Crime Interval Prediction and Classification using Logistic Regression, Random Forest & C5 Algorithm for Denver Crime Data Set

Laveena D'Costa [#1], Ashoka Wilson D'Souza [$2], Prajwal K Augustine [*3], Dillon Arwin Ashwal[@4]

[#1]MCA, St Aloysius Institute of Management and Information Technology, India
[$2] Senior business intelligence analyst Atlantic Data Bureau Service Mangalore, India
*[*3]MSc, St Aloysius Institute of Management and Information Technology, India*
[@4] MSc, St Aloysius Institute of Management and Information Technology, India

[#1]laveenacrasta@staloysius.ac.in
[$2]ashokdesouza@gmail.com
*[*3] prajwalaugustine97@gmail.com*
[@4]dillonarwinashwal@gmail.com

Abstract**—** The crime time interval for the occurrence of a type of crime can be predicted using different classification models using various features like crime type, crime category, district ID and more. In this paper we have used Logistic regression, Random Forest and C5.0 classification model to predict the crime time interval.

Keywords: Logistic Regression, Crime Interval, Random forest

I. Introduction

Crime rate has seen a drastic increase in the past couple of years making it hard for law enforcement to keep track of them. But with the advancements in statistics and technology we can now look for crime patterns so that we can be ready for any future incidents. There is a deep relation between crime type and time of occurrence. Certain crime types happen only during the night while others don't really have a fixed time interval. Some crime types take time to get reported from its first occurrence while others are reported immediately. Since we know that crime and time has a unique connection with each other we can look for ways to exploit this and fight crime before it even occurs.

II. Literature Review

In their paper, the authors Tahani Almanie et al[4] have taken the Denver and Los Angeles crime dataset and applied Apriori algorithm on both the datasets. Using Apriori algorithm they found that Denver has 62 interesting frequent patterns while Los Angeles has 59 patterns. It was clear that Five-Point, Capitol Hill, CBD, Montebello, Union Station, Stapleton, and Westwood are the hotspots that have most crimes frequent patterns in Denver. Additionally, they found that Wednesday was the peak day of crimes occurred in CBD while Union Station has frequent patterns only on weekend days i.e. four hours before and after midnight. They implemented the algorithm on location and time features and excluded the crime type feature to obtain frequent patterns. The authors have also applied Multinomial Naïve Bayes and Decision tree classifiers to predict the crime type in each of the two cities. The features used for these two classifiers are month, day, time and location. They achieved highest accuracy of 51% on Denver dataset and 54% on Los Angeles dataset.

Mustafa Gök and Mehmet Sait performed classification algorithms such as Naïve Bayes and Decision tree to find the most probable criminal of a particular offense incident when the suspected list of lawbreakers are provided with the criminal data which is generated unnaturally using Gaussian Mixture Model. The authors compared the accuracy of two algorithms and concluded that the Naïve Bayes Classifier consumed less execution time and performed better with 78.05% accuracy. [1]

The author P. Monish has used J48 Decision tree classifier which gives a higher accuracy compared to most of his other models. To classify a new object, J48 algorithm initially creates a decision tree based the attribute values available in the training dataset. The J48 model has an accuracy of 97.59%.[3]

A.Agarwal et al[2] analyzed various offenses done by offenders and predict the chance of each offense that can again be performed by that offender. The authors have used Apriori algorithm for frequent item set generation that can be done by the offenders.

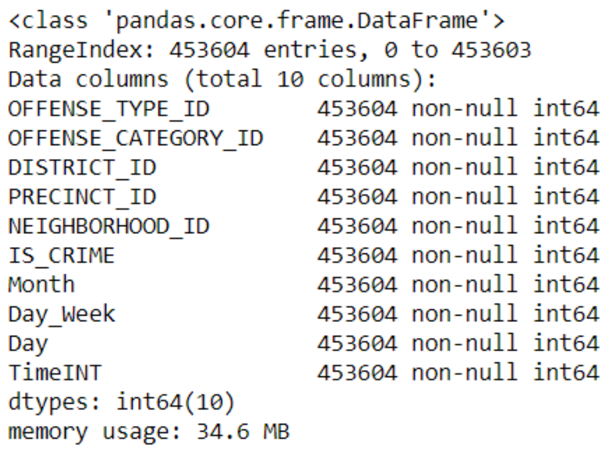
III. Data and Empirical Analysis

A) Denver Crimes Dataset

This dataset consists of the real-world crimes in Denver, Colorado. It includes criminal offenses and crime incidents in the city for the years 2014 to 2019.

B) Procedure and Data Analysis

The data had issues with the date and time as some of the dates were in 24-hour format while others were in the 12-hour format. This issue was corrected using Excel and all the data was converted to a common 24-hour format. The features used for model are listed in the table T1.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 453604 entries, 0 to 453603
Data columns (total 10 columns):
OFFENSE_TYPE_ID        453604 non-null int64
OFFENSE_CATEGORY_ID    453604 non-null int64
DISTRICT_ID            453604 non-null int64
PRECINCT_ID            453604 non-null int64
NEIGHBORHOOD_ID        453604 non-null int64
IS_CRIME               453604 non-null int64
Month                  453604 non-null int64
Day_Week               453604 non-null int64
Day                    453604 non-null int64
TimeINT                453604 non-null int64
dtypes: int64(10)
memory usage: 34.6 MB
```

T1: Features

TimeINT is the Binary Class Variable which is of two intervals. Interval 9am to 8:59pm and Interval 9 pm to 8 59 am. The SMOTE method of python's imblearn package was used to balance the classes.

Logistic Regression

Logistic regression the method used for binary classification problems. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth. It's curve that can take any value between 0 and 1, but never exactly at those limits.

Random Forest Classifier

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then collects the votes from different decision trees to pick the final class of the test object.
Basic factors to Random Forest Classifier can be total number of trees to be generated and decision tree related parameters like minimum split, split criteria etc.

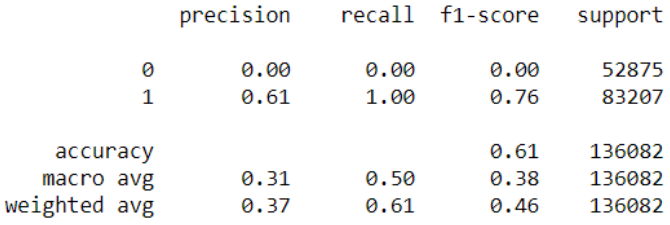
C5.0 classification model

C5.0 algorithm is used to build either a decision tree or a rule set. A C5.0 model works by dividing the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then divided again, usually based on a different field, and the process repeats until the subsamples cannot be divided any further. Finally, the lowest-level divisions are analysed, and those that are not significant are removed.

IV. Results

Logistic Regression

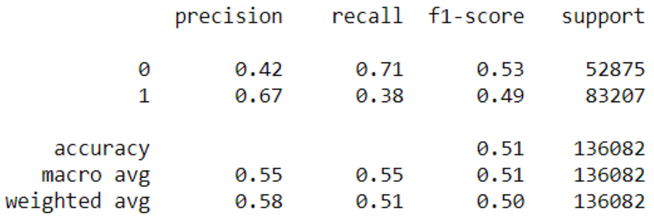
The classification report of the fitted model before balancing the class variable is given in the table T2.



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00     52875
           1       0.61      1.00      0.76     83207

    accuracy                           0.61    136082
   macro avg       0.31      0.50      0.38    136082
weighted avg       0.37      0.61      0.46    136082
```

T2: Classification report before balancing

Even though the model's accuracy is fairly OK there is a lot of misclassification as one of the classes has more records compared to the other. After resolving the class imbalance, the model's performance measures are listed in the table T3. This decreases the accuracy measure which is due to balancing performed on the class variable. The Logistic Regression gives a fairly Decent Accuracy of 51%

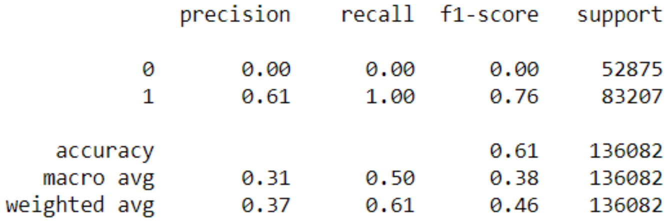


```
              precision    recall  f1-score   support

           0       0.42      0.71      0.53     52875
           1       0.67      0.38      0.49     83207

    accuracy                           0.51    136082
   macro avg       0.55      0.55      0.51    136082
weighted avg       0.58      0.51      0.50    136082
```

T3: Classification report After balancing

Random Forest Classifier

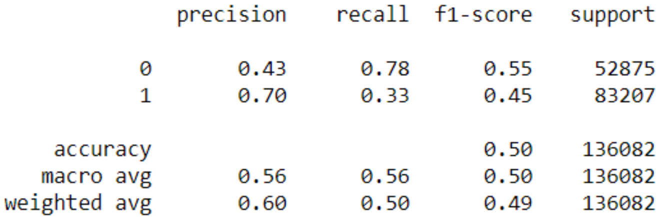
The classification report of the fitted model before balancing the class variable is given in the table T4.



```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00     52875
           1       0.61      1.00      0.76     83207

    accuracy                           0.61    136082
   macro avg       0.31      0.50      0.38    136082
weighted avg       0.37      0.61      0.46    136082
```

T4: Classification report before balancing

After resolving the class imbalance, the model's performance measures are listed in the table T5. The Accuracy of the Random Forest Classifier decreases from 61% to 50% but the amount of misclassification also decreases.

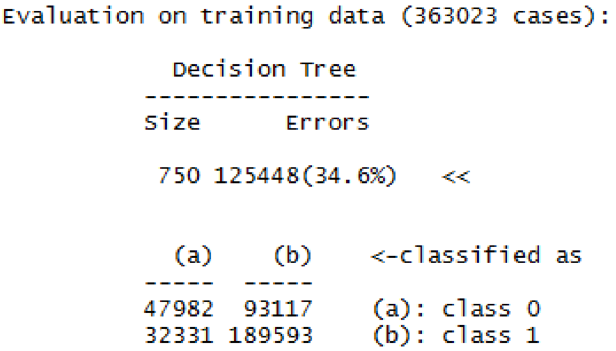


```
              precision    recall  f1-score   support

           0       0.43      0.78      0.55     52875
           1       0.70      0.33      0.45     83207

    accuracy                           0.50    136082
   macro avg       0.56      0.56      0.50    136082
weighted avg       0.60      0.50      0.49    136082
```

T5: Classification report After balancing

C5.0 classification model

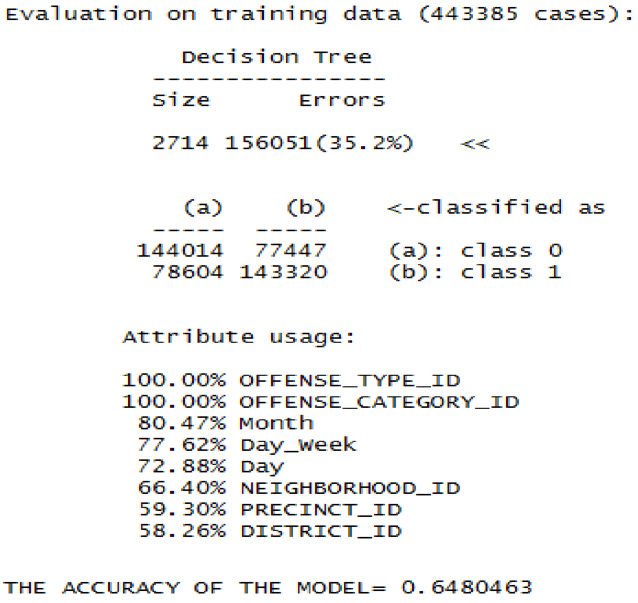
The classification report of the fitted model before balancing the class variable is given in the table T6. The model was fitted using R.



```
Evaluation on training data (363023 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	   750 125448(34.6%)   <<


	   (a)   (b)    <-classified as
	  ----- -----
	  47982  93117    (a): class 0
	  32331 189593    (b): class 1
```

T6: Classification report before balancing

After resolving the class imbalance, the model's performance measures are listed in the table T7. The Accuracy of the C5.0 classification model remains close to 65% even after balancing. The accuracy and classification are better when compared to the other two models

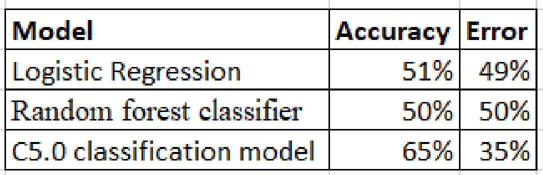


```
Evaluation on training data (443385 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	  2714 156051(35.2%)   <<


	     (a)    (b)    <-classified as
	    -----  -----
	   144014  77447    (a): class 0
	    78604 143320    (b): class 1


	Attribute usage:

	100.00% OFFENSE_TYPE_ID
	100.00% OFFENSE_CATEGORY_ID
	 80.47% Month
	 77.62% Day_Week
	 72.88% Day
	 66.40% NEIGHBORHOOD_ID
	 59.30% PRECINCT_ID
	 58.26% DISTRICT_ID

THE ACCURACY OF THE MODEL= 0.6480463
```

T7: Classification report After balancing

V. CONCLUSION

The table below gives the accuracy measures of the various models implemented in this paper. Based on this we conclude that C5.0 classification model is a better fit for predicting the time interval for the occurrence of a crime. The model gave an accuracy of 65%.



Model	Accuracy	Error
Logistic Regression	51%	49%
Random forest classifier	50%	50%
C5.0 classification model	65%	35%

T8: Models

VI. REFERENCES

[1] Mehmet Sait, and Mustafa Gök. "Criminal prediction using Naive Bayes theory." Springer 28.9 (2016): 2581-2592.

[2] A.Agarwal, D. Chougule, and D. Chimote. Application for analysis and prediction of crime data using data mining. *International Journal of Advanced Computational Engineering and Networking*,(2016)

[3] P. Monish, K. R. Ranjith , G. Varun, S. Sridhar,"CRIME ANALYSIS: PROBABILISTIC PREDICTION",

[4] *Tahani Almanie, Rsha Mirza and Elizabeth Lor.*, Crime prediction based on crime type and using spatial and temporal criminal hotspots.(2015)