Comparative Analysis of Machine Learning Models for Predicting Stock Prices in High-Volatility Scenarios

Jiashuo Xing

Department of Computer Science and Technology, Beijing Institute of Technology, Zhuhai Campus, Zhuhai, Guangdong 519088, China
email: stu2103399@cgt.bitzh.edu.cn

Abstract. This work aimed at predicting Tesla's stock price with focus on high volatility environments by the aid of machine learning algorithms. The research work covered the period between January 2019 to April 2024, the dataset sourced from Yahoo Finance and contained historical stock prices of the selected company, which was Tesla. Data preparation steps included date to ordinal transformation, calculating technical indicators using TA-Lib such as SMA20, RSI14, and many more, and ordering missing observations. The time series was divided into training and testing series from January 1, 2019 to January 1, 2024 and April 29, 2024, respectively. Three machine learning models were implemented using scikit-learn: Linear Regression, Random Forest and Support Vector Machine (SVM) it is type of machine learning techniques. Linear regression uses a straight line to model the variables, Random Forest builds many numbers of decision trees and provides average value, SVM identifies a hyper plane in a higher dimensional space such that the distance between the closest data points is maximized. Model effectiveness was assessed through the MSE and R^2 coefficient, which gave an indication of how well each model was performing in predicting the stock prices of Tesla in a highly volatile market. The results also indicated that SVM had a higher accuracy and sensitivity in the high volatility environment than both Linear Regression and Random Forest in predicting Tesla's stock prices. This shows the need to apply enhanced machine learning method for ANS for efficient financial forecasting in volatile environment.

Keywords: Stock Price Prediction, Machine Learning, High-Volatility Markets, Financial Forecasting, Comparative Model Analysis.

1 Introduction

Stock price prediction plays an important role as it helps investors decide and ensure the stability of the markets. As technology advances, the application of machine learning algorithms often supplements rather than replaces traditional methods, given their efficiency in adapting to rapidly changing and increasingly complex markets. These algorithms serve as the foundation of enabling investors to operate in increasingly complex financial structures, which can help them effectively predict future market trends. Machine learning means the forecast of stock prices is a new

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_46

direction different from the historical methods of linear models and simple statistics. In the present times, employing modern algorithms, it is possible to analyze all the existing large-scale data starting from the market trends and scales of the global economy up to the sentiments in the social networks and significant geopolitical events which are impossible under the traditional approach.

Previous studies have primarily focused on conventional statistical and easy to apply Machine Learning (ML) algorithms. Some of the modern-day developments employ deep learning as well as reinforcement learning, which demonstrate signs of helping to manage data patterns. Nevertheless, the further quantitative comparison of these models in uncertain circumstances, for example, the Tesla stock exchange [1], is still largely unknown. For apparent reasons, comparing with direct counterparts under similar volatility conditions is rare in most studies and this points to a major area of research gap. Instead, the analysis that this research has shown is that there is a need for new theories and frameworks that can guide the practical trader especially from the position of unpredictability and volatility of the market. In the prior works, potentials of stock price prediction models at various stages in the market, along with their basic frameworks and methodologies have been introduced. These are good starting points for developing more advanced understandings of an extended stock market positioning [2]. They pave way for the creation of accurate models which reflect the normality of stocks with high volatility such as Tesla to provide for situations when the abnormal conditions arise. In addition, machine learning is not limited to application in the financial industries only but also in fields like software engineering, geoscience and genetics among others where it can perform numerous analytical as well as prediction tasks. For instance, in software engineering, predictive modeling and automatic code generation have benefited from the machine learning techniques as crucial tools appropriate for fields without clear algorithmic approaches and substantial adaptability to altering environments [2]. Likewise, in geosciences and remote sensing, machine learning enables complicated regression and classification, the essential procedures for earth science data [3]. In genetics and genomics, it helps in processing large data sets, including genome sequencing, and increases awareness of the genetic roots of diseases as well as support the concept of precision medication [4, 5].

This paper conducts an empirical analysis of Tesla stocks, thoroughly comparing the performance of three models: The three algorithms used in this study are linear regression, random forests, and support vector machines for a high-volatility market. The study also underlines the significance of dynamic and resourceful forecasting techniques especially when the market is volatile or unpredictable, it further offers profound understanding into the functionality of the models in relation to the market trends. This work is intended to provide research findings to complement the literature in the fintech field, as well as provide recommendations for investors and market analysts, and contribute to the enhancement of the existing and the development of new specific approaches to the financial analysis in changing markets.

2 Method

This section outlines the methodology employed in this study, encompassing data preparation and the implementation of three distinct machine learning models to predict Tesla's stock price.

2.1 Data Preparation

The data used in this research was obtained from Yahoo Finance [6], the period considered for this analysis was from January 1, 2019, to April 29, 2024. This dataset contains information of each company's daily stock prices such as the opening price, closing price, highest price, the lowest price, and the trading volume. It was done mainly on the closing prices because they are a sum of the market opinion for the respective day.

Some of the data preprocessing done included transforming the date column into an ordinal format that would be suitable for regression analysis. The indicators including 20-period Simple Moving Average (SMA20) and the 14-period Relative Strength Index (RSI14) were calculated using TA-Lab libraries and also included in the dataset for the prediction. This paper employed back filling technique in handling missing values with an aim of having continuity in the technical indicators. Last but not least, the dataset was divided into the training set, including information from 01/01/2019 to 01/01/2024, and the testing set for evaluation containing data from 02/01/2024 to 04/29/2024.

2.2 Machine Learning Models

The study utilized machine learning models including Linear Regression, Random Forest, and Support Vector Machines, implemented via the 'scikit-learn' library [7]. Model performance was evaluated based on Mean Squared Error (MSE) and R^2 score, which helped in assessing the accuracy and the predictive reliability of the models on the testing dataset.

Linear Regression. A linear regression analysis is one of the uncomplicated estimates employed in statistics to relate the dependent variable with one or several predictors– an approach that uses straight lines. These include employing least squared to estimate the coefficients of the line known as line of best fit as well as exploiting the residuals to measure the total sum of the difference between the actual values and the predicted values. This line, known as the least squares line, best fits the data set by reducing these differences to their smallest possible value [8].

As its name suggests, linear regression is a method for modeling the relationship between a continuous dependent variable and one or more regressor variables. It accomplishes this by defining a straight-line relationship between the dependent and independent variables, which can be represented mathematically as follows:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \tag{1}$$

where Y represents the dependent variable, X_1 through X_n are the independent variables, and β_0 to β_n denote the coefficients [9].

The linear regression principle involves selection of coefficients for the linear equation which give the lowest possible sum of squared differences between the actual values and the forecasted values of the dependent variable. The maximum likelihood method provides the best unbiased estimators for the parameters in the model, when all the assumptions of the model are valid.

Random Forest. Random forest is a machine learning technique that constructs multiple decision trees during training as part of an ensemble learning method. In classification tasks, it selects the class with the highest frequency among the trees' predictions. For regression tasks, it calculates the average of the predictions from all the trees [10].

Random forest constructs a large number of decision trees and combines them to achieve more accurate and stable predictions. Each tree is created using a bootstrap sample from the training set, which involves selecting a subset of the training data where some data points may be repeated, and others excluded. During tree construction, the optimal split for each node is chosen from all input features or a random subset of them. This method ensures the uniqueness of each tree, effectively reduces variance, prevents overfitting, and enhances the overall performance and stability of the model. The core idea of random forest is to improve predictive accuracy and control overfitting by aggregating the results of multiple decision trees. Each tree is trained on different subsets of the same training data, and randomly selecting features for each split helps to further reduce variance and avoid overfitting.

Support Vector Machine. Support Vector Machine (SVM) is one of the most popular, powerful, and flexible methods that can be used for linear classification, linear regression, and even outlier detection. It is particularly effective in classifying large but only small or medium-sized data sets [11].

SVM generally seeks a hyperplane in the N-dimensional space, where N is the number of features, to clearly classify the data points. The option of separating hyperplanes is not unique and may comprise a large number of hyperplanes with different characteristics for the separation of the two classes of data points. In other words, the classifier is aimed at identifying a plane that has the largest margin, or, in other words, it is the largest distance between the instances of different classes.

Support Vector Machine (SVM) maps data to a high-dimensional feature space to categorize data points effectively, even when the original data points are not linearly separable. By transforming the data into a higher dimension, SVM aims to find an optimal hyperplane that can clearly separate different classes, making it easier to classify complex datasets. This technique allows SVM to handle cases where the relationships between data points are intricate and not easily distinguishable in lower dimensions. Identifying the separator between categories and transforming the data so that this separator becomes a hyperplane allows for predicting the category of new data based on its characteristics.

3 Results and Discussion

Compare the performance of three machine learning models—Linear Regression, Random Forest, and Support Vector Regression (SVR)—using the test dataset. The results are presented in Table 1, which gives Mean Squared Error (MSE) and coefficient of determination (R^2) for each of the models.

Model	MSE	R^2 Score
Linear Regression	9.38	0.98
Random Forest	25.01	0.95
SVR	8.25	0.99

Table 1. Model performance evaluated by different metrics.

The above research proves that the SVR model has a better performance when it comes to Tesla's stock price prediction. It has the least amount of mean squared error and the highest coefficient of determination thus implying its ability in capturing the price movements. On the other hand, Random Forest model performs reasonably well at the higher-level trends, but it scored slightly higher MSE reflecting its lesser accuracy during the fluctuating periods. The trained Linear Regression model has a fairly average mean squared error and R^2 score. It is useful in predicting general trends, but it has issues when it comes to fluctuations in prices.

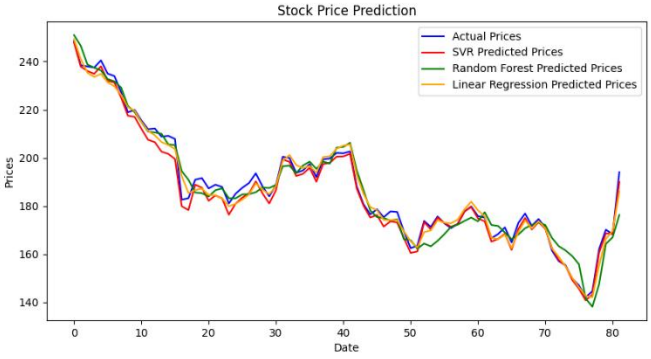

Fig. 1. Actual vs Predicted Closing Prices for Tesla Stocks (Photo/Picture credit : Original).

Fig. 1 gives a more intuitive comparison of the differences between each model. From all the models developed in SVR, the actual chart of Tesla's stock price resembles most to the developed models. The Random Forest model, in general, has relatively higher deviations despite the fact that it also is capable of capturing the holistic trend. The Linear Regression model has the disadvantage of not being able to hand elasticity well even though it can give a good estimate of the general trend. In

summary, it can be seen in Fig. 1 that the SVR model employed in this study achieved a higher level of performance in each of the evaluation metrics identified above.

The evaluation of the three models of machine learning which include Linear Regression, Random Forest and SVR brings out important observations on the machines performance in predicting the stock prices of Tesla under volatile environments. From the above evaluations, it is evident that SVR has the maximum accuracy among the models. This characteristic enables this model to fit the actual stock prices closely, a feature that is advantageous when capturing the stochastic nature of the Tesla's highly volatile stock price. However, the Random Forest model, though capable of approximating trends to a certain extent, shows larger fluctuations in periods of high variability. This is due to the nature of ensemble learning methods, which build numerous decision trees and then average their answers. This approach may make it less sensitive to rapid price fluctuations. Hence, though effective in stable markets, it may not be optimally suited for the swinging which has been depicted in the case of volatile companies such as Tesla. Linear Regression has an average sort of accuracy. It only holds linear assumption, which reduces its ability to capture non-linear patterns in stock prices during rapid market changes, which results in either over or under estimation. Though it is quite easy to understand and apply, there is a decrease in its efficiency in the volatile markets.

The outcompeting of the SVR model also supports the application of models that are able to learn non-linear and complicated structures in the data for volatile markets. The more developed models such as SVR are more accurate and dependable, this is why they are used in the financial fields. Nevertheless, Random Forest, and Linear Regression models can be beneficial, although they demonstrate weak results in considering high-frequency sharp fluctuations, and thus, may require integration with other approaches. It is important to understand the specific market conditions and then select the adequate models. Investors and analysts want to make the best choices, and that is when a good predictive model comes in handy. Therefore, having such an accurate model as the SVR in moments of fluctuation of the markets, could be of great help in the management of risks and the decision-making processes of investments. Thus, future research can use a broader range of price fluctuations and meld SVR with Random Forests or a neural network. Thus, the additional, such as sentiment analysis and macroeconomic indicators, could contribute to the enhancement of model accuracy. The utilization of these models in real-time trading environment would also be beneficial for real-life experience and tuning of these models. Generalizable and robust evidence of such relationships might be gathered by longitudinal studies on different stocks and under different market conditions.

4 Conclusion

This paper's empirical analysis of three models—Linear Regression, Random Forest, and (SVR—established that these models could predict Tesla's stock prices in volatile markets. As elaborated earlier, the SVR model yielded the smallest MSE and the largest R-squared, signifying its strong capacity of accurately modelling the variation

in the prices. In terms of the results, the model indicated that the SVR model not only meant a higher level of accuracy but also holds a great level of flexibility to deal with the non-linear data as evidenced in Tesla's stock prices, making it viable in volatile risky business environment. This work has a limitation in the way it used only three models and in the way the market data that was used was limited to price data only without incorporating other data sources such as news sentiment and social media data that affect prices. Future research can concentrate on developing hybrid models that combine SVR with other advanced machine learning techniques to enhance predictive accuracy [12, 13]. Additionally, expanding the models to incorporate broader data, such as macroeconomic indicators and market sentiment, may further improve their efficiency. Longitudinal testing across different stocks and under varying market conditions are critical to ensure that the model is accurate and valid in general.

References

1. Bao, W., Yue, J., Rao, Y.: A deep learning framework for financial time series using stacked autoencoders and long-short term memory. PLOS ONE 12(7), e0180944 (2017).
2. Patel, J., Shah, S., Thakkar, P., Kotecha, K.: Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications 42(1), 259-268 (2015).
3. Zhang, D., Tsai, J.: Machine Learning and Software Engineering. Software Quality Journal (2004).
4. Lary, D.J., Alavi, A., Gandomi, A.H., Walker, A.L.: Machine learning in geosciences and remote sensing. Geoscience Frontiers (2016).
5. Libbrecht, M.W., Noble, W.S.: Machine learning applications in genetics and genomics. Nature Reviews Genetics (2015).
6. Yahoo Finance. Available at: https://finance.yahoo.com (2024).
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830 (2011).
8. Su, X., Yan, X., Tsai, C.: Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics 4 (2012).
9. Lebanon, G.: Linear Regression. (2010).
10. Breiman, L.: Random Forests. Machine Learning 45, 5-32 (2001).
11. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20, 273-297 (1995).
12. Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., Guo, L.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 103323 (2022).
13. Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M.A., Pendyala, R.M.: Machine learning approaches for estimating commercial building energy consumption. Applied Energy 208, 889-904 (2017).