International Journal of Research in Engineering, Science and Management
Volume-3, Issue-7, July-2020
journals.resaim.com/ijresm | ISSN (Online): 2581-5792

Analyzing and Predicting Cyber Hacking with Time Series Models

C. Soundarya[1*], S. Usha[2]
[1]*PG Scholar, Department of Computer Science and Engineering, Anna University, BIT-Campus, Tiruchirappalli, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, Anna University, BIT-Campus, Tiruchirappalli, India*
*Corresponding author: csoundarya1995@gmail.com

***Abstract*: Cyber hacking implies utilizing PCs to fake action, for instance misrepresentation, security intrusion, taking corporate, individual information, and so on. Analyzing and predicting cyber hacking knowledge is a very vital technique for deepening our understanding of the evolution of the threat scenario. The different algorithms utilized for forecasting can be classified into linear models (AR, ARIMA, ARIMAX) and non-linear models (ARCH, GARCH, Neural Network). ARIMAX is one of the notable models dependent on time series which was utilized in the existing framework. In this system, SARIMAX (Seasonal Autoregressive Integrated Moving Average with exogenous) and RNN (Recurrent Neural Network) are proposed for forecasting the cyber hacking breaches from 2014 to 2017(4years) of a cyber-hacking dataset. These models are trained with the human health care services hacking breach dataset to forecasting the next eight years of the breach size and the incident arrival years based on the past values and also analyze the trend of the breach. The result obtained was compared to the SARIMAX and RNN models and it's been discovered that the recurrent Neural Network is outperforming the present linear model. This forecasting is very useful for making a decision about the security of an organization. In conclusion in phrases of these models to evaluate the superiority in forecasting performance.**

***Keywords*: Cyber hacking, SARIMAX, Time series, Forecasting, Predicting, RNN.**

1. Introduction

Data analytics is a crucial field in cyber security analysis as a result of deepening analyzing and predicting cyber hacking on our understanding of the evolution of the threat state of affairs. It refers to qualitative and quantitative methods and tactics used to lengthen productiveness and enterprise gain [1]. Data is extracted and labeled to pick out and analyze behavioral statistics and patterns, and strategies differ in accordance with organizational requirements. It includes 4 labeled they are descriptive, diagnostic, predictive and prescriptive analytics. Each kind has an extraordinary intention and an exceptional vicinity in the information analysis process. Predictive analytics helps reply questions about what will manifest in the future [1]. These methods use historic records to discover developments and decide if they are possibly to recur. Predictive analytical equipment furnishes treasured perception into what may additionally show up in the future and its methods consist of a range of statistical and laptop studying techniques, such as neural networks, selection trees, and regression.

Cyber Hacking is the method of gaining unauthorized access into a computer system or cluster of laptop systems. Hacking is additionally termed as a kind of cyber or web crime that is punishable by law. The cyber protection incidents can be attributed to the truth that our on-line [2] world is challenging to secure, which can be in addition attributed to the asymmetry that an attacker can be successful by using exploiting a single vulnerability however the defender has to block all vulnerabilities in order to guard a device against assaults. In the actual world hacking generally achieved on businesses due to the fact, the company invests tens of millions of bucks on perimeter safety purposes such as firewalls, anti-virus, and intrusion [3] detection device to decrease protection breaches from attackers such as hacking, phishing and spanning. The attackers typically focal point on private information breaches as a subset of cyber dangers the place a massive quantity of non-public records is exfiltrated from companies commonly for use in identification fraud.

Time series evaluation is broadly used in meteorological information analysis, monetary analysis, electromagnetic signal analysis, and many different fields. The most goal of time series modeling is to rigorously collect associated acute study the past observations of a statistic to [4] develop an applicable model that describes the inherent structure of the series. Statistics unremarkably use the forecast approach. It used to make predicts future information factors primarily based on located statistics over a period. It determines the overall performance of the time sequence one of them is the anticipated diploma of accuracy and the difference is anticipated demand [5]. The accuracy of time collection forecasting constitutes the most vital stage of many decision-making techniques. Proper care has to be taken to shape an ample model of the underlying time series. It is understood that a profitable time collection forecasting relies upon on a terrific model fitting. A lot of tries have been accomplished via scientific researchers over many years for the improvement of environment-friendly fashions to improve forecasting accuracy. As an outcome, more than a few essential time collection forecasting fashions have been advanced in the literature.

Many analysis studies on cyber hacking predictions have conducted with numerous resolution techniques proposed over the years [6]. The necessary methods fall into two vast categories particularly statistical and soft computing techniques. The SARIMAX model as a statistical approach additionally is aware of as the BOX-JENKINS model is generally used in evaluation and forecasting. The Seasonal Autoregressive Integrated Moving Average with exogenous (SARIMAX) is the extension of ARIMAX [7]. It is extensively viewed as the most environment-friendly forecasting approach in social science and is used substantially for a time series. This model is pretty bendy in that they can symbolize a number of special sorts of time sequences their main trouble is the pre-assumed linear structure of the model. This is a linear correlation shape that is assumed amongst the time series values and therefore, no non-linear patterns can be captured by using the SARIMAX model [8]. The approximation of linear fashions to complicated real-world issues is now not constantly satisfactory. RNN has to been proposed to handle these problems.

ANN's as a soft computing approach are the most correct and extensively used as forecasting models in many areas, along with social, engineering, economic, business, finance overseas alternate and inventory issues [9]. ANN's first introduced by Mcculloch and Pitts in 1943 has been undergoing an extended amount of development and has become the foremost distinguished design of deep learning networks for the [11] last 10 years. RNN's are one kind of neural community which due to one or greater connections between the neurons types cycles. It's the cycles in RNN that store the information and depart this world neuron's feedback to a different. These mechanisms repose on internal memory and this facilitates the educational of information that is successive in nature [10]. The RNN works best for short-termed time-sequence statistics and the decision of a suitable fee for the damping element can acquire notably correct and particular forecasting. LSTM could be a category of RNN capable of learning Long contained the long run memory of the previous decision state, that successively preserved the cell state one step before, all the manner back to the initial time steps [12]. Finally, RNN has been fund to be very environment friendly in fixing non-linear issues which includes these in the actual world. This is an awesome distinction to many typical strategies for time-series predictions.

2. Related Work

According to the PRC, over ninety reviews and articles reference the information used in our study. However, solely a few of these reviews performs quantitative analysis, and most do now not check out developments in the measurement or frequency of the statistics breaches.

Thomas Maillart et al. (2016). This study evaluated the personal data breaches data set from the organization using the model extreme heavy-tailed truncated Pareto distribution. It is a simple and very popular model with a power low probability tail. The tail exponent parameter decreasing linearly [13]. The risk is decomposed into two components one is large breaches and small breaches it's only depends on the threshold value if breaches size should be above the threshold will be referred to as large breaches and the if breaches size should below the threshold will be referred to as small breaches.

Maochao Xu et al. (2018). This learns about determined the end result of the mixture of facts of statistics breaches. It learn about investigates a new statistical mannequin multivariate cybersecurity risks, which are manifested via multivariate dependence between cyberattacks and predict the incoming assaults and the losses they incur. Because of its flexibility and functionality in estimating a giant range of parameters however the vine copula has been broadly used in many applications. Analyzing cybersecurity risks, which has three steps they are modeling marginal processes, modeling the dependence structure, estimating mannequin parameters. It determined the outcomes on the represent multivariate dependence between cyber-attacks has a large impact on the complete loss. characterized multivariate dependence between cyber-attacks has a vast impact on the complete loss.

Maochao Xu et al. (2107) This study proposed a novel application of marked point processes to fit and predict extreme cyber-attacks rates while using the value-at-risk(VaR) to measure the intense attacks in particularly using the point-over-threshold method to model the quality of extreme attack [1] rates and use the autoregressive conditional duration approach to describe the arrival of extreme attack rates. This analysis based on the two real-world data sets they are network telescope and honeypot that marked point processes offer accurate in -sample fitting performance and out-of-sample prediction performance.

Yuanhua et al. (2007) This study investigates the different financial time series data set using the SEAIFAR-GARCH model to found the results are trends, difference stationarity, and short and long-range dependence and heteroskedastic model errors. It is a very [14] powerful model for financial time series. The main disadvantage of only has a long memory in the mean but does not have a long memory in the volatility.

Rishabh et al. (2018) This paper research the computer network traffic to proposed the DWT (Discrete wavelet), ARIMA and RNN. The DWT decomposed the traffic data into linear and non-linear component after that component reconstructed using inverse DWT and then forecasting the future traffic using [10] ARIMA and RNN. The ARIMA for the linear data forecasting and the RNN for the non-linear data forecasting. The DWT techniques it can be easily employed at data center.

Hemantha et al. (2011) This paper developed a copula based it is a simulation approach for determining the annual net premiums for three different types of first-party damage cyber-insurance policies. [15] This study does not do predicting and analyzing the time series it just analyzing the loss of the insurance.

Lei Hua et al. (2017) This paper measure and predict the effectiveness of cyber defense mechanism. It proposed a vine copula model for analyzing the non-exchangeable rotationally symmetric [6] dependence structure. It accurately predicts the effectiveness of early-warning in real-time.

Martin et al. (2017) This study proposed the multidimensional scaling and goodness of fit test to analyze the distribution of data breach it found the two results one is a different type of data breaches need to be modelled as distinct risk categories and the second that skew-normal is a good distribution to the [17] amount of the data breach. It working with the cyber insurance policies data set. The analyzer focuses only on the number of data breaches and the amount of loss data but not on the real loss of data.

Arunabha et al. (2013) proposed a popular public dataset to model the Bayesian [3] Generalized Linear Models there has been little focus on trends in data breaches.

Shushan Li et al. (2010) This study found the results of the correlation function by [4] using the BP neural network can be a higher correlation coefficient but the BP neural does not predict the future.

Diane Ahrens et. al. (2016) This study evaluates a SARIMAX model to forecast the daily income of perishable ingredients in a retail store. SARIMA model regressed with the external variables such as charge reductions, holiday effects, and month results are proposed to conquer the disadvantage of the traditional SARIMA model. The comparison of SARIMAX adjusted R2 price to the SARIMA models. It shows the [22] end result from the overall performance of the SARIMAX version is highly accurate and is higher than the benchmark and SARIMA model. The benefit of the SARIMAX version lies in its explanation of the outlying data, which isn't accomplished through the SARIMA model.

3. System Architecture

The primary goal of this paper is to develop a SARIMAX and RNN modelling schemes for forecasting. It can be used for effectively and efficiently model for a large number of time-series data. Specifically devoted our effort toward the following six objectives in designing modelling schemes. i) Data collection, ii) Data pre-processing, iii) Split the data, iv) Model fitting, v) Forecasting, and vi) Trend analyzing.

A. Data Collection

The Human Health care service application dataset collected from the PRC website. This dataset covered the period of 2014 to 2017 and more responsible parameters for the dataset are Incident Dates, Records, and Categories. Next following procedures are adopted at this stage of the research.

B. Data Pre-processing

This technique is used to transform the raw data into a useful efficient format of the dataset which means cleaning, transformation, and reduction. Cleaning unwanted columns, missing data, and noisy data into the dataset. Transformation is taken in order to normalize the data values in a specified range. Reduction means to set the p-value to the attribute. If the p-value is less than a significant level 0.5 so it does not discard the attribute.

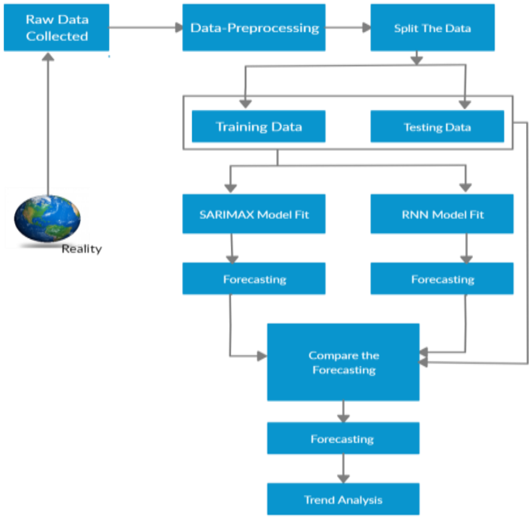


Fig: 1. Architecture Diagram

C. Spliting the Data

Split the data around 20% to 80% between the testing and training stages. The training is actually used to build the models and the testing data are used to improve the model hyper parameters.

D. Model Fitting

First SARIMAX model fitting using the command like model.fit(). The Akaike information criterion is used to measure the goodness of fit of a statistical time series model. If the AIC value is less on the training of the model is a good fit for the processing. Second RNN model fitting using the command like model.fit(). The batch size is divided into a sample. If the sample loss value is low the model is a good fit for the processing.

E. Forecasting

Forecasting the breaches for the next eight years through the two models based on past observation. Then compare the two model forecasting performance via Root mean square value.

F. Trend Analysis

The trend Analysis technique is used to predict the future breach movement based on the recently observed trend data. The categories attribute is used to analyze the trend on the dataset. This attribute contains two variables they are HACK and PORT. The HACK means to breach the personal information through the online and the PORT means to breach the personal information through the Pen drive, USP Port, etc.

4. Methodology

A. SARIMAX

In this study, the SARIMAX model is used to forecast the time collection. The use of the Box-Jenkins SARIMA method and a couple of linear regression (MLR). The SARIMAX model is a SARIMA model with exterior variables, referred to as SARIMAX (p,d,q) (P,D,Q)S (X), the place X is the vector of exterior variables. The exterior variables can be modelled by using multilinear regression equation is expressed as

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \omega_t \qquad (1)$$

where, $X_{1,t}$, $X_{2,t}$, … . $X_{k,t}$ are observations of k number of exterior variables corresponding to the structured variable Y_t ; β_0 , β_1 , … . β_k regression coefficients of exterior variables; ωt is a stochastic residual, i.e. the residual sequence that is unbiased of enter series. The residual sequence ωt can be represented in the shape of the ARIMA model as follows

$$\omega_t = \frac{\theta_q(B)\theta_Q(B^s)}{\varphi_p(B)\varphi_p(B^s)(1-\beta)^d(1-B^s)^{D^{\in_t}}} \qquad (2)$$

The popular SARIMAX model equation can be acquired with the aid of substituting equation 1 in equation 2

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + (\frac{\theta_q(B)\theta_Q(B^s)}{\phi_p(B)\phi_p(B^s)(1-\beta)^d(1-B^s)^{D^{\varepsilon_t}}} (1.3)$$

In this case of the model, the regression coefficient can be interpreted as in the common and simpler way. The universal SARIMAX model consists of 5 iterative steps.

a) Model-identification: This step entails the choice of the order of differencing (d), the order of seasonal differencing (D), seasonal size (S), non-seasonal autoregressive order (p), seasonal autoregressive order (P), non-seasonal transferring common order (q), and seasonal shifting common order (Q). Autocorrelations characteristic (ACF) and partial autocorrelations feature (PACF) are used to become aware of the model.

b) Parameter estimation: The parameters of the recognized model from step 1 are estimated.

c) Diagnosis the health of model: The model is recognized the use of Ljung-Box Q statistic to take a look at the adequacy. If the residuals are now not generally distributed, go to step 4.

d) Inclusion of exterior variables: The applicable exterior variables are protected in the SARIMA model the usage of linear regression. To diagnose the model, go to step 4.

e) Forecasting and validation: The identified model is validated the use of an out-sample. The validated model is used for forecasting future values.

f) Performance Measures: For the reason of evaluation, the out-sample facts is used to grant truthful cross-validation. In this study, root imply squared error (RMSE) is used as the overall performance measure of forecast accuracy.

$$\sqrt{\frac{\sum_{t=1}^{n}\left(Y_t - \overline{Y}_t\right)^2}{n}} \qquad (4)$$

where, Y_t is the forecast and n is the number of observations.

B. Recurrent Neural Network

RNNs are used to process sequential data, a sequence of inputs x_0,...., x_m, indeed, at the time t they calculate their output according to the input x_t but also the state of the hidden layer at the previous time. Thus, they develop an internal state that acts as short-term memory and makes it possible to take into account the temporal dependencies that the inputs manifest. The simplest RNNs are described as follows. Let: n, k, p ∈ N, x0, · · · , x_m where $x_t \in R^n$, $W \in M_{kn}(R)$, $W_h \in M_{kk}(R)$, $O \in M_{pk}(R)$ and $h-1 \in R^k$. Then, the dynamics of the associated RNN is described by:

$$\left.\begin{aligned} h_t &= \sigma\left(W_x + W_h h_{t-1}\right) \\ Y_t &= Oh_t \end{aligned}\right\} \qquad (5)$$

with $h_t \in R^k$ the state of the hidden layer and $y_t \in R^p$ the state of the output layer. In practice, we often take $h_{-1} = 0$. Long short-term memory is a gated memory unit for neural networks. It has 3 gates that manage the contents of the memory. These gates are simple logistic functions of weighted sums, where the weights might be learned by back propagation. It means that, even though it seems a bit complicated, the LSTM perfectly fits into the neural network and its training process. It can learn what it needs to learn, remember what it needs to remember, and recall what it needs to recall, without any special training 2 or optimization. The input gate (1) and the forget gate (2) manage the cell state (4), which is the long-term memory. The output gate (3) produces the output vector or hidden state (5), which is the memory focused for use. This memory system enables the network to remember for a long time, which was badly missing from vanilla recurrent neural networks.

$$\left.\begin{aligned} i_t &= sigmoid\left(W_i x_t + U_i h_{t-1} + b_i\right) \\ f_t &= sigmoid\left(W_f x_t + U_f h_{t-1} + b_f\right) \\ o_t &= sigmoid\left(W_o x_t + U_o h_{t-1} + b_o\right) \\ c_t &= f_t \Theta c_{t-1} + i_t \Theta \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\ h_t &= o_t \Theta \tanh\left(c_t\right) \end{aligned}\right\} \qquad (6)$$

Then going to obtain a train and a test set by splitting the time series at one point. It means that we always test the future. We would expect the algorithm to tell the future, so this choice of validation is natural for such forecasting problems. Yet, the test set consists of a single time period, so this method may not be entirely sufficient for evaluating the model performance. Bootstrapping provides an alternative validation set. An average bootstrap sample contains about 63.2% of the individual observations, or in our case, it contains about 63.2% of all available data sub sequences. The remaining subsequence do not participate in the training process, so we may use them for validation purposes. We are going to call the bootstrap left-out dataset validation set, and the future dataset test set it just for the sake of distinction. They have the same evaluation purpose. Both sets are going to be wont to evaluate the one-step-ahead forecasting ability of our recurrent neural networks. The separate test set, being a complete chronologically ordered series of subseries, may also be used for iterative multi-step-ahead forecasting. Regression and classification metrics are going to be applied in order to evaluate the forecasted values and the predicted changes of direction as well. To calculating the root mean squared error. The RMSE is expressed as

$$(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (7)$$

5. Experimental Results

A. Data Set

In this study, the human health care service application is used between 2014 to 2017. The average monthly hacking and also monthly trend of the breaches are calculated. The yearly breaches records of the product are shown in the fig. 1.

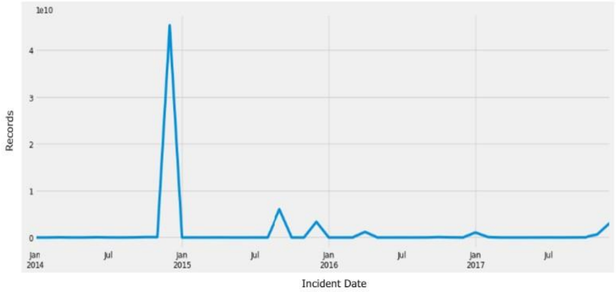


Fig. 1. Data breach sizes over a past 4year period

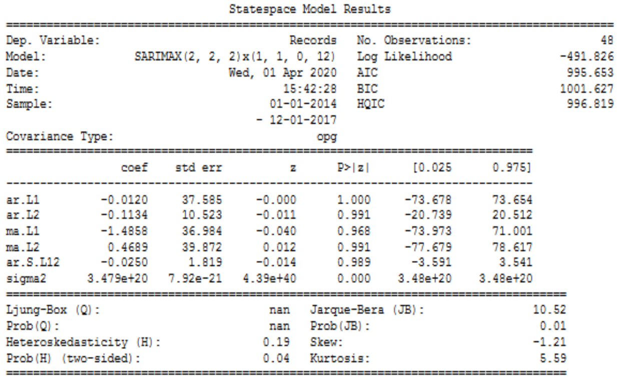
The breaches record over the past 4 years of data obtained from a popular dataset published by the Privacy Rights clearinghouse website. The PRC is a California non-profit corporation focused on issues of privacy. The dataset contains an attribute including the date the breach was responsible for the data, the type of entity breached, the total number of records breached, the location where the entity operates information on the source of the data, and a short description of the breach of the 4.602 breaches in the dataset, only those involving have associated record counts. To address these issues, adopt a statistical modelling approach and showing that in this dataset frequency of breaches has increased over time. We use a SARIMAX and RNN approach to construct accurate models without overfitting.

B. Sarimax Model Fitting and Diagnostics

The function SARIMAX was applied to fit the SARIMAX model. Relevant diagnostics packages were also used to test the model performance.

1) Model-Fitting

The SARIMAX model was classified by a different selection of orders, SARIMAX(p,d,q)(P, D, Q). The breach data have obvious yearly seasonality so s constantly equal 12 months.



```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                          Records   No. Observations:                     48
Model:             SARIMAX(2, 2, 2)x(1, 1, 0, 12)   Log Likelihood                -491.826
Date:                          Wed, 01 Apr 2020   AIC                            995.653
Time:                                  15:42:28   BIC                           1001.627
Sample:                              01-01-2014   HQIC                           996.819
                                   - 12-01-2017
Covariance Type:                            opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.0120     37.585     -0.000      1.000     -73.678      73.654
ar.L2         -0.1134     10.523     -0.011      0.991     -20.739      20.512
ma.L1         -1.4858     36.984     -0.040      0.968     -73.973      71.001
ma.L2          0.4689     39.872      0.012      0.991     -77.679      78.617
ar.S.L12      -0.0250      1.819     -0.014      0.989      -3.591       3.541
sigma2      3.479e+20   7.92e-21   4.39e+40      0.000    3.48e+20    3.48e+20
===================================================================================
Ljung-Box (Q):                       nan   Jarque-Bera (JB):                 10.52
Prob(Q):                             nan   Prob(JB):                          0.01
Heteroskedasticity (H):             0.19   Skew:                             -1.21
Prob(H) (two-sided):                0.04   Kurtosis:                          5.59
===================================================================================
```

Fig. 2. Summary of the SARIMAX Model

The likelihood techniques find the values of the parameters which maximize the probability of obtaining the data that have been observed. The logarithm of the probability of the observed data from the estimated model for given values of p,d, and q r will try to maximize the log-likelihood when finding parameter estimation. The multiple significant models are ranked by Akaike's information criterion (AIC) (2, 2, 2)x(1, 1, 0, 12)12 yields the lowest AIC value of 995.653. Therefore, it should consider this to be an optimal parameter. The Coefficient values and these standard errors of the fitted model. Divided the coefficient by their standard error to get the z-statistics and then calculate the p-values. Then compare the p-value for term significance level to assess the null hypothesis. The significance level of 0.05 indicated a 5% risk. P-value $\leq$ the term is statistically significant. If the p-value is less than or equal to the level that the coefficient statistically significant. The autoregressive term encompasses a p-value that's but the importance level of zero 0.05 that the constant for the autoregressive term is statistically important.

2) Model Diagnostics

Diagnostic checks for SARIMAX were conducted to research the residuals from the model fit to establish any signs of non-randomness. The residuals area unit uncorrelated and normally distributed with zero-mean. In this case, model diagnostics propose that the model residuals are normally distributed.

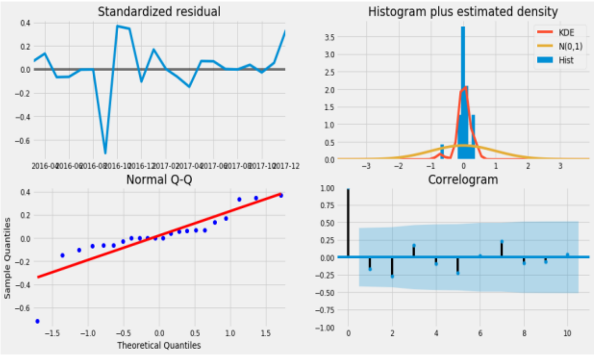


Fig. 3. SARIMAX Model Diagnostics

In the fig. 3, top-right plot, that the red KDE line follows closely with the N(0,1) line (where N(0,1) is that the standard notation for a traditional distribution with mean 0 and variance of 1. This is often an honest indication that the residuals are normally distributed. In the fig. 3 normal Q-Q-plot on rock bottom left shows that the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a typical Gaussian distribution with N(0, 1). Again, this is often a robust indication that the residuals are normally distributed. In the fig. 4, top left plot the residuals over time don't display any obvious seasonality and appear to be noise. This is often confirmed by the autocorrelation (i.e.correlogram) plot on rock bottom right, which shows that the statistic residuals have low correlation with lagged versions of itself. The diagnostic can visualize important information because of the distribution and therefore the Autocorrelation function ACF (correlogram). Values upward the "0" has some correlation over the statistic data. Values almost "1" demonstrate the strongest correlation.

3) Sarimax Model Forecasting

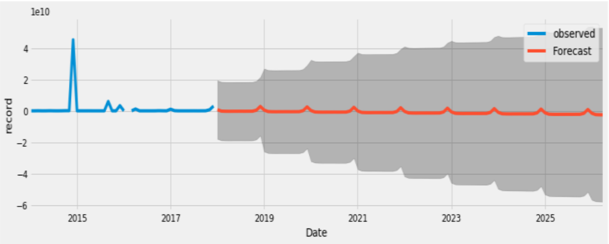


Fig. 4. Graph of observed value vs. forecast values for breach data using SARIMAX

The Human Health care Service series, shown in fig. 4, represents the monthly total number of breaches from January 2018 to December 2025. The plot shows the blue line indicates the observed data and the red line indicates the forecasted values from 2019 to 2025 within the confidence intervals of our forecast.

C. RNN Model

1) Model fitting

Training the RNN model by modifying the number of hidden layers to obtain the optimal result in forecasting, the optimization method used in training is Adam. The optimal lead to forecasting this will be demonstrated by evaluating the value of MSE on each model of training. To fit the random number seed to ensure our results are reproducible to extract the Numpy array from the data frame and convert the integer values to floating-point values which are more suitable for modelling with the neural network. The LSTMs are sensitive to the scale of the input data, sigmoid function to rescale the data to the range of 0 to 1 also called the normalizing. Then fit the LSTMs network it has a visible layer with 1 input, a hidden layer with 4 neurons and an output layer make a single value prediction. The network is trained for 40 epochs it denotes one full forward and backward pass through the whole dataset. Therefore, the number of epochs denotes how many such passes across the dataset are required for the optimal training of the RNN. Even within each epoch, the dataset is traversed the number of times denoted by the epoch size. A batch size of 1 is used it denotes the number of time series considered for each full back propagation in the RNN. This is a more limited version than using all the available statistic directly to perform one back propagation, which poses a big memory requirement. On the opposite hand, this might even be considered a more generalized version of the acute case which uses only one statistic per each full back propagation. Adam is an adaptive learning rate optimization algorithm is used for deep neural networks which estimate the first and second moments of the gradient to adapt the learning rate for each weight of the neural network.

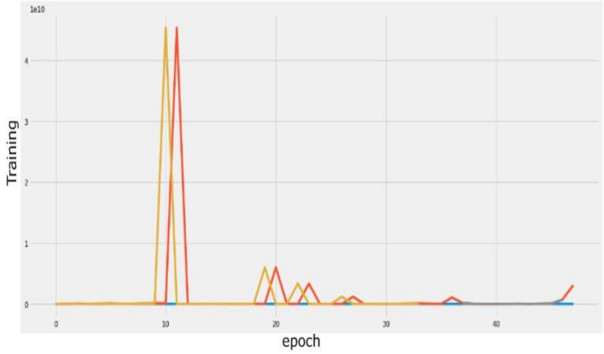


Fig. 5. Result of Iteration

In the fig. 5, the line plot of model accuracy on the train and test sets is created, showing the change in performance over all 15 training epochs. The red line show the training data and the yellow line is the testing data and the blue line is the normal observation of the data and the y-axis is training loss value and the x-axis is epoch.

2) RNN forecasting

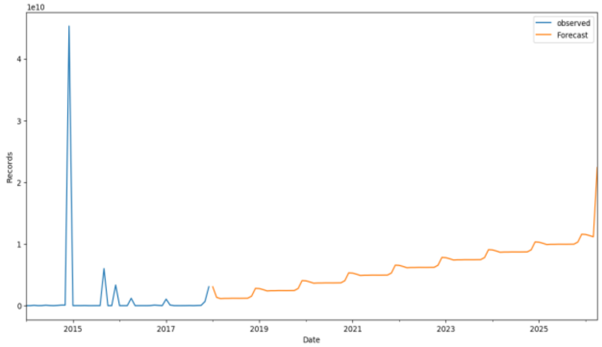


Fig. 6. Graph of RNN model of forecast values against observed values for Dell stock index

In the fig. 6, show, the result of RNN model forecasting the blue line is total breach records of the past data and the red line is future breach records of the data up to 2018 to 2025 years.

3) Comparison of SARIMAX and RNN models

The empirical result presented in fig. 7, show the comparison forecast of two model the red line is the SARIMAX model forecasting and the blue line is the RNN model forecasting the x-axis is breach records size and the y-axis is incident day of 100days.

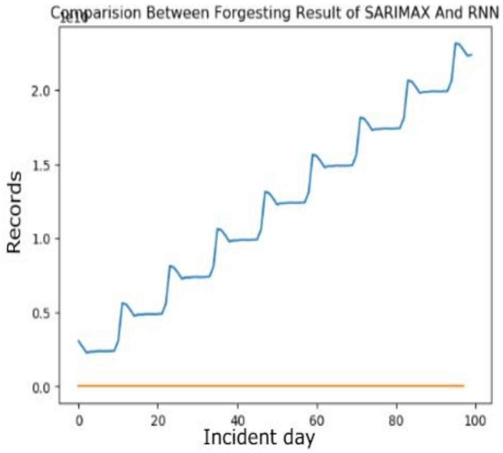


Fig. 7. Graph of forecast values of SARIMAX and ANN model

4) Performance Calculation

From the empirical results presented in Table 1 and Figure 8, It can be argued that RNN models achieved good forecast performance judging from the forecast root mean square error of both models which are quite low. The performance of the RNN model is better than the SARIMAX model in terms of forecasting accuracy on many occasions from the test data.

Table 1
Performance of the two model forecasting

	SARIMAX	RNN
RMSE	5671844275.38081	2.5233915188123723

In the table 1, the SARIMAX time series model RMSE value is too much higher than the RNN model so the Recurrent Neural Network is outperforming to the linear model.

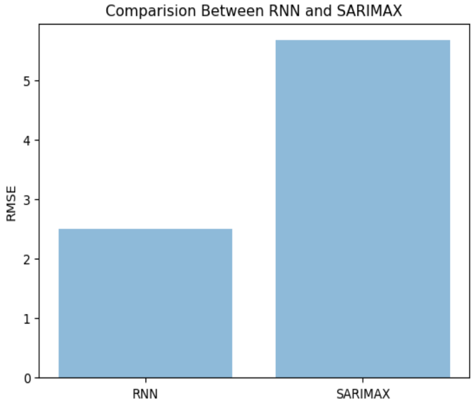


Fig. 8. Accuracy calculation

The results of fig. 8, show that the RNN model is better than the SARIMAX model for cyber hacking prediction.

5) Trend Analysis Results

Trend Analysis may be a statistical technique that tries to work out future movements of a given variable by analyzing historical trends. It would separate a time series into trend and seasonality, which might contain yearly. Analysis of the trend of human health care service breach dataset. It has categories attribute that includes hack and port. If the hack means the movement of breach the personal information through the online and the Port means the movement of breach the personal information through the pen drive, USP Port, etc.

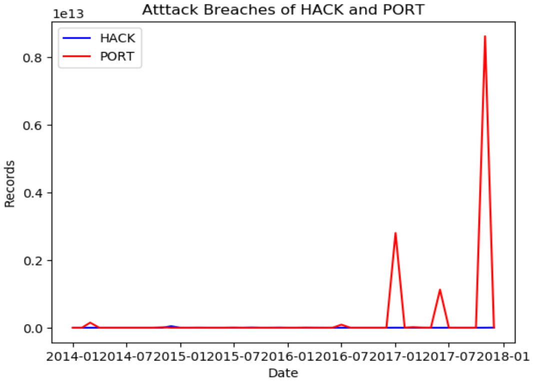


Fig. 9. Attack breaches on the dataset

In fig. 9, red line is the attacker highly breaches the records from 2014 to 2017 on the dataset through the port. The blue line is attacker breaches the records very low from 2014 to 2017 on the dataset through the online

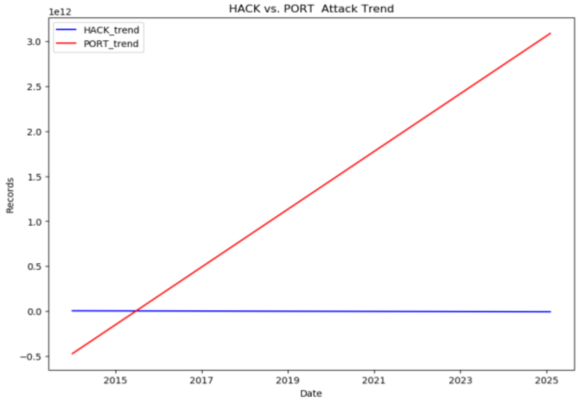


Fig. 10. Future Trend of attacker

In the fig. 10. future trend movement of the attacker from 2019 to 2025 breach records through the port is the high frequency on the dataset but the breach records through the online are low frequency on the dataset.

6. Conclusion

Cyber hacking breach forecasting plays an important key role in the success of business enterprises. This study proposed a time series forecasting techniques to predict the future breach size and Incident time based on past value. The ARIMA model is used extensively in the literature for making forecasting but it is linear in nature. The result obtained with published breach data on the performance of SARIMAX and RNN model to cyber hacking breach prediction and analysis of the trend of the breach has been presented in this study. The forecasting performance of the RNN model was compared with the SARIMAX model, the both SARIMAX and RNN model can achieve a good forecast in application to real-life problems. There are many interesting problems that are left for future studies. for example, it is both interesting and challenging to investigate how to predict the extremely large values and how to deal with missing data, etc.

References

[1] Peng, Chen, Maochao Xu, Shouhuai Xu, and Taizhong Hu. "Modeling and predicting extreme cyber-attack rates via marked point processes." Journal of Applied Statistics 44, no. 14 (2017): 2534-2563.
[2] Condon, Edward, Angela He, and Michel Cukier. "Analysis of computer security incident data using time series models." In 2008 19th International Symposium on Software Reliability Engineering (ISSRE), pp. 77-86. IEEE, 2008.
[3] Mukhopadhyay, Arunabha, Samir Chatterjee, Debashis Saha, Ambuj Mahanti, and Samir K. Sadhukhan. "Cyber-risk decision models: To insure IT or not?." Decision Support Systems 56 (2013): 11-26.
[4] Wang, Lei, and Shushan Li. "A revised approach to detect time series correlation." In 2010 Second International Conference on Intelligent Human-Machine Systems and Cybernetics, vol. 1, pp. 321-324. IEEE, 2010.
[5] Adhikari, Ratnadip, and Ramesh K. Agrawal. "An introductory study on time series modeling and forecasting." arXiv preprint arXiv:1302.6613 (2013).
[6] Adebiyi, Ayodele Ariyo, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. "Comparison of ARIMA and artificial neural networks models for stock price prediction." Journal of Applied Mathematics 2014 (2014).
[7] McKenzie, E. D. "General exponential smoothing and the equivalent ARMA process." Journal of Forecasting 3, no. 3 (1984): 333-344.
[8] Zhang, G. Peter. "Time series forecasting using a hybrid ARIMA and neural network model." Neurocomputing 50 (2003): 159-175.
[9] Khashei, Mehdi, and Mehdi Bijari. "An artificial neural network (p, d, q) model for time series forecasting." Expert Systems with applications 37, no. 1 (2010): 479-489.
[10] Madan, Rishabh, and Partha Sarathi Mangipudi. "Predicting computer network traffic: a time series forecasting approach using DWT, ARIMA and RNN." In 2018 Eleventh International Conference on Contemporary Computing (IC3), pp. 1-5. IEEE, 2018.
[11] Kuan-Cheok, L. E. I., and Xiaohua Douglas Zhang. "An approach on discretizing time series using recurrent neural network." In 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2522-2526. IEEE, 2018.
[12] Katarya, Rahul, and Shubham Rastogi. "A study on neural networks approach to time-series analysis." In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 116-119. IEEE, 2018.
[13] Wheatley, Spencer, Thomas Maillart, and Didier Sornette. "The extreme risk of personal data breaches and the erosion of privacy." The European Physical Journal B 89, no. 1 (2016): 1-12.
[14] Feng, Yuanhua, Jan Beran, and Keming Yu. "Modelling financial time series with SEMIFAR–GARCH model." IMA Journal of Management Mathematics 18, no. 4 (2007): 395-412.
[15] Herath, Hemantha, and Tejaswini Herath. "Copula-based actuarial model for pricing cyber-insurance policies." Insurance markets and companies: analyses and actuarial computations 2, no. 1 (2011): 7-20.
[16] Xu, Maochao, Lei Hua, and Shouhuai Xu. "A vine copula model for predicting the effectiveness of cyber defense early-warning." Technometrics 59, no. 4 (2017): 508-520.
[17] Eling, Martin, and Nicola Loperfido. "Data breaches: Goodness of fit, pricing, and risk measurement." Insurance: mathematics and economics 75 (2017): 126-136.
[18] Hiransha, M., E. Ab Gopalakrishnan, Vijay Krishna Menon, and K. P. Soman. "NSE stock market prediction using deep-learning models." Procedia computer science 132 (2018): 1351-1362.
[19] Chen, Yongsheng, and Stevanus Tjandra. "Daily collision prediction with SARIMAX and generalized linear models on the basis of temporal and weather variables." Transportation Research Record 2432, no. 1 (2014): 26-36.
[20] Petneházi, Gábor. "Recurrent neural networks for time series forecasting." arXiv preprint arXiv:1901.00069 (2019).
[21] Adebiyi, Ayodele Ariyo, Aderemi Oluyinka Adewumi, and Charles Korede Ayo. "Comparison of ARIMA and artificial neural networks models for stock price prediction." Journal of Applied Mathematics 2014 (2014).
[22] Arunraj, Nari Sivanandam, Diane Ahrens, and Michael Fernandes. "Application of SARIMAX model to forecast daily sales in food retail industry." International Journal of Operations Research and Information Systems (IJORIS) 7, no. 2 (2016): 1-21.
[23] Fathi, Oussama. "Time series forecasting using a hybrid ARIMA and LSTM model." In Velvet Consulting. 2019.
[24] Yi, Dokkyun, Sunyoung Bu, and Inmi Kim. "An Enhanced Algorithm of RNN Using Trend in Time-Series." Symmetry 11, no. 7 (2019): 912.
[25] Zhang, G. Peter. "Time series forecasting using a hybrid ARIMA and neural network model." Neurocomputing 50 (2003): 159-175.