CRIME ACTIVITIES TRENDS ANALYSIS WITH AMACHINE LEARNING TECHNIQUE

Francis Ayiah-Mensah[1] and Sadia Ibrahim Abugbilla [1]

*[1] Department of Mathematics, Statistics and Actuarial Science, Faculty of Applied Sciences, Takoradi Technical University, Ghana

Abstract

Criminal activity is a major problem faced by societies worldwide. Identifying factors and patterns that contribute to criminal activity can help law enforcement agencies to prevent and reduce crime. The study purposewas to identify the trends in the rate of criminal activities over a given period of time and proposed a model for the data. It has been noted that both ARIMA(3,0,0) and SARIMA (3,0,0)(1,0,0)12 models are good for the crimedata. The model Root Mean Square Error (RMSE) of 2.813 shows that that the predicted values are about 97% close to the actual dataset, indicating a good model. Again, the Mean Absolute Percentage Error (MAPE) of 2.048 means the model prediction accuracy is about 98%. Thus, we are certain that our forecasted values with themodel are a true reflection of the nature of the crime rate for some periods ahead. In addition, the result shows anup-rise in the crime rate. This means that criminal activities will continue to rise if proper measures or policies are not implemented to curb them.

Keywords: Criminal, Effective strategies, Criminology, Differencing, Auto Regressive

1 Introduction

Criminal activities refer to any behavior that is prohibited by law and can result in punishment by the state. Criminal activities can range from minor offenses such as traffic violations to serious crimes such as murder and terrorism.Understanding criminal activities is an important aspect of criminology which helps in developing effective strate-gies to prevent and control crime. This essay discusses criminal activities and their typologies (Brantingham, P.
J. (2016). Criminal activities can be categorized into different typologies based on their nature, level of harm, and impact on society. The most common typologies of crime include violent crimes, property crimes, white- collar crimes, drug-related crimes, and organized crimes (Barkan, Bryjak, 2011). Violent crimes involve the use of physical force or threat to cause harm to others and include crimes such as homicide, assault and rape. Propertycrimes involve the stealing or damaging of someone else's property and include crimes such as theft, burglary, and vandalism. White-collar crimes are non-violent offenses committed by individuals in business or government positions for personal gain, such as fraud, embezzlement, and corruption. Drug-related crimes involve the illegal production, distribution, and use of controlled substances. Criminal organizations plan, coordinate, and execute organized crimes for profit, power, or control. Criminal activities significantly impact society and can result in severe consequences for both victims and offenders. Criminal activities lead to a loss of life, property, and social stability, affecting individuals' and communities' quality of life. Criminal activities also threaten national securityand economic development by creating fear, disrupting social order, and increasing social costs (Siegel, 2013). Inaddition, the criminal justice system plays a crucial role in controlling crime by enforcing laws, punishing offend-ers, and deterring others from committing similar offenses. Criminal activity analysis and forecasting in Africais largely focused on understanding the factors and patterns contributing to the region's crime. This involves a combination of statistical analysis, mapping, and social and cultural research to identify the underlying causes of criminal behavior and predict future trends. Some of the key factors that have been identified in the literature in- clude poverty, unemployment, inequality, corruption, political instability, and weak law enforcement systems. Onestudy conducted in South Africa found that higher crime levels were associated with poverty, a lack of economicopportunities, and a culture of violence and social dislocation (Lund and Sinclair, 2016). Another study conductedin Lagos, Nigeria, identified factors such as education, employment, and family structure as determinants of crimi-nal activity (Osinubi and Dada-Adegbola, 2015). Regarding forecasting, research has shown that criminal activityis often concentrated in particular areas or neighborhoods and that predictive modeling can be used to anticipate where crime is likely to occur next (Gill and Wong, 2016). Similarly, spatial analysis and mapping techniques can be used to identify hotspots of criminal activity and predict where crime is most likely to occur in the future (Brantingham and Brantingham, 2016). Overall, criminal activity analysis and forecasting in Africa is an impor- tant area of research that can inform policy and interventions to reduce crime rates and improve public safety. It requires a multifaceted approach that considers social, economic, political, and cultural factors, as well as the useof cutting-edge research methods and technologies. The study intended to identify the trends in the rate of criminalactivities over a given period of time and assessed a model with a machine learning technique.

In recent times, it is seen that artificial intelligence has shown its importance in almost all fields, and crime prediction is one of them. However, it is necessary to maintain a proper database of the crime that has occurred, as this information can be used for future reference. The ability to predict the crime that can occur in the future can help law enforcement agencies in preventing the crime before it occurs. The capability to predict any crime on the basis of time, location, and so on can help provide useful information to law enforcement from a strategic perspective. However, predicting crime accurately is challenging because crimes are increasing at an alarming rate.Thus, crime prediction and analysis methods are very important to detect and reduce future crimes. Recently, many researchers have conducted experiments to predict crimes using various machine-learning methods and particularinputs. KNN, Decision trees, and some other algorithms are used for crime prediction.

The economic factor is one of the most prominent factors attributed to increased criminal activity. Davis and Dossett (2018) argue that poverty and unemployment are the primary factors that drive individuals to engage in criminal activities. The study further emphasizes that many people who turn to crime have limited employment op-portunities, forcing them to commit crimes, particularly for those with dependents. Similarly, Wilson et al. (2017)posit that the lack of stable employment opportunities can lead to frustration, which, in turn, leads to alienation andcriminal behavior. Therefore, unemployment and poverty create a conducive environment for criminal activities.

In addition to economic factors, social factors also contribute to increased criminal activity. According to Jensen and Brownfield (2016), individuals who are exposed to antisocial behavior are more likely to engage in criminal activities. For instance, children who experience abuse, neglect, or dysfunctional family structures are more likely to turn to a life of crime. Similarly, peer pressure and social networks can influence an individual's behavior and lead them to engage in criminal activities, as Piquero et al. (2016) noted. Furthermore, social factors,such as cultural values, lifestyles, and religious beliefs, can also shape an individual's criminal tendencies.

2 Material and Methods

Time-series analysis focuses on the time-related factors influencing a variable, and it is instrumental in forecastingfuture trends. It entails studying data for past trends, patterns, and fluctuations to make predictions for future behavior. Again, it can be analyzed in either a univariate or multivariate manner, using statistical models such as Auto Regressive Integrated Moving Average (ARIMA), Vector Auto regression Model (VAR), or GeneralizedAutoregressive Conditional Heteroscedasticity (GARCH) models.
A Random Walk is an example of a nonstationary process which reduces to a stationary one after differencing.
Therefore, it is nonstationary AR(1) process with the value of the parameter $\psi = 1$ to give the model,

$$Y_t = Y_{t-1} + Z_t, \qquad where Z_t \sim WN(0, \sigma^2). \quad (1)$$

The autocovariances of Equation (1) depend on time as well as on lag. However, the first difference $\nabla Y_t = Y_t - Y_{t-1}$ is a stationary process, as it is just the White Noise Z_t. An inclusion of WN in the ARMA class then ∇Y_t is an ARMA(0,0) process, or in ARIMA notation it is ARIMA(0,1,0) process as it is obtained after first order differencing of Y_t. Hence, a general time series model can be written as

$$Y_t = p_t + X_t, \quad (2)$$

where p_t is a polynomial of order k and X_t is a stationary process. Then Y_t is nonstationary which has a polynomial trend. Notwithstanding, we can detrend such a process by calculating the difference of order k. Thus, a polynomial $p(t) = \theta_0 + \theta_1 t + \theta_2 t + ... + \theta_k t^k$ gives

$$\nabla^k Y_t = k!\theta_k + \nabla^k X_t, \quad (3)$$

where $\nabla^k X_t$ is a linear combination of a stationary process, so it is stationary.
Assuming that a dataset has a seasonality without a trend, can be modeled as

$$Y_t = a_t + X_t, \quad (4)$$

where X_t is a stationary process. Also, the seasonality component is such that $a_t = a_{t-h}$, where h is the period length is written as $\sum_{k=1}^{h} a_k = 0$. The removal of the seasonal effect from the dataset by differencing at lag h result in the introduction of the lag-h operator written as

$$\nabla_h Y_t = Y_t - Y_{t-h} = Y_t - \Theta^h Y_t = (1 - \Theta^h) Y_t, \quad (5)$$

which when sunstituting Equation (4) yields

$$\nabla_h Y_t = a_t + X_t - a_{t-h} - X_{t-h} = \nabla_h X_t. \quad (6)$$

Therefore, the seasonality effect is taken away from the operation and leads to the introduction of the SeasonalAutoregressive Integrated Moving Average (SARIMA) model, denoted by ARMA(P,Q)h, which is of the form

$$\Gamma(\Theta^h) Y_t = \Lambda(\Theta^h) Z_t, \quad (7)$$

where $\Gamma(\Theta^h)Y_t = 1 \quad \Gamma_1(\Theta^h) \quad \Gamma_2(\Theta^{2h}) \quad ... \quad \Gamma_p(\Theta^{ph})$ and $\Lambda(\Theta^h) = 1 + \Lambda_1(\Theta^h) + \Lambda_2(\Theta^{2h}) + ... + \Lambda_p(\Theta^{ph})$ arethe seasonal AR operator and the seasonal MA operator, with seasonal period of length h respectively.

3 Results and Discussion

Table 1: **Descriptive Statistics of Crime Rate**

Crime Rate		Statistics
N	Valid	36
	Mission	0
Mean		90.73
Std. Error of Mean		3.381
Median		94.80
Model		46
Std. Deviation		20.285
Variance		411.473
Kurtosis		-.633
Minimum		46
Maximum		122
Range		76
Skewness		-.474

Source: Field data, (2023)

The result from Table 1 includes details on the characteristics of the data for the three-year span. In general, thedataset's mean (average) crime rate is estimated to be about 90.73, with a standard error of 3.381 and a median of 94.80. However, the overall data set's standard deviation is around (20.285). On the other hand, the minimum andmaximum documented crime rates were 46 and 122, respectively. Kurtosis measures the shape of the distribution.The crime rate-based time series graphic in Figures 1 and 2 show how the complete dataset behaves across evenly spaced time intervals. It was noted that there is a distinct, systematic pattern in the incidence of crimeamong people. Additionally, it is evident from Figure 1 that the number of illegal acts only exhibits an increasingtrend. The narrative, which explores the idea of time sequence, delivers a good idea. To enhance the stationarityand statistical features of the series, we apply differencing and logarithm to the crime rate data. This enable a more thorough examination of the underlying patterns, trends, and anomalies in the crime rate. It also improvesthe precision and dependability of forecasting models constructed using the converted data.

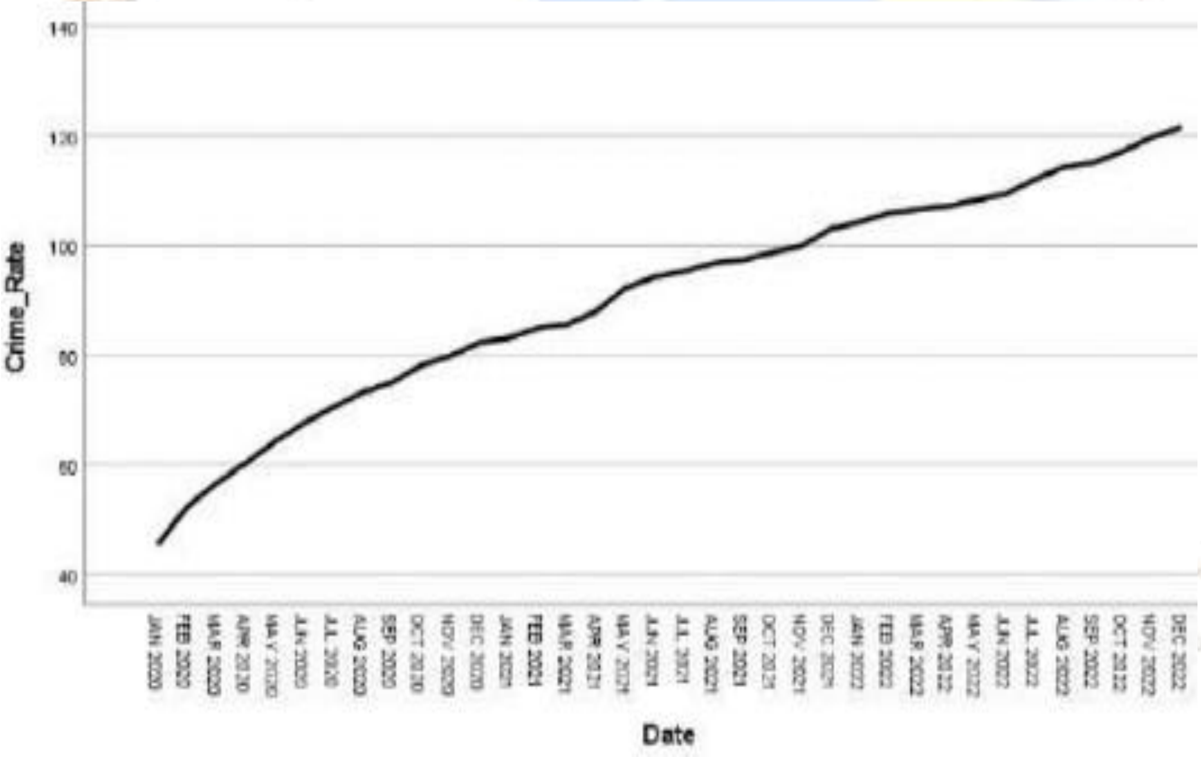

Figure 1: Trend of crime rate for 2020-2023.

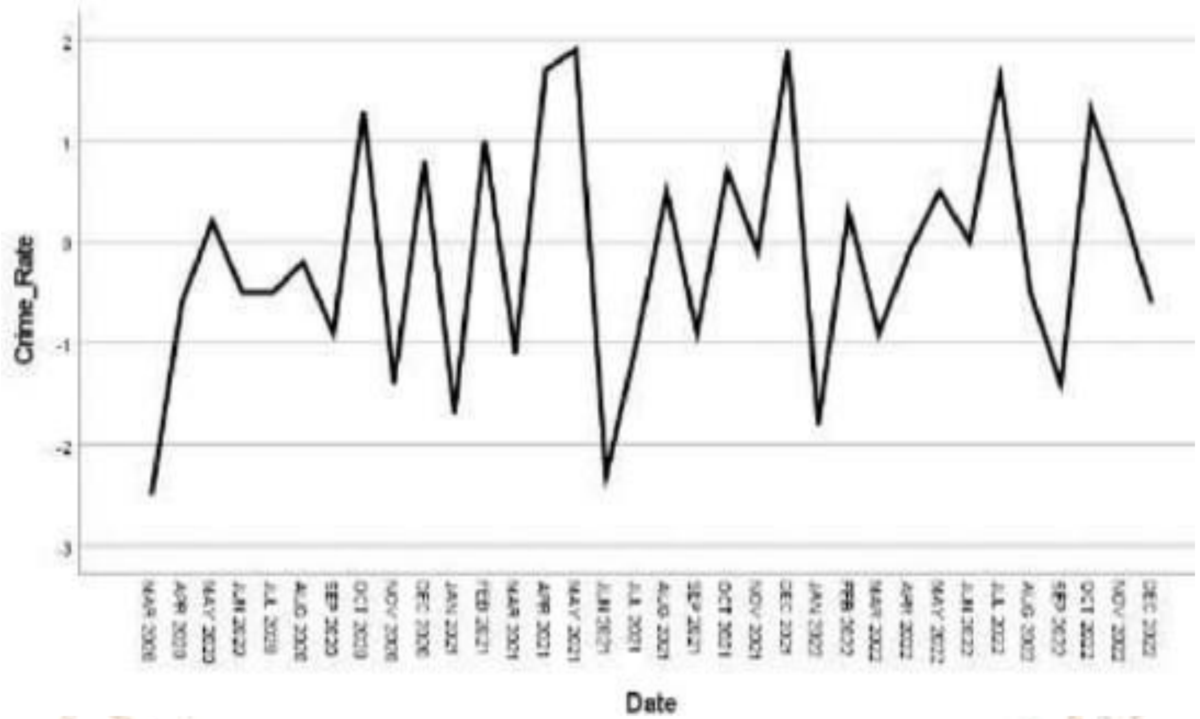

Figure 2: Plot of Natural log of crime rate for 2020-2023.

The Figure 2 is evident that the crime rate began to rise in February 2020 and continued to rise from May to July 2021 until another spike occurred from November to January 2022. As a result of this investigation, it was demonstrated that the trend in the crime rate was in line. Feb. 2020, Mar. 2021, Jul., Nov., and Dec. 22 were the months with the highest crime rates.

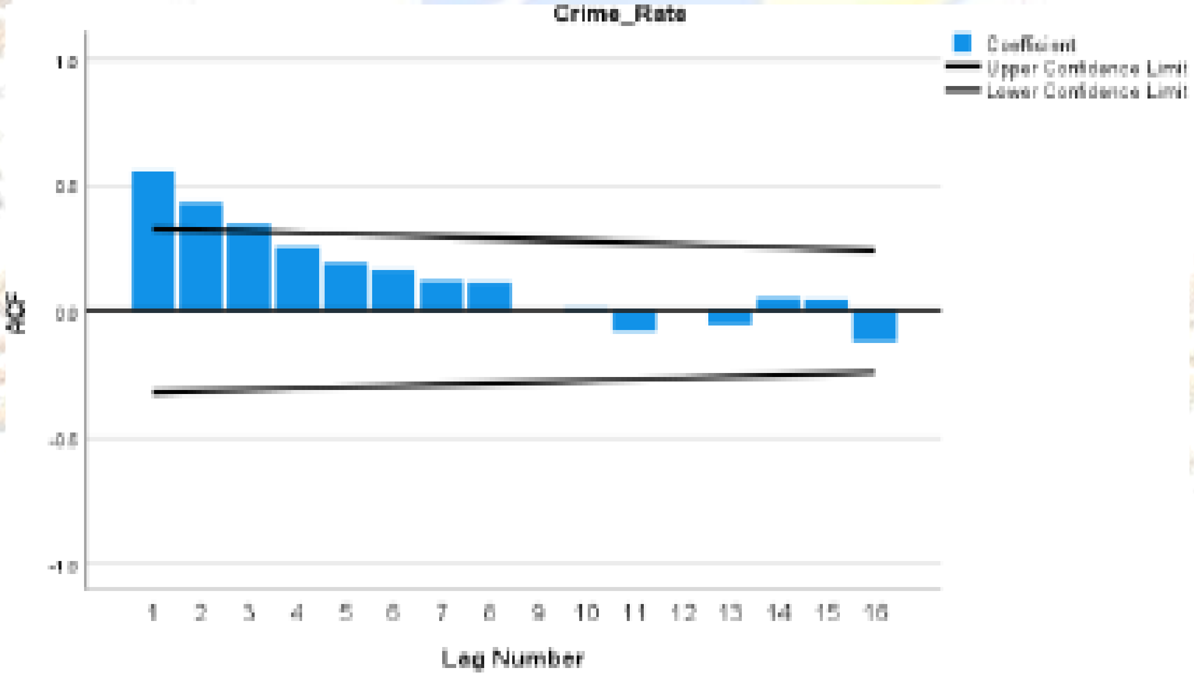

Figure 3: Autocorrelation Function Plot of Crime Rate

The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of crime Rate correl-ograms are in Figures 3 and 4. These allow us to measure the time series against its previous lagged values. Whena time series model contains its own previous values, it distorts the model's prediction accuracy. This causes the Mean Absolute Percentage Error (MAPE) to be bigger. In other words, the accuracy of the model in relation to the actual values and the predicted values as significantly bigger. This should be avoided because errors in the previous lagged values could be transferred to the model, making prediction with the model somehow inaccurate.

Table 2: **Model Statistics of Crime Rate**

Characteristics		**Statistics**
Number of predictors		3
Model Fit Statistics	Stational R squared	.926
	Normalized BIC	4.306
Ljung-Box Q(18)	Statistics	2.552
	R squared	.980
	RMSE	6.078
	DF	15
	sig	1.000
Number of outlieRS		0

Hypothesis Testing:
H_0: Model does not contain previous lagged errors.
H_1: Model contains previous errors.

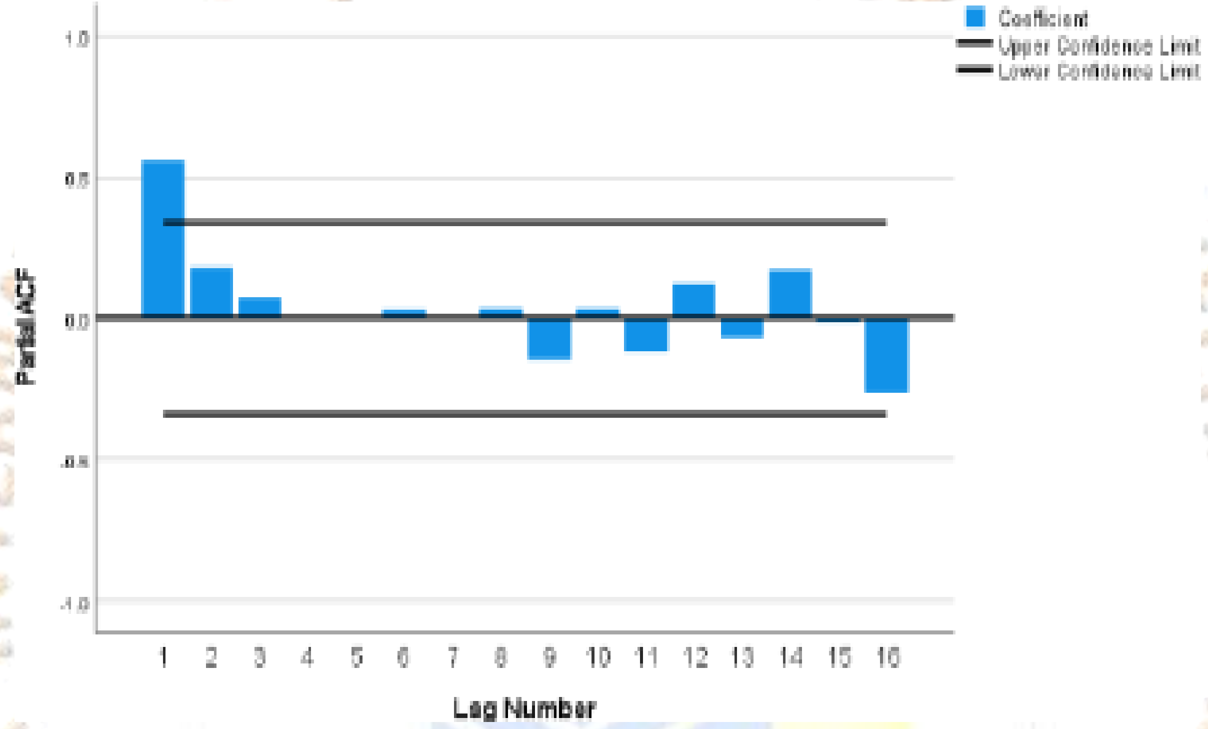

Figure 4: Partial Autocorrelation Function Plot of Crime Rate

The model identification supplied in the model description table is used to label the results for the model. Three predictors are included in the model. Therefore, it would seem that the Expert Modeler has discovered three independent variables that might be beneficial for predicting crime rate. Despite the fact that the Time Series Modeler provides a variety of goodness-of-fit statistics, we only chose the stationary R-squared value. In comparison to regular R-squared (.926), this statistic offers an estimate of the percentage of the overall variance inthe series that the model explains. A better fit is indicated by higher stationary R-squared values (up to a maximumvalue of 1). The model performs a fantastic job of describing the observed variance in the series, as shown by a value of 0.926. The modified Box-Pierce statistic, commonly referred to as the Ljung-Box statistic (18), gives a clue as to whether the model is properly described. A significance level of 0.05 or below suggests that the model cannot fully explain the observed series' structure. We can be sure that the model is accurately described becausethe value of 1.000 shown here is not significant. That's why we fail to reject H0 and conclude that the Model is appropriate for the prediction crime rate. The Lyung-Box Q test statistics significant value is compared with an alpha value of 0.05. From the model statistics output, the Lyung-Box Q significant value (0.717) is greater than the alpha 0.05 which means that we fail to reject the null hypothesis or there is no significant evidence to reject thenull and conclude that the model does not contain any previous lagged errors.

Table 3: **Model SummaryModel Fit**

Fit Statistic	Mean	SE	Minimum	Maximum
Stationary R-squared	.926	.	.926	.926
R-squared	.926	.	.929	.926
RMSE	6.078	.	6.078	6.078
MAPE	3.013	.	3.013	3.013
MaxAPE	69.3826	.	69.3826	69.382
Normalized BIC	4.306	.	4.3069	4.306

Source: Field data, (2023)

The model summary in Table 3 shows the summary of a statistical model. The model's performance is eval- uated using several metrics. However, the above table gives almost all the summary statistics. But the most important one is that the model fit indicated higher stationary R-squared values. This indicates that the model describes the observed variance in the series, as shown by a value of 0.926. The R squared value, which is 0.926,represents the proportion of the variance in the dependent variable explained by the model's independent variables.The Mean Absolute Percentage Error (MAPE) measures the average percentage difference between predicted andactual values. The value of 3.013 indicates that, on average, the model's predictions deviate by approximately 3.013% from the true values. Root Mean Square Error (RMSE) measures the average magnitude of the model's forecasting errors. The value of 6.078 indicates the average difference between the predicted and actual valuesof the variable being forecasted. A lower RMSE value indicates better accuracy. Also, from the table, the model statistics output gives a high Stationary R-squared of 0.980 and R-squared of 0.926, showing that the model and theindependent variables explain the majority of the variances in the model. The Root Mean Square Error (RMSE) isthe difference between the actual and predicted values; this indicates how far or close the predicted values deviatefrom the observations. The model has Root Mean Square Error (RMSE) 3.023, which suggests that the predicted values are about 97% close to the actual dataset, indicating a good model. Again, the Mean Absolute Percentage Error (MAPE) is a measure of the accuracy of the model for prediction. The MAPE 2.161 from the output meansthe model prediction accuracy is about 98%. Thus, we are certain that our forecasted values with the model area true reflection of the nature of the crime rate for some periods ahead. The $SARIMA(3, 0, 0)(1, 0, 0)_{12}$ is used,

Table 4: **SARIMA Model Statistics**

Characteristics		**Statistics**		
Number of predictors		0		
Model Fit Statistics	Stational R squared	.983		
	Normalized BIC	2.566		
Ljung-Box Q(18)	Statistics	12.755		
	R squared	.983		
	RMSE	2.813		
	MAPE	2.048	DF	14
	sig	546		
Number of outlieRS		0		

Source: Field data, (2023)

which has Stationary R-squared and R-squared values of 0.983, implying that 98% of the variabilities in the data are explained by the model together with the independent variables. SARIMA model has a smaller BIC value, alsothe Root Mean Square Error (RMSE) of the model is 2.813, which is a strong indication of the closeness of the predicted values to the actual data. It can be said that the predicted values of the SARIMA model deviated from the actual data only less than 3%, which gives enough confidence about the model's small margin of error.

Again, the SARIMA model; more MAPE of 2.048 or 98%. The Mean Absolute Percentage Error (MAPE) indicates the percentage of deviation of the model's prediction accuracy. It can therefore be concluded that we are98% confident about the prediction accuracy of the model. From the output, the model is not serial.

Correlated (the Lyung-Box Q test against α (0.05), which means no lag errors are associated with the model.

4 Conclusion

Criminal activity analysis and forecasting are critical areas in crime prevention. By identifying the factors and patterns associated with specific types of crime, can lead to effective prevention strategies and tailor interventionsto the needs of our communities. The model RMSE of 2.813 shows that that the predicted values are about 97% close to the actual dataset, indicating a good model. Again, the Mean Absolute Percentage Error of 2.048 means the model prediction accuracy is about 98%. Thus, we are certain that our forecasted values with the model are a truereflection of the nature of the crime rate for some periods ahead. This means that criminal activities will continueto rise if proper measures or policies are not put in place to curb its rise. A multidisciplinary approach in addressingcriminal activity analysis and prediction is recommended. In addition, a collaboration between different experts is necessary for successful forecasting and intervention. Ultimately, criminal activity analysis and forecasting offergreat potential to reduce crime in our communities.

References

[1] Brantingham, P. L., Brantingham, P. J. (2016). Crime pattern theory. Routledge.

[2] Gill, M., Wong, K. (2016). Handbook of security. Springer.

[3] Lund, F., Sinclair, M. (2016). Poverty and crime in South Africa: Some observations and preliminary analy-sis. La Trobe University

[4] Osinubi, T. S., Dada-Adegbola, Y. A. (2015). Factors influencing criminal behavior in Lagos metropolis,Nigeria. Journal of Law and Criminal Justice, 3(1A), 10-18.

[5] Barkan, S. E., Bryjak, G. J. (2011). Fundamentals of criminal justice. Jones Bartlett Publishers

[6] Davis, M. L., Dossett, D. L. (2018). The impact of economic conditions on criminal participation. Journal ofLabor Research, 39(1), 20-35.

[7] Jensen, G. F., Brownfield, D. (2016). Crime, delinquency, and offender treatment. Routledge.

[8] Piquero, A. R., Farrington, D. P., Jennings, W. G. (2016). Criminal behavior: A psychological approach.Routledge.

[9] Wilson, D., Roque, M. B., Lumbad, A. M. (2017). Exploring the relationship between poverty and crimerates in the Philippines. International Journal of Police Science Management, 19(1), 34-45width=!,height=!,