CHAPTER 1 INTRODUCTION

1.1 GOAL

Predict the category of crimes that occurred in the city.

1.2 DATA

Our data is San Francisco Crime Classification which from Kaggle (https://www.kaggle.com/c/sf-crime/data). In there, we have seven variances: timestamp of the crime incident, category of the crime incident, detailed description of the crime incident, the day of the week, name of the Police Department District, how the crime incident was resolved, and the approximate street address of the crime incident, where number of category of the crime is thirty-nine, and number of data is 878,016.

CHAPTER 2 CLEAN DATA

2.1. INTEGRATION

We found some category of crimes be similar, so we let those category of crimes merge, including SEX OFFENSES FORCIBLE and SEX OFFENSES NON FORCIBLE, RECOVERED VEHICLE and VEHICLE THEFT, FRAUD and BAD CHECKS, BURGLARY and LARCENY/THEFT.

2.2. DELETION

We delete less number some category of crimes, including ARSON, BRIBERY, DRIVING UNDER THE INFLUENCE, EMBEZZLEMENT, EXTORTION, FAMILY OFFENSES, GAMBLING, KIDNAPPING, LIQUOR LAWS, LOITERING, OTHER OFFENSES, PORNOGRAPHY/OBSCENE MAT, RECOVERED VEHICLE, RUNAWAY, SUICIDE, and TREA. In particular, OTHER OFFENSES is too difficult to analyze, because of too many uncertain category of crimes, we also delete OTHER OFFENSES.

2.3. OUTLIER

We first draw scatter plot of longitude and latitude, and we can found there is points at (-120.5,90), it is impossible because the data is in San Francisco, how could it happened crime at 90 degrees latitude.

2.4. CONCLUSION OF CLEAN DATA

By the above, we remaining number of data is 737,589 and number of category of the crime is twenty.

CHAPTER 3 ANALYSIS

3.1. METHOD1--PCA

We can found choose first three principle components had explain 90% of variation. So we draw PC1 and PC2 two dimension plot and PC1,PC2 and PC3 three dimension plot, but we can look any information.
Improve: maybe we can let had relationship category use same color, maybe we can look something.

3.2. METHOD2--HISTOGRAM OF VARIABLE

First, we plot histogram of the total criminal number corresponded to each year, each month, each week, and each hour. We observe that there is no difference between these classification methods, but we find out that criminal cases at 3-7 a.m. is less than other time for one day. And criminal cases increased significantly at 11-12 p.m. and 6-8 p.m. We guess there are less loitering people out of the street in the early morning, so the criminal cases also decreased. There are more people on the street during lunch time and dinner time, so the criminal cases also increased.

3.3. LINE CHART

Second, we want to understand every different situation for the criminal cases. We picked 20 colors to represent 20 categories of crimes. We use line chart to see the change process of the number of each categories to each year. We notice that the LARCENY/THEFT is the highest case in all category, and the VEHICLE THEFT, NON-CRIMINAL and ASSAULT are also higher one, but VEHICLE THEFT started decreasing after 2005. For classifying by month and week, the LARCENY/THEFT still the highest case in all category, and others change process is also similar to the former. For classifying by hour, we find out that it has no change with the total criminal times we observed before.

Finally , we added a new variable “PdDistrict” to observe the changes. In the aspect of the year, we can see that "LARCENY/THEFT" in "CENTRAL", "NORTHERN" and "SOUTHERN" have a very large number of criminal cases in the “PdDistrict”. In proportion, "NORTHERN" is San Francisco's richest district, "SOUTHERN" is a tourist attraction, and "TENDERLOIN" is a gathering center for social services and underground bars. This community in downtown San Francisco is a gathering place for homeless people with a relatively dense population. In the chart, we can also see that the number of drug categories is much higher than that of other districts, and it is declining year after year. Under the aspect of the month, we can see that the difference in quantitative proportions outside the fluctuations is in each district' s month by month and it is a little steady. In the aspect of the week, we can't see any regular changes. In the hour, it can be seen that there are regular changes in all districts. Generally, there are lowest points between 4 am to 5 am in the morning. After which the number rose until a peak at midday followed by a slight drop, began to rise again until a higher number of peaks appeared at 6 pm, the only difference being The number of cases in "CENTRAL", "NORTHERN" and "SOUTHERN" varies greatly in each category. It appears that the curve fluctuations are larger than those in other districts, and it is different from the highest number of criminal categories in other districts is "LARCENY/THEFT". The highest number of cases in the "TENDERLOIN" is "DRUG/NARCOTIC".

3.4. Xgboost

The result is not good, the one of reason is the data is not balanced, so some category will choose so much, and some category even less than three thousands sample, and we can choose train data balanced, maybe it will improve the result.

3.5. Map

We want to use map to observed distribution of crime, but if we draw all sample then it will not easy to observed variation, so maybe can use heat replaced points or else method. And if we want to observed more variation, we can through the time of crime to draw map.

CHAPTER 4 RESULT OF REPORT

We notice that the total criminal cases do not have apparently change for classifying by year, month, and week in chart. But it's apparently decreasing at 2-6 a.m. and increasing at 11 a.m.-12 p.m. and 6-8 p.m. We can give these graphs to the police as a reference in order to reducing the crime times by strengthening patrols at the time we observed that criminal time is increasing.

Actually, it's usually not balanced in training data and testing data by xgboost. Therefore, it's necessary to consider the balance of the data when we constructing the model.