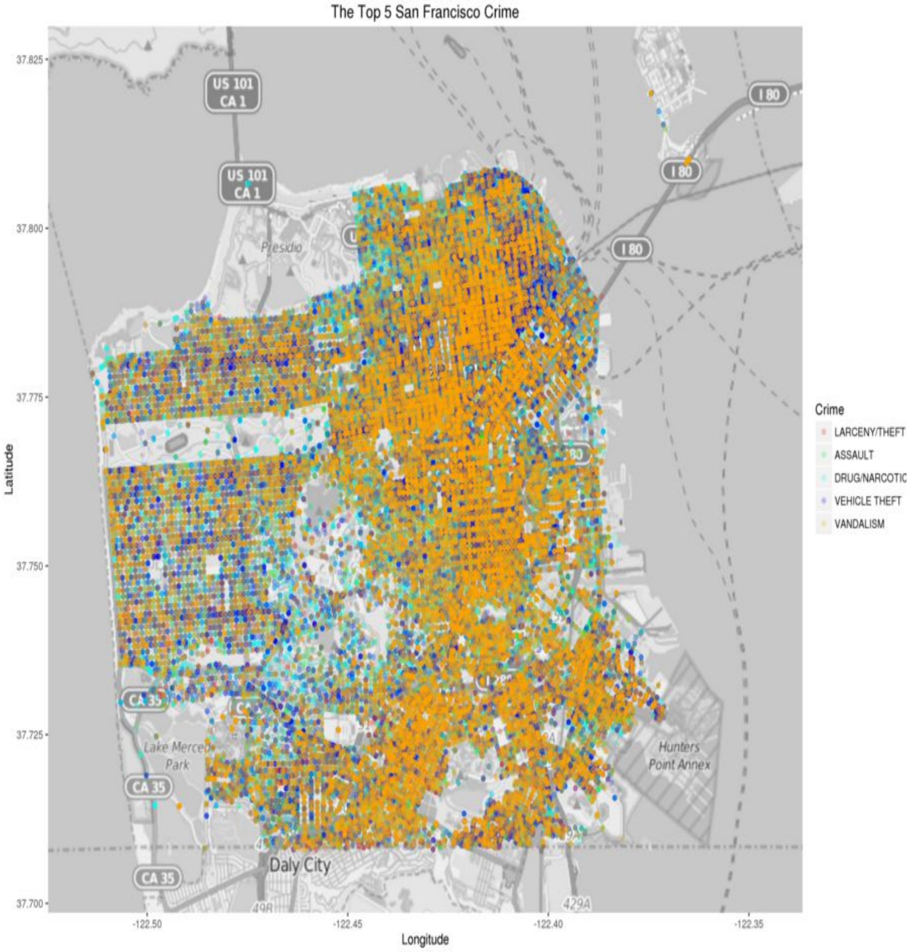


Figure 2: Top 5 San Francisco crimes

We then took into consideration the district of the police department (PdDidtrict) and compared it against the crimes that were committed in the various categories. We observed that there were a drastically high number of crimes in all categories except a few such as "Drug/Narcotic", "Burglary" in the "Bayview" district as compared to the other districts.ANother generalized assumption we could make is that the crimes in "Park" and "Richmond" district are relatively lower than the other districts. The Figure (3) represents the bar chart for each crime category against the district in which it took place. The Figure (4) represents the same information for the top 5 categories to provide clarity for later deductions.

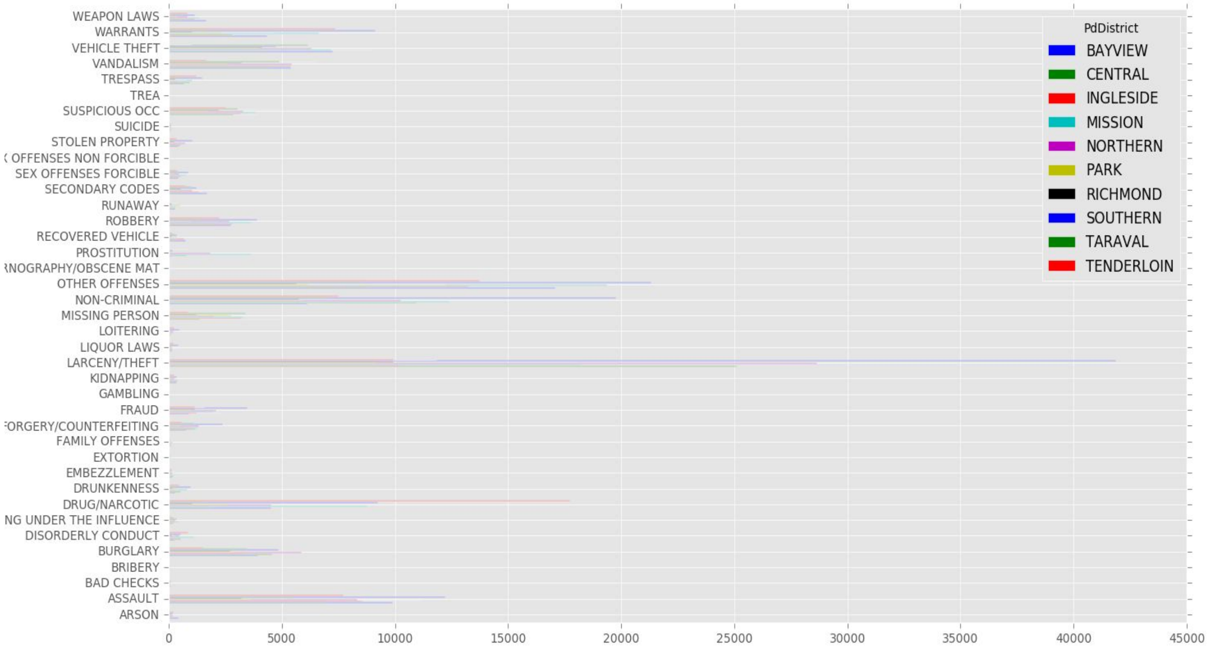


Figure 3: Crime category by PdDistrict

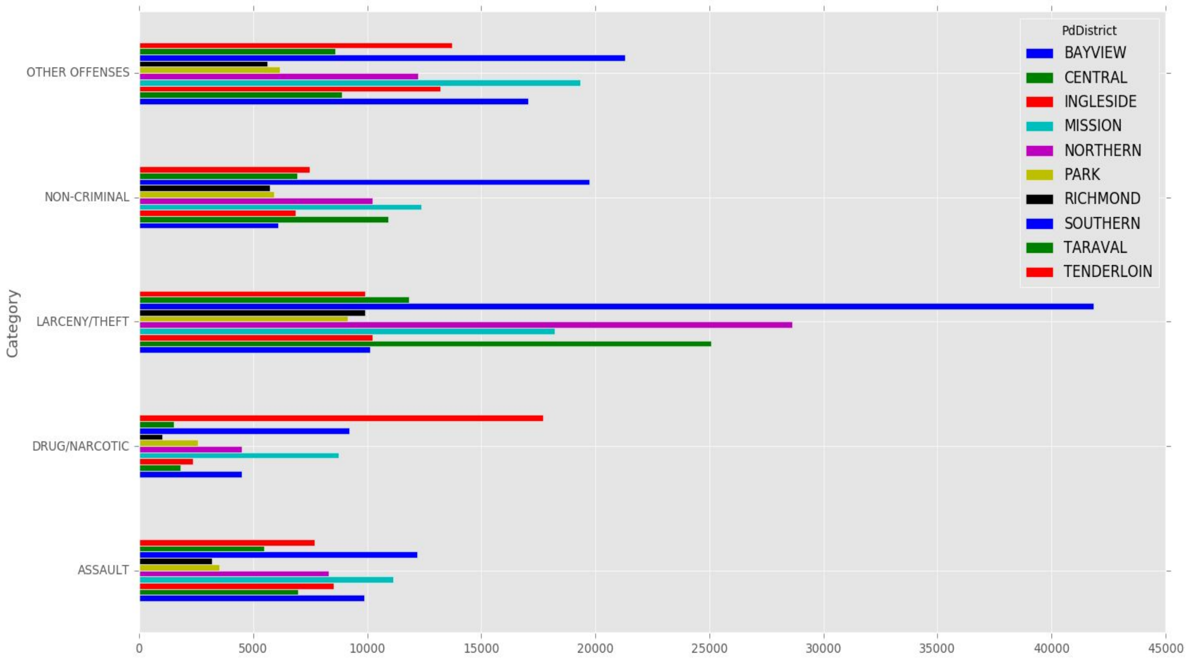


Figure 4: Top 5 crimes by PdDistrict

The next attribute which we considered was the day of the week (DayOfWeek). Given below is the graph which represents the days of the week in increasing order of crimes recorded on those particular days. Now, we can safely say that the rate at which crimes occur is maximum on Fridays and minimum on Sundays.

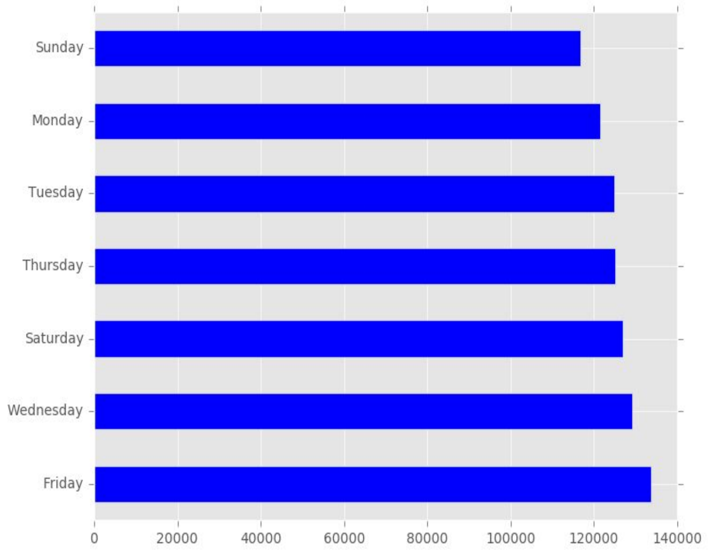


Figure 5: Number of crimes by DayOfWeek

The attribute 'Dates' specifies the exact time and date at which the crimes took place. Utilizing this attribute,we analyzed the number of crimes (segregated for each category) that took place during each hour of the day. We can make various conclusions from this data such as the maximum number of 'larceny/theft' crimes occur from a time period from 6:00 pm. to 7:00 pm (ie. the 18th hour of the day).

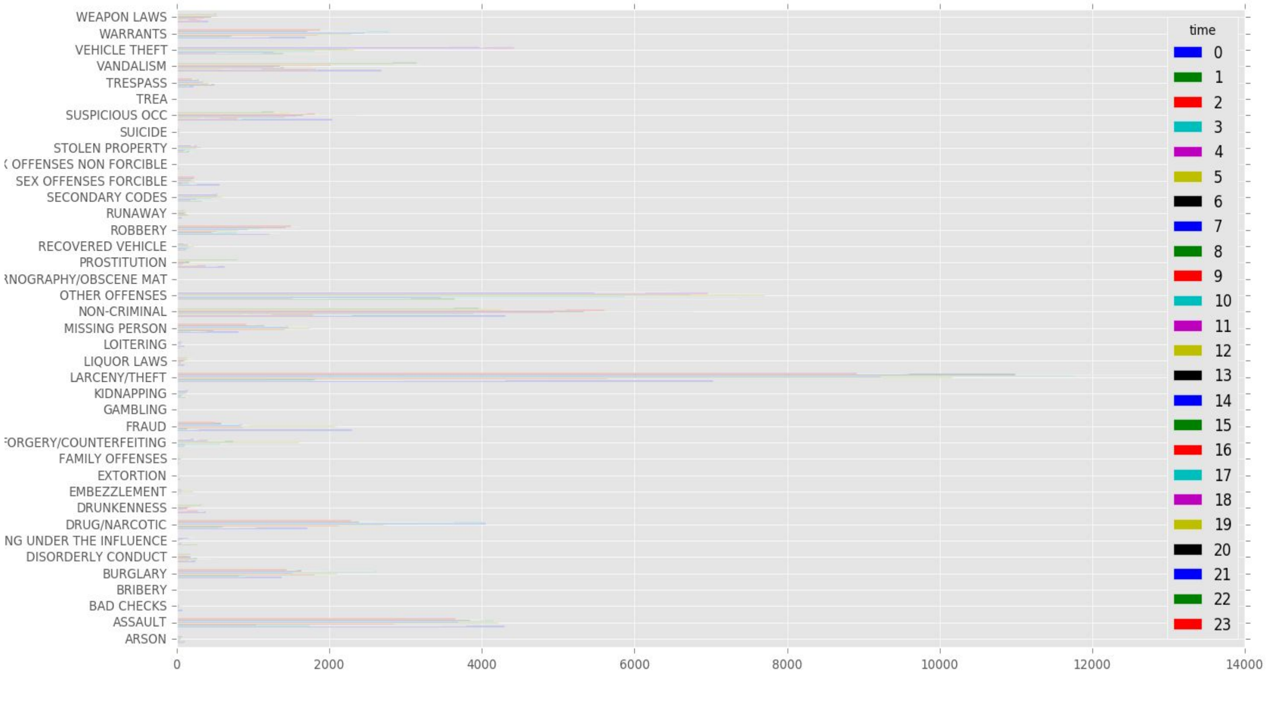


Figure 6: Number of crimes by hour of the day

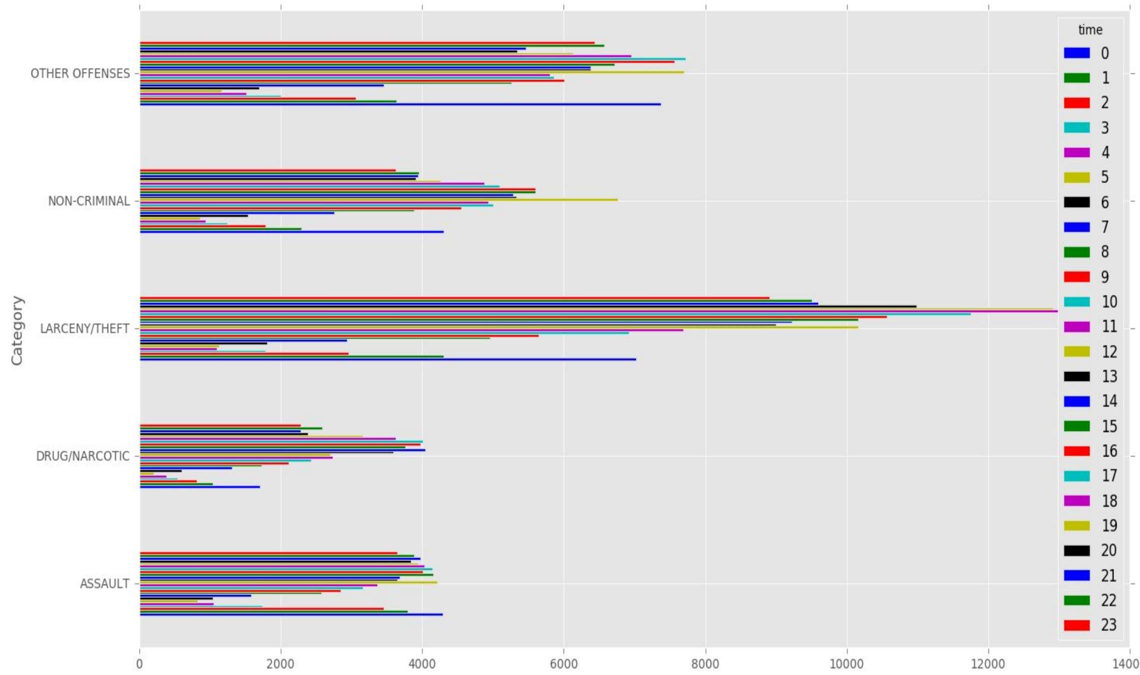


Figure 7: Number of crimes per hour of the day for top 5 categories

Finally, we created a heatmap representing the total number of crimes occurring at a particular time on a particular day of the week. This plot helped us deduce that the rate of crimes does vary tremendously in accordance to the day/time. One of the deductions we can make from the map is that the maximum number of crimes happen on Fridays from 6 pm to 7 pm, which reinstates the previous assumptions we made in the previous two graphs for day and time.

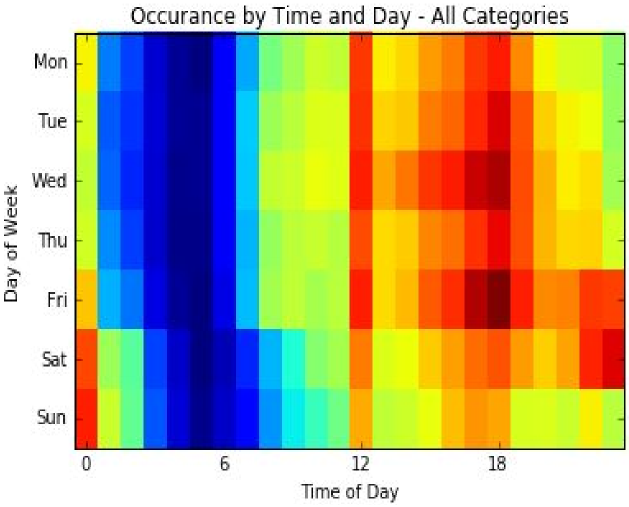


Figure 8 : Heatmap for time of the day vs day of week

Feature Extraction

Based on the exploration that we performed on the data, we deduced the factors which were correlated with the number/category of crimes that occurred in SF(in a particular location, at a particular time).

The attributes which represent those factors are :

(1) Dates
(2) DayOfWeek
(3) PdDistrict
(4) X (Latitude)
(5) Y (Longitude)

All these attributes are **categorical** hence we extracted the requisite features by preprocessing.

DayOf Week :
We used One Hot Encoding to expand the attributes to each represent each day of the week.
We assigned binary (1/0) values to the attributes based on the day on which a particular crime occurred.
Example: Sunday is represented as [1,0,0,0,0,0,0]

PdDistrict :
Similar to the approach we used for DayOfWeek, we applied One Hot Encoding to repreesnt the 10 police districts in San Francisco.
Example: Ingleside was represented as [0,0,1,0,0,0,0,0,0,0]

Dates:
The dates were a feature which represented the time, day, month, year of the crime. We parsed through this data and generated 4 new columns consisting of each these numerical values.

Example: The given entry 10-05-2015 23:59:00 was represented as :
Time: 23 (hour of the day when the crime occurred)
Day: 10
Month: 05
Year: 2015

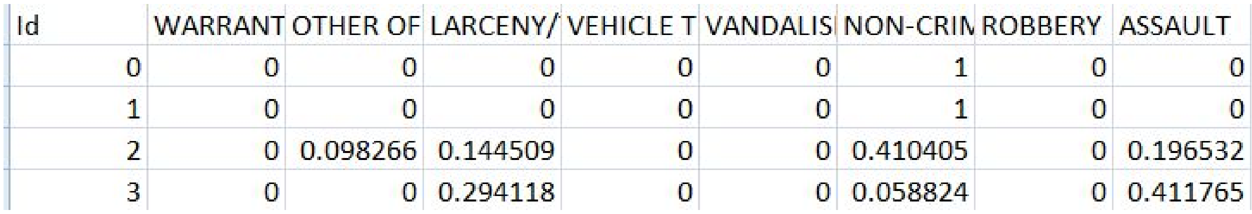
X and Y were numerical attributes hence they required no preprocessing.

Evaluation Dataset:

Since, Kaggle did not provide the categories of the crimes in the test data, we created an evaluation dataset from the training data to test the accuracy of our results. This dataset contained 20% of the values of the training dataset.

Output Format:

We followed the Kaggle format for generating the output for each crime. This was represented by the probability of a crime belonging to each of the 39 categories. The sum of the probabilities in each tuple/row of the output equals 1.



Id	WARRANT	OTHER OF	LARCENY/	VEHICLE T	VANDALIS	NON-CRIN	ROBBERY	ASSAULT
0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0
2	0	0.098266	0.144509	0	0	0.410405	0	0.196532
3	0	0	0.294118	0	0	0.058824	0	0.411765

Evaluation Technique:

We used a multi-class logarithmic loss (logloss) function to evaluate the accuracy of our predictions.

The logloss function is calculated as follows:

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where,

N : Number of cases in the test set

M : Number of class labels

y_{ij} : 1 ; if observation i is in class j

0 ; otherwise

P_{ij} : Predicted probability of observation i that belongs to class j

Technologies Used:

Python Libraries - pandas, numpy, scipy, sklearn, matplotlib.

Methods :

Classification without Clustering

Initially we experimented on the data by running generic machine learning algorithms in on the inherently numeric attributes of the data which were X and Y for the latitude and longitude. The results we obtained were not precise enough due to the inadequacy of training information. The following were the Classification ALgorithms we applied independently :

1. K Nearest Neighbours: The logloss calculated on the evaluation dataset was 25.5457. The best accuracy was obtained when the values of K was 40. The graph below plots the logloss function of the evaluation dataset against the various values of k.

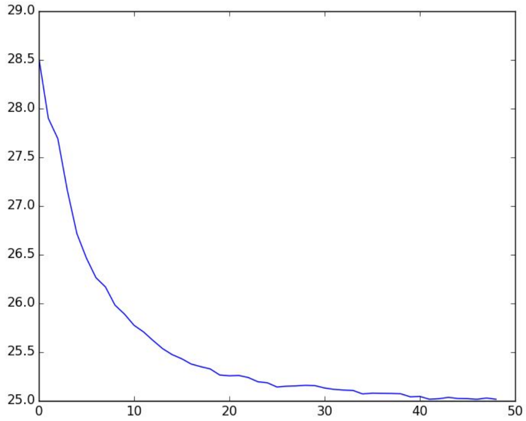


Figure 9: Log loss vs K

2. SVM

We used the Linear SVM for multiclass classification with the one-vs-rest strategy. The results obtained were not as expected. The log loss values were very high and the accuracy was low. The classifier could predict only two class labels. Exploratory analysis suggests that the dataset being sparse, and LARCENY/THEFT and OTHER/OFFENCES being the top 2 crimes, a large majority of tuples had these two labels.

Best log-loss: 34.3241

3. Naive Bayes:

We used a multi class classification algorithm which is probabilistic in nature. The underlying assumption for the Naive Bayes algorithm is satisfied for our dataset since the

two numerical attributed that we are using for the classification are independent of each other. The log loss values improved over SVM classifier.
Best log-loss: 26.8343

4. Decision Tree Classifier:

Decision Tree classifier is another algorithm for multiclass classification. Since we were using only 2 attributes for the baseline classification, and would work with more features in further phases we wanted to find baseline results for this classification technique too. Using the same set of features the log loss metric improved as compared to the Naive Bayes classifier but not as much as expected. We discovered the reason for low accuracy could be overfitting. We repeated the experiments by varying the maximum tree depth attribute for the algorithm and go the best results with max_tree_depth = 18.
Best log loss: 24.7645

5. Random Forests

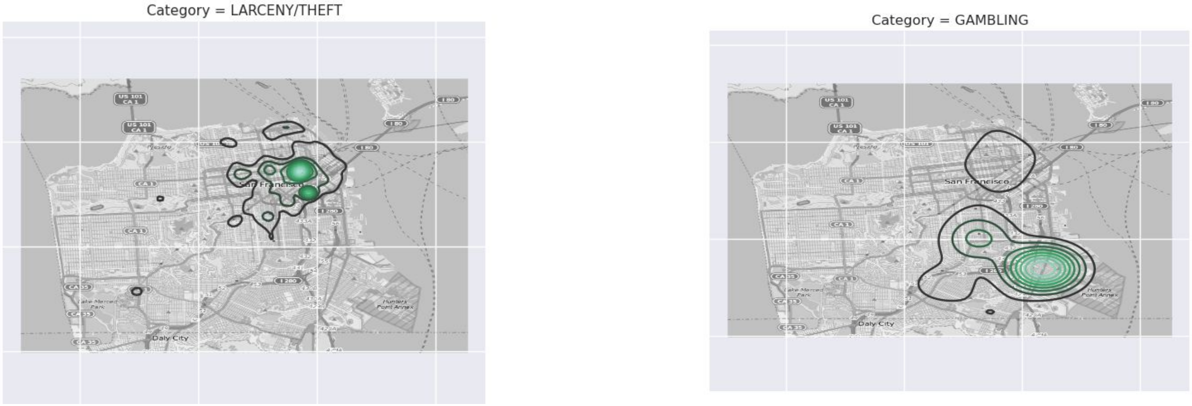
Since decision trees with pruning gave us the best results till now, we decided to use the random forests ensemble technique. Random forests gave us the best results of all the algorithms that we tried. The number of decision trees that formed were used to make predictions in the random forests were 50.
Best log loss: 23.0341

Summary of log loss for all methods:

METHOD	**LOG LOSS**	
	EVALUATION	**TESTING**
KNN	25.5457	N/A
SVM	34.3241	N/A
NAIVE BAYES	26.8343	N/A
DECISION TREES	24.7645	N/A
RANDOM FORESTS	23.0341	24.02

Classification with Clustering:

Exploratory data analysis helped us finding interesting patterns with the spatial features that we had.



(a)

(b)

Figure 10: Crime density by location for (a)Larceny/Theft and (b)Gambling

We clustered the data using the latitude and longitude features using K-means clustering technique. We repeated the experiments for different K values ranging from 30 to 100. As the number of clusters increased, the clusters centres came in close proximity of each other. Optimal clustering was obtained with K =40 which agreed with the results of KNN classification since the number of crime categories was also 40. We performed classification within these clusters with the expanded dataset which is elaborated in the following section.

One of the main challenges we faced was the dataset had fewer numerical attributes. We performed additional preprocessing of the data and expanded the feature set. We expanded the following features:

Dates	Date
	Time
DayOfWeek	7 levels
PdDistrict	23 levels

With the expanded dataset we repeated the classification experiments to predict top 5 category labels.

All the above methods - SVM, decision trees, random forests naive bayes were applied for the clustered data. An extra column with the cluster number was appended to the feature list. This improved the accuracy of prediction and lowered our log loss to a great extend. The below table shows all the log loss values for evaluation and testing data.

Logistic Regression: We applied logistic regression to all the features i.e latitude, longitude, cluster number, one hot encoded values for dayOfWeek and PdDistrict and time of day, date itself. Binomial variance and logit link options were selected in the generalised linear model. This gives us our best performance and least log loss of 2.60025 on the testing data.

Logistic regression with clustering based on latitude and longitude resulted in best results as we make use of cluster information and also all other numerical attributes for the classification. Also, logistic regression does not overfit the data, leading to better results.

METHOD	**LOG LOSS**	
	EVALUATION	**TESTING**
SVM	13.8976	13.9878
NAIVE BAYES	12.551	12.421
DECISION TREES	10.93	10.89
RANDOM FORESTS	6.1498	6.26
LOGISTIC REGRESSION	2.6523	2.60025

Comparison with the Kaggle Leaderboard:

The best log loss obtained on Kaggle is by a PhD student from University of Ontario. He has obtained a log loss of 2.05079. Since this is an ongoing competition and there is still 3 more months to the deadline, these might not be the best results in a few days. We stand close with the log loss of 2.60025 for the Logistic regression with binomial variance.

Since the the kaggle competition has not yet ended, none of top few have shared the techniques used by them to reduce the log loss. Hence we were unable to a analysis of the top few techniques. Below is the log loss for Kaggle Top#3 and our team Outliers.

Name	**Log Loss**
mehran	2.05079
Jghjgfgh	2.06702
papadopc	2.11607
Outliers	2.60025