Stock market analysis with the usage of machine learning and deep learning algorithms

Seethiraju L. V. V. D. Sarma¹, Dorai Venkata Sekhar¹, Gudipatu Murali²
¹Department of Information Technology, Annamalai University, Chidambaram, India
²Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur, India

ABSTRACT

In this work we are focusing on listing out various works in the understanding of various parameters and context to get the overview of stock market analysis in the context of machine learning (ML) and deep learning (DL) models. The work focusses on the stock market analysis along with methodologies and algorithms used to understand the trends and the corresponding results as part of those studies. The importance of this work is to summarize and analyse the parameters which are highly influenced the understanding of the stock market trends. The outcome of the work is understanding the important factors which directly and indirectly influences the stock value raise and drop. The work highlights the methodologies and the algorithms used to stock market data analysis and efficient and effective recommendation of stable stocks to the customers. Further we are listing out the research gaps and future enhancements of the studies which are left over in the earlier works. The work pops up the limitations of some of the works in the existing works along with significance of hyper parameter techniques to clearly identify the features through which we can get more possibilities of better analysis of the data.

This is an open access article under the CC BY-SA license.

Corresponding Author:
Seethiraju L. V. V. D. Sarma
Department of Information Technology, Annamalai University
Chidambaram, Tamil Nadu, India
Email: slvvdsarma@gmail.com

1. INTRODUCTION

In the study of stock data analysis there is always a confusion as the market is highly volatile and not only that as the analysis involves numerous parameters such as market conditions, company background, customer retention rate, political conditions, price hikes in the country. Other than these things many aspects influence the stock market trading like foreign exchange, ownership changes in multinational companies or corporations (MNC) companies, policies issued by governments, securities and exchange board of India (SEBI) and reserve bank of India (RBI) regulatory decisions, changes in the interest rates, foreign investors and domestic investors decisions, pandemics, natural disasters, prices of gold, and bonds the list goes on. Even the stock market analyst who are having many years of experience in trading could not be able to suggest the investors to opt for the shares which gives benefits to the customers. To study and analyse the stock market data the method followed should be qualitative to identify the internal patterns of the data and result the highly influential parameters to observe the highly volatile prices of stock market data.

Most of the customers and market is revolving around online trading which is accessible to everyone, and the members can easily invest form anywhere and with little knowledge in the stock market can get benefitted but the reality is way differently. So, the opportunity here is if we could suggest some better predictions for the customers, they can get more benefits such observations and recommendations. But
as mentioned earlier the stock market movement is not sequential, complicated, and very volatile in nature. Thus, expectation of stock market with strong methods and algorithms to guide the investors that when can they sell their stocks at most elevated prices and purchase the stocks with least expensive prices is very much needed. In general, for making the decisions the professional traders follow traditional methods like the reviews of the company based on the revenue, market position of the company, and growth rate. In case of some technical observation of the stock market data involves relation between stocks and the correspond average the results to better understand value which is helps us to highly influential parameters in the data and reducing the unused dimensions and what can be employed while creating –

2. LITERATURE REVIEW

Nabipour et al. [1] employed an approach of continuous data for the features is used, the metrics opted for quantifying the results are F1-score and in case of machine learning (ML) algorithm usage observed that naïve bayes and decision tree are given least performance which is about 68%, but in case of DL models such as RNN and LSTM the results were encouraging with 86% predictions. In line with the above work our analysis is this, as Naïve Bayes always look for the equality of all the features and all the weak learners were part of the prediction but in the stock market the case is different so the less predictions were recorded due to this property of model the solution for this is in normal conditions, we can opt for this but in the abnormal conditions the naïve bayes is not recommended. The fact here that even the usage of decision tree model also ends up with 68% F1-score which is a combination of precision and recall where there is a scope of accuracy and sensitivity of the model, we can make use of those measures and leverage the results to better understand the source data. In case of DL models, the results are better with 86% which shows that RNN of LSTM could be able to get the betterment due to the time series nature of the source data and the property of retention of the data, but there is an issue with this approach is vanishing gradient issue which we are going to discuss extensively in this research.

The second approach opted is binary data consideration which has given the better predictions when compared with the first approach, the quantification of results was like in Naïve Bayes and decision trees the predictions were 85% and in case of RNN and LSTM it is around 90%. The second approach has recorded the better predictions and faster completion of the model running is observed. The approach followed here is adding of additional layer to identify the trend by scopeing previous and current continuous value which is having the notion of predicting the trends in the newly recorded data. Shah et al. [2] proposed a framework based on the DL architectures. The methodology used in this approach is applicability of convolutional neural network (CNN) and LSTM, the novelty here is usage of hybrid CNN on top of LSTM and dense layers to predict the prices of the stock market index on NSE. The measures observed here are like R-square value is 0.989, mean average error (MAE) is 168.558, mean absolute percentage error (MAPE) is 0.0234 and root mean squared error (RMSE) is 199.076 all these measures were observed in training of the model. In case of the model testing R-square is 0.943, MAE is 242.418, MAE is 0.0310, and RMSE is 413.902.

Our observation from this work is that in general RNN and LSTM were two commonly used approaches but in this work the strategy is opposite like CNN and LSTM combination, the basic idea behind the LSTM usage is to process and make predictions based on the sequence of the data. CNN helps us to exploit the spatial correlation [3] in data and works good with images and speech. The notable property here is CNN can remember much longer sequence and competent enough with LSTM even better than [4], employed a prediction model using DL models. The results were obtained on top of multilayer perceptron (MLP), RNN, LSTM and CNN on the stock market data of NSE–tata motors which belongs to the automobile sector. The MAPE values using auto regressive integrated moving average (ARIMA) were like in case of Maruti 20.66, for HCL it is 24.69, and Axis Bank it is 19.64. Where as the MAPE observed with

Stock market analysis with the usage of machine learning and deep ... (Seethiraju L. V. V. D. Sarma)
respect to DL models were observed like with RNN, Maruti-5.82, HCL-5.40, and Axis Bank-11.64. In case of LSTM applicability Maruti-6.37, HCL 6.97 and Axis Bank 8.13. In case of CNN applicability Maruti-5.36, HCL-6.42, Axis Bank-7.94.

The observation here is in ARIMA usage the results shows that there is no identification of underlying dynamics, where as the usage of CNN done well compared with other 3 models (RNN, LSTM, and ARIMA) in terms of capturing the underlying dynamics in a better way. The MAPE can be referred as absolute percentage deviation which measures accuracy of forecast system this is a poor accuracy indicator and always low value indicates the good results. The work could have been better in case if we use weighted MAPE and symmetric MAPE. Prerana et al. [5] suggested various aspects to consider in the betterment of predictions in the stock market analysis with the usage of DL algorithms. We have observed that while applying the random forest (RF) algorithm to predict the stock market data more parameters can be elected to understand the underlying relations in a better way. As RF can adopt additional randomness to the data and search for the best feature of the data. The other research gap observed in this work is that the work is not scoped the social media data and news articles, tounderstand the markets in a near possible best way.

The generic approach used in the stock market trading is to depend on the existing prices data and to predict the future changes in the prices, financial expectations is like a signal processing problem which is not simple because of less samples, high noisy data, lack of stationarity and non-linear category of the data [6]-[8]. In the consideration of long periods of the data to analyse stock market data for real times transactions may not always give the right results as the conditions over the period changes abnormally. But the helpful observation here is if the stock market analysts could be able to observe the features over the period of the time, they can get valuable information to suggest the elevations and downfall rates of the stocks. To understand the stock market data behaviour various researchers has given their ideas such as applicability of feed forward neural networks [9], SVM [10], and RNN [11] for stock market predictions.

The SVM is suitable for multi-dimensional data processing with kernel trick methodology and done well with separation of the data and if we are not having any idea about the source data. The benefit of electing SVM is the algorithm is having better accuracy in the results. At the same time the main problem with SVM is not suitable for large data sets with noisy data so much time require to pre process the data. The research works [12]-[14] shown the applicability of neural networks to forecast the closing prices of the stock market data for the next day, the short span of the observations also sometimes required to take instant decisions. The observation related to this research work is the selection of DL model is any kind of the functionality can be easily achievable, the visual presentation is explainable to the functional users. The usage of activation functions and cost functions are simple and powerful to exploit the results in a structured way. The methodology in the work focus on comparison of efficiency of predictions in the context of simple artificial neural network (ANN) model and RF for various and proved that the ANN outperforms compared with RF. The implementation aspects focused on applicability of RNN, LSTM, CNN, and multilayer perceptron. In some cases, LSTM outperformed compared with other models, in specific cases CNN outperformed compared with other models.

The election of the algorithms such as RNN and LSTM with finetuned parameter tuning techniques leads to better understand the stock market data. The text-based approach is required as the origin companies provided a set of customers who are continuously following and reading the news published in those papers. An Ai-based stock market prediction using ML algorithm, the authors proposed a method of recommending certain sticks based on the closing price of the stock values [15]. The algorithms utilized in this work are holt-winter algorithm (HWA), RNN and recommendation system to help the investors about the best stocks to consider. In HWA triple exponential smoothing 3 basic things are taken into consideration such as base level, trend level and seasoning factor. In case of RNN usage specifically used LSTM approach, in the recommendation system which a subclass of information filtering system that seeks to predict the rating of the stocks based on certain parameters. The RMSE value as a part of RNN implementation is calculated and used for accuracy in predictions, along with this the approach used in this research is weighted recommendations which will consider various factors to compare the stock prices.

The observations we made from this approach are, usage of RNN on top of LSTM is a common idea and the research is not drawing any significant results from this approach. The usage of RMSE to suggest the stocks to the investors is a good measure, but we are proposing the usage of $R^2$ and Adjusted $R^2$ to penalize the features which are not really contributed to the analysis. The work by Anani and Samarabandu [16] of MLP usage in stock market analysis with greatest average directional accuracy results 65.87% gives the future stock moments with hybrid model by integrating fundamental and technical analysis of stock market. The observation here is the work can only be focused on down trending of the stocks, we could revise the approach to capture the raise in the stock prices. The research works [17]-[21] highlighted usage of new articles for forecasting the stock values keeping in mind the sentiments of the customers. Table 1 shows the comparative analysis of various research works in stock market analysis.
Stock market analysis with the usage of machine learning and deep ... (Seethiraju L. V. V. D. Sarma)
3. OBSERVATIONS

The summary of the above works can be observed in few points as below, there some tool tip aspects we have observed and projecting here. Various characteristics of the stock market data like noisy data nonstationary, no linear nature over the high time duration may not be helping the researchers to understand the patterns of the data and expecting the good predictions and accurate results so need some focus on this aspect for better predictions. Applicability of feed forward networks, SVM and RNNs to analyse and estimate the stock market data in terms of predictions and accuracy were used majorly as they are good fit for multi-dimensional data and recursive processing of the data.

The common point observed in many of the works is in the implementation of models, the common feature used is the closing prices of the stock market data to forecast the immediate day prices instantly which is not suitable for all the company stock, so the integration of other factors in the prediction is mandatory here. The comparative analysis of ANN models with RF in the study of stock market data and there is a slight better performance in ANN compared with RF. The other context focused on RNN, LSTM, CNN and multilayer perceptron. In specific cases LSTM outperformed compared with other models, in some other specific cases CNN outperformed compared with other models. The proposal here is there should be some mechanism like the suitability of these CNN, RNN, and LSTM in such a way that the researchers should be able to adopt these algorithms with some automation or with solid proven assumptions for the acceptable range of predictions. Due to this lot of time and technical resources can be optimized in the process flow.

The other important aspect we have observed from the works done by various researchers is correlation among multiple stock prices may be positively or negatively. We are emphasizing on this point also which is really helping the stock market analyst to suggest some hidden factors while investing on some stocks (at least with prior correlations). The following observations we are making use in our proposed stock market data analysis for better predictions with correlation analysis, ML algorithms, DL algorithms and sentiment analysis based on the current market conditions and news feeds.

a. Provision of huge and various source of stock market data we can expect better predictions.
b. Integration of various statistical measures such as correlation, covariance matrix analysis might give better understanding of the stock market data.
c. The usage of k-fold cross validation in model implementation, along with gridsearch CV will provide the extensive understanding of the feature importance and parameter validation in the stock market data analysis to suggest the better predictions.
d. We can factor the predictions and by integrating the other source of the data like news articles, and feeds, to identify the sentiments of the people which is influencing the stock market data.
e. The second context is study of the correlation among the stocks could really impact a particular stock, which helps to have a conclusion on positive or negative correlations among the stocks.
f. The third aspect is do the stock exchange announcements really impact all the stock values or else few companies are not at all impacted by these conditions, there should be some evidential approach to reach to the conclusion.
g. The fourth context is investor age, employment level and income really having impact on the stocks buying and selling at a particular duration.

By scoping the above four aspects, we are believing that we can come up with better predictions and accuracy values along with understanding the parameters and conditions which are really having high influence on stock market analysis as a driving force. The Figure 1 here gives the understanding of performance estimation with various algorithms like usage of continuous data in the analysis, usage of binary data, RNN and LSTM in the context of binary data usage, and CNN with LSTM usage.

![Figure 1. Various methods and performance](image_url)
4. EXPERIMENTAL SETUP AND RESULTS

To quantify our proposed methodology, we are using spark with python implementation with the following configuration of data bricks cluster of single nodes.10.4 LTS (includes apache spark 3.2.1, scala 2.12). The driver type we have opted for is 15.3 GB memory with 2 cores and 1 DBU. The dataset we have taken consists of stock data with date, open, high, low, close, adjusted close, volume, dividend, and split coefficient around 5273 records were there between 1998 and 2018 duration. We have observed the results on top of this data with the help of MLLib of pyspark library. We have elected decision tree regression and selected the measure RMSE with the result of 0.499624.

The Figure 2 shows the spark job scheduling and other parameters such as the stages information, tasks running over the job and time taken to complete each task upto the level of job, the developer can estimate the performance parameters such as time requirements and number of tasks established by spark to complete each job. The entire spark jobs logically referred as directed acyclic graph (DAG) as mentioned in Figure 3, which gives the understanding of number of transformations and actions elected by spark engine so as to complete the submitted job.

---

4. EXPERIMENTAL SETUP AND RESULTS

To quantify our proposed methodology, we are using spark with python implementation with the following configuration of data bricks cluster of single nodes.10.4 LTS (includes apache spark 3.2.1, scala 2.12). The driver type we have opted for is 15.3 GB memory with 2 cores and 1 DBU. The dataset we have taken consists of stock data with date, open, high, low, close, adjusted close, volume, dividend, and split coefficient around 5273 records were there between 1998 and 2018 duration. We have observed the results on top of this data with the help of MLLib of pyspark library. We have elected decision tree regression and selected the measure RMSE with the result of 0.499624.

The Figure 2 shows the spark job scheduling and other parameters such as the stages information, tasks running over the job and time taken to complete each task upto the level of job, the developer can estimate the performance parameters such as time requirements and number of tasks established by spark to complete each job. The entire spark jobs logically referred as directed acyclic graph (DAG) as mentioned in Figure 3, which gives the understanding of number of transformations and actions elected by spark engine so as to complete the submitted job.
The pyspark usage on top of databricks helps us to perform various pre-processing techniques such as missing value imputations, feature scaling and encoding techniques, usage of machine learning algorithms such as classification and regression along with unsupervised algorithms to reduce the dimensions and to cluster the similar data along with rule generation in the estimation of stock feature correlations. We are exclusively working on all these properties and algorithms available in pyspark libraries with the support of spark architecture, pyspark usage in the implementation with RDD and data frames usage. Most exciting part is the implementation can be done with spark SQL which provides simple yet powerful approach of getting the outcome in most optimized way.

5. CONCLUSION

In this work we have listed out various works done so far in the stock market analysis. The pointers here are study of the stock market data, applicability of ML algorithms such as SVM, KNN, RF and decision trees. The applicability of DL techniques such as CNN, RNN, and ANN like feed forward networks were observed. The typical models in the article processing like LSTM and multipler perceptron-based analysis. The further work we are focusing on 4 pillars like the first one is integration of news articles data in the study of the stock market data really helps us the sentiments associated with the people. The second point we were working out here, the possibility of any correlation kind of the measure can be found in the study of the data, is there any share values which are not at all effected by the market conditions and which are the parameters influence these aspects. The third point we are focusing on the features of community of the customers such as age, level of the employment and income ranges, in buying and selling of the stocks. The fourth point is that the common parameters or a strategic recommendation can we suggest to the investors of course it is too early and too complex to perform and achieve but we are researching on that also. The additional dimension we are focusing is on optimizations of the lines of code, estimation of the time and space utilization while running the jobs on the cluster of databricks are few factors to mention in the quantification of stock market analysis results.

REFERENCES


BIOGRAPHIES OF AUTHORS

Seethiraju L. V. V. D. Sarma 🇮🇳 🇫🇮 🇮🇳 ⚪️ is pursuing his PhD in IT department in Annamalai University. Completed B. Tech in IT from JNTUH affiliated College and M. Tech in IT from JNTUK affiliated College. His area of research includes machine learning and data science. He can be contacted at email: slvvdsarma@gmail.com.

Dr. Dorai Venkatasekhar 🇮🇳 🇫🇮 🇮🇳 ⚪️ has received the B-Tech Degree in Electronics and Communication Engineering from JNTU College of Engineering Kakinada in 1987, MS (Engg) by Research in Information Technology from Annamalai University in 2002, and Ph.D in Computer Science and engineering in 2015 from Annamalai University. He has teaching experience of 30+ years. His field of interest include among others big data, data analytics, data mining, image processing, computer networks, and data communications. He is currently guiding 8 Ph.D scholars. He can be contacted at email: ramavenkatasekar@yahoo.co.in.

Dr. Gudipati Murali 🇮🇳 🇫🇮 🇮🇳 ⚪️ has completed his Degree from Nagarjuna University, in 1996. Master’s degree in Computer Science and Engineering from JRNRU, Rajasthan in 2005 and Ph. D in Computer Science and Engineering in 2013 from ANU, Guntur. Currently, working as a professor in Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences, Guntur, AP. Has teaching experience of 20 years and published several papers in various national and international journals. Currently nine members of scholars are working under his guidance. Four of his scholars has been awarded Ph.D degree from Acharya Nagarjuna University. His field of interest includes computer networks, data mining, big data, data analytics, digital streaming, image processing, VANET’s, and internet of things. He can be contacted at email: m_gudipati@yahoo.com.