Financial Time Series Analysis with Machine Learning

Submitted as Research Report / Honours in SIT723

30-September-2022

T2-2021

Satvik Sharma

STUDENT ID 218595095

COURSE - Bachelor of Software Engineering Honours (S464)

Supervised by: Prof. Maia Angelova

Abstract

This article surveys the publications available on the financial time series analysis with machine learning and aims to answer what machine learning models are most commonly used, how accurate the models can be and what is the future like for the field. Different machine learning techniques have been discussed. A number of papers have been read and a technical analysis on the basis of each paper has been done and according to which a conclusion has been drawn, which is, each of the machine learning model has its own pros and cons, but still is not precise to rely on the forecast. The paper follows the workings of the Statistical machine learning, Deep learning and Hybrid models, a total of 10 different algorithms to find the best among them. The data set for the stock has been chosen from HDFC stock from Nifty-50, and their results for 10-years, 5-years and 1-year has been discussed. The model was trained for 70% of dataset and gave prediction for the rest. The optimizer used for the deep learning models was Adamax. For the short term results, it was seen that Statistical models are the best performing models, but as the dataset increases, the performance of them decreased, especially for ARIMAX model. For the deep learning models, thy are working well for all types of dataset. The bidirectional and the hybrid models did not perform well for the 1-year dataset, but as the dataset increased, the performance increased as well. Overall, most reliable model were CNN and SARIMAX, which showed consistent results.

Contents

List of Figures

List of Tables

1 Introduction

Time Series can be defined as the study of dynamic consequences over a period of time. This can be represented with a simple first order linear equation, for example in the equation
yt=Θyt-1+c
yt represents the value to be taken out at time t in respect to the changes (Θ), external factors(c) and previous value at time t-1 [6]. This article focuses on the financial time series, where the study is more concerned with the financial assets like stocks, shares, currency evaluation, etc. Even though the study of financial time series is a part of time series, but it is highly logical area, where the uncertainty is extremely high, let's say for example the asset returns in the stocks' time series cannot be observed directly. The addition of uncertainty, statistical theory, methods and high volatile market makes financial time series analysis different from regular time series analysis [18].

1.1 Introduction to Financial Time Series

Financial time series has always been of interest of business and financial analysts and when machine learning started to gain popularity, more and more publications have kept adding to the finance literature as determining more effective ways of prediction is important for right investments. This paper will survey the publications available used for financial time series analysis with machine learning and will try to answer the following questions

- What machine learning models are used for financial time series analysis?
- Difference between the most commonly used machine learning algorithms
- Future of machine learning in the field of financial time series analysis

1.2 Objectives

After answering those questions in the literature review, we can start working on the machine learning models and learn how the different models work. For the project, I am choosing the data set from the stock from the Nifty-50 for the time period of ten years, five years and one year between the time period of 2011 to 2020 and working on the previous close value for each of the stock chosen. The previous close value is defined as the price of a stock or market index was when the market closed on the previous trading day. The models that have been compared for the research analysis are from Statistical, Deep learning and Hybrid models, and see how they are working and how accurate its prediction is. The accuracy

will be determined with the evaluation criteria based on root mean squared error, absolute mean error and R2 score and will try to answer the following questions.

- What can be the similarities and differences in the algorithms used for financial time-series analysis?
- How does it tackle the problem of customers who want to invest in that particular stock
- What can be done in the future in order to tackle the problem of higher accuracy for the algorithms?

2 Literature Review

There are hundreds of publications on stock market, trading systems, Forex, etc that use the machine learning technologies like Artificial Neural Network, Evolutionary Computations, Genetic programming, Hybrid techniques or some other [16].
Some of the papers also studied the traditional time series models and compared them with the machine learning techniques, for example in one of the papers written by Xin-Yao Qian, ARIMA was used in comparison with the Logical regression, SVM, and De-noising Auto encoders and it was found that the other machine learning models performed better than ARIMA because the other machine learning models took externals factors into account as well [14]. One of the papers, compares the performance of LSTM and ARIMA, and in conclusion it was found that LSTM based models run much better than the ARIMA based models even if the data set was for one month [17].
There are number of papers that have been reviewed and thus analysed what machine learning algorithms are commonly used and what are the main gaps in them.

Table 1: Methodologies used in the reviewed literature.

Model Name	Total Papers	Paper
ARIMA	5	[14], [17], [12], [3], [2]
ANN	7	[5], [20], [12], [19], [13], [9], [7]
LSTM	5	[17], [12], [19], [9], [8]
LR	2	[14], [8]
SVM	3	[14], [5], [1]
KNN	4	[1], [21], [13], [7]
Hybrid Models	3	[11], [10], [4]
others	9	[14], [1], [20], [3], [11], [10], [21], [9], [7]
End of Table		

2.1 Gaps in current literature

Financial time series is a very popular and complex branch of time series

analysis. The machine learning techniques have taken the branch to new levels. The aim of the paper was to answer what machine learning techniques that are being used for financial time series analysis, finding the best one and what void still needs to be filled. The common algorithms that are used for the financial time series analysis are ARIMA, ANN, LR, LSTM, SVM, KNN, MLP, decision tree, random forest and the hybridisation of these. Even though the study is not just limited to these machine learning algorithms, but these are the most commonly used ones due to their higher accuracy than the other algorithms or their ease of implementation. Moreover, the accuracy of each of the model also varies, depending on the data set as well, for example if the data set is huge, LSTM will work better. All the models have their own pros and cons, but after review of number of papers, hybrid models outperform the basic machine learning models. The gap that needs to be filled for the betterment of the topic is to study and develop more machine learning algorithms, whether it be by creating new ones, or creating new hybrids of existing models for higher accuracy, thus a great opportunity for the researchers in this field. The next part of the dissertation will consist of the development of the artefact, it's evaluation and future work regarding the topic.

3 Research design

The main objective of the project is to model financial time series with machine learning algorithms available. The first part of the project is done with the help of existing machine learning algorithms available and working on them. each of the model tested will be evaluated with RMSE, R2, and mean absolute error. Since, as mentioned in the literature review that each of the model has its own pros and cons, it is a better idea to compare the models in the first phase of the research. For the research I have chosen ten models, which will be compared to find out the best. the benefit of doing this would come in handy when hybridising them with the other models. The second phase of the research will be to hybridise them and work on the client portfolios.

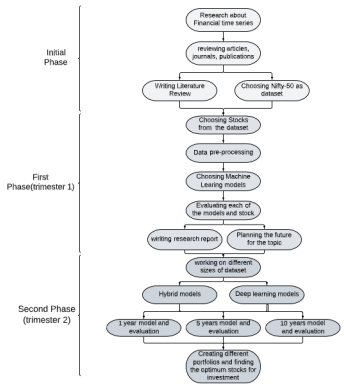

Figure 1: Development Phases for the Project

3.1 Experimentation design

For the analysis of financial time series, the project was divided into number of phases, that has been followed in order to create an optimal research thesis. The steps taken are as follow

- Read and review a number of literature, articles, book, journals etc to find out the current approach and gap in the analysis of financial time series.
- Finding and researching about different machine learning algorithms that can be used to fix the gap in the current learning.
- Create a number of machine learning models and evaluate them and find out if they cover the gaps or not.

For the project, ten years of the Nifty-50 data from 2011-2020 has been chosen [15]. The whole project works on the previous close value which will be extremely helpful in finding out if it the valuation is performing according to the curve or not. The statistical models have been chosen for this part of the project and are down as follow

- Auto Regressive integrated moving average with Exogenous variables (ARIMAX)
- Seasonal Auto Regressive integrated moving average with Exogenous variables (SARIMAX)
- Artificial Neural Network (ANN)
- Convolutional Neural Network (CNN)
- Long Short Term Memory (LSTM)
- Gated Recurrent Unit (GRU)
- Bidirectional-LSTM
- Bidirectional-GRU
- Hybridized LSTM+CNN
- Hybridized LSTM+GRU

All of these models have been tested and compared against each other and evaluated. this is the basis on what the whole research design has been created.

4 Evaluation of Results

The stock that will be used is HDFC, which is a financial corporation stock from the Nifty 50, National Stock exchange, India. There are ten models in total, as defined above, each of them have been compared against each other. The training and testing between the data has been split in the ratio of 70 and 30, meaning 70 percent of the data is trained to predict the 30 percent of the data. There are

three time frames which are short term, 1-year, mid term 5-years and long term 10-years. To keep things simple, the graphs are combined based on the type they belong to, for example there are four deep learning algorithms and these four have been combined in one. Let's look at the next part of the thesis where the graphs of the algorithms are compared against each other.

- 1-year (Short-term predictions)

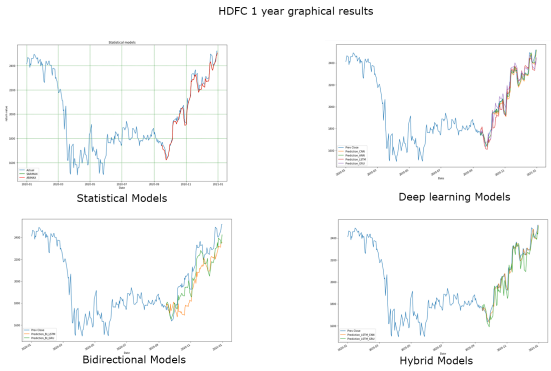

Figure 2: HDFC 1-year graphical results

- 5-year (Mid-term predictions)

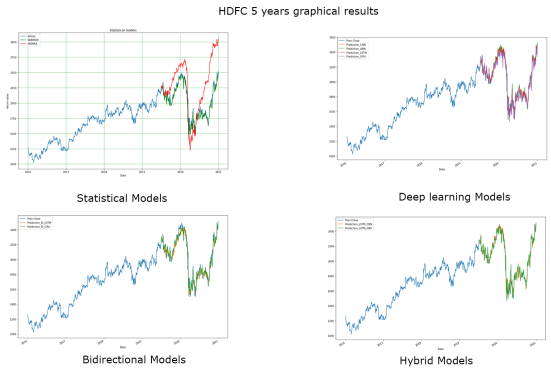

Figure 3: HDFC 5-years graphical results

- 10-year (Long-term predictions)

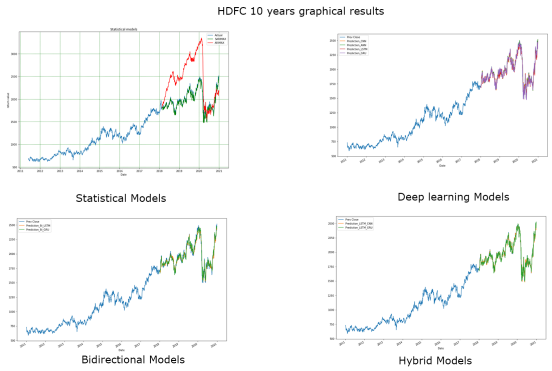

Figure 4: HDFC 10-year graphical results

For the short term predictions, the graphs tells that the statistical models and the deep learning models are very reliable. The Hybrid and the Bidirectional models were showing very high deviations and thus were not reliable to use, with just the exception of LSTM-CNN which was following the trend fairly closely.
By evaluation, it is seen that the worst models are Bi-LSTM as it was having very high RMSE, MAE and R2. The best model of all was ARIMAX. The deep learning models were showing great but interesting results as ANN was performing better than LSTM and GRU.
For mid term predictions, the ARIMAX model stopped being the best and started to show higher deviations. The SARIMAX and deep learning models were performing well according to the graphs. Now that the dataset has increased, the performance of the Bidirectional and Hybrid models have increased as well. It can be seen that there is some deviation, but the overall

trend is fitting the actual trend. When evaluating, it was seen that ARIMAX showed a very high MAE and RMSE as well as an out of bounds value for the R2 score. Interestingly, CNN model was the best model out of all and it was interesting to see this result as it was expected that other models will perform better. Furthermore, the bidirectional models were performing lesser than the regular deep learning models.
For long term predictions, it was interesting to see that for the ARIMAX model, initial predicted values did not fit the graph, but for later on it predicted the values more accurate and closer to the actual values. The SARIMAX and the deep learning models fit the model well and follow the actual values. Since it was for long term predictions, the predictions for long term Hybrid and Bidirectional models was showing great results. After seeing the results for mid term prediction, it was obvious that due to increased dataset, the results for the ARIMAX will not be good. The deep learning models were performing equally well, but it was seen that ANN was performing much better than other deep learning models as well as Hybrid models. The bidirectional and the hybrid models were expected to perform better, but these models are not performing as well as deep learning models. The lowest value for MAE was for SARIMAX, as well as the lowest value for RMSE. For the long term predictions, it was obvious that the best results was for SARIMAX, which was interesting as the ARIMAX model had to be discarded due to very extreme values.

Table 2: HDFC stock evaluation

	Short term			Mid term			Long term		
	MAE	RMSE	R2 score	MAE	RMSE	R2 score	MAE	RMSE	R2 score
ARIMAX	26.420	34.473	98.158%	294.900	369.842	value out of bounds	484.01	568.858	value out of bounds
SARIMAX	25.441	34.497	98.157%	32.097	47.516	96.913%	25.972	38.676	96.900%
ANN	31.861	42.957	97.108%	37.443	53.540	96.074%	28.324	41.004	96.484%
CNN	35.110	47.171	96.513%	36.024	52.417	96.237%	29.758	42.552	96.214%
LSTM	48.970	61.953	93.986%	42.333	57.347	95.496%	95.132	48.253	95.132%
GRU	45.003	56.373	95.021%	47.831	63.014	94.562%	32.923	45.670	95.639%
Bi-LSTM	158.69	178.180	49.971%	50.874	64.82	94.288%	40.152	52.540	94.228%
Bi-GRU	91.400	105.813	82.458%	47.098	62.636	94.627%	36.645	49.006	94.979%

Continuation of Table 2									
	Short term			Mid term			Long term		
	MAE	RMSE	R2 score	MAE	RMSE	R2 score	MAE	RMSE	R2 score
LSTM-CNN	32.584	44.175	96.942%	35.268	51.668	96.344%	29.692	42.192	96.278%
LSTM-GRU	54.121	44.175	91.632%	40.004	51.668	95.784%	30.356	42.192	96.050%
End of Table									

5 Conclusion

This paper researches about different algorithms that can be used for the financial time series analysis. Different techniques have been used. The first methodology is for statistical models, where it is seen that when the data set starts to increase, the models start to perform worse and worse. It can be concluded that statistical models are best suited for short term predictions, with short term data set and will not be reliable with mid term or long term data. When the deep learning models were analysed, it could be clearly seen that the models are performing well for all three datasets, that is short term, mid term and long term. The deep learning models are extremely reliable and can be concluded as the most reliable among all other models and among deep learning models, CNN is the most reliable. During the analysis of statistical models, it was concluded that for short term, the models work best and for big data sets, they lose their performance. The opposite thing can be said for the bidirectional models. It is clear that the model do not work best when the dataset is small as could be seen in the short term predictions, but their performance increases once the dataset has enough values for the predictions. It can be concluded that the Bidirectional models do not perform well for short term predictions and it performs well for long term predictions. The hybrid models are very interesting, as they work well and can be reliable, but at certain times, it performs better than other models, but other times it can be seen that they do not perform as well as others. Further research can answer the questions that will be required by the analysts in the time for the financial time series analysis with machine learning.

References

[1] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, An empirical comparison of machine learning models for time series forecasting, Econometric Reviews, 29 (2010), pp. 594–621.

[2] A. A. Ariyo, A. O. Adewumi, and C. K. Ayo, Stock price prediction using the arima model, 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, (2014).

[3] G. W. Crawford and M. C. Fratantoni, Assessing the forecasting performance of regime-switching, arima and garch models of house prices, Real Estate Economics, 31 (2003), pp. 223–243.

[4] D. S. de O. Santos Júnior, J. F. de Oliveira, and P. S. de Mattos Neto, An intelligent hybridization of arima with machine learning models for time series forecasting, Knowledge-Based Systems, 175 (2019), pp. 72–86.

[5] D. Ersan, C. Nishioka, and A. Scherp, Comparison of machine learning methods for financial time series forecasting at the examples of over 10 years of daily and hourly data of dax 30 and samp;p 500, Journal of Computational Social Science, 3 (2019), pp. 103–133.

[6] J. D. Hamilton, Time Series Analysis, Princeton University Press, 2020.

[7] A. Khazaee Poul, M. Shourian, and H. Ebrahimi, A comparative study of mlr, knn, ann and anfis models with wavelet transform in monthly stream flow prediction, Water Resources Management, 33 (2019), pp. 2907–2923.

[8] T. Kim, S. Sharda, X. Zhou, and R. M. Pendyala, A stepwise interpretable machine learning framework using linear regression (lr) and long short-term memory (lstm): City-wide demand-side prediction of yellow taxi and for-hire vehicle (fhv) service, Transportation Research Part C: Emerging Technologies, 120 (2020), p. 102786.

[9] S. Kouadri, C. B. Pande, B. Panneerselvam, K. N. Moharir, and A. Elbeltagi, Prediction of irrigation groundwater quality parameters using ann, lstm, and mlr models, Environmental Science and Pollution Research, 29 (2021), pp. 21067–21091.

[10] M. Kumar and M. Thenmozhi, Forecasting stock index returns using arima-svm, arima-ann, and arima-random forest hybrid models, International Journal of Banking, Accounting and Finance, 5 (2014), p. 284.

[11] H. Liu, H.-q. Tian, and Y.-f. Li, Comparison of two new arima-ann and arima-kalman hybrid methods for wind speed prediction, Applied Energy, 98 (2012), pp. 415–424.

[12] Q. Ma, Comparison of arima, ann and lstm for stock price prediction, E3S Web of Conferences, 218 (2020), p. 01026.

[13] S. S. Poorna, P. M. V. D. S. Baba, G. L. Ramya, P. Poreddy, L. S. Aashritha, G. J. Nair, and S. Renjith, Classification of eeg based control using ann and knn — a comparison, 2016 IEEE International Conference

on Computational Intelligence and Computing Research (ICCIC), (2016).

[14] X.-Y. Qian, Financial series prediction: Comparison between precision of time series models and machine learning methods, 2017.

[15] R. Rao, nifty-50 stock market data (2000 - 2021), 2022.

[16] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, Financial time series forecasting with deep learning : A systematic literature review: 2005–2019, Applied Soft Computing, 90 (2020).

[17] S. Siami-Namini and A. S. Namin, Forecasting economics and financial time series: Arima vs. lstm, 2018.

[18] R. S. Tsay, Analysis of Financial Time Series, John Wiley and Sons inc, 2 ed., 2006.

[19] V. H. Wentz, J. N. Maciel, J. J. Gimenez Ledesma, and O. H. Ando Junior, Solar irradiance forecasting to short-term pv power: Accuracy comparison of ann and lstm models, Energies, 15 (2022), p. 2457.

[20] C.-C. Young, W.-C. Liu, and W.-L. Hsieh, Predicting the water level fluctuation in an alpine lake using physically based, artificial neural network, and time series forecasting models, Mathematical Problems in Engineering, 2015 (2015), pp. 1–11.

[21] V. Ülke, A. Sahin, and A. Subasi, A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the usa, Neural Computing and Applications, 30 (2016), pp. 1519–1527.