A LSTM APPROACH FOR MODELING AND FORECASTING FINANCIAL TIME SERIES DATA

LAKAVATH KAVITHA, SARAT CHANDRA NAYAK

Abstract: financial prediction of time series data is a most challenging work in present generation, as it is affected by several socio economical factors, uncertain behaviour of stake holders, global economic scenario, and so on. Long Short Term Memory (LSTM) is an advanced method belongs to Recurrent Neural Network (RNN) category. It has shown established performance across various data mining problem such as forecasting of financial time series data. In this article we have done comparative analysis on traditional and LSTM-based forecasting model to explore the capability of Long Short Term Memory in predicting such time series. Machine Learning and Deep Learning are very popular in forecast modelling. A real Datasets of several companies or markets were chosen to convey the efficiency of Machine Learning model for getting better forecasting accuracy. The performances of models are measured by error measures: Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Mean Squared Error (MSE).

Keywords: Financial time series, Recurrent Neural Network, Long Short Term Memory.

I. INTRODUCTION

Sequence of numerical data is known as time series data. Forecasting of time series data is a prediction of any event by analysing the historical data collected from various sources. Forecasting of time series is complex task because uncertain changes in economic condition in one hand as well as incomplete information on the other hand. Market volatility in recent years has introduced serious concerns for economic and financial time series forecasting. Therefore, assessing the accuracy of forecasts is necessary when employing various forms of forecasting methods, and more specifically forecasting using regression analysis as they have several limitations in applications. The main objective of this article is to investigate which forecasting methods offer best predictions with respect to lower forecast errors and higher accuracy of forecasts. In this regard, there are varieties of stochastic models in time series forecasting. The most well-known method is univariate “Auto-Regressive Moving Average (ARMA)” for a single time series data in which Auto-Regressive (AR) and Moving Average (MA) models are combined. Univariate “Auto-Regressive Integrated Moving Average (ARIMA)” is a special type of ARIMA where differencing is taken into account in the model. Multivariate ARIMA models and Vector Auto-Regression (VAR) models are the other most popular forecasting models, which in turn, generalize the univariate ARIMA models and univariate
autoregressive (AR), model by allowing for more than one evolving variable.

Machine learning techniques and more importantly deep learning algorithms have introduced new approaches to prediction problems where the relationships between variables are modeled in a deep and layered hierarchy. Machine learning-based techniques such as Support Vector Machines (SVM), Decision Tree, Multi Layer Perceptron and Random Forests (RF) and deep learning-based algorithms such as Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) have gained lots of attentions in recent years with their applications in many disciplines including finance. Deep learning methods are capable of identifying structure and pattern of data such as non-linearity and complexity in time series forecasting. Comparing traditional and advanced machine learning model LSTM (Long Short Term Memory) out performs traditional models.

To calculate performance of a model by using various statistical metrics such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) applying on different models to get error value which determine performance of the classifier. Minimal error rate shows highest accuracy of the model

II. LITERATURE STUDY

Several researchers proposed different models and analyzed the predicted values of financial market.

In [1] Aakanksha Sharaff, Meenakshi Choudhary this paper involves the comparative study of various stochastic models such as ARIMA model, Artificial Neural Network and Recurrent Neural Network to predict the closing stock indices of S&P Bombay Stock Exchange Sensex in 2012.

In [2] Sima Siami-Namini, Neda Tavakoli and Akbar Siami Namini did experimentation and comparative analysis on ARIMA and LSTM in Forecasting Time Series data. It was observed that the number of training times, known as Epoch in deep learning, had no effect on the performance of the trained forecast model and it exhibited a truly random behavior.

In [3] Richa Handa, A.K. Shrivas, H.S. Hota(2018) using Financial Time Series Forecasting using Back Propagation Neural Network and Deep Learning Architecture. In this study they have used three time series data i.e.BSE30 stock data, INR/USD Foreign Exchange (FX) data and Crude Oil Data for prediction. Many linear and non-linear models have been developed for these time series prediction.

In [4] Zhang Guohui (2017).Research on Time Series Prediction and Its Application Based on Deep Belief Network [DBN].Harbin Institute of Technology .They conducted experimentation on Western reserve university to build training data. A fault diagnosis method combining DBN and LSTM was proposed in this paper using LSTM network has good adaptability to sequential data.


In [6] Yang Yujun, Yang Yimei, Li Jianping research on financial time series forecasting based on SVM. The Research on financial time series forecasting based on the support vector machine. Although the speed of prediction process is slow, it can improve the prediction accuracy of the financial time series. The experimental results show the prediction accuracy of this approach based on the support vector machine.
In [7] Hongbo Sun, Jing Xu (2018) in these paper improved approach for financial market forecasting based on stationary time is explained. This paper approaches several groups of solution for financial data forecasting. On one hand, On the econometric analysis based methods propose an improved vector auto regression model named A-VAR. This paper uses several methods for forecasting the stationary time series data and result of this study explain different methods in different datasets.

In [8] K.Kanchyamalay, Roselina Sallehuddin, Naomie Salim, Ummi Rabaah Hashim (2008) “Time series based forecasting for crude palm oil price utilizing neural network algorithms” This study aims to present time series based forecasting for Malaysian crude palm oil price. The main aim of this research is to forecast the crude palm oil prices based on two important predictors, namely the price of soy bean oil and currency exchange rates. Overall results demonstrate that an increase in the number of input features would influence the accuracy and performance model’s neural network.

In [9] Bo-Sheng Lin, Wei-Tao Chu, & Chuin-Mu Wang(2018) application of stock analysis using Deep learning this paper uses a neural network with memory capability: Recurrent neural network (RNN). In order to improve its performance, Long Short Term Memory (LSTM) architecture was used. This paper combines neural networks and stock history data to predict stock prices.

In [10] Wei Wang, Hong Zhao, Qiang Li, Zhixiong Liu (2009). A Novel Hybrid Intelligent Model for Financial Time Series Forecasting and Its Application. The study shows that the performance of financial time series prediction can be significantly enhanced by using EMD-SVR in comparison with single SVR.

III. METHODOLOGY

A. Linear Regression

Linear regression is a Machine Learning technique used to build relationship between the forecast variable Y and single predictor variable X i.e relationship between dependent and independent variable from the dataset.

Equation for linear regression is

\[ Y = a_0 + a_1 x + e \]  

Where \( a_0 \) represents intercept such as predicted values of Y when \( X=0 \), \( a_1 \) represents slope of the line such as average predicted change in Y when one unit increase in X.

b. Naïve Bayes Classifier

Naive Bayes classifier is a Machine Learning algorithm based on probabilistic Bayes Theorem .It has various applications such as spam filtering, classification documents and sentiment prediction etc. The name Naïve indicates that changes value in one feature does not influence on other features in the dataset. It is very power model because it makes prediction quickly.

The dataset was randomly shuffled, and then it was divided into two subsets Training dataset and testing dataset .Training dataset was used to train the naive Bayes classifier. Test dataset was used to get the results.

\[ P(A | B) = \frac{P(B | A) P(A)}{P(B)} \]  

P (A/B) represents posterior probability P(B/A) is probability of likelihood of evidence P(A) indicates probability of evidence, P(B) indicates prior probability.
C. Support vector Machine

It is a supervised machine learning algorithm which can be used for both classification and regression challenges. SVM used for classification, regression and time series forecasting tasks. The main objective of SVM is to solve non-linear regression estimation problems that makes SVM is successful in time series forecasting. In this algorithm we plot each data as a point in n dimensional space. We perform classification by finding the hyper-plane that differentiates the two classes. Measuring accuracy of a model by applying Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE) are used for evaluating results for time series forecasting.

\[ Y = b + \sum \alpha (1)y(i) \cdot x(i) \cdot x \]  
where \( y(i) \) is the class value of t training example (i), \( \cdot \) represents the dot product. The vector \( x \) represents a test example and the vectors \( x(i) \) are the support vectors, \( b \) and \( \alpha (1) \) are parameters that determine the hyperplane.

D. Multilayer Perceptron

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to refer to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptron (with threshold activation). Multilayer perceptron it learn a function that’s maps series of past observation to predict future value. This model is more flexible depending on hidden layer. Single input layer nodes used for training to feed, hidden layer for changing its weight and output layer for prediction.

\[ Y = \phi(\sum w(i)x(i) + w(0)) \]  

Where \( Y \) represents output, \( x(i) \) is input of the neuron, \( w(i) \) is the weights of the neuron

E. Auto Regressive Integrated Moving Average

Another common Time series model that is very popular among the Data scientists is ARIMA. It stands for Autoregressive Integrated Moving average. While exponential smoothing models were based on a description of trend and seasonality in the data, ARIMA models aim to describe the correlations in the data with each other. An improvement over ARIMA is Seasonal ARIMA. Various ARIMA models get accuracy for several models. ARIMA combines Autoregressive (AR) process and Moving Average (MA) process and builds a composite model of the time series. Autoregressive (AR) is a regression model which is calculated relationship between previous observation and number of lagged observation. Moving Average (MA) is an approach that used to calculated relationship between dependency observation and residual error.

\[ X(t) = C + \sum a_r x_t + e \]  
Where \( c \) is constant, \( a_r \) is autoregressive coefficient, \( e \) is residual error and \( x_t \) is stationary variable.

F. Decision Tree

Decision Tree has various applications in Machine Learning. It works for both continuous as well categorical data. Decision Trees that are grown really deep to learn highly irregular patterns tend to overfit the training sets. Noise in the data may cause the tree to grow in a completely unexpected manner. Random Forests overcome this problem by training multiple decision trees on different subspaces.
of the feature space at the cost of slightly increased bias. This means that none of the trees in the forest sees the entire training data. The data is recursively split into partitions. At a particular node, the split is done by asking a question on an attribute.

G. Random Forest

It is a machine learning techniques used for predicting time series data. This model uses bootstrapping techniques to find observations. In these random accessing of several samples in training dataset and perform classification task according to size of the dataset. This model works well when features are properly prepared.

H. Long Short Term Memory

It is a popular and powerful model for time series forecasting. Time series had more complexity when compare to regression predictive modelling because of sequence depends on input variable. Long Short Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that can model sequence of data so that each sample can be assumed to be dependent on previous ones with the capability of remembering the values from earlier stages for the purpose of future use. It has very large architectures that can be trained easily. Artificial Neural Network (ANN): A neural network consists of at least three layers namely: 1) an input layer, 2) hidden layers, and 3) an output layer. The number of features of the data set determines the dimensionality or the number of nodes in the input layer. These nodes are connected through links called “synapses” to the nodes created in the hidden layer(s). The synapses links carry some weights for every node in the input layer. The weights basically play the

LSTM network that is trained using Backpropagation Through time series and vanishing gradient problem. It has memory blocks that are connected through layers. A block has components that make it smarter than classical neuron and memory or recent sequences. A block contains gates that manage state of the block and output. LSTM network contains three gates which are forget gate, input gate and output gate. Forget gate decides conditionally what information throw away from block, input gate decides which values are taking or input to update memory state and output gate decides output based on input and memory of the block. These gates consist of weights that are learned during the training procedure. Recurrent Neural Networks are very powerful in handling the dependency among the input variables. LSTM network that can hold and learn from long sequence of observations. The algorithm developed is a multi-step univariate forecast algorithm. To implement the algorithm, Keras library along with Theano were installed on computer. These model is required learn from the series past observation to predict next value in the sequence. LSTM network output layer decides prediction for time series dataset

LSTM is a kind of RNN which is a powerful time series model which predict random number of steps in future. LSTM is a combination of five special components, called gates which can model both long term as well as short term data they are: Cell State, Hidden State. Input Gate, Forget Gate and Output Gate. Cell State (It represents the internal memory of cell in which LSTM have
the ability to remove or add information. Hidden State (This is output state information calculate with respect to current input, previous hidden state and current cell state to predict the future data. Input Gate (Input Gate consists of input and decides the information

IV. EXPERIMENTAL RESULT

Dataset is collected from kaggle website and also get from UCI repository. The datasets contain several attributes they are daily opening cost, highest price, lowest price, closing price and transaction volume. And adjustment close price etc.

For experimental analysis we are using Python software, Jupyter Notebook platform and Windows 10 operating system.

Data preprocessing is converting data from raw form to structured or desired form by using Rescaling and Label Encoder of Sklearn library. Split dataset into training and testing. Training set used in model selection and parameter optimization. Testing set used for compare proposed work with other models and apply machine learning techniques to predict appropriate value such machine learning techniques are Support Vector Machine, Naïve Bayes Classifier, Decision Tree, Random Forest, Multilayer Perceptron, Linear Regression etc. Using various performance metrics such as Mean Absolute Percentage error, Mean Squared Error and Root Mean Squared Error.

Mean Absolute Percentage Error (MAPE) is calculated by using absolute error for each and every observation then apply percentage like difference between actual and predicted value divide by number of data points then multiply with hundred. This approach is useful when size of predicted value is important in evaluating.

\[
\text{MAPE} = \frac{\sum (\text{actual} - \text{predicted} + n)}{\sum n} \times 100\
\]

(6)

Mean Squared Error (MSE) is an estimator for unobserved data points. It can be measured as average squared difference between estimated and real value. In these \(n\) is a number of observation, \(y\) values are one for real and other for predicted observations.

\[
\text{MSE} = \frac{1}{n} \sum (\text{actual} - \text{predicted})^2 \quad (7)
\]

Root Mean Squared Error (RMSE) is a frequently used measure square root of the difference between values predicted by model and observed values. RMSE is a measure of accuracy to compare forecasting error of different models for particular dataset and it is scale dependent.

\[
\text{RMSE} = \sqrt{\sum \frac{1}{N} (\text{actual} - \text{predicted})^2} \quad (8)
\]

Below Flowchart shows various steps in experimental analysis.
Fig. 1: Flowchart for forecasting time series data.

Below table shows Performance metrics for different classifiers.

Table -1: Comparison for various metrics

<table>
<thead>
<tr>
<th>Models</th>
<th>MAPE</th>
<th>MSE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>3.37</td>
<td>29.69</td>
<td>5.44</td>
</tr>
<tr>
<td>Naïve Bayes Classifier</td>
<td>3.56</td>
<td>27.68</td>
<td>5.32</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>3.24</td>
<td>24.56</td>
<td>5.21</td>
</tr>
<tr>
<td>Multi-Layer perceptron</td>
<td>3.91</td>
<td>27.43</td>
<td>6.36</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>3.82</td>
<td>25.78</td>
<td>6.02</td>
</tr>
<tr>
<td>Random forest</td>
<td>3.97</td>
<td>25.84</td>
<td>6.54</td>
</tr>
</tbody>
</table>

V. CONCLUSION

the financial time series is complex, the traditional forecasting models are less reliable. The dataset related to the Financial Time Series or Stock Market show that Rooted Mean Squared Error (RMSE) using Long Short Term Memory (LSTM) got accurate result when compare with traditional models. The Neural Network along RNN with memory and uses the LSTM architecture. LSTM can be used to solve long-term time-dependent problems. Effectiveness of forecasting model has never been stopped. Further we can use advanced Deep Learning models to analyse in better way.

REFERENCES

[7] Hongbo Sun, Jing Xu “Improved Approaches for Financial Market Forecasting Based on Stationary Time Series Analysis”.

AUTHORS PROFILE:

**Author1**: Lakavath Kavitha, M Tech student, Department of CSE, CMR College of Engineering and Technology, Hyderabad, India. Email: kavithadofjayaram@gmail.com

**Author2**: Sarat Chandra Nayak, Associate Professor Department of CSE, CMR College of Engineering and Technology, Hyderabad, India. Email: saratnayk234@gmail.com