CRIME TYPE PREDICTION

Anand Jain
Parikshitha Sampigecool Manjunath
Sarthak Pathak

INTRODUCTION

Crime is neither methodical nor entirely random. The high volume of crime datasets and the complexity of relationships among the types of crimes and locations has made it an emerging field to apply data mining techniques. Human intelligence plays a vital role in solving crimes but data mining techniques have been proved helpful in predicting crime across the globe. Law enforcement authorities need excellent mining techniques to ensure the security of people. In this paper, we will be analyzing various changes in crime patterns and trends with respect to the locations, neighborhood and time over the years (2011 to 2016) in the area of Baltimore.

We want to classify and predict the type of crime based on time and location. We also want to analyze the change in trend over the years to understand how the safety and security of different neighborhoods has been affected. We will implement two algorithms to compare the accuracy performance. We used R(ggmap, gplot, rparty, s), Weka and Microsoft Excel data analytics tools.

MOTIVATION

With the crime rate of 62 per one thousand residents, Baltimore has one of the highest crime rates in India. The crime rate in Baltimore has increased at an unprecedented rate in the past years. There are many factors responsible in the increment which will be highlighted in the paper. There are statistics which justify the type of crimes and their average in the past 5 years (2011-2015). Such kind of reports validates the reason behind Baltimore being one of the top 100 dangerous cities in the USA.

	2010	2011	2012	2013	2014	PREVIOUS FIVE-YEAR AVERAGE	2015	PERCENT CHANGE
TOTAL HOMICIDE, NON-NEGLIGENT MANSLAUGHTER:	**20**	**30**	**23**	**20**	**25**	**24**	**30**	**25.0%**
TOTAL RAPE:	**126**	**142**	**145**	**124**	**97**	**127**	**330**	**159.8%**
FORCIBLE RAPE	113	132	126	109	86	113	307	171.7%
ATTEMPTED RAPE	13	10	19	15	11	14	23	64.3%
TOTAL ROBBERY:	**1,335**	**1,451**	**1,367**	**1,510**	**1,512**	**1,435**	**1,525**	**6.3%**
HIGHWAY, STREETS, ETC	686	695	655	805	672	703	650	-7.5%
COMMERCIAL HOUSES	238	258	241	257	300	259	325	25.5%
GAS STATION OR SERVICE STATION	21	31	26	28	39	29	67	131.0%
CONVENIENCE STORE	67	85	82	102	125	92	170	84.8%
RESIDENTIAL (ANYWHERE ON PREMISE)	213	228	226	216	273	231	238	3.0%
BANKS, ETC	30	45	23	25	43	33	28	-15.2%
MISCELLANEOUS	80	109	114	77	60	88	47	-46.6%
FIREARM	530	582	516	574	659	572	718	25.5%
KNIFE OR CUTTING INSTRUMENT	144	129	118	166	165	144	138	-4.2%
OTHER DANGEROUS WEAPON	69	82	98	110	90	90	75	-16.7%
STRONGARM (HANDS, FIST, FEET, ETC.)	592	658	635	660	598	629	594	-5.6%
TOTAL AGGRAVATED ASSAULT:	**2,824**	**2,627**	**2,611**	**2,530**	**2,390**	**2,596**	**2,618**	**0.8%**
FIREARM	280	223	239	258	244	249	241	-3.2%
KNIFE OR OTHER SHARP INSTRUMENT	643	641	594	556	541	595	607	2.0%
OTHER DANGEROUS WEAPON	1,259	1,200	1,289	1,172	1,050	1,194	1,201	0.6%
HANDS, FIST, FEET, ETC	642	563	489	544	555	559	569	1.8%
TOTAL HUMAN TRAFFICKING*:	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**36**	**N/A**
COMMERCIAL SEX ACTS	N/A	N/A	N/A	N/A	N/A	N/A	36	N/A
INVOLUNTARY SERVITUDE	N/A	N/A	N/A	N/A	N/A	N/A	0	N/A
TOTAL PART I VIOLENT CRIME:	**4,305**	**4,250**	**4,146**	**4,184**	**4,024**	**4,182**	**4,539**	**8.5%**
TOTAL BURGLARY:	**4,089**	**4,269**	**4,061**	**3,930**	**3,475**	**3,965**	**3,569**	**-10.0%**
RESIDENTIAL (NIGHT)	361	386	439	656	934	555	892	60.7%
RESIDENTIAL (DAY)	1,049	1,189	1,025	1,523	1,852	1,328	1,548	16.6%
RESIDENTIAL (UNKNOWN)	1,106	1,263	1,630	954	4	991	3	-99.7%
NON-RESIDENTIAL (NIGHT)	391	376	356	384	416	385	609	58.2%
NON-RESIDENTIAL (DAY)	142	129	75	156	268	154	515	234.4%
NON-RESIDENTIAL (UNKNOWN)	1,040	926	536	257	1	552	2	-99.6%
FORCIBLE ENTRY	2,619	2,649	2,402	2,608	2,329	2,521	1,954	-22.5%
UNLAWFUL ENTRY	1,014	1,138	1,212	890	784	1,008	1,250	24.0%

Fig1. Statistics for crime in Baltimore

Our motivation for this project is to observe how time and space alone and the combination of the two can be used to predict the type of crime.

RELATED WORK

Data Mining is the study and analysis of criminology can be categorized into primary areas , crime control and crime suppression. De Bruin et. al. [1] presented a system for crime patterns utilizing another separation measure for contrasting all individuals based on their profiles and then clustering them accordingly. Nazlena Mohamad Ali et al.[2] discuss on a advancement of Visual Interactive Malaysia Crime News Recovery System (i-JEN) and portray the approach, user studies and planned, the framework engineering and future arrangement. Their primary targets were to develop crime based event; explore the utilization of crime based event in enhancing the clustering and classification; build up an interactive crime news recovery framework. A. Malathi et al.[3] look at the use of missing value and clustering algorithm calculation for a data mining way to deal with anticipate the crime patterns and quick up the procedure of unraveling crime. Malathi. An et. al.[4] utilized a clustering/classify based model to foresee crime trends. The aftereffects of this data mining could possibly be utilized to decrease and even counteract crime for the inevitable years.Dr. S. Santhosh Baboo and Malathi. A [5] research work focused on building up a crime analysis tool for Indian situation utilizing distinctive data mining techniques that can help law enforcement office to proficiently handle crime investigation. The proposed tool empowers organizations to effectively and monetarily spotless, describe and analyze crime data to distinguish significant examples and patterns.

These research work focused on analyzing crime trends using techniques like classification and clustering. These clusters helped to plot it on geospatial plot to identify the hot-spot or helped to predict the crime in a specific location. But, these papers didn't have any findings for prediction of crime given the time. The exact locations and streets were not considered which play a pivotal role in determination of results with type of crime event at a specific location.

DATASET

The humongous dataset is acquired from Baltimore Police Department which has 13 attributes summing up to 25,83,064 records. The nominal type of dataset has attributes listed as follows :-

- Crime Data
- Crime Time
- Crime Code
- Location
- Description
- Inside/Outside
- Post
- District
- Neighborhood
- Latitude
- Longitude

- Weapon
- Total Incidents

CrimeDate	CrimeTim	CrimeCod	Location	Description	Inside/(	Weapon	Post	District	Neighborhood	Longitude,Latitude	Total Incidents
12/3/2016	4:00:00	7A	4400 GROVELAND AVE	AUTO THEFT	O		621	NORTHWESTERN	West Arlington	(39.3407700000, -76.6942900	1
12/3/2016	5:05:00	3B	500 E PATAPSCO AVE	ROBBERY - STREET	O		913	SOUTHERN	Brooklyn	(39.2366800000, -76.6034200	1
12/3/2016	5:37:00	3AK	AV & BRUNSWICK ST	ROBBERY - STREET	O	KNIFE	841	SOUTHWESTERN	Millhill	(39.2780900000, -76.6587100	1
12/3/2016	6:30:00	5C	3400 OLD YORK RD	BURGLARY	I		543	NORTHERN	Waverly	(39.3300300000, -76.6083000	1
12/3/2016	7:45:00	4E	700 LINNARD ST	COMMON ASSAULT	I	HANDS	844	SOUTHWESTERN	Edgewood	(39.2966300000, -76.6750500	1
12/3/2016	8:00:00	5A	300 S PAYSON ST	BURGLARY	I		934	SOUTHERN	Carrollton Ridge	(39.2834700000, -76.6484500	1
12/3/2016	8:00:00	4E	600 BRIDGEVIEW RD	COMMON ASSAULT	I	HANDS	922	SOUTHERN	Cherry Hill	(39.2495200000, -76.6218900	1
12/3/2016	8:00:00	4E	600 BRIDGEVIEW RD	COMMON ASSAULT	I	HANDS	922	SOUTHERN	Cherry Hill	(39.2495200000, -76.6218900	1
12/3/2016	8:03:00	3CF	2200 E MONUMENT ST	ROBBERY - COMMERCIA	I	FIREARM	321	EASTERN	CARE	(39.2986500000, -76.5862300	1
12/3/2016	8:20:00	6D	300 CIDER ALY	LARCENY FROM AUTO	O		111	CENTRAL	Downtown	(39.2881700000, -76.6194100	1
12/3/2016	9:20:00	6C	3200 TIOGA PKWY	LARCENY	O		731	NORTHWESTERN	Mondawmin	(39.3180000000, -76.6582100	1
12/3/2016	9:20:00	6J	100 W UNIVERSITY PKWY	LARCENY	I		541	NORTHERN	Tuscany-Canterbury	(39.3353800000, -76.6220400	1
12/3/2016	9:58:00	6C	1300 E NORTH AVE	LARCENY	I		342	EASTERN	East Baltimore Midway	(39.3119200000, -76.6002400	1
12/3/2016	10:00:00	5C	0 S FULTON AVE	BURGLARY	I		933	SOUTHERN	Union Square	(39.2870100000, -76.6451000	1
12/3/2016	10:00:00	5A	2800 HOLLINS FERRY RD	BURGLARY	I		923	SOUTHERN	Lakeland	(39.2571600000, -76.6449500	1
12/3/2016	10:00:00	6D	1200 HAVERHILL RD	LARCENY FROM AUTO	O		832	SOUTHWESTERN	Violetville	(39.2670600000, -76.6757600	1

Fig 2. Original Dataset

DATA PREPROCESSING

The dataset was preprocessed in various steps to remove the various missing values and make it much cleaner for analysis and interpretation.

- Data Cleaning – The dataset had many missing values in different attributes. The blank tuples were ignored in the dataset in order to avoid discrepancy and make the data more consistent. The Crime Time attribute was binned into four bins of Night (00:00:00-05:59:00), Morning (06:00:00-11:59:00), Afternoon (12:00:00 – 17:59:00) and Evening (18:00:00 – 23:59:00).

- Data Reduction – In this step, five irrelevant attributes were removed and dataset was reduced to 8 attributes and latitude, longitude were split into two columns. The removed attributes were Crime Code, Post, Inside/Outside, Weapon, Total Incidents. The street numbers from the Location attribute was removed to avoid overfitting and then the dataset looked as shown below from year 2011 to 2016 :-

CrimeDate	CrimeTim	Location	District	Neighborhood	lat	lon	Description
12/31/2013	Morning	TIOGA PW	NORTHWESTERN	Mondawmin	39.318	-76.6582	LARCENY
12/31/2013	Morning	LIBERTY HGTS AV	WESTERN	Mondawmin	39.31868	-76.654	LARCENY
12/31/2013	Morning	IRIS AV REAR	EASTERN	Orangeville	39.30207	-76.5625	COMMON ASSAULT
12/31/2013	Morning	SUNSET RD	NORTHERN	Levindale	39.35199	-76.6647	AUTO THEFT
12/31/2013	Morning	MAYFIELD AVE	NORTHEASTERN	BelairEdison	39.32142	-76.5714	AGG. ASSAULT
12/31/2013	Morning	N LINWOOD AV	EASTERN	Berea	39.30749	-76.5768	BURGLARY
12/31/2013	Morning	HEIGHTS AV & HILTON RD	NORTHWESTERN	Ashburton	39.32418	-76.6715	COMMON ASSAULT
12/31/2013	Morning	N FREMONT AV	WESTERN	Harlem Park	39.29738	-76.6336	LARCENY
12/31/2013	Morning	N CAROLINE ST	EASTERN	DunbarBroadway	39.29732	-76.5975	LARCENY
12/31/2013	Morning	FEDERAL ST	EASTERN	Berea	39.30906	-76.5809	ROBBERY - RESIDENCE
12/31/2013	Morning	WHITE AV	NORTHEASTERN	Frankford	39.34033	-76.5346	BURGLARY
12/31/2013	Morning	N CHARLES ST	NORTHERN	Charles North	39.31237	-76.6167	AGG. ASSAULT
12/31/2013	Morning	FEDERAL ST	EASTERN	Berea	39.309	-76.5825	ROBBERY - RESIDENCE
12/31/2013	Morning	QUEEN ANNE RD	NORTHWESTERN	Windsor Hills	39.31551	-76.688	BURGLARY
12/31/2013	Morning	W BALTIMORE ST	WESTERN	Poppleton	39.28901	-76.6284	LARCENY

Fig3. Preprocessed Dataset

The 'Description' attribute was chosen as the Class for further application of algorithms to find the results. To build a better class, the description attribute was grouped to similar types of crime. The tuple of Homicide present in this attribute was removed due to the missing values. The way we grouped types of crime are:

Earlier Crime Description	New Crime Description
Agg. Assault, Assault by threat, Common assault	Assault
Arson	Arson
Auto Theft	Auto Theft
Burglary	Burglary
Larceny, Larceny from Auto	Larceny
Rape	Rape
Robbery - Carjacking, commercial, Residence, street	Robbery
Shooting	Shooting

Table1. Grouping of Crimes

METHODOLOGY

Since our objective was to predict the crime type given the time and space alone and the combination of time and space. We chose Naive Bayes algorithm to predict our model. As our training dataset is high Bias and Low variance that is why we choose Naive Bayes algorithm to support the hypothesis. As Naive Bayes algorithm work well to detect good accuracy for this type of huge data set. But we also tries different algorithms like Decision tree J48, Simple Logistics to compare the accuracy with Naive Bayes.

So we carried our project in different iterations to apply different combination in order to check the accuracy and find out the good result and prediction.

Iteration 1:

In this iteration we applies the classification algorithm on the training set with 8 different classes. Below is the detail of iteration:

Classification Method : Naive Bayes Algorithm

Description(Classes) : Assault, Arson, Auto Theft, Burglary, Larceny, Rape, Robbery, Shooting.

Cross Validation : 10 folds

Accuracy : 41.118%

We also tried different combination of attributes with class attribute to check the accuracy and by which we can find any relationship.

Crime time	**Description**	**38.389%**
Crime date	**Description**	**36.799%**
location	**Description**	**40.158%**
District	**Description**	**38.225%**
Neighborhood	**Description**	**40.369%**

Table 2. Different attribute combination accuracy

As we observed that the accuracy is less when we took all the attribute and it is even less when we did all the attribute with class. To check the reason for this less accuracy can be large amount of dataset,

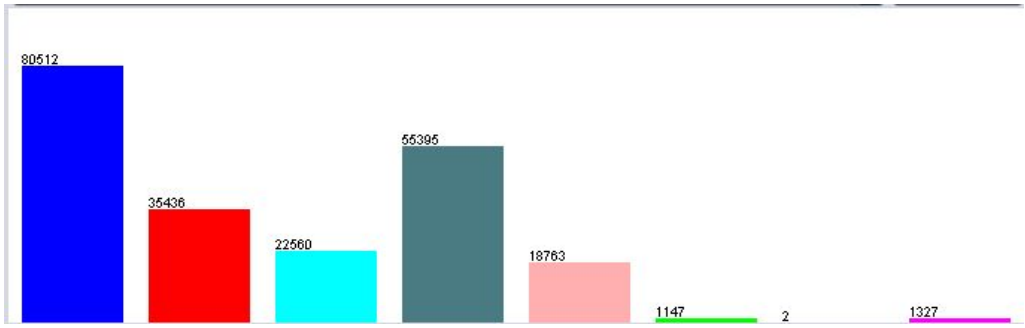
unequal class distribution, more number of classes, high biases. We were unable to find any relationship with crime description(class)

Class distribution :

No.	Label	Count	Weight
1	LARCENY	80512	80512.0
2	BURGLARY	35436	35436.0
3	AUTOTHEFT	22560	22560.0
4	ASSAULT	55395	55395.0
5	ROBBERY	18763	18763.0
6	ARSON	1147	1147.0
7	SHOOTING	2	2.0
8	RAPE	1327	1327.0

fIG 4 . CLASS DISTRIBUTION

After the missing values were removed , there were less number of instances in the Crime Types Shooting and Rape which were overshadowed by the class types Larceny and Assault which has the high amount of instances. As we see the Larceny has larger amount, so if try to train our model, then



it will train the model based on larceny data and did not take the other class into consideration. So it will always classify the other class into larceny. So the given class is misclassified, this is the problem of overfitting here. To examine the reason for low accuracy we went further and saw various results like crime time distribution, confusion matrix.

The Crime Time was binned into four bins and the distribution is shown below:

No.	Label	Count	Weight
1	Night	29395	29395.0
2	Morning	43865	43865.0
3	AfterNoon	69834	69834.0
4	Evening	72048	72048.0

Fig 6. Time attribute distribution

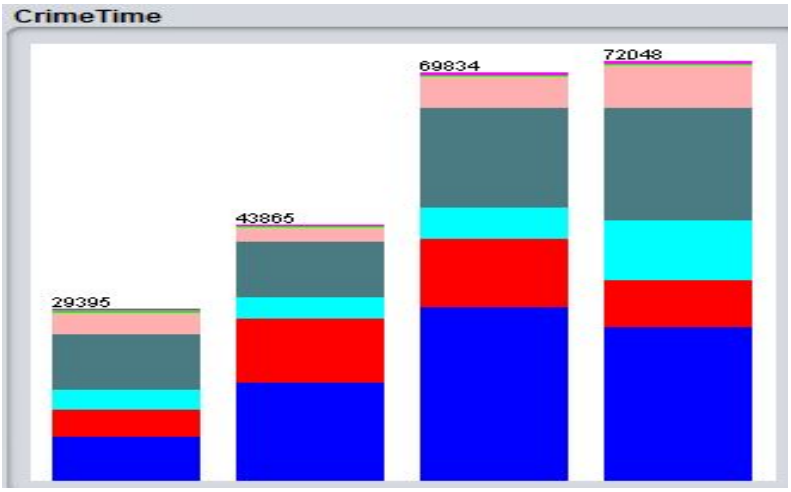


Fig 7. Crime Time visualize distribution

The distribution shows that the crime happened maximum during the Evening and the crime type that happened in excess was Larceny and Assault. The distribution for the District attribute is shown as below:

No.	Label	Count	Weight
1	CENTRAL	24694	24694.0
2	NORTHERN	24492	24492.0
3	SOUTHEASTERN	30205	30205.0
4	NORTHEASTERN	34364	34364.0
5	SOUTHERN	24576	24576.0
6	EASTERN	17814	17814.0
7	WESTERN	17345	17345.0
8	SOUTHWESTERN	19754	19754.0
9	NORTHWESTERN	21898	21898.0

Fig 8. District Distribution

The distribution for the District attribute showed us that Northeastern district was affected with high crime rate.

Below, the figure shows the summary of the approach that was applied:

```
=== Summary ===

Correctly Classified Instances       88464               41.1189 %
Incorrectly Classified Instances    126678               58.8811 %
Kappa statistic                          0.1744
Mean absolute error                      0.1678
Root mean squared error                  0.3024
Relative absolute error                 89.7256 %
Root relative squared error             98.8938 %
Total Number of Instances           215142

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.613    0.376    0.494      0.613    0.547      0.230    0.665     0.559     LARCENY
                 0.279    0.113    0.328      0.279    0.302      0.178    0.673     0.298     BURGLARY
                 0.150    0.056    0.238      0.150    0.184      0.116    0.659     0.181     AUTOTHEFT
                 0.458    0.270    0.370      0.458    0.409      0.176    0.652     0.375     ASSAULT
                 0.025    0.009    0.213      0.025    0.044      0.045    0.631     0.139     ROBBERY
                 0.002    0.000    0.049      0.002    0.003      0.008    0.647     0.011     ARSON
                 0.000    0.000    0.000      0.000    0.000      0.000    0.903     0.000     SHOOTING
                 0.000    0.000    0.000      0.000    0.000      -0.001   0.556     0.007     RAPE
Weighted Avg.    0.411    0.236    0.378      0.411    0.383      0.177    0.659     0.386
```

Confusion Matrix for the approach is shown below:

```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
 49338  8566  4262 17715   610    16     0     5 |     a = LARCENY
 12701  9904  1890 10713   222     6     0     0 |     b = BURGLARY
  8068  3326  3391  7494   271     9     0     1 |     c = AUTOTHEFT
 19757  6355  3321 25365   590     3     0     4 |     d = ASSAULT
  9262  1715  1151  6165   464     5     0     1 |     e = ROBBERY
   320   174   131   510    10     2     0     0 |     f = ARSON
     0     1     0     1     0     0     0     0 |     g = SHOOTING
   504   157    87   567    12     0     0     0 |     h = RAPE
```

In the above matrix as we can see the a given larceny the classified instances as larceny is higher. So this skewed dataset will always gives us incorrect prediction. With this False positive rate is also high. Which is the reason to classify instances incorrectly.

With all this study , we studied the data by just relating every single attribute to the Description (class) to know how strongly every single attribute is related to the class. This Iteration had no accuracy with which we could predict anything or learn much from the results because of the uneven distribution of class types.

Iteration 2:-

In this iteration we decreased some classes. We had 8 types of crime. So, we clubbed Robbery, Shooting, Rape, and Aggravated Assault into a single crime name called Violent Crime. This was clubbed with reference to the definition provided by FBI's Uniform Crime Reporting (UCR) Program which says, Violent crime is composed of murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault [6]. Also, Auto Theft and Burglary was clubbed to Theft, and Assault by Threat and common assault was clubbed to just Assault.

Classification method : Naive Bayes Algorithm

Description : Violent Crime, Arson , Theft, Larceny, Assault.

Cross Validation : 10 folds

Accuracy : 41.046%

Class distribution:

No.	Label	Count	Weight
1	LARCENY	80612	80612.0
2	THEFT	57996	57996.0
3	VIOLENT CRIME	39772	39772.0
4	ASSAULT	36560	36560.0
5	ARSON	1147	1147.0

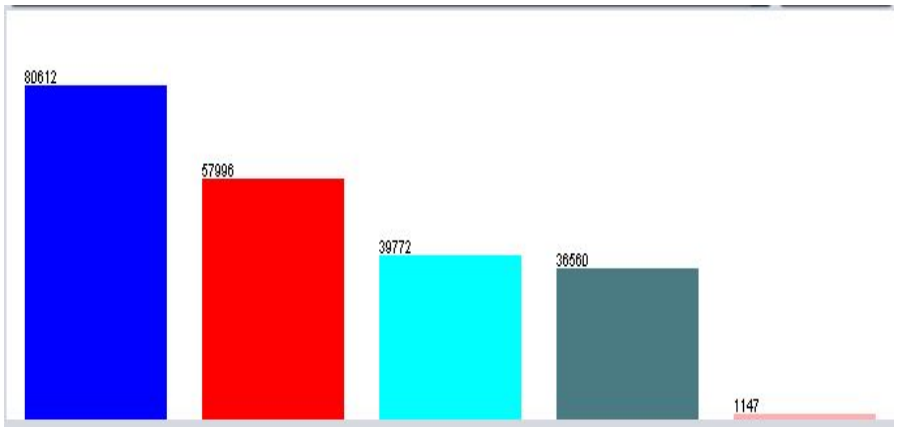


Fig 9. Crime description distribution

This above distribution gave lesser accuracy than the iteration 1. Arson with just 1147 instances got overshadowed by Larceny which has the highest number of instances.

We again tried different combinations of attribute and classes to test the accuracy and predict a relationship between them.

Attributes	Class	Accuracy
Crime time	Description	36.60%
Crime date	Description	37.57%
location	Description	40.50%
District	Description	37.62%
Neighborhood	Description	40.08%

Table 4. Different attributes accuracy with class

Iteration 3:-

In our iteration 3 we undersampled the Larceny data to get better results through which we could read and understand the data. We also clubbed Arson and Theft into Property Crime

Classification method : Naive Bayes Algorithm

Description : Violent Crime, Property Crime,Larceny, Assault.

Cross Validation : 10 folds

Accuracy : 40.24%

Class Distribution:

No.	Label	Count	Weight
1	PROPERTY CRIME	59143	59143.0
2	VIOLENT CRIME	39772	39772.0
3	ASSAULT	36560	36560.0
4	LARCENY	69379	69379.0

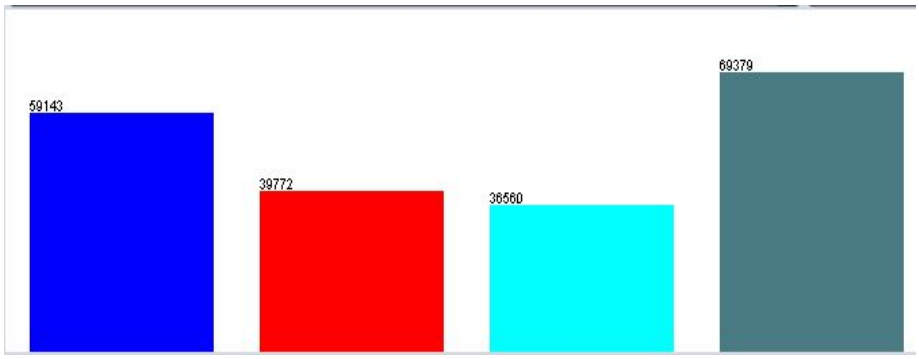


Fig 10. Class new distribution

Iteration 4:-

In our iteration 4 , we reduced our classes to just two classes Violent crime and Property crime.

Description : Violent Crime, Property Crime

Cross Validation : 10 folds

Accuracy : 61.32%

Hypothesis Baseline - The majority label classifier accuracy : 0.58.

Class Distribution:

Fig 11. Class Distribution with Iteration 4

As we can see the there is only two classes where larceny is under sampled and merged into property crime.

We also split the Crime Date to Date , Month and Year separately to check their accuracy.

CrimeMonth	CrimeYear	CrimeTime	Location	District	Description	**Accuracy= 61.174%**

CrimeYear	Location	Description	**Accuracy= 60.346%**

The accuracy is higher for this iteration from our baseline which is the sum of all the higher instances divided by the total number of instances. The reason for this accuracy can be decreased data set and decreased equal class distribution.

Comparison Between all iterations:

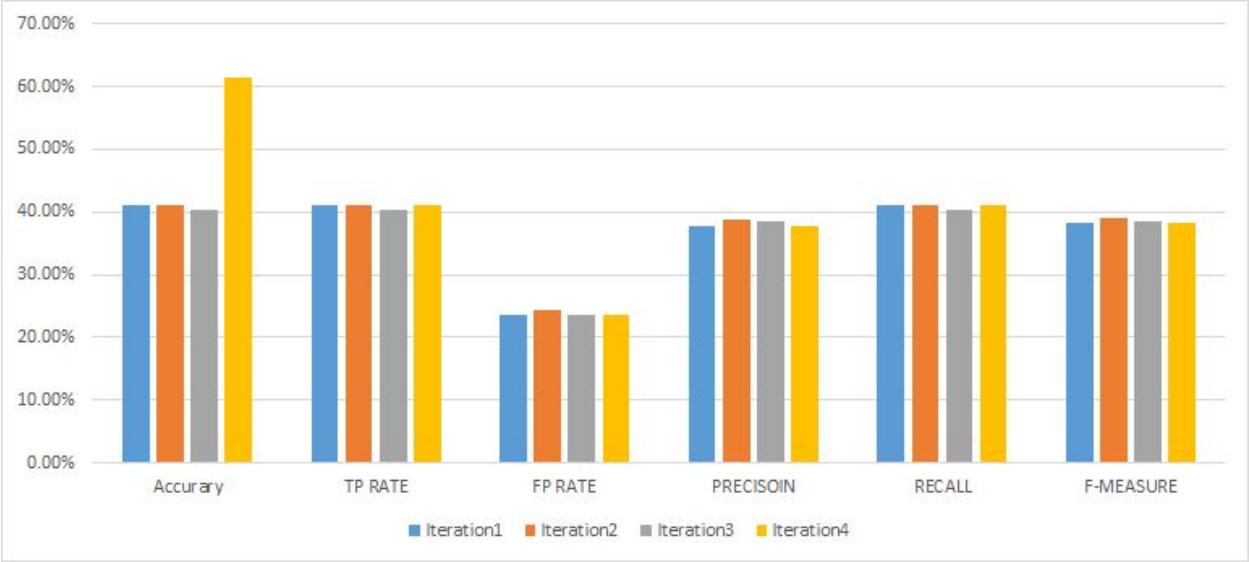


Fig 12. Comparison between All iterations

The accuracy is the higher for the iteration 4. The false positive rate is also smaller in compare to all this means that it will correctly classify the instances.

We tried to run different algorithm on the same iteration. We ran j48 decision tree and simple logistics.

Algorithm	Accuracy
Naive Bayes	61.32%
Simple logistics	59.512
Decision tree- J48	58.55

Table 5. different algorithm accuracy on dataset

Clustering:

We went little bit further and did some clustering on previous dataset. Simple K Means algorithm used with number of bins = 2. We got clusters here, we were expecting the each crime will have its own cluster but it didn't happened. Even we tried to run Kmeans for initial dataset, but each cluster had equally distributer classes.

Below is the clusters and Crime description in each cluster:

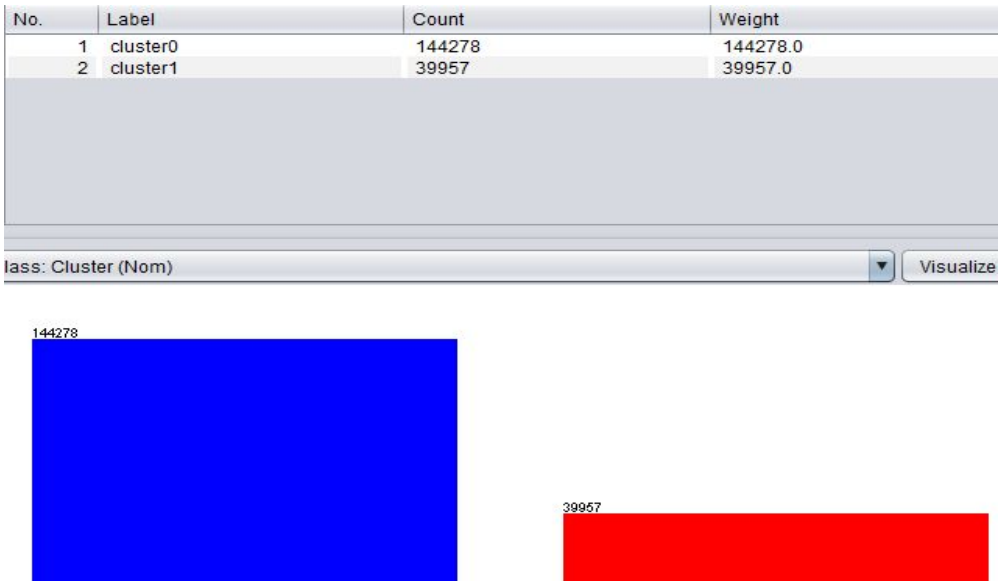


Fig 13. Clusters

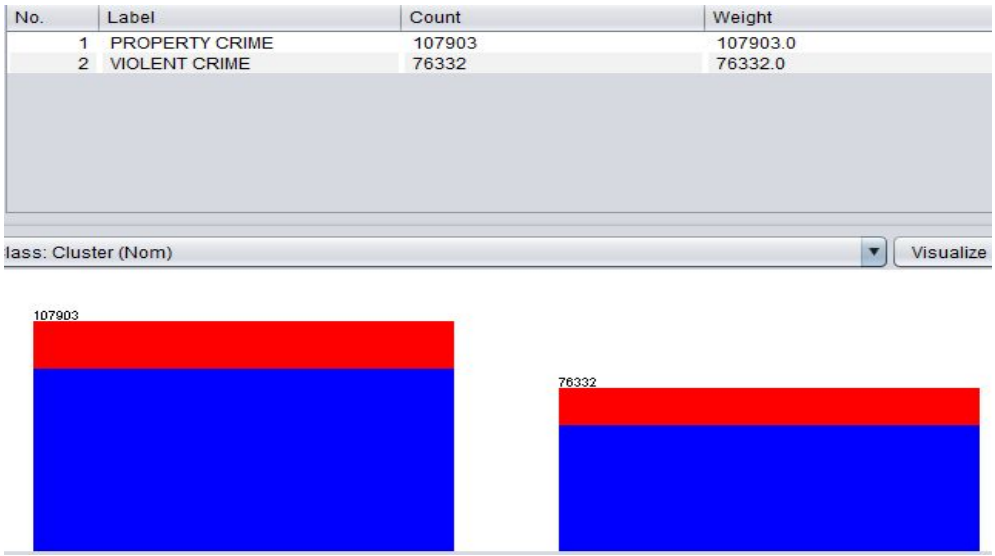


Fig 14. Crime distribution in each cluster

According the figures above the cluster 0 has both the crimes and it is all equally distributed which we are not expecting.

		Cluster#	
Attribute	Full Data	0	1
	(184235.0)	(144278.0)	(39957.0)
CrimeDate	4/27/2015	4/27/2015	4/28/2015
CrimeTime	Evening	AfterNoon	Night
Location	BELAIR RD	REISTERSTOWN RD	BELAIR RD
District	NORTHEASTERN	SOUTHEASTERN	NORTHEASTERN
Neighborhood	Downtown	Downtown	Frankford

Fig 15. Clusters with centroids

The above figure is the clusters having different centroids. This can be used as evaluation.

ANALYSIS & FINDINGS

We did some analysis on what type of crime happened the most at what time(morning, afternoon, evening or night) at t what location, district of baltimore. W used the initial dataset which is having large number of classes so that we can see what type of crime is happening exactly.

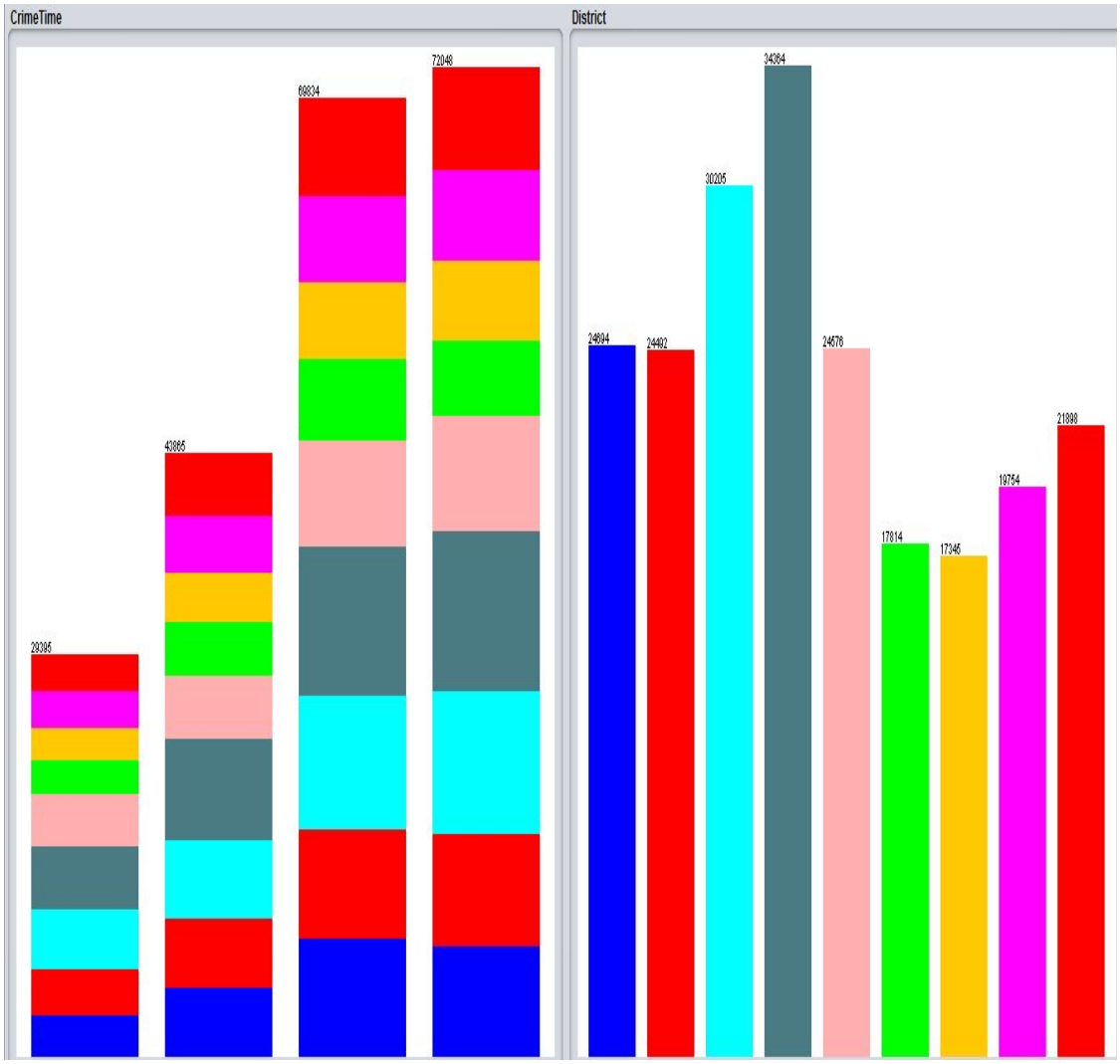


Fig 16. Crime time and Crime District bar graph

As we can see the the highest spike is of Northeastern district and evening time. So we can say that Northeast district has most number of crime. And evening time has most number of crime.

We can also see it from the numbers below:

No.	Label	Count	Weight
1	Night	29395	29395.0
2	Morning	43865	43865.0
3	AfterNoon	69834	69834.0
4	Evening	72048	72048.0

No.	Label	Count	Weight
1	CENTRAL	24694	24694.0
2	NORTHERN	24492	24492.0
3	SOUTHEASTERN	30205	30205.0
4	NORTHEASTERN	34364	34364.0
5	SOUTHERN	24576	24576.0
6	EASTERN	17814	17814.0
7	WESTERN	17345	17345.0
8	SOUTHWESTERN	19754	19754.0
9	NORTHWESTERN	21898	21898.0

Fig 17. Crime time and District crime counts

We still not sure that what type of crime is most in evening in northeast district and at what location exactly.

So we filtered based on district. And we found that the neighbourhood in which most number of crime happen are:

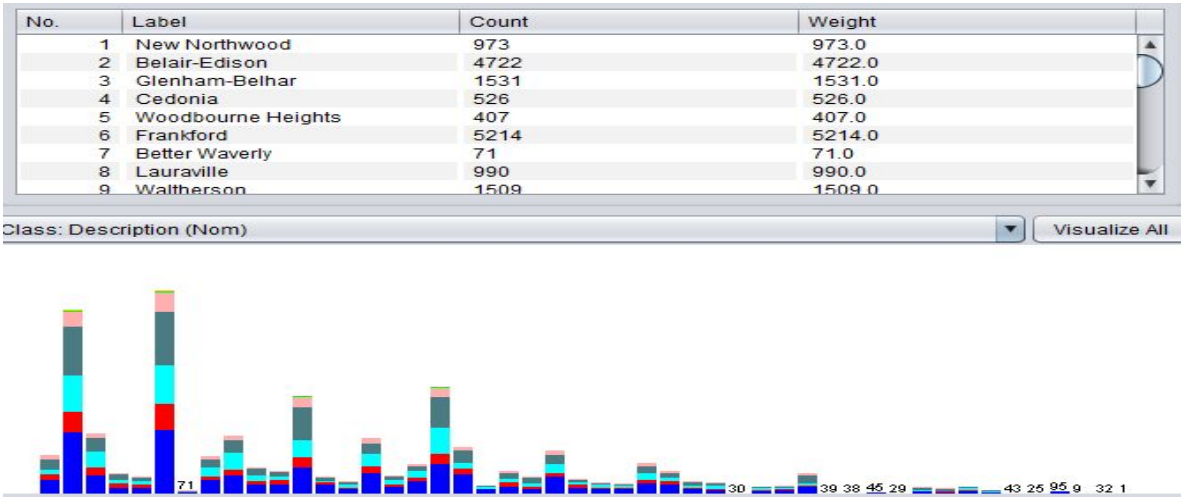
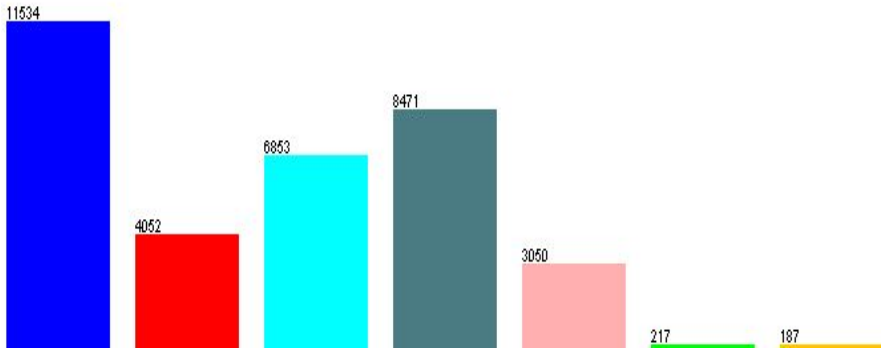


Fig 18. Neighbourhood having most crime

As we can see the FRANKFORD has most number of crime in Northeast district.

No.	Label	Count	Weight
1	LARCENY	11534	11534.0
2	AUTOTHEFT	4052	4052.0
3	BURGLARY	6853	6853.0
4	ASSAULT	8471	8471.0
5	ROBBERY	3050	3050.0
6	ARSON	217	217.0
7	RAPE	187	187.0

Class: Description (Nom) Visualize All



No.	Label	Count	Weight
1	Night	4596	4596.0
2	Morning	7354	7354.0
3	AfterNoon	10814	10814.0
4	Evening	11600	11600.0

Class: Description (Nom) Visualize All

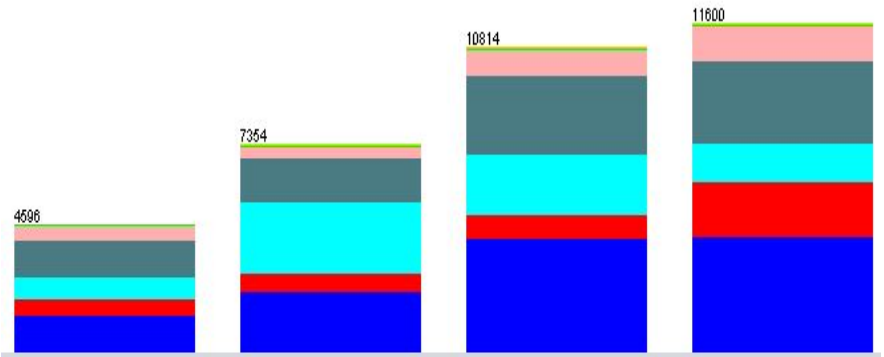


Fig 19. Type of crime and Time of crime

As from above we can see the larceny happen the most in evening time. But we went further to know at what location what crime happen the most. So we filtered from neighborhood and take out the most crime location.

No.	Label	Count	Weight
1	BELAIR RD	391	391.0
2	HAZELWOOD CR	21	21.0
3	ANNTANA AV	36	36.0
4	BLUERIDGE AV	6	6.0
5	GOODNOW RD	311	311.0
6	WHITE AV	23	23.0
7	DENWOOD AV	12	12.0
8	LASALLE AV	38	38.0
9	SEIDEL AV	15	15.0

Class: Description (Nom)

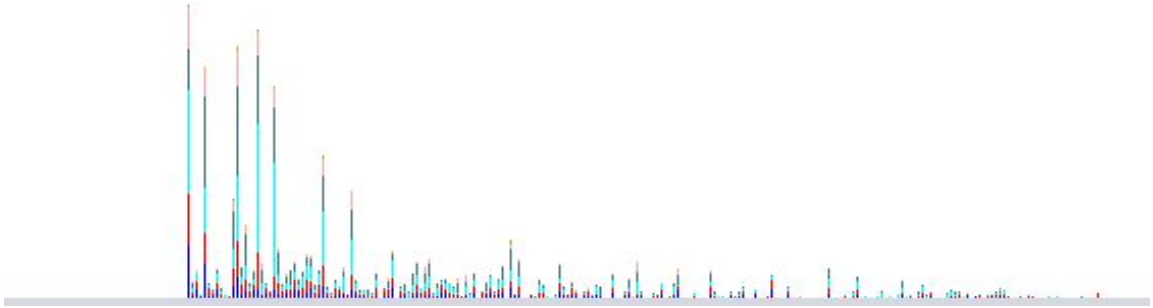
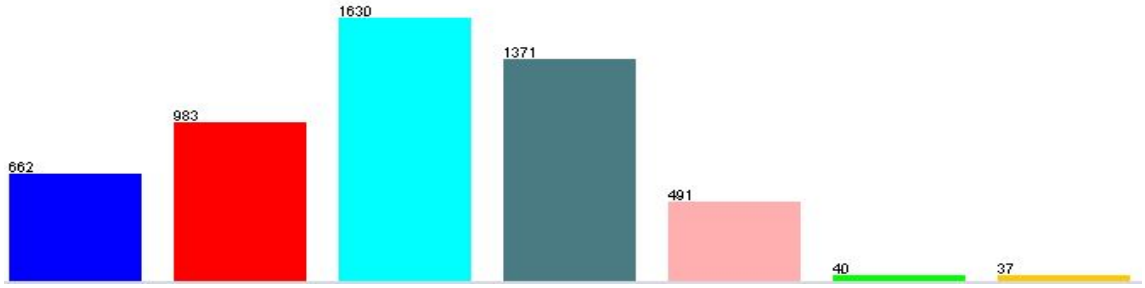


Fig 20. Most number of crime based on location

This shows that most number of crime happened is in Belair Road which is in northeast district and neighborhood is Frankford. And most number of crime at this location is Larceny.

No.	Label	Count	Weight
1	AUTOTHEFT	662	662.0
2	BURGLARY	983	983.0
3	LARCENY	1630	1630.0
4	ASSAULT	1371	1371.0
5	ROBBERY	491	491.0
6	ARSON	40	40.0
7	RAPE	37	37.0

Class: Description (Nom) Visualize All



Findings are:

District : NorthEast

Crime Type: 1) Larceny

2) Assault

3) Burglary

Crime Time: Evening

Neighborhood: Frankford

Location: Belair Road

So according to our analysis and prediction we can say that the Belair road which is in Northeast district has most number of crime happen in evening time and type of crime will be larceny and with 61% accuracy prediction.

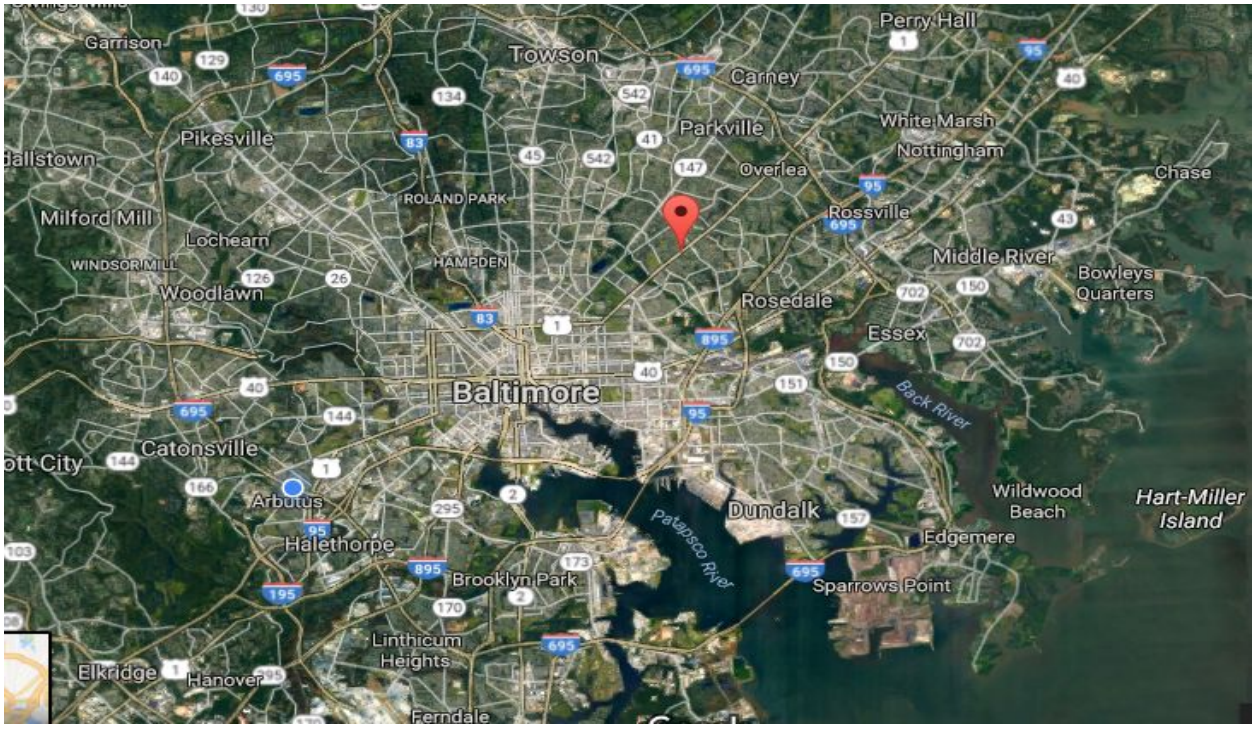


Fig 20. Most Unsafe Place(Belair Road)

map of crime density in Baltimore

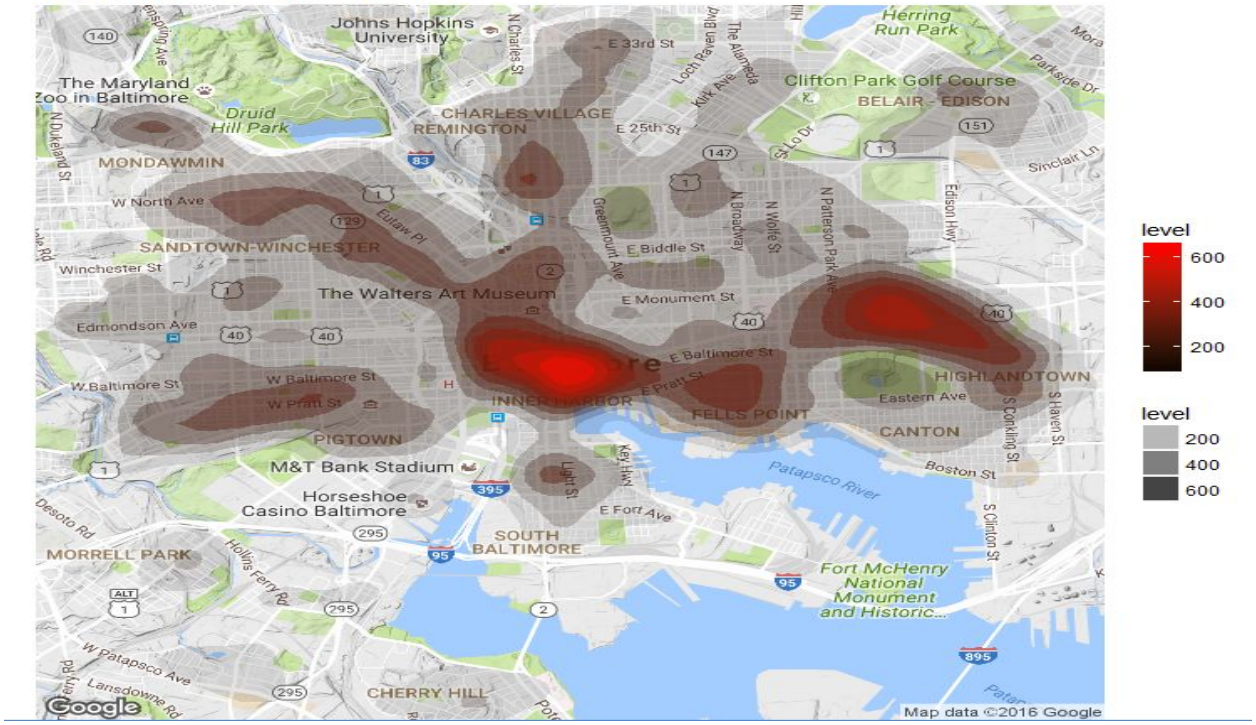


Fig 22. Hot Spots

This is the hotspots of most unsafe(crime oriented) places in Baltimore. Red area is the mostly rated crime and white are is kind of safe place.

VALIDATION

- Baltimore police has reported increment in property crimes in the period of 2015-2016
- The average crime rate of Baltimore is 662 per 100 people more than the average crime rate of US 235 per 1000 people.
- The neighborhood of Frankford contributes in increment of property crime in the city of Baltimore.
- We also found some article based on survey which also shows that belair rd has most crime rate

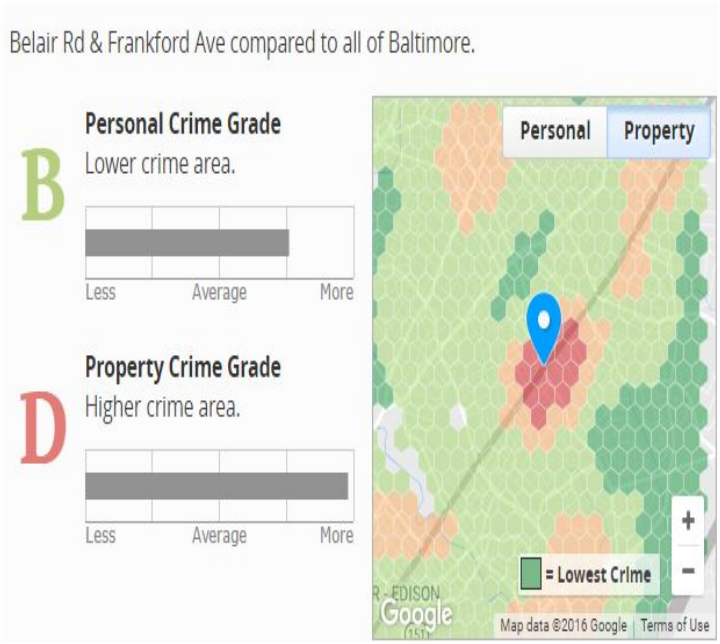


Fig 23. Crime Grades

CONCLUSION & LESSONS LEARNED

- Through the analysis the Northeast district was affected the most with high rate.
- Property crime (Larceny, Arson, theft, Burglary) is more likely to happen during Evening
- The affected area can be provided with more patrolling for safety.
- Property investments can be made in the safer areas (Western Baltimore)
- The data set is large and class distribution is scattered. We tried to check accuracy in very small training set but still didn't get the good accuracy.
- We tried several process starting from preprocessing to data mining and we have to do to

and fro every time. which is kind of lesson learned for us.

- We learned how to process the big data which is kind of spares.

REFERENCES

[1] De Bruin ,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J. and Kok,J.N(2006) Data mining approaches to criminal carrer analysis ,"in Proceedings of the Sixth International Conference on Data Mining (ICDM"06) ,Pp. 171-177

[2] Nazlena Mohamad Ali1, Masnizah Mohd2, Hyowon Lee3, Alan F. Smeaton3, Fabio Crestani4 and Shahrul Azman Mohd Noah2 ,2010 Visual Interactive Malaysia Crime News Retrieval System

[3] A.Malathi ,Dr.S.Santhosh Baboo. D.G. Vaishnav College,Chennai ,2011 Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters.

[4] Malathi.A 1 ,Dr.S.Santhosh Baboo 2 and Anbarasi . A 31 Assistant professor ,Department of Computer Science ,Govt Arts College ,Coimbatore , India . 2 Readers , Department of Computer science , D.G. Vaishnav Collge ,Chennai , India , 2011 An intelligent Analysis of a city Crime Data Using Data Mining

[5] Malathi , A; Santhosh Baboo , S, 2011 An Enhanced Algorithm to Predict a Future Crime using Data Mining

[6] https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime