Stock Market Price Forecasting using ARIMA vs ANN: A Case Study from CSE

Abstract — Main purpose of this study is to forecast stock price using Artificial Neural Network (ANN) and traditional time series approach such as ARIMA and identify most suitable approach for stock price prediction. Time series forecasting is regarded as the most successful criterion among several factors involved in the decision-making process to pick a correct prediction model. Improving predictability has become crucial for decision-makers and managers, especially time series forecasts, in various fields of science. Using K-mean clustering and Principle Component Analysis, dataset is clustered based upon a central point selection and the Euclidian distance measurement. According to clustering results identify the most contribution sector for the CSE and most contributing Company in selected sector within time period of 2008 to 2017. In particular, ARIMA has shown its success in accuracy and precision in predicting the next time-series lags. As part of the literature, very few studies have focused on Colombo Stock Exchange (CSE) to find new predictive approaches for the forecasting of high volatility stock price indexes. This article explores best sectors and company to invest according to previous contributions and whether and how the newly developed deep learning algorithms for the projection of time series data, such as the Back Propagation Neural Network, are greater than traditional algorithms. The results show that Deep learning algorithms like BPNN outperform traditionally based algorithms like the model ARIMA. For ARIMA and ANN, MAPE values are 0.4672206 and 0.1783333 respectively. MAE values are 29.6975 and 4.708423 respectively results for ARIMA and ANN. The MAE and MAPE values relative to ARIMA and BPNN, which suggests BPNN ’s superiority to ARIMA and we suggest ANN model best for forecast stock prices

Keywords — Artificial Neural Network, Auto Regression Integrated Moving Average, Colombo Stock Exchange

I. INTRODUCTION

Financial markets are the location of exchange of financial assets, such as securities, shares, moneys and goods. Some people are driven to trade because they expect their investment to make a return. The best and cheapest way to create a strong future for the country is to invest in stocks. Shares are financial assets that show a company's ownership (Simon, 2012). At every time when the market does their daily trading, millions of investors and trade firms, both locally and internationally, are directly interested. A stock market is a place where aggregation both buyers and sellers in a single platform for offering shares to the general public to raise their capital needed for restructuring, an expansion for new operations. The decisions of investors in the financial markets and the unpredictable existence of stocks make them very difficult and challenging to forecast.

(Ratnayaka et al., 2015) introduce ARIMA-ANN based hybrid approach for stock price prediction in CSE. In this research we focused to identify most contributing sector and company to the CSE that differ from the above research. And the MAPE and MAE of ARIMA and ANN model is relatively less compared to above research. There was very few studies to forecast stock price fluctuations on CSE. (Samarakwicrama et al., 2017) selected few companies randomly and using Recurrent Neural Network types such as Simple Recurrent Neural Network (SRNN), Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) architectures and Feedforward Neural Networks they predict the stock price. Results shows feedforward networks produced highest forecasted accuracy.

In time series predicting, ARIMA is commonly used. Certain methodologies under high volatility are not entirely suited. Linear models, such as linear regression, ARIMA, SARIMA, SARIMAX have the advantage of being very fast, but not precisely so. If a market reaction has a price level and begins to rise rapidly, the linear model does not quickly respond or predict the end of the trend. Likewise, if the price rise to a sudden fall called bubble, a linear model does not do well in this case. This article aims primarily to examine the best forecast model with lower forecast errors and higher performance of forecasting. Machine learning methods have introduced new approaches to problem prediction with deep, layering hierarchy based on the relationships between variables and more precisely deep learning algorithms.

Following research questions are addressed during this research. The way we can predict the stock price in some various sectors like banking sector, construction, plantation, etc. And which sector and company suitable for invest and best model for predict stock values.

The accuracy and consistency of traditional forecasting methods are important research questions compared to deep learning algorithms. To the best of our knowledge, there is no strong empirical evidence that the data production of the economic and financial time series has been predicted using BPNN to fit traditional econometric methods such as ARIMA.

II. OBJECTIVES

Goals and objectives of this research are;

- Use the section vice indices to predict like banking sector, plantation, construction etc.
- Identify best sector and company to invest.
- Finding a suitable approach to predict the long term and short-term price prediction.

III. METHODOLOGY

The main goal of this research is to apply a suitable approach to stock market prediction based on the Artificial Neural Network (ANN) and ARIMA. ARIMA and ANN methodologies are first defined in the methodology. We then find the model for the prediction of the stock price in CSE using different test precisions.
A. Autoregressive Integrated Moving Average Model (ARIMA)

ARIMA is a linear approach that is supposed to be linear in the future value of a variable for which past observations were supposed. The method consists of three main steps: parameter specification, calculation and prediction. Akaike information criterion (AIC), Schwarz criterion (SBIC) and Hannan-Quinn information criterion (HQIC) methods mainly used to select the best model (Hiransha et al., 2018).

The methodology is going under following steps.

B. The Artificial Neural Network (ANN) approach

In here we use 100 epochs to train the model. For hidden layers ReLu is the activation function and output layer use sigmoid as an activation function. We change number of neurons are between 10 to 100 and select the optimal number of neurons for each layer. The Back Propagation Neural Network manages the relation between the input variables very powerfully also that can hold and learn from long sequence of observations.

C. Data Set

In this text, we plan to use a modified pricing of the twenty CSE sectors. Price details of S&P SL20 companies between 2008-01-01 and 2017-03-31 Are downloaded.

D. Data preparation

Data set was preprocessed to check for missing values and to identify outliers. In each set of financial time series, there are a number of openings, high, low, close, changed ASPI and volume adjustments. The authors have chosen an ASPI variable as the only function in financial times for ARIMA and BPNN model training (Selvamuthu et al., 2019). The Economic and Financial Time Data Series has been split into two sub-sets: training and testing, with 80% of the data used in training and the 20% used to test models.

E. K-Mean Clustering

The dataset is clustered with a technique named k-mean based upon a central point selection and the Euclidian distance measurement. In K-mean clustering we have to decide what are the number of clusters that we choose for clustering. For that we use Elbow method. In here 20 features are cluster in to given clusters. We use K-Mean clustering and PCA to select the most trending sector among the 20 sectors and most trending company from the selected sector.

F. Model Accuracy Testing

We use mean absolute percentage error (MAPE), mean absolute error (MSE) for accuracy testing.

\[
\epsilon_{\text{MAPE}} = \frac{1}{M} \sum_{j=1}^{M} \frac{|X_A - X_P|}{X_A}
\]

\[
\epsilon_{\text{MAE}} = \frac{1}{M} \sum_{j=1}^{M} |X_A - X_P|
\]

In here actual value and predicted value represent \(X_A\) and \(X_P\) respectively in time t.

<table>
<thead>
<tr>
<th>MAPE Value</th>
<th>Level of Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE&lt;10%</td>
<td>Very accurate</td>
</tr>
<tr>
<td>10%&lt;MAPE&lt;20%</td>
<td>Accurate</td>
</tr>
<tr>
<td>20%&lt;MAPE&lt;50%</td>
<td>Medium</td>
</tr>
<tr>
<td>51%&lt;MAPE</td>
<td>Less accurate</td>
</tr>
</tbody>
</table>

IV. RESULTS AND DISCUSSION

First doing data preprocessing part by dropping unnecessary columns of data, checking for missing values, data cleaning and remove strings in columns and find the outliers. To use a time series data model data set should be stationary. We use differencing techniques to remove trend and seasonality. To check for stationarity, we use Augmented Dickey-Fuller Test, Plotting Rolling Statistics, and ACF and PACF plots.

Results of Dickey-Fuller Test:

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>p-value</th>
<th>Critical Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.765369e+04</td>
<td>8.253286e-30</td>
<td>0.04668</td>
</tr>
</tbody>
</table>

In here, test statistic > critical value, which means data set is non-stationary. So we can use this data into time series forecasting. Differenceing results shows that there is no trend or seasonality in the given data set. And also PACF and ACF plot shows that is no trend, seasonality or missing values.

Results of identified optimal number of principal components for sector wise shows showed that 5 out of 20 sectors give considerable amounts of contribution of some features to cover the entire data set during the PCA (Principal Component Analysis) phase.
In here we can see the variance is gradually increase in PC-0 to PC-4. After that there is no considerable change. Thus, we can conclude entire data are covered by 5 components out of 20.

Table 2. Average contribution to clusters

<table>
<thead>
<tr>
<th>Sector</th>
<th>Average contribute to clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banks and finance</td>
<td>3.16402233</td>
</tr>
<tr>
<td>Diversified</td>
<td>2.51563666</td>
</tr>
<tr>
<td>Hotels and travels</td>
<td>2.50811733</td>
</tr>
<tr>
<td>IT</td>
<td>2.20941733</td>
</tr>
<tr>
<td>Motors</td>
<td>2.66154333</td>
</tr>
</tbody>
</table>

Results shows that “Banks and finance”, “Diversified”, “Hotels and travels”, “IT”, “Motors” sectors are highly contribute to the principal components. “Banks and finance” sectors are mostly contributing to the clusters. Thus, we suggest investors to invest in “Banks and finance” sector.

Table 3. The models Accuracy for coming week

<table>
<thead>
<tr>
<th>ASPI</th>
<th>Forecasting Models</th>
<th>MAPE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARIMA</td>
<td>0.4672206</td>
<td>29.6975</td>
</tr>
<tr>
<td></td>
<td>BPNN</td>
<td>0.1783333</td>
<td>4.708423</td>
</tr>
</tbody>
</table>

Results for coming week for ASPI value shows that less MAE and MAPE value have to BPNN model. ANN is extremely exact (less than 10%) with lowest MAPE error levels, based on error analysis data. In addition, MAPE results for precision tests show that BPNN exceeds ARIMA.

Fig. 4. True and testing plot for ASPI using BPNN

This graph shows true testing price in red color and predicted testing price in blue line using BPNN. Most of the predicted values are nearly close to true testing value than ARIMA.

V. CONCLUSION

By using K-mean and PCA we discover the most contributing sector to the ASPI and among that sector again we select the best company that most averagely contribute to clusters. Among 20 sectors “Bank and Finance Sector” contribute 3.16402233 to the clusters and Beyond the companies in Bank sector “Sampath Bank” contribute 2.397040 for clusters. Thus, we suggest investors to invest in Sampath Bank in Banking sector.

This paper contrasts ARIMA’s and BPNN’s accuracy as symbolic techniques when estimating data from time series. The implementation and application of these two strategies on a number of financing data showed that the BPNN is superior than ARIMA. In particular, the MAE value by 4.708423 compared to ARIMA by the BPNN-based algorithm. We propose the advantages of the use of deep learning algorithms and techniques for predict economic and financial data rather than ARIMA.

REFERENCES


