Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241568

Stock price prediction of PayPal by Linear Regression, SVM, Random Forest and LSTM

Xueer Zhang

School of Mathematics and Statistics, Wuhan University, Wuhan, China

2020302011099@whu.edu.cn

Abstract. This research paper undertakes an extensive and detailed examination of four distinct machine learning models, specifically Linear Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). The stock values of PayPal are forecast employing these techniques, and Python serves as the fundamental tool in facilitating a comprehensive and thorough evaluation of their efficacy in predicting trends in the financial market. In the realm of methodology, this study encompasses a multifaceted approach. Python forms the cornerstone for data analysis, model development, and testing. The data collection process encompasses historical stock price data for PayPal, alongside an array of relevant economic indicators. The core of the study lies in the comparative analysis of the four machine learning models. Each model is rigorously tested against historical data, allowing for a nuanced understanding of their strengths and weaknesses. Linear Regression, remarkably, emerges as the standout performer in terms of predictive accuracy and consistency. This firmly establishes Linear Regression as the optimal choice for forecasting stock prices. The significance distinction of this research extends beyond its findings. It advances the discipline of financial forecasting by illuminating the comparative effectiveness of different models, offering valuable insights for guiding future research initiatives within this domain. The integration of advanced methodologies and the clear-cut conclusion regarding Linear Regression's superiority underscore the pivotal role of this study in enhancing the precision of financial market trend forecasting.

Keywords: Linear Regression, SVM, Random Forest, LSTM.

1. Introduction

1.1. Background

Stock price prediction is a crucial aspect of financial markets with significant importance. Accurate forecasts of stock prices play a pivotal role in guiding investment decisions, risk management, and the overall stability of financial markets. One primary reason for the necessity of stock price prediction is to assist investors and traders in making informed choices. Investors rely on these predictions to decide when to buy or sell stocks, optimizing their returns. Furthermore, financial institutions and asset managers use these forecasts to design investment strategies, allocate assets, and balance portfolios effectively. Stock price predictions are also essential for risk management. By anticipating price movements, investors can implement risk mitigation strategies like stop-loss orders and hedging, reducing potential losses in case of unexpected market fluctuations. In the broader context, stock price forecasts contribute to market efficiency and stability. They help prevent sudden market crashes or bubbles by allowing regulators and market participants to identify and respond to abnormal price movements in a timely manner. In a nutshell, stock price prediction serves as the cornerstone of informed decision-making, risk mitigation, and overall financial market stability. It provides a foundation for efficient and stable market operations, benefiting investors, financial institutions, and the economy as a whole.

1.2. Problem Statement

Firstly, stock price prediction is a critical and highly debated topic in the financial world due to its potential impact on investment decisions and market stability. PayPal, as a distinctive company, presents its own set of challenges and uncertainties that have yet to yield a consensus in the realm of stock price forecasting. Another major area of controversy lies in the dynamic and multifaceted nature of financial markets. A number of factors, such as economic statistics, corporate performance, market attitude, and world events, can affect how complicated stock price swings are. The argument revolves around how well-suited conventional predictive models—like support vector machines and linear regression—are for precisely describing these complex processes. Furthermore, while deep learning models like Long Short-Term Memory (LSTM) networks and machine learning techniques like random forests have garnered interest, there is still debate about how well they function at accurately predicting stock values. PayPal's uniqueness as a leading digital payments and financial technology company further intensifies the challenge. Its business model and industry position set it apart from traditional financial institutions and other tech companies, making it a distinctive subject for stock price prediction. Determining how these distinctions impact its stock price remains an open question, with no clear consensus among financial experts. In this context, the problem statement revolves around the need to develop more accurate and reliable stock price prediction models, especially for PayPal, to address the complexities and uncertainties that surround its stock valuation. Achieving a unified conclusion regarding the most effective methods for forecasting PayPal's stock price remains a compelling challenge and a subject of ongoing research and debate within the financial community.

1.3. Hypothesis

This research hypothesis asserts that Linear Regression is the most suitable model for predicting PayPal's stock prices. This hypothesis is based on several compelling reasons. Initially, it should be noted that Linear Regression is an easy-to-understand model that shows how input characteristics relate to the target variable—in this case, the stock price. This simplicity is advantageous as it allows for a direct interpretation of the influence of individual factors on stock price movements. Furthermore, Linear Regression can effectively capture historical trends in stock prices, despite assuming a linear relationship between features and stock prices. This is crucial for investors and analysts who rely on identifying long-term trends to make informed decisions. Lastly, the historical success of Linear Regression in financial and economic forecasting across various domains lends support to the hypothesis that it may also excel in predicting PayPal's stock prices. In summary, by assuming that Linear Regression is the optimal model for PayPal stock price prediction, this research aims to explore the trade-off between model complexity and predictive accuracy while shedding light on the distinctive factors affecting PayPal's stock performance.

2. Literature Review

In the world of finance, stock price prediction has been extensively studied and has produced a plethora of study findings. The continual growth of financial markets and the quick progress in information technology over the years have led to considerable improvements in methodologies and approaches for stock price prediction. A synopsis of some recent developments and trends in the field of stock price prediction research is provided below. Gangwar, Kumar, and Bijpuria conducted study on the prediction of a company's future stock prices using machine learning techniques. This study primarily used regression and LSTM-based machine learning approaches to forecast stock prices, accounting for characteristics like volume, low, high, open, and close. And its results demonstrated that employing the algorithmic approach of Support Vector Machines (SVM) in conjunction with Principal Component Analysis (PCA) for feature selection could lead to profitability [1]. In the study of Vijh, Chandola, Tikkiwal, and Kumar, the following day's closing stock price was predicted using ANN, and RF was used as a comparison. An analysis of comparative data based on RMSE, MAPE, and MBE values revealed that ANN is superior to RF in the prediction of stock prices. The comparison of the ANN model's best values with RMSE, MAPE, and MBE revealed the results [2]. Reddy predicted changes in stock indices using machine learning algorithms and data from several foreign financial marketplaces. Upon using the SVM algorithm to a substantial dataset obtained from global financial markets, overfitting problems are successfully avoided. A number of highly efficient machine learning models were put forth to forecast daily stock market trends [3]. Khaidem, Saha and Dey used a random forest classifier to predict stock market trends, and their model demonstrated robustness in forecasting stock movements. For long-term prediction, the model obtained an accuracy range of 85-95% across several datasets, including AAPL, MSFT, and Samsung. The implementation of ensemble learning techniques and the evaluation of factors including accuracy, precision, recall, and specificity proved the effectiveness of the model. Moreover, ROC curves confirmed the model's stability. This method challenged the usefulness of linear discriminant machine learning algorithms by taking a novel approach to the non-linear character of stock market forecast [4]. Panwar, Dhuriya, Johri, Yadav and Gaur employed web scraping to gather stock data and utilized SVM and Linear Regression for stock price prediction. The findings suggest that Linear Regression outperforms SVM for stock market analysis [5]. While these papers offer valuable insights into stock price prediction, research on PayPal stock price prediction remains relatively limited. This reflects the unique nature of PayPal as a financial technology company, and its specific challenges in the realm of stock price forecasting. In order to fill this knowledge gap, the present research investigates the distinct advantages and difficulties linked to forecasting the value of PayPal shares.

3. Data Description

The daily stock price data for PayPal Holdings, Inc. covering a significant eight-year period from September 2015 to September 2023 makes up the dataset used in this study. Many important considerations, ramifications, and benefits drive this dataset selection, as do its longer duration and daily frequency. First, Yahoo Finance, a reliable and well-known source of financial data, was the source of the information. Utilizing this source ensures data reliability and accessibility for research purposes. Second, the extended time span of eight years was selected to provide a comprehensive historical perspective. This prolonged duration enables researchers to capture and analyze long-term trends, market cycles, and a wide range of economic conditions. Meanwhile, the choice of daily frequency offers several benefits. Daily data offers a fine-grained perspective on price changes, enabling a thorough examination of intraday trends and oscillations. On the other hand, it facilitates the detection of seasonal and periodic trends, shedding light on potential seasonality effects in stock price movements. After preprocessing, including handling missing values, these data consist of 2014 rows and 7 columns, encompassing the seven elements: Date, Open, High, Low, Close, Adj Close, and Volume. The Figure 1 below depicting the evolution of PayPal's stock price from September 2015 to September 2023 exhibits several significant phases. The initial phase (2015-2020) is characterized by stable and gradual growth. During this period, PayPal's stock price experiences consistent upward movement, reflecting the company's increasing prominence in the digital payments and fintech industry. However, in early 2020, a sharp and abrupt decline occurs. This plunge is probably response to the global economic impact of the COVID-19 pandemic. Following the initial dip, there is a remarkable recovery and subsequent surge in the curve (2020-2021). This phase symbolizes PayPal's rapid adaptation to the changing digital landscape during the pandemic, leading to a substantial uptick in digital payments and, in turn, the company's stock price. Subsequently, from mid-2021 into 2022, the curve depicts a gradual decline. This phase is best described as a 'correction and consolidation' period, where the stock price corrects from its record highs, providing investors with an opportunity to reevaluate their positions and the stock's valuation. Moving into 2022 and 2023, the curve continues to display a declining trend. However, the rate of descent notably slows down, indicating a more stabilized period. Investors can use this extended timeframe to make informed decisions based on evolving market dynamics. The curve illustrates the historical performance of PayPal stock, showing various market conditions and distinct phases of growth, correction, and consolidation.

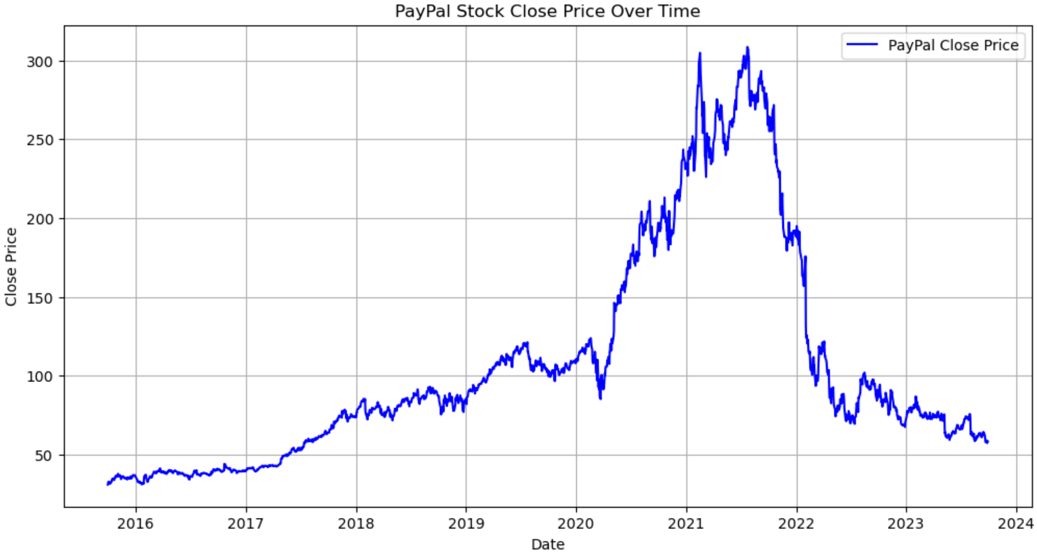

Figure 1. PayPal Stock Price Over Time

4. Methodology

This research leverages a diverse array of machine learning techniques, specifically Linear Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM), to forecast PayPal stock prices. This section furnishes an exhaustive exposition of each approach, accentuating their foundational principles, underlying concepts, strengths, and their respective applications within the realm of stock price prediction.

4.1. Linear Regression

Linear Regression represents a foundational regression analysis method with the objective of delineating a linear connection between input features and the target variable. In the context of this investigation, Linear Regression is harnessed to potentially unveil linear associations between PayPal stock prices and pertinent input characteristics. This methodology endows transparency and is aptly suited for modeling straightforward linear relationships. An exploration into the intricacies of feature selection and the operational mechanics concerning mean squared error and coefficient interpretability will be conducted [6].

4.2. Support Vector Machine (SVM)

Support Vector Machine stands as a formidable, supervised learning algorithm, renowned for its versatility in tackling both classification and regression tasks. Within this research framework, SVM is exploited to apprehend non-linear relationships and intricate patterns latent within the dataset [7]. The employment of the kernel trick empowers SVM to project data into higher-dimensional spaces, consequently augmenting predictive precision. Furthermore, an evaluation of SVM's proficiency with non-linear relationships and high-dimensional data is conducted.

4.3. Random Forest

As an ensemble learning algorithm, Random Forest combines the effectiveness of several decision trees to reduce overfitting and improve prediction accuracy. In this particular study, Random Forest is brought into play to grapple with intricate data patterns and gauge the significance of feature attributes. A detailed dissection of Random Forest's ensemble nature, the intricacies of hyperparameter tuning for optimization, and the import of feature importance scores in fathoming the driving forces underpinning PayPal stock price prognostications is conducted [8].

4.4. Long Short-Term Memory (LSTM)

A deep learning model that has been meticulously developed for the study of sequential data is called Long Short-Term Memory (LSTM), rendering it ideally suited for the domain of time-series analysis. LSTM is enlisted to decipher time-dependent patterns and enduring dependencies residing within PayPal stock price data. An elaborate exposition on the architectural nuances inherent in LSTM networks, the pivotal role played by sequence length and batch size during training, and the strategic deployment of dropout layers to preempt overfitting is expounded upon [9]. Furthermore, a meticulous evaluation of LSTM's efficacy in capturing temporal dynamics and its aptitude in predicting future stock prices is undertaken.

4.5. Implementation

Each of these methodologies will undergo meticulous implementation to provide a comprehensive assessment of their performance in predicting PayPal stock prices. The insights gained from these empirical trials will serve as the guiding compass for drawing ultimate conclusions and recommendations in this study. The "training-test split" step is the first critical step in the process, where the data is carefully separated into two primary components: the test set and the training set. This partitioning plays a critical role, enabling the model to learn from historical data and evaluate its performance. The training set, comprising 80% of the data, serves as the foundation for model learning, allowing it to glean patterns from historical data to make predictions. Meanwhile, the test set, constituting the remaining 20% of the data, serves as the litmus test to assess the model's ability to generalize beyond its training data. This ability to generalize is pivotal for effective stock price prediction. The "feature selection" stage follows, where thoughtful consideration goes into choosing the input features. In this phase, historical prices of the preceding ten days are selected as features [10]. This decision is grounded in the assumption that historical price trends hold essential information about future price movements. Leveraging these historical trends equips the model with the potential to uncover patterns that influence forthcoming stock prices [2]. With the features prepared, the "model construction" phase commences. In this stage, four distinct machine learning models are employed: linear regression, support vector machines (SVM), random forests, and long short-term memory (LSTM) networks. These models are diligently trained on a carefully crafted training dataset, where the model absorbs the nuances of past price data. Their task is to unveil hidden patterns within historical information and use this knowledge to predict future closing prices. Moving on to the "performance evaluation" stage, two key metrics are utilized: mean squared error (MSE) and the coefficient of determination (R-squared, R^2). A lower MSE signifies higher accuracy, while R^2 assesses the model's fit with historical data. A value closer to 1 indicates a more harmonious alignment between the model and the data [11]. Finally, in the "model comparison" phase, comparative analysis is performed using the MSE and R^2 values obtained from each model's performance. This analysis helps identify which model excels in the domain of PayPal stock price prediction, making certain that the best model is chosen for this particular application.

5. Results

The final evaluation of the four models, based on the Mean Squared Error (MSE) and the Coefficient of Determination (R^2) values, reveals a clear trend of diminishing model performance. The MSE increases progressively, signifying a growing discrepancy between predicted and actual values, while R^2 gradually decreases, indicating a diminishing fit between the models and the historical data. In light of these findings, it can be inferred that the models exhibit varying degrees of predictive accuracy. The model with the lowest MSE and the highest R^2 value demonstrates the closest alignment with historical data and offers the most accurate predictions. Conversely, the model with the highest MSE and the lowest R^2 value indicates less precise predictions and a weaker fit with the historical dataset. The choice of the most suitable model for PayPal stock price prediction will depend on the specific trade-off between model complexity and predictive accuracy, taking into consideration the importance of minimizing prediction errors and maximizing the explanatory power of the model. So, it is evident that linear regression yields the best performance. Additionally, the following Table 1, which lists the MSE and R^2 values, along with the Figure 2, which provides a visual representation of the models' performance. These visuals will further aid in drawing comprehensive conclusions regarding the choice of the most effective model for PayPal stock price prediction.

Table 1. MSE and R^2 values

Name	MSE	R^2
Linear Regression	65.762	0.655
SVM	68.104	0.643
Random Forest	69.812	0.634
LSTM	140.437	0.245

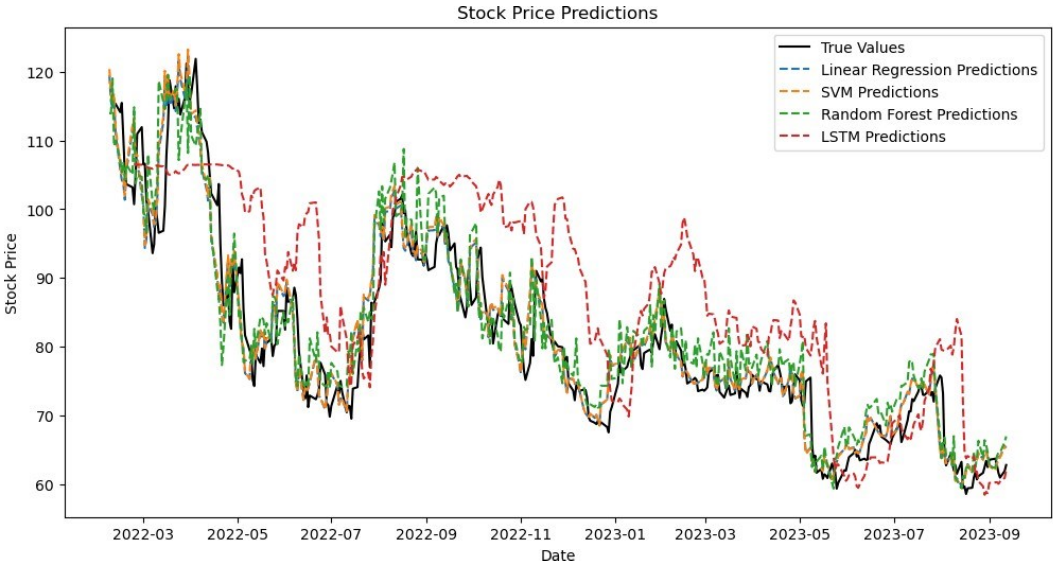

Figure 2. Stock Price Predictions

6. Conclusions

This research paper conducts an extensive and thorough investigation of four distinct machine learning models: Linear Regression, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM). Using Python as the main tool, the effectiveness of these models in anticipating trends in the financial markets is evaluated in detail, and they are employed to forecast the price of PayPal's stock. The linear model possesses the highest R-squared (R2) and the lowest mean squared error (MSE) among the four models that were selected, suggesting that it offers the most accurate prediction of PayPal's stock price. However, it's essential to acknowledge some limitations in this study. Data quality and quantity can significantly impact the models' performance, and limited access to additional relevant data sources or the quality of historical data may have influenced the results. The success of machine learning models depends on selecting appropriate hyperparameters, and further optimization of hyperparameters for each model could potentially enhance their performance. Unpredictable factors that impact stock prices include sentiment in the market and world events. The models' inability to account for these external factors may limit their predictive capabilities. Additionally, potential overfitting issues should be further explored to ensure the model's robustness under different market conditions. To sum up, this study offers insightful information about utilizing machine learning models to predict stock prices. While Linear Regression stands out as a strong performer, addressing the identified limitations and exploring more sophisticated modeling techniques may improve the precision of stock price prediction for PayPal and similar financial assets.

References

[1] Gangwar, A., Kumar, A., & Bijpuria, E. (2021). Stock Price Prediction using Machine Learning. 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 189-193.

[2] Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. Procedia computer science, 167, 599-606.

[3] Reddy, V. K. S. (2018). Stock market prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 5(10), 1033-1035.

[4] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.

[5] Panwar, B., Dhuriya, G., Johri, P., Yadav, S. S., & Gaur, N. (2021, March). Stock market prediction using linear regression and SVM. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 629-631). IEEE.

[6] Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017, April). Stock market predication using a linear regression. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 2, pp. 510-513). IEEE.

[7] Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. The Journal of finance and data science, 4(3), 183-201.

[8] Illa, P. K., Parvathala, B., & Sharma, A. K. (2022). Stock price prediction methodology using random forest algorithm and support vector machine. Materials Today: Proceedings, 56, 1776-1782.

[9] Sunny, M. A. I., Maswood, M. M. S., & Alharbi, A. G. (2020, October). Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. In 2020 2nd novel intelligent and leading emerging sciences conference (NILES) (pp. 87-92). IEEE.

[10] Mehtab, S., Sen, J., & Dutta, A. (2021). Stock price prediction using machine learning and LSTM-based deep learning models. In Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2 (pp. 88-106). Springer Singapore.

[11] Panwai, S. (2021). Artificial neural network stock price prediction model under the influence of big data. Review of Integrative Business and Economics Research, 10, 33-58.