Stock Market Prediction Using Machine Learning

Qingyi Chen(✉)

Department of Statistics, University of Warwick, Coventry CV4 7ES, UK
Qingyi.Chen@warwick.ac.uk

Abstract. Stock market analysis and prediction has always been a challenging problem for finance experts because it is so volatile and susceptible of external factors that deeply affect the sentiment of investors. Machine learning, which produces forecasts based on the values of current stock market indices by training on their prior values, is a recent trend in stock market prediction technologies and it shows great promise. However, the prediction methods and algorithms are still developing and the results seem to be volatile and unstable. Making the predictions fast and accurate can greatly impact the financial markets and it is necessary to further develop the models and expand the scope of machine learning approaches. This paper focuses on comparing the use and effectiveness of various models of stock market prediction including Support Vector Machine (SVM), Convolutional Neural Network (CNN), Regression based model and Long Short-Term Memory (LSTM) by summarizing qualitatively the results obtained from several existing sources and experiments. According to the research results of this paper, SVM and the combination of CNN and LSTM performed well in making accurate predictions in the stock market.

Keywords: Stock Market · Machine Learning · Predictive Modeling · Comparative analysis

1 Introduction

Financial market is one of the greatest inventions of all times that plays an imperative role in the whole economy system. Stock market prediction is the process of forecasting and determining the prospective values of a company stock. The stock market is a key pivot in a prosperous and growing economy of all the countries. Making accurate forecasts is of great interest for investors and financial experts because it has strong implications for trading strategies and helps generate significant profits for both the seller and the broker by predicting market behavior and making correct decision either to sell or hold the stocks they possess, yet it is challenging to conduct accurate predictions due to the dynamic and chaotic nature of the stock market [1]. The financial market forecasting has been explored extensively in the past and most stockbrokers tend to use fundamental and technical analysis to make forecasts on stock prices but these traditional methods can not be trusted fully due to the nature of stock market data. A lot of the factors need

D. Qiu et al. (Eds.): ICBBEM 2022, AHIS 5, pp. 458–465, 2023.
https://doi.org/10.2991/978-94-6463-030-5_47

to be take into consideration when making stock price forecasts and things like politics and economic environment can affect the actions and sentiments of the investors, hence leading to stock market movements [9].

In the last few decades, many innovative machine learning approaches have been trained and tested for the forecast of stock prices. They are proven to be much more efficient because the predictions can be made by carefully analyzing the historical data, which are well performed by the machine learning methods [1]. Although many studies and works have focused on building and testing the effectiveness of models in machine learning, the purpose of this article is to summarize some of the machine learning prediction results obtained. It can help people form an overall understanding of various machine learning methods. And provide a constructive reference for the potential improvement of machine learning models in stock market forecasting.

2 Overview of Model Results

2.1 Regression Based Model and LSTM on 9 Lakh Data Records

2.1.1 Overview and Introduction to Dataset

In the Paper "Stock Market Prediction Using Machine Learning" written by Ishita Parmar et al., supervised machine learning methods Regression-based model and LSTM are employed to forecast stock market prices using a dataset of 9 lakh records obtained from Yahoo Finance. The dataset represents the stock prices at different time intervals each day for one company and it includes variables of open, close, low, high and volume. As the variable names suggest, open, close, high and low are the direct stock bid prices at different times and volume means the total number of shares traded during a specific period of time [1]. The dataset was used for simulation purpose only and thus only the data of one company was extracted. The dataset was divided into training and testing sets for the machine learning models to use and predict.

2.1.2 Regression-Based Model and Results

The Regression-based model in general used linear function to predict continuous values based on previous given values [2]. Siew and Nordin used an algorithm called the gradient descent linear regression algorithm and the general process starts from extracting input data from stock movement, then the machine learning algorithm is applied to build the regression-based model which provides the final prediction of the stock prices. The formula is usually $V = a + bK + \text{error}$, where V is the predicted continuous value, K represents the previous given value, a and b are coefficients. The purpose and advantage of this algorithm is to minimize the error given in the formula, all the five variables are considered in regression and R-square test was used to test the confidence level.

After applying the linear regression algorithm, the resulting graph of batch size 512 and 90 epochs were generated and the confidence score was 0.86625. The confidence score means the probability of an accurate prediction is 86.625%. This result shows that the technique is promising and it improved the accuracy of the prediction. The author did well in choosing the dataset that is appropriate in size and the variables contained, as well

as using a machine learning algorithm that is computational efficient that produces the error gradient in a stable way [10]. However, the study result seems to be a bit shallow and not in-depth because the model is not structured and trained in a complicated way, only one confidence score was given for reference and the number of simulations were not specified. Making a solid conclusion needs more evidence and simulations. In general, this is an understandable method that gives a quite high confidence score. It also can be seen that statistical methods are strong supporting pillars for machine learning models till now due to their performance.

To conclude, amongst all the other models, linear regression model is widely utilized in various aspects due to its simplicity and robustness [3], however, the estimation accuracy varies according to the variables considered, association among variables and the random error component.

2.1.3 LSTM Model and Results

The Long Short Term Memory is an advanced recurrent neural network (RNN) that is capable of learning long-term dependencies rather than only current and recent information [4]. This is a wide-spread and effective method used in stock market prediction because an accurate prediction in the stock market heavily relies on the long term history data of the market and this is exactly what long-term short memory is designed for [5]. LSTM controls its error by holding information of older stages that will make the results more accurate. In general, LSTM deals with the problem of vanishing gradient caused by the processing of large amount of data. Moreover, it contains a remembering cell that deals with long-term propagation very well [1].

Parmer applied a model that stacked two LSTM layers with an output value of 256, they took care of the overfitting and efficiency problems by making 0.3 of the total nodes frozen during the training. The compiling process involves a mean square cost function to hold the error and accuracy is at stake during this process.

As it can be seen from Fig. 1, this prediction is compared to the actual value and it measures the real trend of the stock market data when time passes. The resulting train score is 0.00106 MSE and test score is 0.00875 MSE.

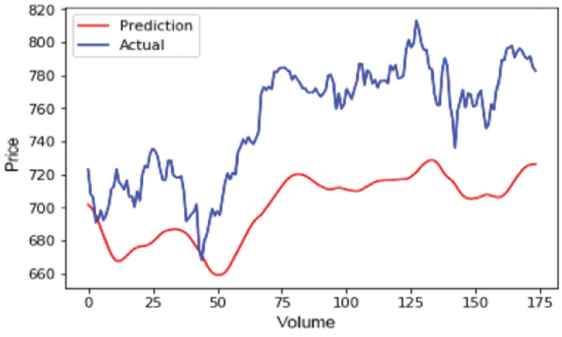

Fig. 1. Plot between Actual & Predicted Trend of LSTM [1]

By comparing the confidence score of the regression model and the resulting MSE score of the LSTM model, the accuracy of LSTM is proven to be better at stock market prediction. It is worth mentioning that larger datasets and more training trails do influence the accuracy of the prediction results. The use of the stacking in this model makes the model deeper and therefore more accurate in making complicated predictions such as that on the stock market. Another thing that Parmer did well is that the effort in setting dropout value as it definitely increased the speed and combats the overfitting issue. However, the model was not trained thoroughly and systematically, and the dataset was not large and comprehensive enough in this case.

2.2 LSTM, Convolutional Neural Network (CNN) and Support Vector Regression (SVR) on Time Dependent Data

2.2.1 Overview and Introduction to Dataset

In the paper "Using machine learning algorithms on prediction of stock price", Li-Pang Chen investigated the performances of different machine learning methods that include LSTM, CNN and SVR in predicting the future stock price on four stocks chosen from Yahoo Finance.

The datasets are extracted from the historical stock data chosen from Apple, Mastercard, Ford and ExxonMobil, these are time-depend data that has variables of open, close, daily high, daily low, adjusted close and volumes as explained in the part 3.1.1. Chen uses datasets that are much larger from the last one, it includes data from January 1st 2002 till March 11th 2020 to ensure the variations will allow the models to perform at their real value with no exceptions. Before building the model, Chen also carefully investigated the properties of the data that includes distribution, variability clustering, linear correlation and long-range dependence, which also helps greatly with the modeling part later on. It is always necessary to investigate and visualize the datasets before doing any practical modeling work because it will ensure the data is trainable and does not have any features that might influence the performance of the models and the accuracy of the results. Most importantly, the Mean Absolute Percentage Error (MAPE) is used to evaluate the performance and accuracy of each model.

2.2.2 Support Vector Machines (SVM)

In the research community, support vector machines have been of high interest for decades, It was Zhang and Shen who introduced its use of the prediction of stock market for the first time in 2009 [6].

Support vector machine is an effective binary classifier that creates a decision boundary where the majority of the points in one category falls in one side and most points of another category falls on the other side. The points are called support vectors and the boundary hyperplane is known as the support vector classifier that places every support vector into the classes [7]. Different from other methods, the support vector machine is applied to the four stocks with a lookback window of 30 days. All the variables listed in the previous session are considered here.

2.2.3 Long Short Term Memory

As mentioned before, LSTM is a variation of RNN that avoids long term dependence problem in predicting financial market data. The memory cells in the hidden layer is the key feature in the model that allows the data to keep updated in the gate structure [8]. During the training process, the LSTM model is trained on the four datasets obtained from Yahoo Finance, namely AAPL, MAST, FORD and EXON. Specifically, a model trained on the whole AAPL can make accurate forecasts of other stock prices as well.

2.2.4 Convolutional Neural Networks (CNN)

Convolutional Neural Networks is a type of neutral networks that is poplar in machine learning and widely used in image processing. Chen takes the one-dimensional input for making predictions solely based on historical data and it turns out to be capable of predicting stock market prices due to its multi-layer nature that facilitates pattern recognition. The adjusted close price is the variable considered here. Similarly, the model is trained separately on the four datasets and trained for all datasets at once. A model trained on the whole dataset of Apple using CNN is used to predict stock market prices as well. For the purpose of comparison, the performance is rated again by MAPE.

2.2.5 Results

It can be seen from Table 1, Chen summarizes the performances of LSTM, CNN and SVR in predicting future market prices on the four stock datasets using MAPE. SVR is

Table 1. Summary of the performance of different methods [9]

Model	Trained on Data from	Mean Absolute Percentage Error MAPE of each Machine Learning's prediction On			
		AAPL	MAST	FORD	EXON
LSTM	AAPL	13.75	6.67	8.68	9.71
LSTM	MAST	-	19.64	-	-
LSTM	FORD	-	-	2.66	-
LSTM	EXON	-	-	-	1.34
CNN	AAOL	2.18	4.78	6.77	7.93
CNN	MAST	-	4.17	-	-
CNN	FORD	-	-	0.55	-
CNN	EXON	-	-	-	0.77
SVR	AAOL	0.67	0.11	0.51	0.21
SVR	MAST	-	0.86	-	-
SVR	FORD	-	-	0.097	-
SVR	EXON	-	-	-	0.085

(Data Sources: http://www.xpublication.com/index.php/jmo/article/view/411)

the most accurate one among the three in this case. And it is interesting to see that the combination of LSTM and CNN will increase the efficiency comparing to LSTM alone. Forecasting other stock prices using the model trained on one dataset only is an effective method that Chen used, it shows the stock market data has striking similarities and it can be useful if people can apply this idea in stocks that have positive relationships, it is both efficient and accurate. However, as the data has a large time span of almost 20 years, many external circumstances may have changed and influenced the current data. Therefore, it is recommended that Chen can look at the data of different time intervals differently and make further analysis based on how the predictions vary in different intervals and digging out the reasons behind it.

3 Discussion

Based on the holistic overview and the detailed analysis of the results of the two papers, The LSTM model is proven to be more accurate than the Regression based model in the paper written by Parmer when predicting the stock market price for one specific company throughout the year. Choosing the correct datasets for building different models and a computational efficient algorithm is what the author did well and these are important in the stock market prediction process as well. Though LSTM performs better than the Regression-based model in this case, it is irrefutable that statistical based methods are the building blocks of many machine learning models and they are widely applied in other fields as well.

The second paper compares and analyzes the performances of SVR, LSTM and CNN in the stock market prediction of four main datasets from Apple, Mastercard, Ford and ExxonMobil. It is found that SVM performs better and makes better predictions when trained on larger datasets and the combination of LSTM and CNN is proven to be of better prediction than the LSTM model alone. The finding that the model trained based on one stock can also be applied to predict stocks of similar types that have a positive relationship with the trained dataset.

A number of things that were not considered or done properly in these two papers include: 1. datasets too large or too small 2. lack of further analysis of the datasets based on the time span 3. limited number of simulations 4. overfitting and efficiency problems 5. influence of external factors like politics, social media posts and financial news.

Hence, it is recommended that people: 1. choose proper datasets of different time spans based on the models they use; 2. try to use larger datasets and more simulations; 3. train the models in a more complete and deeper way; 4. take care of the overfitting and efficiency problem in the training process; 5. assess the role of external factors in stock market prediction and machine learning algorithms.

These recommendations could be a strong boost in making the whole stock market prediction process more complete and accurate.

4 Conclusion

Financial markets provides an essential platform for financial transactions and investments to happen and the it allows people to have the opportunity of making their investment grow. Therefore, stock market prediction is essential in helping the investors make

correct and profitable decisions. The attempt of this paper is to give a systemic review of the modern machine learning methods that are commonly used in the stock market prediction process by analyzing two academic papers that includes the building, testing and analyzing of LSTM, Regression-based model, CNN and SVM. LSTM performed better than the Regression-based model when training on a relative small dataset, SVM performed the best among SVM, CNN and LSTM in predicting future prices based on large stock datasets and the combination of CNN and LSTM also produces better results. It is recommended that people try to choose suitable data sets with different time spans and investigate the characteristics of the data before building a model. At the same time, a computationally efficient method is used when training the model, and it is recommended to apply the model to the current and training experiments. So as to avoid the over-fitting problem in the market forecasting process, so as to improve the process efficiency and forecasting accuracy.

Examining the role of external factors in stock market prediction can be an important aspect in the future study and machine learning definitely plays and important role in it.

Acknowledgment. I would like to show my gratitude to Professor Coggeshall and TA Yuhui for their assistance on the direction and framework of this paper and the comments that greatly improved the manuscript. I am also immensely grateful for sharing their knowledge during the course of this research.

References

1. Parmar et al., "Stock Market Prediction Using Machine Learning," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 574–576, doi: https://doi.org/10.1109/ICSCCC.2018.8703332.
2. H. L. Siew and M. J. Nordin, "Regression techniques for the prediction of stock price trend", 2012 International Conference on Statistics in Science Business and Engineering (ICSSBE), pp. 1–5, 2012.
3. Klein, M.D.; Datta, G.S. Statistical disclosure control via sufficiency under the multiple linear regression model. J. Stat. Theory Pract. 2018, 12, 100–110.
4. S. O. Ojo, P. A. Owolawi, M. Mphahlele and J. A. Adisa, "Stock Market Behaviour Prediction using Stacked LSTM Networks," 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), 2019, pp. 1–5, doi: https://doi.org/10.1109/IMITEC45504.2019.9015840.
5. D. Wei, "Prediction of Stock Price Based on LSTM Neural Network," 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), 2019, pp. 544–547, doi: https://doi.org/10.1109/AIAM48774.2019.00113.
6. Shen, W.; Zhang, Y.; Ma, X. Stock return forecast with LS-SVM and particle swarm optimization. In Proceedings of the International Conference on Business Intelligence and Financial Engineering (BIFE'09), Beijing, China, 24–26 July 2009; IEEE: Piscataway, NJ, USA, 2009.
7. S. Madge, Predicting Stock Price Direction using Support Vector Machines, Independent Work Report Spring, 2015.
8. S. Liu and G. Liao and Y. Ding, "Stock transaction prediction modelling and analysis based on LSTM", 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 2787–2790, 2018.

9. Chen, L. (2020, December 15). Using Machine Learning Algorithms on Prediction of Stock Price | Journal of Modeling and Optimization. Xpublication http://www.xpublication.com/index.php/jmo/article/view/411
10. Donges, N. (2021, August 1). Gradient Descent: An Introduction to 1 of Machine Learning's Most Popular Algorithms. Built In. https://builtin.com/data-science/gradient-descent