

A Comparative Analysis of the Application of Machine Learning Algorithms and Econometric Models in Stock Market Prediction

Zhehan Ni^{1,*†} and Weilun Chen^{1,†}

¹Department of Statistical Science, University of Toronto, Toronto, Canada

*Corresponding author: zhehan.ni@mail.utoronto.ca

[†]These authors contribute equally.

Abstract. Forecasting the future price trend of a stock traded on a financial exchange is the aim of stock market prediction. In recent decades, stock market prediction has been a fascinating topic in the domain of Data Science and Finance. In reality, the stock movement is ambiguous and chaotic due to various influencing factors such as government policy, current events, interest rates Etc. At the same time, accurate enough forecasting of stock price movement leads to substantial benefits for investors. This paper provides a comprehensive review of the application and comparison of Machine Learning (ML) algorithms and Econometric Models in stock market prediction. The mentioned models are categorized into (i) ML algorithms, including Linear Regression (LR), K-nearest neighbors (KNN), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). (ii) Econometric Models, including Autoregressive Integrated Moving Average (ARIMA) Model, Capital Asset Pricing Model (CAPM), and Fama-French (FF) Factor Model.

Keywords: Machine Learning, Econometric Model, Stock Market Prediction, ARIMA, LSTM.

1. Introduction

The stock market has long been one of the most popular platforms attracting all kinds of investors and a fascinating topic in the field of Data Science and Finance. The goal of stock market prediction is to predict the future price trend of a stock traded on a financial exchange. To maximize the investment profits, all buyers want to buy stock shares at the lowest price while selling at the highest price. However, in reality, the stock movement is volatile, ambiguous, and chaotic due to various influencing factors such as government policy, current events, interest rates, Etc., making the pattern hard to predict. Although investing in the stock market is risky and challenging, accurate forecasting of stock price movement simultaneously leads to substantial benefits for investors. Based on research papers, the main-stream methodologies with the highest stability and accuracy for stock market prediction can be classified into (i) Machine Learning (ML) algorithms and (ii) Econometric Models.

ML is defined as a subfield of Artificial Intelligence (AI) that can constantly update the solution without any human intervention and can constantly learn from previous mistakes. There are three primary categories of ML: Supervised, Unsupervised, and Reinforcement Learning (RL). Each type of ML algorithm is applied in a corresponding situation and field to promote innovation and improve operational efficiency. On the other hand, Econometric Models are established from historical data on price and quantities using statistical techniques [1]. The characteristics of econometric models are the strict pre-assumptions of optimizing the behaviour of sellers and buyers as the models depend on economic theories depicting an idealized world [1]. This review paper provides a radical analysis of employment and a comparison between ML algorithms and Econometric Models in stock market prediction.

The paper's architecture is listed as follows. The methodology and comparison of ML algorithms and Economics Models in the categories below are summarized and concluded separately. (i) ML algorithms (section2) include Linear Regression (LR), K-nearest neighbors (KNN), Support Vector Machine (SVM), and Long Short-Term Memory (LSTM). (ii) Econometric Models (section3) include Autoregressive Integrated Moving Average (ARIMA) Model, Capital Asset Pricing Model (CAPM), and Fama-French (FF) Factor Model. Section 4 presents the disadvantages and advantages

of the corresponding ML algorithms and Econometric Models, along with future trends in the stock market forecasting domain.

2. ML Algorithms in Stock Market Prediction

2.1 LR Model

2.1.1 LR Model Explanation

LR is one of the ML algorithms in the field of regression in Supervised Learning. The algorithm models a linear relationship between one dependent variable and one or more explanatory variables.

The general equation of the regression line is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i$. LR is frequently applied in predictions and forecasting for continuous variables. It can also be used to quantify the relationships between various variables. For example, in the stock exchange, the demand and supply of the stock can be used as independent variables to determine the closed price of the stock, which is the dependent variable. The least square method is applied to the data to determine the best-fitting line.

2.1.2 LR Literature Review

Ghani et al. stated the volatility in the global stock market and wished to utilize ML algorithms to help people to take advantage of the stock market with minimum effort [2]. The authors predicted Amazon, Apple, and Google historical stock data using LR, where the close price is the independent variable. The obtained equation was $Y = 1.8511X - 78860$ with R^2 of 0.6938. Similarly, Emioma et al. created a least-square LR model that yields a 1.367% absolute percentage error, and 0.512 roots Mean Squared Error (MSE) [3]. Moreover, Agrawal et al. analyzed the model performance of Deep Learning (DL) algorithms and LR on the 2010 to 2020 stock data from the New York Stock and Tesla [4]. The work concluded that LR with an RMSE of 5.76 had the worst performance among all other algorithms.

Table 1. Summarizes all mentioned papers above under the category of LR model.

Year	Author	Dataset	Time Range	Algorithm	Result
2019	Ghani et al.	GOOGLE, FB, AMAZON, AAPLE data	2019	LR	Prediction is improved by exponential smoothing over LR.
2021	Emioma,etc	Bank of America stock from yahoo finance	June 2012 to June 2019	LR	The LR model has an MSE of 1.3647% and RMSE of 0.512.
2020	Agrawal, et al	Tesla Stock Price and New York Stock	From 2010 to 2020	LR	Non-linear DL methods have better performance on LR

2.2 KNN & SVM Model

2.2.1 KNN & SVM Model Explanation

KNN being a non-parametric, Supervised Learning algorithm, is used to solve classification problems such as identifying the class or category of the data. The K in KNN stands for the number of nearest neighbors needed in the calculation. The algorithm first calculates the k neighbors with a

minimum distance from the new data point. Then the classification of the query example will be done by checking the features from the k neighbors.

2.2.2 KNN & SVM Literature Review

Bhardwaj et al. constructed and compared the performance of the following ML algorithm, random forest, KNN, and logistic regression on the stock price movement of five different datasets extracted from Yahoo finance [5]. The authors created a binary indicator variable, "status," to illustrate whether the open price is greater than the close price. As a result, the KNN algorithm performed best on the AAPL dataset with 74% of accuracy and, on average, 64% accuracy on all five datasets. The literature also highlighted that the KNN was only better than ARIMA and had issues with dataset standardization and inadequate handling of categorical variables. However, a recent study from Qian proposed an improved KNN algorithm because the traditional KNN approach only forecasted changes in trend for the following day using data from the most recent day [6]. Qian suggested that the next day's stock price prediction should depend on the stock price from the previous N Day to achieve better accuracy. Hence Qian grouped and embedded the first N days' data into the KNN model as input. The standard error of the result prediction using the advanced KNN was around 3.66, while the traditional KNN had an error of about 3.97. The work concluded that the improved KNN showed better performance with the extended information. On the other hand, Parray et al. stated the popularity of SVM, which is also a classification model, and compared the predictability of SVM with logistic regression and perceptron neural network on NIFTY 50 stocks from Jan 2013 to Dec 2018 [7]. The method of K-folds cross-validation with ten splits was used to construct the non-time series and time series SVM models. Overall, the SVM demonstrated a better performance than logistic regression. Similarly, Das et al. [8] mentioned the use of SVM in Indian financial stock future price prediction. The study concluded that SVMs with lower normalized MSE and Mean Absolute Error (MAE) forecasted more accurately predicted than BPN. Furthermore, in predictions for the Indian stock market, Nayak et al. proposed the idea of the hybrid model of SVM and KNN [9]. The SVM was used first to make a profit and loss prediction, then the SVM's result assisted in computing the optimal closest neighbors. The hybrid model was compared to the FLIT2Ns and CEFLANN. The result showed that the combination of SVM and KNN had a lower error and better prediction capability, especially in high-dimensional data.

Table 2. Summarizes all mentioned papers above under the category of KNN and SVM model.

Year	Author	Dataset	Time Range	Algorithm	Result
2019	Bhardwaj et al.	Apple Inc., Tata Consultancy service, Infosys, Verizon, and Johnsons & Johnsons.	2015-2019	KNN	Average of 65% accuracy.
2020	Yunneng Qian	Historical price data of stock Neimengyiji	2004 to 2020	SVM	Traditional KNN has Standard error(SE) of 3.97. Improved KNN has SE of 3.66.
2020	Parray	NIFTY 50 stocks of National stock.	Jan 2013 to dec 2018	SVM	SVM yields higher accuracy than logistic regression
2012	Das et al	National stock exchange of India limited	Jan 2007 to Dec 2010	SVM	SVM outperform the BPN
2020	Ravikumar et al	S&P500	2010-2017	SVM, KNN	The accuracy of SVM: 68.41%(linear) 64.8%(poly), 58.65%(sigmoid) accuracy KNN has accuracy of 61.5%

2.3 LSTM Model

2.3.1 LSTM Model Explanation

LSTM is a particular kind of recurrent neural network (RNN) in Deep Learning (DL). It performs well in learning long-term dependencies such as sequence prediction. LSTM has a recurring module where each contains one cell state and three gates. The forget state applies a sigmoid function to determine whether the information in the previous cell should be saved based on the output value of 0 or 1. The input gate controls the information that should be processed into the cell state. Lastly, the output state selects the memory that will be passed to the next hidden state.

2.3.2 LSTM Literature Review

In the study, Balaji et al. indicated the challenge of predicting the nonlinear behavior of stock prices [11]. Nikou et al. also stated the difficulties of stock prediction due to the stock market's nonlinearity and nonstationary [12]. Balaji et al. analyzed the problem by utilizing the various DL models such as LSTM, GRU, and CNN on the dataset S&P BSE-BANKEEX from 2015 to 2017. The two and three-hidden layer LSTM with linear activation functions were assessed in the paper. The result showed the two-layer LSTM, with an average of around 58% accuracy, had better performance

than the three-layer LSTM. Moreover, Nikou et al. compared the performance of the ML algorithm from the following four fields, which are Artificial Neural Networks (ANN), SVM, Random Forest, and DL, on the data of daily closed stock price from Jan 2018 to June 2018 [12]. In the study, they constructed the RNNs and included the LSTM blocks because of the issues of vanishing gradient. The work concluded the DL algorithm with LSTM blocks yields the lowest RMSE and best prediction among all other algorithms. Nabipour et al. constructed nine ML models, including LSTM and SVM, to predict a binary indicator where +1 meant an upward trend and -1 represented a downward trend on four stock market groups from Nov 2009 to Nov 2019 [13]. This observation was made: RNN and LSTM were the best models among all others, and the algorithms from the field of deep were the best for the binary data assessment. In addition, Kim et al. applied LSTM, MLP, CNN, and GCN to predict individual stock prediction tasks. They constructed LSTM consisting of 2 layers and a hidden size of 12 by RMSprop optimizer [14]. On average, all the models had similar accuracy. Overall, the DL methods generated accurate forecasts for the movement of stock prices.

Table 3. Summarizes all mentioned papers above under the category of LSTM model.

Year	Author	Dataset	Time Range	Algorithm	Result
2020	Balaji et al.	S&P BSE-BANKEX index	2005-2017	LSTM	Average around 58% of accuracy
2019	Nikou,et al	iShares MSCI United Kingdom exchange-traded-fund	Jan 2015 to June 2018	LSTM	LSTM yields better prediction random forest and SVM.
2020	Nabipour, et al	Four Stock market groups	2009-2019	LSTM	LSTM outperforms the KNN
2019	Kim et al	S&P 500 companies	Feb 2013 to June 2019	LSTM	All LSTM, MLP, CNN, GCN have similar performance.

3. Econometric Models in Stock Market Prediction

3.1 ARIMA Model

3.1.1 ARIMA Model Explanation

ARIMA model is defined as a model in statistics specialized in forecasting and analyzing time-series data, and its capability in short-term stock market prediction outperforms many complex models [15]. ARIMA model is a combination of three different models: The autoregressive Model (AR), the Moving Average (MA) Model, and Integration [16]. The general form of ARIMA (p, d, q) is expressed as:

$$\phi_p(B)(a - B)^d Y_t = 0 + q(B)e_t \quad (1)$$

where p, d, q, and $\{e_t\}$ stand for the order of AR, Integration, MA, and White Noise Process, respectively [16]. The ARIMA model construction process is significant, and the most robust ARIMA model can be obtained by repeating the construction process several times [17].

3.1.2 ARIMA Literature Review

In this subsection, six papers related to the ARIMA model are reviewed, analyzed, and compared. The team of Khan et al., Wati et al., and Dassanayake et al. focused on the application of the traditional ARIMA model based on different datasets [15-17]. In detail, Khan et al. were trying to determine the most accurate Netflix Stock forecasting ARIMA model [15]. In the research process, the weekly MA ($K=7$) was applied to reduce the uncovered data patterns and white noise. In order to enhance the stationarity of the sample time-series dataset, Seasonality was neglected. As a result, the ARIMA (1,1,33) model was selected as the top model not only by the measurement of Mean absolute percentage of error (MAPE), Partial Autocorrelation Function (PACF), and Autocorrelation Functions (ACFs), but also by its continuity in prediction value. Moreover, Wati et al. and Dassanayake followed a very similar research process to Wati et al. and computed the results of ARIMA (1,1,33) and ARIMA (1,1,0) to be the most accurate ARIMA model for Stocks from Amman Stock Exchange (ASE) and New Zealand Stock Market (NZX50) respectively. On the other hand, papers from Qin et al., Xiong et al., and Zeng et al. took Neural Networks (NN) into account as a comparison to the ARIMA model in forecasting [18-20]. In detail, Qin et al. selected Backpropagation in NN (BPNN) with three layers to compete with ARIMA based on PDD and JD stocks in China Stock Market [18]. In the process, the optimal ARIMA (1,2,2) was determined similarly as mentioned above by the result of AIC, ACFs and PACF. BPNN's primary input was the closing price of the previous four days, which led to 4 neurons in the first layer. The second and third hidden layers were determined by the formula and minimum Mean Squared Error (MSE) in prediction. Thus, BPNN's structure was 4-3-1. MSE and Average Absolute Error (MAE) were applied to evaluate each model's effectiveness in forecasting closing prices. The results indicate that BPNN did not necessarily have higher accuracy than the ARIMA model. While the ARIMA model predicted less accuracy than BPNN for JD, the conclusion was the opposite for PDD stocks. However, in the real stock market, the market volatility can lead to highly nonlinear floating forex data, which results in the ARIMA model forecasting undesirably [19]. As an improvement, the team of Xiong et al. not only researched ARIMA and BPNN in the Chinese Stock Exchange (CSE) but also introduced the innovative hybrid ARIMA-BPNN model as a combination [20]. ARIMA-BPNN combined the advantages of both models as the ARIMA model was applied to identify the linear structure, while BPNN was used for capturing the nonlinear structure and forecasting more accurately. Results showed that ARIMA-BPNN outperformed other models but with a relatively higher training time. Lastly, the team of Zeng et al. presented the Attention-based Recurrent NN (ARNN)-ARIMA model in the paper with the ARIMA model to fit the linear correlation and ARNN best to capture nonlinear structure as well as the robustness test [19]. In detail, the gap between the actual values and forecast using ARNN was reduced significantly with the help of the ARIMA model. The conclusion proved that ARNN-ARIMA boosted the model accuracy while the training time also increased.

Table 4. Summarizes all mentioned papers above under the category of ARIMA model.

Year	Literature Reference	Method/ Approach	Dataset	Time Range	Result	Limitation
2020	Khan et al.	ARIMA model	Netflix Stock Data from Yahoo Finance	From 7 April 2015 to 7 April 2022, 5 years	ARIMA (1,1,33) model performed best	More model types should be considered
2018	Wati et al.	ARIMA model	Stock data from ASE	From Jan.2010 to Jan.2018	Best model is ARIMA (2,1,1) with RMSE=4.00	This model is only for short term forecasting
2021	Dassanayake et al.	ARIMA model	Stock data from NZX50	From 2009 to 2017	ARIMA(1,1,0) performed best	Result may not be duplicate in other stock markets
2021	Qin et al.	ARIMA model BPNN	PDD & JD stock data from Yahoo Finance	From July 26, 2018 to January 29, 2021	BPNN performs better on JD stocks ARIMA performs better on PDD stocks	Models did not combine with the trend of future stock price changes
2017	Xiong et al.	Hybrid ARIMA-BPNN model / ARIMA model/ BPNN	Stock data from CSE	From Sep 2, 2015 to Nov 10, 2016	ARIMA-BPNN performs better than other models	Model accuracy of 78.79% requires further improvement
2020	Zeng et al.	ARNN-ARIMA / ARIMA model / ARNN	MetaTrader5 Platform	From Jan 1, 2019 to Dec 31, 2019	ARNN-ARIMA has the highest accuracy 75.7%	Training time of ARNN-ARIMA is significantly higher than other models

3.2 CAPM

3.2.1 CAPM Model Explanation

CAPM was the first econometric model that provided a coherent framework for the fundamental finance problem: how the investment risk affected the expected return [21]. The basic form of CAPM in equilibrium was expressed as follows:

$$E_s = r_f + \beta(E_M - r_f) \quad (2)$$

where E_s , E_M , r_f , β , refer to the anticipated return on a particular asset, the market portfolio, the risk-free rate, and how sensitive the asset's return is to the market portfolio's return, respectively [21]. Though CAPM was only a single-factor model based on how returns and market factors are related, it followed four strict pre-assumptions that represented a highly idealized world [21]. First, investors must be risk-averse and rational. Second, capital markets are assumed to be perfect in several aspects. In detail, information is available to every investor; Taxes, trading restrictions, or transaction fees are not considered; assets are perfectly divisible; a risk-free rate is applied to all trades. Third, all investors share access to the same investment resources. Fourth, investors all have homogeneous estimates of returns.

3.2.2 CAPM Literature Review

This subsection reviewed and analyzed six papers regarding CAPM in stock market prediction. The team of ES et al. investigated CAPM's suitability in BSE and valuation accuracy for 30 stocks in India [22]. MAPE, Simple Regression Analysis and Correlation Matrix were statistical tools for measuring each Sensex Stock's valuation accuracy, proving a significant relationship between Intrinsic value and stock price, and finding the relationship between CAPM's variables, respectively. The results showed that CAPM yielded an average MAPE of 25.40, and the variables of CAPM did not have multicollinearity issues. In conclusion, a positive and strong relationship between India's intrinsic value and stock price and the high valuation accuracy of CAPM was proven. However, such optimal research results were hard to duplicate in other stock markets as any violations of the pre-assumptions of traditional CAPM would lead to a tremendously lower valuation accuracy and insignificant relation between investment return and market price. The conclusion was supported by the papers of Wei and Hur et al. [23,24]. In detail, Wei focused on analyzing the relation between expected investment return and investment risks based on the closing price of 51 A-share stocks [23]. The research process was very similar to that of ES et al., while the sample data rejected CAPM. The failure was explained by Wei by the restrictive factors in the Securities Market of China and the big gap between China's Stock Market and the mature western Stock Market [23]. On the other hand, Hur et al. investigated the extended CAPM β 's effect on the Security Market Line (SML) and the non-traded risk's effect in the incomplete stock market of Korea [24]. As a result, the extended CAPM β (True β) diverged from the traditional CAPM β (Perceived β) with the equilibrium conditions in the incomplete stock market, and the traditional CAPM had lots of limitations in real-world stock markets. As an improvement in valuation accuracy for the traditional CAPM, the team of Bao et al. and Kumar et al. introduced the generalized CAPM with IAPD (independent and identically asymmetric power distributed) error & the generalized CAPM with IAEPD (independent and identically asymmetric exponential power distributed), and Liquidity-adjusted CAPM (LCAPM) respectively [25,26]. To be more specific, Bao et al. concentrated on whether the CAPM with abnormal errors (CAPM-IAPD, CAPM-IAEPD) exhibited superior performance to the traditional CAPM in return prediction based on the EURO STOCK 50 index [25]. Maximum Likelihood test and the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm were employed on candidate models, and model performances were evaluated by the methods of the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). After testing, two generalized CAPM performed better than the normal CAPM, and CAPM-IAPD outperformed other models with the highest return of 125.36%, which also proved that two tail-shape parameters were not necessarily needed. Also, the results from the paper of Kumar et al. evidence that LCAPM produced a more accurate prediction than the traditional CAPM in forecasting the returns because the idiosyncratic risk and systematic risk were partly formed by liquidity, which led to the explanation of cross-sectional returns in India [26]. At last, the team of Chong et al. conducted a comprehensive comparison between CAPM and FF Factor Model [27].

3.3 FF Factor Model

3.3.1 FF Factor Model Explanation

FF Factor Model and the mentioned CAPM are all factor models. While CAPM is an individual factor model, FF 5-Factor Model is founded on the relationship between average returns and not only one but five risk factors: Value, Size, Market, Profitability and Investment [28].

The FF 5-Factor Time-Series Regression Model was introduced in 2015 and is shown as follows:

$$R_{it} - R_{Ft} = a_i + b_i(R_{Mt} - R_{Ft}) + s_iSMB_t + h_iHML_t + r_iRMW_t + c_iCMA_t + e_{it} \quad (3)$$

where R_{it} refers to the return on decided portfolio i over a chosen t time period; R_{Ft} refers to the risk-free rate; $R_m - R_f$ refers to the difference in return between the cash and stock markets. ; SMB_t refers to the returns on small-cap stocks minus the returns on large-cap stocks; HML_t refers to the contrast between high- and low-valued stock returns; RMW_t refers to the difference between stock

returns with strong profitability and those that have poor profitability; CMA_t refers to the return on conservative firms' stocks minus of aggressive firms' stocks [28]. The only difference between the 5-Factor Model and the 3-Factor Model is the additional factors of Investment and Profitability.

3.3.2 FF Factor Literature Review

This subsection reviewed and analyzed five papers related to the stock market prediction by using the FF Factor Model. To begin with, the team of Dirkx et al. investigated whether the FF 5-Factor Model with an additional momentum factor would produce more accurate forecasting of the German stock market based on monthly return data from CDAX [29]. In contrast, the US stock market has a more significant number of available entities, leading to calculating 4x4 weighted value portfolios instead of 5x5. During the process, Dirkx et al. used four sets of 4x4 portfolios to evaluate the performance of the FF 3-Factor Model and proposed FF 6-Factor Model. However, the additional three factors in the FF Six-Factor model do not lead to a significant change in the 16 Size/BM portfolios' coefficients and test statistics. The improvement for FF 6-Factor Model and 3-Factor Model was marginal. The same results could also be found in the paper by the team of Kubota et al., Jiao et al. and the team of Guo et al. based on the stock market of Japan and China [30-32]. The additional two factors of investment and profitability were statistically insignificant for most portfolios. This proves the conclusion from Fama-French's paper that FF Five-Factor Model worked poorly for most developing countries and some developed countries as maturity between each stock market was different [28].

Table 5. Summarizes all mentioned papers above under the category of FF Factor Model.

Year	Literature Reference	Method/ Approach	Dataset	Time Range	Result	Limitation
2020	Dirkx et al.	FF 3-Factor/ FF 5-Factor plus Momentum	Monthly returns of the CDAX constituents for German	From Jun 2002 to Jun 2019	FF 6-Factor has slight improvement then 3-Factor	More models need be added for comparison
2018	Kubota et al.	FF 5-Factor	Long run stock data for Japan	From Jan 1978 to Dec 2014	Profitability and investment factors are statistically insignificant	Model does not consider time period after 2015
2016	Guo et al.	FF 5-Factor	CSI300	From Jul 1995 to Jun 2014	The investment factor is statistically insignificant	The model does not consider alternate factors for the Chinese stock market
2017	Jiao et al.	FF 3-Factor/ FF 5-Factor	China Stock Market	From 2010 to 2015	Additional two factors are statistically insignificant	The model does not consider alternate factors for the Chinese stock market
2018	Huang	CAPM/ FF 3-Factor/ FF 5-Factor	CSMAR	From Jul 1995 to Jun 2014	FF 5- Factor has the highest accuracy	Investment and Profitability factors offer incremental but small explanatory power for stock returns

4. Results & Discussion

The discipline of data science and finance has long been interested in the problem of stock market forecasting to estimate the stock's future price behavior on a financial exchange. The application of ML algorithms and Econometric Models boosts the accuracy and stability when it comes to stock market forecasting. ML algorithms and Econometric Models are confirmed to be two of the most efficient general methodologies with high accuracy and stability in the stock market prediction domain. In the review paper, 13 and 18 research papers were analyzed and compared under the categories of ML algorithms and Econometric Models, respectively. In this section, the analysis of each model is presented separately, with an overall comparison of ML algorithms and Econometric Models at last.

Overall, a total of four ML algorithms are discussed in previous contents, which are LR, KNN, SVM, and LSTM. LSTM outperforms the other three models, especially when analyzing binary data, in stock market prediction with the highest accuracy. The main reason is that LSTM owns the ability to store information over long periods and to remember previous significant information while ignoring less critical information. Although LSTM is the optimal model, it is essential to exercise caution when deciding how the gates and number of layers will be implemented. For instance, increasing the number of stack layers in an LSTM might boost prediction accuracy and make calculation and training more difficult. In addition, combining LSTM with appropriate clustering methods or another ML algorithm can lower running times and prediction errors.

Moreover, SVM has a slightly better performance than KNN in most cases since SVM performs relatively well on high dimensional data but is not suitable for large datasets. It is noticed that the combination of SVM and KNN exhibits an even better prediction capability than the SVM or KNN. Due to the requirement of independence and linearity, using standalone linear regression models in stock prediction is unusual; instead, combining them with other ML methods is more common. Hence, the effectiveness of linear regression in forecasting stock still needs more investigation.

From the perspective of Econometric Models, three models are discussed and summarized in total: the ARIMA model, CAPM and FF Factor Model. None of the three models necessarily performs better than the other two in stock market prediction. CAPM and FF are all classified as factor models. The traditional CAPM, as a single-factor model, has four extremely strict pre-assumptions for an idealized stock market, which are hard to be fulfilled. However, once the assumptions are met, the model accuracy would boost. As an improvement in valuation accuracy, the generalized CAPM models (CAPM-IAPD /CAPM-IAEPD) and LCAPM are introduced as the idiosyncratic risk, and systematic risk was partly formed by liquidity. Moreover, the FF 5-Factor Model does not necessarily predict more accurately than the FF Three-Factor Model as the additional factors may cause model overfitting, resulting in additional factors being statistically insignificant. Also, as the FF Factor Model was first established based on the US stock market dataset, the exact optimal result may not be duplicated in other stock markets. Such conclusions can be proved by several research papers based on the stock market of Germany, China, Japan, India, Etc. Lastly, the ARIMA model is identified to work more efficiently while being combined with ARNN/BPNN as the ARIMA model, and ARNN/BPNN is good at capturing linear structure and nonlinear structure along with the robust, respectively.

Based on the contents of the cited paper, the further conclusions on comparison between ML algorithms and Econometric Models are (i) ML algorithms tend to have relatively higher accuracy than Econometric Models with a much longer training time in most cases as LR is an exception. (ii) The combination of ML algorithms and Econometric Models shows the best performance as the combined model best uses each model's characteristics.

The main limitations of the paper lie in the lack of model types. Only four models in ML algorithms and three models in Econometric Models are considered. Also, all papers selected are based on relatively mature stock markets (Japan, China, India, USA). Stock markets of more countries need to be considered for testing candidate models better. As several papers reviewed in the previous contents indicate, the future trend in the domain of stock market prediction should focus on the further

combination of ML algorithms and Econometric Models to better utilize models' characteristics for maximizing model accuracy. As an ending word, the paper presents a comprehensive review of the application and comparison of ML algorithms and Econometric Models in the stock market prediction domain, reflecting the writers' enthusiasm and expertise in the corresponding field.

References

- [1] Kneese, Allen V., and James Sweeney. Handbook of Natural Resource and Energy. Vol. 3. Elsevier, 1993. Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.
- [2] Umer, Muhammad, Muhammad Awais, and Muhammad Muzammul. "Stock market prediction using machine learning (ML) algorithms." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 8.4 (2019): 97-116.
- [3] Emioma, C. C., and S. O. Edeki. "Stock price prediction using machine learning on least-squares linear regression basis." Journal of Physics: Conference Series. Vol. 1734. No. 1. IOP Publishing, 2021.
- [4] Agrawal, S. C. (2021, April). Deep learning based non-linear regression for Stock Prediction. In IOP Conference Series: Materials Science and Engineering (Vol. 1116, No. 1, p. 012189). IOP Publishing.
- [5] Bhardwaj, Neha, and M. A. Ansari. "Prediction of stock market using machine learning algorithms." International Research Journal of Engineering and Technology 6.5 (2019).
- [6] Yunneng, Q. (2020, December). A new stock price prediction model based on improved KNN. In 2020 7th International Conference on Information Science and Control Engineering (ICISCE) (pp. 77-80). IEEE.
- [7] Parray, Irfan Ramzan, et al. "Time series data analysis of stock price movement using machine learning techniques." Soft Computing 24.21 (2020): 16509-16517.
- [8] Das, Shom Prasad, and Sudarsan Padhy. "Support vector machines for prediction of futures prices in Indian stock market." International Journal of Computer Applications 41.3 (2012).
- [9] Nayak, Rudra Kalyan, Debahuti Mishra, and Amiya Kumar Rath. "A Naïve SVM-KNN based stock market trend reversal analysis for Indian benchmark indices." Applied Soft Computing 35 (2015): 670-680.
- [10] Ravikumar, S., & Saraf, P. (2020, June). Prediction of stock prices using machine learning (regression, classification) Algorithms. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.
- [11] Balaji, A. Jayanth, DS Harish Ram, and Binoy B. Nair. "Applicability of deep learning models for stock price forecasting an empirical study on BANKEX data." Procedia computer science 143 (2018): 947-953.
- [12] Nikou, Mahla, Gholamreza Mansourfar, and Jamshid Bagherzadeh. "Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms." Intelligent Systems in Accounting, Finance and Management 26.4 (2019): 164-174.
- [13] Nabipour, Mojtaba, et al. "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis." IEEE Access 8 (2020): 150199-150212.
- [14] Kim, Raehyun, et al. "Hats: A hierarchical graph attention network for stock movement prediction." arXiv preprint arXiv:1908.07999 (2019).
- [15] Khan, Shakir, and Hela Alghulaiaikh. "ARIMA model for accurate time series stocks forecasting." International Journal of Advanced Computer Science and Applications 11.7 (2020).
- [16] Wadi, S. A. L., et al. "Predicting closed price time series data using ARIMA Model." Modern Applied Science 12.11 (2018): 181-5.
- [17] Dassanayake, Wajira, et al. "Effectiveness of Stock Index Forecasting using ARIMA model: Evidence from New Zealand." 2021 3rd International Conference on Advancements in Computing (ICAC). IEEE, 2021.

- [18] Qin, Jiaqi, et al. "Stock price forecast based on ARIMA model and BP neural network model." 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2021.
- [19] Zeng, Zhiwen, and Matloob Khushi. "Wavelet denoising and attention-based RNN-ARIMA model to predict forex price." 2020 International joint conference on neural networks (IJCNN). IEEE, 2020.
- [20] Xiong, Li, and Yue Lu. "Hybrid ARIMA-BPNN model for time series prediction of the Chinese stock market." 2017 3rd International conference on information management (ICIM). IEEE, 2017.
- [21] Perold, André F. "The capital asset pricing model." *Journal of economic perspectives* 18.3 (2004): 3-24.
- [22] ES, Dr, and Jeena Antony. "Is CAPM still alive for sensex stocks in Indian stock market?-An empirical analysis." *International Journal of Management* 11.4 (2020).
- [23] Wei, Deng. "The Practicability Test of CAPM Model in China Security Market." *The Frontiers of Society, Science and Technology* 2.9 (2020).
- [24] Kumar, Gaurav, and Arun Kumar Misra. "Liquidity-adjusted CAPM—An empirical analysis on Indian stock market." *Cogent Economics & Finance* (2019).
- [25] Bao, T., Diks, C., & Li, H. (2018). A generalized CAPM model with asymmetric power distributed errors with an application to portfolio construction. *Economic Modelling*, 68, 611-621.
- [26] Hur, Seok-Kyun, and Chune Young Chung. "Revisiting CAPM betas in an incomplete market: Evidence from the Korean stock market." *Finance research letters* 21 (2017): 241-248.
- [27] Chong, Terence Tai Leung, et al. "Profitability of CAPM Momentum Strategies in the US Stock Market." (2017).
- [28] Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, 116(1), 1-22.
- [29] Dirkx, Philipp, and Franziska J. Peter. "The Fama-French five-factor model plus momentum: Evidence for the German market." *Schmalenbach Business Review* 72.4 (2020): 661-684.
- [30] Kubota, Keiichi, and Hitoshi Takehara. "Does the Fama and French five-factor model work well in Japan?." *International Review of Finance* 18.1 (2018): 137-146.
- [31] Guo, Bin, et al. "The five-factor asset pricing model tests for the Chinese stock market." *Pacific-Basin Finance Journal* 43 (2017): 84-106.
- [32] Jiao, Wenting, and Jean-Jacques Lilti. "Whether profitability and investment factors have additional explanatory power comparing with Fama-French Three-Factor Model: empirical evidence on Chinese A-share stock market." *China Finance and Economic Review* 5.1 (2017): 1-19.
- [33] Huang, Tzu-Lun. "Is the Fama and French five-factor model robust in the Chinese stock market?." *Asia Pacific Management Review* 24.3 (2019): 278-289.