Safe Chicago – Safe Place for Everyone using Data Mining Techniques

Shriram Bidkar
Graduate Student @UCR
sbidk002@ucr.edu

1. Abstract

Chicago has been in prime focus for its gun violence for many years now. Few year's back Chicago was responsible for nearly 50% of that's years increase in homicides in entire US, though the nation's crime rates remain near historic lows. While everyday someone loses his/her near and dear ones, there is every little relief which has been provided till today by law enforcement. Now it is time to get some assistance from technology which will help identify some trends through Data Mining Techniques. Build two-fold solution, one – better vigilance which will provide insight to law enforcement and second – give control to users who will get notification if the destination location is safe with given age, gender, time and geo-location.

2. Introduction

Let me confess, I am a father of 5 year old and constant news of gun violence bothers me and makes me thing, is this city good for me and my family? The city which has so far given me everything I ever wanted. So, after good thought on this subject I decided to fight back instead of running away from the situation. I decided to do something about it. First, step I took was see what CPD (Chicago Police Department) is doing to bring gun violence in control. After looking at everything they are doing I found that CPD is doing everything they can in their capacity with the limited tools to control it, but gun violence is out of reach. This is where I decided to help through technology like Data Mining Techniques where we will provide analytics to CPD which will help them identify gun violence crime sooner than before and avoid if not minimize fatalities. As part of analytics we will provide trend analysis using Association Rules technique of data mining which will provide correlation of one feature / attribute with another. Example – relation between victim and his/her age; victim and geo-location; victim and time of crime. We will provide these correlations through visual graphs using scatter plots and also through Apriori algorithm by providing support, confidence and lift. Another component of our solution to provide control to user instead of completely relying on CPD thereby reduce gun violence by notifying the user if they are tagged as potential victim with their given age, gender, time & geo-location. For this we will use binary classification methods like decision tree and SVM (Support Vector Machines) to identify if user of the application is potential victim. We will also evaluate performance of the algorithm using confusion matrix and AUC (Area Under Curve).

3. Related Work

The journey of solution to the problem starts with what exists today. While I kept hearing and seeing news of gun violence on television, I started exploring what are we as community including police department are doing to handle this situation. During my analysis I went through many articles and recommendations and two of them stand out.

Paper 1: Chicago Police Department – Statistics & Data (Annual Reports)

One article coming from Chicago Police Department in form of yearly report which provided good insight of summary data on numerous topics including violent and property crime; murder; firearms; arrests; domestic violence; traffic safety; juveniles; and hate crimes. These reports are available under statistics & data section of website. [source: https://home.chicagopolice.org/]

Below are some of the snippets from the report collected by Chicago Police Department which highlights statistics involved in gun violence –

Day	2018	% of Incidents
Monday	307	12.86%
Tuesday	284	11.89%
Wednesday	278	11.64%
Thursday	281	11.77%
Friday	333	13.94%
Saturday	432	18.09%
Sunday	473	19.81%
Total	2,388	

Table: 1 [source]

Type of Injury	2017		2018	
	Total	%	Total	%
Gun Shot Wound	604	91.52%	475	84.07%
Stab Wound	26	3.94%	48	8.50%
Injury From Assault	11	1.67%	9	1.59%
Blunt Force Injury	14	2.12%	19	3.36%
Strangulation	3	0.45%	6	1.06%
Other Injury	2	0.30%	8	1.42%
Total	660		565	

Table: 2 [source]

The above Table: 1 provides insight on shooting insight by day of the week and Table: 2 provides criminal homicide by injury type. There are few more tables from the report –

Location	2018	% of Incidents
Sidewalk	774	32.41%
Street	874	36.60%
Alley	167	6.99%
Residence Porch/Hallway	74	3.10%
Residence	86	3.60%
Vehicle Non-Commercial	46	1.93%
Residential Yard (Front/Back)	83	3.48%
Apartment	50	2.09%
Parking Lot/Garage (Non-Residence)	56	2.35%
Gas Station	29	1.21%
Park Property	20	0.84%
Other	129	5.40%
Total	2,388	

Table: 3 [source]

Weapon Type	2017	2018
Firearm - Shotgun	1	0
Taser / Stun Gun	0	2
Firearm - Rifle	2	2
Chemical Weapon	1	8
Mouth (Spit, Bite, etc)	13	17
Feet	20	18
Firearm - Revolver	40	21
Vehicle - Used To Strike Officer	37	32
Blunt Instrument	20	32
Hands/Fists	63	56
Knife/Other Cutting Instrument	65	61
Firearm - Semi-Automatic	271	236
Other (Specify)[2]	118	72
Unspecified[3]	41	0
Total	692	557

Table:4 [source]

Table: 3 provides valuable information on shooting location where as Table: 4 provides information on people identified as subject were physically armed. There are few other statistics like gender wise crime committed and geo-location of the crime scene collected by police department which is very helpful.

Short coming from this report to my best ability was correlation or association of these statistical data. There is no insight on which parameter influences other and this information would be vital for police department during their vigilance. This information would help CPD identify gun violence crime sooner than before and avoid if not minimize fatalities.

Paper 2: WTTW Chicago Public Media [source]

During my research I was able to find another source where good amount of analysis was done on gun violence by University of Chicago – Crime Lab and was published by WTTW Chicago Public Media.

Source: FBI UCR, Crime Lab analysis of CPD records

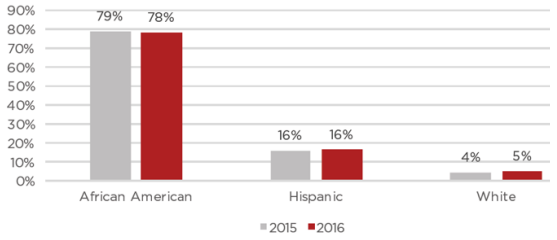


Chart:1 Homicide Victims by Race

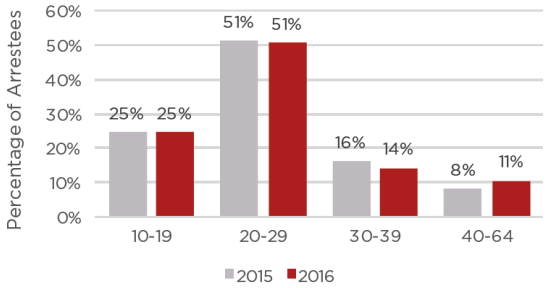


Chart: 2 Age of Homicide Suspects

As seen in the report (chart: 1) African Americans continue to be overrepresented among Chicago's homicide victims. Although African Americans comprise about one-third of the city's population, they made up almost 80 percent of homicide victims in both 2015 and 2016. This phenomenon is even more acute among African American men aged 15 to 34, who made up over half of the city's homicide victims despite accounting for just 4 percent of the city's population.

Based on the report (chart: 2) adolescents make up a larger share of homicide arrestees in Chicago than in other U.S. cities, a majority of those arrested for homicide in Chicago are in their 20s. Similar to what I had observed in the first case / report, there is good amount of statistical data

provided with difference in this report being more explanation on what may be triggering those trends. What I have not seen again in this report if there is any correlation between different trends or different attributes / features.

4. Proposed Method

With good understanding of what is being worked upon by Chicago Police Department and News outlets, I started my work on two folds as discussed earlier,

- Incorporate better vigilance where I worked on identifying unknown association of one feature / attribute to another feature / attribute using Association rule – Apriori algorithm.
- Provide proactive altering by giving control to user and NOT solely rely on CPD for protection. In this we built a model which will tag a user if he/she is potential victim given his/her age, gender, location and time. This model will be built on decision tree algorithm and SVM (Support Vector Machine) algorithm.

For both steps I used Data Mining methodology of Data Preparation, Data Preprocessing, Data Visualization for my first fold where we have better vigilance. For Proactive Altering we continued the same Data Mining methodology but this time we extended it to Model Building, Model Testing and Model Evaluation.

4.1 Data Preparation:

First step in this project was of data preparation and the data is to be collected from GVA website. While exploring this effort I realized the data pulled from the website through their export CSV functionality is not pulling all the necessary attributes / features required for the analysis. I had two options –

1. Either build the dataset manually which would consume lot of time
2. Search for aggregated dataset which will have all the necessary features within dataset

Evaluating pros and cons of both options and choose option 2 which would allow me to concentrate on analytics and to save time by not reinventing the wheel. I was able to find the dataset built by James Ko from Kaggle. From this dataset we have total of 29 attributes which will further undergo data preprocessing.

4.2 Data Preprocessing:

As part of data processing the first step, I took was to clean the data. As part of cleaning I went through every column and determined which column was not necessary and was not providing any significant value on visualization or association. With this step we reduced the attributes from 29 to 13. Most of the attributes which were removed were like '***incident_id***', '***incident_url***', '***source_url***', '***participant_name***'. As part of preprocessing we also identified null / blank values in attributes and to have better accuracy we filled blank values with appropriate values. For this we did it 2 steps –

1. Identify measure of central tendency and identify if the data is symmetrical or skewed.
2. Knowing the data is skewed we used median instead of mean to populate the missing/blank values

Example of attributes where we found that missing values were on ***'congressional_district', 'state_house_district', 'state_senate_district', 'latitude' & 'longitude'.*** Examples of these attributes are illustrated in the diagram below with help of histogram which was plotting for the entire dataset with specific attribute in mind. In the below chart 4 and chart 5 we have attributes – 'state_house_district' and 'longitude' in which are attribute values are skewed and hence we decided to go with median values to populate missing values instead of mean.

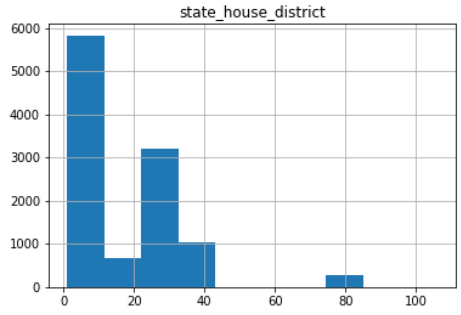


Chart: 3 Age of Homicide Suspects

As part of preprocessing we also utilized label encoder and PCA (principal component analysis) to ensure all features are in line for model building and provides better performance and accuracy.

Furthermore, we wanted to see if there are any underlying relations between different attributes for this, we used Association Rules mining technique especially Apriori algorithm. For this we removed all continuous variables and replaced numerical discrete data with meaningful data example: congressional district column with value of 1 to "Congressional District 1". To utilize Apriori algorithm we need to understand how that works. Apriori has three major components –

- **Support** - Support refers to the default reoccurrence of an attribute and can be calculated by finding number of transactions containing a particular attribute divided by total number of transactions.
- **Confidence** – Confidence refers to the likelihood that an attribute gender is also linked if attribute participant_status label is killedORInjured. It can be calculated by finding the number of transactions where gender is male and participant_status is KilledORInjured, divided by total number of transactions where gender is male.
- **Lift** –refers to the increase in the ratio of gender being male when someone is impacted (injured or killed). Lift(someone impacted –> gender -male) can be calculated by dividing Confidence(someone impacted –> gender -male) divided by Support(gender - male). Lift basically tells us that the likelihood of someone getting hurt or killed and the gender being male is 1.0003 times more than the likelihood of just someone getting killed or injured.

For this apriori algorithm we have used available package which provides the results of support, confidence and lift all together from this link. During implementation the package takes 4 parameters and to get the results which below we need to implemented values for those parameters –

```
apriori(
        apriori_list,
        min_support=0.0670,
        min_confidence=0.5,
        min_lift=4,
        min_length=5
       )
```

Picture 1: Apriori Algorithm Implementation

where,

min_support which denotes default popularity of an item and in this case we are assuming most of the crimes happens during Friday, Saturday and Sunday for last 5 years of data and hence we are calculating = ((52*5)*3)/11639 = 0.0670; where 52 is weeks, 5 is years and 3 is days (Friday, Saturday and Sunday).

min_confidence denotes likelihood that one feature (gender) is influencing another one feature (killedORinjured). In this case we are assuming we have high likely that we have 50% chance that would happen hence 0.5.

min_lift denotes increase in the ratio of one parameter/feature on another parameter / feature. In this case it is we are taking it as 4

Last parameter is **min_length** refers to number of features we want to include in the rule and in our case, we wanted to be 5.

After execution below is the result of code execution, details of the result are discussed after the output –

```
Rule: KilledORInjured -> State House District - 10.0
Support: 0.152447060891
Confidence: 1.0
Lift: 4.23085501859
=====================================
Rule: KilledORInjured -> State House District - 6.0
Support: 0.0814515420438
Confidence: 1.0
Lift: 6.56722446624
=====================================
Rule: KilledORInjured -> State Senate District - 5.0
Support: 0.0747737457165
Confidence: 1.0
Lift: 4.23085501859
=====================================
```

Picture 2: Apriori Algorithm Results

Based on the output of Apriori Algorithm the insight provided is very valuable and which was not shared (to my best ability) in any work survey. The output can be read as there is some correlation between victim and state house district 10 and 6 and state senate district 5. This insight can be shared with Chicago Police Department who in turn can put additional vigilance in and around that area. Details of this result is also shared in conclusion section.

4.3 Data Visualization:

Based on this processed (data cleaning and data reduction) below are some of the plots –

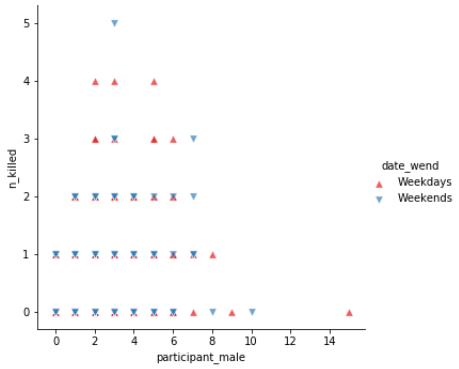


Chart 4: Male Vs Killed (Weekdays Vs Weekend)

In the above chart it depicts men more men get killed during the weekdays as against the trend in the survey which states more participants get killed during weekends. This need to be validated again before this trend is shared with CPD.

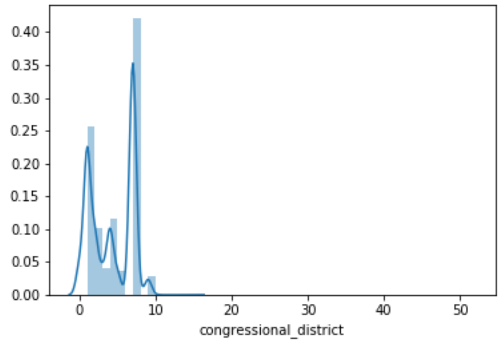


Chart 5: Congress District Vs Crime Rate

Based on the graph plotted it seems crime rate is significantly higher between congressional district 1 to 10. This gives insight to CPD to get more vigilance around these congressional districts from 1 to 10.

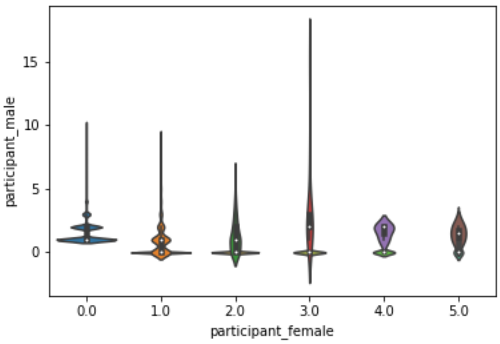


Chart 6: Male vs Female

This chart depicts male are more prone to gun violence than women.

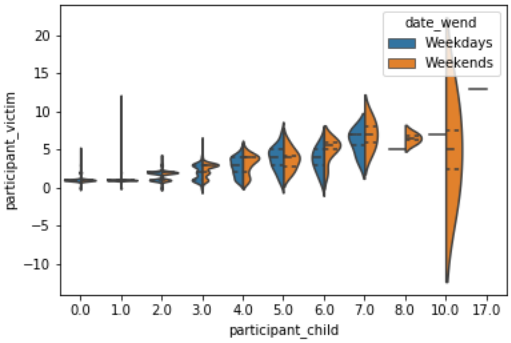


Chart 7: Child Victim Vs Weekdays & Weekends

This chart shares the details of child being a victim and the violance happened during weekend or weekday. It seems childern get hurt more during weekends as against weekdays.

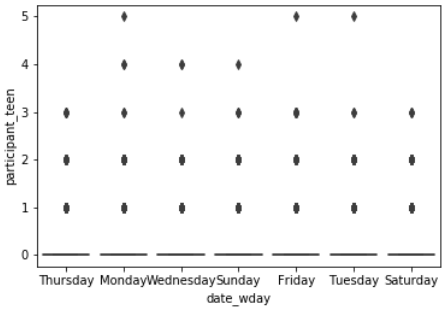


Chart 8: Teen Vs Weekdays

This chart shows teens get hurt most during Friday, Monday & Tuesday.

4.4 Model Building

For our second part of solution where we build model for proactive alerting is built on 2 algorithms. One is decision tree and another is SVM (Support Machine Vectors). For us to build the model we need first had to divide the dataset into test and train dataset to validate performance of model. For this we used holdout method with 70/30 ration, where 30% was used for testing the dataset.

4.4.1 Decision Tree –

In decision tree a tree structure is constructed that breaks the dataset down into smaller subsets eventually resulting in a prediction. There are decision nodes that partition the data and leaf nodes that give the prediction that can be followed by traversing simple IF..AND..AND….THEN logic down the nodes.

The root node (the first decision node) partitions the data based on the most influential feature partitioning. There are 2 measures for this, Gini Impurity and Entropy.

Entropy

The root node (the first decision node) partitions the data using the feature that provides the most information gain.

Information gain tells us how important a given attribute of the feature vectors is.

It is calculated as:

Information Gain=entropy(parent)–[average entropy(children)]

Where entropy is a common measure of target class impurity, given as:

$$Entropy = \Sigma_i - p_i \log_2 pi$$

where i is each of the target classes.

Gini Impurity

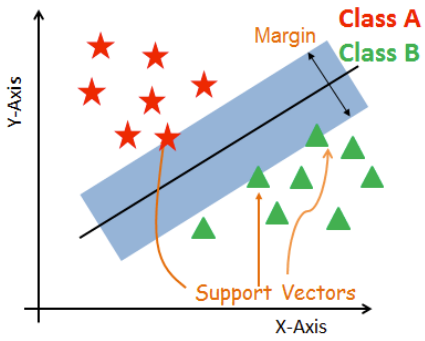
Gini Impurity is another measure of impurity and is calculated as follows:

$$Gini = 1 - \Sigma i\, p^2 i$$

Gini impurity is computationally faster as it doesn't require calculating logarithmic functions, though in reality which of the two methods is used rarely makes too much of a difference.

4.4.2 SVM (Support Vector Machines)

SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.



Picture 3: SVM

Support Vectors

Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

Hyperplane

A hyperplane is a decision plane which separates between a set of objects having different class memberships.

Margin

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.

Cost Function and Gradient Updates

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss –

$$c(x, y, f(x)) = (1 - y * f(x))_+$$

Hinge loss function

The cost is 0 if the predicted value and the actual value are of the same sign. If they are not, we then calculate the loss value. We also add a regularization parameter the cost function. The objective of the regularization parameter is to balance the margin maximization and loss. After adding the regularization parameter, the cost functions look as below.

$$min_w \lambda \parallel w \parallel^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+$$

Loss function for SVM

Now that we have the loss function, we take partial derivatives with respect to the weights to find the gradients. Using the gradients, we can update our weights.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

Gradients

When there is no misclassification, i.e our model correctly predicts the class of our data point, we only have to update the gradient from the regularization parameter.

$$w = w - \alpha \cdot (2\lambda w)$$

Gradient Update – No misclassification

When there is a misclassification, i.e our model makes a mistake on the prediction of the class of our data point, we include the loss along with the regularization parameter to perform gradient update.

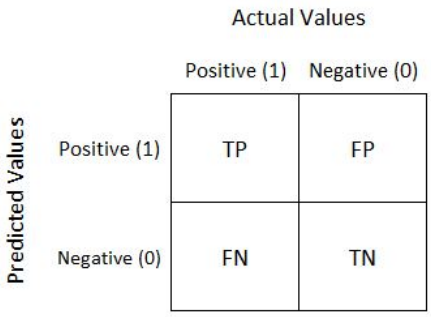
$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient Update – Misclassification

5. Experimental Evaluation

In order to select which model should be used, a selection metric is chosen upon which different models are scored. We will be using AUC-ROC (Area Under Curve - Receiver Operating Characteristics) curve which is one of the most commonly used metrics in industry. Rationale for using ROC curves is easy to understand and evaluate once we have good understanding of confusion matrix and errors.

Confusion matrix is used to showcase the predicted Vs actual class labels from the models. In our example where we are predicting if the user is a potential victim is example of binary class classification problem and confusion matrix can be –



Picture 4: Confusion Matrix

The class labeled 1 is the positive class in our example. The class labeled as 0 is the negative class here. As we can see, the Positive and Negative Actual Values are represented as columns, while the Predicted Values are shown as the rows. Where –

- TP = True Positive – The model predicted the positive class correctly, to be a positive class.
- FP = False Positive – The model predicted the negative class incorrectly, to be a positive class.
- FN = False Negative – The model predicted the positive class incorrectly, to be the negative class.
- TN = True Negative – The model predicted the negative class correctly, to be the negative class.

And,

- Type 1 Error: The model predicted the instance to be a Positive class, but it is incorrect. This is False Positive (FP).
- Type 2 Error: The model predicted the instance to be the Negative class but is it incorrect. This is False Negative (FN).
- Recall (TP / (TP + FN): Out of all the positive classes, how many instances were identified correctly.
- Precision (TP / (TP + FP): Out of all the predicted positive instances, how many were predicted correctly.
- F-Score = (2 * Recall * Precision) / (Recall + Precision): From Precision and Recall, F-Measure is computed and used as metrics sometimes. F – Measure is nothing but the harmonic mean of Precision and Recall.

AUC–ROC curve is the model selection metric for binary & multi class classification problem. ROC is a probability curve for different classes. ROC tells us how good the model is for distinguishing the given classes, in terms of the predicted probability.

A typical ROC curve has False Positive Rate (FPR) on the X-axis and True Positive Rate (TPR) on the Y-axis. The area covered by the curve is the area between the orange line (ROC) and the axis. This area covered is AUC. The bigger the area covered, the better the machine learning models is at distinguishing the given classes. Ideal value for AUC is 1.

Based on all the information provide when we trained and tested the model below are the ROC curve from Decision Tree and SVM (Support Vector Machine) –

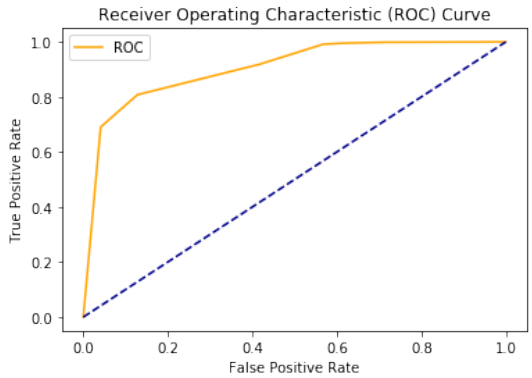


Chart 9: ROC from Decision Tree Model (Accuracy **90%**)

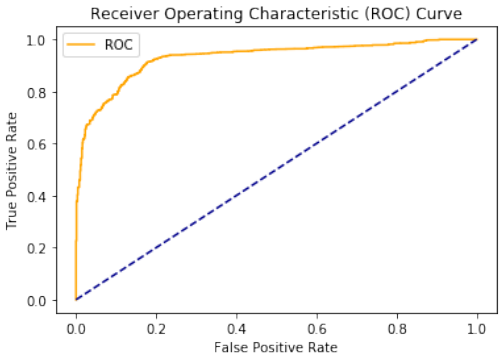


Chart 10: ROC from SVM Model (Accuracy **94%**)

Based on the ROC curve the value of decision tree was 0.91 which is 91% accuracy and ROC curve for mode run from SVM was 0.94 which is 94%. By this we can we ensure that model built using SVM is better than the model built using decision tree.

6. Discussion & Conclusions

In effort to make Chicago safe again, we had identified two fold solution which will first enable better vigilance using Apriori algorithm and second provide proactive alert using decision tree or SVM algorithm.

On better vigilance using Apriori algorithm we were able to identify new insights like association of impacted user with district house 10, district house 5 and district house 6. These insights were not visible in the reference work surveys and can be shared with CPD (Chicago Police Department). In turn CPD can put extra checks and balances in form of enhanced vigilance in these districts to enabling them to identify crime sooner than before and avoid fatalities if not minimize, making it safer than before.

Also, on proactive alerting we have built a model, in fact 2 models and out of which we are recommending using SVM (Support Vector Machine) due to its advantages on accuracy and scalability. This proactive alerting model will give control to user to make a choice by alerting them if the destination is safe given their age, time, location and gender.

Since the model built is limited to evaluation, I would like to pursue it further to productionize it where we build a mobile application which takes inputs from the user and provides alert by utilizing the model built.

7. References

1. Data Mining Concepts and Techniques by Jaiwei Han | Micheline Kamber | Jian Pei
2. WTTW Chicago Public Media
3. Chicago Police Department – Statistics & Data (Annual Reports)
4. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python
5. https://stackabuse.com/understanding-roc-curves-with-python/
6. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
7. http://benalexkeen.com/decision-tree-classifier-in-python-using-scikit-learn/
8. https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/
9. https://www.gunviolencearchive.org/
10. https://www.kaggle.com/jameslko/gun-violence-data