Predicting Potential Hotspot for illicit activities and Forecasting crime rate using Machine Learning

A DISSERTATION

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR
THE AWARD OF DEGREE
OF
MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEMS
Submitted by:

BRIJESH BAJPAI

2K20/ISY/06

Under the supervision
of

Dr. Jasraj Meena

(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE &
ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY, 2022

DELHI TECHNOLOGICAL UNIVERSITY

STUDENT DECLARATION

I, Brijesh Bajpai of Information Systems department, hereby declare that the thesis titled "Predicting Potential Hotspot for illicit activities and Forecasting crime rate using Machine Learning" is submitted by me in partial fulfilment of the requirement for the award of degree in Information Systems (Masters) is original and not copied from any source without citation. This work has not previously formed the basis for the award of and Degree, Diploma Associate ship, Fellowship or any other kind of similarity in title or recognition.

.

Date: May, 2021 (Name of student)
Place: New Delhi Brijesh Bajpai(2k20/ISY/06)

DELHI TECHNOLOGICAL UNIVERSITY

CERTIFICATE

This is to certify that Ms. Brijesh Bajpai (2K13/ISY/06) has finished the major project titled "**Predicting Potential Hotspot for illicit activities and Forecasting crime rate using Machine Learning"** as a partial fulfillment for the award of Master of Technology degree in Information Systems by Delhi Technological University. Under my supervision and guidance during the academic session 2020-2022. Dr. Jasraj Meena

Date: May, 2022

(Name of Guide)

Place: New Delhi

Dr. Jasraj Meena

DELHI TECHNOLOGICAL UNIVERSITY

ACKNOWLEDGEMENT

I want to thank my major project guide Dr. Jasraj Meena, Assistant Professor, lT Dept., Delhi Technological University, for her kind support and guidance that she has provided me in order to finish this major project. I would also like the to take this opportunity to thank the institute for providing us the infrastructure. At last I want to thanks my family and friends for the throughout support.

Date: May, 2022

Brijesh Bajpai

Roll No. 2K20/ISY/06

ABSTRACT

Crime is the biggest threat to the dream of a peaceful and safe society. Data analysis has a significant role in overcoming many issues related to the past and predicting future scenario based on past data. These advancements of these technologies can be taken into account for analyzing crime in a region.

These will result in tracking the crime events and also helps the police department to recognize the hotspots, the structure and the occurrence of crime. The exploratory analysis and data mining is the important part of this system since it helps to store and use the data to recognize the hotspots and will help the security and police agencies to deploy their force and patrolling vehicles in crime areas and manage their resources efficiently to the prevent crime incidence, maintain law and order and ensures safety of the citizens. In this project I will be using Los Angeles crime data set and regression models, ARIMA, SARIMAX, Facebook prophet are used for forecasting and in the end a comparison will be made to see which model works best to predict the occurrence of a crime incidence.

TABLE OF CONTENT

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Crime is an immoral act which is against the laws of the state and if a person commits a crime he shall be punished by the state or judiciary if found guilty. The police have the responsibility to ensure the safety of the citizens and to prevent crime and civil unrest. It has the special powers provided by the constitution with objective to maintain the law and order in the state. Thus, with the increase in technologies in every domain, Police Department also needs to have advantage of these technologies which will help them to be more efficient and breaking down the plans of criminals. Crimes are of different type but the ones which create a panic, fear and sense of insecurity among the citizens are related to theft, kidnapping, shooting, gang wars, rape and molestations. All occurrence of such crimes not only impacts the society but also have a negative effect on industries, governments and other development projects.

“Time series and Forecasting” is the main backbone technique of this project. This will help Police how to break the chain of the gang and a lot the bounty score to the prime members of a gang. Time series and Forecasting will also help the police to act efficiently and manage their resources and will also be beneficial for citizens as they can decide which period of time is the safest, whether they should buy a property or should prefer to live in a specific area. The Forecasting will help to predict the future occurrence of the crime and it will help students and business person to plan their trip accordingly. The dataset is being used and according to [1] the crime index of Los Angeles is 14 (100 is safest) which means it is safer than 14% of American cities, so by forecasting crime rates can also be forecasted for the future.

1.1 Motivation

Moving to the new cities for work, education is a very common phenomenon followed now by the society. Every year whenever a student completes his education then they need to join the join the jobs hence they need to move to the respective cities or even countries. The first and most important task is to find a place of living in a new city. No person wants to stay in such areas where the criminal activities are very frequent or there is high chance of becoming a victim.

Another scenario is with respect to women and children safety. Crimes like molestation, abuse, kidnapping are increasing and has become a problem for the public, police and for government as well. So the main motivation behind this project is to analyse the crime data and project the high risky areas which are termed as hotspots and predict the future of any crime to be happen to take place. By identifying the hotspots, police can be deployed more in numbers with advanced resources to take the criminals down.

This will help society, police and governments in many aspects as people will able to choose safe location to stay at, Schools and offices can have a safe environment so that it helps society to develop.

1.2 Organisation of dissertation

In the further chapter 2 is the background, which describes the literature survey and brief introduction to time series and its concepts. It also gives review how time series is usefull for making future predictions.

Chapter 3 presents the approach used to make the forecasting.

Chapter 4 contains the details of data set used and how data set is preprocessed. Also the methods used to make the predictions of crime forecasting.

Chapter 5 contains the results of the areas which are identified as hotspots and shows how forecasting models have forecasted the crime scenarios and shows the trend

Chapter 6 drives us to the conclusion and future scope of this project.

CHAPTER 2

BACKGROUND

2.1 LITERATURE SURVEY

In past few years data analytics became very popular technique in coping up with the real-world problems like in transportation, stock market analysis, etc. Using the approach, we have the opportunity to extend it in analysing crime events as well, this leads to have a pattern in data which can be further analysed to extract more information of criminal activities. Forecasting will help the organizations to prepare themselves and planning a strategy so that these events can be prevented. The identifications of hotspot display the picture of the areas of the city which are at high and least risk.

In the journal paper by In IEEE paper by Suhong Kim [2] two different models of machine learning which are d-tree model and KNN to analyze the crime pattern and the type of crime took place, the accuracy of the findings are 39-44 percent. In the paper [3] “Crime tracer: Activity space-based crime location prediction,” by M. Tayebi, has used a random walk-based approach to in his research to observe mapping details and predict the locations of crime occurrence. The areas which are known for the crimes are termed as activity spaces, the areas are observed by the activity of known criminals of the activity spaces. The study showed that the criminals often avoid the unknown locations and take benefit of activity spaces to commit crimes more frequently.

In the paper [4]**,** “Using Machine Learning Algorithms to Analyze Crime Data,” the authors have used WEKA, a software for data mining available publicly to study the different crime patterns. They have also implemented three algorithms, Linear Regression, Additive regression and decision stump algorithms under the features over data set. Overall, they found in the study that Linear Regression was most accurate algorithm.

Abraham and Chandy in the paper [5] have used cloud computing for feature extraction and proposed random forest classifier for the same. The features that are extracted are like user identification, request number, time of expiry, time of arrival and memory required. Post feature extraction the learning is performed which learns the details of the features extracted from the request number. This trained data is used to perform the prediction task of predicting work load.

In the paper [6] the researcher has used forecasting method on Pittsburgh data of over a short period of time nearly of a month, Monthly seasonality has been used for forecasting and has been found effective over the data of a shorter period of time. Another IEEE paper [7] uses the K means clustering algorithm based on the data provided by 356 users and their reports regarding the harassment. K means clustering has been used to mark the geographical locations to determine the area is a hotspot for harassment. Several studies focus on the dual-involved population in Los Angeles to provide a more in-depth analysis of the crime prediction among this population.

Young et al.[8] conducted research on the report that was on data desk and being published on the times of Los Angeles, which was comprised where both the team of media house and the team of web engineers. The purpose of this research was to investigate the technological shifts that occurred in the newsroom at the beginning of the twenty-first century. The relationship between the authorized stores licensed by the state medical boards and the rate of intense crimes and ownership crime in Los Angeles is investigated in greater depth by Contreras in a study that was published in [9]. The results of their investigation suggest that marijuana dispensaries should be regarded as a potential source of criminal activity. An additional study that is very close to this one was carried out by Dierkhising et al. [11], and their findings indicate that there was a significant amount of female involvement within the sample. Orsogna et al. tackled the difficult problem which is analyzed for this complex data by employing the modelling research using its tools, many mathematicians some other specific scientists in order of the creation of making predictions about the measures which are taken for the safety and all the prediction for the crime[12]. Almanie summarize the dataset of it, provide the information about it in order to make a potential crime type prediction in the year 2014, than they implement the algorithm which is Apriori, the NB algorithm, and the DT algorithm [13]. Considering robbery as the primary attempted crime, the outcome achieved an accuracy of prediction of 54 percent. Through the use of LR Analysis as the result provided from the years of 2000 where it was done in 8-12 for the city of Chicago, Christian et al. [14] establish a connection in the two system of the model that are socioeconomic and the second one is sustainable development which have different indicators such as the rate of the poverty and the rate of the unemployment was being provided from the given data and criminal activity.

In one of the paper author presented a method that depends on space and time to discover the crime in red zone areas that are primarily metropolitan and related active area are there for every red zone by making use of the clustering technique [15]. “Computational mechanism to classify the crime using machine learning techniques [16] proposed a flexible computational implementation tool to analyse the crime rate in a country”. This helps classify cybercrimes. Sometimes population also plays a significant role to fetch the estimations which can influence the prediction,H. Wang, [17] proposed an bootstrap method to see the same provided when the data used is too large.

2.2 TIME SERIES ANALYSIS

2.2.1 Introduction

This procedure of extrapolation is carried out by a forecasting model, which is in turn a collection of algorithmic operations termed a forecasting method.

The series which is of Time has a collection of statistical information which has been stored or seen as a variety of different time stamps. It is possible for these time stamps to be continuous or discrete. A "Time Series" presents a type of data that is comprised of two components: units of time and some value or collection of values that are associated with that unit of time. When analyzing time series data, the axis that is given the most weight is time.

Since it was mentioned earlier that time stamps can either be discrete or continuous, this suggests that there are two different ways in which a time series can be stored. The first method for storing a time series involves the intervals of time being stored in a discrete format. As a result of the fact that time is deemed to be the primary axis rather than a metric, certain values are recorded at specific timestamps, regardless of whether they occur regularly or consistently. These kinds of time series are referred to as discrete time series. The information regarding sales is an illustration of a discrete series of the time.

A time series can also be stored by recording the values of connected variables in successive order along the major axis. This is the second method (time). The most popular examples of continuous time series are the datasets received from a variety of devices, such as signalling devices, mobile internet data usage trackers, and so on and so forth.

2.2.2 Classification of Time Series

Time series is being grouped and discriminated in two different ways, one of which is based on the dependency between new values of the variable associated with the time stamps and the previously recorded values in the past. The other classification is dependent up of stationarity as that are given in time series. Let's discuss about these two in more detail.

As mentioned above that the one of the classification types of the time series is based on the dependency between newly recorded values and the previous values which were recorded in the past. Under this type of classification there are two categories of values:

- **Short-Term memory time series*:*** The recent data or the data recorded over a short period of time is observed in this is time series' STM. Characteristic is dependent on this series is that the autocorrelation has fast decrease and therefore the correlation when plotted and compared with different lags falls very fast at much higher rate. The financial plots such as of sales and from other finance sectors are typical examples of STM time series.
- **Long-term memory time series*:*** These series have slow decrease in the autocorrelation function. This type of series has high correlation when plotted and compared with large set of lags show slow decrease in autocorrelation. The climate change reports, data obtained from weather forecasting organizations collect data over a large period of time, such type of data records are examples of Long-term memory time series.

The second way of classification of time series is based on the stationarity. Stationarity means the properties of the time series such as mean, variance, covariance which are called statistical properties.

Two types are identified based on changes in these properties:

- *Stationary Time Series:* the series of the given time is being called stationary as they do not contain the statistical properties because change over time. It does not mean that it should be constant throughout but the way it changes should not change over the time. We can say that it should be comparatively constant.

- *Non-Stationary Time Series:* The series do not exhibit the constant behavior of the statistical properties over the period of time. The mean, variance and covariance fluctuate over time. Hence known by the name of Time Series of this Non-Stationary subject.

In most of the time series some of the primary subject which is the key factor is non stationary behavior is observed which is not suitable for forecasting and hence these time series are made stationary by applying various methods (differencing, transformation and detrending) on these series to make them stationary so that forecasting can be done for future predictions.

2.2.3 Uses of Time Series

There are two main uses of time series:

1. To analyze the previous records of the associated variable over time to identify the trends and patterns in the given dataset. This is considered as exploratory data analysis which shows the responses over some different time periods such as weeks, months, years.

2. Forecasting the values which may occur in future based on current and previously recorded values. The forecasted data may change depending upon the lag used to study the data, in some cases like short-term memory time series the lags are not very large unlike the long-term memory time series.

2.2.4 Components of Time Series Analysis

Understanding the structure of any underlying pattern of data recorded across time is crucial to analyzing a time series. The patters which are obtained by analyzing a time series contains different components which cumulatively gives some values of the variable associated with the timestamps. Each component can be split individually and can be modified to such an extent that it can change the form of the values observed. The basic ingredients of time series can be defined as:

- Trend: the movement of observed values over the timestamps are shown by the trend, trend can be upward (increasing slope) or downward (decreasing slope) depending on the values recorded. Figure 1 shows the linear upward trend, this trend can also be downward.

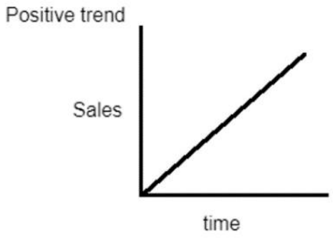


Figure 1: graph showing upward trend

- Cyclical: A repetitive upward and downward movement around the trend is known as cyclical. In figure 2 we can see that the plot is repeating after some time intervals. Mostly it is observed in finance related or climate related data where a phase repeats at least once over a calendar year.

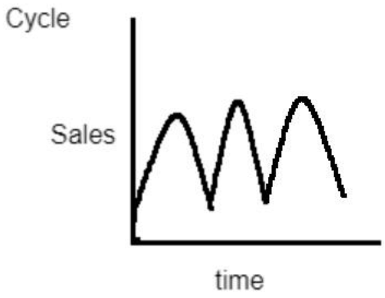


Figure 2: graph representing cycle in TSA

- Seasonality: The movement in the recorded values of the variable associated with time tends to repeat itself at every unit of time, may be weeks, months or years. For example, the sales of air conditioners rise during the summers and falls during the winters. This keeps on repeating every year. Figure 3 shows the seasonality pattern, we can see it in graph represented below.

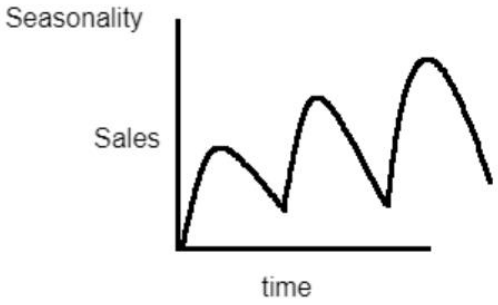


Figure 3: graph representing seasonality in TSA

- Irregularity: Figure 4 shows the undefined or random pattern which is neither cyclic nor repetitive over the time span. Such patterns are considered as irregular.

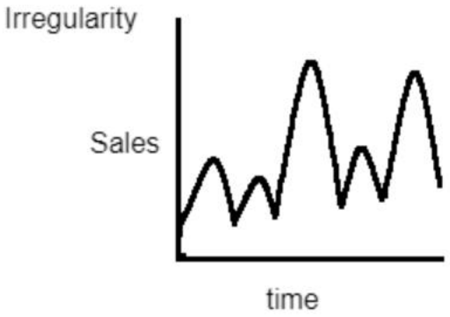
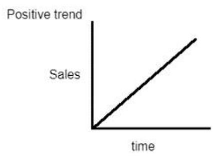
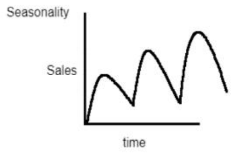
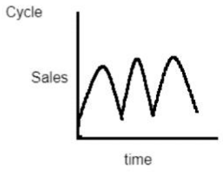
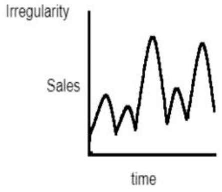


Figure 4: graph representing irregularity in TSA

Table 1 below shows a comparison among the 4 components of time series analysis which are the form of trend, seasonality, cyclical and irregularity.

Table 1: Components of Time series Analysis

	Trend	**Seasonality**	**Cyclical**	**Irregularity**
Time	Fixed time interval	Fixed time interval	Not fixed time interval	Not fixed time interval
Duration	Long and short	Short	Long and short	Irregular

Visualization				
Nature-I	Gradual	Swings between up or down	Repeating up and down	High fluctuation
Nature-II	Upward/Down trend	Pattern repeatable	No Fixed Period	Short and not repeatable
Prediction capability	Predictable	Predictable	Challenging	Challenging

The components of time series, mainly the trend, cyclical and seasonality are used to study the pattern of the observed data recorded over the time. All datasets are not characterized by all the components but the main objective of these components is to split these components from the data and observe how much they impact the values of the secondary axis variable. This leads to the more accurate results in forecasting future values.

2.3 Types of Time Series models

The most crucial step in analyzing the time series data is to analyze the relation between the trend and the seasonality. This can be studied by understanding the components of time series and the way they relate to other components like as said above to understand the how trend is interacted to seasonality. Models of time series are differentiated in two ways:

- **Additive model:** In additive model all components along with the error component are combined all together. This means that the additive model is combination of trend, seasonal and cyclical components and the error components is added to it as well. By this we can understand that all the components are mutually exclusive in nature and independent of the changes done to any other component.

additive model time series = (seasonality)+(cyclical)+(trend)+ (error component)

All four components do not affect the remaining components since they are mutually exclusive.

- **Multiplicative model:** In the multiplicative model no component is mutually exclusive, since all the components are multiplied and hence any change done in any of the component will affect the entire series and the components. Any change takes place is in form of percentage unlike the additive time series model where the change is in absolute value.

multiplicative time series = (seasonality)*(cyclical)*(trend)*(error component)

When analyzing the time series, it is observed that the moving mean is highly sensitive as and when compared to moving median. If the time series consists of very often time periods with null value which are known as inactivity period, then many unexpected zeroes can be part of the trends. These trends may also result in irregularity of the pattern observed in the time series.

2.4 Time Series Forecasting

In time series data, the time is at primary axis while the dependent variable is along the y-axis or secondary axis. As the time moves the data recorded over the time also changes. Various sectors such as finance sector, industrial sector need to study the past records and plan for the upcoming term in prior. For such scenarios forecasting ways can be predefined to rearrange the utility as our upcoming term depends upon the previous values.

The time series forecasting means to observe the historical information of the given subject can be predicted as our value which can be upcoming may be defined by it as the variable which is dependent upon time. There are number of algorithms which performs the functions which lead to the expected results. These algorithms represent the models which are known as forecasting models. A time series may simply be a sequence of values stored in a variable that is indexed by time, or it may be a multivariate sequence of values in which the effects of the time series may be better explained by using a wide variety of external factors. Either way, a time series may be univariate or multivariate. For instance, the locality, the weather, and the number of people in the area are all examples of external elements that might contribute to crime.

Time series analysis and forecasting have been utilized for a considerable amount of time in many different facets of human life. Airlines have been using it to analyze the pattern of passengers travelling between two airports in order to better match the supply and demand of the airlines and to deploy their aircraft in the most efficient manner possible. If we take Delhi and

Mumbai as an example, we can say that these two cities have the most frequent travellers and that the demand for flights between these two cities is the highest. On the other hand, the frequency of travellers between Lucknow and Patna is not the same as the case that of Delhi and Mumbai. Because of this, increasing the number of planes that fly between Lucknow and Patna will lead to a reduction in profits for the airline and will be a waste of resources. In a similar fashion, this model can also be applied in monitoring the crimes that occur in the city in order to ensure proper planning of patrolling vehicles of the police in those areas that are highly sensitive and also to make travel advisory for business people and students so that they can plan their travel safely during the most prominent times of the day. The police will be able to use this information to figure out how to break the chain of the gang and increase the bounty score for the most important members of the gang. Citizens will also benefit from time series analysis and forecasting because it will allow them to determine which time of year is the safest, whether or not they should invest in real estate, and whether or not they should prefer to live in a particular area. Time series analysis and forecasting will also help the police to act more efficiently and manage their resources. The Forecasting will help to predict the future occurrence of the crime, which will assist students and business people in appropriately planning their trip.

CHAPTER 3

Proposed Approach

3.1 Proposed Approach

In this work we have proposed that the ensemble of one or more models have the capability of optimizing the evaluation matrix or RMSE in case of regression models. The main issue lies in the fact that the weightage of the models should be multiplied by an optimized or tuned value which can be decided or specified by several experimental by continuous experimentation.

These optimum weightage value will be multiplied with the outputs of different models and will give the final output. Now with the help of different evaluation matrix we can find out the performance of this ensembled model. The performance of these ensemble models are not sure to be better than the existing ones, the result can be better or it can have more error. So the parameters should be decided accordingly.

CHAPTER 4

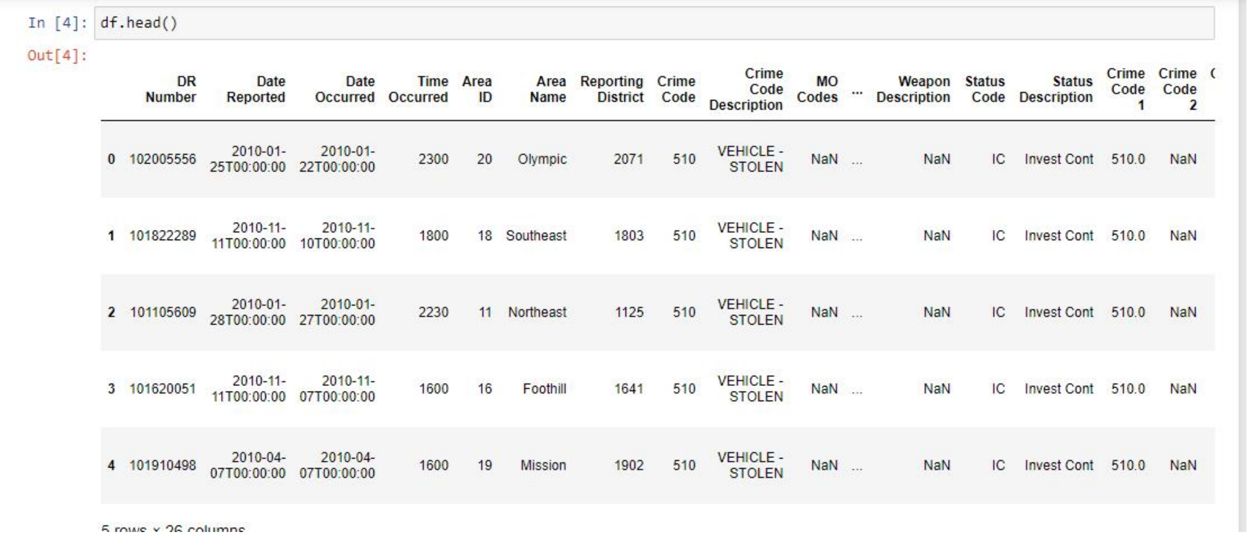
EXPERIMENTAL SETUP

4.1 Data Set Used

Data set used in this project is crime data set of city of Los Angeles, data is obtained from kaggle[19]. The data set has 1993259 rows and 26 columns. It has crime records of the of Los Angeles since 2010 to 2018.

4.2 Data Preprocessing

This chapter discusses dataset preprocessing. Data Preprocessing is crucial in analytics/ML projects. After extracting the needed data, it's important to get the key attributes. Below in figure 5 is the snapshot of raw dataset:



```
In [4]: df.head()
Out[4]:
```

	DR Number	Date Reported	Date Occurred	Time Occurred	Area ID	Area Name	Reporting District	Crime Code	Crime Code Description	MO Codes	...	Weapon Description	Status Code	Status Description	Crime Code 1	Crime Code 2
0	102005556	2010-01-25T00:00:00	2010-01-22T00:00:00	2300	20	Olympic	2071	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest Cont	510.0	NaN
1	101822289	2010-11-11T00:00:00	2010-11-10T00:00:00	1800	18	Southeast	1803	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest Cont	510.0	NaN
2	101105609	2010-01-28T00:00:00	2010-01-27T00:00:00	2230	11	Northeast	1125	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest Cont	510.0	NaN
3	101620051	2010-11-11T00:00:00	2010-11-07T00:00:00	1600	16	Foothill	1641	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest Cont	510.0	NaN
4	101910498	2010-04-07T00:00:00	2010-04-07T00:00:00	1600	19	Mission	1902	510	VEHICLE - STOLEN	NaN	...	NaN	IC	Invest Cont	510.0	NaN

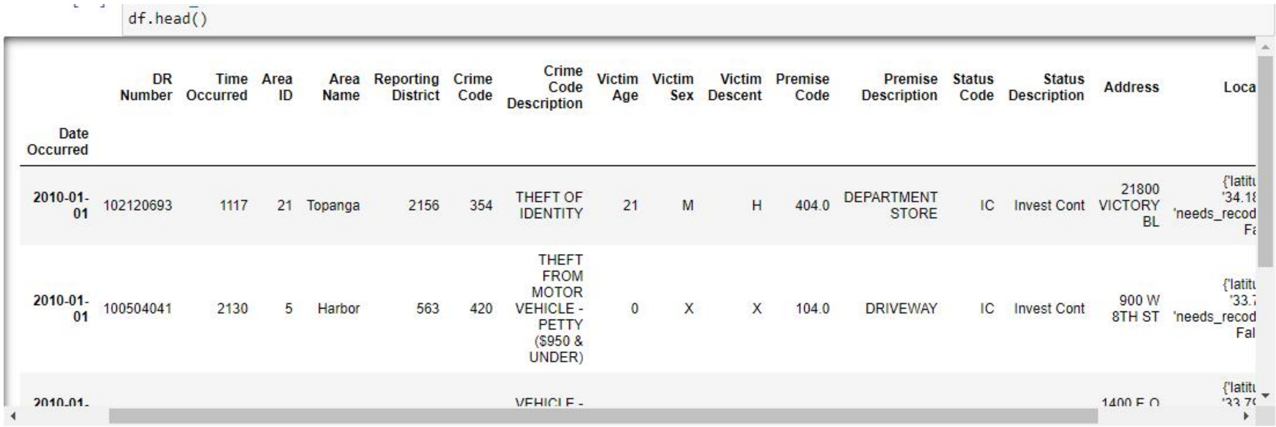
5 rows × 26 columns

Figure 5: snapshot of raw dataset

The various columns of the data set are the time and Date reported, Date occurred also with the Time occurred. It also insures the Area ID with the given Area Name of that particular region weapon description, codes of crime, longitude, latitude, etc.,

Firstly the date occurred is converted to Time Series index by removing the non-valued time stamp from the columns date reported and date occurred and later this is converted to datetime object. All the null valued attributes are scanned and removed since it will cause inconsistency in the dataset and all the columns having the nan values or missing the data are dropped. The size of dataset is reduced to 82191 rows and 16 columns. To reduce the redundancy in the dataset dictionary function is used , this function returns a detailed list of unique values by taking two attributes as input.

The cleaned data is converted to a Pickle catalouge, Pickle is primarily utilised in Python for the purposes of serialising and deserializing a object which belongs to the python structure. Other way to call it out is the procedure which converts a object of the Python by the conversion in byte stream for the purposes of storing it in a file or database, preserving the state of the programme across several assembly, or transporting one's data from the use of certain network. Unknoting the byte stream that is already pickled hence give permission one that reforms and help the OG hierarchy of the certain object. This is possible because the stream was pickled in the first place. Figure 6 below shows the snap of cleaned data set used.



```
df.head()
```

	DR Number	Time Occurred	Area ID	Area Name	Reporting District	Crime Code	Crime Code Description	Victim Age	Victim Sex	Victim Descent	Premise Code	Premise Description	Status Code	Status Description	Address	Loca
Date Occurred																
2010-01-01	102120693	1117	21	Topanga	2156	354	THEFT OF IDENTITY	21	M	H	404.0	DEPARTMENT STORE	IC	Invest Cont	21800 VICTORY BL	{'latitu '34.1 'needs_recod Fa
2010-01-01	100504041	2130	5	Harbor	563	420	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	0	X	X	104.0	DRIVEWAY	IC	Invest Cont	900 W 8TH ST	{'latitu '33.7 'needs_recod Fal
2010-01-							VEHICLE -								1400 E O	{'latitu '33.7

Figure 6: snapshot of cleaned data

The null valued attributes are removed and data is converted to time series index since the non-valued time stamps are removed.

4.3 Methods used for Forecasting

4.3.1 Auto regressive Model

If a statistical model makes predictions about future values based on the values of the past, then the model is said to be autoregressive. For instance, an autoregressive model might attempt to forecast a stock's future pricing by analyzing the company's historical data to determine how well the stock has performed in the past.

The fact that autoregressive models are built on the premise that the values of the past have an effect on the values of the present is one of the reasons why this statistical method is so popular for use in the study of natural phenomena, the economy, and other types of processes that evolve over time. The predictions of multiple regression models are based on the linear combination of predictors, whereas the predictions of autoregressive models are based on the combination of the variable's historical values.

An autoregressive process is said to be AR(1) if the current value is determined by the value that came directly before it, while an AR(2) process is defined as an autoregressive process if the current value is determined by the average of the two values that came before it. White noise is generated via an AR(0) process, which does not include any dependence in the two defined variables we are comparing. For the summation of data to provided variants, in that too there is a great alternative methods for calculating the coefficients that are utilised in these computations, such as the technique that uses the least number of squares.

4.3.2 Moving Average Model

It is a type of computation which is involved in the process data point analysis. This calculation works by generating a summation and denomination with the total number of the subject provided in it for several subsets as the entire set of our data. MA has some type of indicator in the stock that is frequently utilised in the analysis of the technical subject within its realm of finance. When summarizing its average price of that particular stock, the goal being assist in that process of smoothing out the price data by developing an average price that is constantly being updated. When analyzing the rate for the company for a certain amount of time, again summarizing the average of its moving model can help lessen the effects of some of the given short-term also the random swings in the price of the stock. This helps in finding the levels where the stock takes support and where it shows resistance. By this an investor gets the idea of buying, selling or holding the stock.

The lag in the moving average is proportional to the length of the time period used to calculate it. As a result, a moving average of 180 days has to be greater or or higher lag degree from that moving average of 21 days due to the fact that it includes prices over the previous 180 days.The more recent the data that was utilised to calculate the average price, the more responsive it will be to shifts in market conditions. When looking at the average, the larger the time range, the less sensitive it will be.

When calculating moving averages, several time periods with differing lengths can be used, depending on the trading objectives they wish to achieve. Moving averages with shorter time horizons are more common in oer day analysis, whereas moving averages with longer time horizons are utilised more frequently by long-term analysis.

If a security is in an uptrend, the moving average should be increasing, whereas if it is falling, the moving average should be falling, indicating that the security is in a downtrend. A rising verge,has been provided with shorter-term that is now have risen above the LTM average , is another indicator of upward momentum. This crossover occurs when the STMA crosses and than the LTMA. In contrast, the momentum of the downward’s is reinforced by a falling verge, that happen to be at it when a STMA crosses below a LTMA. In this case, the shorter-term moving average is below the LTMA.

Types of Moving Average:

1. **Simple Moving average**: It has to be computed by the input that is taken as a mean of arithmetic computation of a certain set of utility for the predetermined amount of time. This form of the moving average is also the most common type. To put it another way, a group of numbers also the rate of those instruments which are financial, some of them are summed up, also that the total is differentiated by the total amount of rate of this group. Now following is the method that must be used in order to calculate the simple moving average of a security:

$$SMA = \frac{A1 \quad +A2 \quad +\ldots+An}{n} \qquad ..(1)$$

   Where, An = average in the time period n

   n = number of time periods.

2. **Exponential Moving Average:** It lends strengthening significance for more nearby data in an effort so that the one we make has moving average higher sensitive new data.

$$EMA_t = \left[V_t \times \left(\frac{s}{1+d}\right)\right] + EMA_y \times \left[1 - \left(\frac{s}{1+d}\right)\right] \qquad (2)$$

   Where, EMAt =EMA today

   Vt = Value on the day t

   EMAy = EMA a day before EMAt

   s = smoothening

   d = number of days observed

4.3.3 Auto Regressive Integrated Moving Average (ARIMA)

ARIMA model is a technique that has forecasting or envisaging future outcomes by analyzing historical time information. Than being predicated as statistics of serial correlation being an idea, also provide an result that previous data can influence data point of the future too.

One type of regression analysis is called an ARIMA model. This model evaluates the significance of a variable which is being dependent on some other relation to some other variable and to change the subject of that particular model. This model has certain purpose which is to have future movements as a forecast as its securities or can also be states as the discrepancies of that particular model using financial marketbetween the values for a series and not the values which are actual for it.

For a better understanding of the ARIMA, we study every section of the model as shown below in the respective order:

Autoregression (AR): it is a model which helps us to change its variable and it then regresses the values which are already owned in past. This can also be expressed as "lagged" values.

Integrated (I): It indicates that the data are unchanging. The term "stationary data" refers to time-series data that has been "stationed" by removing the observations from the prior values in order to make the data more consistent.

Moving average (MA): automates and sum all the dependency which is an observation also with the error of the residual, when applied the average mover mode to observations of lagged one’s. here we have some type of model which is append for the observations of its lagging.

The parameter in ARIMA are as follows:

- The parameter p: The number of autoregressive terms is denoted by the parameter p, which is also referred to as the amount is "lag observations." It provide us with"lag order," and it impacts its result is being model by delivering data points that are lags behind the current time.
- The value of the parameter denoted by the ‘d’ is referred to as the degree of differencing. It gives an indication of the number of times that the lagging indicators have been eliminated for the stationary data completion.

- The errors in the model's forecast is represented by the parameter q, which is also referred to as the size of the window used for the moving average.

For instance, both the number of terms and the kind of the attributes are accounted for in linear regression models. A value of zero, which is acceptable for usage being the parameter, also it will indicate the specific ingredient has to be utilised for the model. This one is just an AR, I, or MA model they can produce it in the same way that an ARIMA model can be constructed to accomplish the same job.

The fact that ARIMA has founded at price of presumption of the historical utility continue to exert influence on more recent or upcoming values is something that must be taken into consideration.

The autoregressive integrated moving average (ARIMA) model combines characteristics of moving averages with those of the autoregressive model.

AR(1) referes to the model where the present value is related to the value that occurred just before it, but the AR(2) process is known where, present value is dependent on the values that occurred both before it and immediately before it. A computation called moving average is applied for examining data instances by first producing a series of many average of various sub-sets of the whole data provided. The goal of this calculation is to reduce the impact of data points that are considered to be outliers. When producing their projections, ARIMA models are able to consider various components of time series and other non-static categories of data because of the combination of approaches that they employ.

Using the data of time series for the associated variable which are gathered on time basis, ARIMA forecasting can be accomplished simply entering in those numbers. In order to determine whether or not the data are stationary, this analytical program will determine the optimal amount of delays or how many times differencing is applied on data, this will do the analysis. The findings will then be output, and in many cases, they are interpreted in a manner that is analogous as compared to different models of linear.

4.3.4 SARIMAX

ARIMA model has been upgraded to become the SARIMAX model, which stands for “Seasonal Auto-Regressive Integrated Moving Average with eXogenous variables.” The difference between SARIMAX and ARIMA is that SARIMAX comprises of eXogenous components and seasonal impacts where as ARIMA is only autoregressive integrated moving average which does not consider the components considered by SARIMAX, along with the AR and MA component. Hence, we are able to conclude that the SARIMAX model is a seasonal equivalent model just like the SARIMA and Auto ARIMA models.

This model is seasonal equivalent keeps the seasonal rhythm, and it can also deal with the effects caused by external parameters. This aspect of the model is distinctive in comparison to other models. In a time series, the temperature, for instance, exhibits seasonal influences like lower values during the cold and higher values in hot months. Even yet, the temperature in the winter is raised as a result of the influence of external elements such as humidity, and there is also a possibility that the temperature will be lowered as a result of rain. If these factors do not exhibit any cyclical or seasonal behaviour, then we are unable to make an accurate prediction regarding their values. Other models are unable to process this kind of data since they were not designed to do so.

In the parameter for the SARIMAX models, we are required to offer two different sorts of orders. The first one is comparable to the ARIMAX model (p, d, and q), and the second one is to indicate the effect of the seasonality; we refer to this order as a seasonal order, and inside it, we are needed to submit a total of four numbers. (Specification of the Seasonal AR, Order of the Seasonal Integration, Seasonal MA, and Periodicity of the Seasonal) Using mathematics, we are able to represent the model in this way.

4.3.5 Facebook Prophet Forecasting model

The Prophet is a forecasting model for time series data that may take into account yearly, monthly, and daily seasonality, as well as the effects of holidays. Prophet is the most effective model for forecasting when the data is time sensitive and has a significant history of seasonality [10]. The documentation for Prophet that can be found on GitHub states that Facebook utilises prophet for many trustworthy forecasts and because it is resilient to outliers and missing data. Forecasting may be done using both R and Python thanks to the availability of the Prophet API. It is a library that can be installed using Anaconda and is available publicly as it is open source.

One of the most significant distinctions between ARIMA and Prophet is that the latter takes into consideration "change points," also known as particular shifts in the trend of the time series. Even though it is theoretically possible to accomplish this using ARIMA in R, you will need to make use of a different package known as AEDForecasting in order to do so.

The non-linear trends in the series are fitted with the required seasonality in order for Prophet to function properly. This is accomplished through the use of an additive model (whether daily, weekly, or yearly).

From this vantage point, while ARIMA has taken into consideration patterns in the time series such as seasonality and autocorrelation, Prophet has excelled at accurately identifying the crucial change point in the series, and this appears to have boosted the model's ability to make accurate predictions.

We observed fairly clear indicators that this was the case for the dataset that was under consideration, and it's feasible that ARIMA might do very well when projecting data with a more clearly defined trend or one that doesn't really demonstrate seasonality. On the other hand, Prophet might not be appropriate for usage in situations where there is no seasonality involved.

4.3.6 Holt Winters Forecasting

"Triple Exponential Smoothing," is another name for Holt Winter forecasting which is just another name for it. Forecasting time series can be done using the simple or double exponential smoothing method. The use of triple exponential smoothing is preferable when dealing with data that is highly seasonal and trend-driven. According to the Engineering Statistics manual, previous observations carry the same amount of weight in single moving averages. On the other hand, bigger value weights are given to more recent data in the exponential model. This provides a concise summary of the fact that older data have comparatively less weights compared to newer observations [18]. The same idea as Single Exponential Smoothing may be denoted as

$$S_t = \alpha y_{t-1} + (1-\alpha)S_{t-1} \qquad 0< \alpha \leq 1 \ t\geq 3. \qquad ..(3)$$

Smoothing observation is denoted by the Si,

While original observation is denoted by the letter y.

α is known as the smoothing constant.

Sub-scripts are time frames from 1 to n.

Simple Exponential Smoothing method is used for time series data that lacks both a trend and seasonality. In order to account for patterns that emerge in the data, the equation needs to contain two constants. The two equations that make up the double exponential smoothing technique are following:

$$S_t = \alpha y_t + (1 - \alpha)(S_{t-1} + b_{t-1}) \quad 0 \leq \alpha \leq 1 \tag{4}$$

$$B_t = \gamma(S_t - S_{t-1}) + (1 - \gamma)b_{t-1} \tag{5}$$

The smoothing constants are denoted by α and γ

The first equation makes a direct adjustment to St for the pattern of the past periods by merging the trend with the value that was smoothen out from the previous period.

This helps to get rid of the lag, and as a result, St is now an accurate representation of the present values. Equation 5 helps to improve the pattern, that can be seen by looking at the variation between most recent values and its previous values. The inclusion of the trend in this equation marks a significant departure from the conventional single exponential smoothing [18].

4.3.7 Random Forest

Random Forest algorithm is a Supervised ML technique which has found widespread application in the areas of both classification and regression. It constructs decision trees using a variety of samples, then uses the majority vote of those samples to determine classification and uses an average to determine regression. Random Forest Algorithm can handle the dataset which contains non-discrete variables which is a regression task and categorical variables which is a classification task. This capability is what makes the Random Forest Algorithm unique and impactful.

In regression problems, the actual variance is calculated, which is the variance of the true output from the predicted output. If the expected outputs of the tree are too far away from the predicted output, then the tree is considered to be overfitted because it has a high variance. When working with classification problems, a high generalisation error is calculated, and if this error is high, we are able to conclude that the model has been overfitted because it does not perform well when applied to the test dataset.

Bagging is a technique that is used to reduce the variance problem, and as a result, prevent the overfitting of models that was described earlier in this paragraph. The bagging method, which is used to create multiple copies of data with replacement, includes the bootstrapping process as one of its components. Another name for this technique is the resampling method of data.

Because in Random Forest we are training a forest of Bagged D-Trees rather than a single D-Tree, this technique is utilised to address the problem of high variance. In addition to having all of the advantages that D-Trees offer, Random Forest is also able to process different kinds of datasets without the need for any preprocessing. When it comes to the task of splitting, Random Forest uses a method that randomly selects a subset of features rather than all of the features. Because of this, the Tree will look different due to the algorithm's evaluation of a fewer number of outputs.

4.3.8 Logistic Regression

The purpose of logistic regression is to find out the nature of the association that exists between a categorical dependent variable and one or more continuous independent variables. The dependent variable in this case is categorical, and the independent variables are typically continuous. This is accomplished by plotting the probability scores of the dependent variables. A categorical variable is a variable that, as opposed to being continuous, can only take on values that fit into certain predetermined categories.

The results of categorical variables, such as binomial and multinomial values of y, can be predicted using logistic regression. It is a statistical method that is widely utilised for forecasting binary classes and computing the likelihood of an event taking place or a choice being made. For instance, a company might be interested in determining the likelihood that visitors of varying demographics will take advantage of a particular promotional offer posted on their website (dependent variable). In this situation, logistic regression would look at known characteristics of the customers, such as how many times they've been to the company's website (which would be an independent variable) and what other websites they may have come from in order to determine the suitable action. This will be of assistance for the companies developing their process for making decisions concerning promotional content.

Chapter 5

Results

The resultant graphs for hotspot and forecasting are below:

5.1 Hotspot of Los Angeles:

By analyzing the data set used we found the areas of the cities with highest crime incidents that took place which is showed in figure 7 below. We here notice that 77th street, southwest, N Hollywood are top 3 hotspots of Los Angeles.

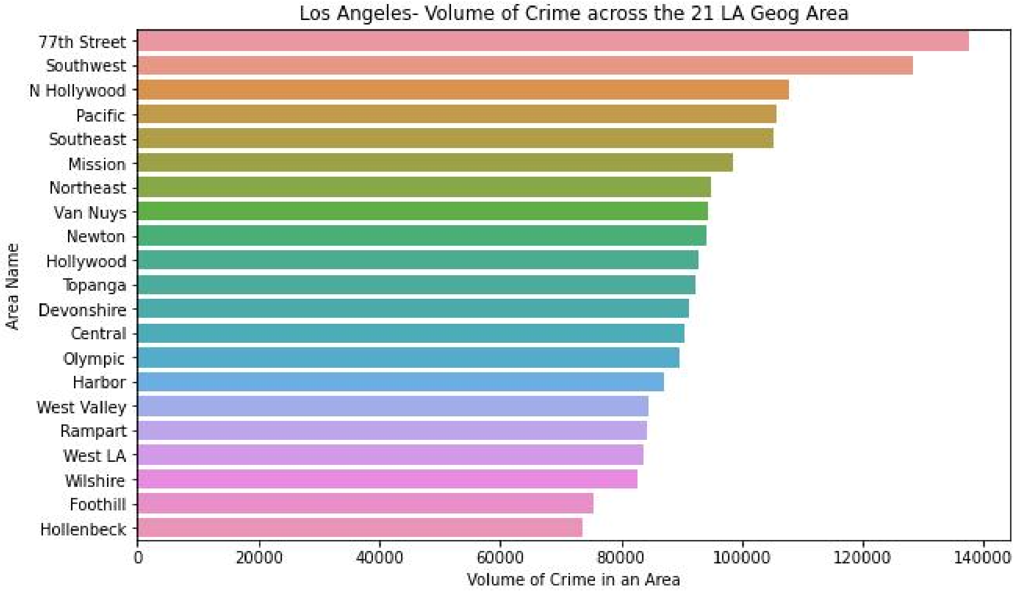


Fig 7: Areas with highest crime incidents

After analyzing the locations of the crime incidents, it is also important to find the premises of the incidents, figure 8 below shows the premises where most of incidents take place

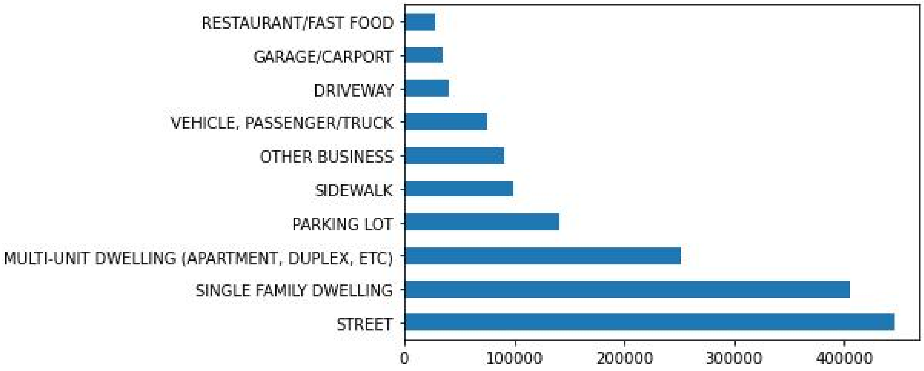


Fig 8: Crime count in the type of Premise

Crime against some particular descents are common among the reported incidents, so figure 9 shows which communities are targeted the most in all areas.

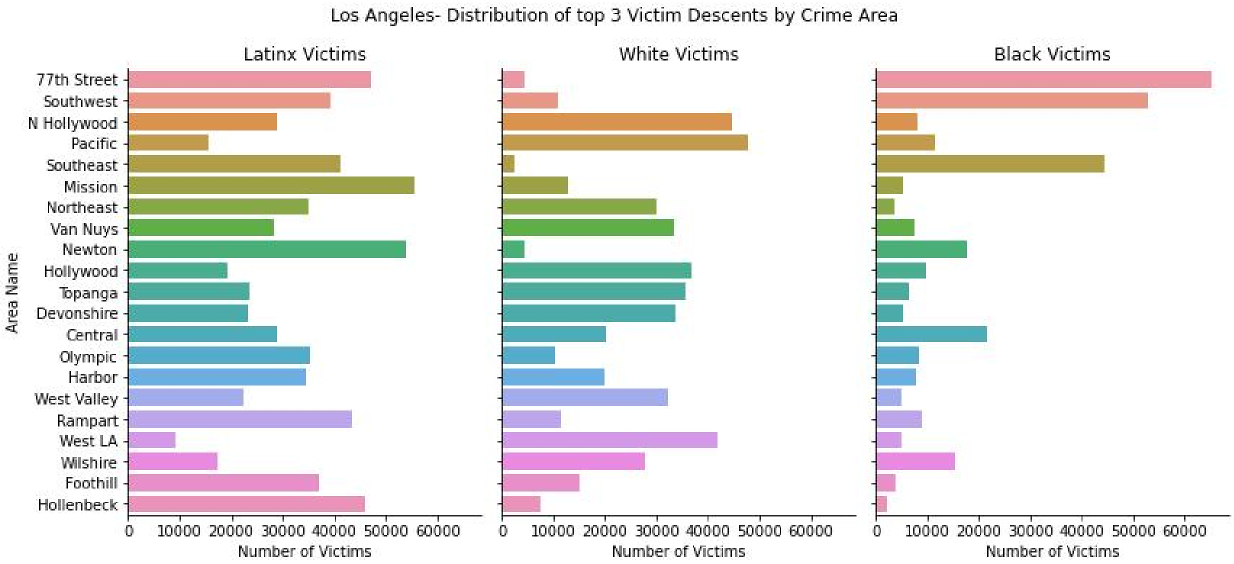


Fig 9: Hotspots based on victim's descent

5.2 Forecasting results

The forecasting graphs are plotted to forecast the crime rate of Los Angeles, figure 10, shows the forecasting results we got from SARIMAX model

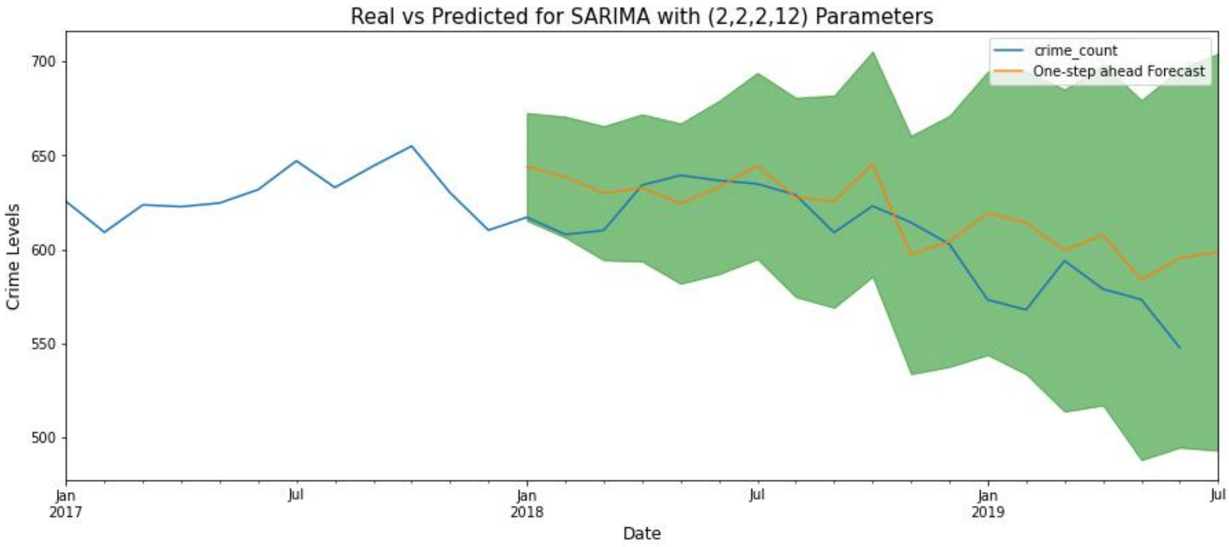


Figure 10: Graph showing future forecast of crime SARIMAX model

Figure 11 below shows the forecasting predictions made by Facebook prophet forecasting model since it also considers the holidays we see the movement in the graph.

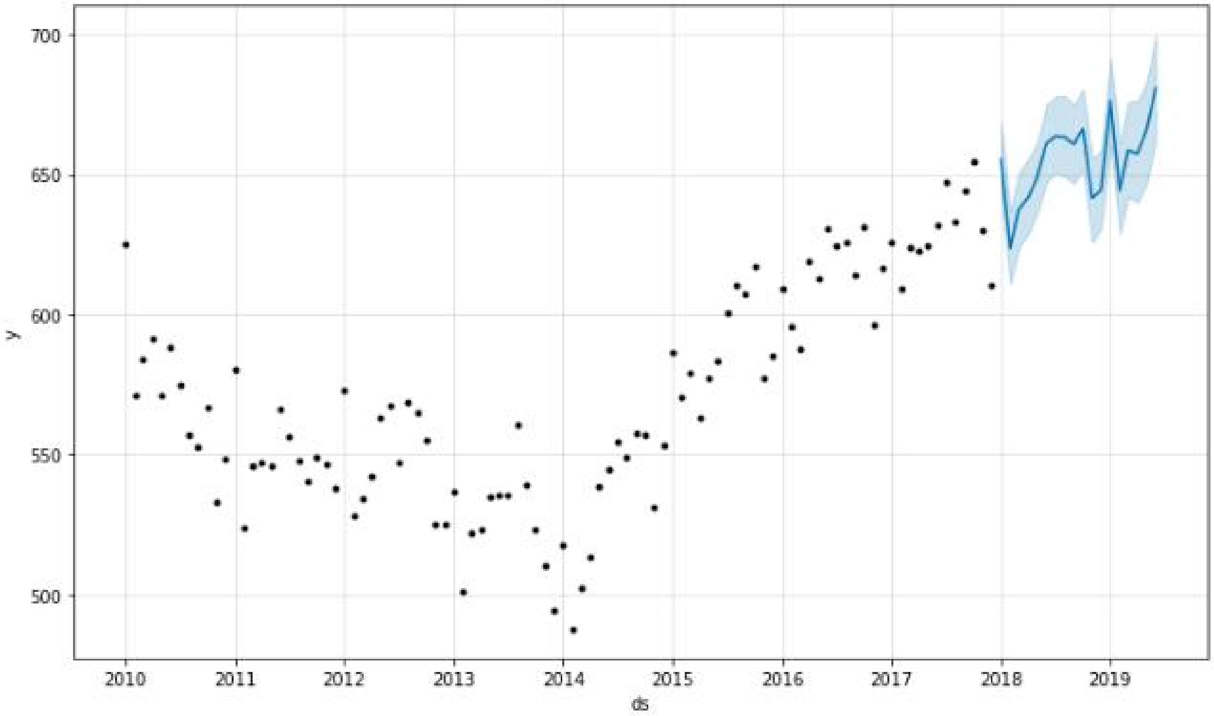


Fig 11: Facebook prophet future forecast

Trend analysis is important in Facebook's prophet model since this model also accounts the holidays, so yearly as well as monthly trend is plotted to see the holiday effect on crimes.

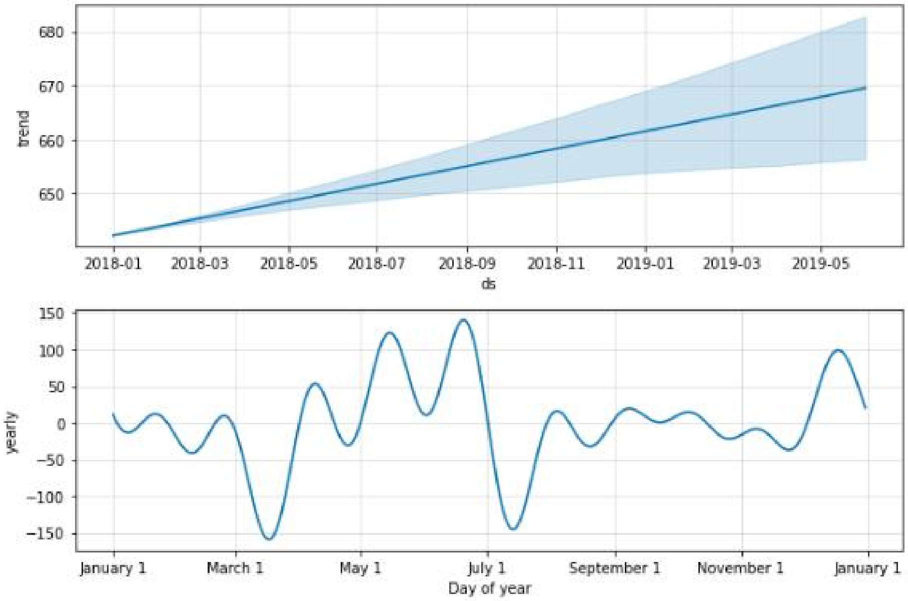
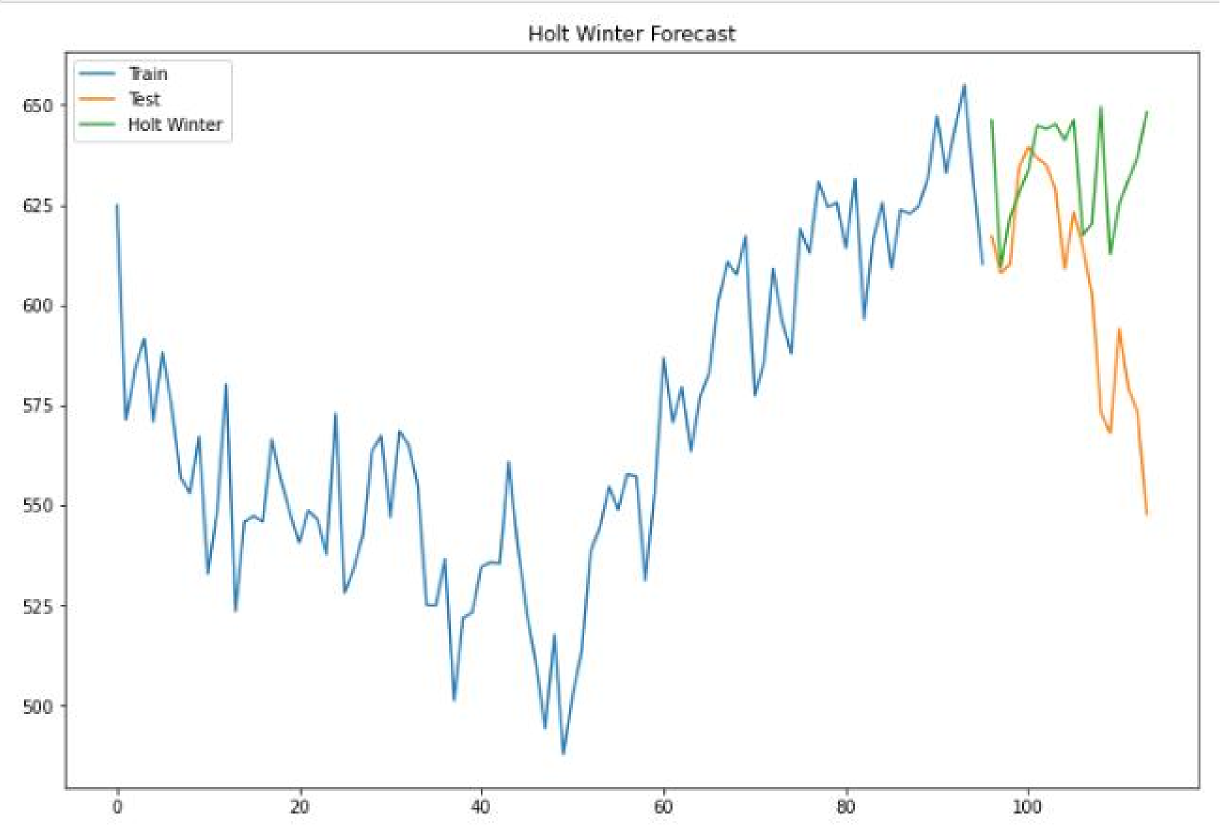


Figure 12: Graphs representing monthly and yearly trends

Figure 13 shows the holt's winter method of forecasting and the result obtained is not very close as we got in SARIMAX and Facebook's prophet model.



```
RMSE = 40.080723295894934
```

Fig 13: Holt's winter forecast

5.3 Error Measurement

Root Mean Square Error (RMSE) : the formula for rmse is as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Where, n=number of observations.$\hat{y}_i$ is the forecasted values, y_i = recorded values.

The table below represents the RMSE of the models

Model	AR	MA	ARIMA	SARIMAX	Holt's winter	Facebook Prophet	Random Forest	Logistic Regression
RMSE	27.53	42.63	23.26	24.65	40.08	60.21	78.27	124.75

Table 2: RMSE of the models

CHAPTER-6

Conclusion and Future work

In this work we have identified the areas where crime incidents are higher in Los Angeles and are identifies as hotspot since the crime count in those areas is higher than other localities. The 77th street, southwest and N Hollywood are among the top 3 hotspot of Los Angeles. For forecasting purpose, we have used various models like auto regressive (AR), moving average(MA), Auto regressive integrated moving average(ARIMA), Seasonal Arima with eXogenous variables(SARIMAX), Facebook's prophet, holt's winter, logistic regression and random forest for forecasting the crime and calculated the root mean square error and observed that traditional ARIMA model has given better performance and logistic regression is not suitable with the used dataset. We can see that the ML methods of forecasting are not surpassing the statistical methods of forecasting of crime. ARIMA in this case is giving the best results.

The future work which can be extended to this is the database of available police resources such as the number of officers, police personnel, vehicles. This will help in deploying the best resources according to type of incidents and the risk level.

REFERENCES

[1] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.307811

[2] Kim, Suhong & Joshi, Param & Kalsi, Parminder & Taheri, Pooya. (2018). Crime Analysis Through Machine Learning. 415-420. 10.1109/IEMCON.2018.8614828.

[3] M. Tayebi, M. Ester, U. Glasser, and P. Brantingham, “Crimetracer: Activity space based crime location prediction,” in Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on, 2014, pp. 472–480.

[4] L. McClendon and N. Meghanathan, “Using Machine Learning Algorithms to Analyze Crime Data,” Machine Learning and Applications: An International Journal(MLAIJ), Vol.2, No.1, March 2015. [Online]. Available: http://airccse.org/journal/mlaij/papers/2115mlaij01.pdf [Accessed 10-March-2019].

[5] Chandy, Abraham, "Smart resource usage prediction using cloud computing for massive data processing systems" Journal of Information Technology 1, no. 02 (2019):108-118.

[6] G.Wilpen, O Andreas and T.Yvonne,”Short-term forecasting of crime,”International journal of forecasting”, Volume 19, Issue 4, pp. 579-594, October–December 2003.

[7] K. Al-Janabi, “A proposed framework for analyzing crime data set using decision tree and simple k-means mining algorithm,” Journal of Kufa for Mathematics and Computer, vol. 1, no. 3, pp. 8–24, 2011.

[8] M. L. Young and A. Hermida, ``From Mr. and Mrs. Outlier to central tendencies: Computational journalism and crime reporting at the Los Angeles times," *Digit. Journalism*, vol. 3, no. 3, pp. 381_397, May 2015.

[9] C. Contreras, ``A block-level analysis of medical marijuana dispensaries and crime in the city of Los Angeles," *Justice Quart.*, vol. 34, no. 6, pp. 1069_1095, Sep. 2017..

[10] Forecasting at scale. [Online]. Available: https://facebook.github.io/prophet/

[11] C. B. Dierkhising, D. Herz, R. A. Hirsch, and S. Abbott, ``System backgrounds, psychosocial characteristics, and service access among dually involved youth: A Los Angeles case study," *Youth Violence Juvenile Justice*, vol. 17, no. 3, pp. 309_329, 2018.

[12] M. R. D'Orsogna and M. Perc, ``Physics for better human societies: Reply to comments on `statiscal physics of crime: A review,''' *Phys. Life Rev.*,vol. 12, pp. 40_43, Mar. 2015

[13] T. Almanie, R. Mirza, and E. Lor, ``Crime prediction based on crime types and using spatial and temporal criminal hotspots," *Int. J. Data Mining Knowl. Manage. Process*, vol. 5, pp. 1_19, Aug. 2015.

[14] S.N. Christian, K. R. Majeed, and S. O. Etinosa, ``Application of data analytics techniques in analyzing crimes," in *Proc. SAIS*, vol. 40, 2018, pp. 1_7.

[15] E. Cesario, D. Talia, and A. Vinci, ``A data-driven approach for spatio-temporal crime predictions in smart cities," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Taormina, Italy, Jun. 2018, pp. 17_24.

[16] Rupa Ch, Thippa Reddy Gadekallu, Mustufa Haider Abdi and Abdulrahman Al-Ahmari, “Computational System to Classify Cyber Crime Offenses using Machine Learning”, Sustainability Journals, Volume 12, Issue 10,Published on May 2020.

[17] H. Wang, H. Yao, D. Kifer, C. Graif, and Z. Li, ``Non-stationary model for crime rate inference using modern urban data," *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 180_194, Jun. 2019.