San-Francisco Crime Classification

Too Lazy To Compete
Aditya Srivastava(IMT2016004)
Pranav Vardia(IMT2016040)
Vaibhav Kumar(IMT2016086)

***Abstract*—**
***War is ninety percent information.* - Napoleon Bonaparte**
Predicting the probability of category of crime that might occur, given a time and location. We analysed relations between possible predictive features and the label(category of crime) using different classification models such as Random Forest, Logistic Regression, Naive Bayes and XGBoost. We implemented the XG-Boost classifier and performed optimizations to further improve our score on the Kaggle competition hosted. Our exploration of the dataset revealed certain interesting phenomenons.

I. Introduction

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. From Sunset to SOMA, and Marina to Excelsior, the dataset provides with nearly 12 years of crime reports from across all of San Francisco's neighborhoods consisting of time, location and other relevant records.

II. The Dataset

The dataset contains 11 years and 5 months of crime incidents reported all across the City and County of San Francisco.

A. *Overview*

The dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set consists of week 2,4,6,8 across the given timeline. The test set contains 10000 rows randomly picked out of the original training dataset provided by SF OpenData. The data of the year 2015 is not complete and only covers data upto 13th May, i.e. 5 and a half months(roughly). In the dataset, each crime record has nine entriesof information related to the crime incident, which are listed in **TABLE I**. In the dataset, there are 36 categories of crime that are classified. **Fig. 1**.
It can be clearly seen that there is a very uneven distribution of crimes. Crimes such as Larceny/Theft, Other Offenses, Non-Criminal, Assault, Drug/Narcotic and Vehicle Theft take huge portions while certain kinds of crime such as Family Offences, Gambling and Extortion are extremely rare.

TABLE I
Information provided in the dataset per incident

Data-Field	Description
Dates	Timestamp of the crime incident
Category	Category of the crime incident(Label)
Descript	Detailed description of the crime incident
DayOfWeek	Day of the week
PdDistrict	Name of the Police Department District
Resolution	How the crime incident was resolved
Address	Approximate street address of the crime incident
X	Longitude
Y	Latitude

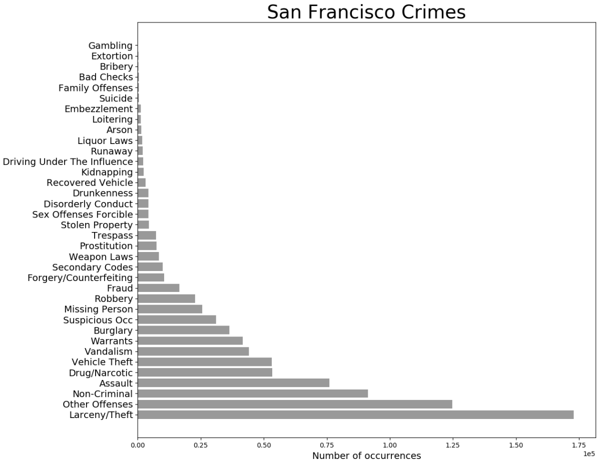

Fig. 1. Number of Crime incidents of Different Categories

B. *Time Features*

Time is the most important factor in a crime. It has the ability to single handedly overturn court rulings, if needed. Here also, time plays has an important effect on crime classification. Time variables related to a crime incident include information in the 'Dates' entry and information in the 'DayOfWeek' entry. The 'Dates' entry in the dataset provides a incident's timestamp with a format of 'year-month-date hour:minute:second'.
We split the timestamp and researched on the following time variables.

1) *Year:*

Fig. II-B1 shows how the total number of crime incidents changes over years. A general trend of reduction can be observed from 2003 to 2011. However, the number of crimes starts to increase since 2012. Another thing to note here is the data of 2015 is not complete since it only covers 5 months record.

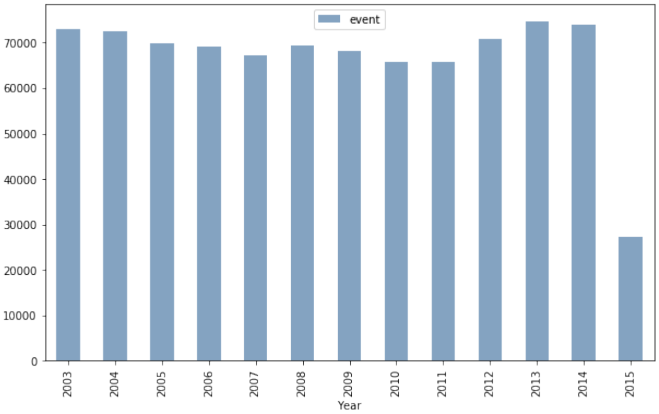

Fig. 2. Crimes in Different Years

Fig. II-B1 shows numbers of top six commonest crimes changes over years. From it we can see that in some degree, the proportion of different kinds of crimes changes in different years. For example, the proportion of Vehicle Theft drops to a low level since 2006, Other Offenses and Drug/Narcotic clearly take a bigger proportion in 2008 and 2009 compared to that in other years, the proportion of Non-Criminal incidents reduces from 2003 to 2008 then start to increase and it finally peaks in 2013. So we can say that the year variable could be a critical feature in crime classification model.

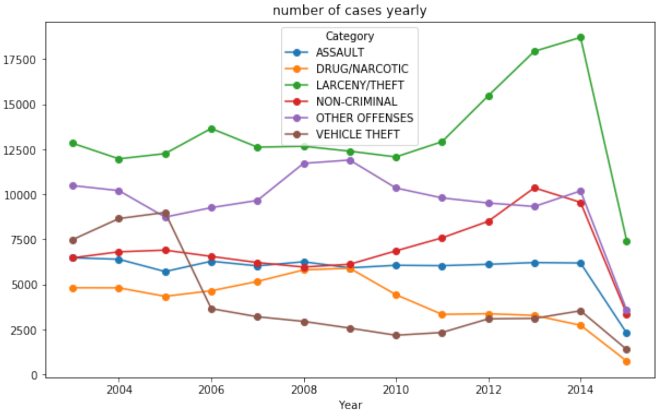

Fig. 3. Top Six Crimes in Different Years

2) *Month:*

Fig. 4 shows numbers of top ten commonest crimes changes over month. Variance of proportion among crime categories here is not very obvious since all categories of crime follow a roughly same trend over different months. So, month may not be a useful variable in the classification model. However, the trend all crimes follow should be noticed. Heres an interesting phenomenon that the total number of crime incidents tends to be lower in summer and winter but higher in spring and fall. We may infer that extreme whether can reduce criminal activities.

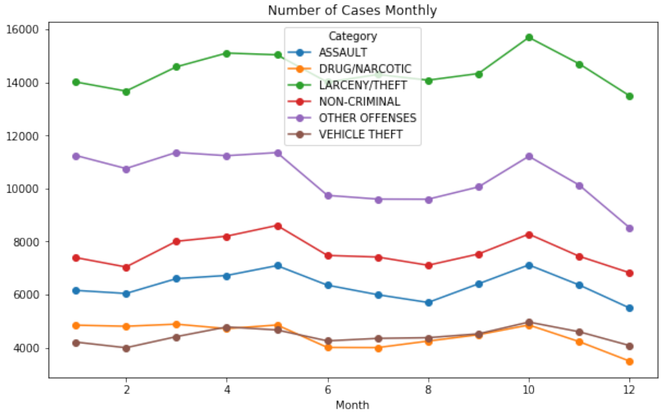

Fig. 4. Top Six Crimes in Different Month

3) *Day Of Week:*

Fig. 5, here we encoded the DayOfWeek by hard-coding them. Our week begins on Monday and ends on Sunday, therefore our encoding is done similarly. It can be found that Drug/Narcotic and Other Offenses both peak on Wednesday and come down at the start and the end of a week. While Vehicle-theft and Larceny/Theft stays low during the weekdays but rises on the weekend. So day of week can be a critical variable in deciding the category of a crime incident.

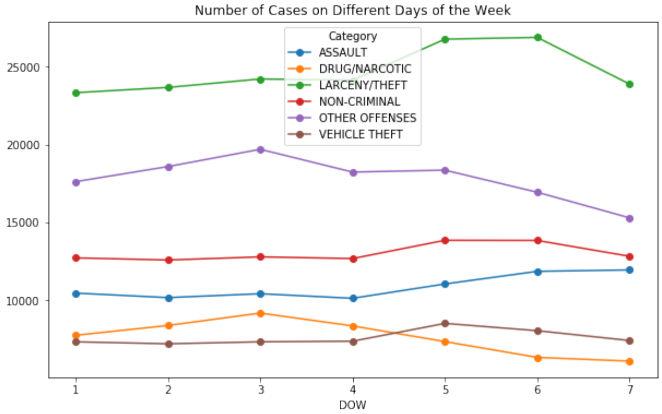

Fig. 5. Top Six Crimes on Different DayOfWeek

4) *Hour:*

Fig. 6 shows the top six most common crimes. In the figure we can observe some proportion changes among different crimecategories: Vehicle Theft rushes to a high level at 6 pm from a rather low one, Drug/Narcotic start to drop from 5 pm and it is the only crime that do not peak at 12 pm while other crimes all tend to do so. Hence we may want to try hour variable in our classification model. We can also easily catch a fun phenomenon that in three to five oclock in the morning the occurrence of criminal incidents is the lowest among the whole day.

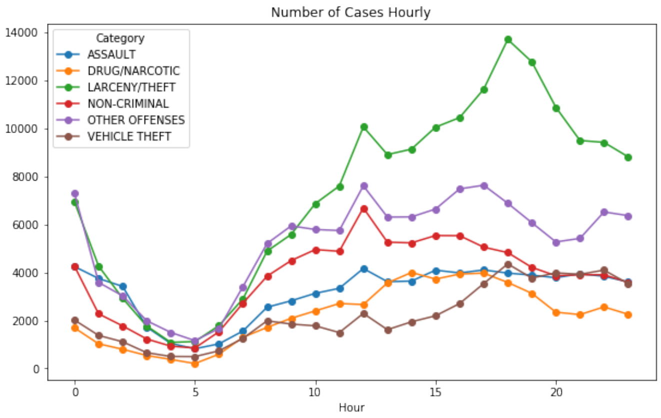

Fig. 6. Top Six Crimes at Different Hours

*5) **WorkDays**:*

We realized that there are certain holidays across the year apart from the weekend. Certain festivals are celebrated on a large scale, like Mardi Gras parade or Memorial day parade. Committing a crime is very easy in such situations as the offender can easily escape. We incorporated the following holidays and the weekend and made a bool feature informing whether the given day was a working day or not. This helped us improve our accuracy. The following days were incorporated:

1. New Years Day: Monday, January 1
2. Birthday of Martin Luther King, Jr.: Monday, January 15
3. Birthday of George Washington (Presidents Day): Monday, February 19
4. Memorial Day: Monday, May 28
5. Independence Day: Wednesday, July 4
6. Labor Day: Monday, September 3
7. Columbus Day: Monday, October 8
8. Veterans Day: Monday, November 12
9. Thanksgiving Day: Thursday, November 22
10. Christmas Day: Tuesday, December 25

C. Location Features

Location is another important feature of a crime incident. It may have crucial effect on predicting crime categories too. In the dataset we have four entries which are PdDistrict, Address, X and Y describing the location of incidents. Here we will research how we can use these four entries to make effective predictions.

*1) **PdDistrict**:*

PdDistrict stands for police department district. According to the dataset, there are ten of them. **TABLE II** shows number of crimes in each PD district in a descending order of numbers. We can see that Southern is a high-occurrence region of crime. Its incidence of crime is nearly 3.5 times of it in Richmond, the region with the lowest occurrence of crime.

TABLE II
CRIME INCIDENT COUNT FOR EACH PD DISTRICT

PD District	Count
SOUTHERN	155408
MISSION	118448
NORTHERN	104118
BAYVIEW	88382
CENTRAL	84461
TENDERLOIN	80878
INGLESIDE	77949
TARAVAL	64783
PARK	48756
RICHMOND	44690

Fig. 7 shows the bar graph of the top 10 most common crimes occuring per PD District. We inferred that Drug/Narcotic crimes have a high concentration in Tenderloin but rarely happen in other places, Larceny/Theft crimes concentrate in Central, Northern and Southern but can hardly be found in Bayview and Tenderloin.

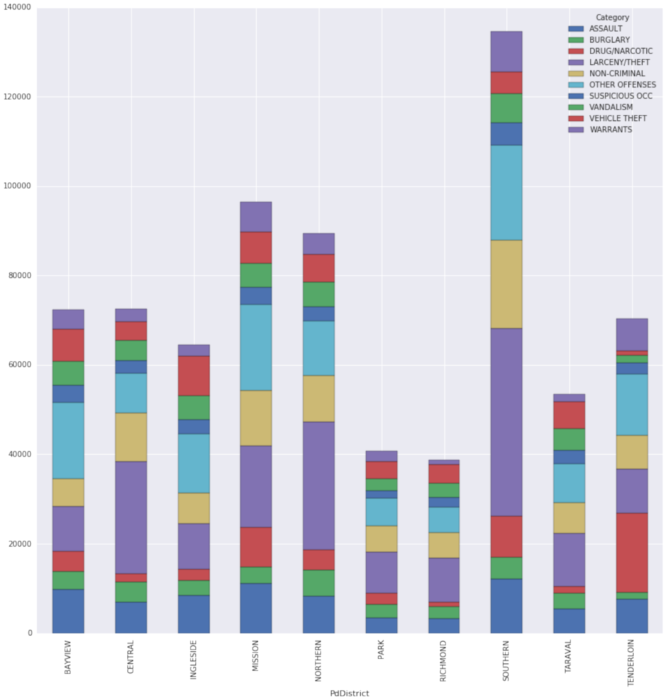

Fig. 7. Top 10 Most Common Crimes across PDs

Regarding above observation, we infer that the PdDistrict entry can play an important role in predicting crime categories.

*2) **Address**:*

Some good information in the address consists of street full name and a street suffix abbreviation. But it is difficult to figure out a good use of the street full name for it could be redundant with the latitude and longitude variable or with the PD District variable. Also, if you treated it as a category feature, there would be a huge vector for there are too many of different street in a city and it could make the model very difficult to be optimized. Thus we didnt work much on the

Address entry and we just split it into two parts, StreetNo and Intersection(Bool).

3) *Latitude and Longitude:*

The entry X and entry Y provide the latitude and longitude information of the crime incidents, which leaves us much space to play with data visualization.
Fig. 8 shows the heat map for all the crime categories (borrowed from scripts on Kaggle) We discovered that these heat spots are very different concerning different crime categories. That means crimes have their characteristic clusters on the map.

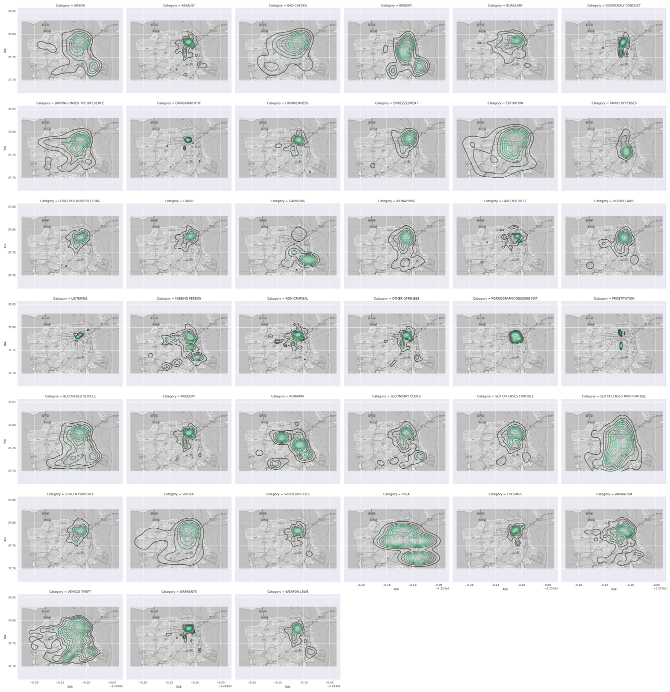

Fig. 8. Top 10 Most Common Crimes across PDs

III. Other Factors

A. *Crime Description and Resolution*

Besides variables related to time and location, the data set also provide us information about detailed description of the crime incident(entry Description), how the crime incident was resolved(entry Resolution). Although, if one consider crime classification as the prediction task, these two entry will not be given. However, we found one interesting thing which tells the efficiency of the various Police Departments tackling top crime of their area which is shown in **Fig. 9**.

B. *Multi Crimes*

We discovered an under-the-surface phenomenon which is very interesting and could stand a good chance to play an important role in the prediction model. Occasionally, multiple crimes occur at the same location, date and time. Those are probably related, i.e. one evil deed consisting of several crimes. We take that into account by adding a new column with the number of Multi crimes (MultCount) **TABLE III**.

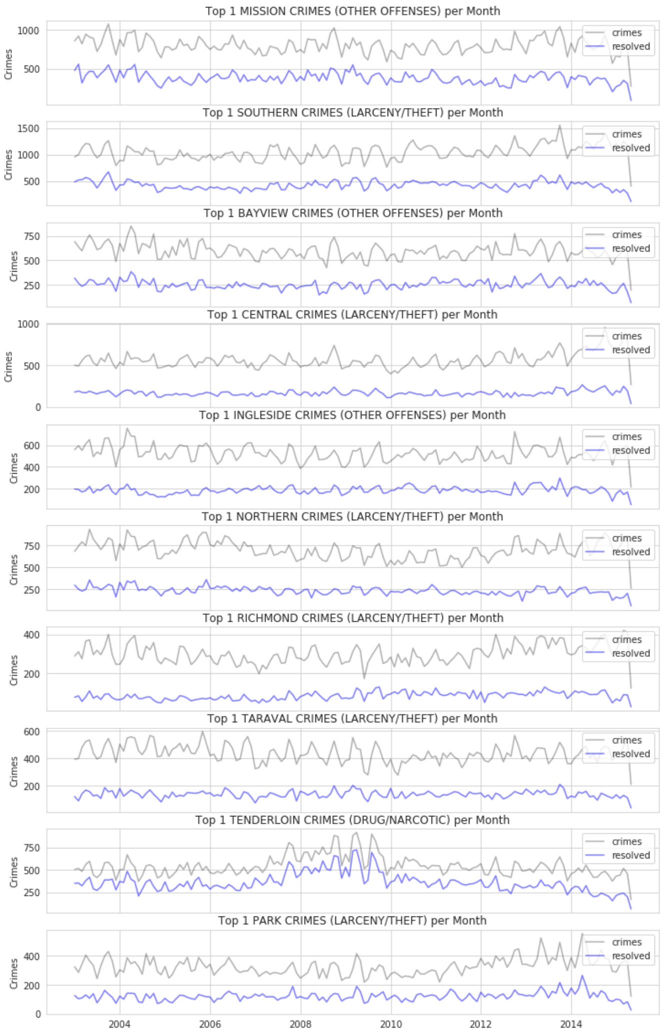

Fig. 9. Crimes Occured and Resolved per District per Month

+
TABLE III
Multiple Crime

Multi-Count	Count
1	546415
2	96770
3	33000
4	4261
5	1328
6	455
7	155
8	82
9	35
10	16
11	11
12	6
13	5
14	1
16	1

We plot the probability of each category being part of different numbers of simultaneous crimes as in **Fig. 10** . On the x axes we have the number of simultaneous crimes in one incident. It can be easily found that certain kinds of crimes do happen almost exclusively in multiple incidents, which includes Disorderly Conduct, Drunkenness, Embezzlement, Extortion, Gambling, Runaway, Kidnapping, Offenses Non Forcible, Vandalism, etc. While at the meantime, some kinds of crimes more tend to happen in isolation, such as Larceny/Theft, Suicide, Arson, Bad Checks, etc.

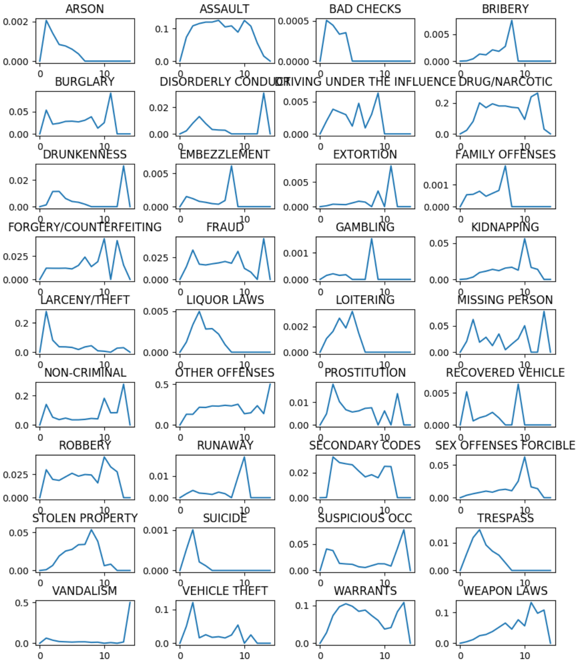

Fig. 10. Probabilities Crimes of Being Part of Simultaneous Crimes

IV. Prediction Task

A. *Description of the Task*

The task is to predict the Category of Crime that occurred given time, location and other related information. As described above, we have to predict a label, so we have to solve a classification problem.

B. *Evaluation*

We split our data into two equal parts, i.e. 50% each. The first part of the dataset was used as the training set while the second part of the dataset was used as the validation set.
Each crime incident has one true label and for each one in the validation set, we calculate the probabilities of it belonging to every category. The prediction results are then evaluated using the multi-class logarithmic loss as described on Kaggle. The formula is as follows:

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

where,
N - the number of cases in the validation set
M - the number of class labels
y_{ij} - 1 if observation i is in class j and 0 otherwis
p_{ij} - the predicted probability that observation i belongs to class j.

C. *Pre-Processing*

We know that features can be extracted from the row data including features related to time, location and a new feature we found, called Multi-Crimes. The following **TABLE IV**. gives a brief description of what all we did and why we did it.

TABLE IV
Summary of Features

Feature	**Format**	**Pre-processing**
Year	Categorical	Parsed from Timestamp
Month	Categorical	Parsed from Timestamp
Hour	Categorical	Parsed from Timestamp
Day Of Week	Categorical	Parsed from Timestamp
PdDistrict	Categorical	Label encoded
Street No	Numerical	Text extraction
Intersection	Boolean	Text Extraction
Co-ordinates	Numerical	Given (Scaled) and Wrong Data Correction
Multi-Crimes	Numerical	Summing up incidents that happen together
WorkDay	Boolean	Classifying each day as Working day or not
Nightime	Boolean	Classifying nightime for each day based on Seasons

Another interesting thing that we realized was the implicit feature of Multi-Crimes. From **TABLE III**. we can see that it is rare to have more than three incidents happen simultaneously. As in **Fig. 11**, for different numbers of multi-crimes, the probability of crimes by their category are plotted. On the x axes we have all 36 crime categories and the y axes represent the probability of a certain kind of crimes happened in the certain number of multi-crimes.

V. Models

Due to the predictive nature of the classification problem, the relevant models are Naive Bayes, logistic regression and tree based learning models. Here we discuss the possible pros and cons of each type.

1) Bernoulli Naive Bayes
   - Description
     A probability based classifier which assumes that features are conditionally independent.
   - Pros
     - Fast convergence even on 800000 rows of data
     - Simple to apply and model
   - Cons
     - Certain features in our model are not independent of each other leading to bad results even after fine-tuning.

2) Logistic Regression
   - Description
     A regression model with its result converted to a probability value in [0,1] using a Sigmoid function.

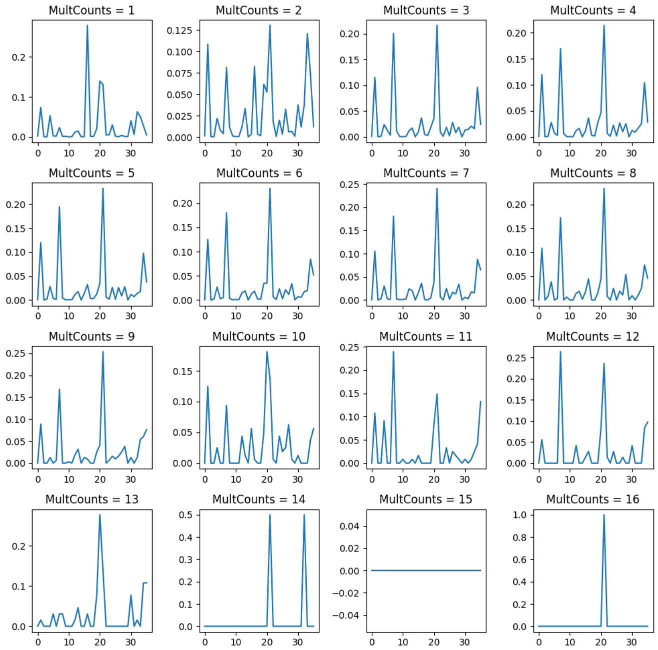

Fig. 11. Probabilities Crimes of Being Part of Simultaneous Crimes

- Pros
  - Considerably slower than Bernoulli NB but is manageable.
  - Good selection if there is a linear relation between features.
- Cons
  - Corner cases not considered and learning on mistakes is not done.
  - Perform poorly when features dont have a linear relation.

3) <u>Ensemble Learning (XgBoost)</u>

- Description
  A gradient boosting framework model where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.
- Pros
  - Since it creates trees, the user does not have to think much on the relation between features.
  - Very accurate and reliable
- Cons
  - Computing time increases exponentially with the size of the dataset.
  - Fine tuning Xgboost is very tough and another time-taking process.

Apart from these 3 main models, we tried Random Forests, KNN and Multinomial NB. The output of KNN and Multinomial NB was below average while Random Forest gave a result similar to XgBoost but converged at a larger value. So we discarded the above mentioned models as well. One model we couldnt try was SVM, which had given a good output on similar datasets.

VI. Model Validation & Results

We cross validated our features with the three selected models using different combinations. We inferred that all our features helped increase our accuracy and lower the log loss. Time feature extraction played an important role as did the implicit information provided via Multi-Crimes. The following **TABLE V** gives a small summary of our outputs:

TABLE V
RESULTS OF THE MODELS(ADDING FEATURES CONSECUTIVELY)

Feature	**Naive Bayes**	**Logistic Regression**	**XgBoost**
Time features+PdDistrict	2.662	2.843	2.554
+XYscaling	2.613	2.721	2.387
+WorkDay+NightTime	2.554	2.68	2.368
+MultiCount	2.523	2.657	2.315
+Address	2.511	2.621	**2.287**

VII. Conclusion

It can be seen that all the variables we have analyzed in the data preprocessing section are effective features in classifying crime incidents. Which in turn has verified our intuitions and hard-work towards the data.
We feel that there are two specific features that standout in improving our results significantly:

- Multi-Count
- Address Features - Intersection, StreetNo

Addition of these two features has led to a significant improvement in the Kaggle scores. In the aspect of model selection, we found that the Ensemble learing model outperformed the other two.
The Naive Bayes model performed poorly when possible interactions between features exist. In terms of speed, Naive Bayes is the fastest, normally it only need few seconds to converge. Logistic regression and XgBoost are rather time taking due to the higher accuracy they provide.
For most part of the training took place on an 8GB RAM system with Quad core. But later on we shifted to a Compute Engine on the Google Cloud Platform for faster training and results. The specifications of the Machine was 16 GB RAM with 8 high performance cores.
Our final score on the InClass competition is 2.24883 which makes us rank 9^{th} on the private leaderboard.

The Google Drive link to the pickle file is here.