San Francisco Crime Classification

Junyang Li
A53210366
jul309@eng.ucsd.edu

Shenghong Wang
A53224867
shw248@eng.ucsd.edu

***Abstract* - We consider the problem of predicting the category of crime that occurred given information regarding time and location. We explored relations between possible predictive features and the label and validated the effectiveness of the features on different classifier models including Naïve Bayes, Logistic Regression and Random Forest. We implemented the K-means clustering algorithm and performed optimizations to further improve the performance of the models. We also discovered some interesting phenomenon beneath the dataset.**

I. Introduction

San Francisco was founded on June 29, 1776[1]. The California Gold Rush of 1849 brought rapid growth, making it the largest city on the west coast at the time. From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz[2]. Today, the tech scene overwhelms its criminal past. However, rising wealth inequality and housing shortages make it not surprising for crime to keep springing up. The San Francisco Police Department published this dataset providing nearly 12 years of crime and their relevant record, to encourage people to mine more facts from it.

II. Dataset Exploration

The dataset contains nearly 12 years of crime reports from across all of San Francisco's neighborhoods, derived from SFPD Crime Incident Reporting system. It was originally provided by SF OpenData[3], the central clearinghouse for data published by the City and County of San Francisco. It was also provided by a machine learning competition known as *San Francisco Crime Classification*[4], which was hosted by Kaggle in 2016.

A. Overview

Time period the dataset covers is from 1/1/2003 to 5/13/2015 with 878,049 data points in total. Especially, it contains only even weeks for Kaggle has separated the whole dataset into training and testing set with every week rotated, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set.

Fig.1 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.

In the dataset, each crime record has nine entries of information related to the incident which are shown in **TABLE I**.

For categories of crimes, in the dataset there are 39 of them. **Fig.2** shows how crime incidents distribute in different categories. It is clear that the distribution is uneven. Crimes such as Larceny/Theft, Other Offenses, Non-Criminal, Assault, Drug/Narcotic and Vehicle Theft take huge portions while at the meantime certain kinds of crime such as Sex Offenses Non-Forcible, Gambling, Pornography/Obscene Mat are extremely rare.

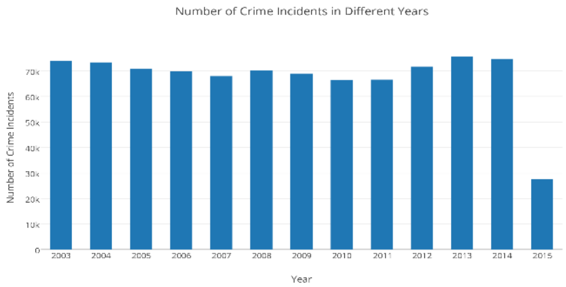

Fig. 1. Number of Crime incidents in Different Years

TABLE I
Information provided by the dataset per incident

Entry	Description
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

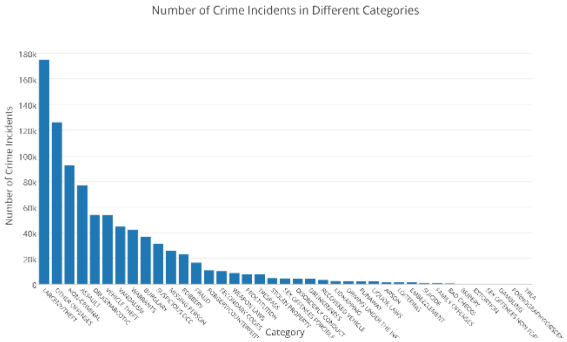

Fig. 2. Number of Crime incidents of Different Categories

B. Variables Related to Time

Time is always the first thing we mention when we report an incident. Here we want to figure out if time variables have effects 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'. Considering date, minute and second are trivial in effecting crime classification, only year, month and hour are researched here. The 'DayOfWeek' provides on which day of week a incident happened, it is another time variable that is researched.

1) Year

Fig.3 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 suppose that the year variable could be a critical feature in crime classification model.

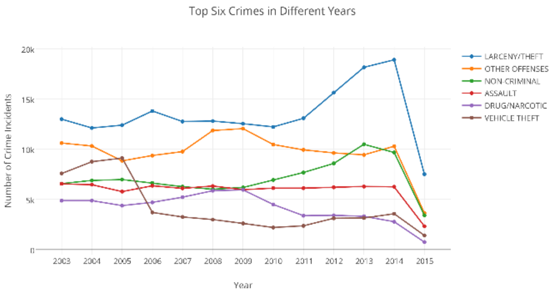

Fig. 3. Top Six Crimes in Different Years

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. Here's 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.

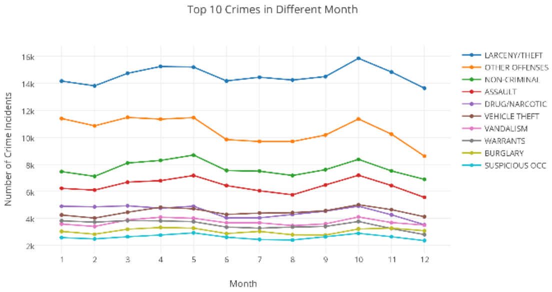

Fig. 4. Top Ten Crimes in Different Month

3) Day of Week

In **Fig.5** , we only let the fifth to tenth top commonest crimes show because their numbers are similar in scale so the change on proportion can be more easy to capture. It can be found that Drug/Narcotic and Warrants both peak on Wednesday and come down at the start and the end of a week while Vandalism is just the opposite. Also, the highest occurrence of Theft, Burglary and Suspicious activity are all on Friday. So day of week can be a critical variable in deciding the category of a crime incident.

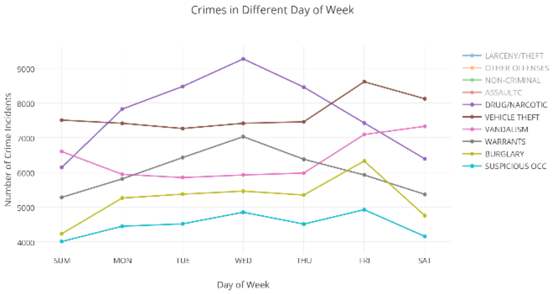

Fig. 5. Crimes in Different Day of Week

4) Hour

Fig.6 shows number of top ten commonest crime with the 1st one shadowed because it is in a huge scale of number compared to others and it would make the change not so obvious. In the figure we can observe some proportion changes among different crime categories: Vehicle Theft rushes to a high level at 6 pm from a rather low one, Warrants and Drug/Narcotic start to drop from 5 pm and they are the only two crimes 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 o'clock in the morning the occurrence of criminal incidents is the lowest among the whole day.

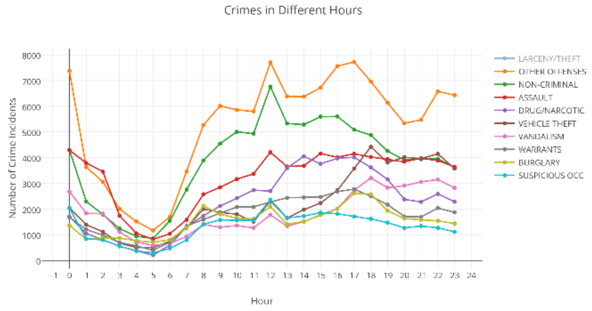

Fig. 6. Crimes in Different Hours

C. *Variables Related to Location*

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're 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. **Fig. 7** is a heap map of all 39 categories of crime's distribution in nine different PdDistrict. Two critical variance can be observed here. The first one is each kind of crime tends to aggregate in a couple of certain PD districts while tends to be very rare in some other PD districts. For instance, 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. The second variance here is that for different PD districts, the proportion of crime categories is various. Bayview and Ingleside mainly suffer from Other Offenses. Central and Northern mainly suffer from Larceny/Theft. And for Tenderloin, it mainly suffers from Drug/Narcotic. Regarding above observation, we infer that the PdDistrict entry can play an important role in predicting crime categories.

TABLE II
Crime Incidents Count for each PD District

PD District	Crime Incidents Count
SOUTHERN	157182
MISSION	119908
NORTHERN	105296
BAYVIEW	89431
CENTRAL	85460
TENDERLOIN	81809
INGLESIDE	78845
TARAVAL	65596
PARK	49313
RICHMOND	45209

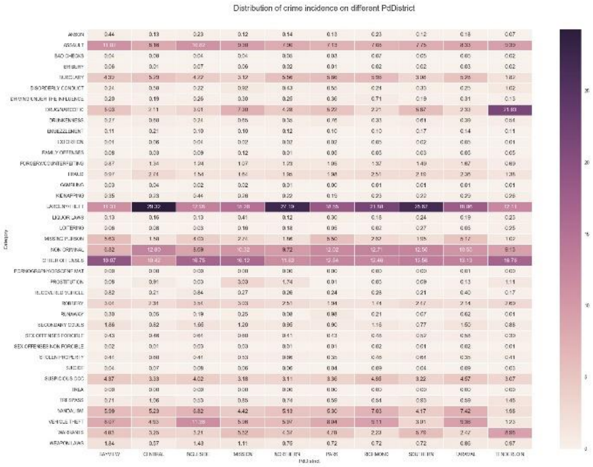

Fig. 7. Distribution of Crime Incidents on Different PD District

2) *Address*

Information in the address consists of a street full name, and a street suffix abbreviation. Some good Information in the address consists of a street full name, and a street suffix abbreviation. Some good instances could be: "1400 Block of GOLDEN GATE AV", "200 Block of EVELYN WY", or "MENDELL ST / HUDSON AV". In above instances, "AV", "WY", "ST" are all street suffix abbreviation. It's 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.

However, in this dataset, we found that there are only 15 different street suffix abbreviations. According to the reference table of US Postal

Service[5], the meaning of these 15 abbreviation are listed below in **TABLE III**. We wondered how different category of crimes distribute on different kind of address, i.e. would a certain crime tends to only happen on highway for example? Or would a certain crime never happen at some place like plaza or terrace? So first we extracted those suffix abbreviations from address text. Besides the condition of one suffix abbreviation comping along, address can also be two abbreviations with a “/” in the middle. We treat that as “Intersection”. Then we studied the variance of crime categories proportion on different kinds of street. **Fig. 8** is a heat map showing distribution of crime incidents on street suffix. Looking at the heat map, some significant trends of certain crime categories are clearly higher or lower in a particular street suffix can be observed. For example, most Suicide incidents happened on highway, Prostitution tends to only happen in intersections and Stolen Property and Drug/Narcotics crimes are concentrated in plaza. The information suggest us that the street suffix in the address entry can be extracted as an important feature. They also could be very interesting facts to be discovered.

TABLE III
Reference Table of the U.S. Street Suffix Abbreviations

Abbreviation	Street Suffix Name
AV	Avenue
BL	Boulevard
CR	Circle
CT	Court
DR	Drive
LN	Lane
PZ	Plaza
PL	Place
RD	Road
ST	Street
TR	Terrace
HWY	Highway
HY	Highway
WY	Way
WAY	Way

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. Borrowed some scripts from Ben Hamner[6] and DBenn[7], we plotted density maps of all 39 crime categories as showing in **Fig.9**. We discovered that these heat spots are very different concerning different crime categories. That means crimes have their characteristic clusters on the map which infers that some clustering algorithms could be used in the prediction task.

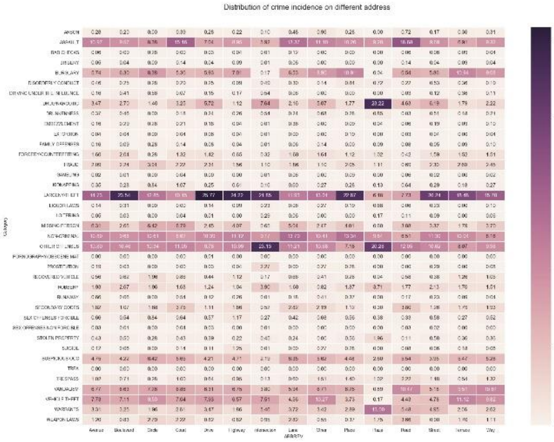

Fig. 8. Distribution of Crime Incidents On Different Address Abbreviations

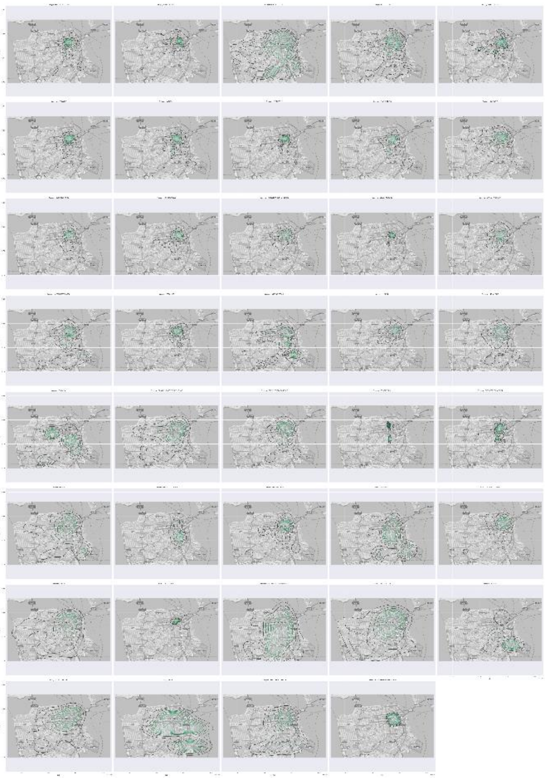

Fig. 9. Heat spots on map for All Crime Categories

D. *Other Factors*

1) *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”). However, if one consider crime classification as the prediction task, these two entry will not be given. So we will not discuss how them would affect crime categories’ proportion.

2) *Simultaneous Crimes*

Besides information provided explicitly in the data set, we have discovered an “under-the-surface” phenomenon which is very interesting and stands a good chance to play an important role in the prediction model. What we have found is that in our data, occasionally, more than one crime incidents could happen at exactly a same time and a same location. We sum up all occurrences of such simultaneous crimes and the result is listed in **TABLE IV**. One explanation of such phenomena could be that sometimes, one commission of crime can be convicted of multiple charges. For instance, say an store was robbed by an armed man, this case could be categorized into “Stone Property”, “Robbery”, “Weapon Laws”, “Trespass” at the same time. If the man was not caught yet, “Runaway” could be added. We then refer that some crimes might only happen in multiple crimes cases for it’s impossible for them to come along, such as “Weapon Laws” and “Runaway”. For those crime categories, other charges seem to be inevitable.

To verify the conjecture, for each crime in all 39 categories in the dataset, we plot the probability of it 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.

TABLE IV
Reference Table of the U.S. Street Suffix Abbreviations

Number of Crimes happened simultaneously	Count
1	550477
2	97879
3	34002
4	4358
5	1392
6	467
7	165
8	80
9	39
10	18
11	11
12	6
13	5
14	1
16	1

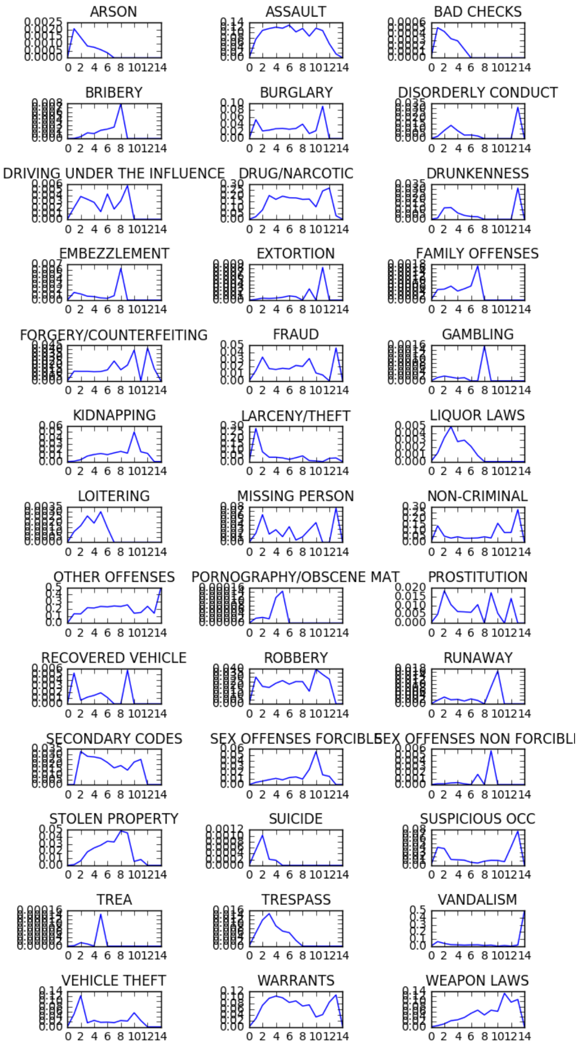

Fig. 10. Probabilities Crimes of Being Part of Simultaneous Crimes

III. PREDICTION TASK

A. *Task Description*

Our predictive task is to predict the category of crime that occurred given time and location information. Because the label we want to predict is categorical, so this task is a classification problem.

B. *Evaluation*

To evaluate our prediction models’ performance, we have split our data into two parts. The first part accounts for 60% of the whole dataset and was used as the training set. The second part accounts for 40% of the whole dataset and was used as the validation set.

Each incident in the data set has exactly one true labeled class. For each incident in the validation set, we calculate probabilities of it belonging to every category. Prediction results are then evaluated using the multi-class logarithmic loss[8] as defined in the following formula:

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

where **N** is the number of cases in the validation set, **M** is the number of class labels, **log** is the nature logarithm, $\mathbf{y_{ij}}$ is 1 if observation **i** is in class **j** and 0 otherwise, and $\mathbf{p_{ij}}$ is the predicted probability that observation **i** belongs to class **j**.

C. Feature Extraction and Data Pre-processing

From the analysis of the data exploration section, we know that features we can extracted from the row data include features related to time, features related to location and an implicit feature related to simultaneous crimes. While the process of we performing some data pre-processing procedure to extract information into useful features and the reason why we did it that way is stated in details in the following part of this section, **TABLE V** is a summary of what features we have at hand.

TABLE V
Summary of Features At Hand

Feature	Format	Size	Pre-processing
Year	One-hot Encoding	13	Parsing to timestamp format.
Month	One-hot Encoding	12	
Hour	One-hot Encoding	24	
Day of Week	One-hot Encoding	7	
PD District	One-hot Encoding	10	None.
Street Suffix	One-hot Encoding	15	Text Extraction.
Simultaneous Crimes	One-hot Encoding	5	Summing up incidents happened at same time and same location.
Coordinate	One-hot Encoding	K	Wrong Data Correction; K-means Clustering.

1) Time Features

For features related to time, we are interested in year, month, hour and day of week and they are provided in the “Dates” entry and “DayOfWeek” entry in the dataset. This part of data is pretty clean. The only thing we needed to do is to parse the timestamps into formatted date variables so later on we can easily extract the part we want. This job was done by using the “pase_dates” function in Pandas.

Moreover, in the data exploration section, we have found that the probability of an incident belonging to a certain type of crime does not change continuously based on the linear change of a certain time variable. In another word, each value of a certain time variable has its particular condition, treating time feature as continuous variables would be impropriate. So finally we treated all time related variables including year, month, hour, day of week as categorical features using one-hot encoding.

2) Location Features

For features related to location, we are interested in PD district, address, latitude and longitude.

PD district is certainly not a linear variable and it certainly does not have any linear relation with the probability. So again we used one-hot encoding and treated it as a categorical feature.

In the data exploration section, we have learned that the street suffix in the address may be an important feature in predicting crime categories. So we extracted street suffixes from the text in the “Address” entry. There are 16 different kinds of suffix and they indicate 14 kinds of place plus the case of intersection. Just like what we did with the PD district feature, we treated this feature as a categorical one and encoded it into 15 classes.

From **Fig.9** the heatmap we have found that different types of crimes tend to have their unique heat spot of density on the map. If we figure out which heat spot an incident happened in, we can make good predictions based on that. Clearly the latitude and longitude variable can help us locate an incident precisely and a clustering algorithm can help us classify the location into one heat spot. Considering the above condition, we decided to perform K-means clustering on all data point in the dataset based on their latitude and longitude information provided in the “X” entry and the “Y” entry. However, according to the satellite map of the city of San Francisco, values of the latitude variable are supposed to be in the range of [-122.5247, -122.3366] and values of the longitude variable should be in [37.699, 37.8299]. But we found that in some data points, the coordinates were way beyond that range. Apparently, some values were recorded by mistake.

Fortunately, there are comparatively few instances of wrong coordinates only 143 cases in 878,049 data points in total. So finally we impute the wrong values with the median for the corresponding PD district. We then used K-means clustering to categorized all incidents into K clusters where K is a variable representing the number of clusters, labeled each incident with the code of the cluster it belongs to and converted the XY coordinate into another categorical feature of clusters. Samey, we used one-hot encoding to represent the feature.

3) The Implicit Feature of Simultaneous Crimes

In the analysis of section 2.4.2, we realized that considering how simultaneous crimes can be used to classify crimes is necessary. From **TABLE IV** we can see that it's rather rare to have more than three incidents happen same time same location. So we plotted figures in another way. As in **Fig.11**, for different numbers of simultaneous crimes, the probability of crimes by their category are plotted. On the x axes we have all 39 crime categories and the y axes represent the probability of a certain kind of crimes happened in the certain number of simultaneous crimes. From the figure, we find significant differences in the distributions of crimes in "crime number = 1", "crime number = 2", "crime number = 3", "crime number = 10" and "Crime number = 13". For crime numbers equal to 4 to 9, their distribution are similar to "crime number = 3". For crime numbers equal to 11 to 12, they have similar distribution with "crime number = 10". And for crime numbers are bigger than 13, we can say their distribution are similar to "crime number = 13". Due to the above analysis, we decided to divide simultaneous crimes cases into five groups: crime number = 1, crime number = 2, crime number = [3, 9], crime number = [10, 12], crime number ≥ 13. Then we labeled all data points with the group it belongs to and finally made simultaneous crimes a categoric feature using one-hot encoding.

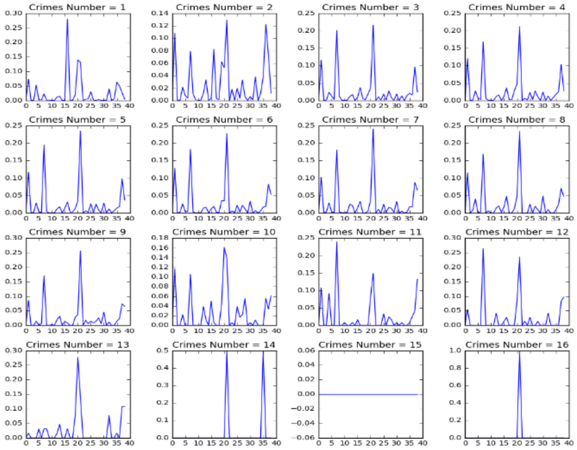

Fig. 11. Probabilities Crimes of Being Part of Simultaneous Crimes

IV. Model Design

A. Options Comparison

Due to the predictive task is a classification task, models from class which are relevant here includes Naïve Bayes, logistic regression and SVM. Here we discuss advantages and possible downsides of these three options.

1) Naïve Bayes

- Description

A simple probabilistic classifier based on the assumption that features are conditionally independent given the label.

- Advantages

- Simple;
- Converge very fast.

- Disadvantages

- Perform badly when features are not independent with each other. Specifically, in our case, PD district and clusters on map may have interactions.

2) Logistic Regression

- Description

A regression model with its result converted to a probability value in [0,1] using a Sigmoid function.

- Advantages

- A little bit slower than NB but still pretty fast;
- The output can be interpreted as a probability;
- Fit for linear features. So it would work with our features for they can all be converted into one-hot encoding format.

- Disadvantages

- No special attention is paid to the difficult cases and the number of mistakes is not optimized;
- When there's huge amount of very easy point and a few number of difficult point, it will ruin the boundary;
- Only fit for linear features. Perform poorly when features don't have linear relation with the label.

3) SVM

- Description

A more sophisticated classifier, focus on very difficult classification problem by minimizing the misclassification error.

- Advantages

- High accuracy.

- Disadvantages

- Not very efficient. Can take a long time to compute when the dataset is rather big.

We had decided to try all these three model at first before we found out that the time SVM required to converge was unbelievably long. And the outcome was very close to the outcome of logistic regression. So finally we discarded the SVM model and only used Naïve Bayes and logistic regression.

Other methods which are not covered in the class includes random forests, XGBoost, neural network, etc. Random forests are an ensemble learning method which are very popular now for classification tasks. They construct multiple decision trees at training time and output the class that is the mode of the classes. Compared to the other two methods(XGBoost and neural network), random forests are easier to understand, fast in speed, won't require huge memory space when operating, also don't have too much parameters to tune. So we selected random forests as the third model for our task.

V. Model Optimization

We did model optimization after we had finished validations towards different feature combinations and had settled one combination which yielded the best result.

For logistic regression, we changed the value of the regularizer, computed log loss separately, and located the sweat point between under-fitting and over-fitting.

For random forest, parameters we have tuned are n_estimators and max_depth. N_estimators is the number of trees to build before taking the maximum voting or averages of predictions. The model is under-fitting when that value is too small and over-fitting when it's too large. Max_depth is a limitation of how deep the decision trees are built. It can prevent the model from over-fitting but the model also suffers from under-fitting when the value is too small. Again we tried different values and located a sweat point.

For Naïve Bayes, because the other two models outperformed it by too much in the feature validation part and the options of optimization towards this model was very limited, we didn't optimized it at last.

VI. Related Literature

A. Related dataset and similar dataset

Compared with the dataset we used hear, other similar dataset that includes more information have been explored. Lam used this dataset and explored weather information from WeatherData library according to time that is given in dataset[9]. He got the relation between weather and the probability of crime, that is, it's more possible for crime to happen in cold weather compared with hot weather.

B. Related methods

We were also considering three state-of-the-art methods that are shared on Kaggle and are enlightened by their analysis on features. However, finally we didn't use these methods due to their drawbacks on this prediction task.

K-nearest neighbor (KNN) is also a classification algorithm. A participant in Kaggle used KNN and made a conclusion that log-loss will reduce with the increase of K. However, he only got log-loss as 25 when k is 50. This is because dataset is highly unbalanced. According to **Fig.2**, the most frequent crime happened the 2×10^4 times more than the least frequent crime, which makes it easily to predict minority crime as majority crime. Besides, KNN is relatively slower than other classification algorithm. For each point, the distance from this point to any other points that are already classified need to be computed, which makes classification process slow.

XGBoost is a fantastic open source implementation of Gradient Boosting Machines, a supervised learning method. A participant in Kaggle used XGBoost and get minimum log-loss as 2.47915[10]. He analyzed the time feature and extracted "season" from "day". Artificial Neural Net(ANN) is also a popularity classification algorithm and is widely used in handwritten and image recognition. A participant in kaggle used Keras package to implement ANN algorithm and got minimum log-loss as 2.55[11].

There also have participants using Naïve Bayes and get better result than the teams who used much more complicated models. And they concluded that on this prediction task, the key to improve predictive accuracy is to analyze the features and extract effective information rather than using fancy model. We have similar conclusion that the improvement of accuracy getting from changing methods is not as obvious as the improvement getting from mining data. The examples we list above indicate that complicated models are not necessary cause higher accuracy.

VII. Results and Conclusions

A. Baseline

The result of baselines for Naïve Bayes, logistic regression and random forests are shown in **TABLE**

VI. The minimum log-loss comes from the random forests model.

TABLE VI
Results for Baseline(Log-loss)

Feature	Naïve Bayes	Logistic Regression	Random Forests
PdDistrict	2.616885	2.625201	2.615285

B. Validation for Time Features

Due to the analysis of the dataset exploration section, our expectation was that the year variable, the hour variable and the day of week variable all would reduce the log-loss to some degree while the month variable might be irrelevant to the task. However, our result shows that all four time features are effective in predicting the crime categories. Results are listed in **TABLE VII**.

TABLE VII
Results for Time Features(Log-loss)

Feature	Naïve Bayes	Logistic Regression	Random Forests
PdDistrict+Hour	2.584778	2.595492	2.581139
PdDistrict+Hour +Year	2.566155	2.577017	2.565990
PdDistrict+Hour +Year+Month	2.565681	2.575701	2.563479
PdDistrict+Hour +Year+Month+ DayOfWeek	2.563626	2.572314	2.548479

C. Validation for Implicit Features

Implicit features which are not given directly in the dataset includes suffix of street and potential simultaneous crimes. Surprisingly, these two features have significantly reduced the log-loss after being added to the model. Relevant results are listed in **TABLE VIII**.

TABLE VIII
Results for Implicit Features (Log-loss)

Feature	Naïve Bayes	Logistic Regression	Random Forests
Baseline+Time +StreetSuffix	2.537182	2.524826	2.484574
Baseline+Time +StreetSuffix +SimultaneousCrimes	2.461018	2.403933	2.337733

D. Validation for the K-means Algorithm

We applied K-means algorithm when clustering data points on the map based on their latitude-longitude coordinate. Results are shown in **TABLE IX.** The clustering did help further reduce the log-loss and we found that 40 was the optimal value of K. The best result is 2.308712 coming from the Random Forests model.

One thing should be noticed is that after adding the cluster feature, both logistic regression and random forest became more predictive. However, Naïve Bayes yielded worse results compared to it without the cluster feature. The reason could be that the this feature is a location-related feature and it is not independent of the other two location features which is the "PD District" and the "Street Suffix". The Naïve Bayes model might double counted their effect and thus made poor classifications. So in the model optimization part we just discarded this model.

TABLE IX
Results for Different Values of K(Log-loss)

K	Naïve Bayes	Logistic Regression	Random Forests
30	2.518361	2.378812	2.316051
35	2.512877	2.378341	2.31708
40	2.505112	2.368901	2.308712
45	2.514870	2.371892	2.314686
50	2.505122	2.366834	2.312148
60	2.497122	2.365937	2.314413
70	2.493330	2.363788	2.314795

E. Model Optimization

We further optimized the logistic regression model and the random forests model with the optimal feature combination which was "PdDistrict + Year + Month + Day of week + Hour + Street Suffix + Number of simultaneous crimes + 40 Clusters".

Results of tuning the regularizer of logistic regression model are plotted in **Fig.12**. When lambda equals to 0.1 the Log-loss is the minimum.

Results of tuning the number of decision tress and the max depth of tree are plotted in **Fig.13** and **Fig.14**. When n_estimator equals to 225 and max_depth equals to 25, the Log-loss is the minimum.

The final optimal results of the two model are shown in **TABLE X**.

TABLE X
Final Optimal Results

	Logistic Regression	Random Forests
Log-loss	2.353283	2.298326

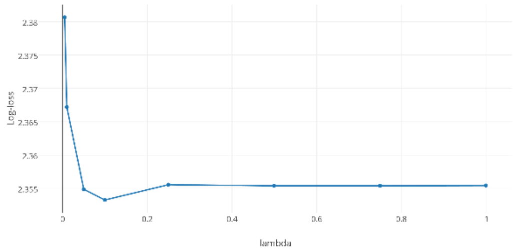

Fig. 12. Process of Tuning the Regularizer of the Naïve Bayes Model

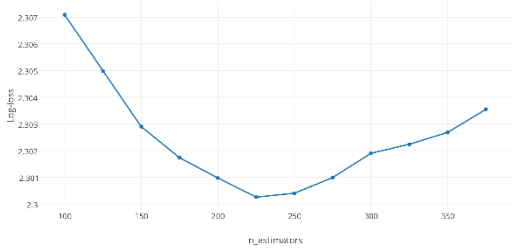

Fig. 13. Process of Tuning the Number of Decision Trees of the Random Forest Model

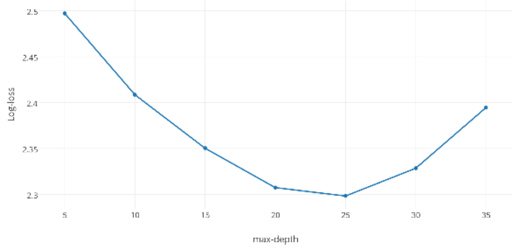

Fig. 14. Process of Tuning the Max Depth of Decision Trees of the Random Forest Model

F. Conclusions

In the aspect of feature selection, it has been proven that all the variables we have analyzed in the dataset exploration section are effective features in classifying crime incidents, which has verified our conjectures towards the data in the dataset exploration section.

The "Number of simultaneous crimes" has been proven to be the most predictive one in the Random Forest Model. That feature was not given explicitly in the dataset and we started to notice it from a minor but interesting phenomenon. We then kept mining it, and have discovered some underlying pattern. We then came up a rational way to convert this phenomenon into a categoric feature and finally made it into our models and have significantly improved the performance of the models. The whole process make us realized the importance of data mining and feature engineering.

The second important feature is the "Street Suffix" feature. And from all suffixes, the one has the largest parameter is "intersection".

In the aspect of model selection, we found that the random forest model outperformed the other two. The Naïve Bayes model performed poorly when possible interactions between features exist. In terms of speed, Naïve Bayes is the fastest, normally it only need less than 2 seconds to converge. Logistic regression and random forests are rather efficient due to their high accuracy. On average, it costs them around 2 minutes and 5 minutes respectively. However SVM was not efficient at all. It may only fit for small-scaled experiments that requires low error rate. We also find that K-means algorithm requires much space of memory when operating it on PCs. To operate clustering with a number of clusters greater than 30, we had to use a PC with 16G RAM.

Our final score on Kaggle is 2.29875, which makes us rank 15% on the leaderboard.

REFERENCES

[1] San Francisco on Wikipedia. https://en.wikipedia.org/wiki/San_Francisco
[2] Background of the San Francisco Crime Classification competition on Kaggle. https://www.kaggle.com/c/sf-crime
[3] SF OpenData. https://data.sfgov.org
[4] Kaggle machine learning competition: San Francisco Crime Classification. https://www.kaggle.com/c/sf-crime
[5] US Postal Service: Street Suffix Abbreviations. http://pe.usps.gov/text/pub28/28apc_002.htm
[6] Ben Hamner: Show Map Image in Python. https://www.kaggle.com/benhamner/sf-crime/show-map-image-in-python
[7] DBenn: Crime Density by Location. https://www.kaggle.com/dbennett/sf-crime/test-map
[8] Multi-class Logarithmic Loss. https://www.kaggle.com/wiki/MultiClassLogLoss
[9] Correlation between weather condition and the type of crime. https://nycdatascience.com/ correlation-between-weather-condition-and-the-type-of-crime
[10] K-nearest-neighbor. https://www.kaggle.com/wawanco/sf-crime/k-nearest-neighbour/code
[11]Neural net using Keras package. https://www.kaggle.com/ smerity/sf-crime/fighting-crime-with-keras.