Statistical Machine Learning Model and Commodity Futures Volatility Information for Financial Stock Market Forecasting

Denis A. Pustokhin[1], Irina V. Pustokhina*[,2]

[1] Department of Logistics and Marketing, Faculty of Economics and Business, Financial University under the Government of the Russian Federation, Leningradskiy Prospekt 55, Moscow 125993, Russian
[2] Department of Logistics, State University of Management, Moscow 109542, Russian
Emails: dapustokhin@fa.ru; iv_pustokhina@guu.ru

Abstract

A country's economy and social structure are greatly influenced by the stock market. It is extremely difficult for investors, expert analysts, and scholars in the financial industry to forecast the stock market accurately because of the pretty unstable, parametric, non-linear dynamical, and unstable character of stock price time series. In the financial sector, stock market forecasting is a critical activity and a prominent study subject because stock market investments carry greater risk. It's conceivable, however, to reduce most of the risk through the development of computationally intelligent approaches. This paper introduces the support vector machine regression to make a model forecasting the stock market financial.

Keywords: SVM; Machine Learning; Forecasting; Regression

1. Introduction

Companies can raise funding for research and innovation, new products, emerging businesses, economic expansion, acquisition of competitors, etc. by selling stock securities in the open market. The ownership of a firm is represented by the ownership of a share. The stock market's performance has a significant impact on a country's economy and social structure[1]–[3]. Many countries' economic development is impacted by financial activity, and the stock market plays an important part in this[4]. To put it simply, investing in stocks is a riskier endeavor than investing in other financial markets, but a well-executed strategy can mitigate most of that risk. As a result, predicting the stock market before investing in it is a new challenge. In the literature on time series and intelligent systems, stock price forecasting is one of the most difficult challenges for financial analysts[5]. Investors who can correctly predict the stock market will be well rewarded.

Stock market forecasts can be made using a variety of methods. These methods can be divided into four broad categories: (1) fundamental analysis; (2) technical analysis; (3) classic statistical methods; and (4) soft computing. The two most prominent methods for studying and predicting stock market activity are quantitative data and technical analysis [6]. Before buying a company's shares, investors who choose the first strategy look at a variety of data that show how healthy the business is. Various variables such as turnover and expenses, yearly and quarterly reports, financial statements, financial assets, balance-sheet, etc. are examined by fundamental analysts. Long-term investors prefer fundamental analysis. The study of market statistics is the basis of technical analysis. In the opinion of the technical analyst, the stock price already contains all of the factors

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

that influence it and can be modeled using previous data. Analysis of time series shapes allows technical analysts to anticipate future behavior based on historical behaviors of time series, assuming that the past may repeat itself [7].

Economic conditions of companies and countries, bank rates, currency exchange rates, commodity prices, gold prices, stock market movement, preconceptions of investors, political rallies, company policies, investor psychology, and many other macroeconomic factors influence stock market volatility [8]–[10]. Many alternative techniques to stock market forecasting and intelligent decision-making have been proposed in the last few years. Statistical techniques and soft computing methodologies [11]are two of the most often utilized ways of predicting stock price time series. ARMA, ES, ARIMA, ARCH, and GARCH are all examples of traditional statistical forecasting methods that use past stock price data to predict future stock price movements [12]. These methods also include exponential smoothing, autoregressive integrated moving averages, and autoregressive conditional heteroskedasticity. To predict future series values, these models assume that the financial sequence under examination is created by a linear process [13]. There is a great deal of noise in stock price time series data because of the non-linear structure of the series and the complexity of the data. To represent the non-stationary and complicated character of stock markets, typical statistical methods cannot be employed at all.

Even though there is no universally accepted definition of machine learning, it may be regarded as a set of techniques that automatically create predictions from complicated data[14]. For the most part, machine learning employs a function-fitting technique to discover a usable approximation of the functionality that underpins the predictive connection between the input and output of data.[15]. Machine learning uses statistical and computational methodologies from computer science in its search for patterns in the data [16]. Additionally, machine learning seeks to deal with high-dimensional data. A high dimensionality problem develops when the number of factors (independent variables, features) utilized to predict the output variable (dependent variable) is considerable relative to the number of predications available for comparison. In this situation, conventional statistical procedures are inapplicable [14].

Supervised learning, unsupervised learning, and reinforcement learning are the three major types of machine learning. In the context of supervised learning, it is possible to anticipate an output variable's value based on a collection of input variables' values. Supervised learning depends on a set of inputs that are jointly monitored for each data point to do this [15]. Using input factors like time of year, price, advertising expenditures, and the existence of competitors, one may anticipate the sales of a product. A collection of input observations whose joint distribution has been determined is the sole need for unsupervised learning. The output, on the other hand, remains a mystery (response). Direct inference of these observations' characteristics is the objective [15]. It is an instance of unsupervised learning where clients are categorized into (previously unknown) client personas based on their observable attributes, such as their purchasing habits. Reinforcement learning is a kind of machine learning in which the algorithm interacts with its environment to find the best way to maximize a numeric cumulative reward [17]. The program generates its data by engaging with its surroundings. Classical examples of reinforcement learning in games include chess, checkers, and go. Semi-supervised learning, which is often referred to as a "fourth" category, is a hybrid of the supervised and unsupervised learning approaches, integrating small amounts of labeled data with a substantial quantity of unlabeled data. Increasing the amount of labeled data available for supervised learning is the goal of this study [18]. Typically, supervised techniques are used for FP&A goals, which are focused on making projections from the number of inputs and assumptions.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

2. Materials and methods

The sample covers the quarterly log returns of Crude Oil, Western Texan Intermediate (WTI), and 74 main S&P 500 share prices. In this research, we use monthly log returns instead of daily ones to try to filter out the impact of unimportant economic variables like political headlines. The 74 stocks were picked based on the magnitude of their market capitalisation. We see a linear average relationship line of experimental log returns, shown against the log returns of 2 fuel prices and 74 stock prices. These occurrences are already well known to the public. All of the information for our sample was gathered from Yahoo Finance. For all of our investigations, we first transformed the original price data into log returns.

It is a nonlinear relationship between a reaction and one or more variables that are described by the support vector machine. The SVM answer might be normal, binomial, or Poisson, unlike the simple linear or exponentially regression, in which the answer has a normal distribution. A linear description is given a connection function f in the SVM. The response exhibits a normal distribution in the following prediction analysis.

For example, the SVM regularisation shrinkage approach appropriate adaptation descriptors, choose descriptors and produce fewer coefficients in a model formula by imposing a penalty function to reduce the model complexity. Cross-validation was carried out five and ten times in this study [19]. A total of 10 folds were randomly partitioned from the overall dataset, and the model was trained and validated on nine of the folds. The model's prediction accuracy was thought to be affected by a very little amount of random data selection.

The nonlinear mapping x was firstly conceived and translated to the m-dimensional feature set in the matching regression of the support vector machine. In the feature set, a linear method was created, which yielded:

$$M(w,e) = \sum_{i=1}^{n} e_i\, g_i(w) + b(1)$$

Nonlinear transformations are represented as $g_i(w)$, where the j values range from 1 to m. Normally, after processing, the data were presumed to be 0 and the deviation term was inconsequential.

The SVM model's capacity to predict is evaluated by the error function. Errors in the wide feature set were not sensitive to linear regression. linear regression decreased model complexity by keeping training samples from deviating from their non-sensitive areas as close to zero as possible The kernel function R of the SVM regression method can be generated by adding non-sensitive parameters.

$$R(w,w_i) = exp(-\frac{||w-y||^2}{2\sigma^2})(2)$$

A direct generalization evaluation can be obtained by using SVMR models. Non-zero adoption coefficients limit the leave-one-out technique, which is based on a sample-to-sample ratio of 0. If you use the SVM model for non-sparse fitting, you can develop a set of generalized error metrics that takes into account all samples and each training set. This boundary is smaller than the standard regression model for support vector machines. To exclude outliers, it's simple when all the training sites fail to make a single error on one boundary. [20]Reached the minimal value of the loss function by assuming that all of the training examples in the repeated computation.

$$y_i = \sum_{j\neq 1} y_i R(w_i, w_j)(3)$$

As a result of setting the goal function's minimal value to 0, the start leaving SVM model can more easily be optimized. The leave-one-out method and the SVM model are particularly well-suited for dealing with the unified regression issue that arises when working with samples of many types. There are still no standardized variables, the approach can provide a deeper understanding of the data sample based on the error limits and repeated repetitions of empirical computations.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

3. Results

As part of our effort to mimic stock market crises, we have gathered data from a variety of sources and databases. We used the data from the Yahoo Financial Website. We collected data from 1 January 2022 to 15 July 2022. Figure 1 shows the visualization data from 1 January 2002 to 15 July 2022. From 2002 to 2004 the marker stock prices decreased. From 2004 to 2008 the prices increased. But the international cries in 2008 the prices decreased from 2008 to 2010. From 2012 to 2016 the prices increased. From 2021 the prices decreased due to the COVID-19 and war in Russia and Ukraine.

The dataset has 5169 rows and 7 columns. The amount of all closes is 5169.000000. The average is 95.477193. The standard deviation is 49.695885. The minimum 33.700001 is and the maximum is 242.970001.

We compute the seasonal trend in the dataset in figure 2. Figure 3 shows the heat map. Figure 4 shows the correlation between datasets in a heat map. Figure 5 shows the relationship between the dataset.

We make statistics probability distribution in the dataset as in figures 6-10.

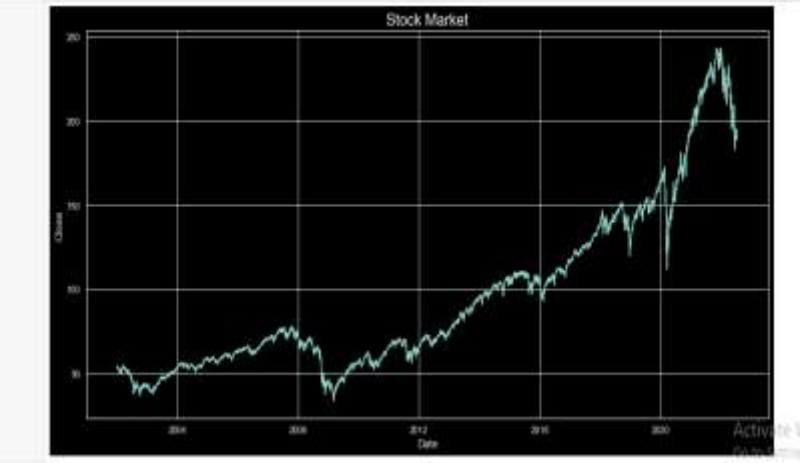

Figure 1: Visualization of stock market prices from 1 January 2002 to 15 July 2022.

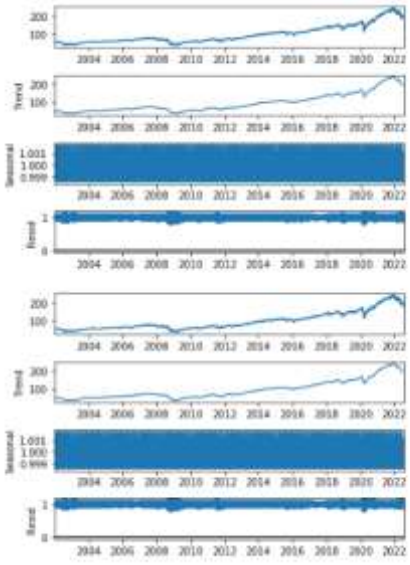

Figure 2: The Trend and Seasonal data.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

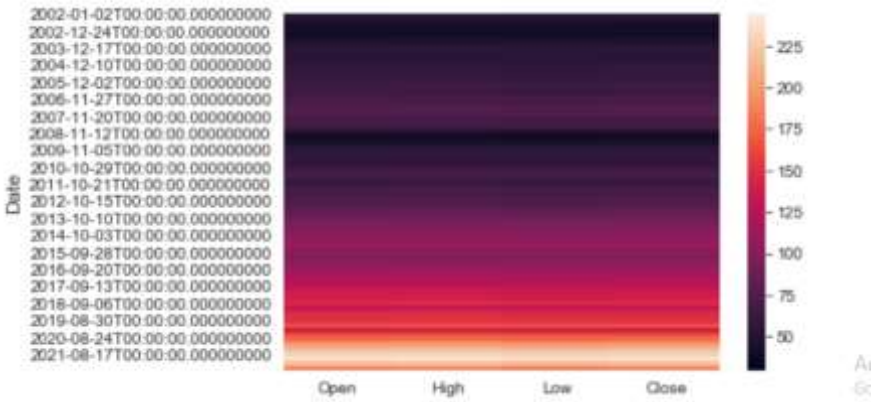

Figure 3: The heat map between dataset.

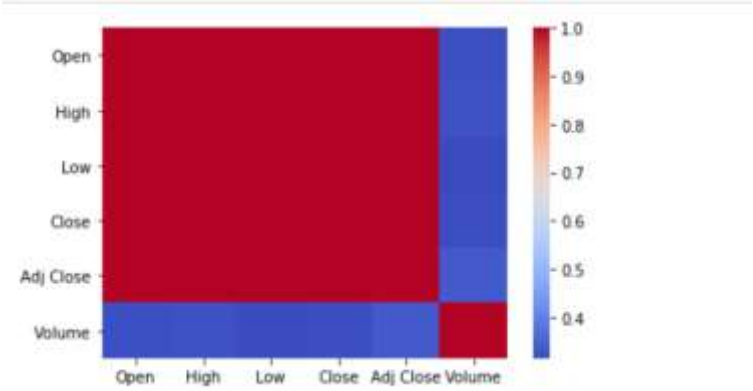

Figure 4: The correlation between dataset.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

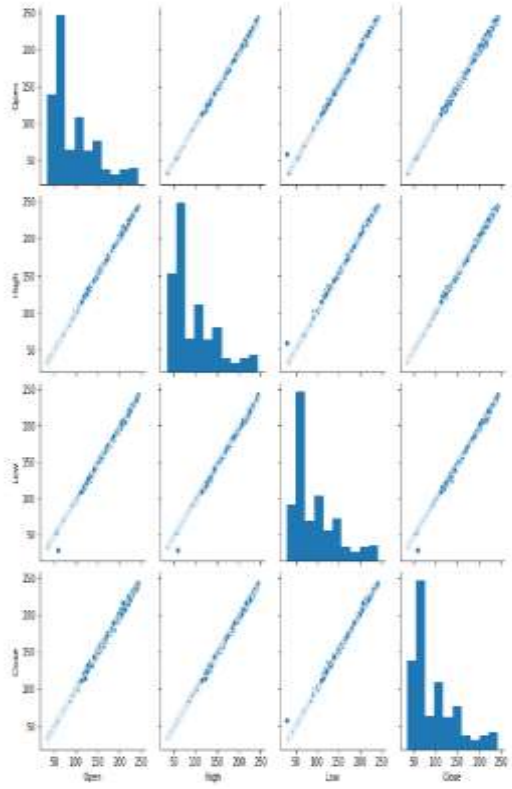

Figure 5: The relationship between datasets.

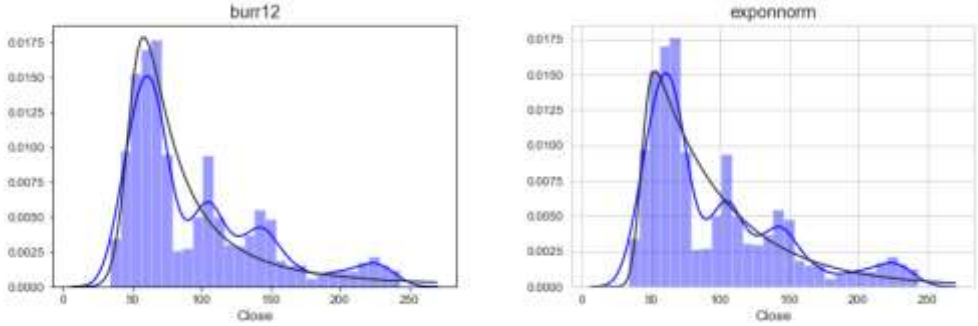

Figure 6: The statistics probability distribution function 1,2.

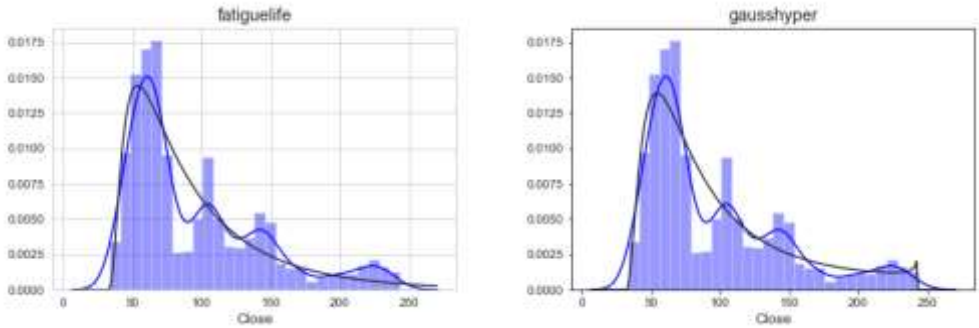

Figure 7: The statistics probability distribution function 3,4.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

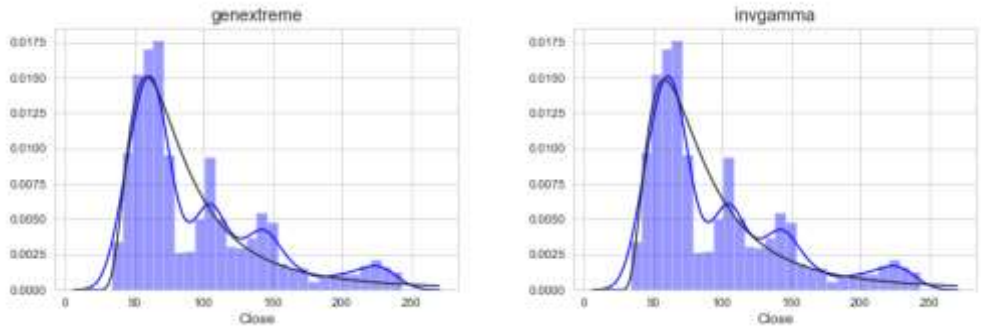

Figure 8: The statistics probability distribution function 5,6.

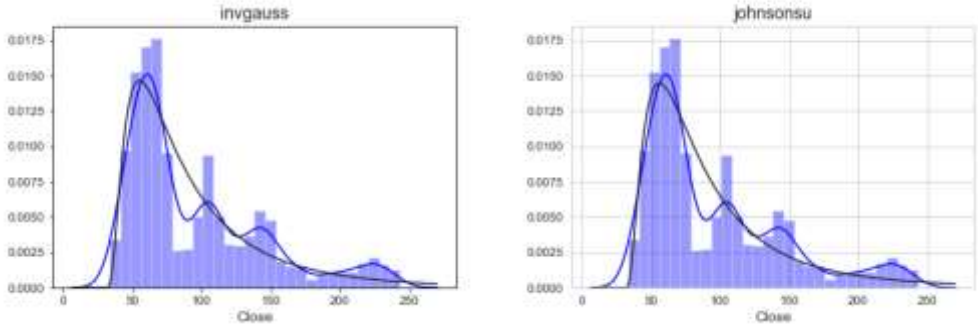

Figure 9: The statistics probability distribution function 7,8.

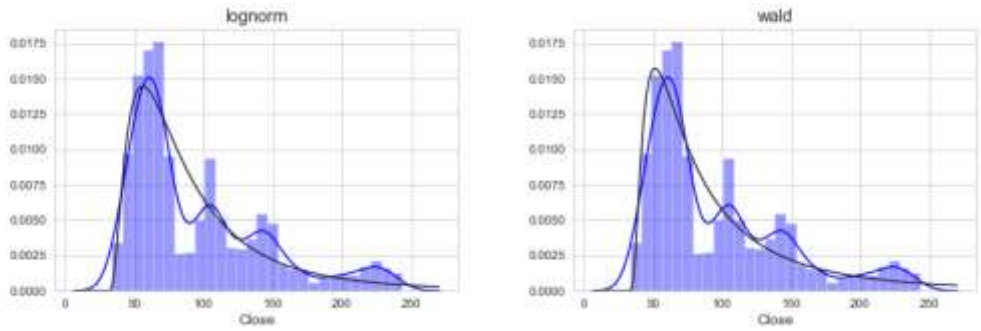

Figure 10: The statistics probability distribution function 9,10.

We divide the dataset into a train and test dataset. The training dataset has 67% of the dataset and the test has a 33% of the dataset. Table 1 shows this division. Then apply the steps of the SVM. We fit the training dataset, then predict the input dataset. Table 2 shows some methods of the error to measure the accuracy of the model.

Table 1: The train and test dataset.

	Train	Test
Total	5169	
Percentage	67%	33%
Amount	4135	1034

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

Table 2: The errors of the prediction.

Method	Value
Mean Square error	4.34392093412326
Mean absolute error	0.911120177551219
r2 score	0.9982092360809871

Measure Performance

The AI model's efficacy is measured by the performance measurements, which may be either statistical or informal. MSE, MAE, MSPE, RMSE,MAPE, normalised NMSE, squared significant relation (R2), and standard deviation are some of the key metrics used in the surveyed articles. While statistical mistakes may give you a sense of the big picture, they aren't enough to evaluate the model on their own since they don't tell you which way the stock market is going to go (i.e., higher or lower next price). Thus, most of the papers include other performance indicators. The HIT rate, which assesses how accurate the suggested model's trend forecasts are, is the most popular of the non-statistical indicators, since it directly relates to the forecast's potential for financial gain. Non-statistical measurements include things like an average return, a trend forecast, yearly profit, break-even information costs, and continuously compounded return. However, only a small number of papers give definitive results concerning which metric is the best for comparing the performance of different models for making forecasts.

The T-test has been used to determine whether or not the Classification model is superior to other classifiers like the support vector machine (SVM). Consequently, the T-test is used to compare the efficacy of the two leading classifiers, "KNN and SVM." When comparing classifiers, it is clear that there is a statistically significant difference. Results are shownas for t test −4.6572 and p value 1.4371E-0.4.

After plugging in the monthly RV of 19 commodities futures into the baseline model, we are able to extract the predictions of the 19 AR individual models. The predictions of the shrinkage specific products are generated using the same indicators as those of the AR individual model. We can see the p-values for the Clark and West MSFE-adjusted statistic, the out-of-sample projecting efficiency of each individual component used to predict the DJIA, and the significance level of these results. Nearly of the negative results show that the classic and elastic net approaches' individual models have weak predictive ability. While many non-lasso models fail to provide good results, we show that most lasso-based models may produce substantial results at the 10% or 5% ranges. In particular, the price volatility knowledge of energy markets demonstrates better prediction performance for predicting future the RV of the DJIA index, as evidenced by the fact that all individual models that included energy primary commodities relevant data (i.e., crude oil, fuel oil, oil and gas, and regular gasoline gasoline) can create good attributes and are meaningful at least at the 10% level.

Statistics on oil and stock prices' monthly log values are shown in Table 3. We found that the macroeconomic features of Brent and WTI are similar: the log return of Brent and WTI pricing are distributed normally with fat tails, and the mean log returned of Esso and WTI pricing are near zero.

Standard errors of BRENT vs WTI log returns are comparable, as seen in Table 1. Oil prices will rise in the future because the skewness values in the log returns of BRENT and WTI over this time period are positive. Furthermore, the log return of BRENT and WTI have levels of kurtosis that are larger than 3, indicating that they possess hefty tails relative to a normal curve. Brent and WTI weekly log returns throughout the time period are shown in Figure 1. Figure 1 depicts the dramatic reduction in Brent and WTI log returns during the 2008 economic crisis. During the first half of that

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

year, these returns were quite high. Brent and WTI's log returns were low in December 2018 because investors were worried about rising interest rates, a weakening economy, and escalating trade tensions between the United States and China.

Table 3: Statistics Analysis

	Mean	Median	Minimum	Maximum
BRENT	–0.003	–0.015	–0.192	0.326
WTI	–0.002	–0.014	–0.215	0.336
BRENT	St.D	Skewness	Kurtosis	
WTI	0.089	0.968	4.476	

4. Conclusion

Financiers are expected to offer accurate financial predictions as well as strategies for efficient and effective resource allocation in today's businesses. Fast and accurate planning and forecasting are critical in volatile or rapidly changing market conditions. Strong financial functions are defined by their ability to provide accurate forecasts. As a result, it's not unexpected that most big corporations have FP&A teams inside their finance department. In this study, we used the Yahoo financial dataset to predict the stock market. We divide the dataset into a train and test dataset. We used the SVM model to make forecast the stock market. We proposed some methods to compute the error of this study. In the future study, the deep learning model can be used to forecast the stock market financials.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022

References

[1] D. P. Gandhmal and K. Kumar, "Systematic analysis and review of stock market prediction techniques," *Computer Science Review*, vol. 34, p. 100190, 2019.

[2] M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, and E. Salwana, "Deep learning for stock market prediction," *Entropy*, vol. 22, no. 8, p. 840, 2020.

[3] B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers," *Applied Intelligence*, vol. 26, no. 1, pp. 25–33, 2007.

[4] C.-S. Lin, S.-H. Chiu, and T.-Y. Lin, "Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting," *Economic Modelling*, vol. 29, no. 6, pp. 2583–2590, 2012.

[5] F. E. H. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," *omega*, vol. 29, no. 4, pp. 309–317, 2001.

[6] M. Lam, "Neural network techniques for financial performance prediction: integrating fundamental and technical analysis," *Decision support systems*, vol. 37, no. 4, pp. 567–581, 2004.

[7] J. J. Murphy, *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[8] K. Miao, F. Chen, and Z. G. Zhao, "Stock price forecast based on bacterial colony RBF neural network," *Journal of Qingdao University (Natural Science Edition)*, vol. 2, no. 11, 2007.

[9] K. N. Arman, Y. W. Teh, and N. C. L. David, "A novel FOREX prediction methodology based on fundamental data," *African Journal of Business Management*, vol. 5, no. 20, pp. 8322–8330, 2011.

[10] H. Haleh, B. A. Moghaddam, and S. Ebrahimijam, "A new approach to forecasting stock price with EKF data fusion," *International Journal of Trade, Economics and Finance*, vol. 2, no. 2, p. 109, 2011.

[11] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[12] D. A. Kumar and S. Murugan, "Performance analysis of Indian stock market index using neural network time series model," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013, pp. 72–78.

[13] J.-Z. Wang, J.-J. Wang, Z.-G. Zhang, and S.-P. Guo, "Forecasting stock indices with back propagation neural network," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14346–14355, 2011.

[14] M. Taddy, *Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw Hill Professional, 2019.

[15] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[16] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4, no. 4. Springer, 2006.

[17] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[18] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[19] J.-H. Kim, "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap," *Computational statistics & data analysis*, vol. 53, no. 11, pp. 3735–3745, 2009.

[20] S. R. Sain, "The nature of statistical learning theory." Taylor & Francis, 1996.

DOI: https://doi.org/10.54216/AJBOR.070203
Received: April 15, 2022 Accepted: August 26, 2022