PREDICTION OF CRIME CATEGORIES IN SAN FRANCISCO AREA

by

KEMIN XU

(Under the Direction of Jeongyoun Ahn)

ABSTRACT

Numerous information and data now are available to us with an increasing development of the Internet. Even though with so much useful and valuable information and data, there are still much to do and to think about how to make use of them. At the very beginning, it is required to discover the useful data for the research because different data are suitable for different researches. Then, what matters is how to put the data and information into reasonable use to construct a model to make prediction.

At the same time, machine learning also plays a very important role for the big data. Machine learning is a subfield of computer science which includes lots of useful methods. I will use both decision tree and random forest method in my analysis. All the three methods will be used for the two datasets from a data science competition website which are regarding the survival from the sinking of Titanic and the crime category of San Francisco respectively. The purpose of Sinking of Titanic is to predict which passengers survived the tragedy and the purpose of the crime category of San Francisco is to predict the category of crimes that occurred in the city. I will combine the all three models' results to see if it is helpful to the accuracy of prediction.

PREDICTION OF CRIME CATEGORIES IN SAN FRANCISCO AREA

by

KEMIN XU

B.S., Arizona State University, 2014

A Thesis submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Athens, Georgia

2016

PREDICTION OF CRIME CATEGORIES IN SAN FRANCISCO AREA

by

KEMIN XU

Major Professor: Jeongyoun Ahn

Committee: Jaxk Reeves

Liang Liu

Electronic Version Approved:

Suzanne Barbour

Dean of the Graduate School

The University of Georgia

May 2016

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

Page

Page

CHAPTER 1

INTRODUCTION

1.1 Introduction of the Thesis

With the recent development of internet, we now have immediate access to different data related to the topic we are interested in. However, even with this increased access there are some important considerations. First is how to determine which part of the data is useful. Another important consideration is how to use the information to construct a good model to make a prediction. We will apply decision tree, random forest and logistic regression to two real problems. The first is regarding the survival from the sinking of the Titanic. The second is predicting crime category of San Francisco. The data set for both problems was taken from Kaggle.com which is a data science competition website. We not only use information from the data provided by website but also collect external data.

1.2 Research objective

The target of the Titanic research was to test which factors play an important role when we predict whether a passenger survived. We also want to use our best model to predict if a person with specific characteristic would have survived from the disaster. The purpose of San Francisco Crime prediction is to find out the most predictive reason leading to a different category of crimes. Also, we will try to visualize the data such as plotting one category of crimes on the map of San Francisco.

1.3 Titanic

The sinking of the luxury steamship Titanic is one of the most famous historical events. As reported, 1,502 people lost their lives among 2,224 people on board in this

disaster owing to the suddenness of this disaster and the limited number of lifeboats. According to the records, when faced with fear of sinking and death, most of the men on the ship sacrificed themselves and actively let the old, women, and children board on the lifeboats.

Whether the passengers on the Titanic could survive or not depended on different elements, for instance, the sacrifice of most men, as mentioned earlier. Thus, we would usually make the prediction that the survival rates of the old, women and children were higher than adult man. From another aspect, we may predict that the people who were close to the location of lifeboats were more likely to get through the disaster.

We will use the information of the passengers which is provided by kaggle.com to predict whether a passenger was survived or not. We can also find out which variable is more predictive during the model optimizing.

1.4 San Francisco crime prediction project

San Francisco, a port city on California Pacific coast, is the cultural, financial, and commercial center of California. Near the well-known high and new tech zone- Silicon Valley, San Francisco has become the most important development and research area for new and high tech, now is considered as the vital financial center of Western American. Although San Francisco has the smallest area in the state, its population density only falls behind New York city and it has the fourth largest population among all the cities in California.

However, San Francisco also has another famous background. Alcatraz Island, one of the most famous federal prisons, was located in the San Francisco Bay from 1933 to 1963. The small island was developed with facilities for a lighthouse, a military fortification and a federal prison.[1] The prison is regarded as inescapable and had locked up some of the most notorious criminals in the world.

But it seems that people pay much less attention to San Francisco's security problem compared to its quick development. In fact, San Francisco's crime rate is higher than 98%

of the communities in California. The crime rate is about 62 incidents per 1000 residents combined with 8 for violent and 54 for property. The crime index is 3 for San Francisco (100 is safest) which means the city is only safer than 3% of the other cities in the US.[2]

Kaggle.com provides data of all crimes that happened in San Francisco in the past 12 years. We are going to use these information to predict the categories of crimes that occurred in San Francisco. Additionally, we made a visualization to see what we can learn about the city from the map of crimes.

CHAPTER 2

METHODS

2.1 Decision Tree

Decision tree is a decision making method, which based on the context of each decision, assigns a probability to each of the possible consequences. [3] It is a basic method of machine learning and also a very popular data analysis method. The name comes from the tree-like graph.

To provide a simple example of how decision tree works, we can consider the following. We have a data set with three independent variables which are sex, high school GPA score, and district. Our target is to predict SAT scores on the independent variables. The decision tree method will pick up the most predictive variable, for example sex, and consider it as the top of the tree. So the data is cut in two parts which are male and female. For each sex, the method will pick another variable, and separate data into smaller groups. For instance, the male students will be divided into three groups if there are three different districts in the data. The same method will also be applied to the last variable, high school GPA. The result will be a tree-like graph.

Two parameters that need to be set for the decision tree method are minimum split and complexity parameter. Minimum split is defined as the minimum number of observations that must exist in a node in order for a split to be attempted. Any split that does not decrease the overall lack of fit by a factor of the complexity parameter is not attempted. Decision tree usually provides a good result for a large amount of data in a short time. But the result also can be biased when variables have many levels.[4] If the data have too many uncertain values, the problem will become complex and the running time will become longer.

2.2 Random Forest

Random Forest is another very popular data mining method. It can be considered as an improvement of decision tree since it will build a number of decision trees simultaneously. The difference is that each time a split is considered, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split is allowed to use only one of those *m* predictors.[5] The size of the sample *m* is equal to square root of total p predictors and will be taken at every split.

Random Forest has many advantages including high prediction accuracy, and is able to deal with large amount of data and works well even if the data has missing values. Comparing with the decision tree method, another improvement is that random forest can estimate the importance of each variable which helps us optimize the model. The biggest problem of random forest is over fitting. Also, the variable with more levels have higher influence to the decision tree when each variables have different levels which may lead to higher weight.

2.3 Logistic Regression

Logistic regression is a widely used regression model in statistics. The logistic regression function is

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X.$$

The log odds will be altered by β_1 when X is increased by 1 unit.

The independent variable can be either continuous or categorical, but the dependent variable is categorical. The dependent variable can have two or more values. The binary logistic regression treats the dependent variable Y as an indicator variable, such as success or failure which is depended on whether or not an event occurred. In the real problem, the outcome is interpreted as "0" or "1" as it leads to the most straightforward interpretation.[6]

Cases with multiple variables are named multinomial logistic regression. The dependent variable has more possible types that are not ordered. Multinomial logistic regression was used to calculate the probability for each category and the sum of probability should equal to 1.

As one of the widely used statistical models, logistic regression not only can predict results under different independent variables, but is also able to find out which variable influences the result the most.

CHAPTER 3

PRELIMINARY REPORT: SURVIVAL PREDICTION OF TITANIC PASSENGERS

3.1 Data Description for Survival prediction of Titanic passengers

There are two data files named Train.csv and Test.csv from Kaggle.com which record the information of the passengers on Titanic. There are 891 passengers in the train data and 418 customers in the test data. Variables include pclass, name, sex, age, sibsp, parch, ticket, fare, cabin and embarked. The only difference that exists between these two data files is that the survival situation of the customers is only showed in the Train data file.

The goal is to use the information from the Train data to predict the survival of the passengers in the test data file. Once the prediction is submitted to the kaggle website, prediction accuracy will be calculated using the true survival information by the website. We compare decision tree , random forest and logistic regression in terms of their prediction performances. The Table 3.1 describes the variables in the Train data set.

Variable	**Description**
survival	Survival (0= No, 1= Yes)
pclass	Passenger Class (1=1st, 2=2nd, 3=3rd)
name	Name of the passengers
sex	Sex (Male, Female) of each passenger
age	Age (0.42 - 80.00) of each passenger
sibsp	Number of Siblings/ Spouses Aboard
parch	Number of Parents/ Children Aboard
ticket	Ticket number for each passenger

fare	Passenger Fare
cabin	Cabin for each passenger
embarked	Port of Embarkation (C=cherbourg; Q= Queenstown; S=Southampton)

Table 3.1 : The description of the variables of Titanic data

The youngest victim was less than 1 year old and the oldest victim was 80. The median of age is 28 which is very close to the mean space(29.70). However, the age information of 177 passengers are missing.

We can use the imputation method to deal with the 177 NA in age variable. However, just using the mean of the age variable instead of all of them is not accurate because the missing proportion is very high that is 177/891 = 0.1987. Since we can use the decision tree to predict if the passenger survived, we can also use the same method on the data with the age values available. Then, this decision tree can be used to predict the age of those passengers with NA in age variable. The only difference is that survival is category variable and age is continuous variable. It is easy for us because the only thing needs to be done is to change the method to Anova in the R code. The comparison between original age variable and new age variable is shown in Figure 3.1. The distribution of the completed age variable is similar to the non-missing values.

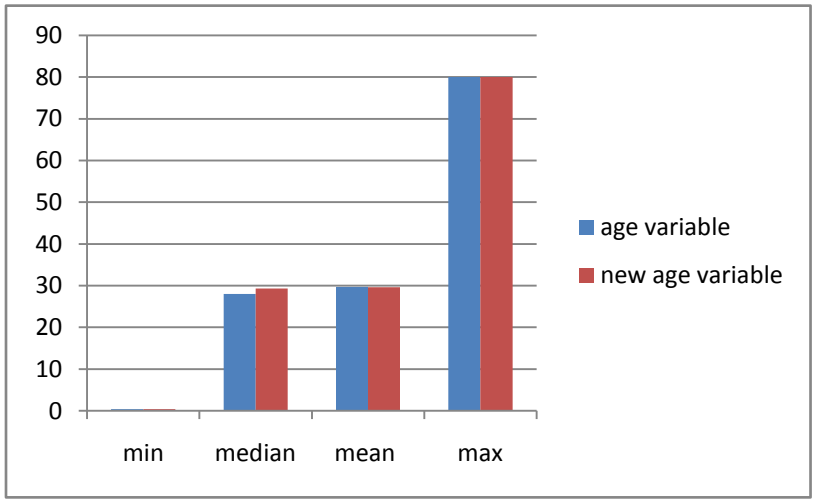

Figure 3.1: The comparison between original age variable and the new age variable after filling the missing values

There was only one missing value in Fare variable and two missing in Embarked variable. The mean value will replace the missing values in both the Fare and Embarked variable.

In order to find more information from the given data, a new variable will be constructed named Title. We obtained useful information from the addition. For example, it is likely to predict that the survival rates of the passenger with title Mrs and Miss were higher than those with title Mr and Master.

In sum, there were 18 different titles in this new variable. Some titles were very common such as "Mr". There are 517 passengers with this title. Some titles were very rare such as "Don" or "Major" which appeared less than 5 times. We do not want to contain all these different titles in our title variable because more levels in a variable means longer running time. So we will combine all the rare titles with similar information together. For example, we will combine "Lady" and "Ms" with "Mrs" because they only appear very few times and have similar meaning with "Mrs".

Table 3.2 describes the number of different titles after combining the unusual titles together.

Master	**Miss**	**Mr**	**Mrs**
60	185	517	129

Table 3.2: Newly created Title variable from train data after combining the unusual title together

We will not include ticket variable and cabin variable in our model. The reason that the ticket variable is not put into use is because the format of ticket number is different. For example, one is named as 17463 while another is named as A/5 21171. The cabin variable has also been deleted because more than half of passengers information is missing.

3.2 Survival prediction of Titanic passengers

3.2.1 Decision Tree

First, the decision tree method is applied to the survival prediction of Titanic passengers. Most of the variables except name, ticket and cabin are included in it. The name variable is transferred to the title variable which will be used in the model. The default setting of decision tree is used in figure 4.1 : minimum split =20 and complexity parameter =0.01. The constructed title variable is considered as the most predictive variable from the fitted tree. The model even stops splitting for the passengers who have a Mr. title.

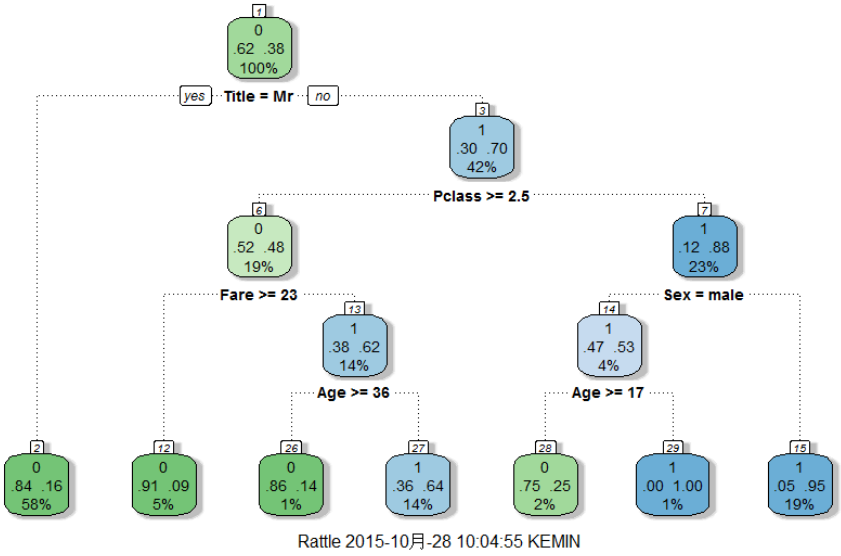

Figure 3.2: Decision tree for Survival prediction of Titanic passengers with default setting

From the Figure 3.2, we can find out the importance of each variable. The variable in the higher level is more predictive than the variable in the lower level. In the bottom of the plot, the decision tree divide all the passengers into 7 groups and predict the probability of survived for each group. For example, 58% of passengers have Mr as title, and 16% of them survived.

By calculating both resubstitution error rate and the cross validation error rate, we can check the accuracy of our decision tree.

In order to obtain the resubstitution error rate, the first step is to use the decision tree to predict how many passengers will survive from the disaster. The confusion matrix will be constructed by combining the predicted result and the actual result together. Table 3.3 is the result with details. The results show 52 passengers were predicted as survived while they actually died in the disaster. Also, 91 passengers were predicted as dead while they actually survived. Thus, the subsequent error rate is equal to (52+91)/891=16.05%.

	Pred. died	**Pred. survived**	**Total**
Actual died	497	52	549
Actual survived	91	251	342
Total	588	303	891

Table 3.3: Confusion matrix for Survival prediction of Titanic passengers

Compared to the resubstitution error rate method, the cross validation method is a resampling approach which reveals a more honest error rate estimate of the decision tree computed on the dataset. A brief introduction of the cross validation method is as follows.

The first step of the cross validation method is to randomly split the data in K folds. Then, we use (K-1) folds to construct a model to compute the error rate on the folded 1/K of the data. After that, we repeat this process by K times and collect K different error rates. At last, we calculate the mean of these error rates in order to get the final error rate for cross-validation.

The total number of train data for the Titanic's problem is 891, and we will set the K equal to 10. The result of 10 different error rates are shown in Figure 3.3. The cross validation error rate is the mean of these ten errors which is 18.31%. This result is also close to the median number of these ten errors.

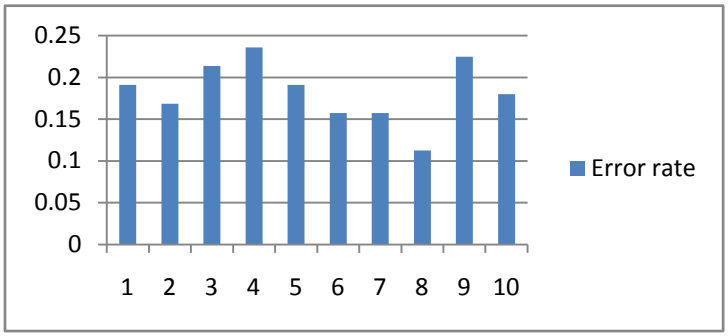

Figure 3.3: The different error rates calculated by cross validation method for checking if the decision tree fits well

The decision tree in Figure 3.2 will be used to predict the survival variable in the test data and the result will be submit to Kaggle.com. As mentioned before, the website has the information whether the passengers in the test data survived. The website reveal a 22.01% error rate for my prediction result. Both resubstitution error rate method and the cross validation error rate method have a lower error rate than the test errors.

In order to decrease the error rate, it is necessary to change the default parameters of the first decision tree and build different trees. Each of the new tree will be uploaded to the website to get the accurate error rate. The result is shown in Table 3.4.

minimum split	**complexity parameter**	**Error rate**
2	0.001	27.75%
5	0.01	22.01%
5	0.02	21.05%
10	0.01	22.01%
10	0.02	21.05%
20	0.01	22.01%
20	0.02	21.05%

Table 3.4: More decision trees by changing the default parameters and the accurate error rate calculated by Kaggle.com

There is no big difference between the results of the error rate for each tree. The result shows that a smaller minimum split number or a smaller complexity parameter does not mean that we can get smaller error rate. Even though we have the smallest error rate 21.05% when we change the complexity parameter equal to 0.02, these new trees do not significantly reduce the error rate compared to the default settings.

3.2.2 Random Forest

The random forest model will be applied next, using the same variables appeared in the decision tree model. We set the parameter ntree to 1000, meaning that we will grow 1000 trees for this model. Figure 3.4 shows important measure from the result of random forest. The picture can show us how well each tree performs without each variable. A variable with a higher decrease in accuracy or a higher score means that it is very predictive and important. Pclass, Sex and Fare are the top three predictive variables from the picture. This is a little different from the decision tree method since our decision tree consider the Title variable as the most predictive variable. However, we still keep the Title variable in our model because it falls in the middle, between the most and least predictive variables.

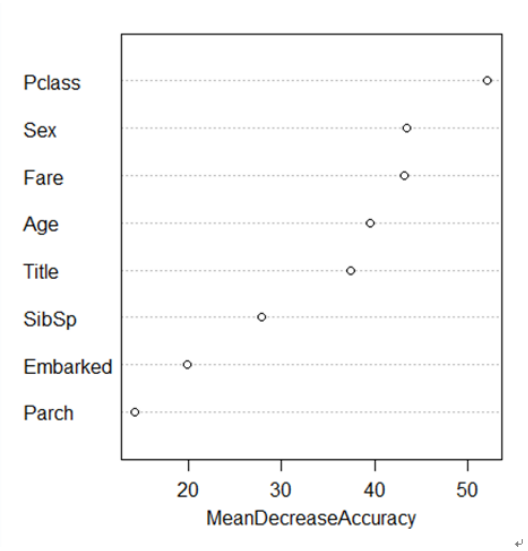

Figure 3.4: Importance measures from the result of random forest for the Titanic problem

The "Out-of-Bag"(OOB) method is used to check whether the random forest method fits well. It is similar to the resubstitution error rate method for the decision tree. Table 3.5 is the confusion matrix from R output. We can see 47 passengers were predicted as survived while they actually died in the disaster. Also, 99 passengers were predicted as dead while they actually are survivors. Thus, the "Out-of-Bag"(OOB) estimate of error rate is equal to (47+99)/891=16.39%.

"Out-of-Bag"(OOB) estimate			
	Pred. died	**Pred. survived**	**Total**
Actual died	502	47	549
Actual survived	99	243	342
Total	601	290	891

Table 3.5: The "Out-of-Bag"(OOB) estimate of error rate for the random forest model

Random forest model will be applied to the test data. The error rate is 20.0% for prediction compared to the true survival information in the website. The error rate decreases about 2% compared to the decision tree method and the ranking of prediction improves from 1800th place to 700th place.

Another random forest model is constructed by deleting the two lowest predictive variables which are Parch and Embarked. Figure 3.5 is the important measure from the result of new random forest. Compared to the Figure 3.4, the Pclass variable is still the most predictive variable. The Age variable becomes the second important variable instead of the Sex variable.

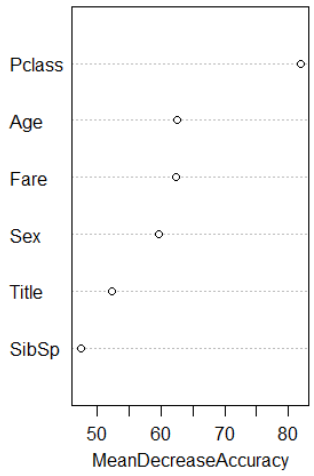

Figure 3.5: Importance measures from the result of new random forest by deleting Parch and Embarked for the Titanic problem

The "Out-of-Bag"(OOB) estimate of error rate for the new random forest model is equal to (50+97)/891=16.5%. This number is very close to the previous random forest model. However, the accuracy of prediction calculated by Kaggle.com has a little improvement which decreases to 19.6%.

From the Figure 3.5, Pclass, Age and Fare variables are considered as the three most predictive variables. The young passengers with higher Pclass level and more expensive fare are more likely survived from the disaster. Also, a female passenger has larger probability to be saved.

3.2.3 Logistic Regression

Our full model of logistic regression still includes all eight variables which were used before. The summary is shown in Table 3.6. Pclass, Age, Sibsp, Parch and fare are continuous variables while Sex, Embarked and Title are considered as class variable. For example, for the continuous variable Pclass, a unit increasing in Pclass means the log odds will reduce by 1.1119. Nevertheless, for the class variable such as Sex, male is regarded as a dummy variable. This means being a man will reduce the odd rate by 16.4642 if female is considered as the base level. The AIC value of this logistic model is 766 and the residual deviance is 741.97.

The P-value for Pclass, Age and Sibsp variables are all less than 0.0001 which means they have strong relationship with whether the passengers survived. The P-value of fare variable is 0.1902 which is not significant in 1% or 5% level. This result is quite different from the two machine learning methods due to they both consider fare variable as one of the most predictive variable.

	Estimate	**P-value**
Intercept	20.8533	0.9662
Pclass	-1.1119	<0.0001
Sexmale	-16.4642	0.9733
Age	-0.0455	<0.0001
Sibsp	-0.4786	<0.0001
Parch	-0.3064	0.0209
fare	0.0033	0.1902
EmbarkedQ	0.0077	0.9843
EmbarkedS	-0.3723	0.1296
TitleMiss	-15.6605	0.9746

TitleMr	-1.8267	<0.0001
TitleMrs	-14.5923	0.9763

Table 3.6: Summary of the logistic model with all variables

The next step is to compare the different variables by analyzing the Deviance Table. Table 3.7 includes the deviance and P-value for each variable. The third column of table is the difference between null deviance and residual deviance. A larger deviance number means more important of the variable because it will significantly reduces the residual deviance. For example, Pclass is a very predictive variable because the Deviance number 102.254 which is large and it reduces the null deviance from 1186.66 to 1084.66.

A variable with smaller P-value means the accuracy of model will decrease without that variable. So the variables with small P-value in our model will be kept. For this model, Pclass, Sex, Age, Sibsp and Title variables are considered as predictive variables because they all have a large deviance number and the P-value of them are all less than 0.00001. The fare variable is still not predictive from this deviance table which is different from the machine learning models.

	Df	**Deviance**	**Resid. Df**	**Resid. Dev**	**Pr(>chi)**
Null			890	1186.66	
Pclass	1	102.254	889	1084.40	<0.00001
Sex	1	257.206	888	827.20	<0.00001
Age	1	24.244	887	802.95	<0.00001
Sibsp	1	17.108	886	785.84	<0.00001
Parch	1	0.528	885	785.32	0.4676
fare	1	1.382	884	783.93	0.2398
Embarked	2	3.203	882	780.73	0.2016
Title	3	38.764	879	741.97	<0.00001

Table 3.7: The Deviance Table for the logistic model

The result of the prediction only shows each passenger's likelihood of survival but not the prediction whether they survived or not. Supposed the passenger survived when the p number is greater than 0.5. Otherwise, the passenger will be marked as died. We apply our logistic model to the test data. The error rate is 23.0% for my prediction compare to the true survival information in the website.

After removing the Parch, fare and Embarked variables, a new logistic regression model is constructed. The error rate calculated by kaggle.com reduces to 22.5% but still is not as good as the two machine learning methods.

All three methods give us a not bad prediction since the error rates are all around 20%. Table 3.8 shows the error rate of all three methods and it is obviously that the Random Forest did best among the three methods.

Method	**Decision tree**	**Random Forest**	**Logistic regression**
Error rate	21. 1%	19.6%	22.5%

Table 3.8: The best error rate for each method

CHAPTER 4

DATA DESCRIPTION

4.1 San Francisco crime prediction project

The data we used are from two different sources. The original data is downloaded from Kaggle.com which is named train.csv and test.csv. These two datasets contain incidents derived from SFPD Crime Incident Reporting system. Time of the data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8... belong to training set. Both of these two files are composed from the following variables: Dates, DayofWeek, PdDistrict, Address, X and Y. The difference is that train.csv file has three more variables named Category, Descript and Resolution. The explanation of each variables is list in the Table 4.1.

Variable	**Description**
Dates	Timestamp of the crime incident
DayofWeek	The day of the week
PdDistrict	Name of the Police Department District
Address	The approximate street address of the crime incident
X	Longitude
Y	Latitude
Category	Category of the crime incident (target variable to predict)
Descript	Detailed description of the crime incident
Resolution	How the crime incident was resolved

Table 4.1: The description of the variables of San Francisco crime data

In order to improve the prediction accuracy, we decide to construct new variables based on the original ones. In addition, integrating the external data with the original data helped mediate this problem.

Obtaining additional variables is not easy in general, since making sure the new variables are closely related to the original variables is not easy. Weather is one of the

external variables that is connected with the original variables. The main reason is that bad weather might increase the probability of crime. Fortunately, weather data for San Francisco between 2003 to 2015 can be found in most websites and it does support much more information of the model.

According to www.wunderground.com/history/, the new variables includes tempAve, tempMax, tempMin, windSpeed and rainFall. The explanations of each new variables are as follows in the Table 4.2 with the four new variables constructed as explained above.

Variable	**Description**
year	The year when crime incident happens
month	The month when crime incident happens
day	The day when crime incident happens
hour	The hour when crime incident happens
tempAve	The average temperature for that day
tempMax	The maximum temperature for that day
tempMin	The minimum temperature for that day
windSpeed	The wind speed for that day
rainFall	The rain capacity for that day
Resolution	How the crime incident was resolved

Table 4.2: The description of the new added variables of San Francisco crime

4.2 Visualization of San Francisco crime prediction project

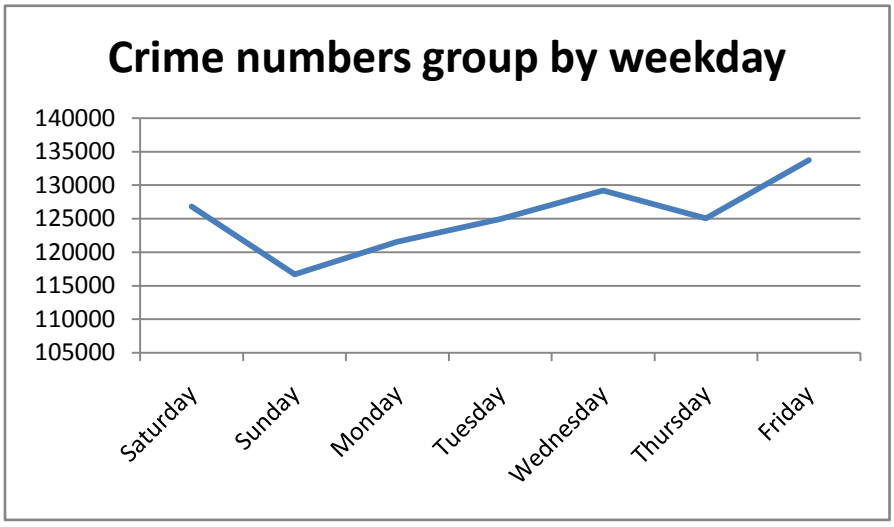

Figure 4.1: Crime numbers group by weekday for San Francisco crime prediction

The Figure 4.1 above shows crime numbers grouped by weekday. There are obvious differences between each weekday. Friday has the most crime and Sunday has the lowest crime. The reason for this I am speculate is that most people come back home on Friday or prefer to hang out with friends. Friday is the day people has many more activities than other days which increase the probability of crimes. Without doubt, Sunday should be the lowest one since most people prefer to stay at home on Sunday. Monday to Thursday are business days and have the average in this picture except Wednesday.

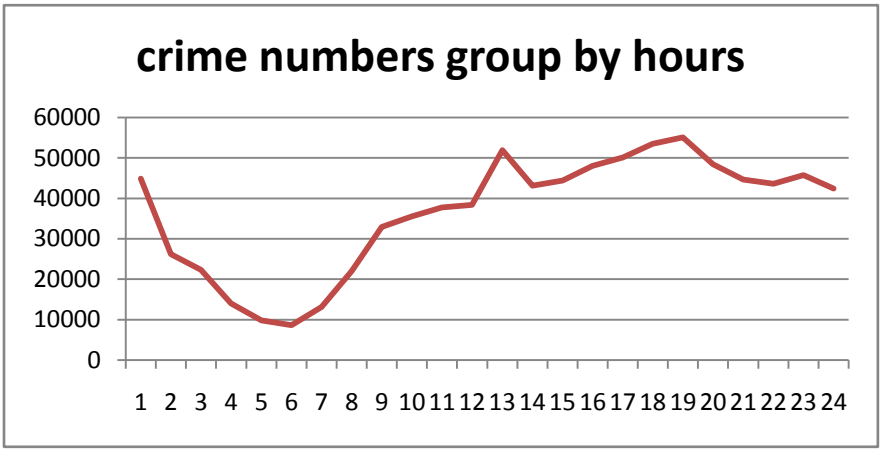

Figure 4.2: Crime numbers group by hours for San Francisco crime prediction

The Figure 4.2 shows crime numbers grouped by time of day. 1am is the first peak from the picture. At that hour, people are not yet straggling out of bars, and parties are for some at a high swing. These people have high blood-alcohol levels which make them likely offenders and targets of crime.[7] We can clearly see that six in the morning has the lowest number of crimes. The number goes down between 1am to 6am because most people stay at home during night but most crimes happen in public. The number goes up between 6am to 1pm since people get up and go out during this time. The second peak is 1pm probably because it is the lunch break and more people will not stay at home or office. The second peak is at 7pm which is also understandable since people get off work at that time. 7pm to 12pm has a very high crime rate since offenders are very active during the night.

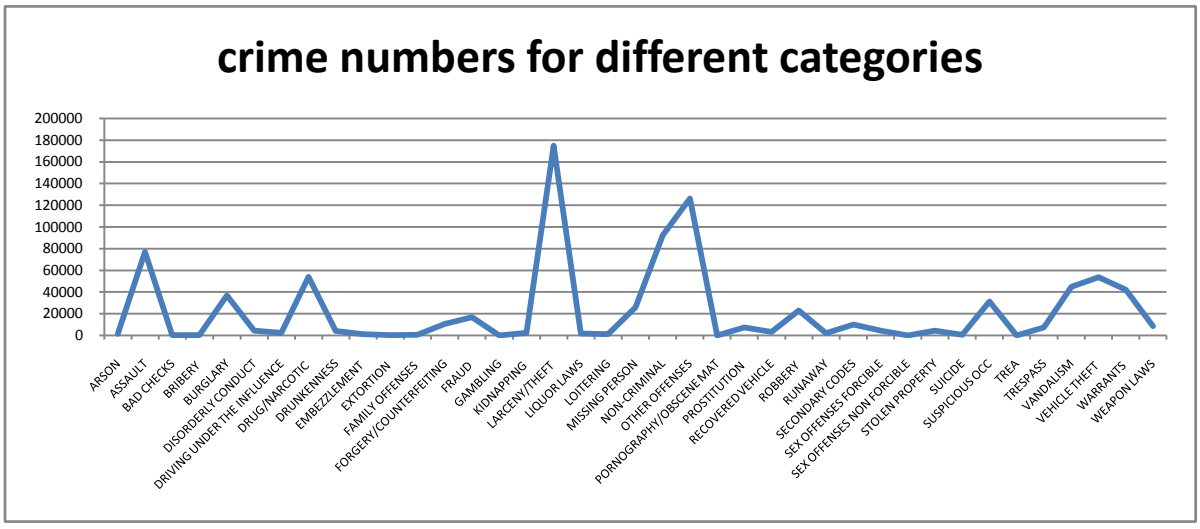

Figure 4.3: Crime numbers for different categories in San Francisco area

The Figure 4.3 shows the crime numbers for different categories. The top five crime ranking categories are Larceny/Theft, other offenses, Assault, Drug/Narcotic and Vehicle theft.

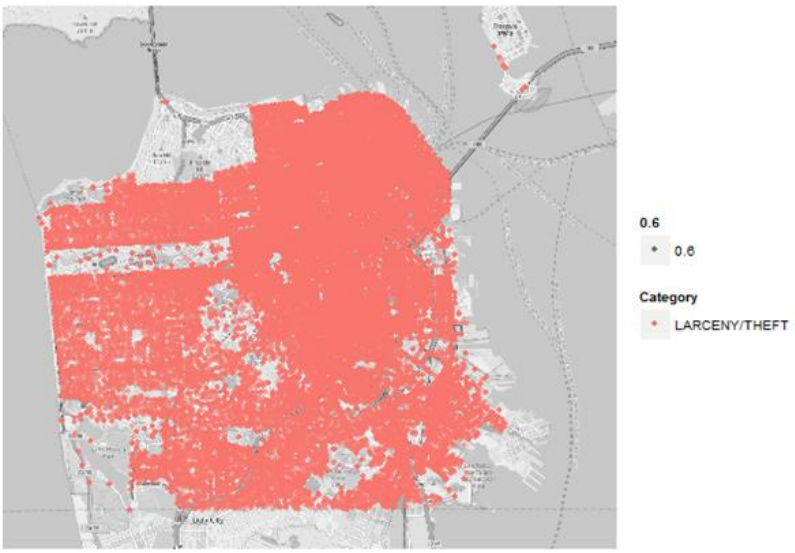

Figure 4.4: The plot for Larceny/Theft in San Francisco area

Figure 4.4 is the plot for Larceny/Theft on the map of San Francisco. However, if the entire data set was used, the picture for Larceny/Theft become unreadable and the other categories meet the same situation. To make the picture clear and readable, I will only use the train data for 2015.

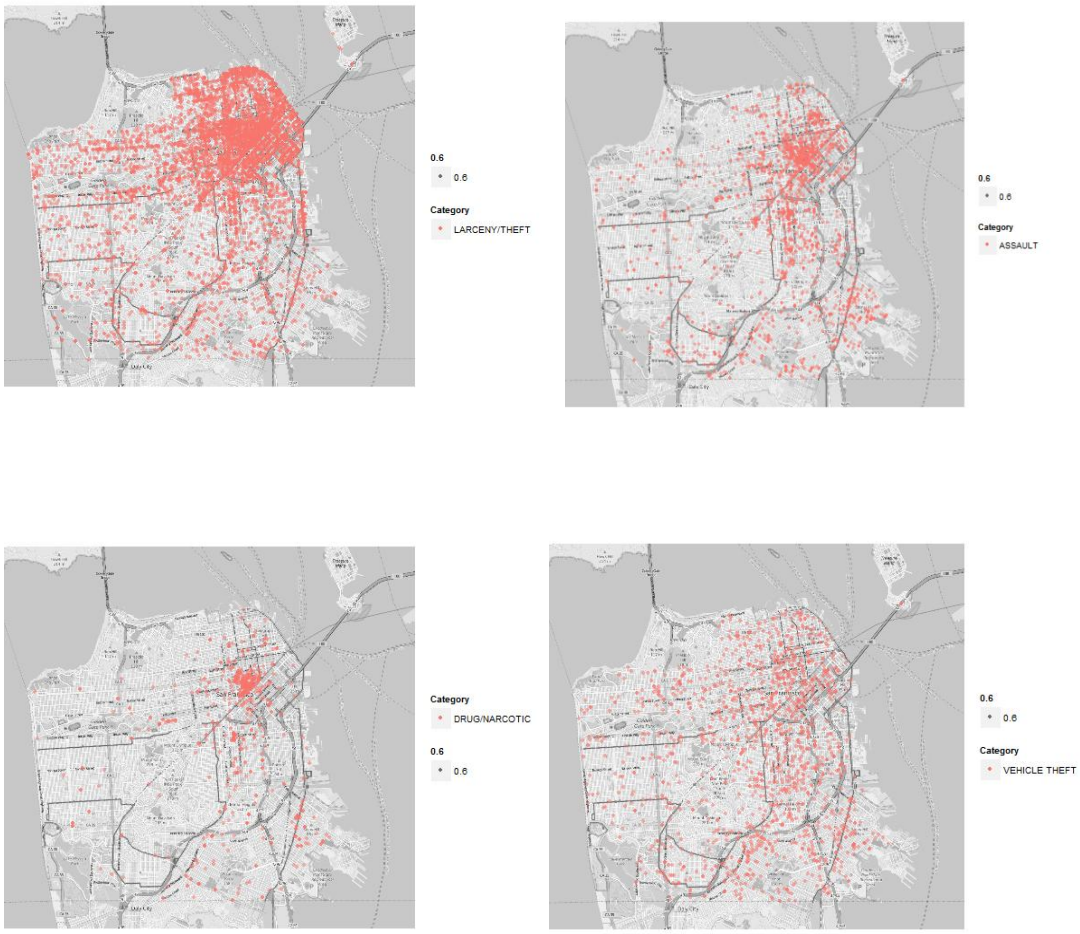

Figure 4.5: The plots for Larceny/Theft, Assault, Drug/Narcotic and Vehicle theft of San Francisco crime in 2015

From Figure 4.5, the vehicle theft look like a random scatter plot. This should be understandable since cars are normal vehicles in the US and exist everywhere. But we can clearly see most of the crimes are concentrate on the right corner of the map for the Larceny/Theft, Assault and Drug/Narcotic. So I check the districts map of San Francisco to see the reason.

Figure 4.6: The main district of San Francisco area

Figure 4.6 shows the main district of San Francisco. Looking at the population of different districts, the left side of the map has a lower population rate. For example, the population of Sunset is about 90,000 and the population of Twin Peaks is only 4,000 comparing to the total population of San Francisco is about 840,000.[8]

The top right corner of the map are the places with very high population. One of the best instances is Chinatown which has about 100,000 people in a very small district. Downtown San Francisco is also located in this corner which means a large amount of the population and business are in this area. This could be the main reason why more offenders tend to stay in this area. Also, the crime numbers of Drug/Narcotic offenders are extremely high in downtown since many vagrants live in this area. Chinatown has a large population with high crime rate which could be another reason why numerous red dots concentrate in the right corner.

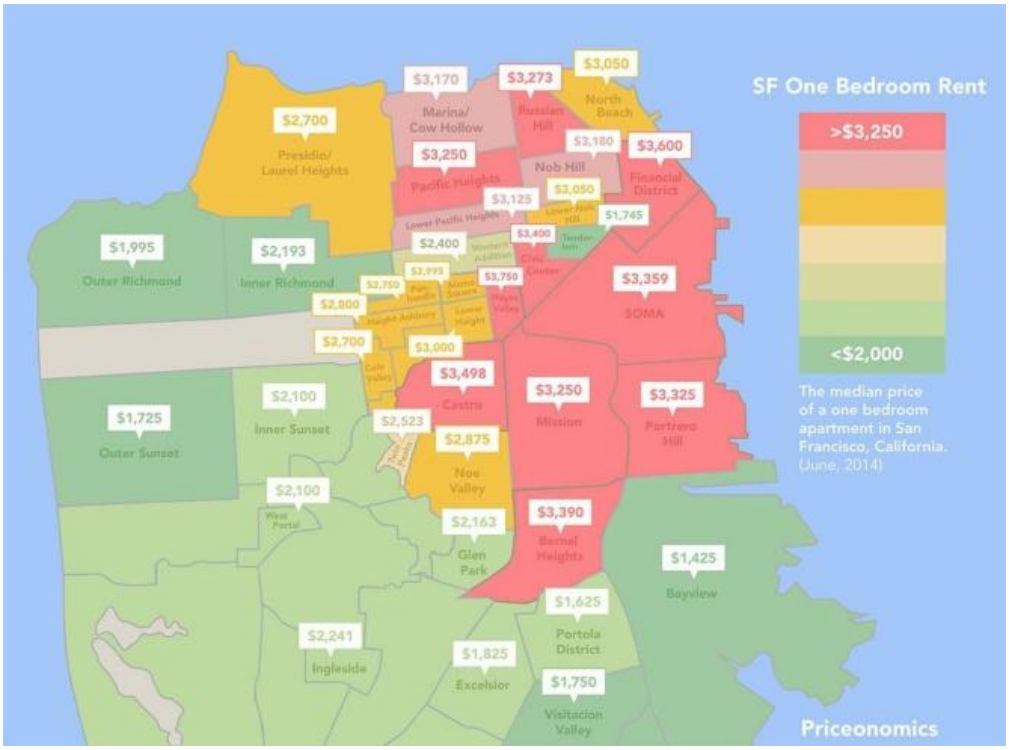

Figure 4.7: One bedroom rent map for different districts in San Francisco

Figure 4.7 is one bedroom rent map for different districts. House rent can be used as a symbol of economy. Higher house rent usually means more rich people or better development in the district with relatively better security. The red area in the picture should be the richest area which corresponds to SOMA/Potrero Hills area. We can clearly see the number of Larceny/Theft, Assault and Drug/Narcotic are much lower than Downtown and Chinatown which has much lower house rent. The reason why the most lowest house rent area also has a very low number of crimes could be fewer population exist there than the rich area.

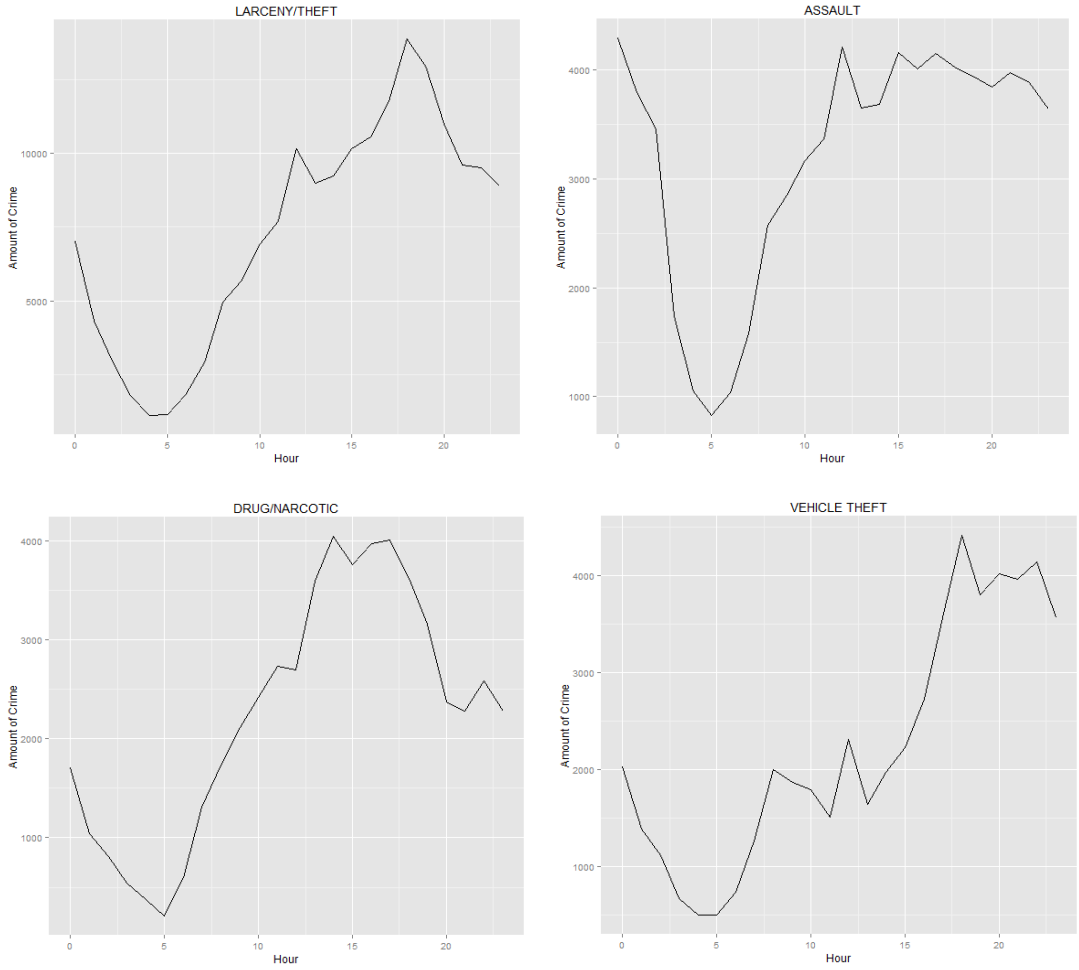

Figure 4.8: The plots of the four different categories in San Francisco (Larceny/Theft, Assault, Drug/Narcotic and Vehicle theft)

From figure 4.8, the peaks of Assault, Drug/Narcotic and Vehicle theft are all about 4000 which are almost the same as the lowest crime number for Larceny/Theft. So I will just compare the crime number for a specific category in different time of one day.

The tender of the four pictures is similar because my assumption is that the rate of crime is decided by people are at home or not. Between 0-8 in the morning, it is clear that the lowest crime rate happened at five o'clock for all four different crime categories. However, the four pictures looks quite different and shows useful information after 8 o'clock. As a result, my assumption is wrong in some aspects.

The crime rate of Larceny/Theft keeps rising up from 5 to 18 and reaches the peak at 18, which indicates the thieves tend to commit a crime during the daytime. Evening is also a risky time comparing to the very early in the morning.

The mean number of Assault is much less than Larceny but it keeps a very high rate. Most of people will stay at home after 20 which will lead to the decreasing of crime rate; nevertheless, in fact, it does not decrease between 20 to 24. The worse public security at night can explain the increasing crime rate. Also, most people stay at home not only means the targets for criminal become fewer but also means fewer people will notice the crime and help the victim. So some criminal would tend to assault at night.

The picture of Drug/Narcotic looks similar to the Larceny/Theft one. The only difference is the peak of Drug/Narcotic moves to left a little. This is interesting information because most people would take drugs in the afternoon is not in my imagination. What I thought before was that taking drugs would increase with the increasing number of parties at night. But actually, the number of taking drugs at night are quite fewer than day time. There are some materials about the reasons why do people take drugs. The main reason is to escape from work pressure or relax but not to have fun.[9] This information is helpful to explain the picture. Most people will confront agitating time in the afternoon because the pressure of work. I guess this might be the reason why the number of taking drugs reaches the peak in the afternoon.

The picture of the vehicle theft can be understandable. The crime number stay in a low level because most people drive their cars to work and the cars are not in the parking lot at home. The vehicle thieves would not commit a crime in public because it is easily to be found by others. The crime number raise up very quickly and reach the peak after the nightfall. This is because people drive back and park the car at home which means fewer pedestrians will pass that parking lot. I think it will indirectly increase the number of thieves. Another reason why car thieves appear at night is that there are less people during this time period and they are easier to escape.

CHAPTER 5

ANALYSIS

5.1 Summary of Variables for San Francisco Crime

The variables for San Francisco crime prediction problem are much more complicated than those in Survival data of Titanic passengers. We do not want to include all variables described in section 4 in our model because some variables are redundant. It was necessary to do variable selection before running different models.

In addition to the original variables in the data, we add information on daily weather in San Francisco. Specificially, we consider the daily temperature, which is given as the minimum and maximum temperature of the day. The other two variables related to weather are WindSpeed and RainFall. Both of these two variables will be included in our basic model in order to see whether they are predictive. The RainFall variable will be transformed to a categorical variable: all records with 0 in the RainFall variable will be regarded as sunny, otherwise, rainy.

The DayofWeek and PdDistrict variables correspond to 7 weekday and 10 police stations, respectively. Both of them are categorical variables, and will be included in the model. The Hour variable is also very predictive because the crime numbers show big differences for different hours as mentioned before.

The Address variable shows the approximate street address of the crime incident. There is no way to convert it to a category variable since it has too many levels. However, we are able to use locations, in terms of X and Y coordinates instead. The combination of X and Y variables shows the exact locations of crimes which is more accurate than the Address variable.

The year variable is one of the most predictive variables because the category of

crimes show big difference by different years. The target variable, crime category, has 40 types of crimes. For each incident, we need to calculate predicted probabilities for every class and the sum of the probability equal to 1. Submissions are evaluated using the multi-class logarithmic loss. The formula is,

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

where N is the number of incident in the test set, M is the number of class labels, y_{ij} is 1 if observation is in class and 0 otherwise, and p_{ij} is the predicted probability that observation belongs to class, ith observation belongs to jth class.

The Kaggle website will provide us a score for each submission. Also, the submission by each participants will be recorded in a public leaderboard. A lower score indicates the result is more accurate, and the scale of this leaderboard is now between 2.05 to 34.53. However, half of the 1371 submissions are located between 2 to 3 which means a good prediction should stay in this scale.

We want to test how each variable in the train data influences the types of crimes. The train data is divided into several groups by each variables in order to simple our plot. Since the incidents number in each group is different, the crime number for each category will be divided by the total number in that group in order to make Figures 5.1-5.8 more accurate. The crime number for the top 6 categories in each group will be calculated.

For example, we divide the train data into 4 groups by maximum temperature. The x-axis in Figure 5.1 shows the range of 4 groups. The number of incidents for each group is different, they are 202502, 237814, 235042 and 202691 which corresponds to 46-59℉, 60-64℉, 65-70℉and 71-99℉. The y-axis of Figure 5.1 is the crime rate for one category in each group. For example, the crime number of Larceny/Theft is 38853 when the maximum temperature is between 46-59℉. So the crime rate of Larceny/Theft is 38853/202502≈19% in this range. The top six crime categories are plotted in Figure 5.1.

Figures 5.1-5.8 show the plots for eight variables which will be used in the model.

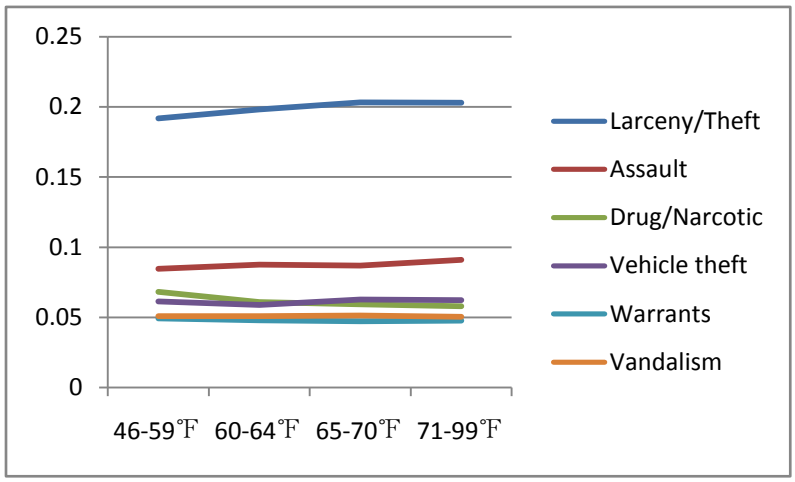

Figure 5.1: the top 6 crime categories grouped by the maximum temperature divided by the incidents number in the group

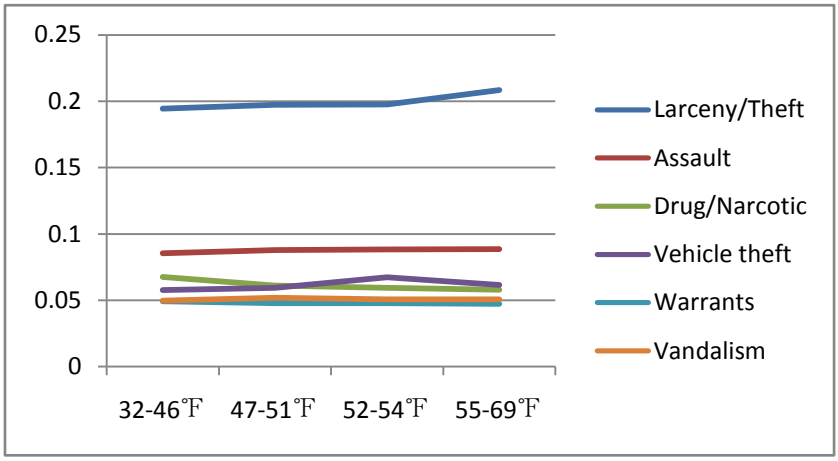

Figure 5.2: the top 6 crime categories grouped by the minimum temperature divided by the incidents number in the group

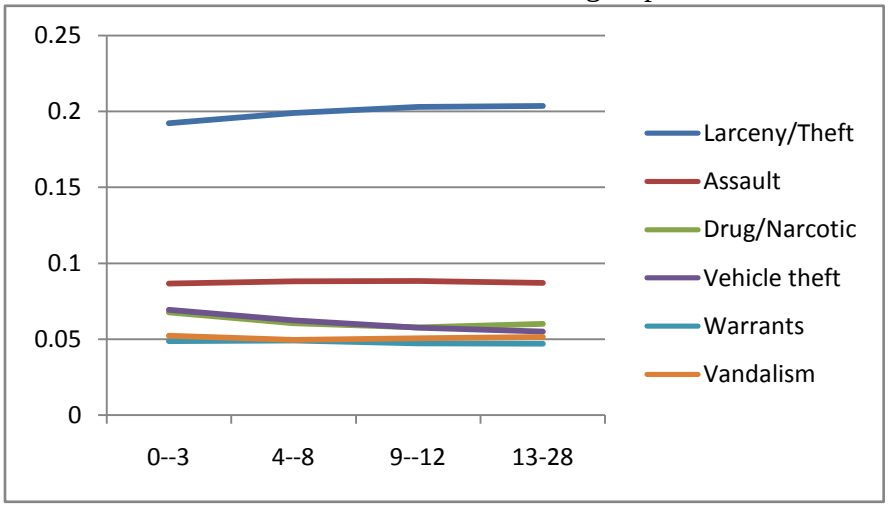

Figure 5.3: the top 6 crime categories grouped by the windspeed (mph) divided by the incidents number in the group

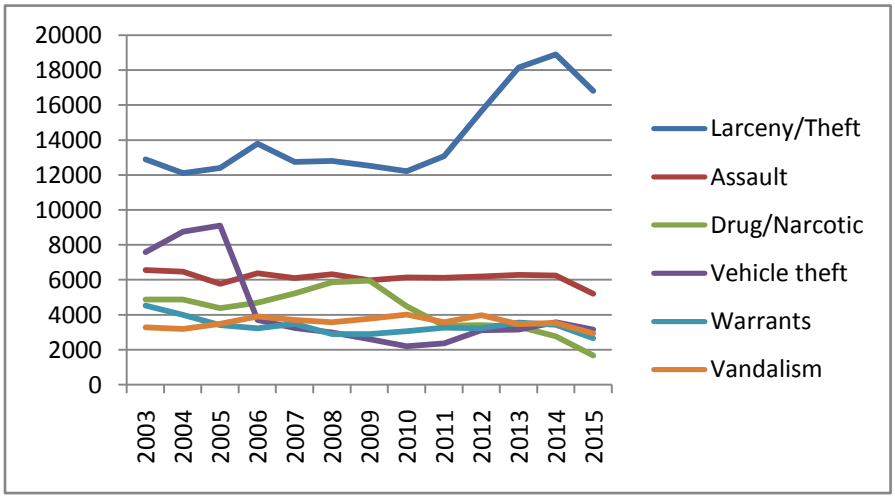

Figure 5.4: the top 6 crime categories grouped by the year

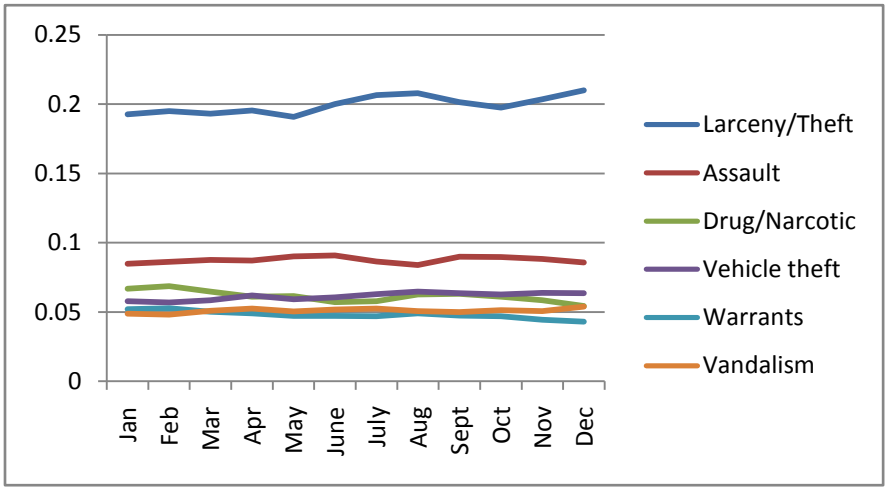

Figure 5.5: the top 6 crime categories grouped by the month divided by the incidents number in the group

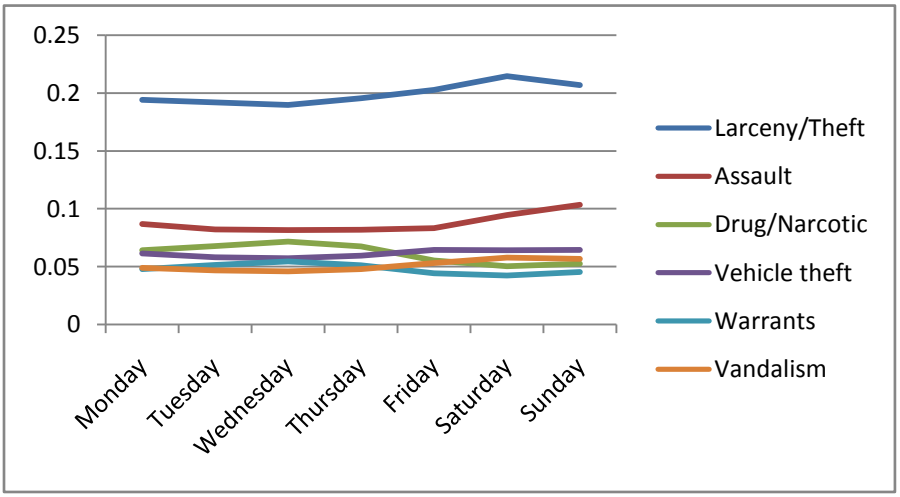

Figure 5.6: the top 6 crime categories grouped by the DayofWeek divided by the incidents number in the group

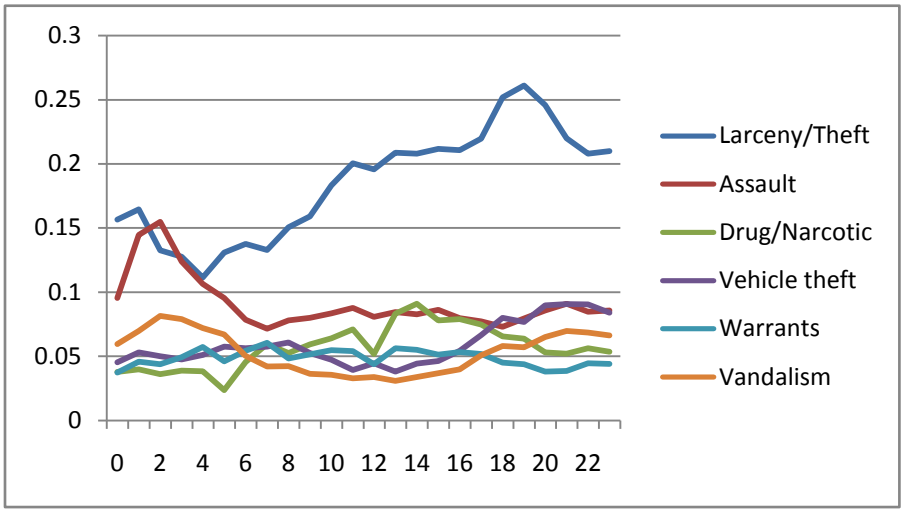

Figure 5.7: the top 6 crime categories grouped by the Hour divided by the incidents number in the group

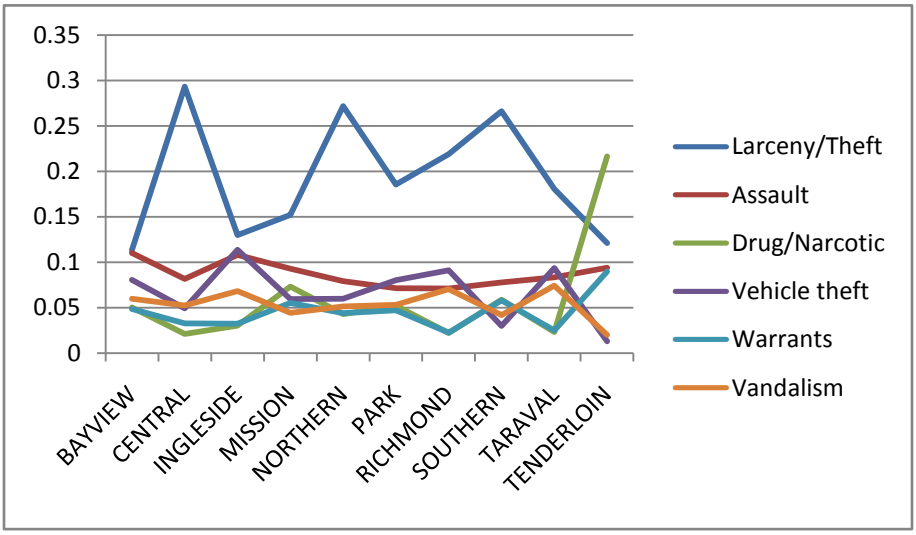

Figure 5.8: the top 6 crime categories grouped by the PdDistrict divided by the incidents number in the group

5.2 Decision Tree

The decision tree was constructed using all variables mentioned in part 5.1 with minimum split equal to 200 and complexity parameter 0. The decision tree assigns a number for each variable, and the larger number means the variable is more predictive.

Due to the large size of the decision tree method, we just show the top part of the decision tree in Figure 5.9. The figure shows one crime will be considered as

Larceny/Theft if the crime happens in one of the Central, Northern, Park, Richmond, Southern and Taraval PdDistrict, otherwise, the crime will be regarded as Other Offenses.

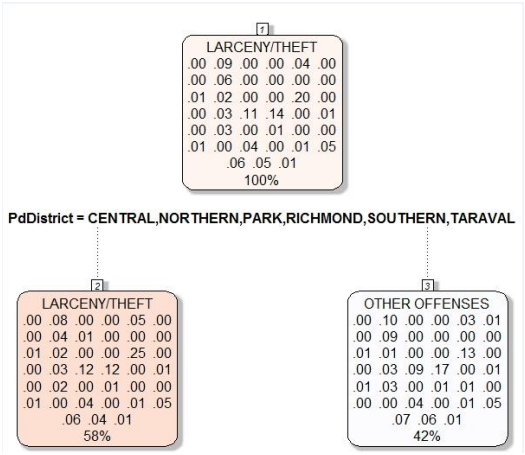

Figure 5.9: The top part of the decision tree using Y, X, PdDistrict variables

Figure 5.10 shows the important measurement from the result of decision tree. Variable importance is computed based on the corresponding reduction of predictive accuracy when the predictor of interest is removed.

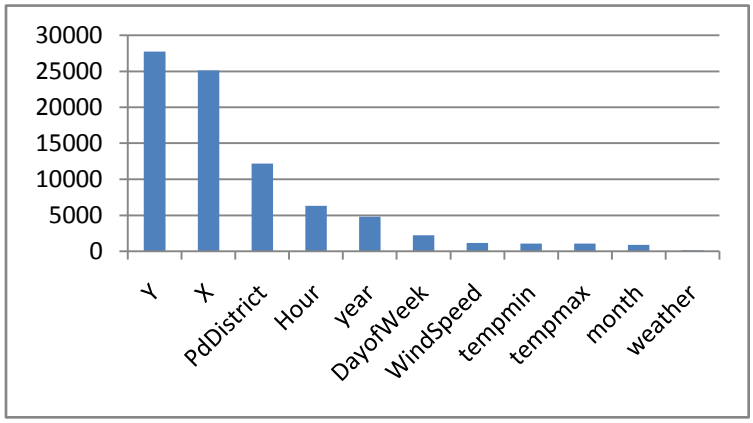

Figure 5.10: The important measurement from the result of decision tree

The variable with the higher score means it stays in a higher position of the decision tree. For example, the tree may assign 20 different categories to the first police station if the PdDistrict is considered as the first level of the tree. Among these 20 different categories, the tree will divide them into more groups in the second level by using the Hour variable.

For example, there is one crime report by the Richmond police station that happened at 1pm. According to figure 5.8, the probability of Warrants is higher than the probability of Vandalism for the Richmond police station. However, according to figure 5.7, Vandalism has a higher probability than Warrants at 1pm. We may prefer to regard this crime as Warrants since the decision tree suggest PdDistrict is a more predictive than Hour variable.

In order to obtain the resubstitution error rate, the decision tree was used to predict our train data. The confusion matrix will be constructed by combining the predicted result and the actual result together. It is a 40 by 40 matrix because the number of categories is 40. The numbers on the diagonal are the correct prediction. Thus, the subsequent error rate is equal to 1-(the sum of the diagonal)/(the number of incidents in the train data)=1-(269146/878049)= 69.35%.

The decision tree method will keep splitting until it reaches the desired complexity parameter number. In this situation, the method starts with CP=0.02 and ends up with CP=0. The x-error in Table 5.1 is the cross-validation error. Each row represents a different height of the tree. In general, more levels in the tree mean that it has lower classification error on the train data. However, the model may run the risk of overfitting. The Rel.error keeps decreasing while the nsplit number increases. The decision tree method split a total of 302 times and parts of the summary are shown in Table 5.1.

	CP	nsplit	Rel.error	xerror
1	0.02	0	1.00000	1.00000
2	0.006	1	0.97891	0.97994
...		...	...	...
298	<0.00001	4875	0.86598	0.7264
...		...	...	..
301	<0.00001	4908	0.86597	0.7264
302	0	4922	0.86597	0.7264

Table 5.1: Decision tree method with different nsplit number

After upload our decision tree model, the website reveals a 3.39 score which is in the

middle of the leaderboard. To make sure the least accurate variables do not influence the decision tree model, we eliminate the month and weather variable, and rerun the model. Both the resubstitution error rate and the cross validation error rate show insignificant change, and the score of the model only improves to 3.38.

5.3 Random Forest

The random forest model is also constructed using all variables mentioned in part 5.1 with tree number equal to 1000. The "Out-of-Bag"(OOB) estimate of error rate is equal to 75.6%. While this is a high percentage, it is reasonable because the there are 40 categories of crimes. Our submitted prediction on the test data received the evaluation score 3.5, which falls in the middle of the leaderboard.

As mentioned before, an important parameter of random forest method is the number of trees built. Usually, 500 or 1000 trees is a normal choice for most models. However, we could not build that many trees for this problem due to computational burden. Figure 5.11 shows the full random forest model baed on 100 trees. This number is not enough for a good random forest method especially when the model includes many variables. This causes the fitted model to have a lower accuracy, and the estimated variable importance may not be correct.

Figure 5.11 shows the variable importance measures from the result of random forest. The hour variable is found to be the most predictive variable. The minimum temperature, which is part of the external data, is ranked in second place. But at the same time, the maximum temperature is not very helpful to the model as shown in Figure 5.11. We also need to construct different random forest models by deleting the least important variable.

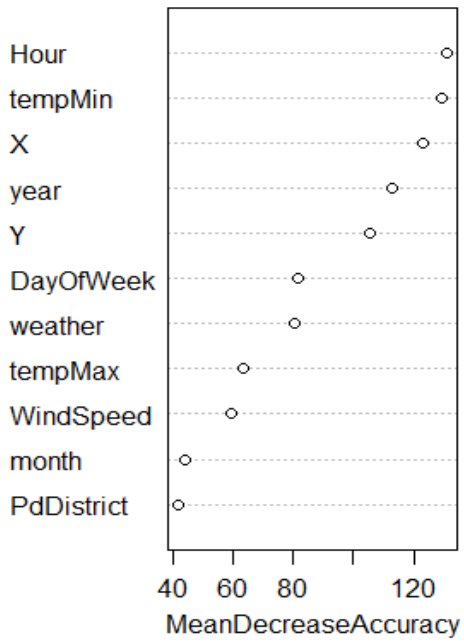

Figure 5.11: Importance measures from the result of random forest with external information for the San Francisco crime prediction project

Therefore we tried separate modeling for each year, within which we can build 1000 trees for random forest. Figure 5.12 shows the variable importance plot for years 2005. With the Year variable removed, the coordinate variables X and Y become more important. Hour and DayofWeek variables still remain in an important position. However, the minimum temperature variable moves to a very low position and becomes even less predictive than the maximum temperature variable which is opposite to our full random forest model. The PdDistrict variable, which was the least important variable in the full model now becomes important in the 2005 model.

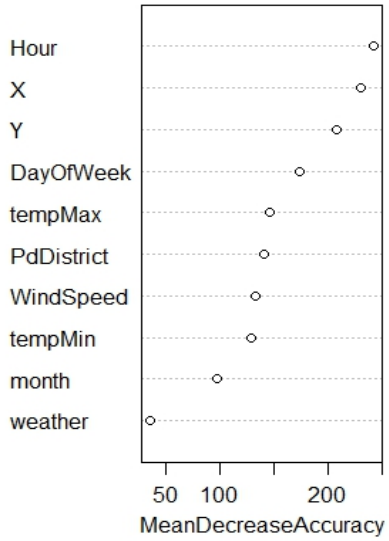

Figure 5.12: Importance measures from the result of random forest for 2005

For example, one of the test data with id number 731131, are considered to have 29% probability of being in the Assault category and 23% probability of being in the Vandalism category by our random forest model, which means the Assault category is considered as the real category for this incident. The basic information for this incident is shown in table 5.2.

tempMax	tempMin	Windspeed	DayofWeek	PdDistrict	Year	Hour	Month	weather
54	48	4	Sunday	Bayview	2005	2am	Jan	Rainy

Table 5.2: The basic information for the test data with Id number 731131

According to Figure 5.12, Hour, DayofWeek, and PdDistrict are three of the most predictive variables. The distribution of Assault category by hour reaches its peak at 2am from Figure 5.7. As shown in Figure 5.8, a crime report by Bayview police station has higher probability of being in the Assault category than the other police stations. Also, Sunday is the weekday with the highest probability of being in the Assault category among the seven days. From the information above, the random forest method predicts the Assault category may be the correct category for this incident.

To make sure which variable can be removed from our model, 13 random forest models are constructed for the all 13 years. The least predictive variable in each random forest model will be recorded, and the one with most occurrence is the worst variable in our basic model. Among all the 13 different plots, the weather variable appears 7 times at the lowest position and is regarded as our worst variable.

The test data are also separated into 13 small data sets by different years. The random forest for each year is used to predict the corresponding test data in that year. The "Out-of-Bag"(OOB) estimate of error rate for each prediction decreases to about 72%. The website also reveals a 3.2 score which shows much improvement compared to our previous result.

As mentioned before, the weather variable is a categorical variable which was transformed by the rainFall variable. Instead of using weather variable, we reran our model with the rainFall variable to ensure our transformation did not influence it's accuracy. However, the rainFall variable is still in the lowest position. Thus, we consider the rainFall variable as our least predictive variable and similarly, the month variable is found to be the second least predictive variable. After deleting the two worst variables rainFall and month, the new random forest model does not show much difference. The "Out-of-Bag"(OOB) estimate of error rate only decreases to 71.6% and the score becomes 3.18.

5.4 Logistic Regression

We also fit multinomial logistic regression, discussed in Section 2.3. The model was used to calculate the probability for each category and the sum of probabilities is 1. In this case, the model assign a probability to all 40 categories for each incident in test data.

For binary logistic regression, we treat the dependent variable Y as an indicator variable, such as success or failure which is depended on whether or not an event occurred. But we cannot simply run the logistic regression since our dependent variable

includes 40 categories. What we did is to construct a loop for these 40 different crimes, and the model will run 40 times. For each time, we set one category in train data as *i*. The dependent variable Y is considered as success if that category *i* occurred. Otherwise, if the other 39 categories happened, the dependent variable Y is regarded as failure.

In comparison to the other two machine learning methods, the multinomial logistic regression provides us more information since we are able to know if one variable is significant to any crime category. For example, we can run a model for the Assault category in order to test which variable is significant for it. Table 5.3 is the result of the chi-square test for the Assault category.

	Deviance	Pr(>chi)
tempmax	0.12	<0.00001
tempmin	0.10	0.0011746
windspeed	0.19	0.3196696
factor(PdDistrict)	1149.25	<0.00001
Factor(DayOfWeek)	5.87	<0.00001
Year	43.39	0.0002958
Month	0.63	0.7845765
Hour	3.72	<0.00001
X	12.51	<0.00001
Y	289.62	<0.00001
weather	4.46	0.0002331

Table 5.3: The summary of the chi-square test for the Assault category

For this case, the dependent variable Y is considered as success only if Assault occurred. A large deviance number and smaller P-value means the variable is more predictive. The P value of the weather variable is very small which means the model considers the weather variable as one of the most predictive variables. This is quite different from the other two machine learning methods since their models conclude that the weather variable is not important. The model indicates Windspeed and Month as the two lowest predictive variables because they both not significant in the 5% level.

Category	tempmax	tempmin	windspeed	PdDistrict	DayofWeek	year	month	Hour	X	Y	weather
Larceny/ Theft	○	○	○	○	○	○	○	○	∆	○	○
Assault	○	○	∆	○	○	○	∆	○	○	○	○
Drug/ Narcotic	○	○	○	○	○	○	○	○	○	○	∆
Vehicle theft	∆	○	○	○	○	○	○	○	○	○	∆
Warrants	○	∆	○	○	○	○	○	○	○	○	○
Vandalism	∆	∆	∆	○	○	○	○	○	∆	∆	○

Table 5.4: Variable importance for the top 6 categories of crime

In order to see if all our 11 variables have same influence to the 40 different categories, we include the results for the top 6 categories of crime in Table 5.4. ○ means the variable is significant in the 5% level while ∆ means the variable is not significant in the 5% level. The PdDistrict, DayofWeek, year, month and Hour variables are significant for all 6 categories, while the other six variables show different results for these 6 categories. Even if both X and Y variables are considered as two of the most predictive variables, they still not significant for every category.

For each category, the model only provides us the probability of whether this category will happen or not. Thus, the probability of being any other of the 39 categories is the same. For example, if the model shown in Table 5.3 reveal the probability of being the Assault category is 61% for the first incident in the test data, the probability of being one of the rest 39 categories is 1% since the probability of not being the Assault category is 39%. After combining all 40 models for the 40 different categories together and dividing by 40, we are able to get our final multinomial logistic regression model.

The result is very good after we submit this logistic regression model to the Kaggle.com. The website reveal a 2.59 score which is much better than the scores for the other two machine learning methods.

We reran our model after deleting the two lowest predictive variables, tempmin and month. The p-value of tempmax variable become much larger while the other variables do not show a big change. This change should have no big influence on the model because tempmax variable is not a important variable for us even though we keep it in our first logistic regression. However, the prediction score calculated by kaggle.com reduces to 2.67 after the submission. Thus, we keep our first logistic regression model as our best model.

CHAPTER 6

CONCLUSION

Figure 6.1 is the rank of the Prediction of Crime Categories in San Francisco Area. Since most of the predictive submissions have a score under 3, a little difference in the score results big difference in the rank. The score of my best model is 2.59779 which is not far from the first place.

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	–	mehran *	2.05079	97	Mon, 11 Jan 2016 15:42:13
2	–	Jghjgfgh	2.06702	12	Sat, 05 Dec 2015 01:44:02
3	–	papadopc	2.11607	111	Tue, 08 Dec 2015 04:54:53 (-32.4d)
742	**↓44**	**KeminXu**	**2.59779**	**18**	**Tue, 01 Mar 2016 16:36:07 (-32.6d)**

Figure 6.1: The rank of the Prediction of Crime Categories in San Francisco Area

Decision tree, random forest and logistic regression are all very useful analysis techniques for the big data. The first two methods belong to machine learning method which require large amount of calculation. The third method is one of the most popular statistical method which can check whether each variable is significant. All three methods are applied to the preliminary test and prediction of crime categories in San Francisco area. The random forest method provides the most predictive model for the preliminary test while logistic regression method is considered as the best model for the San Francisco crime prediction..

For the small data set, we test every variable in order to find out the ones with large P-value. These variables will be deleted to improve the accuracy of the model. However, for the problem with large data set, keeping the variables with large P-value in the model does not decrease the accuracy of the model. Thus, P-value is not always a good

parameter to test the significance of variable. Also, for the multinomial logistic regression case, a predictive variable does not mean it is significant for every category.

To improve the accuracy of the model, it is a good idea to integrate the external data which are closely related to the original variables. Even though there exists overlap between the external data and original data, it still doesn't influence the model as mentioned in previous paragraph.

When people read this paper without strong statistics background , they may prefer to see more figures that can directly see the results than explaining by words. So I try to convert some information on the map which is also the most difficult part for me. For example, I plot all the Larceny/Theft, Assault, Drug/Narcotic and Vehicle theft crimes in 2015 on the map of San Francisco in Figure 4.5.

For the machine learning method, we may not directly use it when facing the problem with large data set. Without a strong computer, it is hard to run the machine learning method such as random forest. The statistical method still perform well with large data set even we cannot conclude it is the best prediction method.

REFERENCES

[1] Odier, Odier (1982). The Rock: A History of Alcatraz: The Fort/The Prison. L'Image Odier. ISBN 0-9611632-0-8.

[2] http://www.neighborhoodscout.com/ca/san-francisco/crime/#data

[3] David M. Magerman, Bolt Geranek and Newman (1995), Statistical decison-tree models for parsing. ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics Pages 276-283.

[4] Deng, H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).

[5] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013), An Introduction to Statistical Learning. Pages 324-325.

[6] Hosmer, David W., Lemeshow, Stanley (2000). Applied Logistic Regression (2nd ed.). Wiley. ISBN 0-471-35632-8.

[7] Marcus Felson, Erika Poulsen (2003). Simple indicators of crime by time of day. International Journal of Forecasting 19 page 595-601.

[8] http://www.city-data.com/neighborhood/

[9] http://www.drugfreeworld.org/drugfacts/drugs/why-do-people-take-drugs.html