Business Analytics Term Project

Proposal Report

‘Crime Forecasting’

11146315 Seungwan Seo

12146305 Jegyeong Kim

Contents

1. Abstract

As crime rate increases and crime method has been complicated, the prevention of crime became important. So we would like to prevent crime by increasing the patrols in the right place and on right time based on data analysis. Basically, we analyze and predict what kind of crime is occurred by crime history data.

2. Analysis Background

Recently, random and violent crime is one of biggest of social problem. According to National Police Agency, the crime against women, children and the elderly is very serious and especially, the crime against woman accounts for 87 percent[1] of crime. The best way to handle crime is prevention, not after treatment. So we thought the way to manage the patrols efficiently and effectively is needed.

3. Purpose

We expect that we can know what kind of crime will be occurred based on our analysis. By using outcomes, increase the patrols in the area where the outcome was predicted crime will occurred. In this way, police can prevent more crime than before. As a results, the crime rate will be reduced. In other word, out purpose is that to reduce crime rate by increase patrol in right place and right time based on our analysis outcomes.

4. Data Acquisition

Before, we planned to use public open data of Korea at proposal stage. However there are only statistical result of crime, so we couldn't find connection between each subjects. For example, there is statistical result about the number of crimes by site and kind of crimes. Also there is the number of crimes by time and kind of crimes. However we couldn't know where and when the crime was occurred. So we inquired to National Police Office of Korea and required to get data of crime, but they said, they can't provide detail crime data for us.

After that, we kept finding other crime data and got the detail crime data with date, time, kind of crime, longitude, latitude and so on from Kaggle. It is data of San Francisco, but the procedure to predict crime is similar. Therefore if we have data of Korea that is similar to Kaggle data, we will be able to predict crime and get similar performance with this data.

[1] http://news.heraldcorp.com/view.php?ud=20150915000605&md=20150915112251_BL

5. Analysis

5.1. Data Exploration

There are 9 columns in data set.

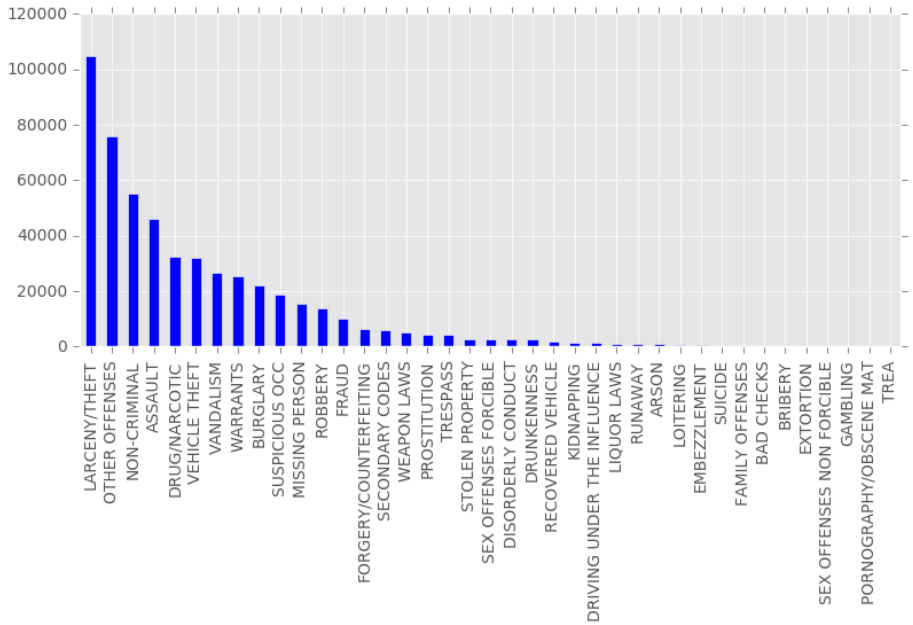
Dates - timestamp of the crime incident
Category - category of the crime incident.
Descript - detailed description of the crime incident.
DayOfWeek - the day of the week
PdDistrict - name of the Police Department District
Resolution - how the crime incident was resolved
Address - the approximate street address of the crime incident
X – Longitude
Y – Latitude[2]

Types of Data is like following matrix.

Type	Continuous Variables	Categorical Variables
Columns	Dates, X, Y	Category, Descript, DayOfWeek, PdDistrict, Resolution, Address

Then, we examined each columns in detail like how they look like, how they are distributed and so on. We made it as graph and the results are as following.

Figure 5-1. The number of crimes by category



[2] These descriptions of data is from Kaggle

Figure 5-1 shows distribution of the number of each crimes. As you can see, there are big gaps between categories of crimes and there are too many ‘larceny/theft’ cases. Because it can overwhelm other variables, we would consider solution for imbalanced data.

Figure 5-2. Crimes by year

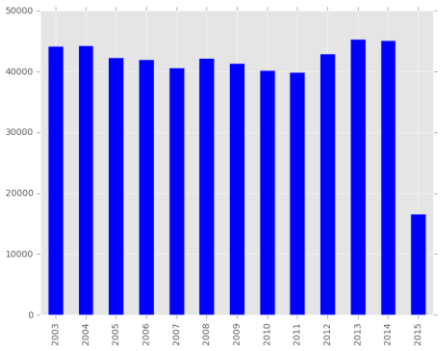


Figure 5-3. Crimes by week of day

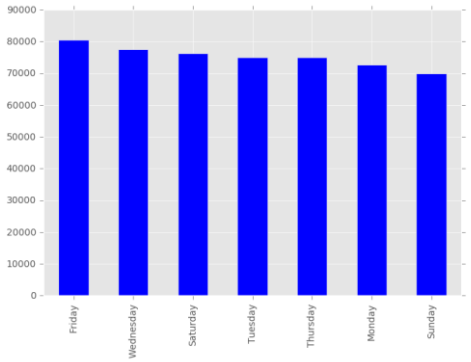


Figure 5-2 represents the number of crimes by year. The number of crimes by year increase slightly, but we thought it is no much influence. So we will compare model with ‘Year’ Column and model without ‘Year’ column later and select the model. And figure 5-2 explains the number of crimes by week of day. There are some differences between week of days, crimes were occurred most frequently on Friday.

Figure 5-4. Crimes by Month

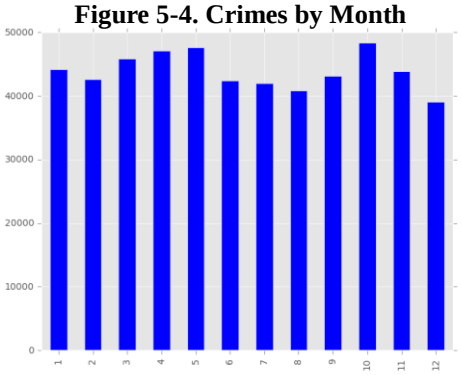


Figure 5-5. Crimes by hour

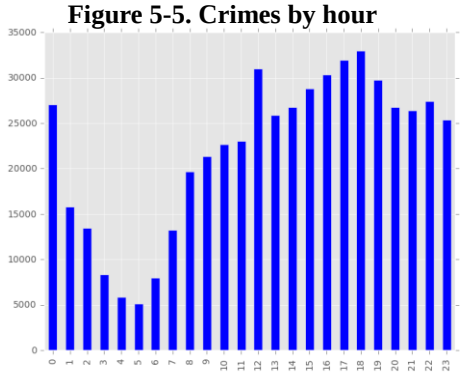


Figure 5-4 shows that the number of crimes by month. Crimes were occurred least frequently on December and most frequently on October. And figure 5-5 represents the number of crimes by hours. In the morning, the number of crimes is small, otherwise, the amount of crimes is very high in dinner time.

As a result of exploration data, we decided to focus on serious criminal offense because our purpose of analysis is to patrol and prevent crimes. So we selected category of crime under conditions that 1)is it serious criminal offense? 2)can we prevent the crime in advance by patrol? Then we replaced ‘Category’ column by binary. If it is included in our condition, it is class 1, otherwise, it is class 0.

The list of target class(1) is as followings; "ARSON", "ASSAULT","VEHICLE THEFT", "ROBBERY", "WEAPON LAWS", "DRUG_NARCOTIC", "TRESPASS", "RUNAWAY", "SEX OFFENSES FORCIBLE", "KIDNAPPING".

After this process, we explorated data again.

Figure 5-6

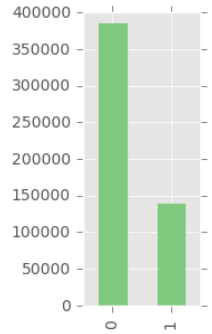


Figure 5-6 shows the number of crime as serious and minor offenses. Most of crimes are minor offenses and serious crime is about 30% of minor crime. The imbalanceness of this data is better than the data before being divided as binary, but we thought there are still slight imbalanceness between class 0 and class1. So we would consider solution for imbalnceness like over sampling and under sampling and compare that models with original model.

Figure 5-7 the number of category of crimes by year

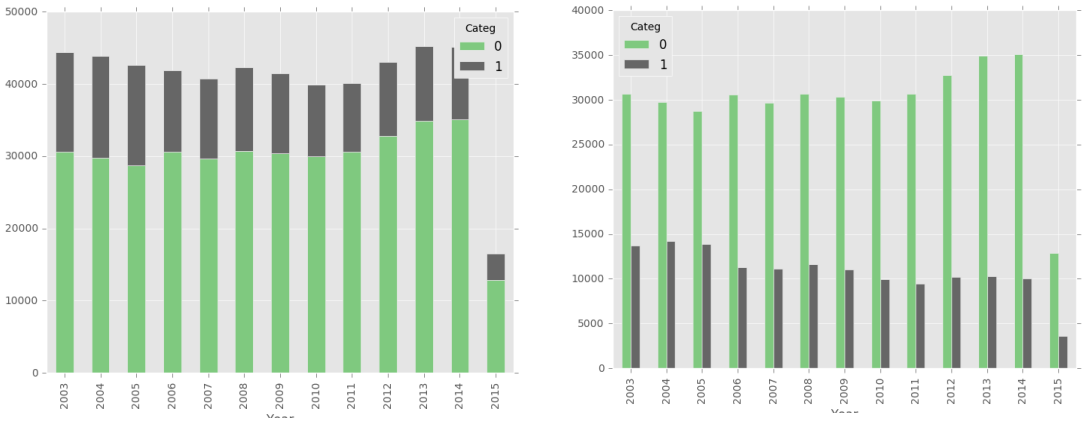


Figure 5-7 shows that the number of crimes as seriousness by year. According to figure 5-7, interesting is that the number of serious crime(class 1) decrease and the number of minor crime increase as time goes by.

Figure 5-8 the number of category of crimes by hour

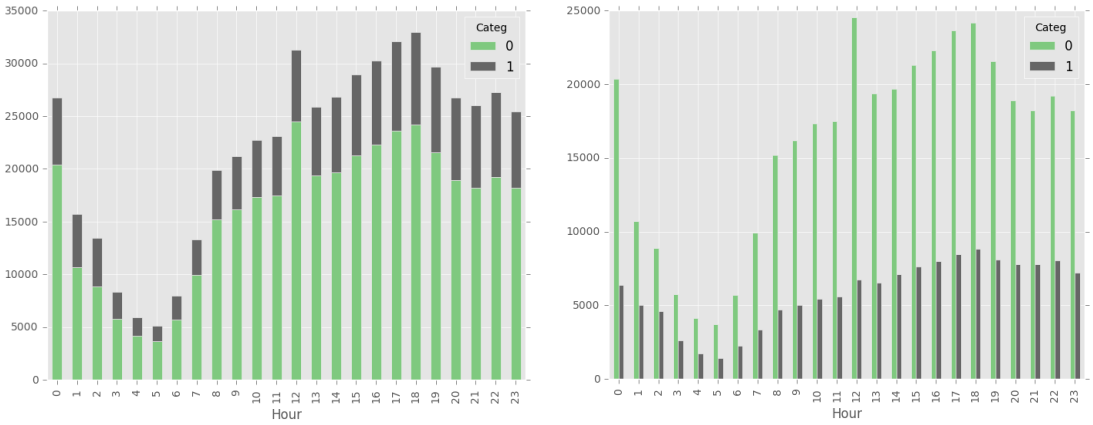


Figure 5-8 shows that the number of crimes as seriousness by year. The figure implies the minor offenses is more influenced by time comparing with serious crime. And figure 5-9 shows the number of crimes by month. The noticable thing is that both of serious and minor crime follows similar pattern and there are no big difference.

Figure 5-9 the number of category of crimes by month

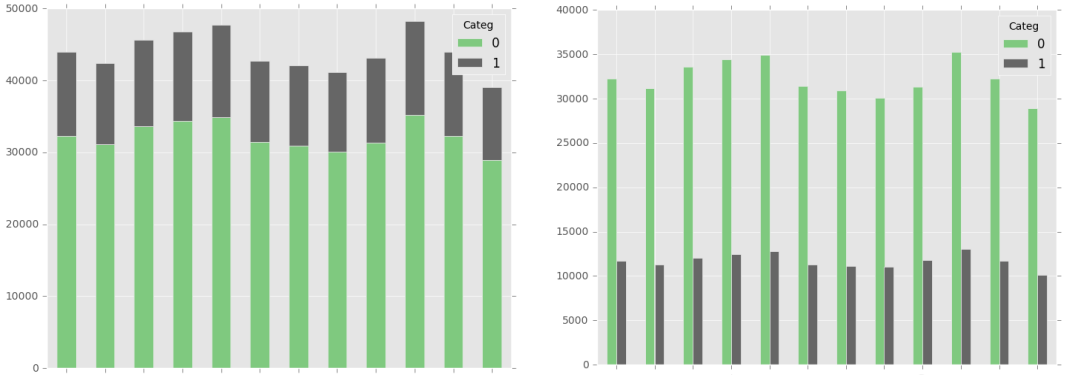


Figure 5-10 the number of category of crimes by day of week

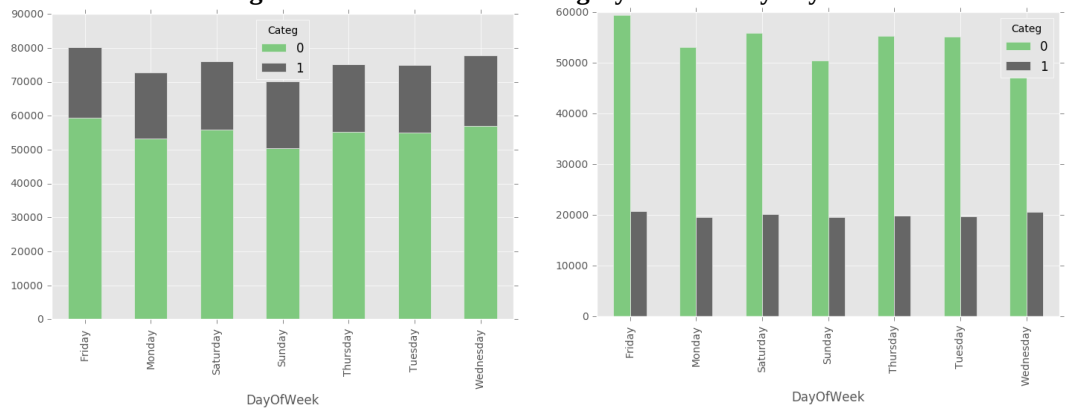
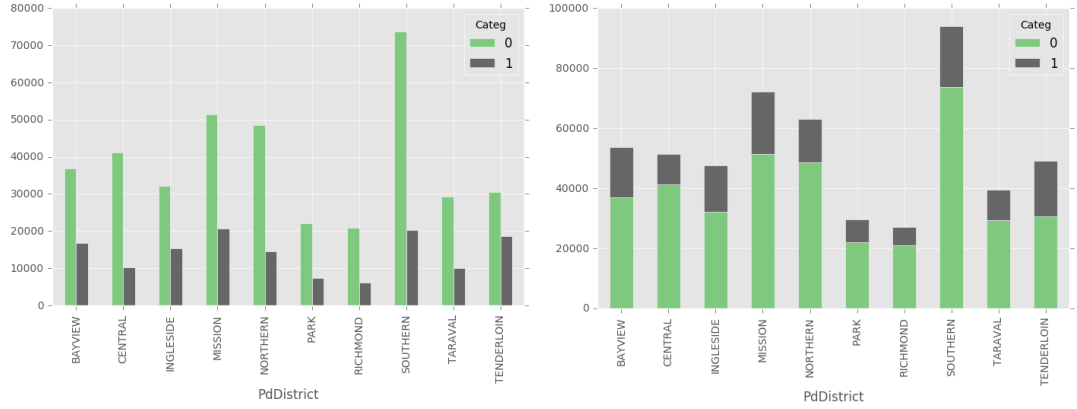


Figure 5-10 shows that the number of category of crimes by day of week. The figure explains that the number of minor crime is influenced by week of day. The interest thing is that serious crimes do not be affected by day of week at all and almost same regardless of day of week.

Finally, figure 5-11 represents that the number of crimes by police department district. There are 10 police department district. And there are characteristics by police department district. For example, 'Southern' has the highest proportion of minor offenses, but the number of serious offenses is the highest in 'Mission' district.

Figure 5-11 the number of category of crimes by Police Department District



5.2. Preprocessing

Some preprocessing procedures are included in 5.2. Data Exploration, so some process can be repeated.

5.2.1. Split Data

For validation and testing, we split data into train, validation and test data set. The proportion of sets are 0.6, 0.2 and 0.2. They were selected randomly among data.

5.2.2. Drop Columns

'Descript' is detailed explanation of category, so we deleted 'Descript' column. Also we thought that 'Resolution' do not affect to predicting crime and dropped the 'Resolution' column. And unique value of 'Address' is over 20,000. So if we make dummy variable for 'Address' column, there would be too many dummies to fit, so we delete 'Address' column also.

5.2.3.

We don't have much knowledge about time-series forecasting, so we divided 'Dates' into Year, Month and Hour (day of week is already in column). Then we will use that columns by categorical variable.

5.2.4. Convert Target Variable to Binary

we decided to focus on serious criminal offense because our purpose of analysis is to patrol and prevent crimes. So we selected category of crime under conditions that 1)is it serious criminal offense? 2)can we prevent the crime in advance by patrol? Then we replaced 'Category' column by binary. If it is included in our condition, it is class 1, otherwise, it is class 0.

The list of target class(1) is as followings; "KIDNAPPING", "ARSON", "ASSAULT","VEHICLE THEFT", "ROBBERY", "WEAPON LAWS", "DRUG_NARCOTIC", "TRESPASS", "RUNAWAY", "SEX OFFENSES FORCIBLE".

5.2.5. Make dummies

For categorical, we made a dummy variables. The variables to be dummy variable are (PdDistrict, Year, Month, DayOfWeek, Hour). We used OneHotEncoder.

5.2.6. Under sampling(Not Choose)

Major class has 644352 data points, while minor class has 233697 data points. We thought this differenciation could be meaningful. Therefore we did under sampling for better performance of our model. However the performance decresed.

5.3. Analyze

For us, Accuracy and Recall is most important score factors. The reason why recall is important than precision is as following explanation. Assume that if polices patrol, the

crime is prevented. Precision means how many times we prevent serious crime among patrols we go to. And recall means how many times we prevent serious crime among really occurred crimes. So we will consider accuracy and recall mainly.

5.3.1. Selecting Variables

Because our data is too big, most of mining method take too long time to fit our data. So we used Naïve Bayse mining for selecting variable because it is respectively faster than other. We decided 'X', 'Y', 'Time', 'Minute' is critical factors to patrol. We compared models by removing and adding rest of variables. As a result of comparing, there were no big differences when we removing and adding 'Year', 'Month', 'DayOfWeek'. So, we chose the model with all of variables because it has slightly high accuracy and other scores.

	Original	Without 'Year'	Without 'Month'	Without 'DayOfWeek'
Accuracy	0.6522	0.6586	0.6516	0.6458
Precision	0.3517	0.3428	0.3502	0.3487
Recall	0.3643	0.3087	0.3615	0.3618
F Score	0.3579	0.3249	0.3558	0.3645

5.3.2. Selecting Sampling Method

	Accuracy	Precision	Recall	F Score
Original	0.8417	0.7197	0.6636	0.6905
Under sampling	0.5786	0.5766	0.5840	0.5802
One class learning	0.7355	0.5782	0.0223	0.0431

We didn't apply over-sampling because our data is too big, so computation effor is considerable. Original model without any sampling has best performance.

5.3.3. Selecting Model

	Naïve Bayes (Gaussian)	Decision Tree	SVM	Logistic Regression	One class learning
Accuracy	0.6518	0.8417	0.6243	0.7335	0.7355
Precision	0.3506	0.7197	0.2777	0.4316	0.5782
Recall	0.3622	0.6636	0.2574	0.0050	0.0223
F Score	0.3563	0.6905	0.2672	0.0100	0.0431

The purpose of our model is prevent crime by patrol system. Therefore better recall score, better performance to us. In this situation, decision tree has highest score on all of diffusion matrix elements, we can choose our model easily.

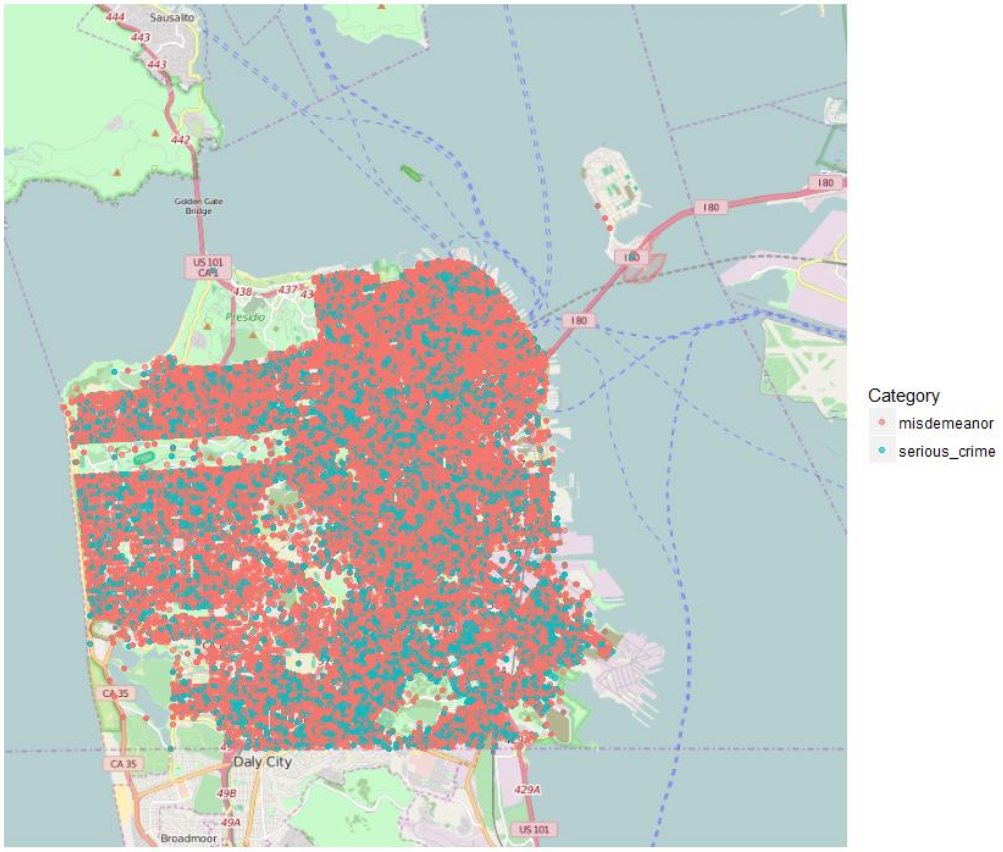
5.3.4. Performance and Visualization output

Final performance of our model is as followings.

Decision Tree	Accuracy	Precision	Recall	F-score
Train	0.8417	0.7197	0.6636	0.6905
Test	0.8396	0.7154	0.6611	0.6872

Score of train set is slightly higher than test set.

The following picture represents the position of crimes that are predicted to be occurred of test data set. Red color means the crime is not serious, the other way, blue color means that the crime is serious, such as assault, kidnap, and sex offense.



6. Conclusion

We can prevent serious crimes more than 7 times among 10 times by patrolling when we use decision tree algorithm. The accuracy of our model is a considerably meaningful especially to prevent crime situation. Police officer would be able to use this algorithm for getting information about where, when, and what kind of crime would be happened. They can get these information in real time and minute by minute. And they would be able to set a plan for prevent serious crime. We expect, using this way, police can prevent crime more than before. If we have data of Korea we will be able to predict crime and get similar performance with this data.