National College of Ireland

Crime Rate Prediction Using Time Series Forecasting

MSc Research Project
Data Analytics

Vishwakh Madhu
Student ID: x21125295

School of Computing
National College of Ireland

Supervisor: Dr.Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Vishwakh Madhu
Student ID:	x21125295
Programme:	MSc in Data Analytics
Year:	2022-2023
Module:	MSc Research Project
Supervisor:	Dr.Hicham Rifai
Submission Due Date:	01/02/2023
Project Title:	Crime Rate Prediction Using Time Series Forecasting
Word Count:	XXX
Page Count:	16

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	1st February 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Crime Rate Prediction Using Time Series Forecasting

Vishwakh Madhu
x21125295

Abstract

One of the most significant social issues in the nation is crime, which has an impact on socioeconomic position, child development, and public safety. In order to decrease crime and improve the quality of life for individuals, officials must have a clear understanding of the variables that contribute to increased crime. In our study, we address a basic issue: our research has undertaken a comparison method to estimate crime rates using the real-world crime data-set of the years 2007 to 2022 acquired from the official statistics web-site of the government. Whereas methods for estimating crime rates using time series data include SARIMA, Random Forest, XG Boost, and Prophet.

1 Introduction

One of the fundamental needs of every member of society is a safe and secure place to call home. Therefore, it's critical to develop strategies for reducing crime. A city's ability to grow economically is hampered by high crime rates. Law enforcement may greatly benefit from knowing the underlying factors that raise the risk of any given crime event happening at any given moment in order to avoid crime incidents. The surrounding environment, including nearby neighborhoods and human mobility, is significant for predicting crime events, according to the criminology idea. Analysis of crime data is required to reduce the crime rate due to a significant rise in crime worldwide. Because of this, it is easier for the general people and the police to take quick action to solve crimes. In order to anticipate characteristics that contribute to a high crime rate, data mining methods are used in this study to analyze crime data. While supervised learning uses data sets to train, evaluate, and produce desired outcomes on them, unsupervised learning categorizes or clusters chaotic, unstructured data. Some of the supervised learning techniques used in data mining and machine learning to forecast the elements that influence crime in an area or locale include regression, decision trees, and naive bayes. The Crimes Record Bureau and Police Department may take the appropriate steps to lessen the likelihood that a crime would occur based on the rankings of the attributes. The literature lacks the relative accuracy for crime prediction from huge datasets for several cities, despite significant research efforts; the Chicago dataset, for example, have only sometimes been employed. None of the researchers has ever combined the usage of all these models into a single study. Which is the crucial part of this study. Because it helps the capacity of investigative authorities to handle crime computationally, crime prediction has grown in prominence in recent years. Better prediction algorithms are required to focus police patrols on criminalsBrantingham et al. (2018). The primary goal of time series analysis is to examine time-series data in order to derive significant statistics and other features.

Based on previously observed data, this knowledge strongly predominates in forecasting future values. Violent crime and mishaps are increasing as a result of population growth and ongoing urbanization. collecting data and looking for hidden patterns Tasnim et al. (2022).

2 Research Question

How well crime rates in Chicago can be predicted using time series approaches?

3 Related Work

We have witnessed tremendous study efforts in the area of crime rate prediction from different academics. Machine learning, deep learning, and a few other combinations or other approaches were characterized in this review of the literature. For real-world problems, this study will be applying time series approaches. In the actual world, policing crime throughout the state or nation is quite tough. While deep learning techniques need a lot of data to be as successful as machine learning algorithms, they may train well from small amounts of data.

3.1 LITRATURE REVIEW

Machine learning and data mining are important tools for preventing and identifying many sorts of crime. The researcher utilized WEKA, an open source tool, to conduct a comparative analysis of community crime patterns and un-normalized crime datasets. Neighborhoodscout.com also contributed real crime data for the state of Mississippi to the University of California-Irvine repository. On a dataset of neighborhoods and crimes, this study employed decision stump, additive regression, and linear regression approaches. The linear regression technique performed the best out of the three options.
This paper's primary goal is to show the effectiveness and accuracy of machine learning algorithms used in data mining analysis for predicting violent crime trends McClendon and Meghanathan (2015). The Decision Stump approach outperformed the other two procedures the least. There are also a few more uses, such as finding "hot spots" for crime, creating criminal profiles, and comprehending crime trends, that were overlooked. Ingilevich and Ivanov (2018) investigate a variety of approaches to the problem of calculating the number of crimes in different metropolitan areas. Gradient boosting, logistic regression, and linear regression were the three types of predictive models that were explored in this work. The constructed models were tested using data on crimes in Saint Petersburg. Gradient boosting is the most effective method for forecasting crime rates in a certain area, according to our analysis of the results of all three models' predictions.
Learning about the reasons of crime has been a key issue for scholars, which is what the author of the work Alves et al. (2018) is concentrating on. A machine learning approach based on ensembles can tackle these problems well. when the importance of urban indicators is reviewed and categorized into equal-influence groups that hold true regardless of changes in the data sample being assessed. We use a random forest regressor, which has a 97 percent accuracy rate, to anticipate crime and quantify the effect of urban characteristics on murders.

Several machine learning algorithms have been used for crime data in this study to examine how the financial crisis is impacting crime in India. According to Mittal et al. (2019), research has been done on the correlation between theft, robbery, and burglary as well as the unemployment rate and gross domestic product. Additionally, Granger causality between crime rates and financial indicators has been calculated. The experimental investigation has shown that the key economic factor influencing the rate of crime is the unemployment rate, paving the path for crime control through enhancing work opportunities.

A DNN-based method for feature-level data's fusion with nature's context is presented in the research Kang and Kang (2017). Their dataset is compiled from many online sources that provide data on the population, climate, and crime rates of Chicago, Illinois. Before providing training data, we choose data that relates to crime by statistical analysis. the DNN, which has layer for joint feature representation, environmental context, spatial and temporal information, and time, is then trained. Our fusion DNN is the outcome of an efficient decision-making process that statistically assesses data redundancy and critical information gathered from other sources. Experimental results show that our DNN model performs better than other prediction models.

In the paper Mohler (2014), the authors show how a marked point process technique may be used to capture both short- and long-term risk patterns and to extend point process models of crime to include leading indicator crime types. Numerous years' worth of data from various crime categories are meticulously integrated to create accurate hotspot maps that may be used for predictive policing of gun-related crime. We leverage a substantial open source data set made available to the public by the Chicago Police Department for our study.

In the study Chen et al. (2008), researchers forecast short-term property crime in one Chinese city using an ARIMA time series model. The given data on property crime over 50 weeks is used to create an ARIMA model, which forecasts the amount of crime for the next week. In terms of fitting and prediction accuracy, it has been shown that the ARIMA model performs better than exponential smoothing.

This study by Kim et al. (2018) examines the subject of machine learning-based crime prediction (2018), This research evaluates crime statistics from Vancouver over the preceding 15 years using two different data-processing approaches. With the use of boosted decision trees and K-nearest-neighbor algorithms, it is feasible to predict crime in Vancouver with an accuracy of 39 percent to 44 percent.

The goal of the work by Bappee et al. (2018)is to create a machine learning model for crime prediction that takes into account the spatial features of different forms of crime. The reverse geocoding technique obtains spatial data from Open Street Map (OSM). The report also advises searching for areas that have been isolated from criminal hotspots. The study presented here established two kinds of geographic characteristics: geocoding and the shortest distance to hotspots. It was discovered that the classifiers put to the test improved significantly in accuracy and AUC when the newly produced features were included.

The study by Adesola et al. (2022) investigates the effectiveness of crime prediction techniques based on machine learning that have previously been used by other researchers. Machine learning techniques for predicting violent crimes are discussed in this study. Five years of historical data, from July 2014 to July 2019, given by the Nigerian Police in Lagos, were collected, evaluated, and used to train the model. Two types of Machine Learning prediction model, Decision Tree and K-Nearest Neighbors, were implemented

using IBM Watson Studio with the real-world dataset gathered from the Nigerian Police Obalende Lagos and the online crime reported portal during violent crime prediction in Lagos, with extreme crime prediction accuracy of 79.65 percent and 81.45 percent respectively.

Mukherjee and Ghosh (2022) In the research, a model that may be used to determine the overall number of crimes committed in a given state, split down by category, was built. Now that machine learning methods are more widely used, it is possible to predict crimes using historical data. The main focus of this work is the development of ensemble models, which outperform earlier algorithms in terms of stability, accuracy, and generating more precise forecasts.If data decomposition techniques that emphasize area-wise accuracy rates are used, the results will outperform the existing approach. Examining geographic crime data is a vital part of this endeavor as well. Regression, random forest, naive Bayes, and decision tree approaches are used here.

In the research Wang et al. (2016), they discuss the problem of estimating the crime rate at the neighborhood level. Demographics and geography characteristics have typically been used to estimate crime rates in a given area. Big data on modern cities has been gathered as a result of the quick development of location technology and the widespread use of mobile devices, and this data may provide new perspectives on crime. We discovered that crime rate inference worked noticeably better than using conventional features. There has been continuous progress over many years. We also use the variable importance analysis to show the relevance of these other qualities.

As an alternative to the present modeling methodologies, Kianmehr and Alhajj (2008) investigate a support vector machine (SVM)-based method for predicting the position. Support vector machines are the most recent generation of machine-learning techniques for determining the optimal class separability within datasets. Two alternative SVMs methodologies, such as two-class SVMs and one class SVMs, are evaluated for efficacy. Additionally compared to SVM were a spatial autoregression-based method and a neural network-based approach. The older technique performs a little bit better, but the more recent strategy yields respectable results, according to analyses of two different geographic datasets. Additionally, they provide a general framework in this research that may be used to jobs in many spatial domains that include information similar to the investigated crime datasets, allowing for the adaptation of geographic data classification.

Researchers Malathi and Baboo (2011) study how attribute-based data mining and clustering algorithms might be used to forecast crime patterns and expedite the criminal investigation process. To fill in the missing value and spot criminal tendencies, we will concentrate on the MV algorithm and Apriori algorithm with certain enhancements. We used these techniques to real crime data. The semisupervised learning technique is also used in this work to increase prediction accuracy and gather data from criminal histories.

Joshi et al. (2017).'s paper's main objective is to look at crime utilizing both qualitative and quantitative methods. This include investigating crimes such as theft, murder, and drug offenses in addition to odd behaviors, noise complaints, and burglar alarms. On the basis of crime data from the New South Wales region of Australia, the K-means clustering and data mining approach has been used to identify crime rates for each type of crime as well as high-crime areas.

The main objective of the study by Kiani et al. (2015) is to classify clustered crimes based on how often they occur over time. Data mining is often used to examine, look into, and identify trends for the occurrence of certain crimes. We used a theoretical model employing data mining techniques like clustering and classification on an actual

crime dataset gathered by authorities in Wales and England between 1990 and 2011. We assigned weights to the attributes in order to improve the model's quality and eliminate components with poor value. The Outlier Detection operator parameters are optimized using the Genetic Algorithm (GA) with the aid of the RapidMiner tool.
Researchers Krishnendu et al. (2020) have used a systematized approach, crime analysis, and pattern prediction to classify and examine crime patterns. Numerous clustering techniques are available. They do not, however, reveal all the requirements. K means algorithm among them provides a better way to predict the results. The main objective of the planned study was to identify places with higher crime rates and age groups that were more or less likely to commit crimes. In order to save time and increase effectiveness, we provide an enhanced K means approach. The K-means algorithm's result lowers the number of iterations while increasing the precision of the final cluster.

4 Methodology

The research on crime rate prediction that is being presented employs a set of stages based on the KDD approach throughout the whole data mining process. The picture below provides further information on each of these phases.

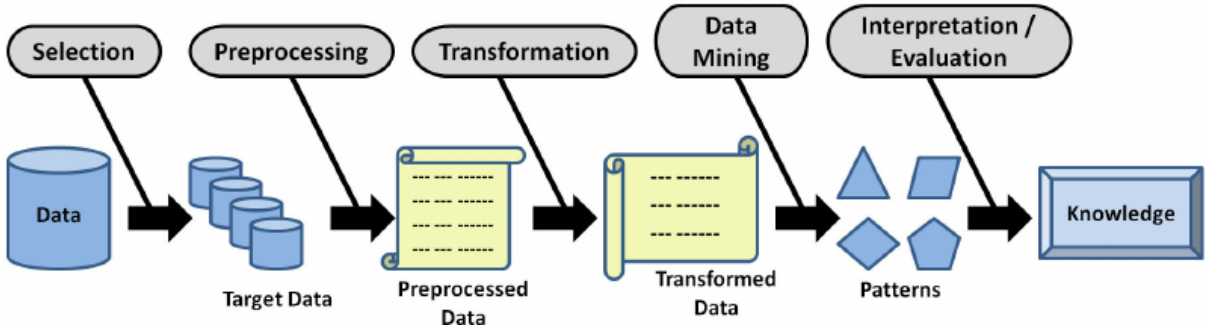

Figure 1: Crime Prediction Methodology

4.1 Data Acquisition

The government's official statistics portal, which is accessible to the public for scientific and academic use, provided the data required for this study. The crimes that were committed in the City of Chicago between 2001 to the present are shown in this dataset. The Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system was used to collect the data. It is accessible through the Chicago Data Portal. (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2).

The dataset includes the following columns: (ID, Case Number, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, Year, Latitude, Longitude, Location, Updated on). All required data is gathered in a Python environment using Anaconda Navigator's Jupyter Notebook.

4.2 Data Pre-Processing and Transformation

Python programming language is used for model application and dataset processing once a number of pre-processing tasks are completed on the datasets.

4.2.1 Dealing with missing data

A few fields in the crime dataset, including Location Description, Ward, Community Area, Latitude, Longitude, and Location, have missing values. These values are eliminated since these columns won't be part of our prediction.

4.2.2 Feature Engineering

The crime dataset contains data on offenses from 2001 to 2022; however, for our estimate, we'll utilize data from 2007 to 2021. The date column is converted to DateTime Format for the sake of our study. The dataset is then filtered based on the top 10 major crime types to provide a clearer picture of the crime statistics. Since time series forecasting calls for a target column for forecasting, we will add a goal column called Count.

For this research, identifying and handling outliers is critical. Using the winsorize python module, we are removing the outliers, and the changed count is now used as the winsorized count.

Features are extracted from the date column, including hour, month, week, and year, since this study focuses on weekly prediction.

4.2.3 Transformation

Once the data are adequate for modeling, we must establish if they are stable. To ascertain if the data is stationary, the Agumented Dickey-Fuller test is employed. The null hypothesis is considered here because the data are not steady, and the alternative hypothesis is that the data are stationary. The Augmented Dickey-Fuller test may be used to produce a P-Value. If the P-Value is less than 0.05, the null hypothesis is rejected and the data are stationary. If the P-Value is more than 0.05, several adjustments must be made before the data can be fit into the model. Data that includes trend and seasonal components is referred to as non-stationary data. We cannot use it to choose the best model since the data are not stable and the P-Value is more than 0.05.

Data transformation is done using first order differencing, and stationarity is confirmed using the ADF (Augmented Dickey-Fuller) test once again. Here, the P-Value was less than 0.05, and our dataset likewise showed a seasonality tendency.

5 Design Specification

This research follows three-stage design flow such as data preparation, modeling, and visualization stage shown in figure 2.

The processes of data collection, combining, exploratory data analysis, feature engineering, and feature selection are all included in the data preparation stage. Each data set was downloaded as a csv file or collected utilizing API connections using Python programming on Jupyter Notebooks, depending on the data source.
The modeling step involves the implementation of several models, including SARIMA, Prophet, XGBoost, random forest, and evaluation using metrics like mean absolute error.

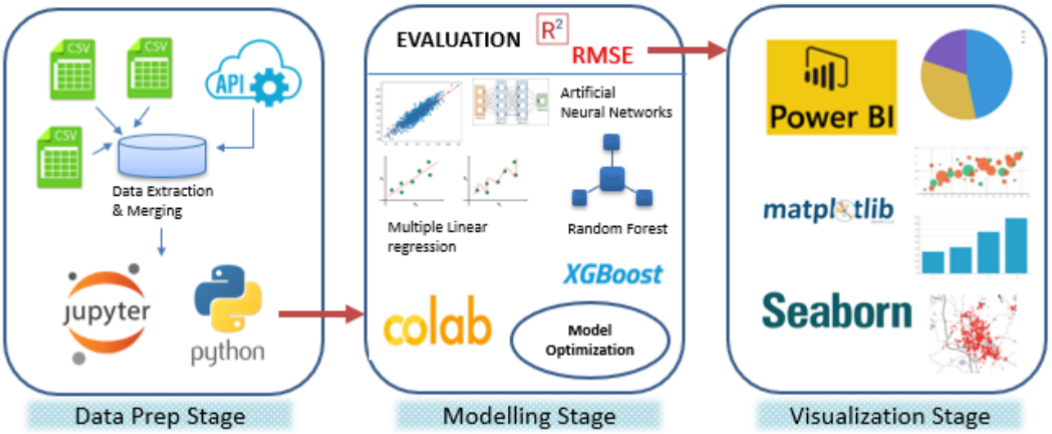

Figure 2: Design Specification

On Jupyter Notebook, modeling and optimization have been done.
Finally, the data were shown in plots and graphs for viewing purposes as needed.

6 Implementation

The whole process and application of the models utilized to accomplish the goals of the research study are covered in this part. The flow diagram of the implementation used during the investigation is depicted in Figure 3 below

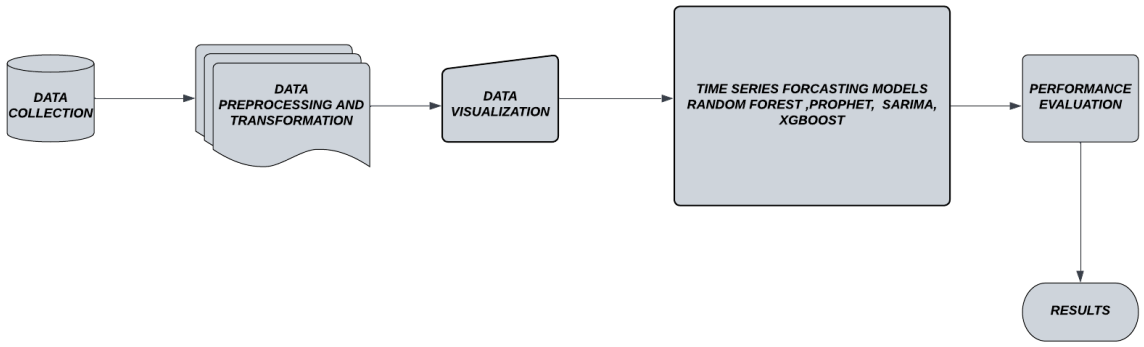

Figure 3: Implementation Flow Chart

6.1 Data Collection

The government's official website is where the data set for this study was gathered. The acquired data is carefully examined in order to comprehend it and derive the necessary insights. The information used in this study was taken from the official government website, which is open to the public for scientific and research reasons. The information is

presented as comma-separated values (CSV) and includes statistics by areas and categories of crimes for the years 2007 through 2022. The data collection has 22 columns and 7690564 rows, including NA values (ID, Case Number, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, Year, Latitude, Longitude, Location, Updated on.). Figure 4 below shows a sampling of the data from the collection.

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	Ward	Community Area	FBI Code	Coordinat
0	10224738	HY411648	09/05/2015 01:30:00 PM	043XX S WOOD ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	True	...	12.0	61.0	08B	1165074.(
1	10224739	HY411615	09/04/2015 11:30:00 AM	008XX N CENTRAL AVE	0870	THEFT	POCKET-PICKING	CTA BUS	False	False	...	29.0	25.0	06	1138875.(
2	11646166	JC213529	09/01/2018 12:01:00 AM	082XX S INGLESIDE AVE	0810	THEFT	OVER $500	RESIDENCE	False	True	...	8.0	44.0	06	NaN
3	10224740	HY411595	09/05/2015 12:45:00 PM	035XX W BARRY AVE	2023	NARCOTICS	POSS: HEROIN(BRN/TAN)	SIDEWALK	True	False	...	35.0	21.0	18	1152037.(
4	10224741	HY411610	09/05/2015 01:00:00 PM	0000X N LARAMIE AVE	0560	ASSAULT	SIMPLE	APARTMENT	False	True	...	28.0	25.0	08A	1141706.(

5 rows × 22 columns

Figure 4: Dataset

6.2 Data Pre-Processing

Cleaning the raw data is essential before adding the real data to the model in order to eliminate unnecessary information such as missing values (NA), outliers, etc., and to ensure that the model operates well. This was accomplished by using the Juypter notebook (Anaconda Navigator) program to carry out the necessary data cleansing for this study. The specific actions used for data pre-processing are listed below.

1) Null values are first eliminated.
2) Information before year 2007 was eliminated.
3) The top 10 most common kinds of crimes are used to filter the crime dataset.
4) Only two relevant columns from the dataset were chosen, and the other columns were dropped.
5) The dataset now includes a count column.
6) For forecasting purposes, the date column from the dataset was transformed to datetime format.
7) utilizing the winsorize python tool to detect and eliminate outliers.
8) Using the Date column as the dataframe's index.
9) dividing the dataset into weekly groups.

Modeling was done using this cleaned data. Similar improvements were performed to gather more insights, and data were analyzed to determine the most prevalent crime category nationwide. The top 10 kinds of crime are clearly illustrated by the analysis's findings.

6.3 Time Series Forecasting Models

6.3.1 SARIMA Model

The crime data is first put into a dataset by parsing the Date column and converting it to a Date Time format in Jupyter Notebook once the necessary libraries have been loaded. Only the Date and count columns are now used as input to the model, with the removal of the following columns: postal code, ID, Case Number, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, Year, Latitude, Longitude, Location, Updated. Pandas, NumPy, Matplotlib, Sklearn, and statsmodel.tsa are the main libraries used to construct/build ARIMA and SARIMA models. If the time series data is stationary or not, it is determined from the beginning. The Dickey-Fuller test findings that indicate a trend in time series data indicate that the time series data is stationary. It is decided whether or not the provided data set is stationary by creating a separate test stationarity function. Using the 'adfuller' function on the time series data, the Dickey-Fuller test is run after plotting the rolling mean and standard deviation. Using test statistics, p-value, lags, and observations, a sequence of output is produced. The p-value was less than 0.05 because the time series data utilized was stationary. Therefore, 1-order differencing (value of d) was carried out, and again the Dickey-Fuller test was carried out, giving a p-value of less than 0.05, which demonstrates that the time series was stationary. Which show the ACF and PACF plots, respectively, aid in determining which parameter should be supplied to the ARIMA model. The graph after the model is trained to fitted vs. the original for the one order differencing is considered. Since the data is in seasonal pattern SARIMA Model performs better. Similar steps are followed to fit SARIMA model. In order to determine the values of p and q, it was also given to verify the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. The data is now split into train and test groups, and the SARIMA model is trained on the train data using the variables p, d, and q.Seasonal Order(1,1,1,52 The model is then fitted using the.fit() method. The values for the test data are predicted using the function.forecast() once the model has been trained. The evaluation metrics RMSE, MAPE, and MAE are used to examine both data frame values and forecasting values. We used the SARIMA (Seasonal ARIMA) machine learning model in this study since the crime dataset exhibits seasonal patterns.

6.3.2 Prophet

The next model used in this study is Facebook Prophet, which the social media giant created and has successfully employed with time series data. The cleaned data is first put into a data frame. The Date column is transformed into a DateTime format using the DateTime function. This model only accepts the inputs from the columns Date and Count. Later, the column Date was changed to "ds" and the column count was updated to "y" in accordance with the prophet model's specifications. As we are predicting weekly crime data, the crime dataset has now been split into training and testing. The frequency is represented as "w," which stands for weekly data. For training data, the prophet model is fitted using the [.fit()] function, and after training, prediction values are predicted. RMSE, MAPE, and MAE metrics are used to compare test and prediction values before the actual and predicted values are shown.

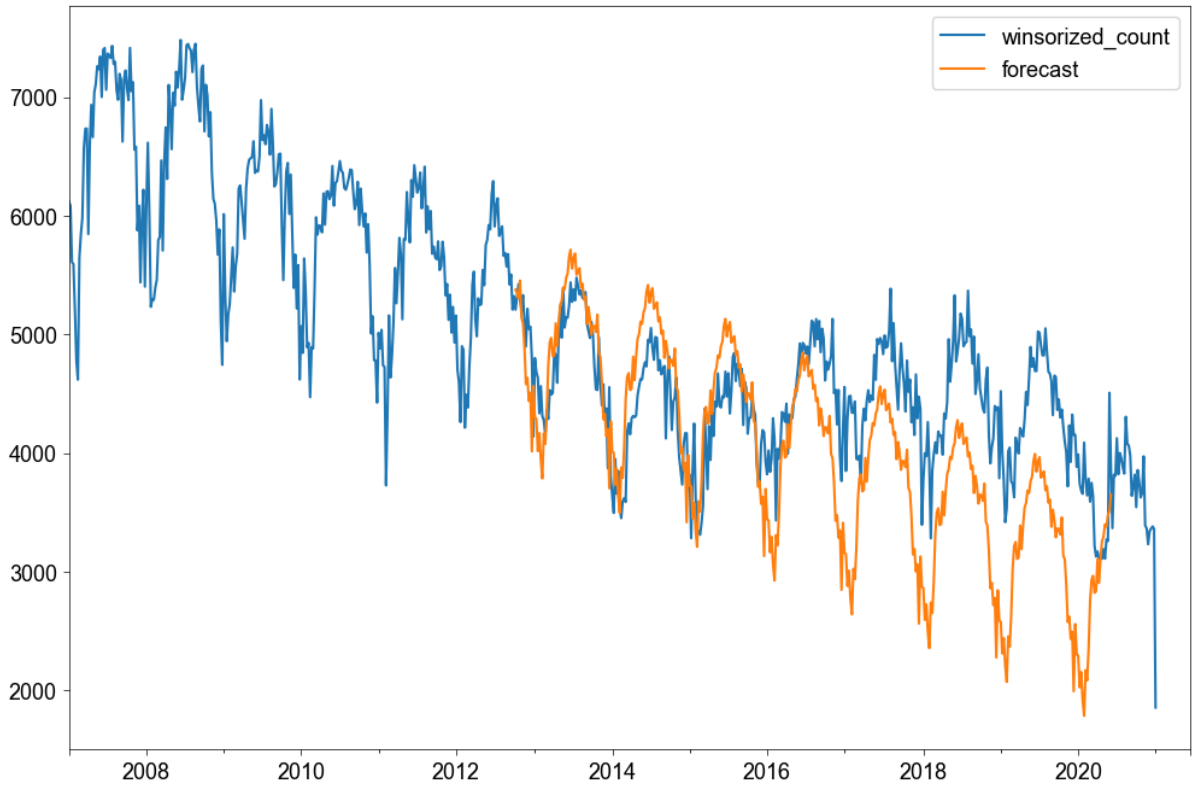

Figure 5: SARIMA Model Forecast

6.3.3 XG Boost

One of the most effective ways to execute the gradient boosting machine learning technique is called extreme gradient boosting (XG Boost). In addition to being used for time series forecasting, XGBoost is developed for classification and regression issues. Installing the XGBoost library is the first step. For the XG Boost model, the cleaned crime dataset is now used as input. The first step is to separate the training dataset from the testing dataset. The training dataset includes time series data from 2007 to 2018 whereas the testing dataset includes data after 2018. Dayofweek, quarter, month, year, dayofyear, dayofmonth, and weekofyear are among the feature functions I have developed. The training dataset has now been divided based on the chosen features and the target value. The training dataset and the testing dataset are then fitted using the XGBregressor model. Finding the significance of a characteristic. The XG Boost model is finally evaluated using mean absolute error. MAE= 66.68.

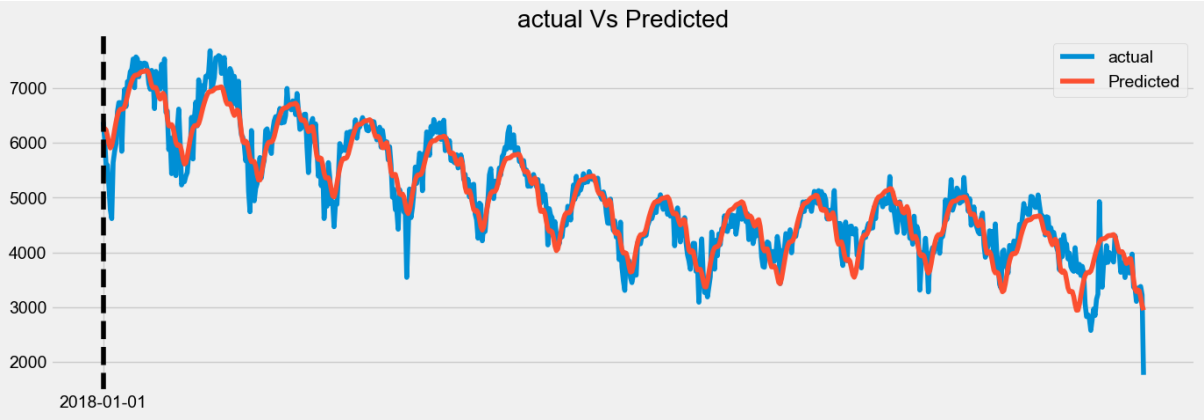

Figure 6: Prophet Model Forecast

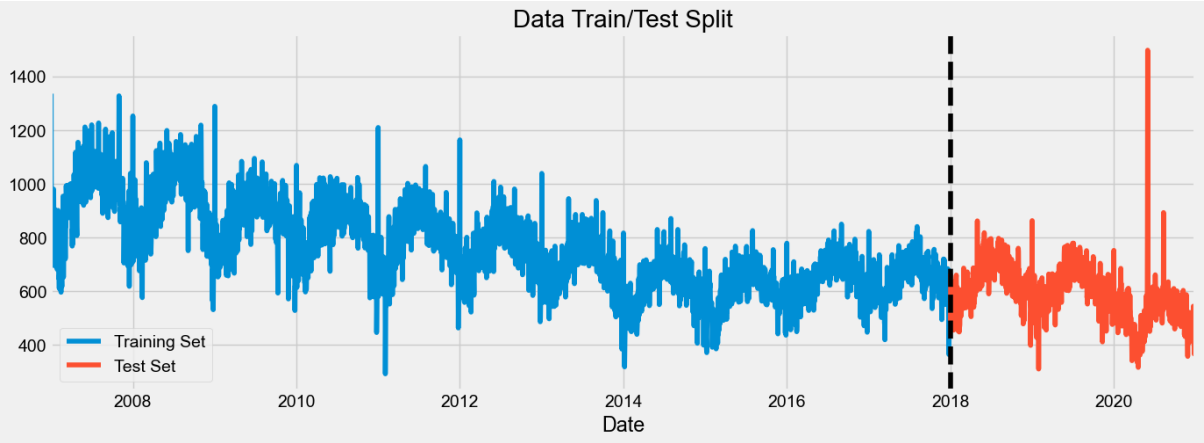

Figure 7: XG Boost Train Vs Test

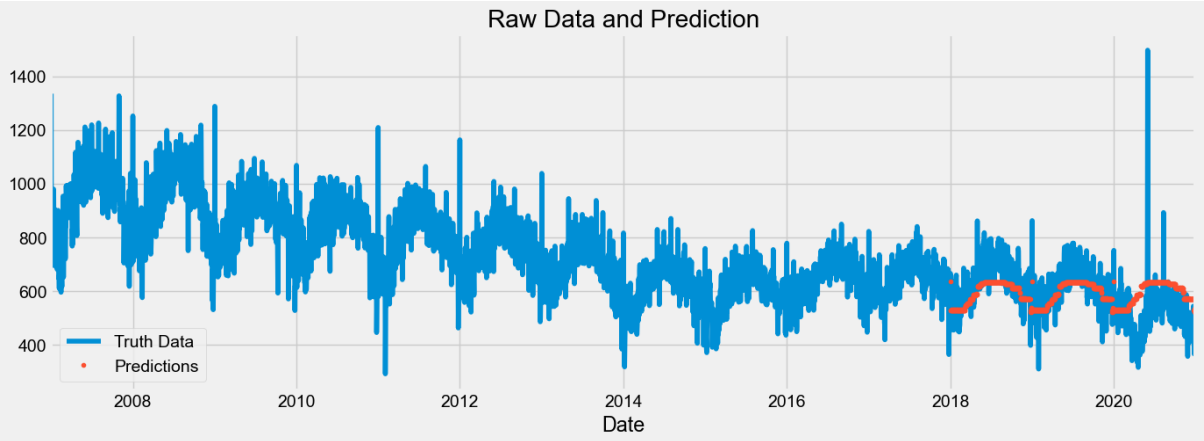

Figure 8: XG Boost Model Forecast

6.3.4 Random Forest

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It is an approach to group learning. It may be used to ML issues involving both classification and regression. However, by manually adding lag variables and seasonal component variables, it may also be utilized in time series forecasting, both for univariate and multivariate datasets. Since the p-value is less than 0.05, the cleaned weekly crime dataset is now obtained and checked to see whether the data is stationary. Because we are projecting for weeks, lag variables are established by setting the value as 54. Keeping and testing data are separated in the dataset. I used the most recent 54 weeks (1 year) as the testing dataset and the remaining data as the training dataset before using the [.fit()] function to fit the model. I have reduced the number of independent variables/features to 7 using RFE (recursive feature elimination). I used 100 estimates for the number of trees in the forest (n estimators). accuracy was 92.84% and mean absolute error was 476.65 after model assessment.

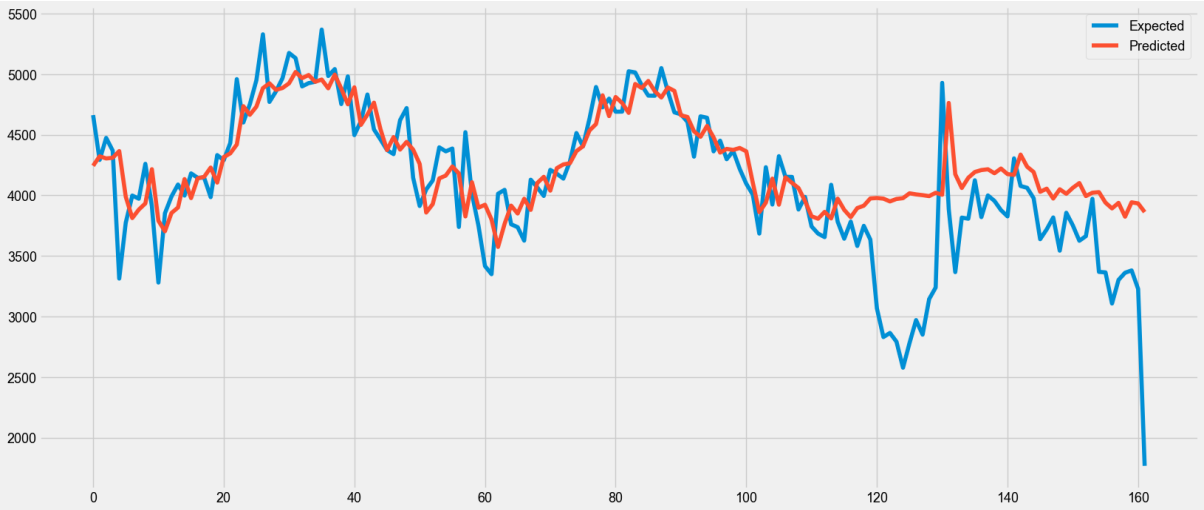

Figure 9: Random Forest Model Forecast

7 Evaluation

This section will examine the thorough examination of all models that have been used. The performance indicators that will be used to evaluate each model's effectiveness in predicting crimes are RMSE, MAPE, and MAE.

7.0.1 RSME

A performance indicator called Root Mean Squared Error (RMSE) is used to calculate the standard deviation of errors. The performance of the model improves with decreasing RMSE values.

7.0.2 MAE

A performance indicator called mean absolute error (MAE) calculates the absolute average by adding the differences between the original and predicted values and dividing the result by the total number of rows.

7.0.3 MAPE

The accuracy of the model after forecasting is demonstrated by the Mean Absolute Percentage Error (MAPE), a performance parameter that assists in calculating the error in the projected values. The model has better prediction/forecasting capacity the lower the MAPE value, which also shows the model to be more accurate.

7.1 Experiment 1 – Evaluation of SARIMA Model

In this experiment, the SARIMA model was trained with order (1,1, 1) and seasonal order (1,1,1,52) using pre-processed data as an input, and the latter model was fitted. A prediction was performed using test data after the SARIMA model had been trained and evaluated, and the results are reported in figure 10 for the metrics RMSE, MAE, and MAPE. Additionally, as shown in Figure, the original and predicted values were compared using the matplotlib package.

Metrics	MAE	RSME	MAPE
Values	299.99	444.21	13.53

Figure 10: Model Metrics Table - SARIMA Model

7.2 Experiment 2 – Evaluation of XG Boost Model

In this experiment, XGBoost model was trained and tested using pre-processed data as input, and the latter model was fitted. A prediction was performed using test data after the XGBoost model had been trained and evaluated, and the results are reported in figure 11 for the metrics RMSE, MAE, and MAPE.

Metrics	MAE	RSME	MAPE
Values	66.6	87.29	12.16

Figure 11: Model Metrics Table - XG Boost Model

7.3 Experiment 3 – Evaluation of Prophet Model

In this experiment, the PROPHET model was trained and tested using pre-processed data as input. A fresh data frame was produced with annual frequency and for durations equal to the length of the test data set once the model had finished training. The forecast

was made using test data that yielded the metrics values in figure 12 for RMSE, MAE, and MAPE, which are below.

Metrics	MAE	RSME	MAPE
Values	239.33	321.97	5.14

Figure 12: Model Metrics Table - Prophet Model

7.4 Experiment 4 – Evaluation of Random Forest Model

In this experiment, the Random Forest model was trained and tested using pre-processed data as input, and the latter model was fitted. A prediction was performed using test data after the Random Forest model had been trained and evaluated, and the results are reported in figure 13 for the metrics RMSE, MAE, and MAPE.

Metrics	MAE	RSME	MAPE
Values	294.51	432.13	8.48

Figure 13: Model Metrics Table - Random Forest Model

7.5 Discussion

The XG Boost Model performed better than the other models, with superior accuracy giving RMSE, MAPE, and MAE metrics, after examining all four of the models mentioned above. When compared to other models, PROPHET provided a decent MAPE value, but it fell short of the XG Boost model's output. In terms of the performance measures employed, the SARIMA model performed poorly in comparison to the other three models.

8 Conclusion and Future Work

Finding the best forecasting model for crime prediction using real-world data was the main goal of this study, and the findings may be used to create new policies or modify those that already exist to combat crime. Data for this study was gathered from the Chicago Data Portal website, which contains information on crimes from 2001 through 2022. Additionally, utilizing the four machine learning models SARIMA, Random Forest, XGBoost, and PROPHET, forecasting for global data was carried out by filtering the data for the top crimes together. RMSE, MAE, and MAPE performance indicators were employed to assess the models' performance. Each model was trained and tested, and it was discovered that the XGBoost model performed the best. For the years 2020 to 2021, weekly crimes were predicted using the XGBoost model. Here, the goal is accomplished by locating the best model and forecasting upcoming crime. Future studies may compare the

performance using more models like SVM and KNN. More historical data, particularly weekly or monthly data, would be necessary.

References

Adesola, F., Azeta, A., Misra, S., Oni, A., Ahuja, R. and Omolola, A. (2022). Spatial analysis of violent crime dataset using machine learning, *Emerging Technologies for Computing, Communication and Smart Cities*, Springer, pp. 183–191.

Alves, L. G., Ribeiro, H. V. and Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning, *Physica A: Statistical Mechanics and its Applications* **505**: 435–443.

Bappee, F. K., Soares Júnior, A. and Matwin, S. (2018). Predicting crime using spatial features, *Canadian Conference on Artificial Intelligence*, Springer, pp. 367–373.

Brantingham, P. J., Valasik, M. and Mohler, G. O. (2018). Does predictive policing lead to biased arrests? results from a randomized controlled trial, *Statistics and public policy* **5**(1): 1–6.

Chen, P., Yuan, H. and Shu, X. (2008). Forecasting crime using the arima model, *2008 fifth international conference on fuzzy systems and knowledge discovery*, Vol. 5, IEEE, pp. 627–630.

Ingilevich, V. and Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors, *Procedia Computer Science* **136**: 472–478.

Joshi, A., Sabitha, A. S. and Choudhury, T. (2017). Crime analysis using k-means clustering, *2017 3rd International conference on computational intelligence and networks (CINE)*, IEEE, pp. 33–39.

Kang, H.-W. and Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning, *PloS one* **12**(4): e0176244.

Kiani, R., Mahdavi, S. and Keshavarzi, A. (2015). Analysis and prediction of crimes by clustering and classification, *International Journal of Advanced Research in Artificial Intelligence* **4**(8): 11–17.

Kianmehr, K. and Alhajj, R. (2008). Effectiveness of support vector machine for crime hot-spots prediction, *Applied Artificial Intelligence* **22**(5): 433–458.

Kim, S., Joshi, P., Kalsi, P. S. and Taheri, P. (2018). Crime analysis through machine learning, *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, IEEE, pp. 415–420.

Krishnendu, S., Lakshmi, P. and Nitha, L. (2020). Crime analysis and prediction using optimized k-means algorithm, *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, pp. 915–918.

Malathi, A. and Baboo, S. S. (2011). An enhanced algorithm to predict a future crime using data mining.

McClendon, L. and Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data, *Machine Learning and Applications: An International Journal (MLAIJ)* **2**(1): 1–12.

Mittal, M., Goyal, L. M., Sethi, J. K. and Hemanth, D. J. (2019). Monitoring the impact of economic crisis on crime in india using machine learning, *Computational Economics* **53**(4): 1467–1485.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in chicago, *International Journal of Forecasting* **30**(3): 491–497.

Mukherjee, A. and Ghosh, A. (2022). Predictive geospatial crime data analysis and their association with demographic features through machine learning approaches, *International Conference on Computational Intelligence in Pattern Recognition*, Springer, pp. 545–557.

Tasnim, N., Imam, I. T. and Hashem, M. (2022). A novel multi-module approach to predict crime based on multivariate spatio-temporal data using attention and sequential fusion model, *IEEE Access* **10**: 48009–48030.

Wang, H., Kifer, D., Graif, C. and Li, Z. (2016). Crime rate inference with big data, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 635–644.