CRIME PREDICTION AND ANALYSIS

A PROJECT REPORT

Submitted for the partial fulfillment

of

Capstone Project requirement of B. Tech CSE

Submitted by

1. Vanshika Gadhwal, 22070521169
2. Arya Rehpade, 22070521154
3. Harshal Meshram, 22070521171

B. Tech Computer Science and Engineering

Under the Guidance of

Dr. Latika Pinjarkar

॥वसुधैव कुटुम्बकम्॥

SYMBIOSIS

INSTITUTE OF TECHNOLOGY, NAGPUR

Wathoda, Nagpur
2025

CERTIFICATE

This is to certify that the Capstone Project work titled "**CRIME PREDICTION AND ANALYSIS**" that is being submitted by **Vanshika Gadhwal- 22070521169**, **Arya Rehpade- 22070521154, Harshal Meshram- 22070521171** is in partial fulfillment of the requirements for the Capstone Project is a record of bonafide work done under my guidance. The contents of this Project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for award of any degree or diploma, and the same is certified.

Dr. Latika Pinjarkar
Name of PBL Guide & Signature

Verified by:

Dr. Parul Dubey
Capstone Project Coordinator

The Report is satisfactory/unsatisfactory

Approved by

Prof. (Dr.) Nitin Rakesh
Director, SIT Nagpur

ABSTRACT

We are interested in applying machine learning methods to datasets regarding crime (crime statistics in particular cities) and possible related factors (such as tweet data, income, etc.). Specifically, we are interested in investigating if it is possible to predict criminal events for a specific time and place in the future (for example, assigning a risk level for a shooting within the next week to different neighborhoods)

To be better prepared to respond to criminal activity, it is important to understand patterns in crime. In our project, we analyze crime data from the Torronto Dataset, scraped from publicly available in kaggle.

The use of AI/ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be "proving" to politicians that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop. Improvements in crime prevention technology will likely spur increased total spending on this technology. We also attempt to make our classification task more meaningful by merging multiple classes into larger classes. Finally, we report and reflect on our results with different classifiers, and dwell on avenues for future work.

Keywords— Python; Machine Learning;Clustering; Time Series; Education; Students; Performance;

CONTENT

Chapter 1

Introduction

Crimes are a common social problem affecting the quality of life and the economic growth of a society. It is considered as an essential factor that determines whether or not people move to a new city and what places should be avoided when they travel . With the increase of crimes, law enforcement agencies are continuing to demand advanced geographic information systems and new data mining approaches to improve crime analytics and better protect their communities Although crimes could occur everywhere, it is common that criminals work on crime opportunities they face in most familiar areas for them. By providing a data mining approach to determine the most criminal hotspots and find the type, location and time of committed crimes, we hope to raise people's awareness regarding the dangerous locations in certain time periods.

Therefore, our proposed solution can potentially help people stay away from the locations at a certain time of the day along with saving lives. On the other hand, police forces can use this solution to increase the level of crime prediction and prevention.for police resources allocation. It can help in the distribution of police at most likely crime places for any given time, to grant an efficient usage of police resources . By having all of this information available, we hope to make our community safer for the people living there and also for others who will travel there.

1.1. Problem Statement

Our study aims to find spatial and temporal criminal hotspots and also forcasting of crime using a set of real-world datasets of crimes. We will try to locate the most likely crime locations and their frequent occurrence time.In addition, we will predict what type of crime might occur next in a specific location within a particular time. Finally, we intend to provide an analysis study by combining our findings of a particular crimes dataset with its demographics information.

1.2. Scope And Objectives:

Our proposed solution can potentially help people stay away from the locations (crime hotspot)at a certain time of the day along with saving lives. On the other hand, police forces can use this solution to increase the level of crime prediction and prevention.for police resources allocation. It can help in the distribution of police at most likely crime places for any given time.

- Technology is noticeable
- Elimination of confusion
- To help people be aware of the crimes and help the society
- Interactivity
- Crime forecasting

1.3. Proposed Model

In this work, we will build a machine learning module. The model works on the concept of Time Series Forecasting and Clustering.After successful running of the module the analysis and the forecasting results are shown through graphs and plots. We intend to provide an analysis study by combining our findings of a particular crimes dataset with its demographics information.

Chapter 2

Literature Review

Machine Learning is undeniably one of the most influential and powerful technologies in today's world. More importantly, we are far from seeing its full potential. There's no doubt, it will continue to be making headlines for the foreseeable future. This article is designed as an introduction to the Machine Learning concepts, covering all the fundamental ideas without being too highlevel.

Machine learning is a tool for turning information into knowledge. In the past 50 years, there has been an explosion of data. This mass of data is useless unless we analyse it and find the patterns hidden within. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making.

We are drowning in information and starving for knowledge — John Naisbitt

Most of us are unaware that we already interact with Machine Learning every single day. Every time we Google something, listen to a song or even take a photo, Machine Learning is becoming part of the engine behind it, constantly learning and improving from every interaction. It's also behind world-changing advances like detecting cancer, creating new drugs and self-driving cars.

To learn the rules governing a phenomenon, machines have to go through a **learning process,** trying different rules and learning from how well they perform. Hence, why it's known as Machine Learning.

Ada Lovelace, one of the founders of computing, and perhaps the first computer programmer, realized that **anything in the world could be described with math**.

More importantly, this meant a mathematical formula can be created to derive the relationship representing any phenomenon. Ada Lovelace realised that **machines had the potential to understand the world without the need for human assistance**.

Around 200 years later, these fundamental ideas are critical in Machine Learning. No matter what the problem is, it's information can be plotted onto a graph as data points. Machine Learning then tries to find the mathematical patterns and relationships hidden within the original information.

2.1. Summary:

The summarized Literature review explains about what is Machine Learning,how it started and the influence of Machine Learning in different area. We also describe the approaches for gaining customers with the help of ML.

Chapter 3

Basic Concepts/Technology Used

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.**Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.

Types of Machine Learning Algorithms:

There are 4 types of Machine Learning today:

1. **Supervised Machine Learning Algorithms**
2. **Unsupervised Machine Learning Algorithms**
3. **Semi-Supervised Machine Learning Algorithms**
4. **Reinfrocement Machine Learning Algorithms**

1. **Supervised machine learning algorithms** can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

2. In contrast, **Unsupervised machine learning algorithms** are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

3. **Semi-supervised machine learning algorithms** fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

4. **Reinforcement machine learning algorithms** is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Classification

In the scope of Machine Learning, classification is an approach of supervised learning where the result set is to be cataloged as one of many existing result classes which is already trained.

Various classification algorithms include-:

Logistic Regression

Random Forest Classification

Support Vector Classification

Decision Tree Classification

Time Series

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time. Interrupted time series analysis is the analysis of interventions on a single time series.

Different classical time series forecasting methods; they are:

Autoregression (AR)
Moving Average (MA)
Autoregressive Moving Average (ARMA)
Autoregressive Integrated Moving Average (ARIMA)
Seasonal Autoregressive Integrated Moving-Average (SARIMA)
Seasonal Autoregressive Integrated Moving-Average with Exogenous Regressors (SARIMAX)
Vector Autoregression (VAR)
Vector Autoregression Moving-Average (VARMA)
Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX)
Simple Exponential Smoothing (SES)
Holt Winter's Exponential Smoothing (HWES)

Chapter 4
System Requirements

Jupyter Notebook Version

Version 5.7.8 is used in our Project.
The Jupyter notebook is a tool which we can use for our machine learning project and statistical analysis.We can download anaconda from the web sourse and within it Jupyter notebook most useful tool for machine learning purpose.

Python Version

Python 3.7.7 is used for this project.
Python is a very useful programming language.It is object oriented and interpreted.It is a high level language.There are lots of in-built libraries in Python for machine learning purpose which we can use easily.

Text Editor

Atom Version 1.47.0 is used here for bettre visualisation of the code structure and understanding of the code.

Windows Version

Jupyter notebook and python 3 can be used is all the operating systems including Windows, iOS and Linux.

It is best useful in Linux but can be used in windows as well.
It can be run on windows xp,vista,7,8 and the latest version windows 10 as well

Chapter 5

DATA DICTIONARY

Sl. no	Feature Name	Feature Description
1	Index	It stores the serial number of the crimes which are reported.
2	event_unique_id	Stores the unique id of the crime
3	occurrencedate	Date in which crime had actually occurred.
4	reportdate	Date in which the complaint was lodged.
5	premisetype	Type of the premise in which the crime took place like Commercial apartment, house, etc.
6	ucr_code	It is the abbreviation of Uniform Crime Reporting which enumerates offense codes. A code list that describes a criminal offense within a code book.
7	ucr_ext	It is a code list which describes the type of offense within a code book.
8	offence	Stores the type of offence such as Assault, robbery, etc.
9	reportedmonth	Month in which the complaint was lodged.
10	reportedday	Day in which the complaint was lodged

	reporteddayofyear	Stores the report day of the year such as 99, 321, etc.
	reporteddayofweek	In which day of the week the crime was reported.
	reportedhour	Stores the hour in which the report was filed.
	occurrenceyear	Year in which the crime took place.
	occurrencemonth	Month in which crime took place
	occurrenceday	Date of the month in which the crime took place.
	occurrencedayofyear	Day of the year in which the crime took place.
	occurrencedayofweek	Day of the week in which the crime took place.
	occurrencehour	The hour in which the crime took place.
	MCI	It is an abbreviation of Major Crime Indicators which stores the type of crime that has been reported such as assault, break and enter, robbery, etc.
	Division	Stores the division number of the neighborhood.
	Hood_ID	Contains the ID of the neighborhood.
	Neighbourhood	Name of the neighborhood where the crime took place.
	Long	Longitude of the place where crime had occurred.
	Lat	Latitude of the place where crime had occurred.

	ObjectId	It is an index of the crime after the reporting to the police.

Chapter 6

Implementation

Implementation is the process of converting the designed system architecture into working modules where it is made sure that all the functional and non-functional requirements are met.

The implementation section is divided into two parts-

- Creating the Core functionality (Mahine Learning Module)

7.1. Creating the core functionality

Dataset Importing

Importing of dataset from .csv file into the jupyter notebook.

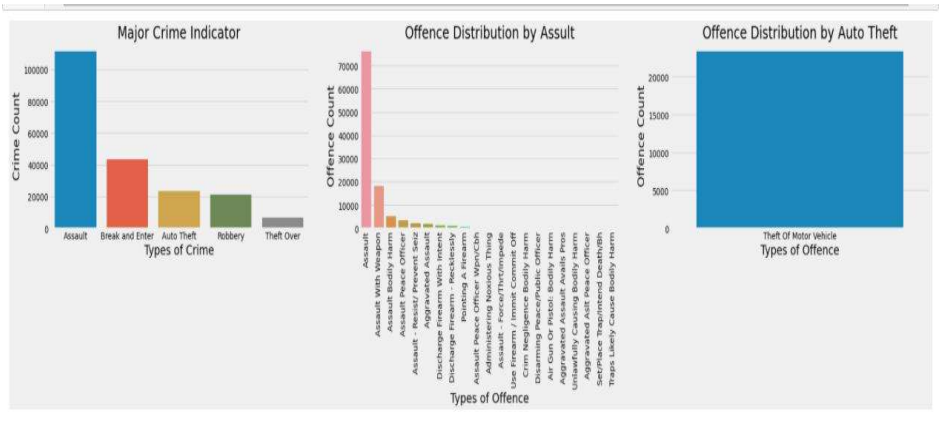
Our team members are well versed with data manipulation. The data analysis was done using Python3 in jupyter notebook.Some data that NULL in the dataset,they were droped. The Data was thoroghly analyzed to get the the minimum error in our Time Series method. Encoding were also done. Then the clustering was done.

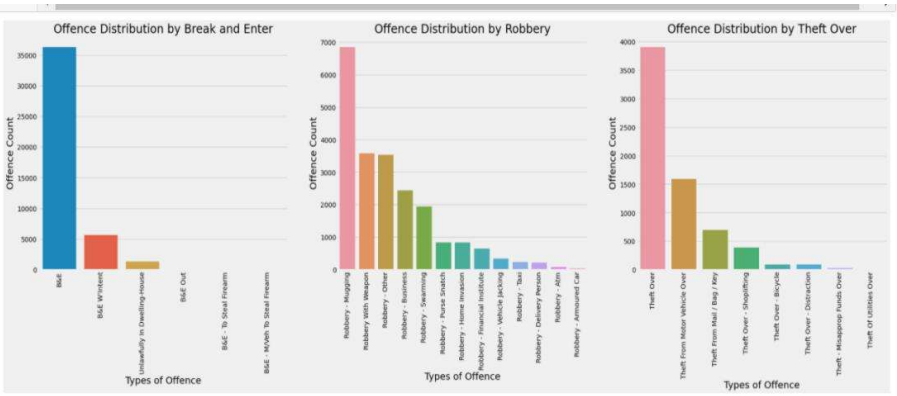
```
[ ]:
1 df=pd.read_csv('/content/drive/MyDrive/Colab Notebooks/Arpan/Copy of MCI_2014_to_2019.csv')
2 df.head()
```

Out[4]:

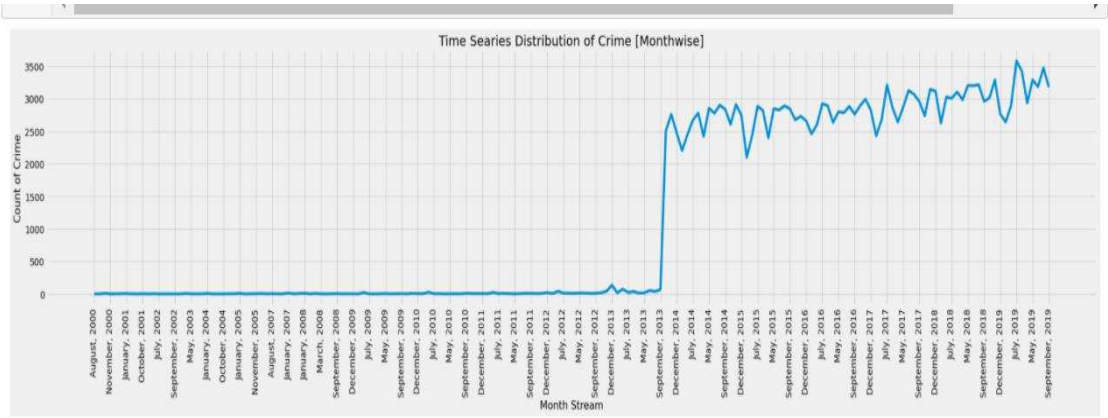
	X	Y	Index_	event_unique_id	occurrencedate	reporteddate	premisetype	ucr_code	ucr_ext	offence	reportedyear	reportedmo
0	-79.405228	43.656982	7801	GO-20152165447	2015-12-18T03:58:00.000Z	2015-12-18T03:59:00.000Z	Commercial	1430	100	Assault	2015	Decem
1	-79.307907	43.778732	7802	GO-20151417245	2015-08-15T21:45:00.000Z	2015-08-17T22:11:00.000Z	Commercial	1430	100	Assault	2015	Aug
2	-79.225029	43.765942	7803	GO-20151421107	2015-08-16T16:00:00.000Z	2015-08-18T14:40:00.000Z	Apartment	2120	200	B&E	2015	Aug
3	-79.140823	43.778648	7804	GO-20152167714	2015-11-26T13:00:00.000Z	2015-12-18T13:38:00.000Z	Other	2120	200	B&E	2015	Decem
4	-79.288361	43.691235	7805	GO-20152169954	2015-12-18T19:50:00.000Z	2015-12-18T19:55:00.000Z	Commercial	1430	100	Assault	2015	Decem

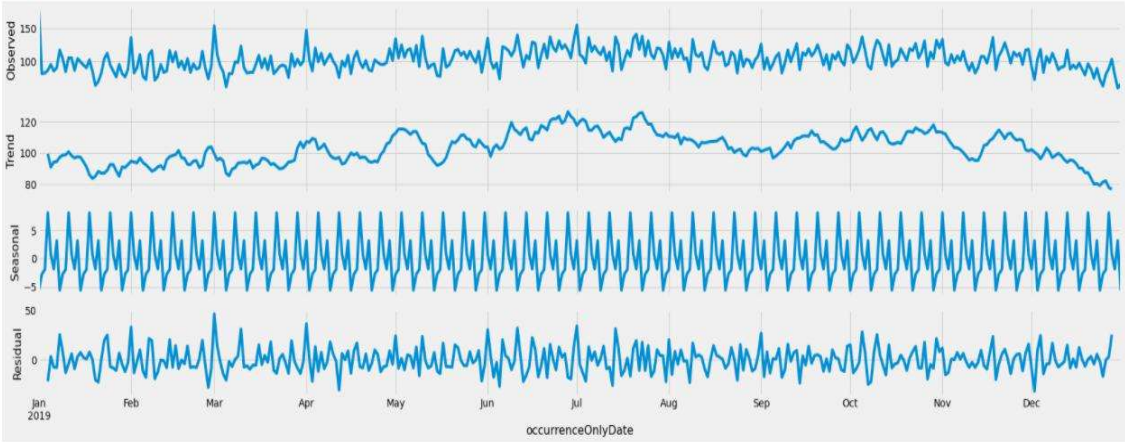
Plotting and analysis of different crime:





Time Series Analysis for Total Crime Count:





ARIMA Time Series Forcasting:

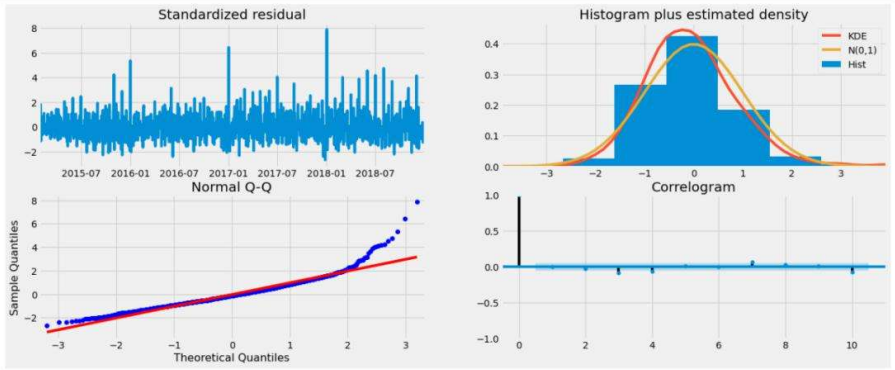
A popular and widely used statistical method for time series forecasting is the ARIMA model. ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data.

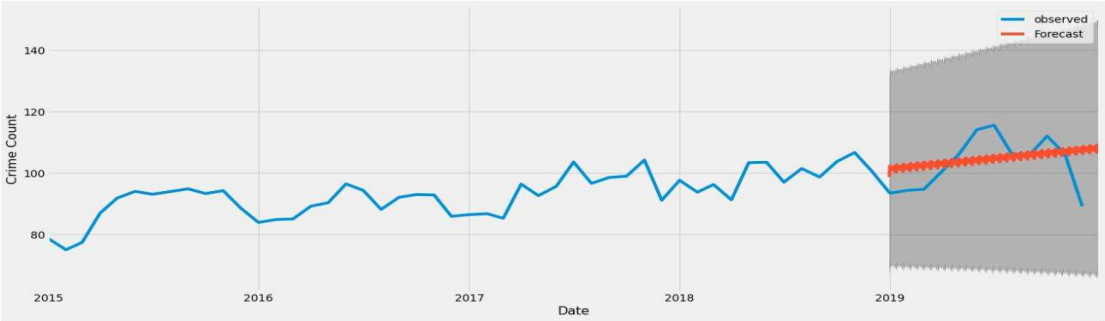
SARIMA for Time Series Forecasting:

An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called SARIMA.

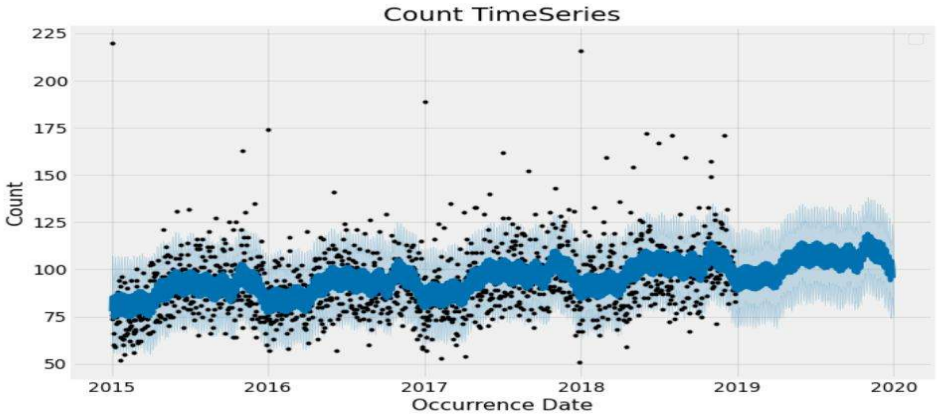
In this tutorial, you will discover the Seasonal Autoregressive Integrated Moving Average, or SARIMA, method for time series forecasting with univariate data containing trends and seasonality.

```
SARIMAXresults.plot_diagnostics(figsize=(20, 8))
plt.show()
```



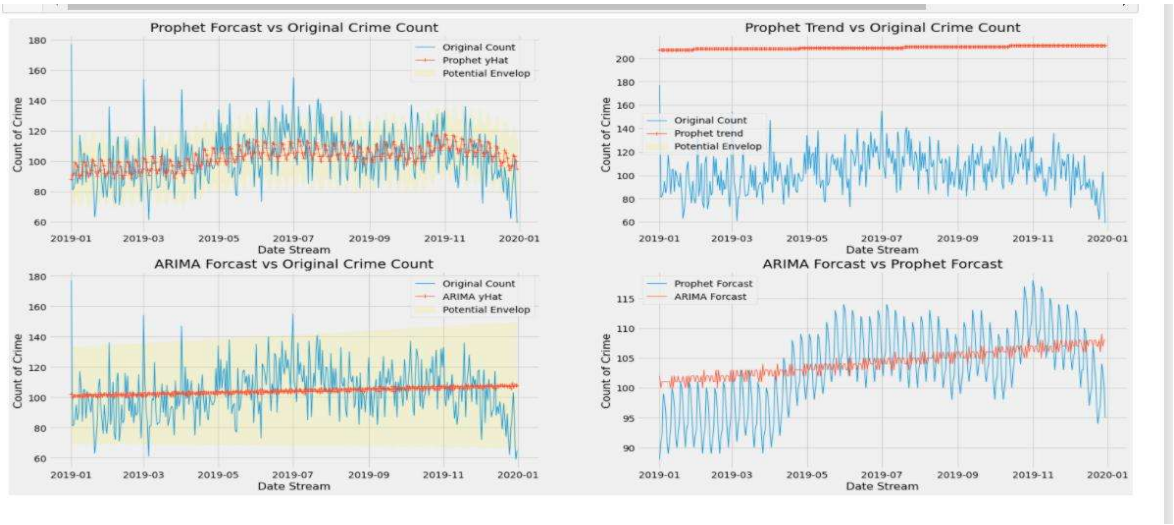


SARIMAX FORECASTING

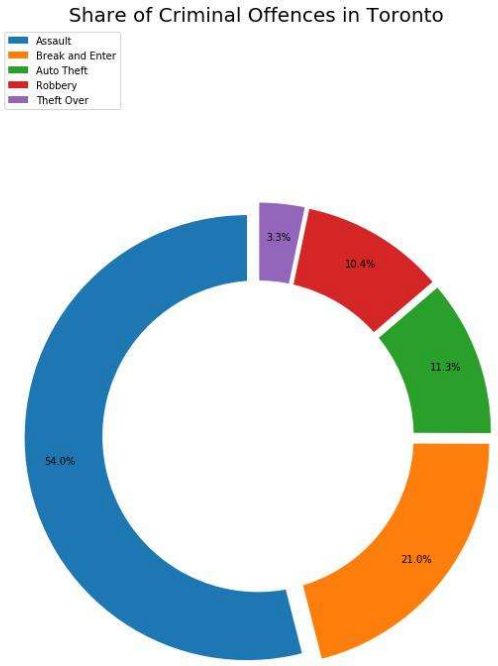


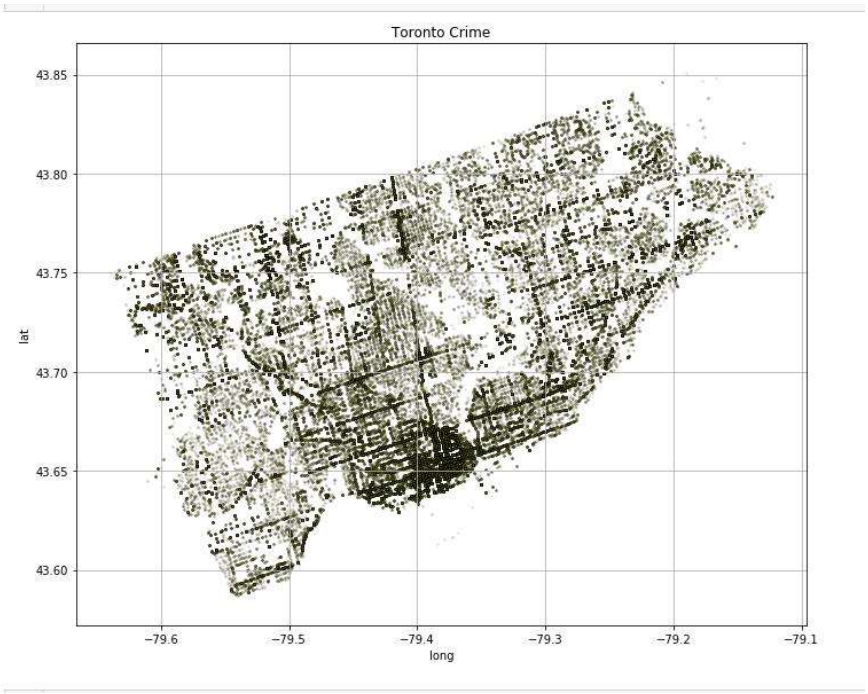
Model Comparison and Actual Value Deviation Analysis

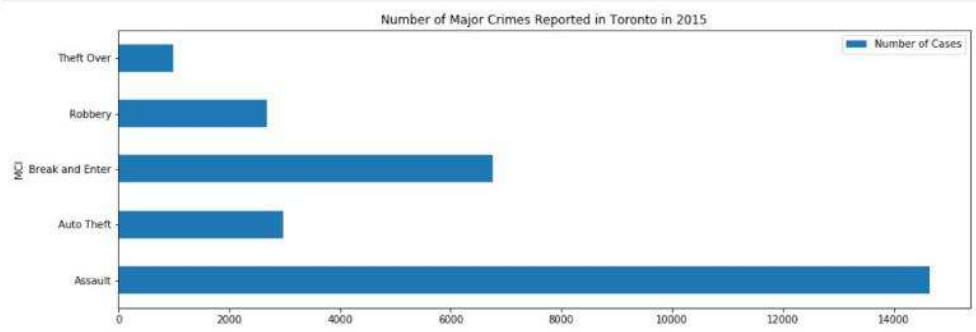
Prophet Time Series Forecasting



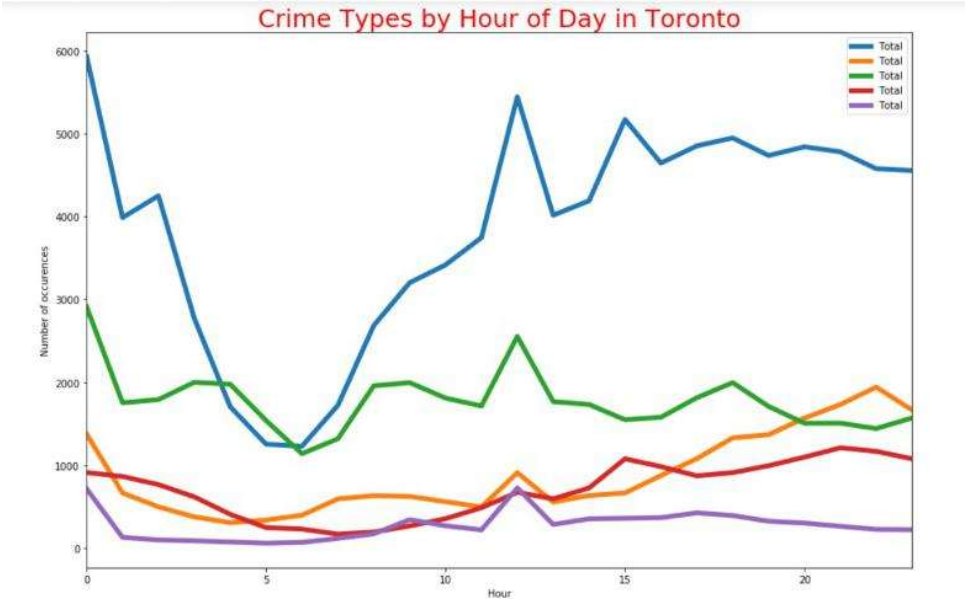
Differences between the original and the predictive model plotted in graphical manner




Toronto Crime
43.85
43.80
43.75
43.70
43.65
43.60
lat
−79.6
−79.5
−79.4
−79.3
−79.2
−79.1
long


Number of Major Crimes Reported in Toronto in 2015
Number of Cases
Theft Over
Robbery
Break and Enter
Auto Theft
Assault
MCI
0
2000
4000
6000
8000
10000
12000
14000


Major Crime Indicators by Month
Assault
Auto Theft
Break and Enter
Robbery
Theft Over
MCI
April
August
December
February
January
July
June
March
May
November
October
September
occurrencemonth
7500
6000
4500
3000
1500

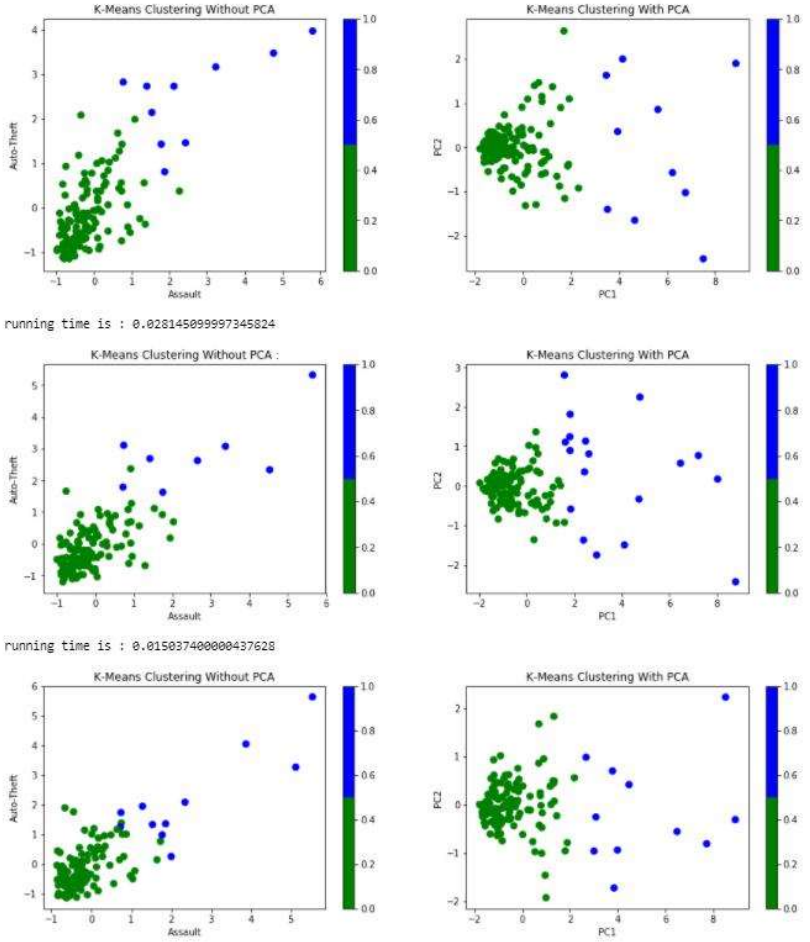

Crime Types by Hour of Day in Toronto
Total
Total
Total
Total
Total
Number of occurences
Hour
0
1000
2000
3000
4000
5000
6000
0
5
10
15
20

K-means Clustering:

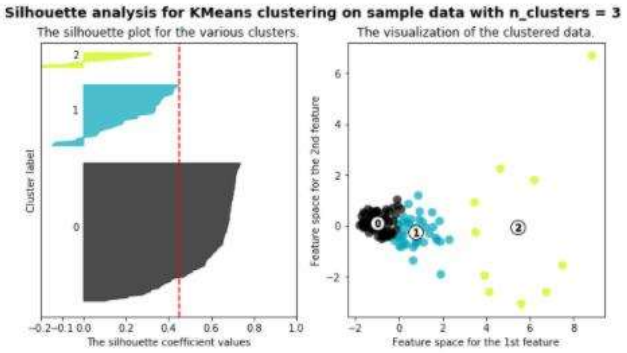
Kmeans algorithm is an iterative algorithm that tries to partition the dataset into *K*pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

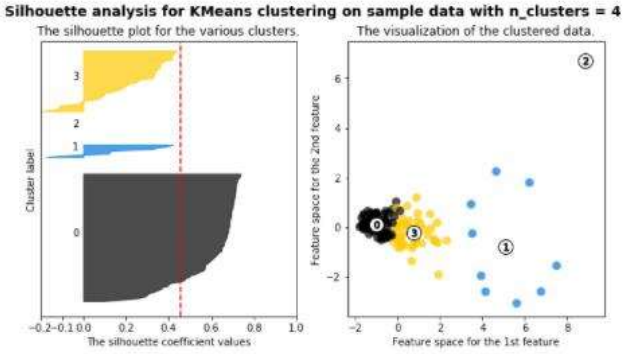
PCA:

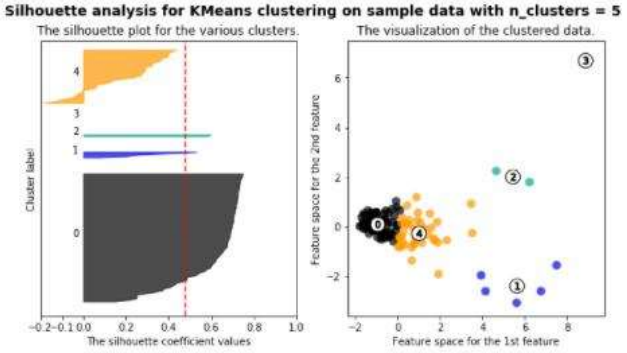
PCA is used in exploratory data analysis and for making predictive models. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

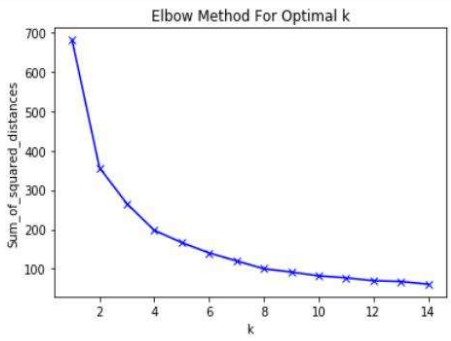


We have done clustering one without PCA and one with PCA and have found that applying PCA Before doing clustering can help in getting better clusters and also vizulazation becomes much better.









Elbow method was used to determine the number of clusters which should be used in order to get better clusters. It consists of plotting the explained variation as a function of the number of clusters. In this case we have considered 2 clusters.

Chapter 8:

Verification Table:

1		SARIMAX		Prophet	
2		MSE	RMSE	MSE	RMSE
3		290.11	17.03	248.27	15.76

So as we applied SARIMA model Prophet model and got this much Mean Sqaure Error and Root Mean Square Error. We have got less error in the Prophet time series model . So we can use this model for further works.

```
print("Accuracy of Random Forest with OneHotEncoder : ",accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test_OH, y_pred_OH))
print(classification_report(y_test_OH,y_pred_OH, target_names=definition_list_MCI))
```

```
Accuracy of Random Forest with OneHotEncoder :  0.5838382934955169
[[19536  2298   228    21   861]
 [ 4883  5581    27    14   211]
 [ 3280   309   260     7   371]
 [ 1164   327    20    17   118]
 [ 2862   634   142     9  1543]]
                precision    recall  f1-score   support

       Assault       0.62      0.85      0.71     22944
Break and Enter       0.61      0.52      0.56     10716
       Robbery       0.38      0.06      0.11      4227
     Theft Over       0.25      0.01      0.02      1646
    Auto Theft       0.50      0.30      0.37      5190

     micro avg       0.60      0.60      0.60     44723
     macro avg       0.47      0.35      0.35     44723
  weighted avg       0.57      0.60      0.56     44723
```

```
For n_clusters = 2 The average silhouette_score is : 0.7307306081429717
For n_clusters = 3 The average silhouette_score is : 0.4511660188781014
For n_clusters = 4 The average silhouette_score is : 0.4529368049556418
For n_clusters = 5 The average silhouette_score is : 0.4776823069308707
```

These are the accuracy of Random Forest Classification and silhouette scores of K-means clustering respectively. We have good better score in K-means clustering using 2 clusters.

Chapter 9:

Conclusion and Future Work:

While, there is little reason to believe that the crime rate will increase dramatically in the first decade of the 21st Century, given the anticipated increases in the globalization, sophistication, and organization of crime, one may conclude that the impact of crime on Western societies may be more severe than the one witnessed under a similar rate of crime in the past. The goal of any society shouldn't be to just catch criminals but to prevent crimes from happening in the first place.

1.Predicting future crime spots.

2.Predicting who will commit the crime.

Chapter 10:

References:

[1] N.Cristianini and J.Shawe-Taylor (2000). *An Introduction to Support Vector Machines*.Cambridge University Press, London.

[2] https://towardsdatascience.com/machine-learning/home

[3] Prophet Time Series : https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/

[4] https://www.ibm.com/industries/government/public-safety/crime-prediction-prevention