Predicting The Most Profitable Stocks of BSE 100 using Bayesian Approach and Neural Network

Report submitted for the

Major Project

STMJ600

By

Soumyadip Bandyopadhyay

Enrolment No: A90579221007

of

Program Name (Batch:2021-23)

Amity University, Kolkata

Name of the Faculty Guide – Mr. Soumya Banerjee

May 2023

Student Declaration

*I, Soumyadip Bandyopadhyay(A90579221007), a student of M.Stat.(2021-23), Department Of Statistics, Amity Institute of Applied Sciences, Amity University Kolkata; hereby declare That the NTCC Major Project, STMJ 600 entitled "**Predicting The Most Profitable Stocks of BSE 100 using Bayesian Approach and Neural Network** " which is submitted by me to the Department of Statistics, Amity Institute of Applied Sciences, Amity University, Kolkata, India, done during the tenure between 31.01.23 to 19.05.23 in partial fulfillment of the requirement for the award of the degree of Master's in Statistics, has not been previously formed the basis for the award of any degree, diploma or other title or recognition.*

Place: Kolkata Signature of the Student

Date:

Certificate by the Faculty Guides

Based on a declaration submitted by *Soumyadip Bandyopadhyay(A90579221007)*, a student of**, the** Department of Statistics, Amity Institute of Applied Sciences, Amity University Kolkata; I hereby certify that the NTCC **Major Project, STMJ 600** entitled **"*Predicting The Most Profitable Stocks of BSE 100 using Bayesian Approach and Neural Network*"** which is submitted to Department of Applied Statistics, Amity Institute of Applied Sciences, Amity University, Kolkata, India, done during the tenure between **31.01.23** to **19.05.23** in partial fulfilment of the requirement for the award of the degree of *Master's in Statistics*, is an original contribution with existing knowledge and faithful record of work carried out by him/them under my guidance and supervision.

To the best of my knowledge, this work has not been submitted in part or full for any Degree Diploma to this University or elsewhere.

Dated:

Signature of the Faculty Guide (Internal)

Name of Faculty Guide with the corresponding affiliation

Acknowledgment

I would like to express my sincere gratitude to Mr. Soumya Banerjee, my advisor, for their invaluable guidance and support throughout this project. Their insights, feedback, and expertise have been instrumental in shaping my work.

I would also like to thank Bikram Chakrabarty for his contributions and assistance in various aspects of this project.

I am grateful to Amity University Kolkata for providing me with the resources and facilities necessary to carry out this research.

Finally, I would like to thank my family and friends for their love, encouragement, and understanding during this process.

Table of content

Abstract

The BSE 100 stock market index is a widely recognized benchmark for the Indian stock market, and predicting the performance of individual stocks within the index can be challenging. Machine learning techniques have been used to predict stock market performance in recent years, and Bayesian approaches have shown promising results. Bayesian filtering techniques can be used to identify the best stocks within the BSE 100 index, by using a probabilistic model to update beliefs about the future performance of stocks based on new information. Bayesian approaches can also analyze the uncertainty associated with stock market predictions and help investors make informed decisions. However, it is important to keep in mind that stock market prediction is inherently uncertain, and past performance is not necessarily indicative of future results.

Introduction

The stock market is a complex and dynamic system that is subject to numerous factors that can influence the performance of individual stocks. Accurately predicting the future performance of stocks is a challenging task, but it is of great importance for investors who are looking to make informed investment decisions. In recent years, machine learning techniques have been used to analyze historical data and identify patterns and trends in the stock market. Bayesian approaches have shown particularpromise in this area, as they allow for the incorporation of new information and the analysis of uncertainty associated with predictions.

The BSE 100 stock market index is a widely recognized benchmark for the Indian stock market, and predicting the performance of individual stocks within this index can be challenging. However, by using Bayesian filtering techniques, it is possible to identify the best stocks within the BSE 100 index based on a range of criteria, such as earnings growth, price-to-earnings ratios, dividend yield, and other financial metrics. This approach allows investors to make informed decisions about which stocks to invest in, based on their probability of future success.

It is important to note that stock market prediction is inherently uncertain, and past performance is not necessarily indicative of future results. Nevertheless, by using Bayesian approaches, investors can incorporate new information and analyze the uncertainty associated with stock market predictions, leading to more informed and potentially profitable investment decisions. This paper will explore the use of Bayesian approaches for predicting the performance of stocks within the BSE 100 index and filtering for the best stocks based on various criteria.

Objective

The objective of this study is to use a Bayesian approach to predict and filter out the best stocks among BSE 100 equity stocks of 2018 Jan – 2022 Dec. The study aims to derive the log return ratio and filter out the top 10 companies based on the positive return ratio average and skewness to get the names of the most profitable stocks. The posterior probability of the selected stocks will be computed by setting the prior probability distribution and using one sigma limit as the filtration technique, then finding the likelihood function. The study will also aim to develop a forecasting model for the top three stocks using GBT and LSTM models and compare their performance using RMSE and MAPE to determine which model is better suited for each company

Methodology

Bayesian Statistics:

A theory of inference called Bayesian statistics deals with how beliefs are changed when new information is discovered. Given the observed data and prior beliefs about the hypothesis, it is a mathematical method for estimating the probability of a hypothesis. There are three main components to Bayesian inference: the earlier likelihood dissemination, the probability capability, and the back likelihood circulation.

The distribution that represents our beliefs prior to observing the data is called the prior distribution. It very well may be any likelihood circulation, however it is normally decided to be instructive or non-useful. With regards to the financial exchange

examination, the earlier dispersion could address our convictions about the verifiable execution of a stock or the presentation of different stocks in a similar area.

The distribution of the data based on the parameters of interest is the likelihood function. Given a set of parameters, it describes the probability of the observed data. The likelihood function could describe the connection between a specific stock and economic indicators like interest rates or inflation rates in the context of stock market analysis.

The back circulation is the dispersion of the boundaries in the wake of considering both the earlier data and the noticed information. It is determined by duplicating the probability capability and the earlier circulation and afterward normalizing the outcome. The back conveyance addresses the refreshed conviction about the boundaries subsequent to noticing the information. The posterior distribution could represent our belief about a stock's future performance after taking into account all of the information that is available in the context of stock market analysis.

Because it provides a methodical approach to analyzing historical data and bringing beliefs about stock performance up to date, Bayesian statistics are utilized in the process of filtering profitable stocks. Investors can incorporate all available information, including prior beliefs and new data, into their analysis by utilizing Bayesian statistics. By calculating the posterior probabilities of various stocks, Bayesian statistics can be utilized to eliminate stocks that are most likely to deliver positive returns. It is more likely that stocks with high posterior probabilities will be profitable and worthwhile investments.

The ability to incorporate prior knowledge into the analysis is one of the main benefits of employing Bayesian statistics in stock market analysis. This prior knowledge, which can help improve the accuracy of the analysis, can be based on previous data or financial models. The relationship between various variables, such as a specific stock and macroeconomic indicators, can be modeled using Bayesian statistics. Investors can predict how changes in one variable will likely affect the performance of various stocks by modeling the relationship between variables.

Additionally, stock performance over time can be examined using Bayesian statistics. The autoregressive integrated moving average model and other time series analysis methods are frequently used to predict stock prices in the future. Bayesian measurements can be utilized to refine these models by integrating earlier convictions about a stock's conduct over the long run.

In conclusion, by incorporating all of the information that is available, Bayesian statistics is a powerful tool that enables investors to make informed decisions regarding stock investments. It can be used to forecast stock performance, filter out profitable stocks, and model the relationship between various variables. Investing can reduce risks and boost returns by utilizing Bayesian statistics.

LSTM model:

RNN (Repetitive Brain Organization) is a kind of neural network that is intended for handling sequential information. Applications like language translation, speech

recognition, and image captioning frequently make use of it. Time series data like stock prices can be processed well with RNNs.

However, RNNs struggle to train effectively over lengthy data sequences due to the "vanishing gradient" issue. To beat this issue, LSTM (Long Transient Memory) networks were created.

A type of RNN called LSTM was created to overcome the drawbacks of conventional RNNs. LSTM networks have a more perplexing design than conventional RNNs, with extra memory components and gating instruments. Applications like speech recognition, image captioning, and stock price prediction frequently make use of them.

There are three main parts to the LSTM network: the input gate, the forget gate, and the output gate.

The LSTM network consists of three main components: the input gate, the forget gate, and the output gate.

1. Adding new information to the cell state is controlled by the input gate. The cell state is updated by adding the new information based on the current input and the previous hidden state.

2. A forget gate removes old information from the state of a cell to an extent controlled by the forget gate. The algorithm takes the current input and the previous hidden state

as inputs and outputs a value between 0 and 1, which indicates how much old information should be retained.

3. The output gate determines how much of the cell state is exposed to the output. By using the current input and the previous hidden state as inputs, it outputs a number between 0 and 1, which indicates how much of the cell state should be used to generate the output.

LSTM networks are particularly useful for stock price prediction because they can learn long-term dependencies from the historical stock price sequence. They can capture the complex patterns and trends in the data and use this information to make accurate predictions. LSTM networks can also be used to analyze the relationship between different variables, such as a particular stock and macroeconomic indicators. By modeling the relationship between variables, investors can forecast how changes in one variable will likely impact the performance of different stocks.

The accuracy of the LSTM network in predicting stock prices depends on many factors, including the quality and quantity of the data, the complexity of the model, and the choice of hyperparameters. However, LSTM networks are effective in predicting stock prices for short-term periods, typically up to a few months.

In summary, RNN and LSTM networks are powerful tools for analyzing sequential data, such as stock prices. LSTM networks are particularly useful for stock price prediction because they can capture long-term dependencies in the data and make accurate predictions. LSTM networks are effective in predicting stock prices for short-

term periods, and they are used extensively in the financial industry for analysis and decision-making.

Material and Software:

From the BSE 100 website data was collected on 100 equity stocks, then used **Python (Jupyter Notebook)** and **Excel** for filtration and forecasting.

Filtration of Profitable Stocks and Forecasting Results

Explanation:

1. at first, I collected data and check whether there is any missing data in any of the CSV files, one company ‘SBI Cards and Payments Service Ltd,’ didn’t have the data for 2018-22 so I have to discard that particular file.

2. Now I ran a Python code for all the CSV files to get their return ratio (in the form of Log (Ct / Ct-1), where Ct is the current period close price and Ct-1 is the lag period close price.)

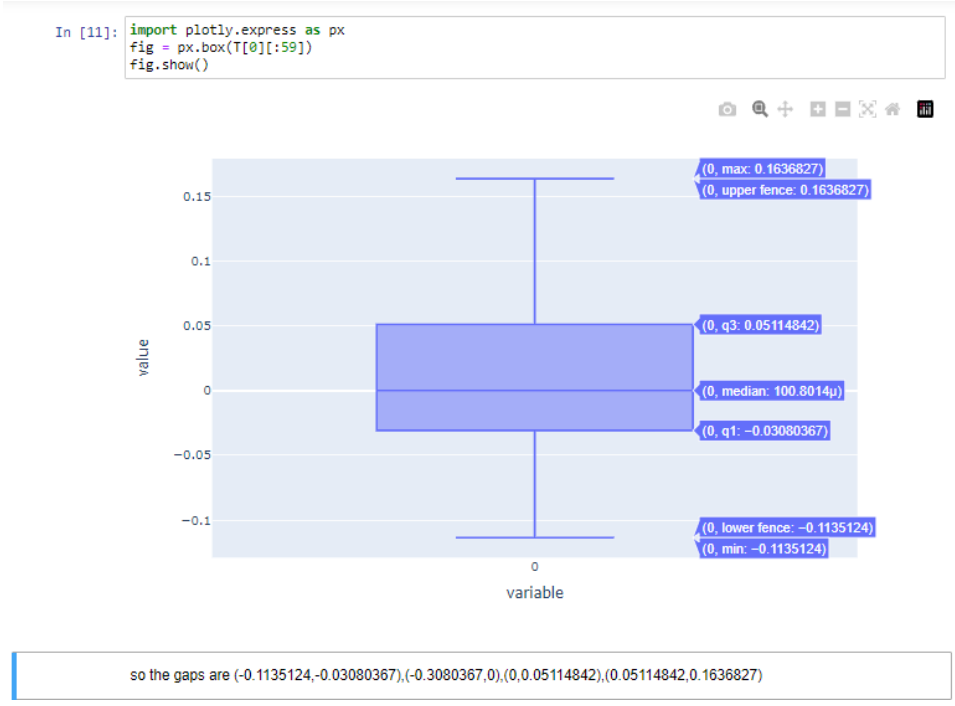
3. Then used Scipy.stats, statistics library to get the average, standard deviation, and skewness concatenated into a data frame which I converted into a .xlsx

File to view and filter in Excel.

4. after getting the data in Excel using the constraint of positive return ratio average and skewness and investors perception map, where in origin there is combined mean and SD, and we chose the stocks on the second quadrant because we want return ratio to be higher than combined average and SD to be lower than combined SD, and we got it down to the top 10 stocks.

And the top 10 filtered stocks are- **Dabur India Ltd., PIDILITE Industries Ltd, Hindustan Unilever Ltd., Asian Paints Ltd, Colgate – Palmolive (India) Ltd, Marico Ltd, ICICI Lombard General Insurance Company Ltd, DR. REDDY's Laboratories Ltd., Siemens Ltd, NTPC Ltd.**

5. now for each company we used the same procedure and it starts with plotting the boxplot of return ratios. Let's see the box plot of "Dabur India Ltd." For example.



so the gaps are (-0.1135124,-0.03080367),(-0.3080367,0),(0,0.05114842),(0.05114842,0.1636827)

6. From the above picture we get the 4 quartile ranges, now the important part is, we didn't have any prior information so We use the relative frequency of return ratio

data points in each quartile range as the prior probability distribution (pi s), and for this model, we took it as the baseline distribution also as it is the current data of stock price movement we have.

7. So, by using the frequency and mid-point we derived E(X) and Var(X), and we did it by hand.
8. Now we are going to use one sigma limit as the filter to discard extremities and outliers, and after computing the remaining relative frequency in the baseline quartile range we are going to use it as the likelihood function to compute posterior distribution.

Bayes' Theorem-

$P(X|D) = P(D|X) * P(X) / P(D)$

Where:

- $P(X)$ = prior probability of X
- $P(D|X)$ = likelihood of D given X
- $P(D)$ = marginal probability of D = sum($P(D|X) * P(X)$)

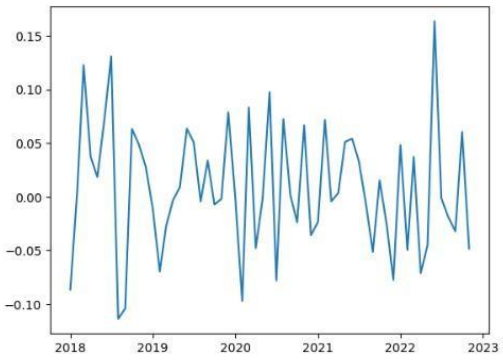
9. So frequency in the first gap is 15, in the second 14, in the third 15, and in the fourth 15. The relative frequencies are respectively, 0.254,0.237,0.254,0.254,these are prior probabilities, and they are taken as probabilities as pi, then we got a probability distribution, then first sigma limit is computed using E(X) – SD(X), E(X) + SD(X), i.e. 4/40, 14/40, 15/40, 7/40 = 0.1, 0.35, 0.375, 0.175. then the Posterior Probabilities are 0.1 x 0.254 /(0.1 x0.254 + 0.35 x 0.237 + 0.375 x 0.254 + 0.175 x 0.254) = 0.102 ,0.334 ,0.383,0.179.

Now we are computing total posterior probability over the last two quartile range because we most efficient profitable stocks, and for all those 10 companies we derived the probability and sorted them in descending order, and the top 3 stocks are,

“Hindustan Unilever ltd.”, “Dabur India Ltd”, “PIDILITE Industries Ltd.”

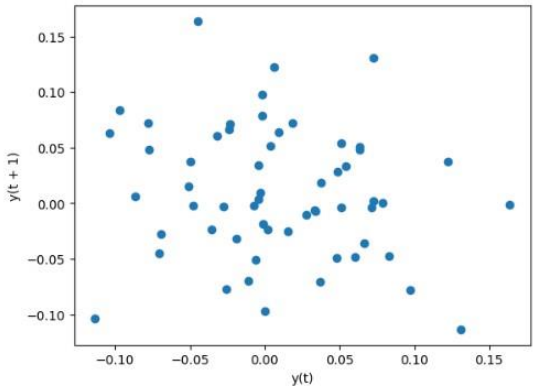
10. Now we are going to collect the data of these 3 stocks day-wise, but we use the monthly data to predict with the GBT model and day-wise data to predict with the LSTM model, but before that, we use the normality test and ACF plot and Augmented Dickey-Fuller test to test the stationarity and linearity of the Data.
11. Dabur India data plot:

```
In [13]: plt.plot(data)
Out[13]: [<matplotlib.lines.Line2D at 0x1b3f107f090>]
```



Lag plot to see the randomness in the data:

```
In [18]: pd.plotting.lag_plot(df['Return Ratio'],lag=1)
         plt.show()
```

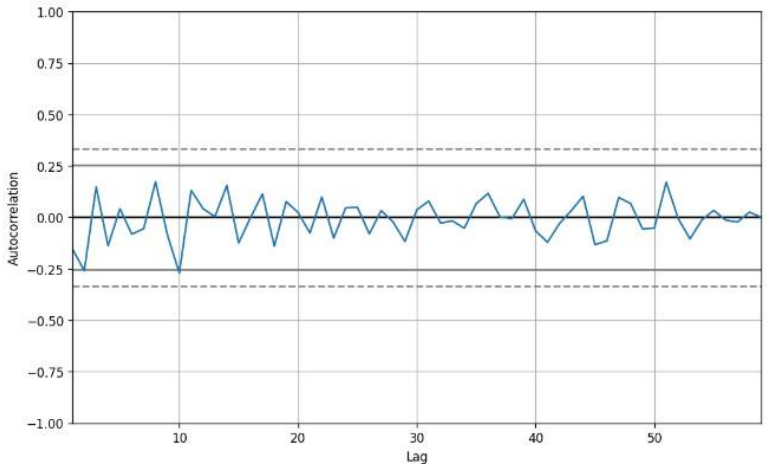


Plot to test the seasonality:

```
In [20]: # Test for seasonality
         from pandas.plotting import autocorrelation_plot

         # Draw Plot
         plt.rcParams.update({'figure.figsize':(10,6), 'figure.dpi':120})
         autocorrelation_plot(df['Return Ratio'][:59].tolist())

Out[20]: <Axes: xlabel='Lag', ylabel='Autocorrelation'>
```



12. To check the normality, we did the Shapiro-Wilk test:

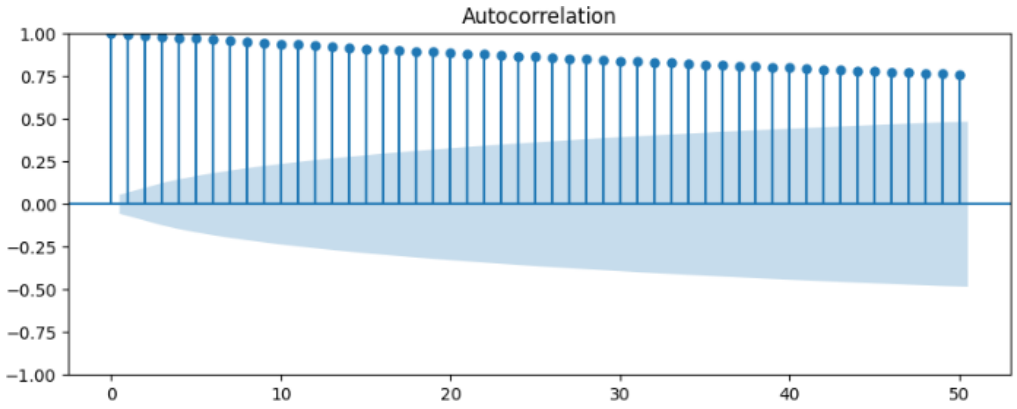
Which gives a test statistic of 0.974 and a p-value of 0.211.

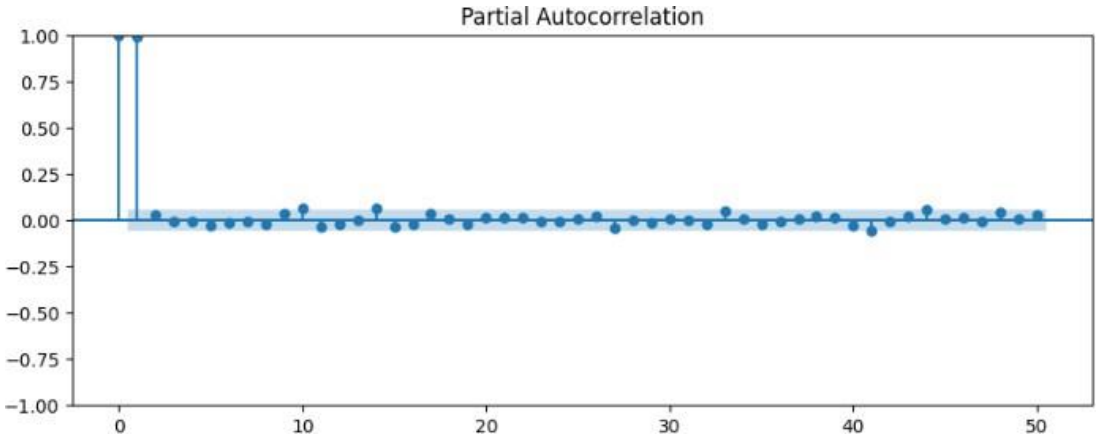
ARIMA assumes that the time series is stationary and follows a normal distribution.

If our data is not Gaussian, it may not be appropriate to use ARIMA for forecasting.

And by looking at the p-value which is >0.05, it's clear that the data is Gaussian.

13. ACF and PACF plot:





Plotting the ACF of the original data can help identify any underlying autocorrelation patterns in the data and also let us determine the value of q in ARIMA here it has significant autocorrelation with lags but it's gradually decreasing and PACF models a linear regression using original data as Dependent variable and lag values as an independent variable, and plots their coefficients, in this case we see only Lag 1 has a significant relation with original data, so it's most likely to follow AR(1) process.

14. ADF test:

Using the ADF test we get a p-value> 0.05, which means it's non-stationary nonlinear data.

15. Using this data, we fitted an LSTM model, first taking the day-wise stock price data for the last 5 years, then using MinMaxScaler to scale the data to (0,1), followed by training and testing.

Next, a 60-time step look-back period is defined. In other words, the LSTM model will predict the next time step based on the previous 60-time steps.

A function called create_dataset is defined to create input-output pairs for the LSTM model. It takes in the time series data and the look-back period as inputs and returns arrays of input sequences (X) and corresponding output values (Y).

Finally, the function is called to create the training and testing datasets, where X_train and X_test contain sequences of the past 60 time steps for each data point, and y_train and y_test contain the next time step value to predict.

LSTM models require input data to be in a 3-dimensional shape with dimensions (samples, time steps, features), where:

samples refer to the number of samples in the dataset

time steps refer to the number of time steps in each input sequence

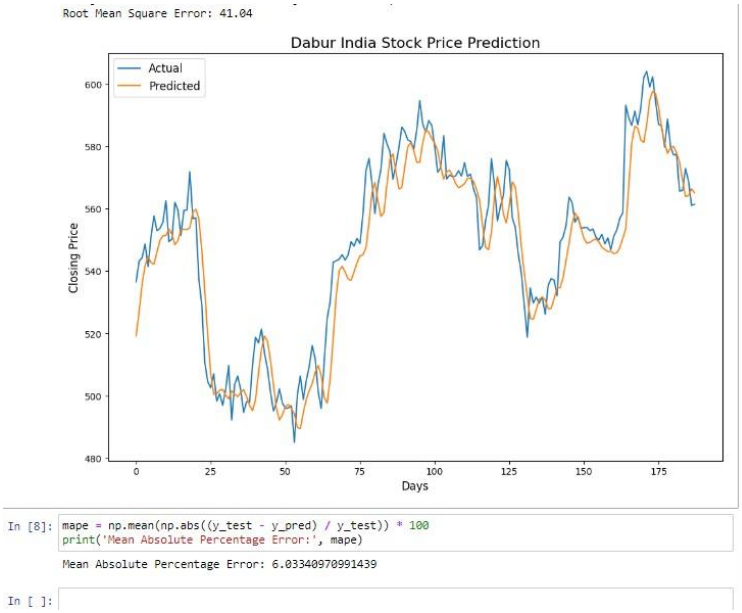
features refer to the number of features (or variables) in each input sequence.

it is an LSTM model for stock price prediction and trains it on training data. It uses the trained model to make predictions on the test data and calculates the Root Mean Square Error (RMSE) between the predicted and actual values. Finally, it plots the actual and predicted values to visualize the model's performance.

The LSTM model has three layers with 50 units each, it gets followed by dropout layers to prevent overfitting. A dense layer with one unit to output the predicted closing price. It is compiled with the Adam optimizer and the mean squared error loss function. The model is trained for 100 epochs with a batch size of 32.

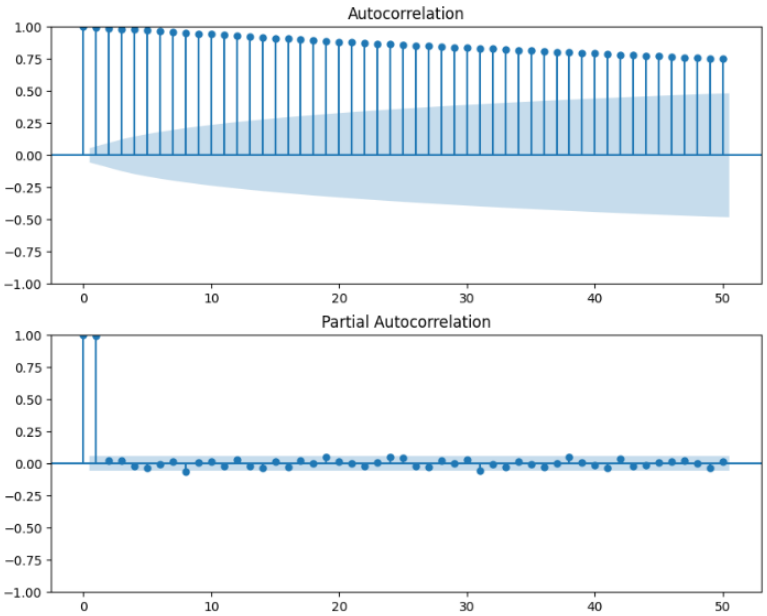
The training data is split into input sequences with a look-back period of 60 days, and the input data is reshaped to be suitable for the LSTM model. The output of the model

is scaled back to the original range of closing prices using the MinMaxScaler used earlier in the code.

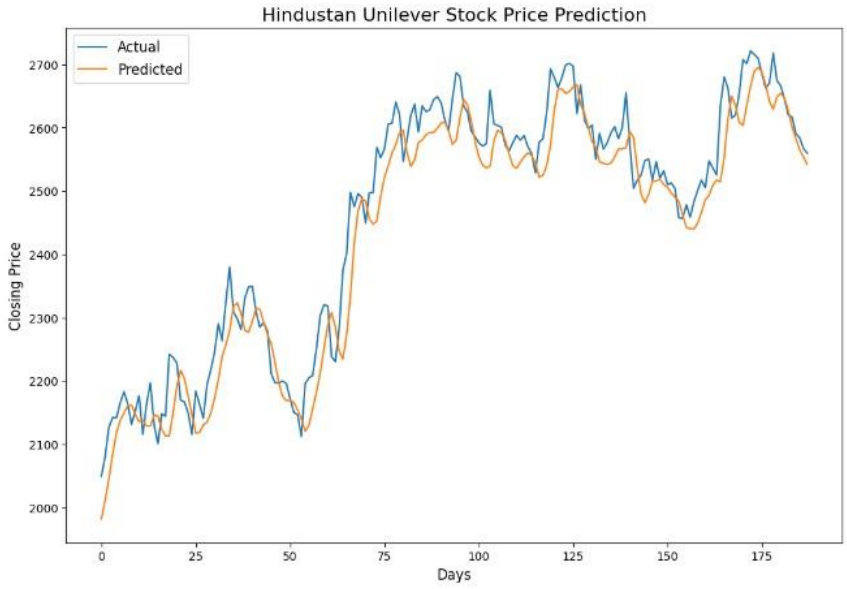


We got a RMSE of 41.04 and MAPE of 6.03 which is pretty low and the plot of the prediction vs actual test data came amazingly.

16. For Hindustan Unilever same tests have been done, so I will just show the results.



```
Root Mean Square Error: 274.27
```




```
mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
print('Mean Absolute Percentage Error:', mape)
```

```
Mean Absolute Percentage Error: 8.84895024267111
```

It performed pretty well with a MAPE of 8.84.

17. For PIDILITE Industries, because the data is not Gaussian.

his code fits different probability distributions to the training data and compares their goodness-of-fit based on the Akaike Information Criterion (AIC). The AIC is a measure of the relative quality of statistical models for a given set of data, where lower values indicate a better fit. And the result is

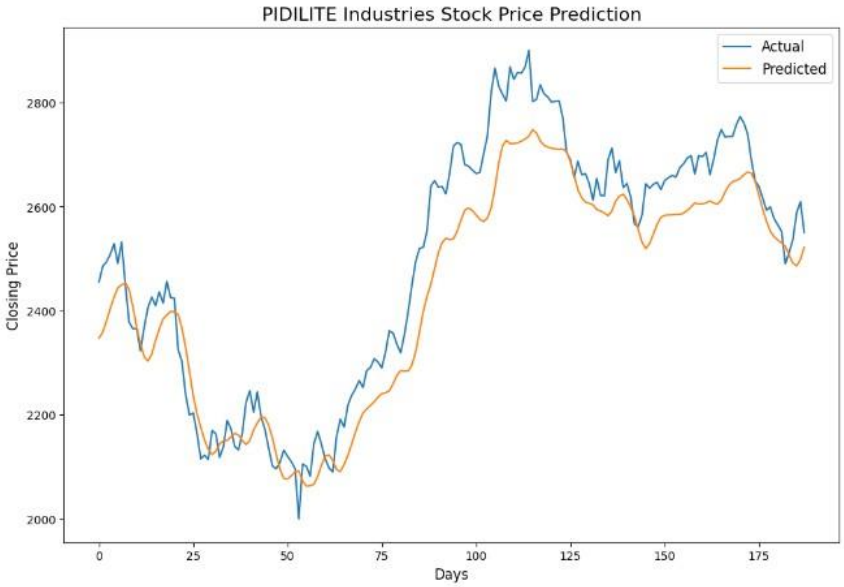
```
Best distribution: vonmises
```

LSTM model which doesn't need normally distributed data.

The result is,

MAPE is 10.66 with RMSE 325.39, seeing the plot and MAPE it's clear the prediction didn't happen better than the previous companies, but still pretty good considering its non-normal data.

```
Root Mean Square Error: 325.39
```




```
mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
print('Mean Absolute Percentage Error:', mape)
```

```
Mean Absolute Percentage Error: 10.660029489718951
```

Conclusion

Using the Bayesian approach and computing the total posterior probability for the third and fourth quartile, we filtered the 3 most Profitable Stocks for the last 5 years, i.e., Dabur India Ltd, Hindustan Unilever Ltd., and PIDILITE Industries Ltd., but in the case of forecasting, we saw LSTM performed very good in predicting the stock price movements,

Future Outlook

For PIDILITE Industries One Can use the Vonmises distribution for modelling for further inspection and using a different length of look-back period for all the LSTM models can give different sorts of results. We used 60 days because A larger lookback

period can capture more complex patterns in the data, but it may also increase the risk of overfitting. A smaller lookback period may be simpler and less prone to overfitting but may not capture all relevant patterns in the data, also

References

- "Bayesian Data Analysis" by Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin
- "Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference" by Cameron Davidson-Pilon
- "Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan" by John K. Kruschke
- "Time Series Analysis: Forecasting and Control" by George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel
- "Applied Time Series Analysis for Fisheries and Environmental Sciences" by J. Hampton and S. J. Smith
- "Introductory Time Series with R" by Paul S. P. Cowpertwait and Andrew V. Metcalfe
- "Deep Learning" by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
- "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems" by Aurélien Géron
- "Recurrent Neural Networks with Python Quick Start Guide: Sequential learning and language modeling with TensorFlow" by N.D Lewis
- ChatGpt free research Version

Predicting The Most Profitable Stocks of BSE 100 using Bayesian Approach and Neural Network

by Soumyadip Bandyopadhyay

Submission date: 12-May-2023 06:25PM (UTC+0530)
Submission ID: 2091311842
File name: NTCC_MAJOR_PROJECT_REPORT_7.docx (486.93K)
Word count: 2901
Character count: 14933

Predicting The Most Profitable Stocks of BSE 100 using Bayesian Approach and Neural Network

ORIGINALITY REPORT

13%	5%	3%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Liverpool John Moores University Student Paper	3%
2	Submitted to Southampton Solent University Student Paper	1%
3	Submitted to UOW Malaysia KDU University College Sdn. Bhd Student Paper	1%
4	Submitted to California Lutheran University Student Paper	1%
5	Submitted to Middle East Technical University Student Paper	1%
6	Submitted to University of Bristol Student Paper	1%
7	arxiv.org Internet Source	1%
8	Submitted to University of Ulster Student Paper	1%

9	Submitted to University of Gloucestershire Student Paper	1%
10	Submitted to University College London Student Paper	1%
11	downloads.hindawi.com Internet Source	1%
12	thesai.org Internet Source	<1%
13	stats.stackexchange.com Internet Source	<1%
14	Submitted to University of Witwatersrand Student Paper	<1%
15	koreascience.or.kr Internet Source	<1%
16	docplayer.net Internet Source	<1%
17	aircconline.com Internet Source	<1%
18	towardsdatascience.com Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		