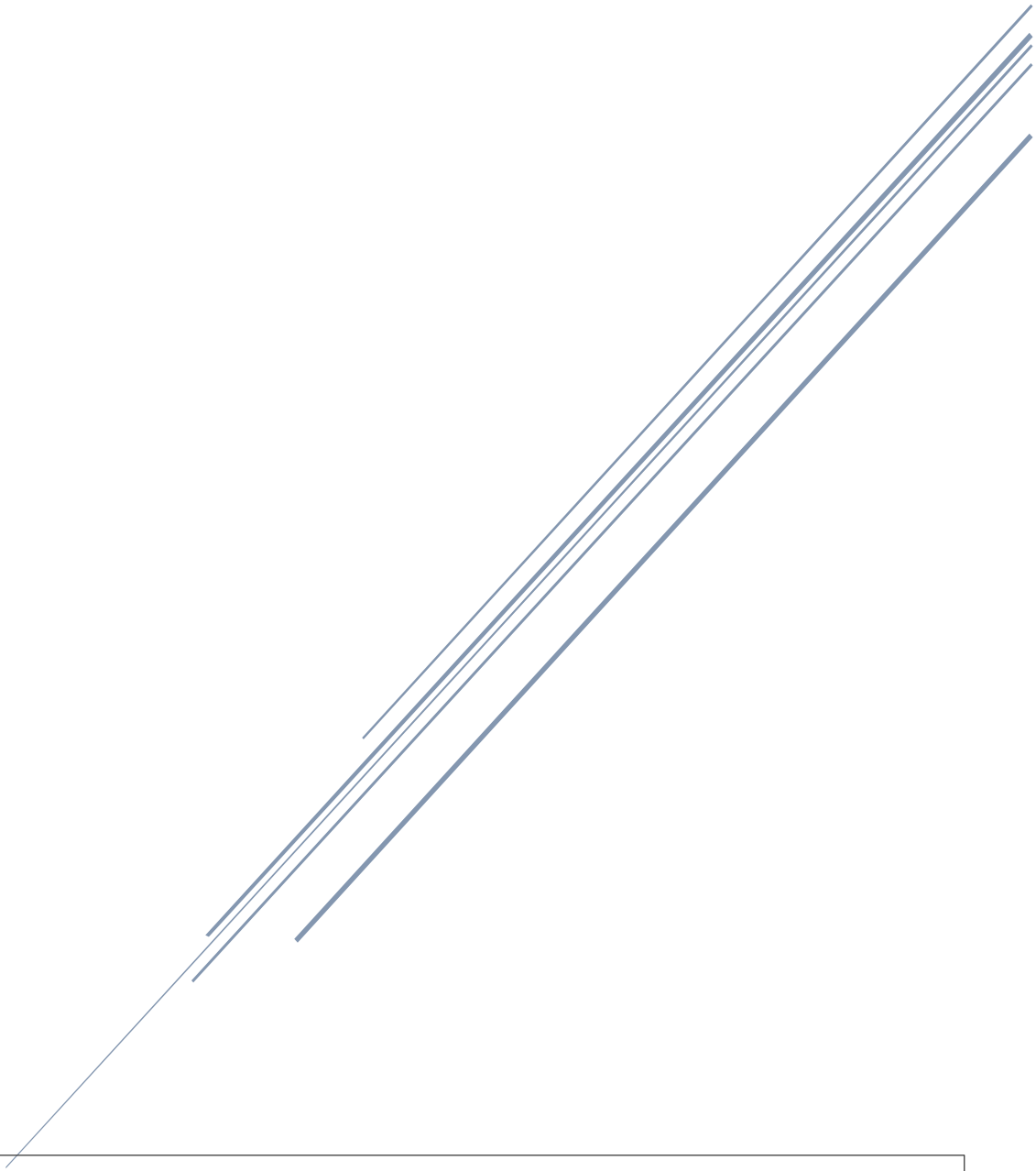
FINANCIAL TIME SERIES WITH MACHINE LEARNING

Literature Review



218595095

Financial time series analysis with Machine Learning: A literature Review

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 more precise models are the hybrid models of the simpler machine learning models.

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

$y_t=\Theta y_{t-1}+c$

y_t 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 [1]. This article focuses on the financial time series, where the study is more concerned with the financial assets like stocks, shares, currency evaluation, et cetera. 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 [2].

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

This literature review will be useful in understanding the analysis of financial time series using machine learning and types of technologies are used for predictions.

Technical analysis based on papers published

There are hundreds of publications on stock market, trading systems, forex, et cetera that use the machine learning technologies like Artificial Neural Network, Evolutionary Computations, Genetic programming, Hybrid techniques or some other [3].

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 Denoising Autoencoders 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[4]. 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 dataset was for one month [5].

There are number of papers which have been mentioned in the table below, that have been reviewed and thus analysed what machine learning algorithms outperforms others. The table will show a small summary of what machine learning models were used and what was the conclusion of the papers

Model Name	**Paper**	**Performance criteria**	**feature dataset**	**Time period**	**environment**	**conclusion**
ARIMA, LR, MLP, SVM, DAE	[4]	-	S&P, Nasdaq, Dow 30	5-years (2012-2016)	Python Statsmodels, Sklearn, Theano	Machine learning models perform better than ARIMA and SVM outperforms all other models
ARIMA, LSTM	[5]	% Reduction in RMSE	Dow Jones Index	1-month	Python with Keras and Theano	LSTM based models work better than ARIMA
SVM, ANN, KNN	[6]	RMSE	DAX 30, S&P500	10-years (2004-2014)	-	KNN outperforms other models
KNN, SVM, Gaussian progress, MLP	[7]	-	M3 time series competition data	Monthly for thousand time series	-	Best two models are MLP and gaussian progression
ANN, ARMAX, 3-D Hydrodynamic model	[8]	MAE, RMSE, R	Water level fluctuation in alpine lake	1-year (2009-2010)	Python Statsmodels	The best working models are ANN and ARIMAX
ARIMA, ANN, LSTM	[9]	P values, graphical observations	Dell's stock price	1 year (2010)	-	ANN model is better than ARIMA model and LSTM, but the hybrid of ARIMA-GARCH model can be used for more accuracy.
ANN, LSTM	[10]	nRMSE, MAPE, R^2	Prediction of Solar irradiance	Hourly	Google Colab, Python scikit-learn, Keras	Due to small dataset, the ANN model works better than the LSTM, but both works much better than the persistence model
ARIMA, GARCH, regime-switching	[11]	RMSE	Home price indices by the OFHEO	20-years (1980-2000)	-	Regime switching models perform better than ARIMA and GARCH.

ARIMA-ANN, ARIMA-Kalman	[12]	MAE, MAPE, MSE	Wind speed prediction	-	-	Both hybrid models have good forecasting accuracy and suitable for wind samplings
ARIMA-SVM, ARIMA-ANN, ARIMA-Random Forest	[13]	-	Indian stock trend	5-years (2004-2009)	MATLAB6.1, SPSS13.0	The hybrid model ANN_ARIMA, was able to predict great values than other models
AR, ARDL, KNN, SVR, Naïve, VAR, MLP	[14]	RMSE, R^2	Inflation forecasting	30-years (1984-2014)	-	SVR and ARDL outperforms other models and machine learning models work best with more volatile and irregular series,
Hybrid ARIMA models	[15]	-	Canadian lynx time series, sunspot time series, airline and star data	Different time periods	-	The hybrid system leads to a higher accuracy in prediction
ANN, KNN	[16]	-	Recorded EEG signals	Patient data set	-	ANN classifier's accuracy and sensitivity was higher than that of KNN classifier.
ANN, LSTM, MLR	[17]	MSE, r, RMSE	Prediction of irrigation groundwater quality	-	Python	ANN and MLR model have highest accuracy in multiple scenarios
LR, LSTM	[18]	RMSE, MAPE	For hire vehicles and yellow taxi	-	-	LR is used to select the important variables and LSTM helps to improve the accuracy.
ARIMA	[19]	-	New York Stock exchange and Nigeria stock exchange	-	python	ARIMA has strong potential for short term prediction
MLR, KNN, ANN, ANFIS	[20]	Mash-Sutcliff coefficient	Stream flow prediction	Monthly	-	The accuracy of each of the model depends on the condition, but the hybridisation was effective.

After observing the table above, it can be clearly seen that the ARIMA is not a good option for the analysis of financial time series even though it is one of the most common methods used. Multiple machine learning methods like LSTM and ANN are also most commonly used as well as have good

prediction accuracy as well. hybrid models have the highest accuracy among the machine learning models. the table below enlists the popular models from the journals and books published.

Model name	Total papers	Paper
ARIMA	5	[4],[5],[9],[11],[19]
ANN	7	[6],[8],[9],[10],[16],[17],[20]
LSTM	5	[5],[9],[10],[17],[18]
LR	2	[4],[18]
SVM	3	[4],[6],[7]
KNN	4	[7],[14],[16],[20]
Hybrid models	3	[12],[13],[15]
others	9	[4],[7],[8],[11],[12],[13],[14],[17],[20]

Machine Learning Techniques

This part of the literature review will give succinct details about some of the most popular machine learning models used for the analysis of financial time series.

- ARIMA
  ARIMA model is basically the integration of AR (Auto regressive) and MA (Moving Average) and is capable of working with the non-stationary data. It is also referred to as box Jenkins models as it was popularized by George Box and Gwilym Jenkins [21]. It can be represented mathematically as shown below
  $y_t = \Sigma^{p}_{i=1}\alpha_i y_{t-l} + \Sigma^{q}_{i=1}\beta_i u_{t-l} + u_t$
  where y_t stands for the goal variable, with which the values of y_{t+1}, y_{t+2}, and so on can be determined [22]. It is one of the most common statistical methods used for financial time series analysis as from the reviewed literature five of the papers were on ARIMA.
- ANN
  ANN or Artificial Neural Network are inspired from the human brain's neurological functions and made in such a manner that it replicates its decisions similar to humans and can be created by programming computer to behave like neurons. Mathematically, it can be represented simply as
  $h_\Theta(x)=1/(1+e^{-\Theta x T})$
  where $h_\Theta(x)$ is the output, x is the input and Θ is the parameter vector [23].
- KNN
  KNN or K-Nearest Neighbour can be called as the one of the simplest algorithms which classifies the data point on the basis of its neighbours. It will be suitable for big datasets for analysis and prediction [24].
- LSTM
  LSTM or long short-term memory belongs to recurrent neural network architecture and consists of memory cells that store information which can be updated from time to time by input, output and forget gate [25].
- LR
  LR or Logistic Regression is statistical analysis method that can be used to predict a binary value on the basis of the data observed. The model works with the binary data, that is 0 meaning the event does not happen, and 1 the event happened [26].
- Hybrid models
  Hybrid models are the combination of two or more algorithms, and this is a major approach towards more accurate and reliable methods, because it benefits from the two methods and

thus reach higher performance. The examples of the hybrid models are DTFNN, ARIMA-ANN, et cetera [27].

These are the most commonly used algorithms that have been reviewed and used for the analysis of the financial time series.

Conclusion

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 dataset as well, for example if the dataset 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.

Bibliography

[1] J. Hamilton, *Time Series Analysis*. Princeton University Press, 41 William St, Princeton, New Jersey 08540, 2020.

[2] T. Ruey S., *Analysis of Financial Time Series*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.

[3] O. Sezer, M. Gudelek and A. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019", *Applied Soft Computing*, vol. 90, p. 106181, 2020. Available: 10.1016/j.asoc.2020.106181.

[4] X. Qian, "Financial Series Prediction: Comparison Between Precision of Time Series Models and Machine Learning Methods", no. 5, 2018. Available: https://doi.org/10.48550/arXiv.1706.00948. [Accessed 1 April 2022].

[5] S. Siami Namin and A. Siami Namin, 2018. Available: https://arxiv.org/ftp/arxiv/papers/1803/1803.06386.pdf. [Accessed 1 April 2022].

[6] 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 S&P 500", *Journal of Computational Social Science*, vol. 3, no. 1, pp. 103-133, 2019. Available: 10.1007/s42001-019-00057-5.

[7] N. Ahmed, A. Atiya, N. Gayar and H. El-Shishiny, "An Empirical Comparison of Machine Learning Models for Time Series Forecasting", *Econometric Reviews*, vol. 29, no. 5-6, pp. 594-621, 2010. Available: 10.1080/07474938.2010.481556.

[8] C. Young, W. Liu and W. 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*, vol. 2015, pp. 1-11, 2015. Available: 10.1155/2015/708204.

[9] Q. Ma, "Comparison of ARIMA, ANN and LSTM for Stock Price Prediction", *E3S Web of Conferences*, vol. 218, p. 01026, 2020. Available: 10.1051/e3sconf/202021801026.

[10] V. Wentz, J. Maciel, J. Gimenez Ledesma and O. Ando Junior, "Solar Irradiance Forecasting to Short-Term PV Power: Accuracy Comparison of ANN and LSTM Models", *Energies*, vol. 15, no. 7, p. 2457, 2022. Available: 10.3390/en15072457.

[11] G. Crawford and M. Fratantoni, "Assessing the Forecasting Performance of Regime-Switching, ARIMA and GARCH Models of House Prices", *Real Estate Economics*, vol. 31, no. 2, pp. 223-243, 2003. Available: 10.1111/1540-6229.00064.

[12] H. Liu, H. Tian and Y. Li, "Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction", *Applied Energy*, vol. 98, pp. 415-424, 2012. Available: 10.1016/j.apenergy.2012.04.001.

[13] 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*, vol. 5, no. 3, p. 284, 2014. Available: 10.1504/ijbaaf.2014.064307.

[14] 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*, vol. 30, no. 5, pp. 1519-1527, 2016. Available: 10.1007/s00521-016-2766-x.

[15] D. de O. Santos Júnior, J. de Oliveira and P. de Mattos Neto, "An intelligent hybridization of ARIMA with machine learning models for time series forecasting", *Knowledge-Based Systems*, vol. 175, pp. 72-86, 2019. Available: 10.1016/j.knosys.2019.03.011.

[16] S. Poorna et al., "Classification of EEG based control using ANN and KNN — A comparison", *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2016. Available: 10.1109/iccic.2016.7919524 [Accessed 6 April 2022].

[17] S. Kouadri, C. Pande, B. Panneerselvam, K. Moharir and A. Elbeltagi, "Prediction of irrigation groundwater quality parameters using ANN, LSTM, and MLR models", *Environmental Science and Pollution Research*, vol. 29, no. 14, pp. 21067-21091, 2021. Available: 10.1007/s11356-021-17084-3.

[18] T. Kim, S. Sharda, X. Zhou and R. 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*, vol. 120, p. 102786, 2020. Available: 10.1016/j.trc.2020.102786.

[19] A. Ariyo, A. Adewumi and C. Ayo, "Stock Price Prediction Using the ARIMA Model", *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, 2014. Available: 10.1109/uksim.2014.67 [Accessed 6 April 2022].

[20] 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*, vol. 33, no. 8, pp. 2907-2923, 2019. Available: 10.1007/s11269-019-02273-0.

[21] "14.5.1 - ARIMA Models | STAT 501", *PennState: Statistics Online Courses*, 2022. [Online]. Available: https://online.stat.psu.edu/stat501/lesson/14/14.5/14.5.1. [Accessed: 08- Apr- 2022].

[22] R. Hyndman and G. Athanasopoulos, *Forecasting*, 2nd ed. Melbourne, Australia: OTexts, 2018.

[23] D. Otchere, T. Arbi Ganat, R. Gholami and S. Ridha, "Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models", *Journal of Petroleum Science and Engineering*, vol. 200, p. 108182, 2021. Available: 10.1016/j.petrol.2020.108182.

[24] P. Soucy and G. Mineau, "A simple KNN algorithm for text categorization", *Proceedings 2001 IEEE International Conference on Data Mining*. Available: 10.1109/icdm.2001.989592 [Accessed 8 April 2022].

[25] J. Cao, Z. Li and J. Li, "Financial time series forecasting model based on CEEMDAN and LSTM", *Physica A: Statistical Mechanics and its Applications*, vol. 519, pp. 127-139, 2019. Available: 10.1016/j.physa.2018.11.061.

[26] J. Cramer, "The Origins of Logistic Regression", *SSRN Electronic Journal*, 2003. Available: 10.2139/ssrn.360300.

[27] S. Ardabili, A. Mosavi and A. Várkonyi-Kóczy, "Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods", *Lecture Notes in Networks and Systems*, pp. 215-227, 2020. Available: 10.1007/978-3-030-36841-8_21 [Accessed 8 April 2022].