JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TS Predictor - An Approach to Analyze and Forecast Time Series Data

Phulen Mahato

Assistant Professor
Department of Computer Science
Barrackpore Rastraguru Surendranath College, Barrackpore, Kolkata-120, India.

Abstract : In this work, a methodology has been proposed for the analysis & forecasting of time series data. The methodology involves collecting historical stock data, preprocessing the data such as handling of incomplete data, duplicate data, and incorrectly formatted data, dealing with outliers, data normalization to ensure all inputs are on a similar scale, splitting the dataset into training and testing data sets, organize the dataset into sequential pattern. Train the proposed model using training dataset. Finally we have used the test dataset to evaluate the performance of the model. Our proposed model produces better accuracy in prediction when compared to the similar work done in the same field.

***IndexTerms* - Time series data analysis, data forecasting, training dataset, testing dataset, noise removal, data normalization, sequential pattern, Long short-term memory.**

I. INTRODUCTION

A time series is a sequence of observations taken at equally spaced points in time say days, weeks, months or years. It is helpful in investigating about changes of an entity over time.

Time series analysis is the process of analyzing the time series data collected over a period of time.

It is widely used in many real world applications such as weather forecasting, healthcare forecasting, retail forecasting, business forecasting and financial market prediction and so on.

The Organizations are using time series forecasting to anticipate upcoming events by analyzing data over regular intervals.

In this proposed work at first a time series data of historical stock market has been developed. Dataset may contain incomplete data, duplicate data, and incorrectly formatted data and outliers. These discrepancies might cause hindrance in producing accurate results. Data cleaning is one of the most important stages in creating an efficient model. Hence at the beginning we have incorporated the procedure for data cleaning.

Long Short Term Memory based models are very sensitive to the scale of the data. If model is fit on unscaled data, it is possible for large input to slow down the learning and convergence of the model, even in some cases prevents the model from effectively learning the problem. So we applied MinMax scalar to scale the data within range (0, 1)..

Next split the dataset into training and testing data sets, organize the dataset into sequential pattern. Train the proposed model using the training dataset.

Finally used the trained model on testing dataset to evaluate the performance of the model.

II. PRELIMINARIES

This section deals with some fundamental concepts used for achieving the goal of time series data analysis and forecasting.

First one concept is time series data. A time series is a sequence of observations taken at equally spaced points in time say days, weeks, months or years. It is helpful in investigating about changes of an entity over time.

Time series analysis and forecasting is the process of analyzing the time series data collected over a period of time to anticipate upcoming events.

Long Short Term Memory (LSTM) , a variant of Recurrent Neural Network(RNN) designed to address the vanishing gradient problem is capable to capture complex patterns and dependencies in historical data. Incorporation of memory cells and gates in LSTM, enable it to selectively retain and propagate information over extended time intervals. This uniqueness allows LSTM based models to capture intricate temporal relationships in sequential data, making them particularly well-suited for predicting time series data.

.

III. METHODOLOGY

The methodology begins with collecting the historical stock data; next, we preprocess the data such to tackle duplicate data, incomplete data and outliers.

Next, we have normalized the data to ensure that all inputs are on a similar scale. Here, we have used MinMax scalar to scale data within range (0, 1).

Next, organized the dataset into the sequential pattern.

Next, split the cleaned and normalized dataset into training and test sets, we have used first 75% of data for training purpose and the rest25% of the data for testing purpose of the model.

The proposed model is designed with LSTM layers, and other layers like dense layers for additional processing. We utilize an appropriate loss function and optimizer during the model training.

Train the proposed LSTM based model using training dataset. We have considered step size as 100, batch size 64, and number of epochs 200. During compilation of the model, we have used Adam optimizer to adjust the parameters of the model to improve the speed and accuracy of the model, mean squared error (MSE) for calculating the loss function.

Finally, we have used the test dataset to evaluate the performance of the model. Use the trained model to make prediction on unseen data.

IV. RESULTS

For execution of the methodology mentioned above we have considered a machine with specifications 4 GB RAM and INTEL Core i3 Processor. The method been implemented using Python Jupyter Notebook IDE.

We have used here the NSE TATA GLOBAL dataset, which is a dataset of Tata Beverages from Tata Global Beverages Limited. Dataset contains data from 21-07-2010 to 28-09-2018. It contains a total of 2035 records. We have used first 75% of the data for model training purpose and remaining last 25% of the data for testing purpose. We compiled the model with an Adam optimizer, calculating the loss using mean squared error.

Table 1 shows the snapshot of first 5 rows of the taken dataset.

Table 1: First 5 rows of the dataset

	Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover (Lacs)
0	2018-09-28	234.05	235.95	230.20	233.50	233.75	3069914	7162.35
1	2018-09-27	234.55	236.80	231.10	233.80	233.25	5082859	11859.95
2	2018-09-26	240.00	240.00	232.50	235.00	234.25	2240909	5248.60
3	2018-09-25	233.30	236.75	232.00	236.25	236.10	2349368	5503.90
4	2018-09-24	233.55	239.20	230.75	234.00	233.30	3423509	7999.55

Table 2 shows the summary of the proposed LSTM based model.

Table 2: Summary of the model

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 100, 50)           10400

 lstm_1 (LSTM)               (None, 100, 50)           20200

 lstm_2 (LSTM)               (None, 50)                20200

 dense (Dense)               (None, 1)                 51

=================================================================
Total params: 50,851
Trainable params: 50,851
Non-trainable params: 0
```

Mean squared error (MSE) and Mean absolute error (MAE) values calculated for the model are 0.0002096385433105752 and 0.011509445495903492 respectively.

Root mean square error (RMSE) calculated for training data and predicted training data: 166.61411425830292

Root mean square error (RMSE) calculated for test data and predicted test data: 106.8973897846642

As shown both the calculated values are very close. It indicates the model accuracy is very good.

After execution of the proposed methodology on the fed time series data, the following plots have been generated.

Figure 1shows the generated plot of actual dataset

Figure 1: Plot of actual Dataset

Figure 2 shows the generated plot of predicted training dataset

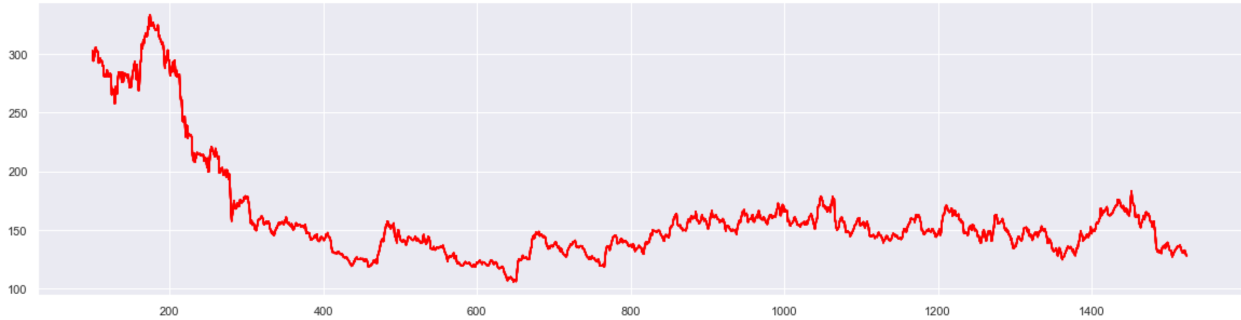

Figure 2: Plot of predicted training dataset

Figure 3 shows the generated plot of predicted test dataset

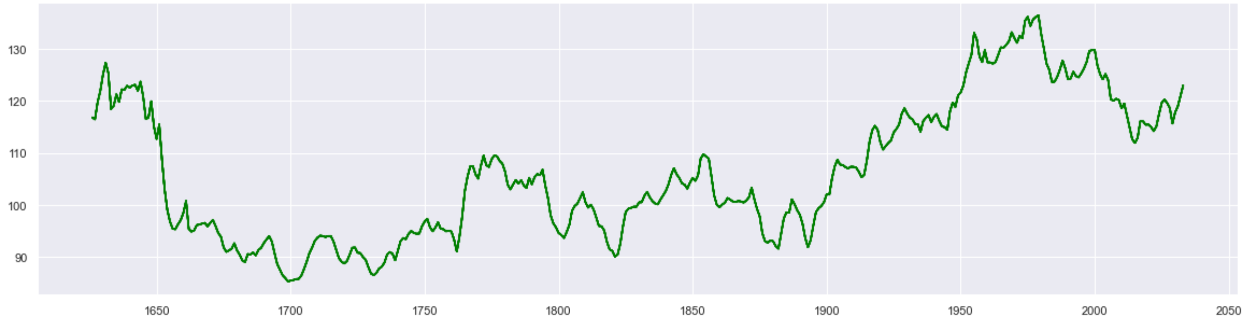

Figure 3: Plot of predicted test dataset

Figure 4 shows the generated combined plot of predicted training and test dataset

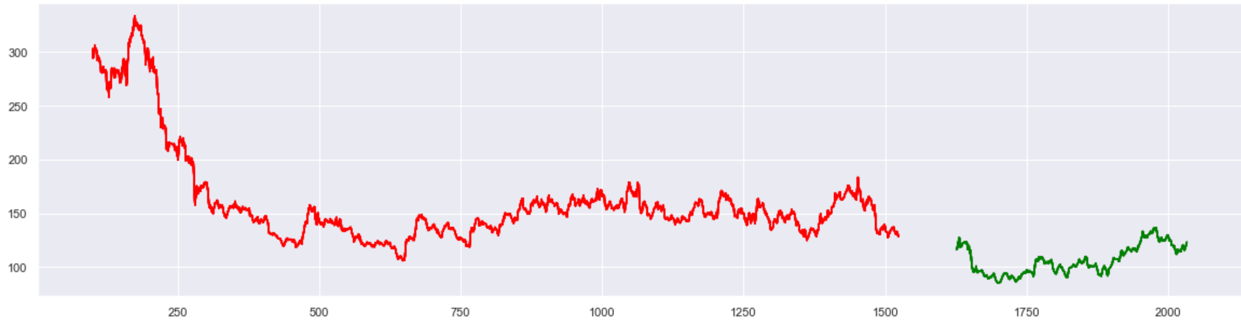

Figure 4: Plot of combined predicted training (red) and test (green) dataset

Figure 5 shows the generated combined plot of actual , predicted training and test dataset

Figure 5: Plot of combined actual (black), predicted training (red) and test (green) dataset.

Figure 5 clearly shows that our proposed model has predicted the stock open price very well.

IV. CONCLUSION

This proposed Long Short-Term Memory based model works very efficiently for analyzing and forecasting the time series data. Deep learning power of LSTM unlocks insights into the unpredictable nature of the time series data. It overcomes the limitations of traditional RNN by incorporating memory cells and gates. The Incorporation of data cleaning makes it an efficient model, data normalization process makes learning and convergence of the model much faster for larger input sequence.

REFERENCES

1. Kumar R, Kumar P, Kumar Y. Multi-step time series analysis and forecasting strategy using arima and evolutionary algorithms. International Journal of Information Technology. 2022;14(1):359–373. doi: 10.1007/s41870-021-00741-8

2. Henrique BM, Sobreiro VA, Kimura H. Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science. 2018;4(3):183–201. doi: 10.1016/j.jfds.2018.04.003

3. Vijh M, Chandola D, Tikkiwal VA, Kumar A. Stock closing price prediction using machine learning techniques. Procedia computer science. 2020;167:599–606. doi: 10.1016/j.procs.2020.03.326

4. Li Y, Bu H, Li J, Wu J. The role of text-extracted investor sentiment in chinese stock price prediction with the enhancement of deep learning. International Journal of Forecasting. 2020;36(4):1541–1562. doi: 10.1016/j.ijforecast.2020.05.001

5. Zhang X, Shi J, Wang D, Fang B. Exploiting investors social network for stock prediction in china's market. Journal of computational science. 2018;28:294–303. doi: 10.1016/j.jocs.2017.10.013

6. Kumar R, Srivastava S, Dass A, Srivastava S. A novel approach to predict stock market price using radial basis function network. International Journal of Information Technology. 2021;13(6):2277–2285. doi: 10.1007/s41870-019-00382-y

7. Shah A, Gor M, Sagar M, et al. A stock market trading framework based on deep learning architectures. Multimedia Tools and Applications. 2022;81:14153–14171. doi: 10.1007/s11042-022-12328-x

8. Lecun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436–444. doi: 10.1038/nature14539

9. Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. Npj Computational Materials. 2018;4(1):1–8. doi: 10.1038/s41524-018-0081-z

10. Dietterich TG. Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems. 2000;p. 1–15. doi: 10.1007/3-540- 45014-9_1 .