Indiana University Bloomington

Machine Learning

# San Francisco Crime Classification

*Author:*
Sujith Shivaprakash
Anup Prasad

*Guided by:*
Prof. Christopher Raphael

April 24, 2016

**Abstract**

In this paper, we are going to discuss our work and findings on "San Francisco Crime data" .The dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods. Given time and location, we must predict the category of crime that occurred. The predictor variables present in the data are of both discrete as well as continuous in nature. The first step towards building a classification model was to plot graphs of the attributes and discover hidden patterns . Next, we used Naive Bayes , multinational logistic regression and feed forward neural network to predict the occurrence of crime at a given location. We evaluated our model using 3-fold cross validation. Original set of predictor variables did not give a good accuracy so, we performed feature engineering, by noticing the different patterns in graph visualizations we added new features. We then performed K-Means clustering to eliminate insignificant labels.With the new predictors, we saw a 5% increase in the accuracy of our model . At the end, we have presented a comparative study of the performance of the three classifiers viz. Naive Bayes , multinational logistic regression and feed forward neural network.

# Contents

content...

# 1 Introduction

From 1934 to 1963, San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today ,the city is known more for its tech scene than its criminal past.But ,with rising wealth inequality ,housing shortages , and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. In this project we have explored and deduced a model that classifies the crime at a particular place given the date, time and other predictors.

# 2 Data Set

Crime incident data of San Francisco is available on kaggle.com. The data contains more than 0.8 million incidents of 39 different types of crime occurred in 10 different counties, at 23000 different locations during the year 2014 and 2015. There are 9 different attributes in the data, they are given in the following table along with their category.

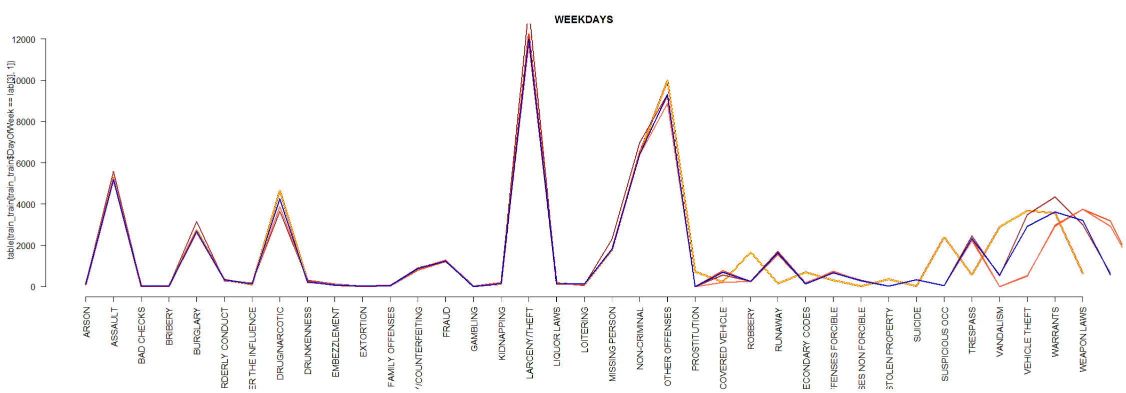
| Attribute name | Type | Description |
|---|---|---|
| Date | Categorical | Date and time at which crime happened |
| Category | Label | Type of Crime |
| Descript | Categorical | Short description of what actually happened |
| DayOfWeek | Categorical | On what day of the week crime took place |
| PdDistrict | Categorical | In which district crime took place |
| Resolution | Categorical | Short description of what action was taken |
| Address | Categorical | Address at which crime took place |
| X | Continuous | Latitude |
| Y | Continuous | Longitude |

**Table 1: Predictor Variables**

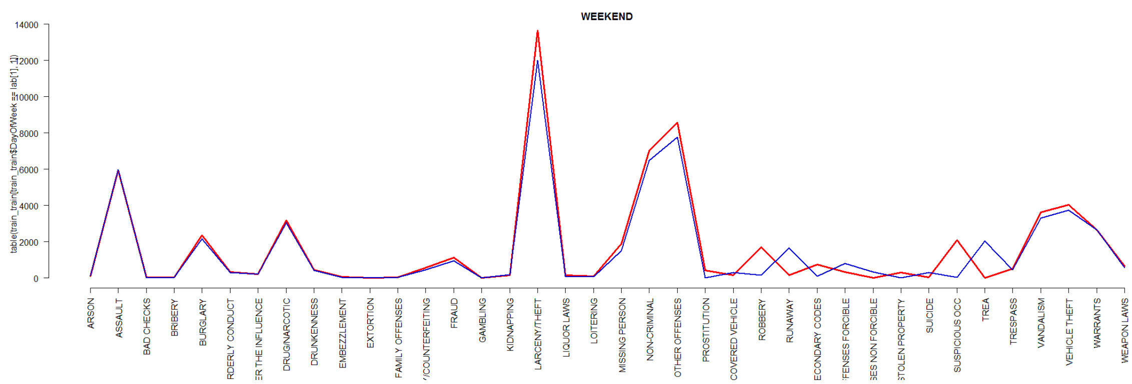
# 3 Feature Engineering

The very first step we took was to discover relationship between day of week and different categories of crime occurring on those days. We divided the days to Weekends and Weekdays, where Weekdays consisted of Monday,Tuesday, Wednesday, Thursday and Friday, while the weekends consisted of Saturday and Sunday. On plotting them on graphs as shown below for different categories of crimes, we found out that there is not significant variation in the peaks of crimes occurring during both these times, in fact the peaks depicted turned out to have similar frequencies. We can see that crime which has high frequency of occurrence on week days are also having high frequency of occurrence during weekend. So creating new attributes based on Weekday and Weekend would not help us getting a better accuracy is what we found out by experimenting.

### Crime frequency for Week Days



### Crime frequency for Weekends

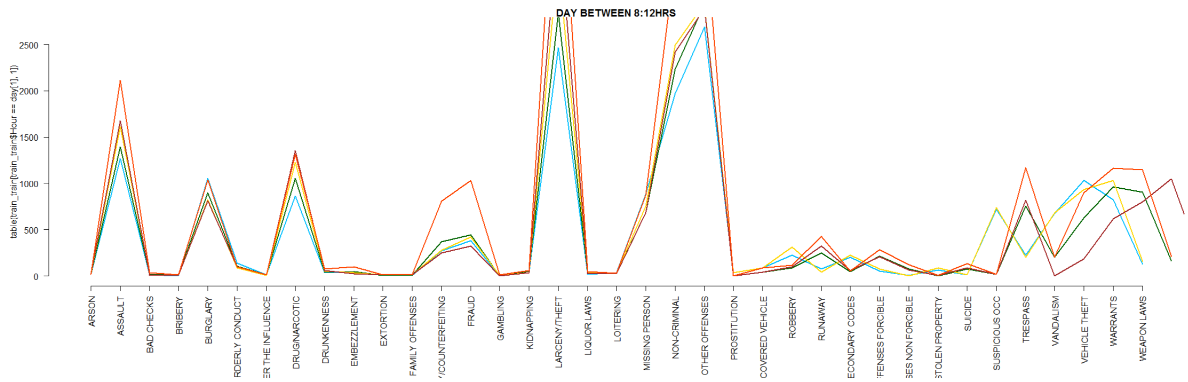



We then tried to find relationship between crimes and time of the day.For this we divided 24 hrs of a day into three categories where the first one consists of Night time from 10PM-7AM, this is usually when people are asleep; Morning-Late Afternoon category from 8AM-5PM this covers the time interval when people go to work and return back by 5PM and finally being the late Evening from 6PM-9PM.

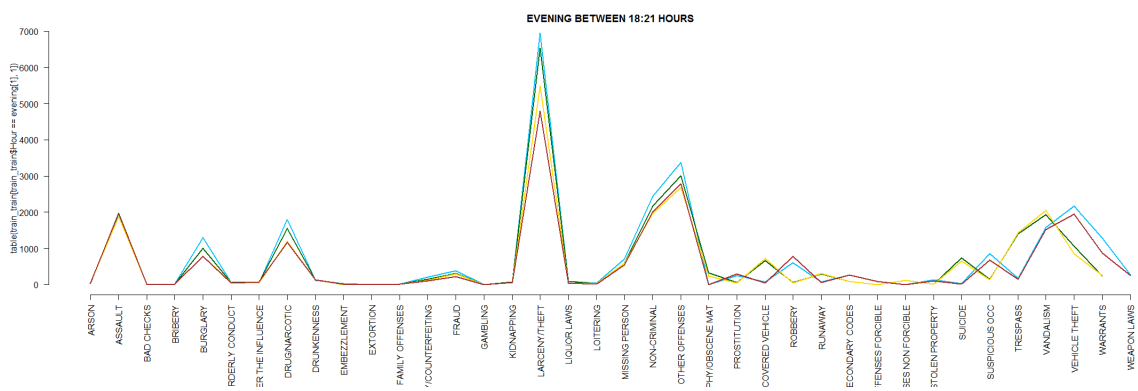
| Feature | Time range |
|---|---|
| Night | 10:00 PM to 7:00 AM |
| Morning-Late Afternoon | 8:00 AM to 5:00 PM |
| Evening | 6:00 PM to 9:00 PM |

**Table 2** Time category

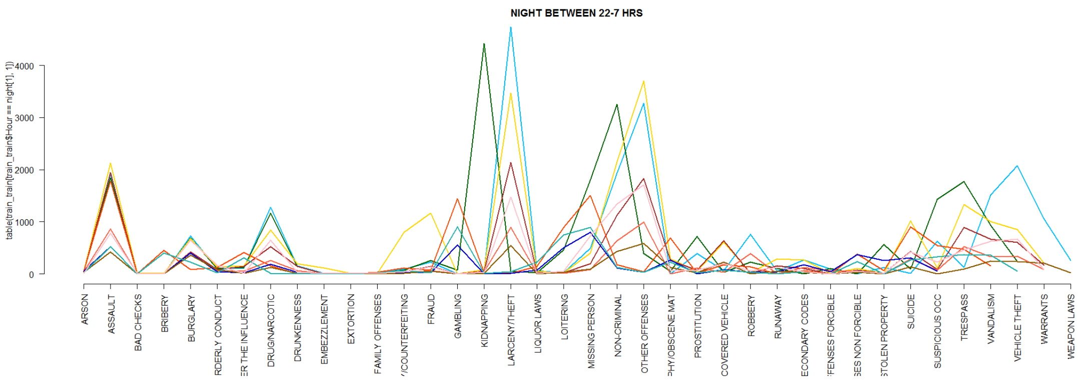
### 10:00 AM to 5:00 PM




### 6:00 PM to 9:00 PM



### 10:00 PM to 7:00 AM




From the above graphs one can see that,

1. Crimes like Gambling and Kidnapping has mostly occurred during night time.
2. While crime like suicide mostly has taken place during evening and some time during night time .
3. Crime like Burglary as mostly taken place during office hours between 8:00 AM to 5:00 PM.
4. One can closely observe that crimes happening during night from 10PM to 7AM are spread out, i.e crimes are not just concentrated to be of a particular types like the other graphs show, rather its spread over all crimes, or the crimes are more diversified.
5. Observing the different patterns in crimes on three different times we went ahead to divide the hours of a day to three categories and added them as new features.

On the plots shown previously, one can see that the most frequently occurring peaks are of 13 different types.So, instead of having 39 different labels which we considered as it might hinder the performance of the classifier, we decided on to reduced it to 8 different labels represented by the peaks in the graphs. The rest of the labels were labeled as **OTHERS**. To verify this we performed **K-means** clustering on the labels w.r.t the different addresses.

This turned out to be in agreement with our result by. K-Means We wanted to create cluster of locations based on crime. For this we used crime vector, structure of which is discussed in Data Transformation for K-Means section. This analysis helped us to filter out some of the Class labels .

**K-MEANS**

euclidean distance were used to calculate distance between cluster centroid and data points.
$d(p,q) = \sqrt{(p_1 - q_1)^2 + .... + (p_n - q_n)^2}$
parameters :
K=40
Max Iteration=40

**Algorithm to label a cluster**

1. for each Cluster
2. $ClassLabel \longleftarrow Max(\sum_{eachAttribute} \sum_{dataPoint} sum(AttributeFrequency))$

**Cluster**

| Cluster Size | Cluster Name |
|---|---|
| 23 | LARCENY.THEFT |
| 112 | LARCENY.THEFT |
| 1 | NON.CRIMINAL |
| 92 | ASSAULT |
| 26 | OTHER.OFFENSES |
| 543 | LARCENY.THEFT |
| 337 | LARCENY.THEFT |
| 1 | MISSING.PERSON |
| 3 | MISSING.PERSON |
| 280 | OTHER.OFFENSES |
| 8 | DRUG.NARCOTIC |
| 44 | DRUG.NARCOTIC |
| 19 | ASSAULT |
| 25 | MISSING.PERSON |
| 44 | NON.CRIMINAL |
| 102 | OTHER.OFFENSES |
| 55 | LARCENY.THEFT |
| 12 | DRUG.NARCOTIC |
| 86 | LARCENY.THEFT |
| 10 | PROSTITUTION |
| 77 | OTHER.OFFENSES |
| 7 | NON.CRIMINAL |
| 896 | OTHER.OFFENSES |
| 1 | LARCENY.THEFT |
| 14 | LARCENY.THEFT |
| 16 | OTHER.OFFENSES |
| 1 | LARCENY.THEFT |
| 28 | ASSAULT |
| 1397 | ASSAULT |
| 10 | MISSING.PERSON |
| 1439 | LARCENY.THEFT |
| 9 | LARCENY.THEFT |
| 315 | ASSAULT |
| 22 | DRUG.NARCOTIC |
| 11851 | OTHER.OFFENSES |
| 26 | LARCENY.THEFT |
| 1 | DRUG.NARCOTIC |
| 262 | LARCENY.THEFT |
| 4936 | LARCENY.THEFT |
| 97 | NON.CRIMINAL |

**Table 5** Time category

# 4 Data Transformation

Most of the feature which we have used are discrete. To make our regression algorithm work we transformed them into numerical values being represented by 0's and 1's. We achieved this through one-hot encoding or Dummy encoding.In this encoding if a categorical attribute takes N different values that particular feature is divided into N different features. Eg.Day of Week which can take 7 unique values "Monday" to "Sunday" is converted into 7 attributes where each value is being represented by either 0 or 1.
$\Phi(WeekDays) = \{Mo, Tu, We, Th, Fr, Sa, Su\}$
E.g. $\Phi(Tuesday) = \{0, 1, 0, 0, 0, 0, 0\}$
Following variables has been converted into numeric value using one-hot encoding.

| **Attribute name** |
|---|
| Months |
| DayOfWeek |
| PdDistrict |
| Year |
| Address |

**Table 4**: Features transformed using Dummy encoding

**Our final set of Feature space now includes:**

**Year** | **Month** | **Day** | **DayOfWeek** | **Late-Night** | **Morning-Evening** | **Late-Evening** | **X** | **Y** | **Address**

# 5 Modeling the data

We have used following algorithm for prediction ,clustering and other statistical analysis.

1. Naive Bayes.
2. Multi-Class Logistic Regression.
3. Apriori Algorithm

## 5.1 Naive Bayes

For Naive Bayes classification we have used following set of attributes. Our assumption is that these predictor variable are conditionally independent given the class label.

$$\mathbf{P}(CV \in \{8crimes\} \mid PV) = max_{8crimes}(\mathbf{P}(CV) \times \prod_{PV=1}^{8} \mathbf{P}(PV \mid CV))$$

With above model we did 10 fold cross validation where we split our data into 10 equal subsets and trained on 9 subsets, then tested it on the 10th subset, and repeated this process.Through this we got an accuracy of around 17%.

## 5.2 Logistic Regression

Initially on performing Logistic regression on raw data we got an accuracy of 21%. However, on performing feature engineering and decreasing the number of labels we got an accuracy of 25% over 10-fold cross validation,which appears to be a significant gain given the number of labels..

$$P(y = j|\mathbf{x}) = \frac{e^{\mathbf{x}^T w_j}}{\sum\limits_{k=1}^{K} e^{\mathbf{x}^T w_k}}$$

**ANUP PLEASE PASTE THE KMEANS PART AND ITS RESULTS IN THE PLACE THAT I HAVE MENTIONED PREVIOUSLY**

## 5.3 Apriori Algorithm

We used Apriori algorithm to find association between crimes. This helped us to know which crime is most likely to happen with other crime . Some of the top association rule are given below.
Parameter used :
Support : 0.4
confidence :0.1

| LHS | RHS | support | Confidence | lift |
|---|---|---|---|---|
| ASSAULT | OTHER.OFFENSES | 0.4042535 | 0.8774061 | 1.358874 |
| ASSAULT | LARCENY.THEFT | 0.4079559 | 0.8854420 | 1.274764 |
| VANDALISM | OTHER.OFFENSES | 0.4169967 | 0.8350720 | 1.293309 |
| VANDALISM | LARCENY.THEFT | 0.4462287 | 0.8936115 | 1.286526 |
| VEHICLE.THEFT | NON.CRIMINAL | 0.4203547 | 0.7054913 | 1.281849 |
| VEHICLE.THEFT | OTHER.OFFENSES | 0.4663337 | 0.7826590 | 1.212135 |
| VEHICLE.THEFT | LARCENY.THEFT | 0.5015068 | 0.8416908 | 1.211776 |

**Table 5** Time category

From above table we can Infer following facts about crimes .

1. Vehicle theft is more closely associated with Larceny . This might be due to the fact that same kind of criminal mentality works to do these crimes.
2. Vandalism is associated with Larceny . This can be explained by the fact that person involved in Larceny doesn't care much about preserving property .
3. Assault is associated with LARCENY . Which is obvious as assault is same often go side by side with larceny

| | Heading 1 | Heading 2 | Heading 3 |
|---|---|---|---|
| Row 1 | Cell 1,1 | Cell 1,2 | Cell 1,3 |
| Row 2 | Cell 2,1 | Cell 2,2 | Cell 2,3 |

# 6 Results

| Algorithm | Accuracy | Base Line |
|---|---|---|
| Naive Bayes | 17 | NA |
| Multi Class Logistic Regression | 25 | NA |

**Table 5** Result

# 7 Conclusion

I believe that we could have achieved a better accuracy on performing Feature Engineering on other attributes like Year and Address. But due to the enormous amount of data and a large number of labels the accuracy is being hindered.

## 7.1 Future Work

# 8 References

## 8.1 PEOPLE

## 8.2 BOOKS