NYPD NYC Crime Data:

Analysis

Sinan Can Soysal

Instructor: Dr. Keziban Seckin Codal

Business Intelligence & Data Mining: MIS402

Contents

Introduction

Problem description

Security and life safety are important aspects while making decisions regarding selection of place for residence, education or work. There are various sources available to get information about the safety of a place such as social media, native people which may or may not provide credulous information. The problem we are analysing is to find out the severity of crimes in a region which can help people in their decision making about their place of choice.

Motivation

Crimes cannot be predicted since it's not systematic or random in nature. Crimes could be rather sporadic in nature. The advancements in technologies and specialized methods of utilizing these technologies have enabled individuals in conducting crimes in various geographic areas.

Brief description of the report organization

The report is organized based on sections dealing with various aspects of the project. The first section gives a detailed introduction by providing an overview on the problem description, motivation behind the project. The second section will help in exploring the data utilized by this project, discuss various aspects of it like distribution, describe the fields/attributes. The final section details the methods devised to handle pre-processing, mining for data, delineating the performance of the model built in Weka and detailing the logic of the problem as well as the conclusion drawn from performing data mining on the dataset.

Data Exploration

(Summary of Checkpoint #2)

Source

The NYPD arrests dataset ranging from 2012-2020 years is taken from the NYC OpenData website which is provided by the Police Department (NYPD). Information in the dataset is manually extracted, reviewed by the Office of Management Analysis and Planning before being entered in the NYPD website. The dataset contains 5.15 million recorded crimes, with 18 columns each containing details of the crime. Few interesting figures from exploring the data are:

- Highest recorded crimes observed are in Kings (1334914) followed by Manhattan (1303786), Queens (895903), Bronx (1097367) and the least in Staten Island (166361).
- The Bronx had the highest crimes for dangerous drugs (1056378) as well as most crimes committed in the age group of 25-44.
- Males have significantly higher crimes (3996247) compared to females (802092).

The attributes of the dataset are (Grouped as per category for quicker understanding) are:

- (Crime identification/description) PD_CD, PD_Desc, KY_CD, OFNS_DESC, LAW_CODE, LAW_CAT_CD
- (Crime region/handle at) ARREST_BORO, ARREST_PRECINCT, JURISDICTION_CODE
- (Crime description/fine details of the perpetrator): AGE_GROUP, PERP_SEX, PERP_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude.

Link to the dataset

https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u

Methodology

Data Pre-processing (Checkpoint #3)

The first step towards data pre-processing was to trim the dataset which consisted of nearly 5.2 million records. Loading this entire dataset on Weka resulted in crashing of the file and therefore this dataset had to be reduced to a certain amount. In order to do so we separate the data based on the number of years and this was done by coding in python. Later, an approach called Simple random sampling was used. The dataset was loaded using the visual studio and a sample of records which consisted of about 31,720 records was chosen. Missing values were eliminated manually.

The dataset consisted of 18 columns from which some were eliminated as well. Columns such as Arrest key, date, KY_CD, Jurisdiction code, X and Y coordinates were removed since they did not contribute much to the classification methods that were to be implemented. We performed PCA (Principal Component Analysis) before removing these attributes to rank them in which these attributes were ranked lowest. Finally, these records were then copied onto the excel file and saved with the .csv extension and then converted into .arff in weka.

Mining the data

The algorithms that are being implemented on this project are done by using Classification and Association Rule Mining techniques. In this project the main focus is to find the outcome of the class values such as AGE_GROUP, PERP_SEX, PERP_RACE, and ARREST_BORO. These class values consist of data that will provide useful insights while developing the model and testing against it.

Classification algorithms

We have implemented several different classification algorithms on the sample dataset and have obtained the following results.

The Naive Bayes classifier is based on the Bayes theorem. The assumption made in this classifier is that features are independent, hence called naive. In our dataset, features such as OFNS_DESC, ARREST_BORO, AGE_GROUP, PERP_SEX, PERP_RACE are independent. So, we chose the Naive Bayes classifier.

J48 is a decision tree algorithm used for classification. The C4.5 algorithm is named J48 in Weka. Decision trees are based on splitting criteria in which instances are classified according to their feature values. A leaf node symbolizes a class label. J48 is well known for good accuracy and high precision, so we chose J48 for classification.

Bagging classifier is an ensemble meta-estimator which gives predictions either by voting or by averaging by introduction of randomization into its construction procedure and then making an ensemble out of it.

Lazy IBK is the K-Nearest Neighbors (KNN) algorithm. The number of nearest neighbors is the decision factor for classification. We need to choose the value of K. Closest similar points are found by using Euclidean distance. Then voting for class labels takes place. Since our dataset has only 10 features and KNN works better with a lower number of features than a large number of features, we chose this with a value of K as 1.
In all of these classification algorithms, we use default settings in weka. Accuracies of the above mentioned algorithms are mentioned in the appendix section.

Association Rule Mining using Apriori

The Apriori algorithm iteratively reduces the minimum support so that the required number of rules are obtained with the given minimum value of confidence.
This algorithm was used mainly to find the relationship between attributes to obtain frequent items sets that mainly provide the following information such as:
1. Age and Gender that dominated in committing crimes for that particular county
2. County with the most recorded criminal activity.
3. Offense committed based on race and gender

In Order to implement the Apriori algorithm on the dataset few parameters have to be taken into account. This algorithm can work only on nominal attributes. Hence, for this dataset columns such as PD_DESC, OFNS_DESC, LAW_CODE, LAW_CAT_CD, PERP_SEX, PERP_RACE, AGE_GROUP and ARREST_BORO were nominal type and the remaining attributes had to be removed. PD_DESC is a police description of crimes and it contained values similar to OFNS_DESC therefore generating rules which were of no importance and similarly LAW_CODE, LAW_CAT_CD were eliminated by using the remove option in the pre-process section. Therefore taking 5 attributes into account the rules were generated.
Following parameters were used:
The Apriori algorithm is bound by lowerBoundMinSupport of 0.1 to upperBoundMinSupport 1.0. This algorithm finds all the rules between these bounds with certain increments of delta value of 0.05.
The metricType was set to Conviction to rank the results
minMetric: 1.1 (default)
numRules: 20

Upon running the algorithm following observations were made:
Minimum support: 0.1 (3127 instances). Number of cycles performed: 18 Three large item sets were generated L(15), L(33) and L(13)

1. Fig [1] displays the best rules generated by the Apriori algorithm. From this result it is noted that a high value for conviction suggests that a consequent is highly dependent on the antecedent. Based on the rules generated it is seen that PERP_SEX being Male will imply that he was in the possession of dangerous drugs and this predominantly occurred in county Bronx B. They were also in the age range of 25-44.

2. It is also noted that White hispanics and those who committed offense such as possession of drugs were also males. The number of males that committed crimes were significantly higher than their female counterparts.
3. American Blacks were seen to have been reported for crimes such as criminal mischief and other related activities. It was noted that males from county B had mostly committed crimes.
4. Male white hispancs were reported for having dangerous weapons.
5. If the gender was male and a Black American then it is likely that he was from county K.

Models Performance

For the performance measurement of the models, various measures were considered such as correctly classified Instances, AUC i.e, Area Under Curve, ROC i.e., Receiver Operating Characteristics, Precision, Recall, Confusion Matrix. These performance measures tell us how well the applied data mining algorithm is performing on a given dataset. Since, in this project classification algorithms are used such as Naive Bayes, Lazy IBK, Bagging and J48. So, to check how many data points are correctly classified, correctly classified Instances measure in weka gives us this information. AUC is used in classification analysis in order to check which of the used classification models gives the best prediction. ROC curves is an example of its application. More the value of AUC, the better is the model. Precision refers to the percentage of the results which are relevant. Recall refers to the percentage of total relevant results which are correctly classified by data mining algorithm. Precision gives exactness of the classifier whereas Recall gives completeness of the classifier. A low value of Precision shows a large number of False Positives and a low Recall shows a large number of False Negatives. Confusion Matrix indicates expected class labels and predicted class labels. All these performance measures indicate the accuracy of the classification algorithms applied on the dataset. In this project, 10 fold cross-validation was used while training the models with approximately 31000 records. So, in this approach out of 10 samples, 9 were used for training and 1 was used for testing. Plus we used a separate test dataset comprising about 19000 records, to check the performance of the models created. Performance of models by using 10 fold cross-validation is mentioned in the above table named Training dataset accuracy and that of separate test dataset is mentioned in table named Test dataset accuracy.

Logic of the Problem

We are keen on finding areas that are prone to crime. We use the critical thinking approach for better analysis of issues. To do so we have come up with metrics and relevant data to analyze the issue at hand and we have applied critical thinking standard learning models to the problem in detail. We have used this model and have considered the point of view of end users and come up with issues that they face. We have used this approach and careful attention was given to the problems of the reader.

Pros of analytical thinking procedure:

1. Analytical Thinking leads us to abandon non-adaptive beliefs.
2. It expands the knowledge base since it requires knowing reasons and every aspect of the problem.

Cons of analytical thinking procedure:

1. It is a more time consuming procedure since it involves thinking in more depth about some problem.

   The standards that the model uses to evaluate elements in Logic of the problem, Complexity standard can be added. While evaluating elements, complexity standards play an important role. Other standards such as Clarity, Accuracy, and Precision are very much related to the Complexity of the problem.

Conclusion

For classification, J48 and bagging algorithms gave best accuracy along with good ROC and precision. Since, the dataset was skewed for PERP_SEX with more number of Male values, so in this case False Positive rate was more in all classification. Also in the case of PERP_AGE based classification, the dataset was skewed with more instances for the age group of 25-44. In this case also the False Positive rate was more.

In the case of ARREST_BORO based classification, although classification accuracy was less but got a good ROC with less False Positive rate since the dataset was balanced for each borough. For PERP_RACE based classification, we achieved on average 50 percent accuracy for each classifier with good ROC values. But got a more False Positive rate of Black class instances.

In summary, due to the skewed dataset we faced issues in classification of PERP_SEX and PERP_RACE based classification. We tried to solve these issues by doing under sampling to make the dataset balanced. After doing this we got better accuracies with less False Positive rates.

Association Rule mining provided values which gave items sets pointing towards male indicating most of the crimes were committed by this gender.

In this project, our main focus was on classification and association rule mining algorithms. In the future, we would like to implement clustering techniques to make clusters according to the severity of crimes in various areas.

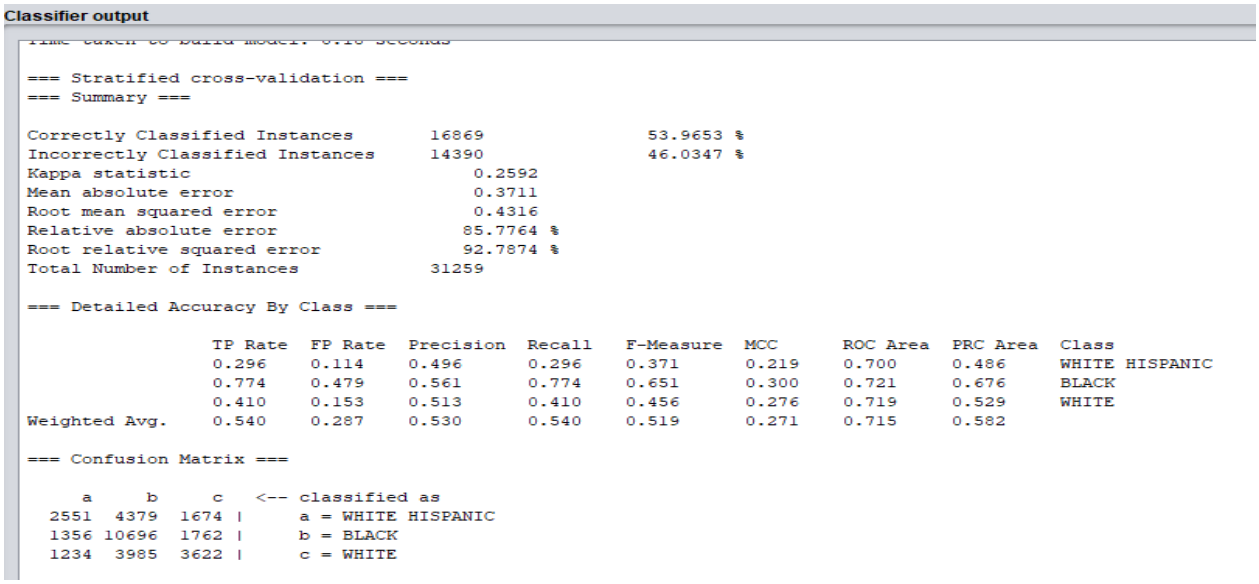
Appendix

Training dataset accuracy:

Algorithm	AGE_GROUP	PERP_SEX	PERP_RACE	ARREST_BORO
Naive Bayes	44.25%	73.99%	42.27 %	33.18%
Lazy IBK	53.94 %	82.44 %	53.71 %	41.54%
Bagging	54.09 %	82.59 %	54.22 %	42.29%
J48	54.05 %	82.47 %	53.97 %	41.98%

Test dataset accuracy:

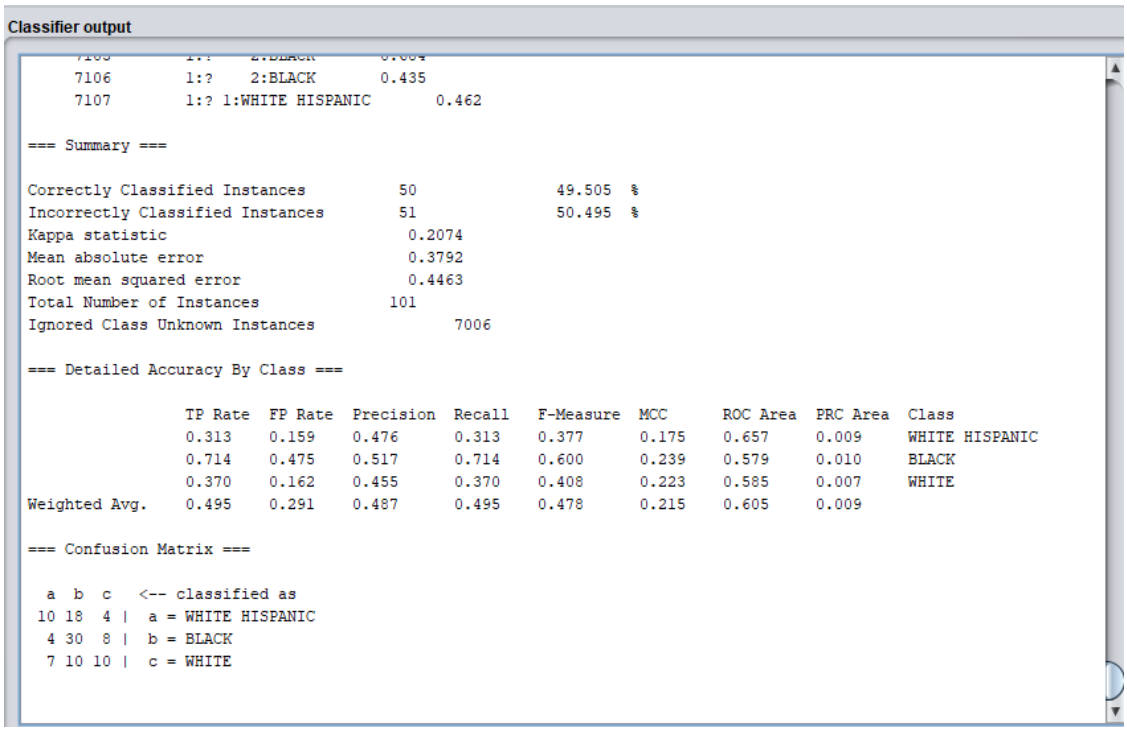
Algorithm	AGE_GROUP	PERP_SEX	PERP_RACE	ARREST_BORO
Naive Bayes	39.53 %	55.81 %	38.61 %	32.94%
Lazy IBK	54.09 %	55.81 %	51.49 %	41.18%
Bagging	47.67 %	56.98 %	46.53 %	42.35%
J48	47.67 %	56.98 %	49.51 %	38.82%



Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       16869               53.9653 %
Incorrectly Classified Instances     14390               46.0347 %
Kappa statistic                          0.2592
Mean absolute error                      0.3711
Root mean squared error                  0.4316
Relative absolute error                 85.7764 %
Root relative squared error             92.7874 %
Total Number of Instances            31259

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.296    0.114    0.496      0.296    0.371      0.219    0.700     0.486     WHITE HISPANIC
                 0.774    0.479    0.561      0.774    0.651      0.300    0.721     0.676     BLACK
                 0.410    0.153    0.513      0.410    0.456      0.276    0.719     0.529     WHITE
Weighted Avg.    0.540    0.287    0.530      0.540    0.519      0.271    0.715     0.582

=== Confusion Matrix ===

     a     b     c   <-- classified as
  2551  4379  1674 |     a = WHITE HISPANIC
  1356 10696  1762 |     b = BLACK
  1234  3985  3622 |     c = WHITE
```

Fig [1] Screenshot of the classification of Perp_Race using J48 algorithm using training dataset



Classifier output

```
     7106        1:?    2:BLACK       0.435
     7107        1:? 1:WHITE HISPANIC       0.462

=== Summary ===

Correctly Classified Instances          50               49.505  %
Incorrectly Classified Instances        51               50.495  %
Kappa statistic                          0.2074
Mean absolute error                      0.3792
Root mean squared error                  0.4463
Total Number of Instances              101
Ignored Class Unknown Instances               7006

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.313    0.159    0.476      0.313    0.377      0.175    0.657     0.009     WHITE HISPANIC
                 0.714    0.475    0.517      0.714    0.600      0.239    0.579     0.010     BLACK
                 0.370    0.162    0.455      0.370    0.408      0.223    0.585     0.007     WHITE
Weighted Avg.    0.495    0.291    0.487      0.495    0.478      0.215    0.605     0.009

=== Confusion Matrix ===

  a  b  c   <-- classified as
 10 18  4 |  a = WHITE HISPANIC
  4 30  8 |  b = BLACK
  7 10 10 |  c = WHITE
```

Fig [2] Screenshot of the classification of Perp_Race using J48 algorithm using testing dataset

```
Best rules found:

 1. OFNS_DESC=DANGEROUS WEAPONS ARREST_BORO=B 4368 ==> PERP_SEX=M 4074    conf:(0.93) lift:(1.13) lev:(0.02) [483] < conv:(2.64)>
 2. OFNS_DESC=DANGEROUS WEAPONS PERP_RACE=WHITE HISPANIC 3759 ==> PERP_SEX=M 3444    conf:(0.92) lift:(1.11) lev:(0.01) [353] < conv:(2.12)>
 3. OFNS_DESC=DANGEROUS WEAPONS AGE_GROUP=25-44 5439 ==> PERP_SEX=M 4977    conf:(0.92) lift:(1.11) lev:(0.02) [505] < conv:(2.09)>
 4. OFNS_DESC=DANGEROUS WEAPONS 11634 ==> PERP_SEX=M 10374    conf:(0.89) lift:(1.08) lev:(0.03) [810] < conv:(1.64)>
 5. OFNS_DESC=DANGEROUS WEAPONS PERP_RACE=BLACK 4599 ==> PERP_SEX=M 4074    conf:(0.89) lift:(1.08) lev:(0.01) [293] < conv:(1.56)>
 6. AGE_GROUP=25-44 ARREST_BORO=B 4515 ==> PERP_SEX=M 3906    conf:(0.87) lift:(1.05) lev:(0.01) [194] < conv:(1.32)>
 7. PERP_SEX=M ARREST_BORO=B 7791 ==> OFNS_DESC=DANGEROUS WEAPONS 4074    conf:(0.52) lift:(1.41) lev:(0.04) [1175] < conv:(1.32)>
 8. ARREST_BORO=K 7581 ==> PERP_RACE=BLACK 4179    conf:(0.55) lift:(1.25) lev:(0.03) [828] < conv:(1.24)>
 9. PERP_RACE=WHITE HISPANIC AGE_GROUP=25-44 4242 ==> PERP_SEX=M 3633    conf:(0.86) lift:(1.04) lev:(0) [145] < conv:(1.24)>
10. PERP_SEX=M PERP_RACE=WHITE HISPANIC 7182 ==> OFNS_DESC=DANGEROUS WEAPONS 3444    conf:(0.48) lift:(1.29) lev:(0.02) [771] < conv:(1.21)>
11. PERP_SEX=M ARREST_BORO=K 6258 ==> PERP_RACE=BLACK 3360    conf:(0.54) lift:(1.21) lev:(0.02) [594] < conv:(1.2)>
12. ARREST_BORO=B 9345 ==> OFNS_DESC=DANGEROUS WEAPONS PERP_SEX=M 4074    conf:(0.44) lift:(1.31) lev:(0.03) [973] < conv:(1.18)>
13. ARREST_BORO=B 9345 ==> OFNS_DESC=DANGEROUS WEAPONS 4368    conf:(0.47) lift:(1.26) lev:(0.03) [891] < conv:(1.18)>
14. ARREST_BORO=K 7581 ==> PERP_SEX=M PERP_RACE=BLACK 3360    conf:(0.44) lift:(1.25) lev:(0.02) [671] < conv:(1.16)>
15. OFNS_DESC=DANGEROUS WEAPONS 11634 ==> PERP_SEX=M ARREST_BORO=B 4074    conf:(0.35) lift:(1.41) lev:(0.04) [1175] < conv:(1.16)>
16. OFNS_DESC=DANGEROUS WEAPONS PERP_SEX=M 10374 ==> ARREST_BORO=B 4074    conf:(0.39) lift:(1.31) lev:(0.03) [973] < conv:(1.15)>
17. OFNS_DESC=DANGEROUS WEAPONS 11634 ==> ARREST_BORO=B 4368    conf:(0.38) lift:(1.26) lev:(0.03) [891] < conv:(1.12)>
18. OFNS_DESC=CRIMINAL MISCHIEF & RELATED OF 8631 ==> PERP_RACE=BLACK 4326    conf:(0.5) lift:(1.13) lev:(0.02) [511] < conv:(1.12)>
19. PERP_RACE=BLACK ARREST_BORO=B 4095 ==> PERP_SEX=M 3444    conf:(0.84) lift:(1.02) lev:(0) [77] < conv:(1.12)>
20. PERP_RACE=WHITE 8841 ==> PERP_SEX=M 7434    conf:(0.84) lift:(1.02) lev:(0.01) [166] < conv:(1.12)>
```

Fig [3] Screenshot of the best rules using Apriori algorithm

References

[1] C. Yu, M. W. Ward, M. Morabito and W. Ding, "Crime Forecasting Using Data Mining Techniques," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, 2011, pp. 779-786.

[2] K. C. Lekha and S. Prakasam, "Data mining techniques in detecting and predicting cyber crimes in banking sector," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 1639-1643.

[3] Bansal, D., & Bhambhu, L. (2013). Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women.