Assignment No : 3

INFO 7390 : Advances in Data Sciences/Architecture

Team 7

Chandni Sharma
Hari Panjwani
Sanath Shetty

Table of Contents

1. Problem Statement 3
2. Data Wrangling & Cleansing 4
3. Model Selection 10
4. Selecting the Best Model 16
5. Analysis on Dataset 18
6. References 21

1. Problem Statement for SFO Crime Dataset

1.1 **Project Title:** San Francisco Crime Classification

1.2 **Domain:** Government Data

1.3 **Description:** San Francisco is currently the cultural, commercial and financial center of Northern California. Today the city is known more for its tech scene but it has a massive criminal past. The sudden growth in the population has brought an inequality in terms of living, housing shortages leading to no scarcity of crime in the city by the bay.

1.4 **Problems to Address:** The project aims at accurately predicting the category of crime based on the twelve years of records. Provided in the dataset.

1.5 **Machine Learning Algorithms Used:** Multi-Class Decision Forest, Multi-Class Decision Jungle, Multi-Class Neural Network.

1.6 **Technologies and Tools:** Microsoft Azure Machine Learning Studio, R Studio, Java, Power BI, Spring Tool Suite, Bootstrap, jQuery, REST API.

1.7 **Business Case:** Our analysis could help the police department to get an overall view on the category of crime occurring in a particular area. Based on our analysis the police department could set up extra patrolling/ checks in notorious areas to avoid criminal activities in the city of San Francisco.

2. Data Wrangling & Cleansing

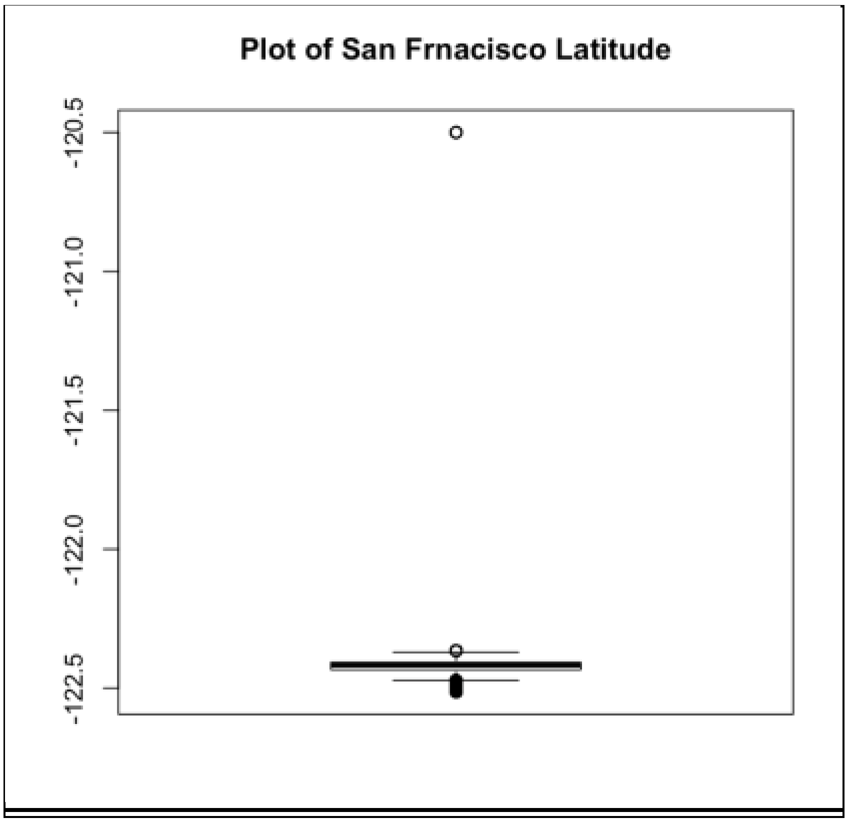
2.1 **<u>About the dataset:</u>** The dataset contains around 8 lakh rows. The dataset contains record about the crime incidents that has occurred in the city of San Francisco. It contains information about the category of crime, date of the incident, latitude and longitude of the location where the crime occurred, district (where the incident has occurred) and time of the incident. Below is the summary of the dataset:

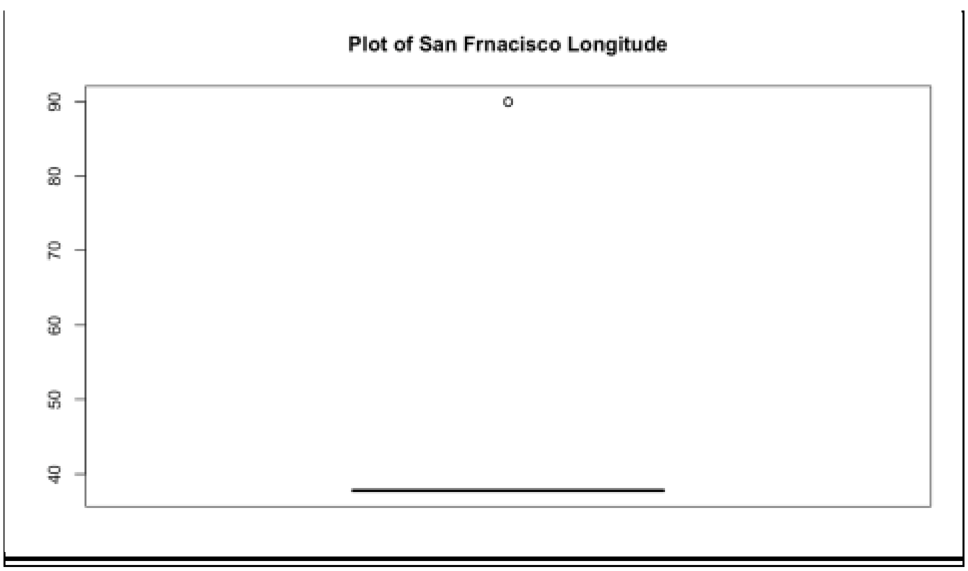
	Category	LARCENY/THEFT :174900
	Category	OTHER OFFENSES:126182
	Category	NON-CRIMINAL : 92304
	Category	ASSAULT : 76876
	Category	DRUG/NARCOTIC : 53971
	Category	VEHICLE THEFT : 53781
	Category	(Other) :300035
	Descript	GRAND THEFT FROM LOCKED AUTO : 60022
	Descript	LOST PROPERTY : 31729
	Descript	BATTERY : 27441
	Descript	STOLEN AUTOMOBILE : 26897
	Descript	DRIVERS LICENSE, SUSPENDED OR REVOKED: 26839
	Descript	WARRANT ARREST : 23754
	Descript	(Other) :681367
	DayOfWeek	Friday :133734
	DayOfWeek	Monday :121584
	DayOfWeek	Saturday :126810
	DayOfWeek	Sunday :116707
	DayOfWeek	Thursday :125038

2.2 <u>Data Cleaning:</u> The initial step was reading the data from the source and creating a dataset with column names in a more readable and verbose form. Then we proceeded with the data wrangling methods taught to us. We first took the summary of the entire dataset which helped us in understanding the nature of the data. Then we started excluding all the redundant data from the dataset.

2.3 <u>Outlier Analysis:</u> By using the boxplot function we plotted the graph for the numeric values of latitude and longitude in the dataset. We then remove these values from our dataset. Then we proceeded with analyzing each column in the dataset and we found that most of the data could be used as it is and there is not a lot of cleaning that is needed. We then removed the NA's present in the Date column of the dataset. As there were only 47 rows containing NA's we could afford to discard them from the dataset as it is not even 1 percent of the dataset.

2.4 Snapshot of Outlier Analysis:





2.5 Feature Engineering:

We have used feature engineering principle and have come up with important features to analyze the dataset.

i. **Category Map:** There were a total of 39 crimes that were mentioned in the dataset. We normalized the Crime category data and came up with percentages of occurrence of each crime in the dataset. Below is the normalized list :

	Var1	Freq	round.CategoryFreqPercentage.
1	LARCENY/THEFT	174894	20
2	OTHER OFFENSES	126179	14
3	NON-CRIMINAL	92297	11
4	ASSAULT	76871	9
5	DRUG/NARCOTIC	53970	6
6	VEHICLE THEFT	53779	6
7	VANDALISM	44721	5
8	WARRANTS	42213	5
9	BURGLARY	36749	4
10	SUSPICIOUS OCC	31413	4
11	MISSING PERSON	25986	3
12	ROBBERY	22998	3
13	FRAUD	16679	2
14	FORGERY/COUNTERFEITING	10609	1
15	SECONDARY CODES	9985	1
16	WEAPON LAWS	8555	1
17	PROSTITUTION	7483	1
18	TRESPASS	7326	1
19	STOLEN PROPERTY	4539	1
20	SEX OFFENSES FORCIBLE	4388	0
21	DISORDERLY CONDUCT	4319	0
22	DRUNKENNESS	4279	0
23	RECOVERED VEHICLE	3138	0
24	KIDNAPPING	2340	0
25	DRIVING UNDER THE INFLUENCE	2267	0
26	RUNAWAY	1946	0
27	LIQUOR LAWS	1903	0
28	ARSON	1513	0
29	LOITERING	1225	0
30	EMBEZZLEMENT	1166	0
31	SUICIDE	508	0

As you can see that from category 19 all the crimes are nearly 0 we can ignore them. So now we have reduced the crime category to 19 crimes which can be seen below:

1	LARCENY/THEFT	174894	20
2	OTHER OFFENSES	126179	14
3	NON-CRIMINAL	92297	11
4	ASSAULT	76871	9
5	DRUG/NARCOTIC	53970	6
6	VEHICLE THEFT	53779	6
7	VANDALISM	44721	5
8	WARRANTS	42213	5
9	BURGLARY	36749	4
10	SUSPICIOUS OCC	31413	4
11	MISSING PERSON	25986	3
12	ROBBERY	22998	3
13	FRAUD	16679	2
14	FORGERY/COUNTERFEITING	10609	1
15	SECONDARY CODES	9985	1
16	WEAPON LAWS	8555	1
17	PROSTITUTION	7483	1
18	TRESPASS	7326	1
19	STOLEN PROPERTY	4539	1

ii. **<u>AddressMap</u>**: In the dataset it can be seen that for a particular crime two street names has been associated. So for crimes related with two street names we have given a value 0 and for crimes related with a single street name we have given a value 1.

iii. **<u>DayOfWeekMap:</u>** So using this variable we are segregating the crimes occurred on weekdays and weekends by giving the DayOfWeekMap as 0 for weekday and 1 for weekend (counting Friday as Weekend).

So finally we have our optimized dataset. Below is the summary:

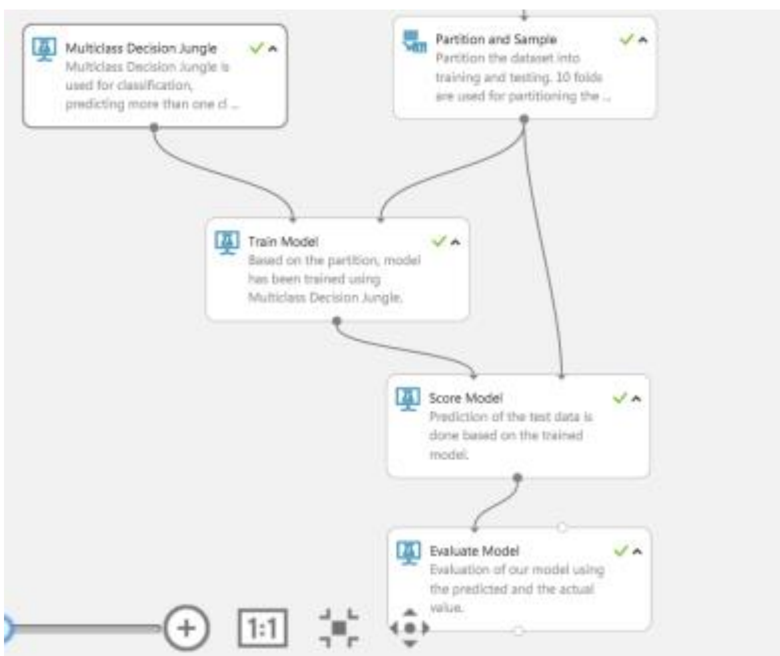
Var1	Var2	Freq
	PdDistrict	SOUTHERN :157171
	PdDistrict	MISSION :119905
	PdDistrict	NORTHERN :105291
	PdDistrict	BAYVIEW : 89425
	PdDistrict	CENTRAL : 85450
	PdDistrict	TENDERLOIN: 81809
	PdDistrict	(Other) :238951
	X	Min. :-122.5
	X	1st Qu.:-122.4
	X	Median :-122.4
	X	Mean :-122.4
	X	3rd Qu.:-122.4
	X	Max. :-120.5
	X	NA
	Y	Min. :37.71
	Y	1st Qu.:37.75
	Y	Median :37.78
	Y	Mean :37.77

3. Model Selection

3.1 Multi-Class Decision Jungle

For this assignment we have used Microsoft Azure Learning Studio to clean the dataset and to run the machine learning algorithms on our dataset. Once done with the cleaning using the Execute R-script option in Azure, we partition and sample the dataset in the ratio 0:70 to 0:30 in order to train our data and test the prediction. We use the train dataset to train our multi-class decision jungle model and the test dataset to predict/classify the crime category in the test dataset based on the model created on the train dataset.

3.1.1 Snapshot of Multi-Class Decision Jungle (Azure)



3.1.2 Performance Metrics

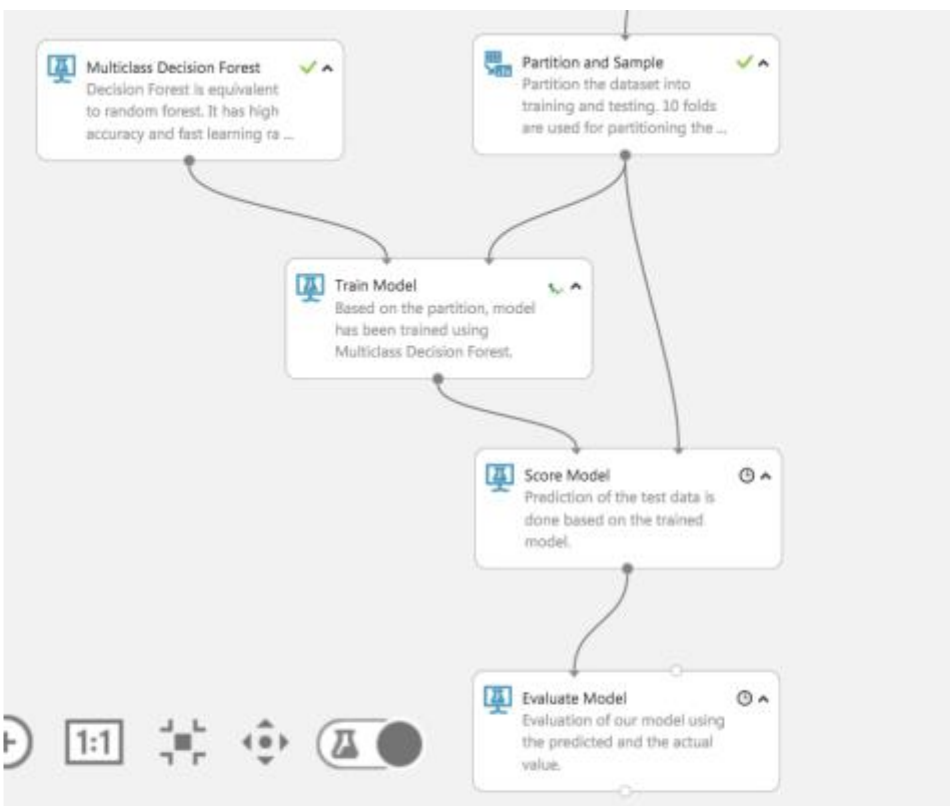
Metrics

Overall accuracy	0.276586
Average accuracy	0.923851
Micro-averaged precision	0.276586
Macro-averaged precision	0.35834
Micro-averaged recall	0.276586
Macro-averaged recall	0.118527

3.2 Multi-Class Decision Forest

For this assignment we have used Microsoft Azure Learning Studio to clean the dataset and to run the machine learning algorithms on our dataset. Once done with the cleaning using the Execute R-script option in Azure, we partition and sample the dataset in the ratio 0:70 to 0:30 in order to train our data and test the prediction. We use the train dataset to train our multi-class decision jungle model and the test dataset to predict/classify the crime category in the test dataset based on the model created on the train dataset.

3.2.1 Snapshot of Multi-Class Decision Forest (Azure)



3.2.2 Performance Metrics:

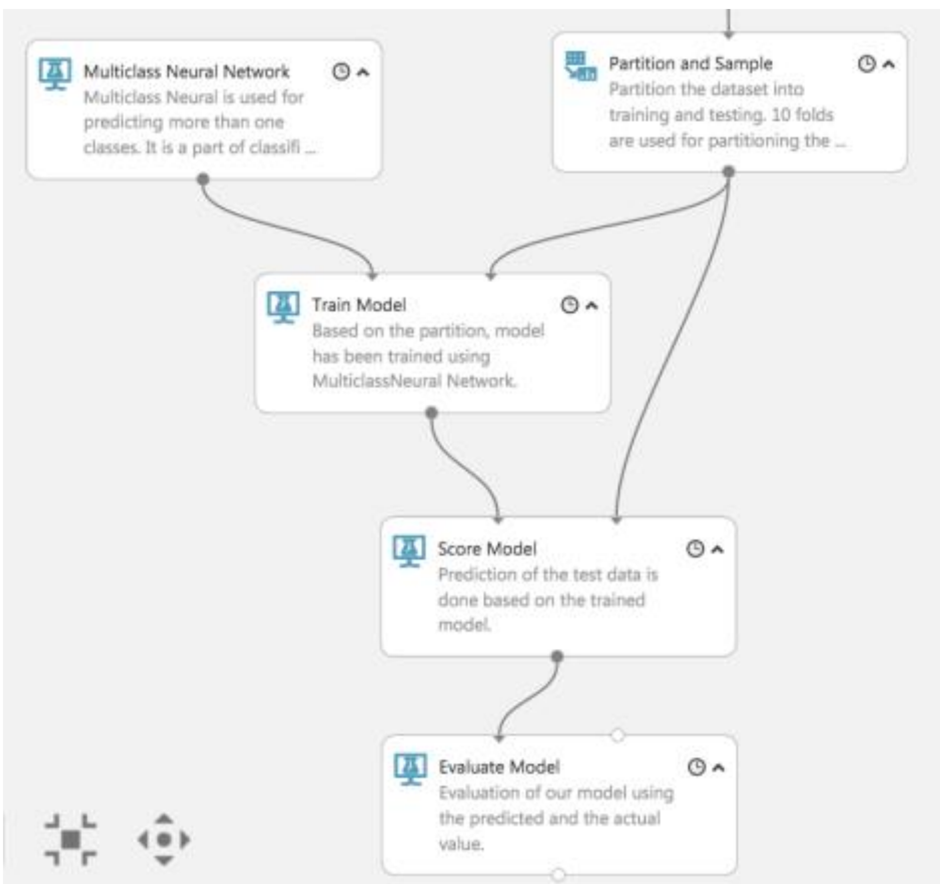
Metrics

Overall accuracy	0.826081
Average accuracy	0.981693
Micro-averaged precision	0.826081
Macro-averaged precision	0.772411
Micro-averaged recall	0.826081
Macro-averaged recall	0.749411

3.3 Multi-Class Neural Network

For this assignment we have used Microsoft Azure Learning Studio to clean the dataset and to run the machine learning algorithms on our dataset. Once done with the cleaning using the Execute R-script option in Azure, we partition and sample the dataset in the ratio 0:70 to 0:30 in order to train our data and test the prediction. We use the train dataset to train our multi-class decision jungle model and the test dataset to predict/classify the crime category in the test dataset based on the model created on the train dataset.

3.3.1 Snapshot of Multi-Class Neural (Azure)



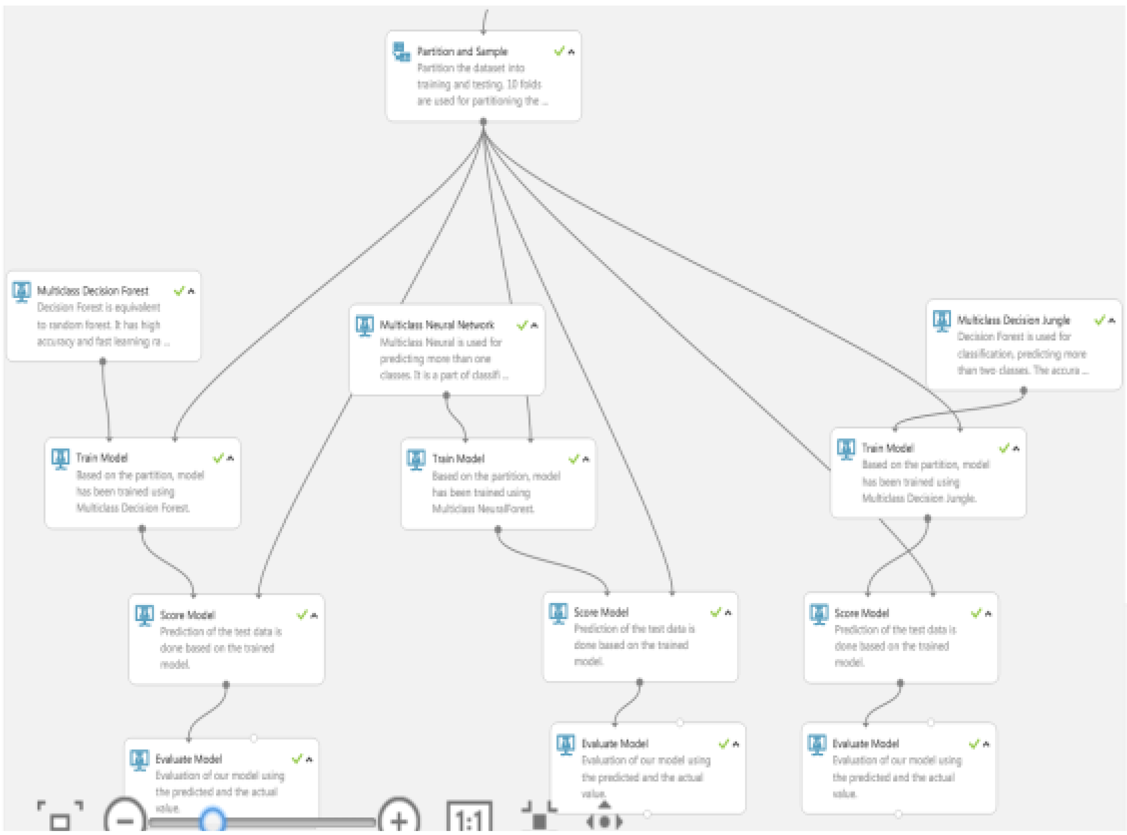
3.3.2 Performance Metrics

◢ Metrics

Overall accuracy	0.242235
Average accuracy	0.920235
Micro-averaged precision	0.242235
Macro-averaged precision	NaN
Micro-averaged recall	0.242235
Macro-averaged recall	0.090237

4. Selecting the Best Model

4.1 Snapshot of All Models:



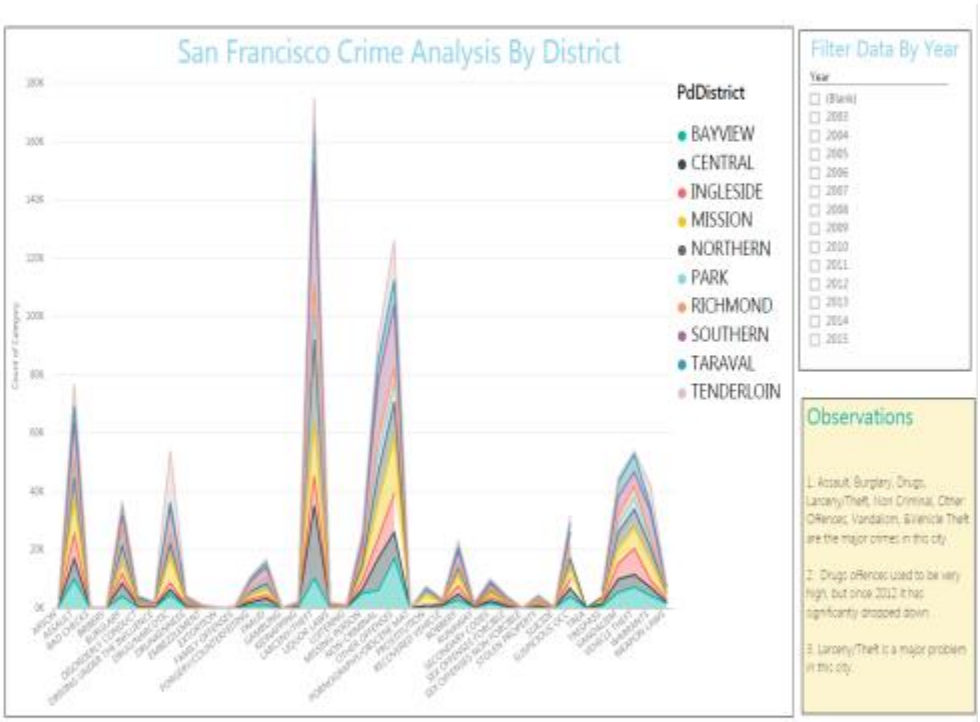
4.2 Comparing Performance Metrics:

Metrics	Jungle	Forest	Neural
Overall Accuracy	0.276586	0.826081	0.242235
Average Accuracy	0.923851	0.981693	0.920235
Micro-average Precision	0.276586	0.826081	0.0.242235
Macro-average Precision	0.35834	0.772411	NaN
Micro-average recall	0.276586	0.826081	0.242235
Macro-average recall	0.118527	0.749411	0.090237

Based on the above performance metrics, we have decided to select the Multi-Class Decision model as our best one as it has a good accuracy as compared to the others.

5. Analysis on the San Francisco Crime Dataset

5.1 Overall Crime Analysis By District



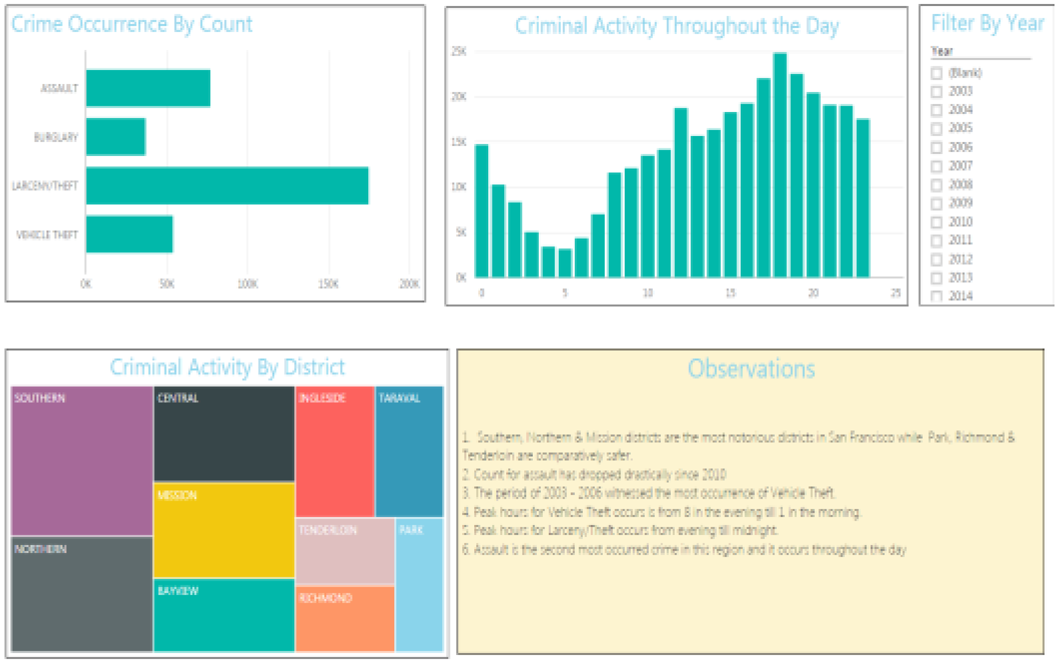
Observations:

1. Assault, Burglary, Drugs, Larceny/Theft, Non Criminal, Other Offences, Vandalism, &Vehicle Theft are the major crimes in this city

2. Drugs offences used to be very high, but since 2012 it has significantly dropped down.

3. Larceny/Theft is a major problem in the city

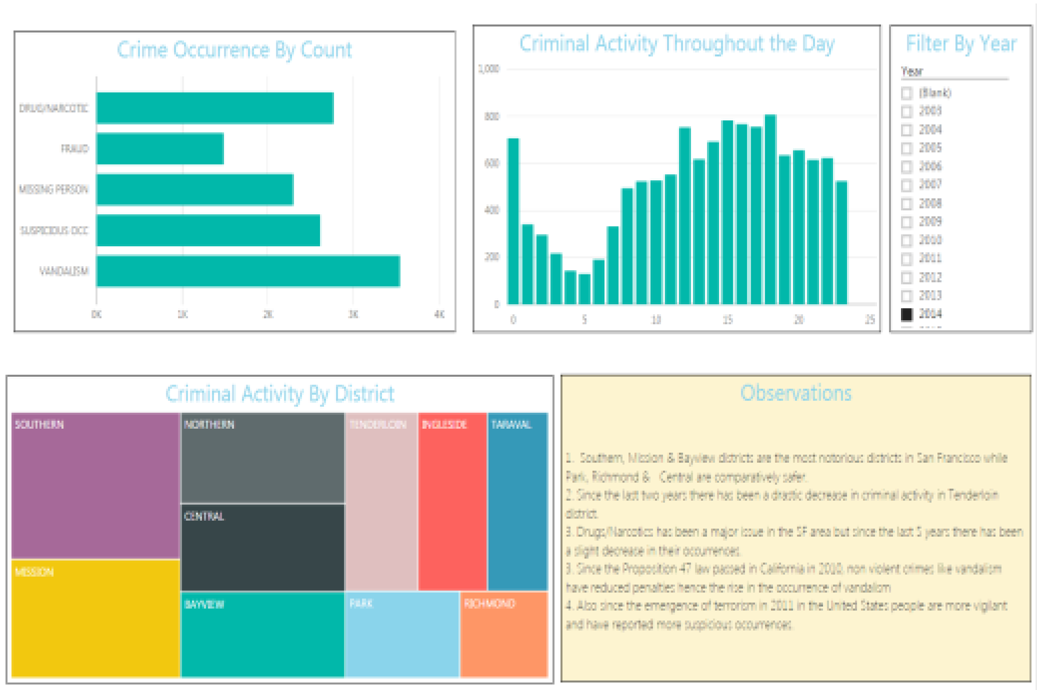
5.2 Serious Offences Overall



Observations:

1. Southern, Northern & Mission districts are the most notorious districts in San Francisco while Park, Richmond & Tenderloin are comparatively safer.
2. Count for assault has dropped drastically since 2010
3. The period of 2003 - 2006 witnessed the most occurrence of Vehicle Theft.
4. Peak hours for Vehicle Theft occurs is from 8 in the evening till 1 in the morning.
5. Peak hours for Larceny/Theft occurs from evening till midnight.
6. Assault is the second most occurred crime in this region and it occurs throughout the day

5.3 Non Serious Offences Overall



Observations:

1. Southern, Mission & Bayview districts are the most notorious districts in San Francisco while Park, Richmond & Central are comparatively safer.
2. Since the last two years there has been a drastic decrease in criminal activity in Tenderloin district.
3. Drugs/Narcotics has been a major issue in the SF area but since the last 5 years there has been a slight decrease in their occurrences.
4. Since the Proposition 47 law passed in California in 2010, non violent crimes like vandalism have reduced penalties hence the rise in the occurrence of vandalism
5. Also since the emergence of terrorism in 2011 in the United States people are more vigilant and have reported more suspicious occurrences.

(P.S.-more analysis present in PowerBI)

6. References

- Kaggle:
  https://www.kaggle.com/c/sf-crime
- Microsoft Azure Machine Learning Studio:
  https://azure.microsoft.com/en-gb/
- Microsoft Power BI:
  https://powerbi.microsoft.com/en-us/