Pattern Analysis and Predictive Policing Using Socio-Economic Factors

Trishla Mishra, Oklahoma State University;

ABSTRACT

According to a recent survey, crime rate in USA has gone down by 2% in the country. Police Department in some of the largest cities such as Los Angeles have started using crime-prediction techniques which is partly the reason why crime rates have declined. Currently, these departments use information related to crime such as day/time of occurrence, premise etc. for prediction. This paper aims at inducing socio-economic factors such as total population, median income, median age, unemployment etc. to check their influence and predict crimes based on the same. To do so, LosAngeles crime data was collected from catalog.data.gov. It was merged with the socio-economic data obtained from American FactFinder. Tableau Public 18.3 was used to explore the dataset. With initial exploration, it was found that the most frequently occurring crime is battery – simple assault. Also, 77th street turned out to be most unsafe location. Other algorithms such as Apriori were used to find patterns between crime and the day/time of occurrence of crime and logistic regression and advanced predictive models such as decision tree, gradient boosting and random forests were created in SAS® Enterprise Miner™ 14 for multi-class classification of crime and the best performing model was chosen to predict crime to predict the same.

KEYWORD

Predictive policing, crime prediction, pattern finding, socio-economic factors

INTRODUCTION

We live in a world that is driven by Artificial Intelligence that answers all our questions given we have the right kind of data for it. Now, its utility has extended to almost all aspects of the society which also covers predictive policing. According to a latimes.com, the LA Police Department took a revolutionary leap in 2010 when it started implementing predictive policing to curb crimes. Eventually, the police departments of the other parts of the country joined hands and started doing the same. This methodology helped them create location-based strategies, and predict who were most likely to become criminals in the next 12 hours. However, the strategy was soon dumped by the department as they realized they have insufficient data (data restricted to location of crime, time the crime happened, victim's demographics etc.). This paper incorporates socio- economic factors such as population of the area, median household-income, employment rate etc. to see how accurately crime can be predicted.

The data for this paper was collected from datagov.in. This dataset has historical data of crime that have occurred in Los Angeles from 2010-2019. Socio-economic factors such as were collected from American FactFinder.

The main areas of research for this paper are listed below

1) Identifying if there is a pattern among crimes, day of the week and location

2) Predicting the category of crime given certain location, day of the week, time of the day, and other socio-economic parameters of the location.

Python 3.6 was used to prepare the dataset and find the underlying pattern. Tableau Public 18.3 and MS Excel 2016 were used for visualizing the dataset. The result of the predictive model is explained using the decision tree model created in SAS® Enterprise Miner™ 14.

RATIONALE

Currently, most of the police departments use predictive policing techniques that aim at predicting potential criminals rather than hot spots of crimes. Such techniques have been allegedly biased against certain section of people and lead to unwarranted atrocious behavior towards them. The technology is at the center of a heated debate about their implications for civil liberties [1]. With this research paper, the aim is to predict criminal hot-spots rather than predicting who is a potential criminal. The objective is to improve the accuracy of predictions and to do so, apart from the historic crime data, socio-economic parameters such as median income, population, literacy rate, proportion of population earning etc. have been joined on zip-code level. In other words, this can also be called RTM (Risk Terrain Modeling) by using socio-economic data.

By applying this technique, any police department can make predictions on which location is going to be a hot spot for what kind of crime based on prior knowledge about the location. This technique can help in developing an unbiased and ethical approach towards predictive policing, thereby leading to a growth in respect for people's rights [2].

DATA COLLECTION AND PREPARATION

1. Data Preparation

The data set contained 1,048,576 records. For better predictions and faster processing of algorithms, data was filtered to contain records of crimes from 2017 to 2019. The total no. of records used are 465,678 records. The co-ordinates in the dataset were reverse geo-coded using geocodio. After obtaining the corresponding zip codes, socio-economic data was obtained from American FactFinder and the data sets were merged.

2. Variable Reduction

All the redundant and leakage variables were removed for more practical prediction. For example, the status code of the crime is a leakage variable that cannot be used for modeling. Similarly, a derived variable, difference in days between when crime occurred and when it was reported was also dropped. Among the redundant variables, it was observed that percent population in labor force and percent population above 16 in labor force also had high correlation. The latter was dropped to deal with redundancy.

3. Variable Creation

Certain new variables were derived to improve the performance of the model and remove the heavily skewed numeric variables. The following gives the list of new variables that were generated from the older ones.

- Percent_educated = $\frac{\boldsymbol{SomeCollege/AssociateDgree(25-64)+BachelorOrHigherDgree(25-64)}}{\boldsymbol{TotalPopulation}}$
- Percent_Male = $\frac{\boldsymbol{MalePopulation}}{\boldsymbol{TotalPopulation}}$
- Percent_Female = $\frac{\boldsymbol{FealePopulation}}{\boldsymbol{TotalPopulation}}$

Time_span, season, day of the week and month were derived from the date of occurrence of crime. The following table contains information on all the variables that were used for the project. This table contains the derived variables as well as the socio-economic information joined on the zip-code level.

Table 1. Variables used for modeling and their description

Variable	Description
Area Name	Area where crime occurred
Day of week	Day of the week when crime occurred
Premise Description	Premise of crime
Time Span	Divided into 6 time-spans (12.00 AM-4.00 AM, 4.00 AM-8.00 AM, 8.00 AM-12.00PM, 12.00PM-4.00PM, 4.00PM-8.00PM, 8.00PM-12.00AM)
Season occurred	Season when the crime occurred
College/Associate Degree	Population in the area with some college/ associate degree
BachelorOrHigherDgree	Population in the area with bachelor/ higher degree
PercentInLaborForce	Percent population above 16 and in labor force
PercentOfPeopleBPL	Percent population below poverty line
HouseholdUnits	Household units in the area where crime occurred
MedianAge	Median age in the location
MedianHouseholdIncome($)	Median household income for the area
TotalPopulation	Population in the area
Percent_Male	Percent population male
Percent_Female	Percent population female
Percent_Pop_Above_16	Percent population above 16
Crime	Type of crime (Theft, Assault/Burglary, Vandalism and Other Crimes)
percent_educated	Percent population above 25 and educated

DATA EXPLORATION

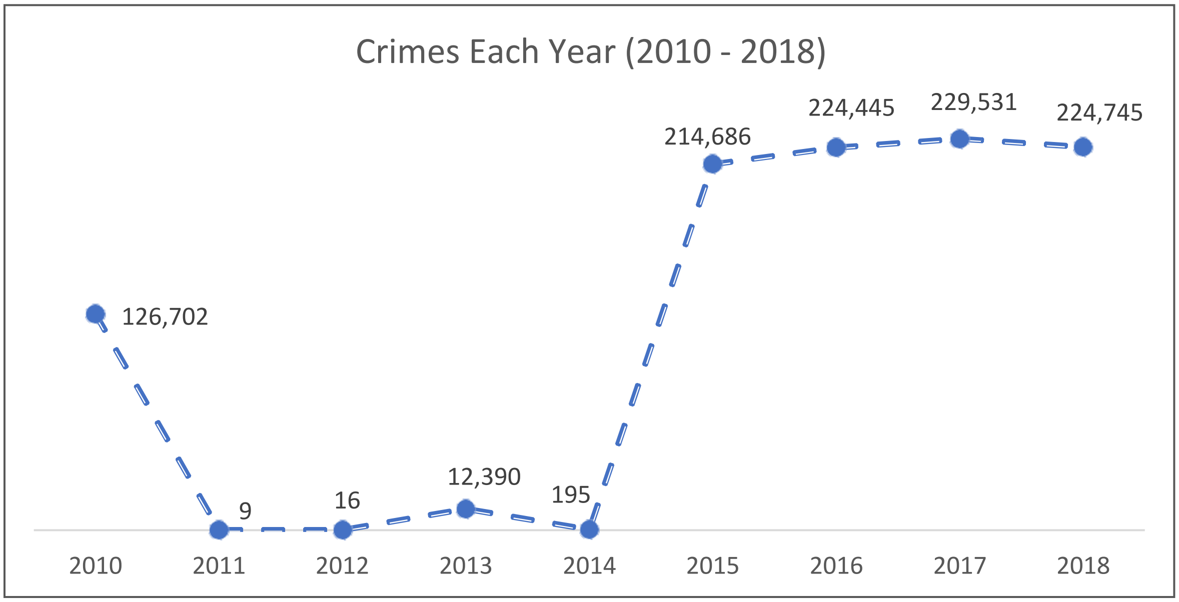


Figure 1. Trend chart for crimes each year

Figure 1 represents the total no. of crimes that occurred each year from 2010 to 2018. As can be seen from the graph, there has been an increasing trend in the total no. of crimes. Since there wasn't enough data from 2011-2014, the total no. of crimes is very low for these 3 years.

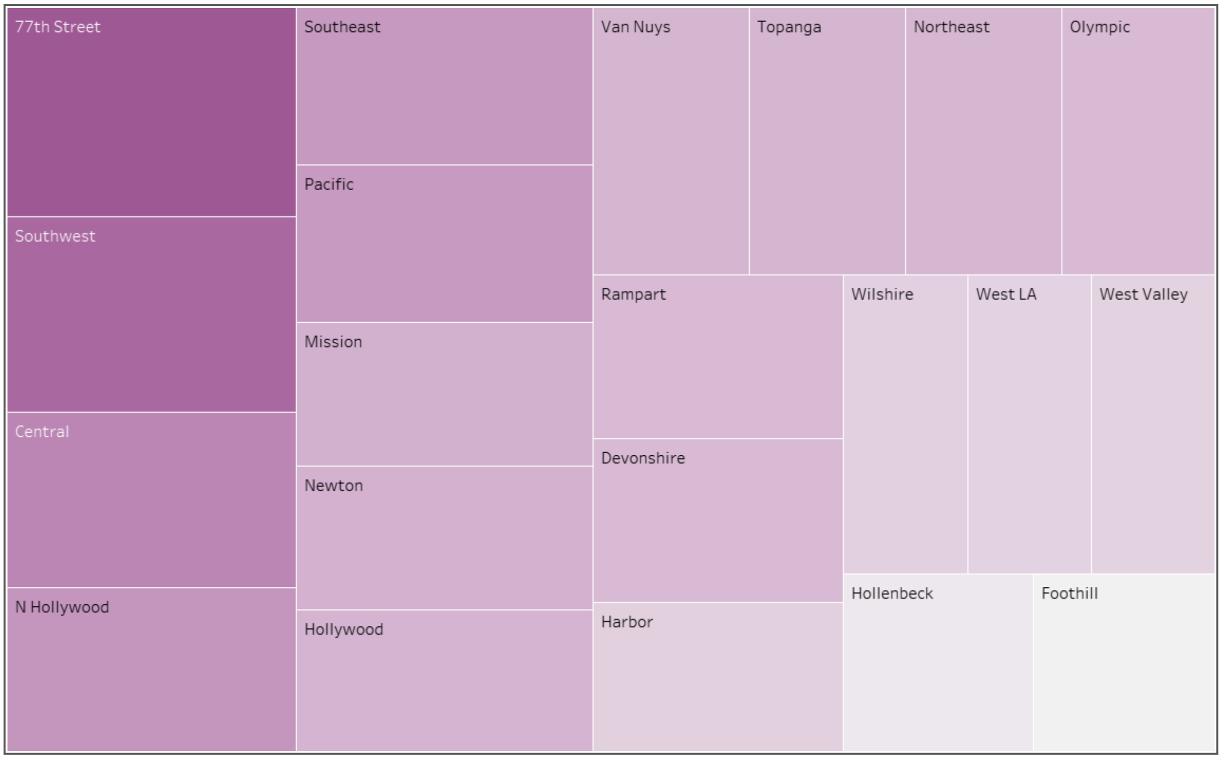


Figure 2. Heatmap to find frequency of crimes in each location

Figure 2 represents the locations where most crimes have occurred from 2010 to 2019. 77th street is the most crime-attractive location followed by Southwest, Central and North Hollywood.

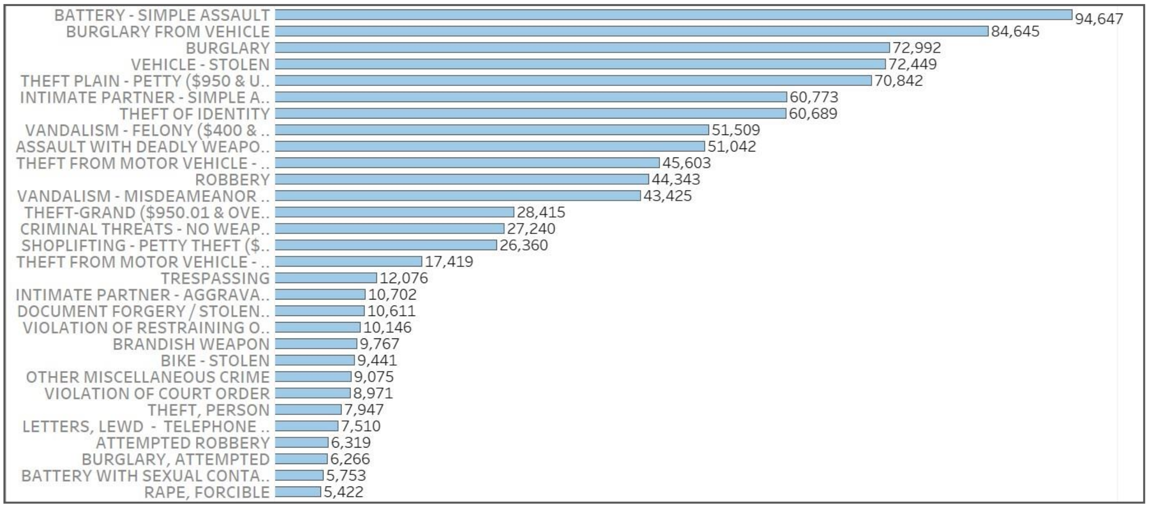


Figure 3. Frequency distribution of crimes

Figure 3 shows the frequency of each crime type. The most common type of crime is battery-simple assault followed by burglary from vehicle and stealing of vehicle.

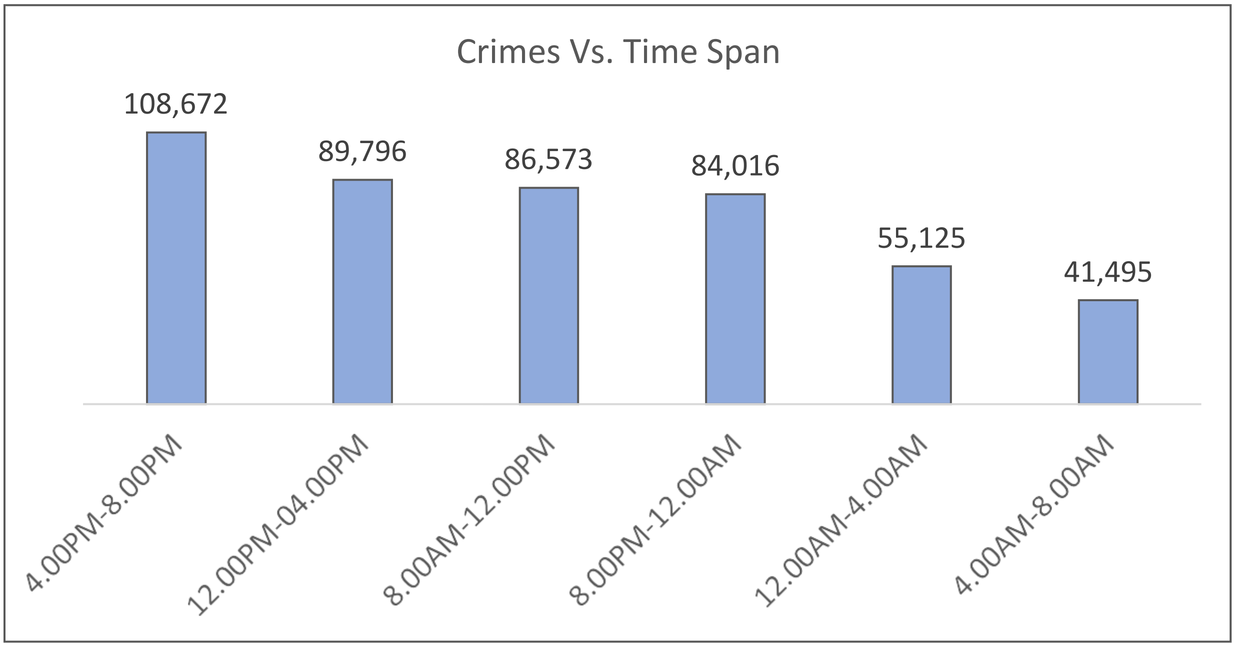


Figure 4. Frequency of crimes for each time span

Figure 4 shows the frequency of crimes across each time span. From the figure, it is evident that maximum crimes occur between 4.00 PM and 8.00 PM in the evening. The early morning hours (4 AM to 8 AM) has the fewest crimes happening.

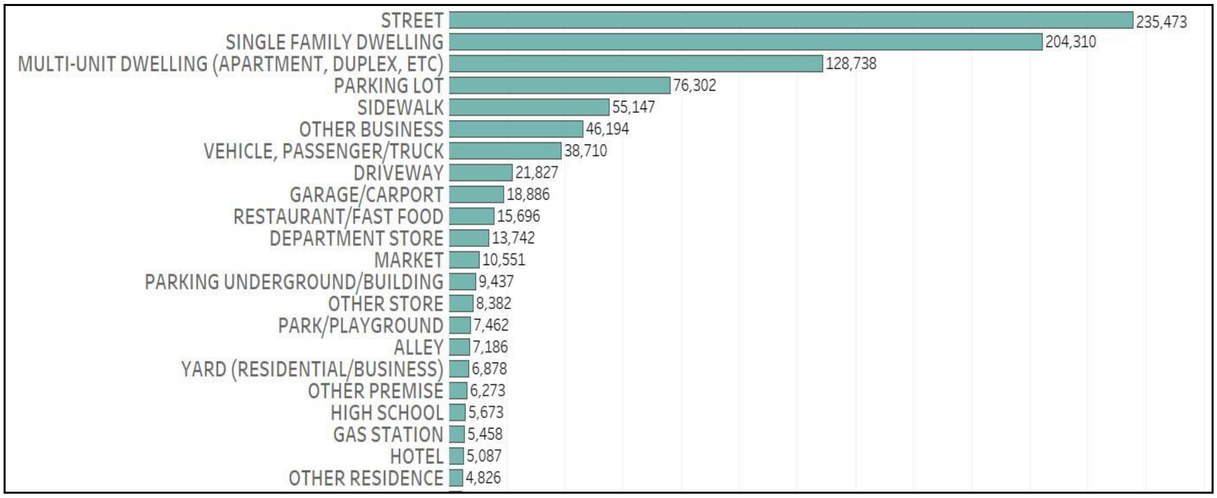


Figure 5. Frequency of crimes for each premise

From figure 5 above, it can be seen that most crimes occur in the street followed by single family dwelling, multi-unit dwelling and parking lot. Based on this distribution, the variable ‘premise description’ was categorized into 5 distinct levels. Figure 6 shows that most of the crimes have happened on Fridays.

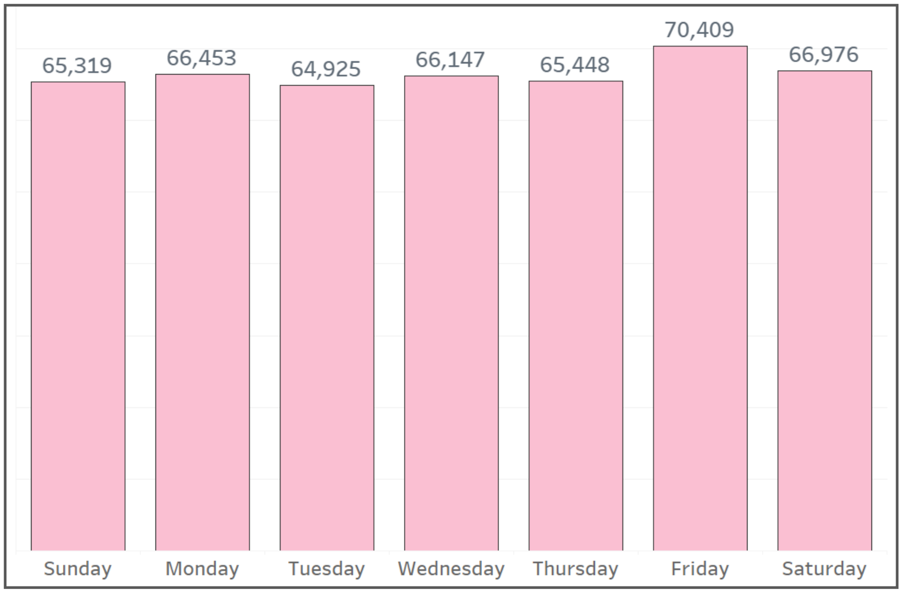


Figure 6. Frequency of crimes for each day of the week from 2017 to 2019

PATTERN FINDING USING APRIORI ALGORITHM

Apriori algorithm was used to accomplish the first objective of the research and that is mining patterns and crime hot spots. It is one of the most basic algorithms for identifying pattern and helped to come up with a list of all crime hotspots irrespective of the committed crime type [3].

This model was implemented using Python 3.6. The dataset was filtered to have only 3 variables and they are crime type, location where the crime occurred and day of the week

A cutoff was provided for the value of lift. The selected cutoff value is 1.01 which means any pattern that has a lift of greater than 1.01 was used as the result of this algorithm. A cutoff value for lift was chosen as only confidence does not take care of the frequency of occurrence of the consequent. Hence, lift tells us if the pattern has occurred by chance or there is a significant association.
Lift overcomes the challenge posed by confidence. Any value greater than 1 vouches for high association between the antecedent and consequent. The following table gives the list of all the associations whose lift value was greater than 1.

Table 2. Table shows the association rules between Crime, area and day of week

Crime	Area Name	Day of Week	Support	Confidence	Lift
ASSAULT	77th Street	Sunday	0.00303	0.32415	1.55135
THEFT	N Hollywood	Friday	0.00344	0.42279	1.17246
THEFT	Pacific	Friday	0.00377	0.46020	1.27619
THEFT	Southwest	Friday	0.00354	0.37435	1.03813
THEFT	Topanga	Friday	0.00319	0.45885	1.27246
THEFT	N Hollywood	Monday	0.00329	0.43127	1.19599
THEFT	Pacific	Monday	0.00343	0.46645	1.29353
THEFT	Topanga	Monday	0.00305	0.45790	1.26983
THEFT	N Hollywood	Thursday	0.00303	0.40696	1.12856
THEFT	N Hollywood	Tuesday	0.00309	0.41592	1.15342
THEFT	N Hollywood	Wednesday	0.00309	0.40517	1.12359
THEFT	Pacific	Saturday	0.00332	0.43200	1.19800
THEFT	Pacific	Sunday	0.00310	0.41932	1.16284
THEFT	Pacific	Thursday	0.00342	0.46422	1.28735
THEFT	Pacific	Tuesday	0.00337	0.44964	1.24692
THEFT	Pacific	Wednesday	0.00339	0.45814	1.27048
THEFT	Southwest	Thursday	0.00317	0.36651	1.01638
THEFT	Topanga	Thursday	0.00307	0.46222	1.28181

Most of the association rules for theft had a lift of over 1. Theft is one of the most frequently occurring crimes in the dataset. A total of 18 such association rules were found using the Apriori algorithm.

PREDICTIVE MODELING

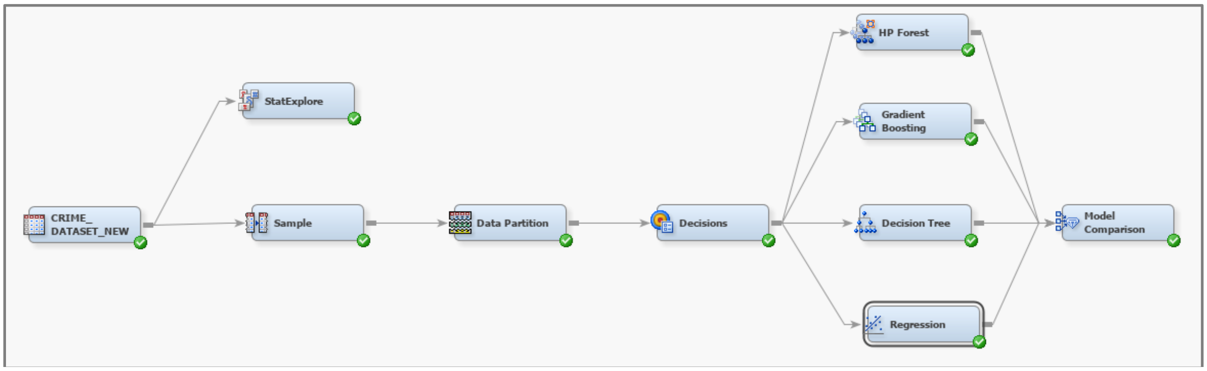


Figure 7. Process flow diagram in SAS® Enterprise Miner™ 14

The dataset had 3 levels of the target variable a) Theft b) Assault/Burglary and c) Vandalism and other crimes.

The dataset had 36% theft, 36% assault/burglary and 28% vandalism and other crimes. Hence, it was under-sampled using SAS's sample node and the distribution was set to 33% for all the 3 levels.

After sampling, the dataset was divided to 75% training dataset and 25% validation dataset to conduct honest assessment. After data partition, random forest, decision tree, logistic regression and gradient boost model were run on the dataset and were compared using the model comparison node.

The selection criteria for model selection node was set to validation-misclassification rate since the objective here is to get better accuracy and capture maximum crimes. With valid-misclassification as the criteria, the following result was obtained.

Table 3. Results of model selection node, with valid-misclassification rate as the criteria

Selected Model	Model	Target	Selection Criterion: Valid Misclassification Rate
Y	HP Forest	Crime	0.614193084
	Regression	Crime	0.616148621
	Decision Tree	Crime	0.616838205
	Gradient Boosting	Crime	0.666663236

Random forest gave the lowest misclassification rate which corresponds to the highest accuracy. The accuracy of the model is about 39% which is higher than the accuracy of all other models. The iteration plot is shown below in figure 8.

Figure 8. SAS output – Iteration plot for random forest model

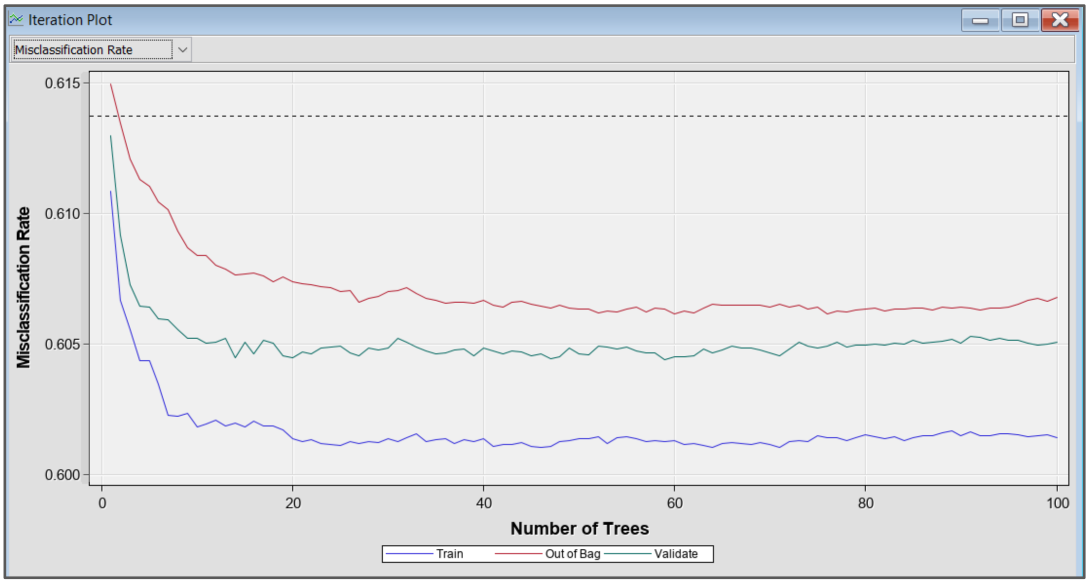
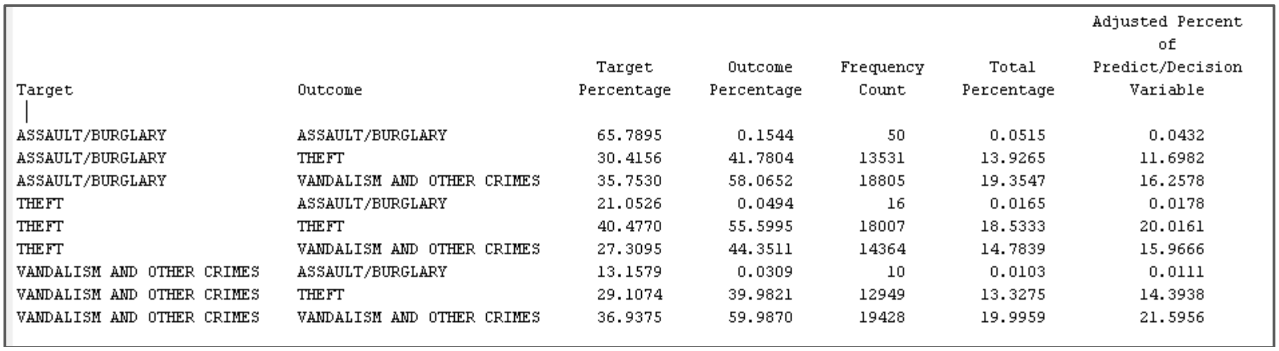


Table 4. Classification chart generated by SAS® Enterprise Miner™ 14 for decision tree model



```
                                                                                                  Adjusted Percent
                                                                                                         of
                                                      Target      Outcome    Frequency      Total     Predict/Decision
Target                  Outcome                    Percentage   Percentage     Count     Percentage       Variable

ASSAULT/BURGLARY        ASSAULT/BURGLARY             65.7895       0.1544         50        0.0515          0.0432
ASSAULT/BURGLARY        THEFT                        30.4156      41.7804      13531       13.9265         11.6982
ASSAULT/BURGLARY        VANDALISM AND OTHER CRIMES   35.7530      58.0652      18805       19.3547         16.2578
THEFT                   ASSAULT/BURGLARY             21.0526       0.0494         16        0.0165          0.0178
THEFT                   THEFT                        40.4770      55.5995      18007       18.5333         20.0161
THEFT                   VANDALISM AND OTHER CRIMES   27.3095      44.3511      14364       14.7839         15.9666
VANDALISM AND OTHER CRIMES  ASSAULT/BURGLARY         13.1579       0.0309         10        0.0103          0.0111
VANDALISM AND OTHER CRIMES  THEFT                    29.1074      39.9821      12949       13.3275         14.3938
VANDALISM AND OTHER CRIMES  VANDALISM AND OTHER CRIMES  36.9375   59.9870      19428       19.9959         21.5956
```

Table 5. Confusion matrix for the optimal model

Confusion Matrix	Actual Positive	Actual Negative
Predicted Positive	19,428	33,169
Predicted Negative	12,959	31,604

The random forest model had 19,428 true positives (TP), 31,604 true negatives (TN), 33,169 false positives (FP) and 12,959 false negatives (FN). Apart from accuracy, the following metrics provided more information on the performance of the model.

1. Sensitivity = $\frac{TP}{TP+FN}$ = 0.59

2. Specificity = $\frac{TN}{TN+FP}$ = 0.49

3. Precision = $\frac{TP}{TP+FP}$ = 0.37

4. F1 Score = $\frac{2*Sensitivity*Precision}{Sensitivity+Precision}$ = 0.4572

Table 6. Evaluation metrics

Measure	Value
Sensitivity	0.59
Specificity	0.49
Precision	0.37
F1 Score	0.4572

Figure 9 below shows the most important variables that determine the level of crime that is going to happen in a certain location. For this model, the most important factor was time span which is a derived variable, followed by area and other socio-economic data such as median household income, population of an area, median age, literacy level and economic standards. This proves that socio-economic factors are crucial in predicting crime and can be used for predictive policing.

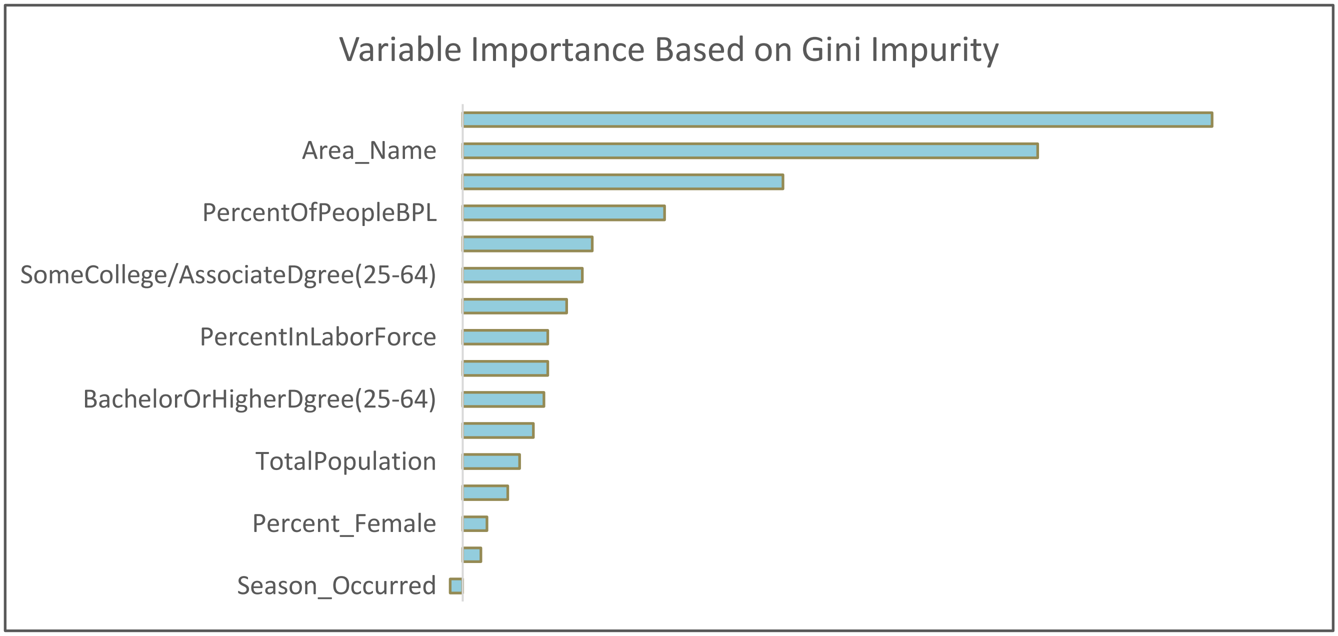


Figure 9. Variable importance for random forest model

SUMMARY

The LA crime data merged with socio-economic factors was analyzed to find some interesting facts. It was found that most of the crimes occurred in 77th street and the time at which they occurred mostly varied between 4.00 PM and 8.00 PM. Also, it was found that most of them happened on the street and Friday was the day when maximum crimes happened.

To accomplish the objective of the research which is association rule mining and predictive policing using socio-economic factors, Apriori algorithm was used and associations with a lift of more than 1 were mined and key results were reported in table 2. After analyzing the pattern, the dataset was prepared for predictive modeling. It was under-sampled to account for the minor variation in the distribution of target variable and run through logistic regression, decision tree, random forest and gradient boost models. Random forest gave the lowest misclassification rate and the highest accuracy rate (approxiately 39%). Hence, it was chosen as the winning model. With the winning model, it was found that socio- economic factors such as median income, median age, population etc. played a pivotal role in predicting crime and were thus reported to be significant variables.

LIMITIATION AND FUTURE SCOPE

With more information on victim's demographics, association can be mined between crime and victim's information that will help to provide better security to potential victims. The current data had too many nulls for victim's demographics as a result of which, it could not be used for modeling.

Along with the victim's demographics, weather data can also be added which will help check for impact of extreme weather on crimes that happen in and around a location. A very generic hypothesis states that warmer weather leads to more violent crimes. This hypothesis can be tested by adding valuable weather information to extend the study to cover its effects too.

Lastly, to conduct effective spatial analysis, spatial data for the location where the crime occurred can also be recorded. With the additional spatial data, crime hot spots can be efficiently identified.

REFERENCES

[1] Justin Juvenal. 2016. “Police are using software to predict crime. Is it a ‘holy grail’ or biased against minorities?”https://www.washingtonpost.com/local/public-safety/police-are-using-software-to-predict-crime-is-it-a-holy-grail-or-biased-against-minorities/2016/11/17/525a6649-0472-440a-aae1-b283aa8e5de8_story.html?noredirect=on

[2] Vice News. 2015.” Places, Not People, Are The Focus Of This New Crime-Fighting Data Analysis Tool, VICE, October 2015” https://news.vice.com/en_us/article/pa4kvv/places-not-people-are-the-focus-of-this-new-crime-fighting-data-analysis-tool

[3] Tahani Almanie, Rsha Mirza and Elizabeth Lor. 2015. “CRIME PREDICTION BASED ON CRIME TYPES AND USING SPATIAL AND TEMPORAL CRIMINAL HOTSPOTS” https://arxiv.org/ftp/arxiv/papers/1508/1508.02050.pdf

ACKNOWLEDGEMENT

I would like to extend my sincere gratitude to SCSUG for giving me the platform to showcase and present my research paper. I would also like to thank my professors Dr. Goutam Chakrabarty and Dr. Miriam Mcgaugh for their unending support and guidance throughout the course of this project.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name : Trishla Mishra

Student at : Oklahoma State University

E-mail : trishla.mishra@okstate.edu