Data Mining Final Project
San Francisco Crime from *Kaggle*

Yu-Han Chen
Department of Computer Science
and Information Engineering
102401053@cc.ncu.edu.tw

ABSTRACT
Crime prevention has always been the major task of the police. However, in reality, due to the limited resource, the performance on these tasks is constrained. San Francisco is notorious of its high crime rate, which is rated the top third dangerous cities in the US. If the police are able to be aware of the crimes at certain location in advance, there will be a great possibility that they can forestall crimes and lower the crime rate. Nevertheless, there are all kinds of crime, and again the resources are limited, thus, trying to prevent all kinds of crime is only a goal but it's not efficient. As a result, I decided to focus mainly on "Index Crimes". In this paper, we utilized the dataset derived from SFPD Crime Incident Reporting System and extracted only the data that are considered as Index Crime by its category to be our training set. Given time and location, we predicted the crime categories that occurred from 2003 to 2015. I investigated classification models including *Random Forest Tree*, *Naïve Bayes*, *C4.5* and compared the performance among them.

Keywords
Classification, Imbalanced data, Under-sampling, C4.5, Random Forest, Naïve Bayes

1. INTRODUCTION
The goal is to predict the crime category from past records. In section 2, I will introduce the dataset I got originally and the manipulated data used in the experiment. In section 3, I'll present the exploration of the data. Section 4 is to introduce the models and the difference between the training data used in detail. Then, take a look at the evaluation of the models in section 5. Last, the conclusion will be discussed.

2. DATA SET
The dataset used is provided by *Kaggle*[1]; whereas my goal is distinct from the goal predefined by *Kaggle*. Thus, some manipulation is done on the original data to create a modified data for further experiment.

- **Features of the original dataset**
  - Dates - timestamp of the crime incident (2003-01-06 ~ 2015-05-13)
  - Category - category of the crime incident
  - Descript - detailed description of the crime incident
  - Day of Week - the day of the week
  - PdDistrict - name of the Police Department District
  - Resolution - how the crime incident was resolved
  - Address - the approximate street address of the crime incident
  - X - Longitude
  - Y – Latitude
- **Statistical data of the original dataset**

Crime Category	Counts
LARCENY/THEFT	174900
OTHER OFFENSES	126182
NON-CRIMINAL	92304
ASSAULT	76876
DRUG/NARCOTIC	53971
VEHICLE THEFT	53781
VANDALISM	44725
WARRANTS	42214
BURGLARY	36755
SUSPICIOUS OCC	31414
MISSING PERSON	25989
ROBBERY	23000
FRAUD	16679
FORGERY/COUNTERFEITING	10609
SECONDARY CODES	9985
WEAPON LAWS	8555
PROSTITUTION	7484
TRESPASS	7326
STOLEN PROPERTY	4540
SEX OFFENSES FORCIBLE	4388
DISORDERLY CONDUCT	4320
DRUNKENNESS	4280
RECOVERED VEHICLE	3138
KIDNAPPING	2341
DRIVING UNDER THE INFLUENCE	2268
RUNAWAY	1946
LIQUOR LAWS	1903

[1] https://www.kaggle.com/competitions

ARSON	1513
LOITERING	1225
EMBEZZLEMENT	1166
SUICIDE	508
FAMILY OFFENSES	491
BAD CHECKS	406
BRIBERY	289
EXTORTION	256
SEX OFFENSES NON FORCIBLE	148
GAMBLING	146
PORNOGRAPHY/OBSCENE MAT	22
TREA	6

- **Features of the modified dataset**
  - Year – 2003~2015
  - Month - 1~12
  - Hour - 00 ~24
  - Time – Morning (4: 00 ~ 12:00) , Afternoon (12:00 ~ 20:00) , Night (20:00~4:00)
  - Category - category of the crime incident
  - Day of Week - the day of the week
  - PdDistrict - name of the Police Department District
  - X - Longitude
  - Y – Latitude
- **Statistical data on the modified dataset**

Crime Category	Counts
LARCENY/THEFT	174900
ASSAULT	76876
VEHICLE THEFT	53781
BURGLARY	36755
ROBBERY	23000
SEX OFFENSES FORCIBLE	4388
ARSON	1513

In the original data set, there are totally 39 categories of crime, among which the top frequent crimes are Theft, Other Offenses, Non-criminal, Assault, Drug /Narcotic according to Figure 1.

However, I want to especially focus on Index Crimes so what exactly is an index crime? Index Crimes are offenses that are more seriously. According to some research conducted, San Francisco is found to be a dangerous city when comparing to other cities in the US. The probability of a violent crime occurs is 8 out of 1000 residents; however, the national median is only 3.8 out of 1000 residents. Index Crimes may be more lethal, so allocating more resources to index crimes will be more cost-efficient. Hence, I extracted the data that the crime types belong only to the following seven categories to form the new modified dataset.

Index Crimes:

1. LARCENY/THEFT
2. BURGLARY
3. ASSAULT
4. ARSON
5. VEHICLE THEFT
6. ROBBERY
7. SEX OFFENSES FORCIBLE

3. EXPLORATORY ANALYSIS

After exploring the data, we can tell the SOUTHERN police District department has the greatest amount of crimes reported from Figure 2. The relation between time series and categories occurrence is shown in Figure3. From Fig 3, it's apparent that the peak of crime committed is during 17:00 to 20:00, thus, the police should strengthen patrols to avoid crimes from occurring. The dataset is also analyzed by DayOfWeek, however, the hours of a day tell better story about when crimes happen most frequently. Moreover, to explore the data in a geometric manner, I drew a scatterplot of the seven different crimes in different colors on the map in Figure 4. In the last figure, Figure 5, the crime category: SEX OFFENESES FORCIBLE occurred in each PdDistrict is shown. Last, take a look at the boundary and the location of each PdDistrict in Figure 6.

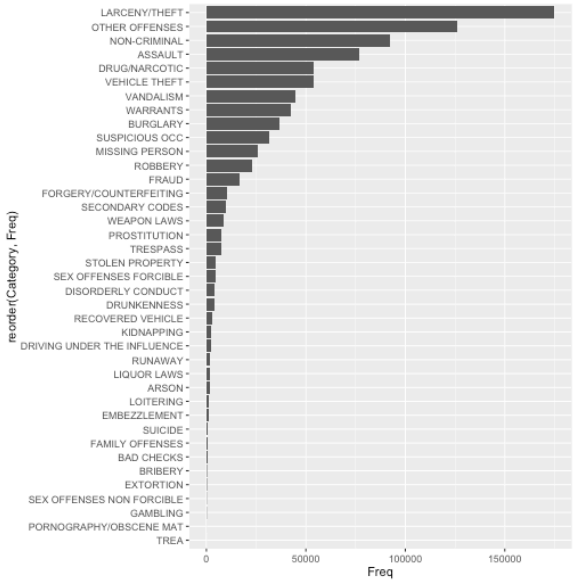


Figure 1: Crime Distribution

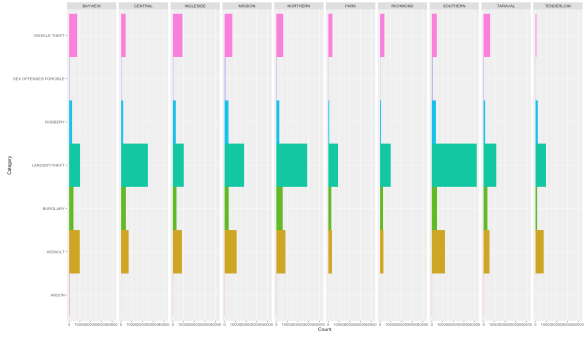


Figure 2: Crimes by PdDistirct

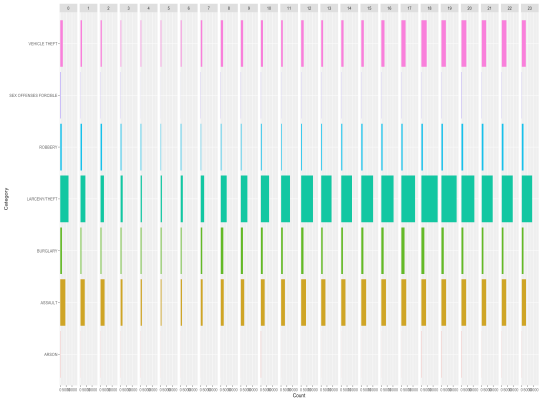


Figure 3: Crimes by Hour

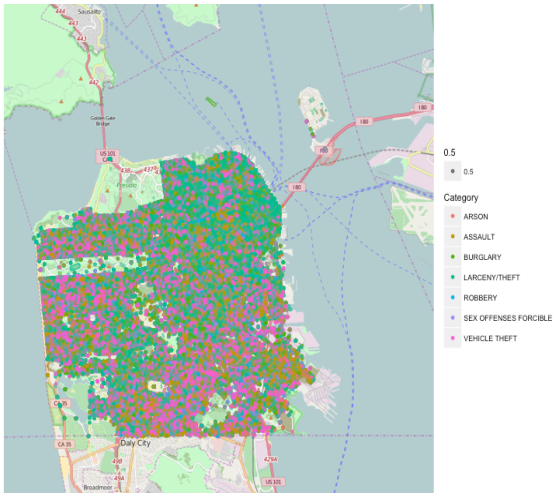


Figure 4: Index Crimes over the city

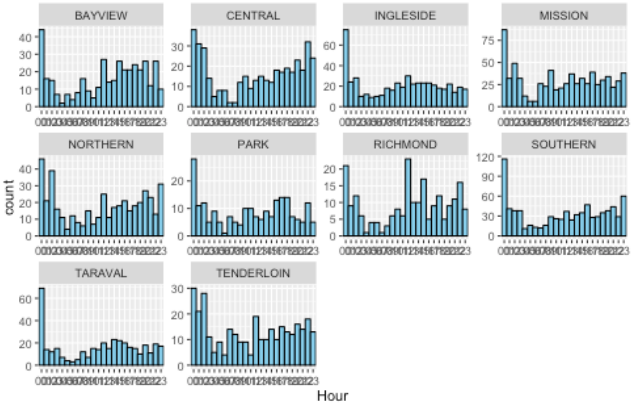


Figure 5: Sex Offenses forcible to PdDistrict

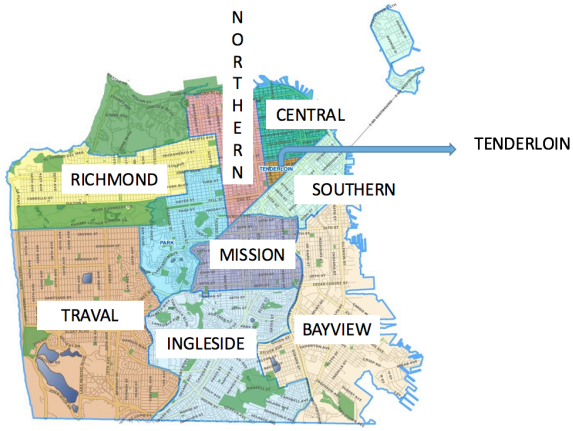


Figure 6: PdDistrict boundary

4. MODELS

In this section, the four models experimented will be introduced. Before introducing the models in detail, I'll first explain the two different training data used.

If you take a look at the number of instances in each crime category in the Index Crime data set, the data set is a highly imbalanced data. Due to the imbalance of the data, the performance of the models is greatly reduced. As a result, I resampled from the Index Crime dataset as another training data to compare the influence of imbalanced data on the model's performance. The purpose of resampling is to decrease the gap between the number of instances in the top crime category (Larceny/theft) and the bottom crime category (Arson).

I divided both the Index Crime dataset and resampled dataset to 66% to 34% as training and testing data. I'll present the evaluation result on both the Index Crime and Index Crime resampled dataset.

- **Index Crime dataset**

Crime Category	Counts
Larceny/Theft	174900
Assault	76876
Vehicle Theft	53781
Burglary	36755
Robbery	23000
Sex Offenses Forcible	4388
Arson	1513

- **Index Crime resampled dataset**

Crime Category	Counts
Larceny/Theft	6783
Assault	5235
Vehicle Theft	4789
Burglary	3849
Robbery	3444
Sex Offenses Forcible	3000
Arson	1513

4.1. *Naïve Bayes* Classifier

Naïve Bayes [2]is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

The reason *Naïve Bayes* is chosen because it not only is a simple but also a reliable classifier. Although it's simple, it often outperforms other complex models in many complex real-world situations.

- **Index Crime dataset**

	Training	Testing
Precision	0.397	0.420
ROC Area	0.598	0.616
Mean absolute rate	0.2197	0.2187

- **Index Crime resampled dataset**

	Training	Testing
Precision	0.533	0.557
ROC Area	0.858	0.860
Mean absolute rate	0.162	0.162**1**

4.2 *Random Forest* Classifier

Random Forest [3] is a classifier that constructs a multitude of decision trees. Its prediction is the mode of all the trees. This feature made the approach not be influenced by the imbalanced data. Therefore, I thought it is suitable for our data, so that this algorithm was selected as one of the experimented models.

- **Index Crime dataset**

	Training	Testing
Precision	0.503	0.499
ROC Area	0.665	0.624
Mean absolute rate	0.1628	0.1604

- **Index Crime resampled dataset**

	Training	Testing
Precision	0.757	0.572
ROC Area	0.947	0.827
Mean absolute rate	0.1049	0.1422

4.3 *C4.5* classifier

C4.5 [4]is an algorithm used to generate a decision tree and is also the top ranked popular classifier. C4.5 uses *information gain* to select features for tree nodes. Since the goal is to predict seven categories, a decision tree seemed like a suitable classifier. However, comparison between the performance of the three models is still needed.

- **Index Crime dataset**

	Training	Testing
Precision	0.516	0.484
ROC Area	0.675	0.629
Mean absolute rate	0.1808	0.1724

- **Index Crime resampled dataset**

	Training	Testing
Precision	0.592	0.601
ROC Area	0.830	0.828
Mean absolute rate	0.1434	0.1417

4.4 *Logistic Regression* classifier

Logistic regression [5] measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

- **Index Crime dataset**

	Training	**Testing**
Precision	0.572	0.357
ROC Area	0.647	0.649
Mean absolute rate	0.1422	0.1905

- **Index Crime resampled dataset**

	Training	Testing
Precision	0.568	0.575
ROC Area	0.879	0.879
Mean absolute rate	0.2805	0.1606

4.5 Proposed model

Because the number of instances of each category is still uneven, so there's a higher tendency for the model to predict the category to a crime category which has relatively dominant number of instances. Thus, precision isn't a good metric for evaluation. If the result all mainly predicted to a category, the model still can receive a high precision. As a result, we decide to use the ROC area of each category to evaluate on the models. Since it's obvious that models trained with resampled data performs better, so the ROC[6] area result of the model trained on the resampled data is presented.

[2] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[3] https://en.wikipedia.org/wiki/Random_forest

[4] https://en.wikipedia.org/wiki/C4.5_algorithm

[5] https://en.wikipedia.org/wiki/Logistic_regression

[6] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

- ***Naïve Bayes* model**

Crime Category	training data ROC Area	testing data ROC Area
Larceny	0.941	0.944
Assault	0.787	0.795
Arson	0.889	0.866
Robbery	0.928	0.919
Burglary	0.769	0.774
Vehicle Theft	0.784	0.788
Sex Offenses Forcible	0.909	0.907

- ***Random Forest model***

Crime Category	training data ROC Area	testing data ROC Area
Larceny	0.953	0.920
Assault	0.950	0.809
Arson	0.948	0.832
Robbery	0.983	0.922
Burglary	0.912	0.685
Vehicle Theft	0.922	0.704
Sex Offenses Forcible	0.966	0.884

- ***C4.5 model***

Crime Category	training data ROC Area	testing data ROC Area
Larceny	0.905	0.905
Assault	0.801	0.802
Arson	0.902	0.898
Robbery	0.689	0.692
Burglary	0.924	0.930
Vehicle Theft	0.839	0.847
Sex Offenses Forcible	0.713	0.713

- ***Logistic Regression model***

Crime Category	training data ROC Area	testing data ROC Area
Larceny	0.948	0.949
Assault	0.846	0.852
Arson	0.896	0.881
Robbery	0.961	0.959
Burglary	0.777	0.775
Vehicle Theft	0.797	0.798
Sex Offenses Forcible	0.922	0.919

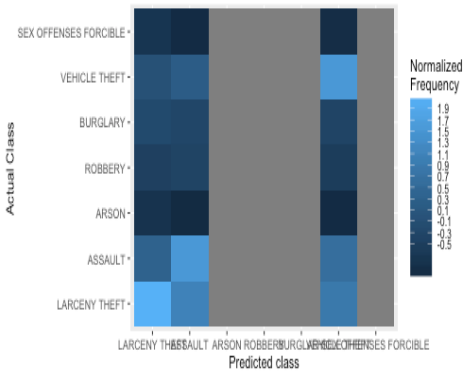
Surprisingly, *logistic Regression* classifier is the model with the best average ROC Area, so the logistic regression model trained with resampled data is the best model.

5. COMPARISON ON MODELS

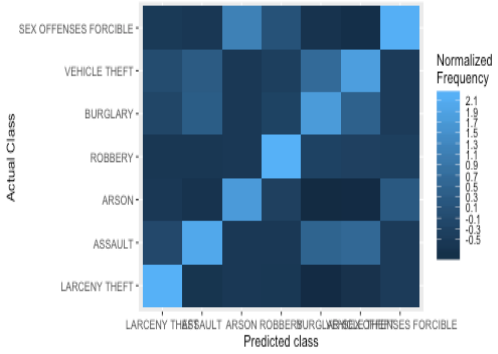
In this part, I present the confusion matrix of logistic regression model trained with Index Crime dataset and Index Crime resampled dataset to show that when it comes to imbalanced data, under-sampling is a reliable approach to improve the performance of the model.

From the below confusion matrixes, we can tell that the model trained with Index Crime dataset only predicted crimes as Larceny, Assault and Vehicle Theft. However, in the model trained with resampled data, the result is distributed over the seven categories.

Logistic Regression model (Index Crime dataset)



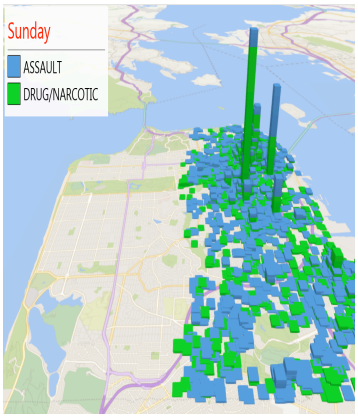
Logistic Regression (Index Crime resampled dataset)

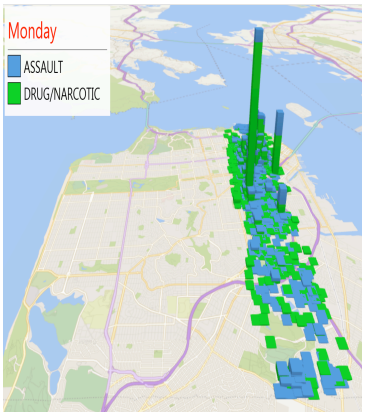


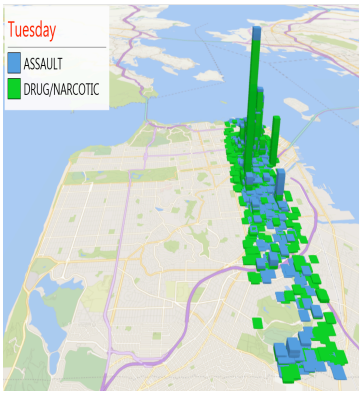
6. RELATED WORKS

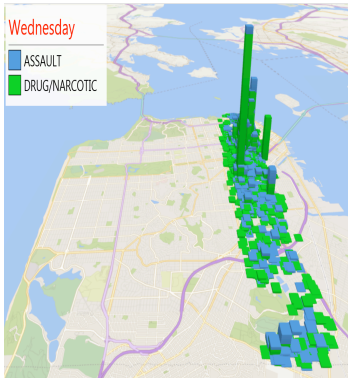
While I was conducting the experiment, I had an interesting finding: Drug/narcotic's occurrences highly gathered in Tenderloin and the usage of drugs seemed to be affect by Day of Week. Also, violent crimes like Assault also occurred frequently in this region.

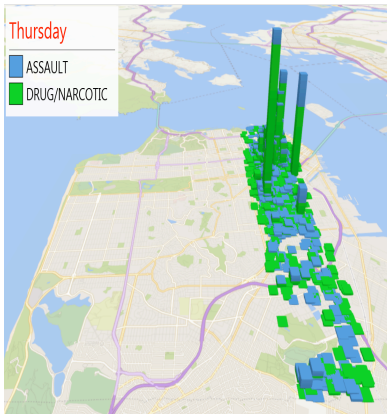
Below are the figures of the two crime categories' occurence with respect to Day of Week.

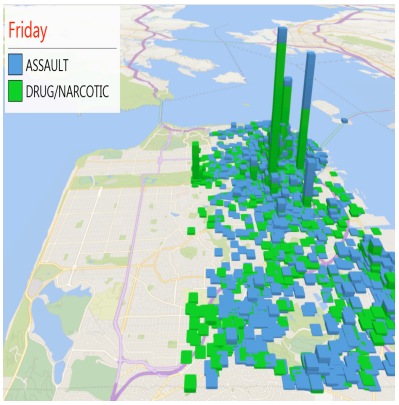


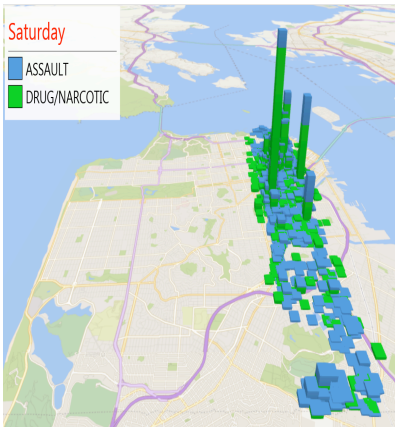












7. CONCLUSION

The prediction and prevention of crimes is an important but complicated task of public order. In order to improve the safety in San Francisco, it is necessary to predict category of crimes rapidly and correctly. To make the prediction more accurate and efficient, I focused on Index Crime in this experiment. However, in many classifiers that are trained with Index Crime dataset, the result was biased due to the imbalance of the data. Therefore, under sampling is conducted from the dataset to create a resampled dataset. After experimenting on the models, under sampling is an effective approach to deal with imbalanced data is successfully verified.

Secondly, while finding the relationship between category and Police District Department I figured that Drug doing especially gathers in the Tenderloin district. Furthermore, the occurrence of Assaults is much higher than average. Out of curiosity, I did a search on the Internet to learn more about the Tenderloin district.

Coincidently, according to Wikipedia, Tenderloin[7] is said to be a high-crime neighborhood in which the first block of Turk Street had one of the highest rates of violence and drug activity in San Francisco. This is an interesting finding in this research.

8. REFERENCES

[1] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[2] https://en.wikipedia.org/wiki/C4.5_algorithm

[3] https://en.wikipedia.org/wiki/Tenderloin,_San_Francisco

[4] https://www.kaggle.com/competitions

[7] https://en.wikipedia.org/wiki/Tenderloin,_San_Francisco