# Crime Data Visualization and Forecasting Using SARIMA Algorithm

**[1]Safiya Mohd, [2]Hithasri Muthyala**

[1,2]Assistant Professor
Computer Science and Engineering (Networks)
Kakatiya Institute of Technology and Science, Warangal, Telangana, India

***Abstract* - Machine learning is the ability to perform specific tasks by using past experience data while measuring the performance, upgrading knowledge and skills by a system for developing the software application. SARIMA (Seasonal Auto regressive Integrated Moving Average) algorithm is a technique to forecast and predicts the approaching data. SARIMA contains seasonal and trend modules in which, it can support a univariate of seasonal module. The year wise crime data set is used to implement and visualize the crime rate in each district from each state. To visualize and analyze the crime data with the help of graphical representation of time series graphs as heat map, trend line and column charts. Crime data is mainly used to reduce the crime rate and predicting the different crimes in a particular location. It is used for police departments, law and social control agencies to analyze the past crime incident data and to reduce the crime activities in the country. The crime data analysis is used for government agencies to speed up the process of finding the incidents in the society and speedup the process with help of time series algorithms. SARIMA model gives better efficiency to predict the crime data to solve problems in easiest way and to reduce the crime rate in the country.**

***Index Terms:* Visualization, Forecasting, Prediction, SARIMA, ARIMA, Prophet.**

## I. INTRODUCTION

Crime data analysis and visualization is mainly helpful for police departments, law enforcement agencies and residents of the cities in many ways. Analyzing the crime data to reduce the crime rate and predicting the various crimes in a particular location based on the past criminal data throughout the country.

Crime is a violation of humanity, often punishable by law. Criminology is a study of crime, interdisciplinary science that investigates and analyzes the performance of criminal data. Criminal activity is very high in various locations. The main responsibility is to reduce the crime rate in the society with the help of visualization and forecasting by police, law enforcement and government agencies. These government sectors are help to hurrying up the process of solving the crimes. The computerized systems, crime data analysts can help to reduce the crime and finding the crime in less time. Crime data forecasting and prediction is the systematic analysis for identification and predicting the structure, seasonal and trended data elements in crime. The prediction is not done by individual victims of crime but can predict the place that has a probability of its occurrence.

The primary goal of this approach is to study and predict the crime rate which can be helpful to investigators, police and law agencies to solve the crime problem from the large amount of information that is stored previously in the database. It focuses on creating a model that helps to detect the number of crimes by its type in a particular in a particular state. The scope of the paper is to prevent crimes and control the crime activities which may occur in the future. Crime analysis is indirectly helps to reduce the crime rate in India and helps to improve the securities in such areas or locations.

Crime data visualization and forecasting is a time series technique in machine learning and various visualization techniques and plots are used which can helps to law, social control and police departments which is to detect and predicts the crime data with higher efficiency and accuracy.

Time series forecasting methods like SARIMA, ARIMA (Auto Regressive Integrated Moving Average) and PROPHET. are employed for crime data visualization. Comparing 3 time series algorithms for the best efficiency to predict and forecast the crime rate and to speed up the process for reducing the crime rate in every area in India.

## II. LITERATURE REVIEW

Mingchen Feng et al. [1], proposed the Crime Data visualization and forecasting analysis with big data technology. The exploratory data analysis is used for forecasting and predicting the crime data with trend element. The use of visualization is to forecast with past data. and trends. The results shows that the future prediction data performs better efficiency with Prophet and LSTM models. Comparing with different neural network models the Prophet and LSTM gives best performance measures.The best quantity of three years data can be trained the in a way to make the better prediction of trend data in position of RMSE(Root Mean Square Error) and spearman statistics. The results explained to provide new visual sense into crime trends and predictions. It uses to make a best decision for both police and law departments to control the crime rate.

Abish Malik, et al. [2], implemented the forecasting method which is supported as with STL (Seasonal trend decomposition based on loss). It can be applied in a spatiotemporal for visual analytics context ,which can be provides for an analysts with predicted levels of future predicted data of such type crime activity. This can be mainly predict the future crime data and analyse the past data to reduce the crime data in each sector. It is mainly used for various investigation sectors, crime bureau investigation departments,

police sector and law enforcement agencies with a natural scale templates and methods of a framework. It provides the focusing and drill down approach which is suitable for geospatial levels and temporal resolution levels. Theprocess for predicting the recent incidents of crime activities is guided by the spatial correlation at close locations using the technique as kernel density estimation.

Abdulrahman Takiddin et al. [3], Proposed the model of embedding system to determine the cyber attacks on electricity thefts in deep recurrent vectors. The data representation method is a model of vector embedding The method is used to express in terms of consumption electricity profiles can be represented in terms of of actual numbers. According to the electricity readings may be difficult to analyze and capturing the patterns with in the reading of customers report. A ordered grid search excitable parameter improvement algorithm is to change the models detection performance of electricity cyber thefts. The average performance metric uses the ISET dataset that gives better performance with 96% DR (detection rate),21% FA (false alarm), and 94% HD (highest difference). The usage of vector embedding on with GRUs gives better performance with 35%–97% is DR, 31%–10% is FA, and 87%–22 is HD%.

WAJIHA SAFAT et al.[4], proposed with deep learning and machine learning models for predicting and forecasting the data of crime incidents. The current calculation of the crime data, variety and hotspots from historical techniques that to create several computational challenges, problems and chance. The main effort of this research is needed for police agencies and crime investigation departments and so on. The requirement is available to improve the predictive algorithm, which directs the police force for the criminal activities. This research study is mainly applied on various machine learning models based on supervised and unsupervised algorithms and time series models as by (Long-Short Term Memory)LSTM and ARIMA model . Among all of those algorithms the ARIMA algorithm is the best fitted model for the crime data set. The execution and the evaluation of LSTM algorithm is to analyze the equal to the root mean square error (RMSE) value and mean absolute error value (MAE) performance measures.

MingJian Tang et al. [5] 2019, proposed the existence of volatility clustering effect implemented with Auto regressive conditional heteroskedasticity (ARCH) and Generalized Auto regressive conditional heteroskedasticity (GARCH) models. TheGARCH model shows the effective multivariate time series datato understand about thechallenging dependency structures. The vulnerability management is more compare to reactive.The statistical analysis models is purely depends on vulnerability and that can be provide the essential element to organizations. It helps them to becomes more active in the management of crime rate and cyber crime incidents. The domain of the real world cyber security problemsare exploiting vulnerability revealing the trends with complex multivariate data of time series. The computational historical vulnerability of cyber data uses a proposed a new implementation and new framework that has alter this new capabilities and challenges to solve the complex problems. The framework uses a case study for handling the continuity unpredictability of data and also used the purpose of multiple variable dependency structures among various liable crime activities attempts. The more general multivariate time series data is to capture their intriguing relationships.The Big Data systems are new in trends which are becomes more attack targets by already proposed and rising threat agents.

Hongjian Wang et al. [6] implemented an AutoRegressive network Models of the non parametric sparse additive for crime data reduction in Chicago country. The applications that are developed with different technologies using the mobile devices for collecting the huge amount of data. The data is collected from urban areas using Big data techniques for understanding and analyzing the crime data.In this research the city of Chicago id used because for thehigh scalability of measurement data and machine data. To identify geospatial location for non-stationary belongings to utilize the geographical adjusted regression on upper layer of the negative binomial model (GWNBR). This implementation performs better on negative binomial model. The $\beta$ and $\varphi$ are the attributes of Markov chains process to observe themethodshas to make the crime rate and incidents of Chicago country the kernel hilbert spaces (RKHSs) approach is used. It provides the statistical experiments that can support the hypothetical outcomes and display the benefits of SPAM framework using non-parametric for a Chicago crime data set.

Suman Rath et al. [7],proposed the cyber attacks on different types of cyber attack like false information solution, flooding the targets with traffic (DoS) and repeat attacks are taking the action of event with the activity of prospective vulnerabilities. The analysis had been carried out in a micro grid operative tool in a way where the connection can be used for crimes in a real world scenario system thorough the (Controller Area Network) CAN connection. The software simulation that are performed in MATLAB environmental area and actual methods. The proposed work can be done with the test bed which can has simulators and physical CAN equipment to form a bus connected links. The approach is proved with various times of attacks on a simple one link an also the cases of simultaneous attacks can performed with many links. In CAN network two transceivers are available, onetransceiver is for transfers and receives the data where it may connect to the Arduino.Second transceiver is used as a harmful system can connected to a Raspberry pi, directs to the networks that can perform the attacks.

Sobia Khalid et al. [8], implemented a framework for crime incident analysis on fuzzy logic sets on formal sociality concepts and criminal data can begrouped in to the mapped data. To ensure that sociality concepts can be unevenly defined based on a set of property that are extracted from the criminal records to reduce the crime rate. The quality of the results is verified and evaluated with panel of experts in criminology and sociology scenarios for thinking of the crime act. The methodology is used to help in law departments to identify the features of various crimes. This research is helpful for identifying the different crimes in society and reduces the crime rate in society in a particular locality. The collection of crime data contains the crossfunction and clustering structure. Cross function mapping of crime incidents and happening to sociality concepts which can be performed byfuzzy logic. Clustering is performed by the tasks of crime incidents to social and law concepts such as data preprocessing, converting the social concepts into structured form. Clustering techniques are natural cluster technique such as dbscan algorithms for preprocessing the data and fuzzy clustering algorithm for predicting the crime data.

Priyanka Das et al. [9], proposed a clustering technique from machine learning approach. It is an approach of unsupervised learning in machine learning which can extract each and every matter from newspaper and articles based on criminal data. The identification and observations of the crime patterns which can helps in crime rates and criminal judiciary system using the

clustering techniques. The group of clusters are most frequent words, simply it may take place that few of the context words are existed in the gathered data which cannot reflect to the same crime data as of the clustered. In such conditions collecting different words with defining the same meaning. The internal cluster evaluation can be performed by Infomap and Fast greedy clustering algorithms using real data sets. The methodology that provide a brief description about the crime activities to explore the type of crime patterns. It will helps to the law enforcement and police departments to examine the crime incidents at a quicker rate.

Meetha V. Shenoy et al. [10],planned a structure on crime effect, analysis, and with intensity levelsfor women security and safethrough the professional bodies and social group predictions which are designed and tested. The requirements are tested with functional reliability and usability of the model. It continuously check for secure and stress before thetime interval of the last preparation is done by the public. In implemented system that can performs the efficient crime investigations and generating of data to plan the preventative performance and evaluations against crime. The field of study will not be effective in providing by time but it can helps in different law social control agencies and personal linkages are concerned in recovery and outcome. The Integrated system can has a factor as Web GIS which can include the geographical area database for storing and retrieving the criminal records and hot spot networking analysis, and visualization of each victim. In another way the mobile application is used for increasing alerts and tracking of the person and locations who are in the risk zone.Observing the crimes with hotspots in the location and to enable for tracking the persons and areas using the precautions to reduce and alerts for the specific crime.

## III. IMPLEMENTATION

In this implementation section, forecasting and visualization of crime data analysis with different time series algorithms using machine learning technology includes as SARIMA, ARIMA and Prophet where the algorithms are used to performance with best accuracy and measures of predicated values while training and testing the datasets.

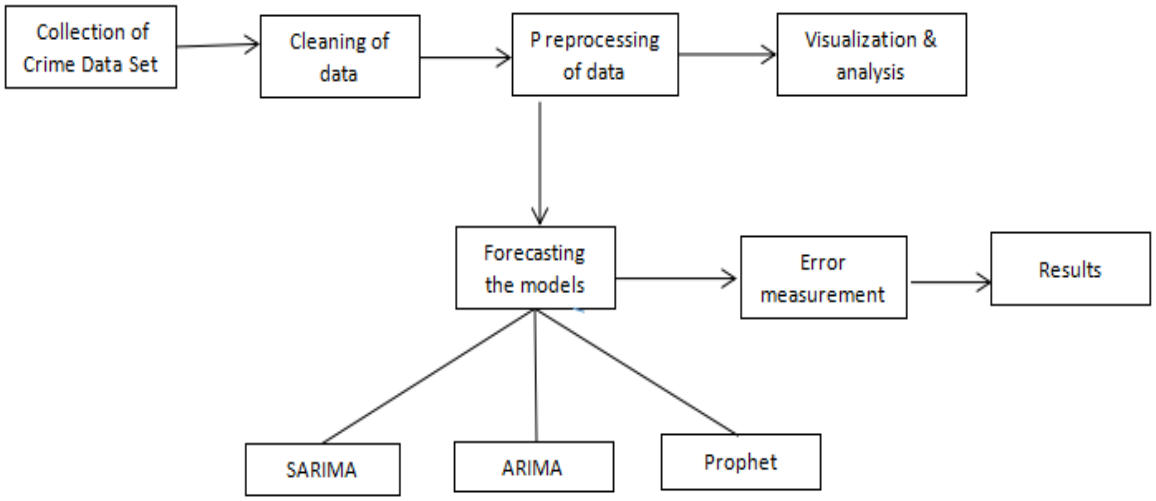


Figure 3 Architecture of time series model evaluation

### 3.1 Data collection

Data collection is a process to perform the specific task in machine learning model. The specific datasets contains unnecessary information so first need to preprocess the data and obtain perfect dataset for the algorithm.

| | S. No | Category | State/UT | 2016 | 2017 | 2018 | Percentage Share of State/UT (2018) | Mid-Year Projected Population (in Lakhs) (2018)+ | Rate of |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | State | Andhra Pradesh | 616 | 931 | 1207 | 4.40 | 520.30 | |
| 1 | 2 | State | Arunachal Pradesh | 4 | 1 | 7 | 0.00 | 14.90 | |
| 2 | 3 | State | Assam | 696 | 1120 | 2022 | 7.40 | 340.40 | |
| 3 | 4 | State | Bihar | 309 | 433 | 374 | 1.40 | 1183.30 | |
| 4 | 5 | State | Chhattisgarh | 90 | 171 | 139 | 0.50 | 284.70 | |
| 5 | 6 | State | Goa | 31 | 13 | 29 | 0.10 | 15.30 | |
| 6 | 7 | State | Gujarat | 362 | 458 | 702 | 2.60 | 673.20 | |
| 7 | 8 | State | Haryana | 401 | 504 | 418 | 1.50 | 284.00 | |
| 8 | 9 | State | Himachal Pradesh | 31 | 56 | 69 | 0.30 | 72.70 | |
| 9 | 10 | State | Jammu & Kashmir | 28 | 63 | 73 | 0.30 | 134.30 | |
| 10 | 11 | State | Jharkhand | 259 | 720 | 930 | 3.40 | 370.50 | |
| 11 | 12 | State | Karnataka | 1101 | 3174 | 5839 | 21.40 | 654.50 | |
| 12 | 13 | State | Kerala | 283 | 320 | 340 | 1.20 | 350.00 | |

Figure 3. 1. 1 Data collection from different states

| | STATE/UT | DISTRICT | YEAR | MURDER | ATTEMPT TO MURDER | CULPABLE HOMICIDE NOT AMOUNTING TO MURDER | RAPE | CUSTODIAL RAPE | OTHER RAPE | KIDNAPPING & ABDUCTION | KIDNAPPING AND ABDUCTION OF WOMEN AND GIRLS | KIDNAPPING AND ABDUCTION OF OTHERS | DACOITY | PREPARATION AND ASSEMBLY FOR DACOITY | ROBBERY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ANDHRA PRADESH | ADILABAD | 2001.00 | 101 | 60 | 17 | 50 | 0 | 50 | 46 | 30 | 16 | 9 | 0 | 41 |
| 1 | ANDHRA PRADESH | ANANTAPUR | 2001.00 | 151 | 125 | 1 | 23 | 0 | 23 | 53 | 30 | 23 | 8 | 0 | 16 |
| 2 | ANDHRA PRADESH | CHITTOOR | 2001.00 | 101 | 57 | 2 | 27 | 0 | 27 | 59 | 34 | 25 | 4 | 0 | 14 |
| 3 | ANDHRA PRADESH | CUDDAPAH | 2001.00 | 80 | 53 | 1 | 20 | 0 | 20 | 25 | 20 | 5 | 1 | 0 | 4 |
| 4 | ANDHRA PRADESH | EAST GODAVARI | 2001.00 | 82 | 67 | 1 | 23 | 0 | 23 | 49 | 26 | 23 | 4 | 0 | 25 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 9752 | LAKSHADWEEP | LAKSHADWEEP | NaN | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 3. 1. 2 Data collection from different states with different crime activity.

| | Area_Name | Year | Subgroup | Rape_Cases_Reported | Victims_Above_50_Yrs | Victims_Between_10-14_Yrs | Victims_Between_14-18_Yrs | Victims_Between_18-30_Yrs | Victims_Between_30-50_Yrs | Vic |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Andaman & Nicobar Islands | 2001.00 | Total Rape Victims | 3.00 | 0.00 | 0.00 | 3.00 | 0.00 | 0.00 | |
| 1 | Andaman & Nicobar Islands | 2001.00 | Victims of Incest Rape | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | |
| 2 | Andaman & Nicobar Islands | 2001.00 | Victims of Other Rape | 2.00 | 0.00 | 0.00 | 2.00 | 0.00 | 0.00 | |
| 3 | Andaman & Nicobar Islands | 2002.00 | Total Rape Victims | 2.00 | 0.00 | 0.00 | 1.00 | 1.00 | 0.00 | |
| 4 | Andaman & Nicobar Islands | 2002.00 | Victims of Incest Rape | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |

Figure 3.1.3 Data collection from different states with different crime activity.

### 3.2 Data pre-processing

It is the collection of the task related to the dataset based on some target variables to analyze and visualize the data to produce some outcome. However, most of the data may be noisy that should contain incorrect and inaccurate values. Data preprocessing can be done bycleaning of data, transformation of data and selection of data.

| | S.No | State | Districts | YEAR | MONTH | DAY | HOUR | MINUTE | VICTIMS OF RAPE | SERIOUS FRAUD | PROPERTY STOLEN | AUTO THEFT | Cases_under_crime_against_women | Unnamed: 13 | poilce housing | Human_ri |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.00 | Maharashtra (MH) | Thane | 2020.00 | 5.00 | 12.00 | 16.00 | 15.00 | 99.00 | 88.00 | 14.00 | 64.00 | 2.00 | 64.00 | 96.00 | |
| 1 | 2.00 | West Bengal (WB) | North 24 Parganas | 2020.00 | 5.00 | 7.00 | 15.00 | 20.00 | 184.00 | 116.00 | 0.00 | 16.00 | 0.00 | 16.00 | 67.00 | |
| 2 | 3.00 | Karnataka (KA) | Bangalore Urban | 2020.00 | 4.00 | 23.00 | 16.00 | 40.00 | 126.00 | 59.00 | 1.00 | 31.00 | 0.00 | 31.00 | 65.00 | |
| 3 | 4.00 | Maharashtra (MH) | Pune | 2020.00 | 4.00 | 20.00 | 11.00 | 15.00 | 98.00 | 0.00 | 3.00 | 25.00 | 0.00 | 25.00 | 68.00 | |
| 4 | 5.00 | Maharashtra (MH) | Mumbai suburban | 2020.00 | 4.00 | 12.00 | 17.00 | 45.00 | 193.00 | 92.00 | 11.00 | 60.00 | 0.00 | 60.00 | 166.00 | |

Figure 3. 2. 1 Data preprocessing with from states with different crime activity.

### 3.3 Data Visualization

Data visualization and analysis of crime which are represented by pie charts, heat maps, column chart and graphs to predict and forecast the crime data in India.

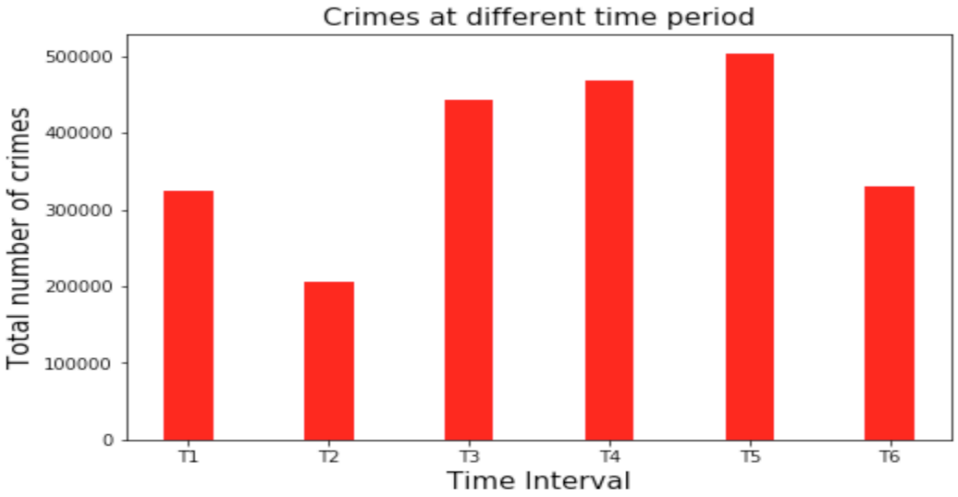


Figure 3. 3. 1 Crime rate visualization of time interval using Bar chart

The above visualization shows the time interval from T1 to T6. Each time interval varies on time period. T1 ,T2, T3 and T4 shows the crime rate at various time periods . T1 shows 4 hours difference to T2, T2 shows the same 4 hours difference to T3, T3 to T4, T4 to T5 and T5 to T6 of 4 hour of time duration from 12 am to 12 am. The analysis with preprocessed data is based on the week data which can be considered as the weekly crime rate in each state. In Python language the importing libraries with the pandas for making week and monthly data which is collected in a easy way on crime data. In python the Matplotlib library is used to create visualizations of time series data with collected data.

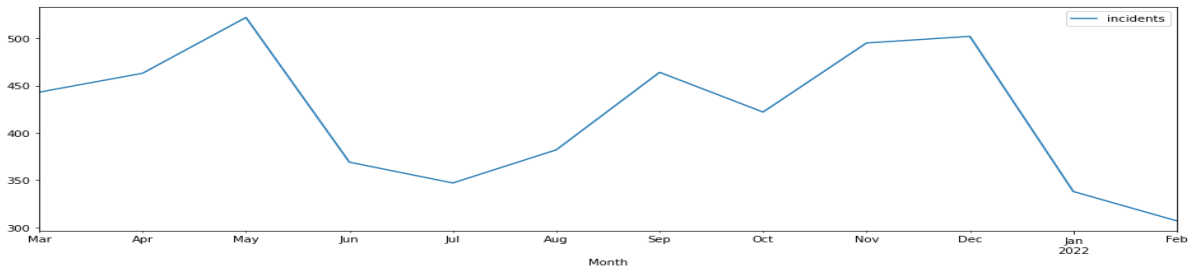


Figure 3. 3. 2 Yearly occurrences of crime in India

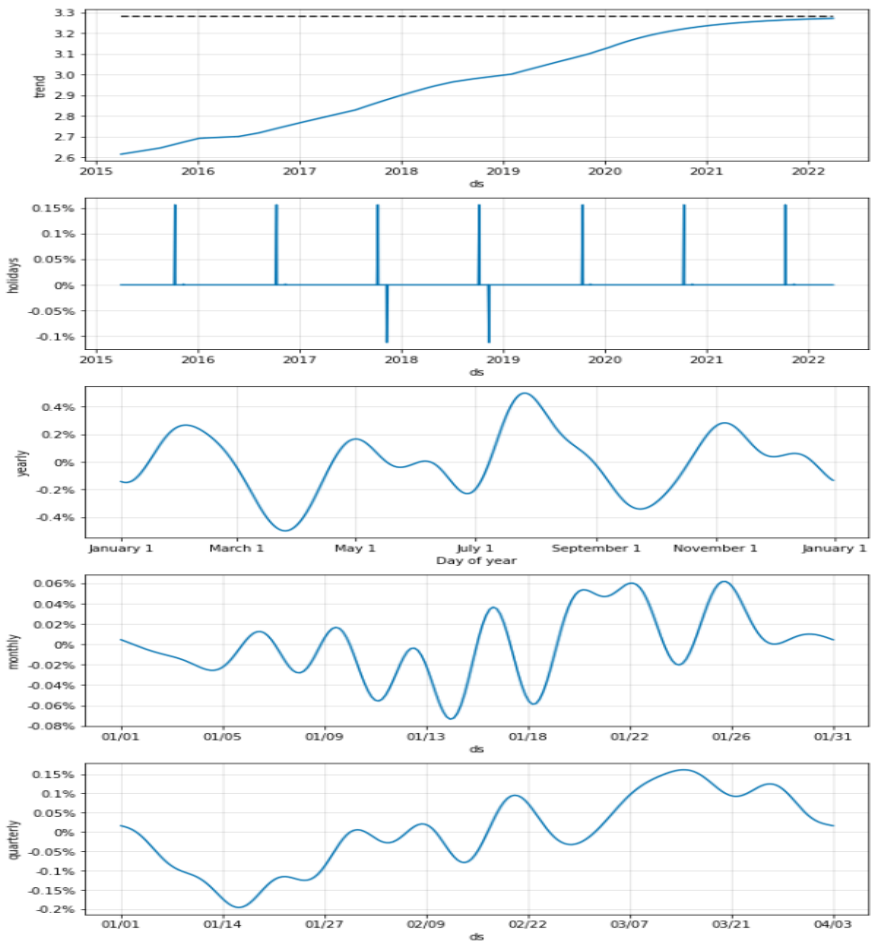


Figure 3. 3. 3 Weekly, Monthly and Yearly occurrences of crime in India

### 3.4 Forecasting and Prediction with time series algorithms

Crime is a violation of humanity, often punishable by law. Criminology is a study of crime, interdisciplinary science that investigates and analyzes the performance of criminal data. Criminal activity is very high in various locations. The main responsibility is to reduce the crime rate in the society with the help of visualization and forecasting by police, law enforcement and government agencies. These government sectors are help to hurrying up the process of solving the crimes. The computerized systems, crime data analysts can help to reduce the crime and finding the crime in less time. Crime analysis and prediction is the systematic formulation for the identifications and visualizations of each the patterns, seasonal data and trend data in crime. The prediction is not done by individual victims of crime but can predict the place that has a probability of its occurrence.

**SARIMA algorithm**

SARIMA algorithm is a time series algorithm for predicts the approaching data using past history in the time series.SARIMA contains seasonal and trend modules in which, it can support a univariate of seasonal module. An SARIMA algorithm is the extension of ARIMA algorithm that assists the straightforward modelling of the seasonality modules is called SARIMA. It can be added newly considered factors is to identify an auto regression (AR), differences (I) and moving average (MA) of the seasonal elements of time series of an extra factors for the historical data for the seasonality. Configuring a SARIMA algorithm that requires to selects the hyper factor for the elements of trended and seasonality of the time series data. In SARIMA algorithm two elements are available such as

1. Trended Element

The three elements of trend that can requires the design as (p, d, q) where p element is auto regression order, d element is difference order and q element is moving average order of trended data.

2. Seasonality Elements

It contains four elements for the seasonal elements such as (P, D, Q, m), where P is autoregressive order, D is difference order, Q is moving average order and m is the amount of single seasonal period in time series.

3.Residual Elements

Residual is an element in machine learning for calculating the differences between observed and future values of data. It is used to measure the model is better or not.

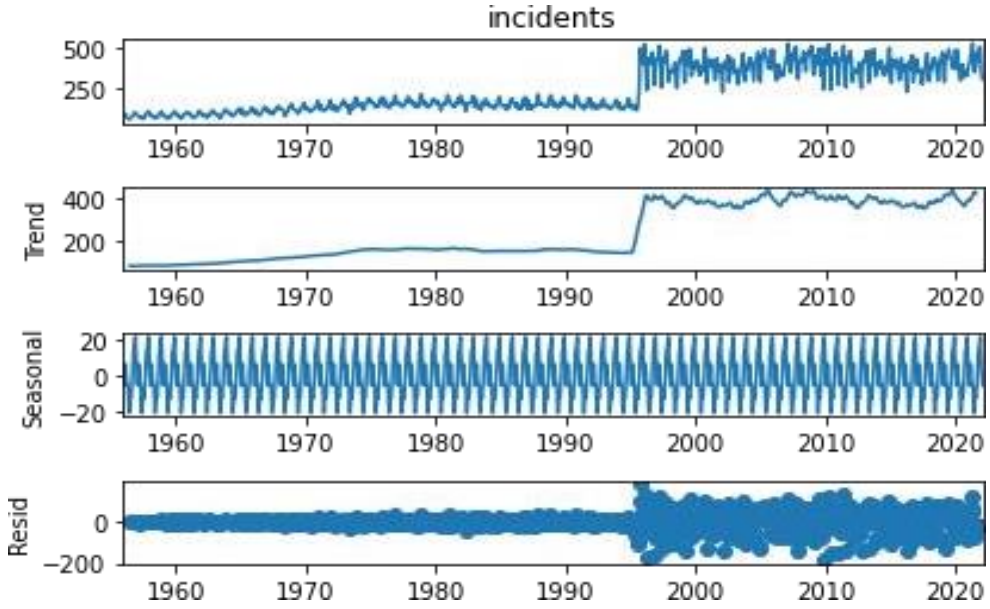


Figure 3.4.1 Forecasting with SARIMA Element on Crime Data

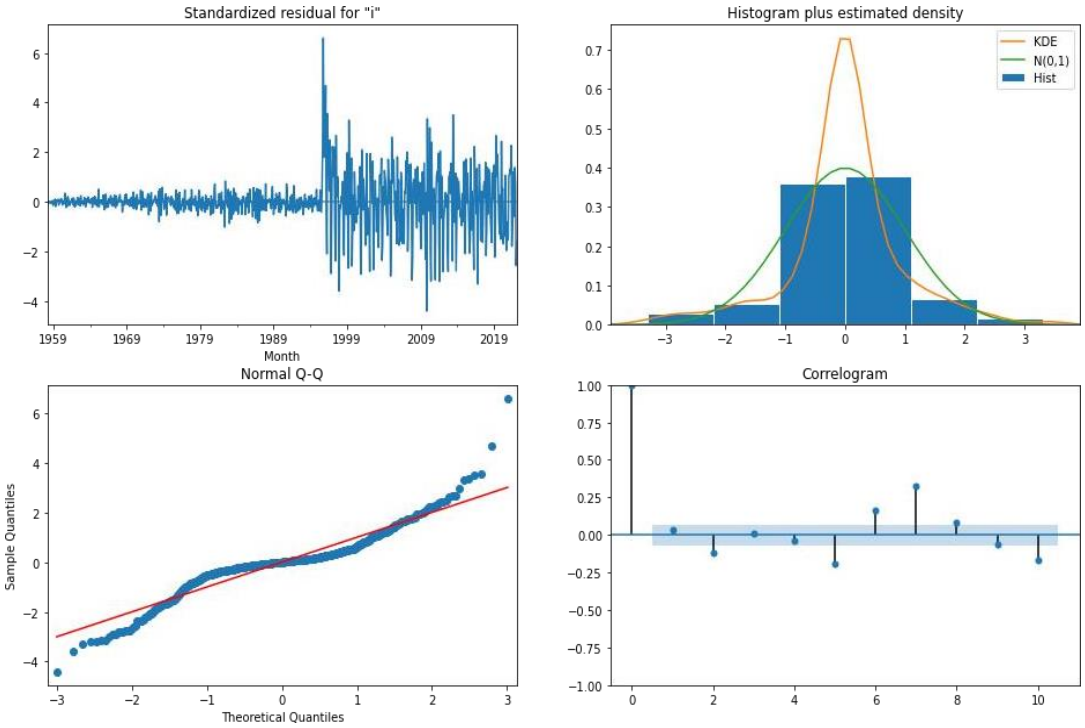


Figure 3.4.2 Forecasting with SARIMA on crime data

Top left is the residual errors seem to fluctuate around a mean of zero and have a uniform variance. Top Right is the density plot suggest normal distributionwith mean zero. Bottom left is for all the dots should fall perfectly in line with thered line. Any significant deviations would imply the distribution is skewed. Bottom Right is the Correlogram, aka, ACF plot shows the residual errors are not auto correlated. Any auto correlation would imply that there is some pattern in the residual errors which are not explained in the model. So you will need to look for more X's (predictors) to the model.

Figure 3.4.3 Predicting with SARIMA on Crime Data

```
sarima_pred = arima_result.predict(start = len(train_data), end = len(df)-1, typ="levels").rename("SARIMA Predictions")
sarima_pred
```

```
2021-03-01   380.74
2021-04-01   393.39
2021-05-01   402.48
2021-06-01   385.15
2021-07-01   352.33
2021-08-01   343.93
2021-09-01   380.76
2021-10-01   401.61
2021-11-01   428.40
2021-12-01   418.53
2022-01-01   380.67
2022-02-01   343.01
Freq: MS, Name: SARIMA Predictions, dtype: float64
```

After prediction SARIMA algorithm the result shown in the summary asshown in the figure as model, date, time and sample of the data set. The number of observations are recorded in the summary with Akaike Information Criterion (AIC), where Bayesian Information Criterion (BIC), where AIC and BIC are mainly used for selecting the model and Hannan Quinn Information Criterion (HQIC) is used for measure for the goodness of the model and fitness of the model for the given data .

SARIMAX Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | y | **No. Observations:** | 794 |
| **Model:** | SARIMAX(0, 0, 3)x(1, 1, [1], 12) | **Log Likelihood** | -4154.514 |
| **Date:** | Thu, 17 Mar 2022 | **AIC** | 8325.028 |
| **Time:** | 14:16:48 | **BIC** | 8362.322 |
| **Sample:** | 01-01-1956 - 02-01-2022 | **HQIC** | 8339.370 |

**Covariance Type:** opg

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **intercept** | 5.6037 | 1.011 | 5.543 | 0.000 | 3.622 | 7.585 |
| **seasonal_index** | -0.0003 | 2.66e+04 | -1.1e-08 | 1.000 | -5.21e+04 | 5.21e+04 |
| **ma.L1** | 0.4838 | 0.021 | 22.907 | 0.000 | 0.442 | 0.525 |
| **ma.L2** | 0.3135 | 0.026 | 11.994 | 0.000 | 0.262 | 0.365 |
| **ma.L3** | 0.3477 | 0.021 | 16.732 | 0.000 | 0.307 | 0.388 |
| **ar.S.L12** | -0.1027 | 0.035 | -2.944 | 0.003 | -0.171 | -0.034 |
| **ma.S.L12** | -0.7632 | 0.025 | -30.272 | 0.000 | -0.813 | -0.714 |
| **sigma2** | 2370.4232 | 71.623 | 33.096 | 0.000 | 2230.044 | 2510.802 |

**Ljung-Box (L1) (Q):** 5.91 **Jarque-Bera (JB):** 863.51

Figure 3. 4 .4 SARIMA Result for Crime Data Prediction

The model has estimated the AIC and the P values of the coefficients look significant. Let's look at the residual diagnostics plot. The best model SARIMAX(0, 0 , 3) x (1, 1, 1, 12) has an AIC of 8325.0 and the P Values are significant.

**ARIMA algorithms**

ARIMA is a mathematics and statistical technique to predict the future data with seasonal and trend data which is collected from the database. A statistical model is auto regressive if it predicts future values based on historical values. The crime data visualization and forecasting steps using SARIMA algorithm

1. Visualization of the Time Series Data.
2. Identify the date,if it is stationary or non stationary.
3. Plot the charts and graphs in Correlation and Auto Correlation function.
4. Based on data to construct the ARIMA Model or Seasonal ARIMA.

The ARIMA model to plot the data for checking the time series whether the data is stationary or not. After converting the stationary and identifies the ordering difference of each degree using the ACF and PACF graph plots from that find different ARIMA model. The diagnostics are good and generatingbetter performance with time series algorithms. The estimated forecasting make with three parameters and fit the model with residual analysis then start the forecasting the best model.

When ever the consideration of an algorithm as ARIMA model first we have to identify the stationary and identify the differencing of plots such as (ACF)Auto Correlation Function) and PACF(Partial Auto Correlation Function). After differencing the plots then find the ARIMA model. Estimation of the p, d, q parameters for the model. Find the ARIMA model is fit and the performs the residualson the ARIMA model for the best result and then start the forecasting and predicting the the model with better accuracy as mean squared error.

Crime analysis and prediction is the systematic formulation for the identifications and visualizations of each the patterns, seasonal data and trend data in crime. The prediction is not done by individual victims of crime but can predict the place that has a probability of occurrence

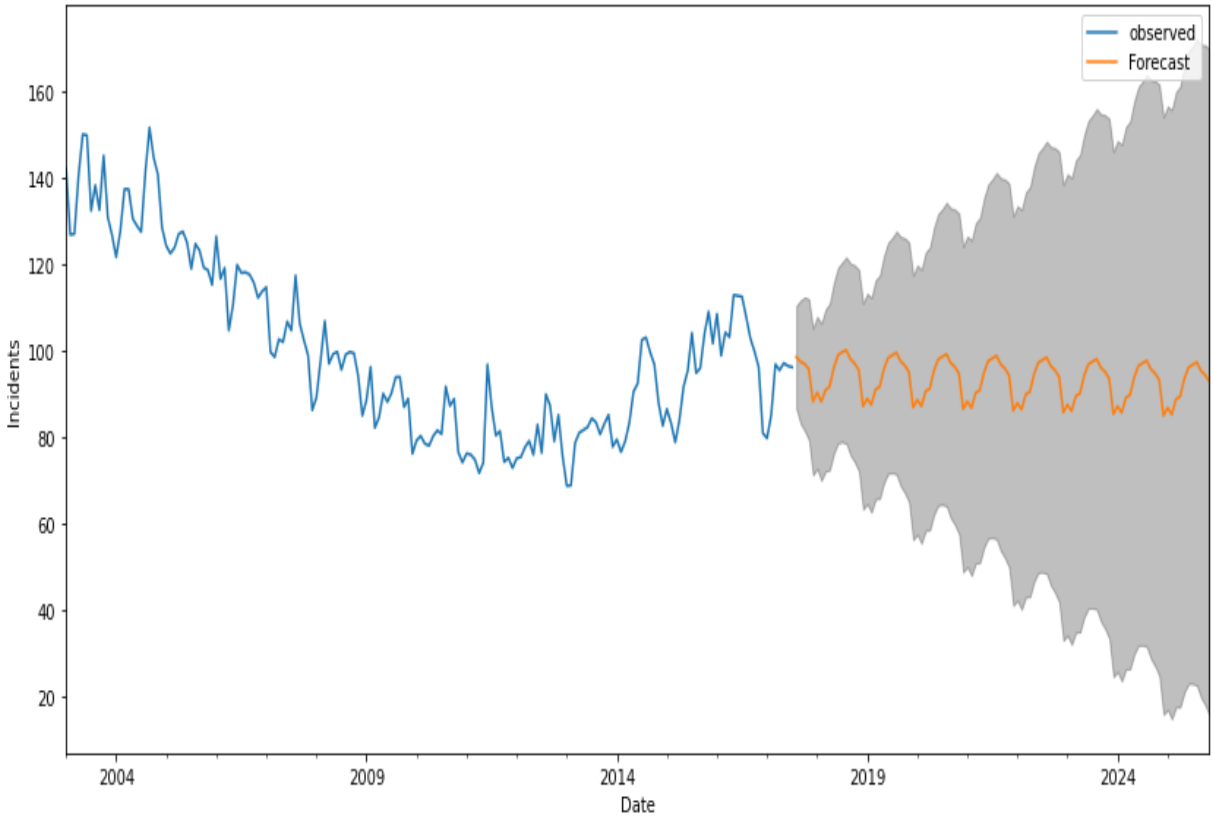


Figure 3.4.5 Forecasting with ARIMA on crime data

**PROPHET algorithm**

Prophet algorithm is a model to forecast the period of time data assists on regression hypothesis. The daily, weekly, yearly, seasonal and holiday effected data are used to fit the model. Prophet time series data implements as an additive time series forecasting model. It supports the implementation as for trends, seasonality, and holidays.

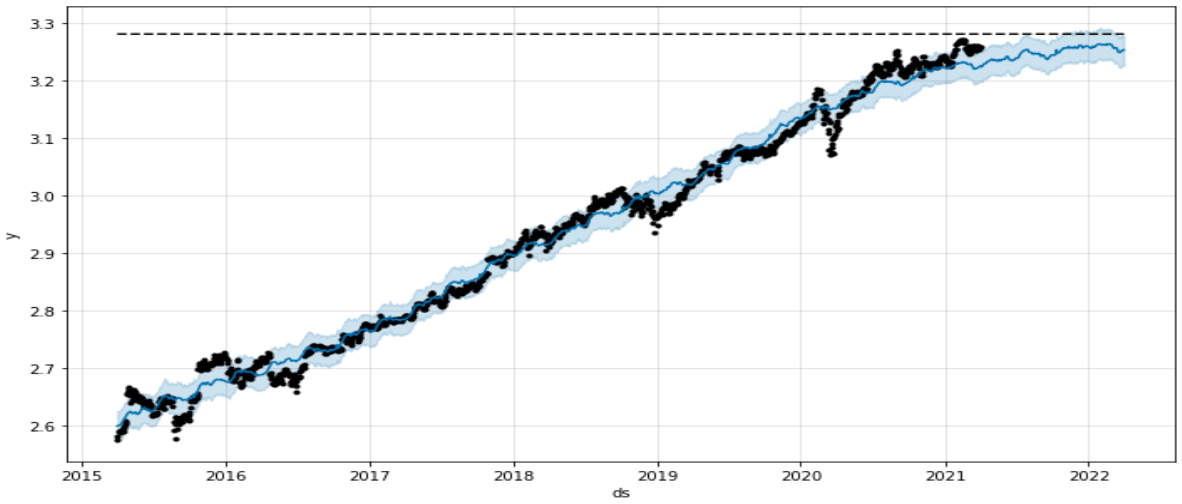


Figure 3.4.6 Forecasting with PROPHET on crime data

### 3.5 Error measurement

The error measurements is to verify the model correctness. In this implementation the crime rate forecasting and predictions with three techniques of time series algorithms such as the collected data can be divided into two parts as 80% of training and 20%of testing data, where the mistake rate can be performed on the test data. The train data is to fit the model and test data is used to evaluate the fit the machine learning

The error is known while forecasting the collected based on their future data to predict the crime data. The efficient and better way to measure the error for forecasting and predicting data using the technique such as MAPE (Mean Absolute Percentage Error). In forecasting techniques the good model has low MAPE and also the forecast value is less than or more than the actual value. The MPE (Mean Percentage Error) finds the separation of forecasting value and actual value and divides the present values by actual value.

SARIMA algorithm gives the MAPE value as 0.26 and MPE 0.10 to verify the model correctly and gives the best measures for forecasting and predicting the crime data.

## IV. RESULTS AND DISCUSSION

The crime data forecasting and prediction can be based on the time series models such as SARIMA, ARIMA and Prophet model. The SARIMA model, which also shows that the combination of the seasonal adjustment process and the ARIMA model can adjust the data in efficient way and predict more accurately.

The crime data forecasting and visualizing using the India crime data set. The State and District Wise Crimes in India data set is used to implement and visualize the crime rate in each district from each state. To visualize and analyze the crime data with the help of graphical representation of time series graphs as heat map, trend line and column charts. Crime data is mainly used to reduce the crime rate and predicting the different crimes in a particular location. Prediction and forecasting of crime data is mainly used for the police , law and crime investigators to investigate the bodies to examine the past crime data to reduce the crime activities in the country.SARIMA model gives better efficiency to predict the crime data to solve problems in easiest way and to reduce the crime rate in the country.The prophet model is to change the points in trends in the time series. ARIMA is a time series model to predict and forecast the future crime data where the criminal data is taken from the past existed details of each crime and elements.

```
crime
```

| | S.NO. | YEAR | CRIME INCIDENT | Unnamed: 3 | CRIME RATE |
|---|---|---|---|---|---|
| 0 | 1 | 1980 | 1368529 | NaN | 206.2 |
| 1 | 2 | 1981 | 1385757 | NaN | 200.8 |
| 2 | 3 | 1982 | 1353904 | NaN | 192.0 |
| 3 | 4 | 1983 | 1349866 | NaN | 187.4 |
| 4 | 5 | 1984 | 1358660 | NaN | 184.7 |
| 5 | 6 | 1985 | 1384731 | NaN | 184.4 |
| 6 | 7 | 1986 | 1405835 | NaN | 183.5 |
| 7 | 8 | 1987 | 1406992 | NaN | 180.1 |
| 8 | 9 | 1988 | 1442356 | NaN | 180.8 |
| 9 | 10 | 1989 | 1529844 | NaN | 188.5 |
| 10 | 11 | 1990 | 1604449 | NaN | 194.0 |
| 11 | 12 | 1991 | 1678375 | NaN | 197.5 |
| 12 | 13 | 1992 | 1689341 | NaN | 194.7 |
| 13 | 14 | 1993 | 1629936 | NaN | 184.4 |
| 14 | 15 | 1994 | 1635251 | NaN | 181.7 |
| 15 | 16 | 1995 | 1695696 | NaN | 185.1 |
| 16 | 17 | 1996 | 1709576 | NaN | 183.4 |
| 17 | 18 | 1997 | 1719820 | NaN | 180.0 |
| 18 | 19 | 1998 | 1778815 | NaN | 183.2 |
| 19 | 20 | 1999 | 1764629 | NaN | 178.9 |
| 20 | 21 | 2000 | 1771084 | NaN | 176.7 |
| 21 | 22 | 2001 | 17690308 | NaN | 172.3 |
| 22 | 23 | 2002 | 1780330 | NaN | 169.5 |
| 23 | 24 | 2003 | 1716120 | NaN | 160.7 |
| 24 | 25 | 2004 | 1832015 | NaN | 168.8 |
| 25 | 26 | 2005 | 1822602 | NaN | 165.3 |
| 26 | 27 | 2006 | 1878293 | NaN | 167.7 |

| | | | | | |
|---|---|---|---|---|---|
| **27** | 28 | 2007 | 1989673 | NaN | 175.1 |
| **28** | 29 | 2008 | 2093379 | NaN | 181.5 |
| **29** | 30 | 2009 | 2121345 | NaN | 181.4 |
| **30** | 31 | 2010 | 2224831 | NaN | 187.6 |
| **31** | 32 | 2011 | 2325575 | NaN | 192.2 |
| **32** | 33 | 2012 | 2387188 | NaN | 196.7 |
| **33** | 34 | 2013 | 2647722 | NaN | 215.5 |
| **34** | 35 | 2014 | 2851563 | NaN | 229.2 |
| **35** | 36 | 2015 | 2949400 | NaN | 234.2 |
| **36** | 37 | 2016 | 2975711 | NaN | 233.6 |
| **37** | 38 | 2017 | 3062579 | NaN | 237.7 |
| **38** | 39 | 2018 | 3132955 | NaN | 236.7 |
| **39** | 40 | 2019 | 3225701 | NaN | 241.2 |
| **40** | 41 | 2020 | 4254356 | NaN | 341.3 |

Figure 4.1 Crime Data incidents and crime rate through every year

```
# Accuracy metrics
def forecast_accuracy(forecast, actual):
    mape = np.mean(np.abs(forecast - actual)/np.abs(actual))  # MAPE
    me = np.mean(forecast - actual)**10      # ME
    mae = np.mean(np.abs(forecast - actual)**10)**2   # MAE
    mpe = np.mean((forecast - actual)/actual)   # MPE
    rmse = np.mean(np.abs(forecast - actual)/actual)**2  # RMSE
    corr = np.corrcoef(forecast, actual)[0,1]   # corr
    mins = np.amin(np.hstack([forecast[:,None],
                              actual[:,None]]), axis=1)
    maxs = np.amax(np.hstack([forecast[:,None],
                              actual[:,None]]), axis=1)
    minmax = 1 - np.mean(mins/maxs)             # minmax
    acf1 = acf(fc-test)[1]                      # ACF1
    return({'mape':mape, 'me':me, 'mae': mae,
            'mpe': mpe, 'rmse':rmse, 'acf1':acf1,
            'corr':corr, 'minmax':minmax})

forecast_accuracy(fc, test.values)
```

```
{'acf1': 0.7989070635246542,
 'corr': 0.8011758928428688,
 'mae': 3.551481544626573e+45,
 'mape': 0.26198236496343735,
 'me': 6.714599832663149e+17,
 'minmax': 0.24182805250315664,
 'mpe': -0.10635223721720918,
 'rmse': 0.06863475955183569}
```

Figure 4.2 SARIMA Model Performance Metrics on Crime Data

In this research the SARIMA model is used to analysis crime visualization and prediction of crime using Year Wise Crimes in India data set is used to implement and visualize the crime rate in each district from each state. To visualize and analyze the crime

data with the help of graphical representation. It contains the information about the crime incident to identify easily with the help of crime datasets. The data is identified with the type of crime incident and crime happened frequently in every year. This is useful for Indian government to understand the distribution of crimes in different incidents, predict future crimes and take actions to prevent them. In future research, the model to apply many types of crimes, such as robbery, intrusion, and premeditated murder, to improve the model's performance.

The main objective of this paper is to save and screening of all the Indian citizens from the harm and incidents, as well as protecting the safety and health. The point contribution of this paper is to reduce and prevent the crime rate and incidents from every state in India. The police officers has the main role to protection the peoples from the harmful conditions. Yes, we can take the various crime which happening now a days. some of the legendary police officers taking crucial decision to eliminate criminals from the earth while observing the past incidents.

This prediction and forecasting is easy to identify the crime rate using the auto correlation of 3 different algorithms . the SARIMA gives best accuracy for the future predictions.

The SARIMA algorithm result can be shown in the below figure

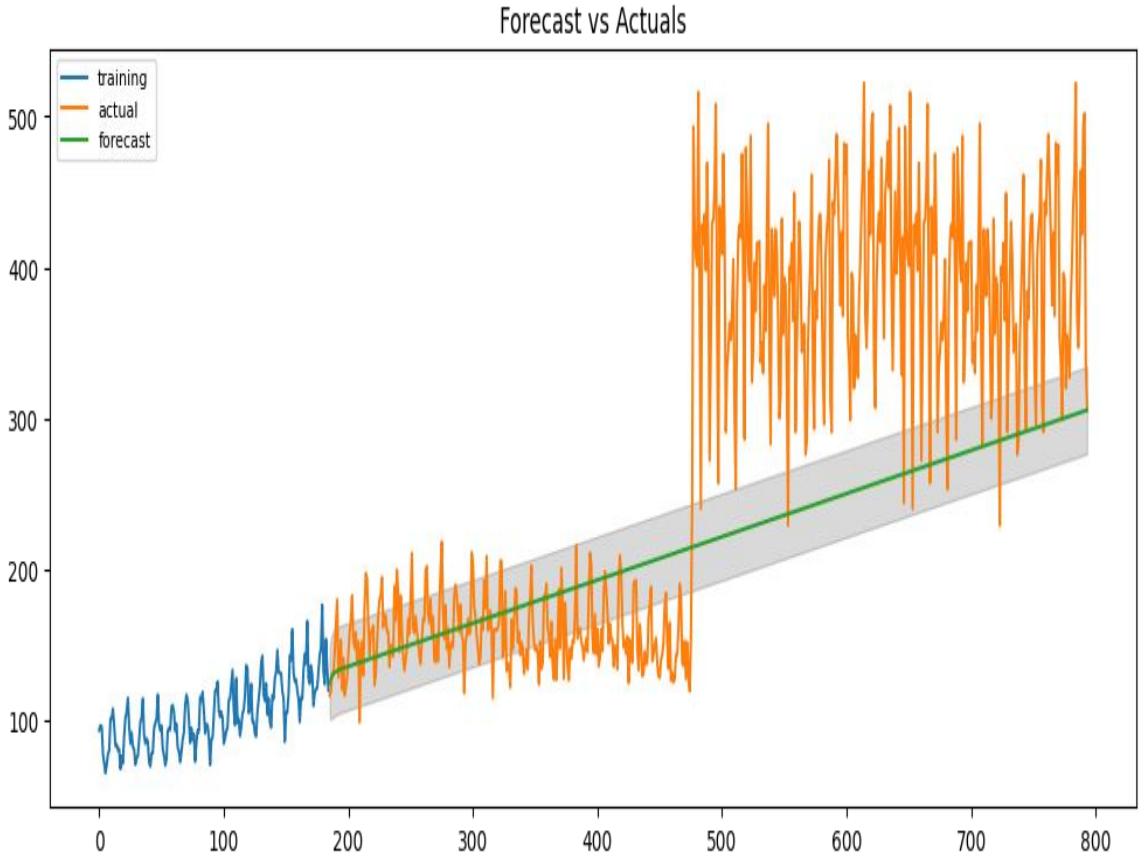


Figure 4.3 Forecasting result with SARIMA on crime data

Data were obtained from many official websites. The trained and validated the proposed SARIMA model using crime data obtained from the official data archive websites. These records were from the India, parameter values were determined according to the collected data.

A min-max scalar function was used to perform data normalization. Scaling the data was an important step to keep the variance values stable. Data normalization generally improves performance and reduces the associated computational complexity. Before starting model training, we normalized all datasets in this study using equation. In this formula, Xi represents the scaled data set, xi refers to the actual data, and the min(xi) and max(xi) terms correspond to the minimum and maximum values of the actual data set.

$$Xi=xi-\min(xi)\max(xi)-\min(xi)$$

According to the correct ARIMA prediction approach based on the actual values of the BIC, AIC, RMSE, MAE and MAPE criteria. Selecting the optimal parameters for an ARIMA model using graphical techniques is not an easy or quick process and takes a lot of time. Then fit a seasonal ARIMA model using the SARIMAX() function from the stats models module for each parameter combination and performed a scoring step to assess the overall quality of the model fit. After examining the full range of parameters, the optimal set of parameters was identified that gave the best performance on the criteria of interest.

The first step in model development is to identify a set of parameters and assign seed values to each. The value of s was set to 12 because we collected data monthly over 12 months. We then performed a grid search to find the best possible model with the lowest possible AIC value. The next step was to select the best combination of parameters that yielded the lowest amount of error (AIC) and assign it to the best model. AIC values for several predictive models are shown in Table 1.Furthermore, in figure 3.4.4 the lowest AIC value determined by the SARIMA model was (1,0,8) × (1,0,0,12). As a result, the best prediction model parameters were determined by combining the parameters (1,0,8)*(1,0,0,12). AIC and RMSE values are commonly used to compare SARIMA models. As can be seen from the comparison table, the predictive ability of the SARIMA (1,0,8)×(1,0,0,12) model was very robust compared to other models during the forecast period. A grid search method solved the problem of determining the optimal parameter values for the proposed SARIMA model.

**4.1 Comparison of different algorithms**

The comparison from SARIMA, ARIMA and Prophet algorithm from the time series model based on univariate seasonal modules. Prophet is the efficient and it is a model for forecasting time series data which trended element that are fit with seasonal data as daily, weekly, monthly, and yearly with holiday effects. Prophet is to change the points in trends in the time series. ARIMA is a model to predict the future crime data and forecast the criminal data based on the historical values of time series. SARIMA is a seasonal ARIMA it is used with time series with seasonality.

| **Model** | **Performance of MSE** | **RMSE** | **MAPE** | **MPE** |
|---|---|---|---|---|
| SARIMA | 0.24 | 0.06 | 0.26 | 0.24 |
| ARIMA | 26.05 | 62.32 | 6.5 | 6.2 |
| Prophet | 14.17 | 73.14 | 14.19 | 14.1 |

Table 1. Comparison of SARIMA, ARIMA and Prophet

SARIMA algorithm gives the MAPE valueas 0.26 and MPE 0.10 to verify the model correctly and gives the best measures for forecasting and predicting the crime data .The below graph shows the comparison between the three models which gives the better efficiency.

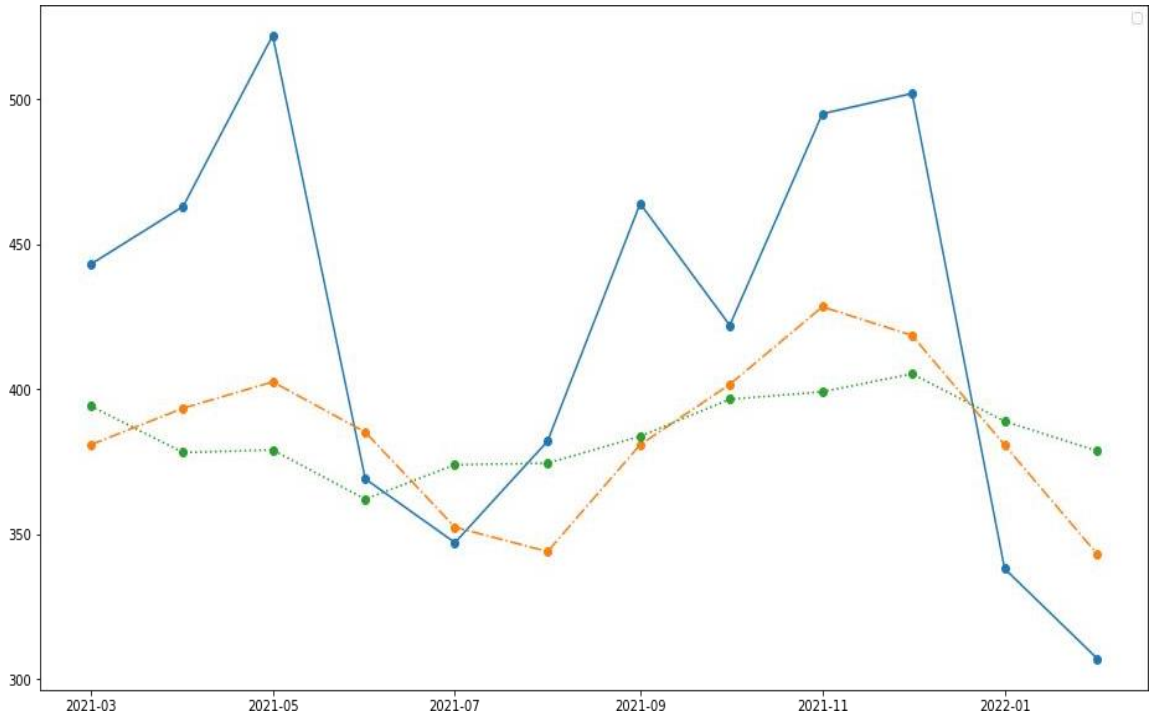


Figure 4.1.1 comparison of SARIMA, ARIMA and Prophet

The comparison between three time series algorithms where blue color indicates the SARIMA , Orange color line indicates as ARIMA and green color line indicates the Prophet models. In this comparison results SARIMA gives the best forecasting technique for the crime data set in India and predicting the future crime data to reduce the crime rate in the society .

## V. CONCLUSION

Forecasting and prediction the future data of crime data analysis which can be mainly used to distinguish types of measures. It improves the accuracy of the location, year and crime. The visualization is to identify the crime prone areas and can be used to design the methods with precautions for the future. Crime data analysis is used for the police officers, law and social control to analyze the past crime data and to reduce the crime activities in the country. SARIMA model gives better efficiency to predict the crime data to solve problems in easiest way and to reduce the crime rate in the country.SARIMA model gives better efficiency to predict the crime data to solve problems in easiest way and to reduce the crime rate in the country. Crime rate is increasing day by day with different issues that results a great loss, human loss and property loss in the society. To overcome this problem the computing era that reduce the crime rate and predicting the crime so that sufficient to measures can be taken to minimize the loss strategies.Forecasting and prediction the future data of crime data analysis is implemented with different time series algorithm with different data methods such as data collection, data cleaning, data preprocessing and data visualization are used for a large mixture of data sets where the data can be observed and revealed the crime data in the particular amount of time that cannot be observed before. This helps so much who are targeted from the various kinds of crime incidents and understanding about the things.The crime statistics that behaved this way in the past and will not behaved this way in the future because the crime rate is decreases day by day by consideration future predicted algorithm which used in this paper. It is not that much simple it depend on the crime.

How the incidents and crime activity being happening with the predicted future data with help of different time series algorithms in machine learning. In my point view analyzing the data, understanding about the crime, developing a positive hope for the upcoming future after seeing crime data in trended elements in the criminal data.This project make me more excitement about the crime data outcome in future safety understanding and analyzing what happened in the future. Future work with crime analysis and visualizations is to predict and forecast with tracking the locations with the geographical maps and tagging the crime activities with different algorithms

## REFERENCES:

[1] Mingchenfeng, Jiangbin zheng, Jinchang ren, (Senior member, Ieee), Amir Hussain, (Senior member, IEEE), Xiuxiu Li, Yue Xi, And Qiaoyuan Liu “Big data analytics and mining for effective visualization and trend forecasting of crime data”, Vol. 10, No.1109, Access, Auguest 2019.

[2] Abish Malik, Ross Maciejewski, member IEEE, Sherry Towers, Sean mccullough, and David S. ebertFellow IEEE, “Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement”,Vol. 20, No. 1 , September 2014.

[3] Abdulrahman takiddin, Muhammad Ismail, senior member, IEEE, Mahmoud nabil, Mohamed M. E. A. Mahmoud, senior memberIEEEand Erchin Serpedin, fellow IEEE,” Detecting electricity theft cyberattacks in ami networks using deep vector embeddings”, IEEE Systems Journal, Vol. 15, No. 3, September 2021.

[4] Wajiha safat, Sohail asghar, (Member, IEEE), and Saira Andleeb Gillani," Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques",Vol. 9, No. 70092, May 2021.
[5] Mingjian Tang, Mamoun Alazab, Senior Member, IEEE, and Yuxiu luo," Big data for cybersecurity: vulnerability disclosure trends and dependencies", IEEE Transactions on Big data, Vol. 5, No. 3, July-September 2019.
[6] Hongjian wang, Huaxiu yao, Daniel kifer, Corina graif, and Zhenhui li, member IEEE, " Nonstationary model for crime rate inference using modern urban data", IEEE Transactions on Big data, Vol. 5, No. 2, April-June 2019.
[7] Suman rath, Diptak pal, Member, IEEE, Parth sarthi sharma,and Bijaya ketan panigrahi, senior member, IEEE, " A cybersecure distributed control architecture for autonomous ac microgrid",IEEE Systems Journal, Vol. 15, No. 3, September 2021.
[8] Sobia khalid, Shoab ahmed khan and Syed Qasim Ifzal, " A Fuzzy logicbased framework for mapping crime data on established sociological hypothesis for societal disorder identification and prevention", Vol. 9, 10.1109/Access.3083542 June 2021.
[9] Priyanka das, (Student member,IEEE), Asit kumar das , Janmenjoy nayak , Danilo pelusi, and Weiping ding, (Senior Member, IEEE), " A graph based clustering approach for relation extraction from crime data " , Vol. 7, 10.1109/Access.2019.2929597 Auguest 2019.
[10] Meetha V. shenoy, (Member,IEEE), Smriti sridhar, Girish salaka, Anu gupta, and Rajiv gupta, (Member IEEE)," A holistic framework for crime prevention,response, and analysis with emphasis on women safety using technology and societal participation", Vol. 9, 10.1109/Access May 2021.
[11] Germain garcıa, Jaqueline silveira, Jorge poco, Member IEEE,, Afonso Paiva, Marcelo Batista nery, Claudio T. Silva, Fellow,IEEE," Crimanalyzer understanding crime patterns in sao paulo", IEEE Transactions On Visualization And Computer Graphics, Vol. 27, No. 4, April 2021.
[12] Germain garcıa zanabria, Marcos M. Raymundo, Jorge poco Member IEEE, Marcelo Batista nery, Claudio T. Silva fellow IEEE, Sergio adorno, luis gustavo nonato member IEEE," Cripav streetlevel crime patterns analysis and visualization",Vol. 14, No. 8, August 2015.
[13] Meng Yue, Member IEEE Tao Hong, Senior Member, IEEE, and Jianhui Wang , Senior Member, IEEE,"Descriptive analyticsbased anomaly detection for cybersecure load forecasting", IEEE Transactions On Smart Grid, Vol. 10, No. 6, November 2019.
[14] Olivera Kotevska, (Student member, IEEE), A. Gilad kusne, Daniel V. samarov, Ahmed Lbath , (Member, IEEE), and Abdella battou1, " Dynamic network model for smart city data-loss resilience case studycity-to-city network for crime analytics", Vol. 5, 10.1109/Access, October 2017.
[15] Binbin zhou , Longbiao chen, Fangxun zhou , Shijian Li, Sha zhao , Sajal K. das , Fellow, IEEE, and Gang Pan , Member, Ieee, "Escort finegrained urban crime risk inference leveraging heterogeneous open data", IEEE Systems Journal, Vol. 15, No. 3, September 2021.
[16] Saiba nazah, Shamsul huda, Jemal abawajy, (Senior member IEEE), and Mohammad mehedi hassan, (Senior member,IEEE), "Evolution of dark web threat analysis and detection a systematic approach", Vol. 8, 10.1109/Access, September 2020.s
[17] Jianming zhou, Zheng Li, Jack J. Ma, and Feifeng jiang, " Exploration of the hidden influential factors on crime activitiesa big data approach", Vol. 8, 10.1109/Access, August 2020.
[18] Hongjian wang, Huaxiu yao, Daniel Kifer, Corina graif, and Zhenhui Li, member, IEEE, "Nonstationary model for crime rate inference using modern urban data", IEEE Transactions on Big data, Vol. 5, No. 2, April-June 2019.
[19] Myung sun baek , (member, IEEE), Wonjoo park, Jaehong park, Kwang ho jang and Yong tae lee, "Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation", Vol. 9, 10.1109/Access, October 2021.
[20] Farkhund Iqbal, Benjamin C. M. fung, (Senior member, IEEE), Mourad debbabi, Rabia batool and andrew marrington, "Wordnet based criminal networks mining for cybercrime investigation", Vol. 7, 10.1109/Access, March 2019.