Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241571

Predict Amazon stock by SVM and Random Forest

Jiawei Li

Department of Social Science, University of Southampton, Southampton, United Kingdom

jl11g22@soton.ac.uk

Abstract. The inherent uncertainties of market dynamics, such as economic data, geopolitics, and natural calamities, make stock market prediction extremely difficult. One increasingly effective method for handling this complexity is machine learning. Using data from the world's largest e-commerce and technology company, Amazon, this study concentrated on supervised machine learning models for stock market prediction. The most successful model was Support Vector Machine (SVM), which achieved an amazing prediction accuracy of 89.11%. Furthermore, Principal Component Analysis (PCA) significantly improved Random Forest's accuracy, enhancing it from 75.25% to 87.13%. In addition, the results show that the SVM outperforms the random forest no matter the PCA is considered. These results underscore SVM's importance in stock price prediction and PCA's value in enhancing Random Forest's performance. This research provides valuable insights into machine learning's role in financial forecasting, empowering investors and decision-makers to make informed choices in the ever-evolving stock market landscape.

Keywords: Amazon, SVM, Random Forest.

1. Introduction

Forecasting stock market movements presents a formidable challenge in the realm of financial time-series forecasting. This complexity arises from the inherent uncertainties surrounding market dynamics. Market dynamics are fraught with uncertainties, and various factors can trigger volatility and unpredictability. For example, economic data releases, such as employment or inflation rates, can sway market sentiment significantly when they deviate from expectations. Also, geopolitical events, like trade tensions or political instability, introduce uncertainty by affecting global trade and economic stability. Lastly, natural disasters, such as hurricanes or pandemics, disrupt supply chains, damage infrastructure, and create uncertainty about long-term economic repercussions. Consequently, achieving a high degree of accuracy in predicting the trajectory of asset prices becomes a formidable and often elusive goal, necessitating the employment of advanced analytical tools, constant vigilance, and a deep understanding of the ever-evolving landscape of global finance [1]. Academic research has unveiled a critical facet of market dynamics – their intrinsic non-randomness characterized by intricate, non-linear, and dynamic patterns. This revelation presents a formidable challenge to the traditional predictive models that rely on linear assumptions. Machine learning has increasingly become the focus of research because of its ability to capture these features.

DOI: 10.54254/2755-2721/46/20241571

Each day, countless individuals worldwide engage in stock market investments, highlighting the escalating demand for reliable machine learning-based stock price prediction models. These models play a crucial role as practical tools for investors, corporate leaders, and decision-makers, empowering them to make well-informed and impactful investment decisions. This study conducts a thorough evaluation of works that concentrate on the use of supervised machine learning models in the field of stock market prediction. The implementation of several supervised machine learning algorithms to improve the accuracy of stock market forecasts is the main emphasis of this work. Support Vector Machine (SVM) is the most often used technology for stock price prediction among the several methods that were studied. SVM's widespread adoption can be attributed to its notable performance and exceptional accuracy in capturing intricate market dynamics. Additionally, the paper highlights the promising results yielded by the Random Forest technique. This approach, known for its ability to handle complex datasets and mitigate overfitting, has demonstrated its potential to contribute significantly to the realm of stock market prediction [2]. Founded in 1994, Amazon is a multinational technology and e-commerce corporation. It has grown to be among the biggest and most powerful corporations in the world, playing a big role in the technological and international e-commerce industries. Amazon's vast and diverse business has generated extensive financial and operational data, making it an ideal source of stock analysis. Therefore, this study collected data from Amazon and used two models, SVM and Random Forest, to predict stock prices in order to assist investors in related fields.

2. Data

The dataset covers the period from January 1, 2021, to January 1, 2023, and it includes the daily stock market data for Amazon, one of the world's leading technology and ecommerce conglomerates [3]. This dataset serves as a valuable resource, providing a holistic perspective on Amazon's stock performance throughout this duration. It offers a rich and diverse set of data that is essential for conducting a wide range of financial analyses and supporting decision-making processes in the realm of investments and finance. This dataset (Table 1) encompasses fundamental components of stock market information, comprising daily low and high price points, opening and closing prices, precise date timestamps, trading volumes, and pertinent details regarding dividends and stock splits.

The difference between the highest and lowest prices for the year 2021–2022 is around 100. This suggests that there may be some trading possibilities or hazards as this stock has had significant price changes over this time. There is some range of movement in the gap between the opening and closing prices, although it is a rather wide range. There is around a 100-dollar difference between the highest and lowest prices. Standard deviation (Std.dev) is a measure of stock price volatility. It measures the dispersion of price data. The standard deviation of this stock is around 73. The standard deviation is relatively high, indicating that its price fluctuates greatly. The stock's average price during this period remained around 135, indicating that the stock's average price was roughly around this level during the time frame of the table.

Table 1. Descriptive statistics of Amazon

	Open	Close	High	Low
Max	187.199997	186.570496	188.654007	184.839493
Mix	82.800003	81.82	83.480003	81.690002
Mean	135	133.873753	135.672253	133.264748
Std.dev	73.8219437	73.6151228	73.8109871	72.9377046

The inclusion of date timestamps in the dataset facilitates in-depth time-series analysis, enabling the identification of trends and patterns in Amazon's stock performance over time (Figure 1).

DOI: 10.54254/2755-2721/46/20241571

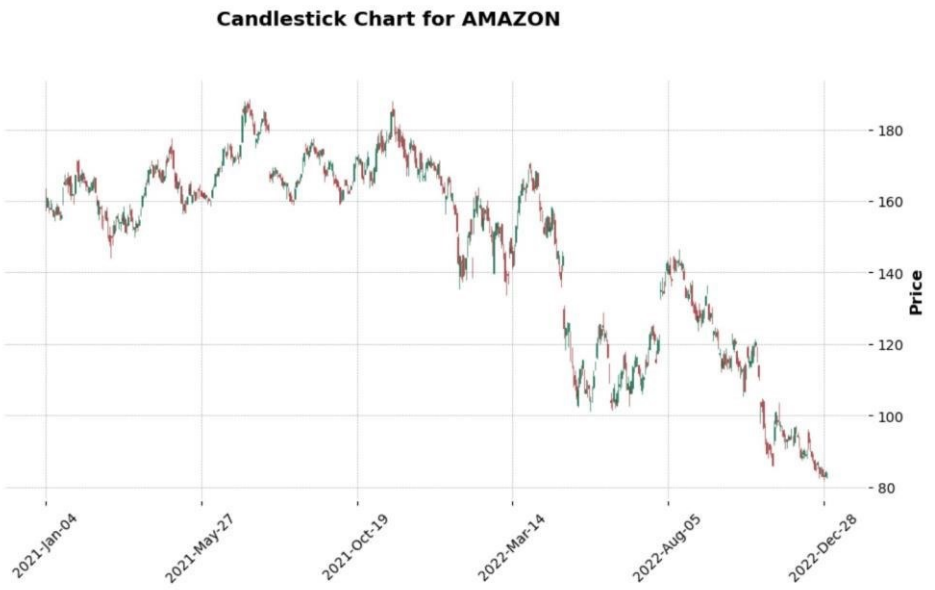

Figure 1. Amazon's stock performance

3. Preprocessing

First, this paper drops the NaN (Not-a-Number) values from the dataset since missing values can lead to incorrect calculations, biased results, or errors in machine learning algorithms. By removing them, this paper ensure that the data is complete and reliable for analysis.

3.1. Normalization

Normalization is a vital preprocessing step in machine learning, standardizing input feature scales and distributions for equitable model training. Its key benefit lies in balancing feature importance, preventing large numeric ranges from unfairly dominating model learning and causing bias. Normalization fosters proportional contribution among features. Moreover, it expedites convergence, especially in gradient-based optimization, saving computational resources and boosting model performance. Additionally, it enhances model interpretability by rendering feature importance scores and coefficients more intelligible [4].

In the paper, this paper employs StandardScaler as a vital preprocessing technique within the machine learning framework. In order to guarantee that the feature values in the dataset have a mean of 0 and a standard deviation of 1, this technique is essential. This conversion turns out to be very helpful, particularly when working with machine learning algorithms that are sensitive to the size of the input features. Notably, Support Vector Machines (SVM) and Principal Component Analysis (PCA) will be used in the research that follow in this investigation. By eliminating scale-related biases and fostering fairness across feature contributions, StandardScaler guarantees that these algorithms run as efficiently as possible, thereby improving the robustness and interpretability of the findings.

3.2. Labelling

Since the primary objective is to investigate the directional trends of stock prices, specifically whether they will rise or fall, it is imperative to label the close values accordingly. To achieve this, this paper adopts a straightforward labeling approach: when the current close value surpasses the previous close value, this study assigns it a label of 1, signifying an upward trend. Conversely, if the current close is lower than the preceding close, this study label it as -1, denoting a downward trend. This labeling scheme provides a clear directional sign for each data point, facilitating the subsequent analysis of stock price movements and aiding in the development of predictive models to discern these trends. By

DOI: 10.54254/2755-2721/46/20241571

Figure 2, it becomes evident that the labeled dataset exhibits a well-balanced distribution, with roughly an equal number of upward and downward trends. This balance in the data is highly beneficial as it contributes to a more robust evaluation process. With a near even split between positive and negative labels, the evaluation of machine learning models and predictive algorithms becomes more reliable. It helps prevent bias towards any particular class and ensures that the model's performance metrics, such as accuracy and precision, are not skewed.

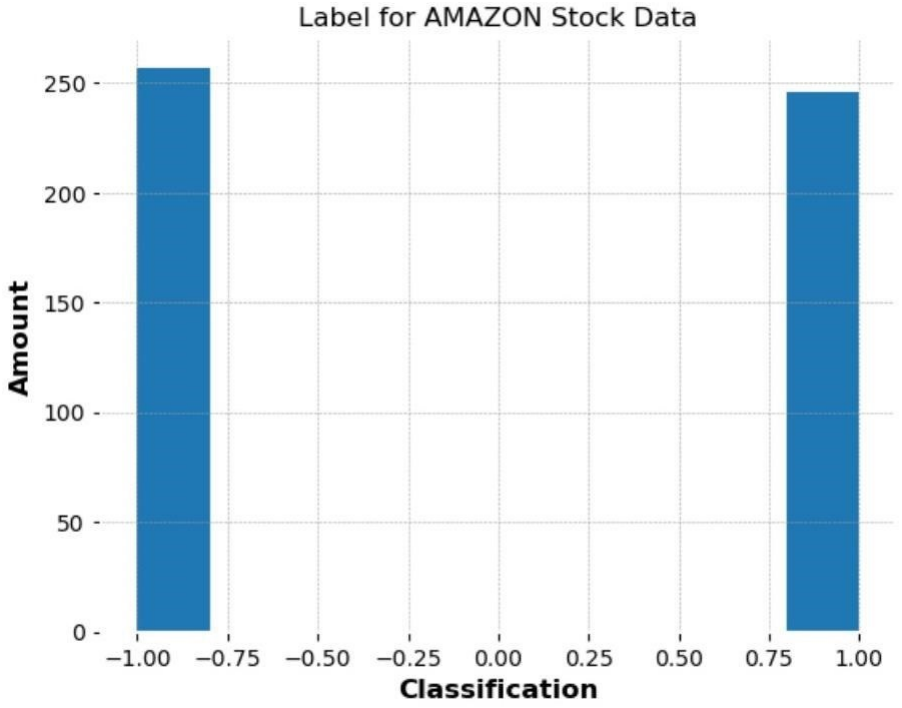

Figure 2. Labeled dataset

4. Feature Engineering

Feature engineering is the process of creating new, meaningful features or modifying pre-existing ones from the raw data to improve the performance of machine learning models. It entails picking, altering, and adding elements to make them more instructive and applicable to the particular job at hand. Effective feature engineering can enhance a model's ability to capture patterns, reduce overfitting, and ultimately lead to better predictive performance.

In feature engineering, PCA is a dimensionality reduction technique that helps minimize the number of features while preserving as much of the original data as feasible. Principal Component Analysis (PCA) is a key approach used in this work. PCA plays a crucial role in feature engineering, enabling us to reduce the dimensionality of the dataset while retaining essential information. This study reduced the number of features and lessened the effects of high dimensionality by converting the original features into a set of uncorrelated principal components. This allows us to streamline the analyses, improve model efficiency, and enhance the ability to discern meaningful patterns and trends in the data [5]. PCA's utility in this study underscores its significance as a valuable tool for data preprocessing and dimensionality reduction in the analytical processes.

5. Methodology

Support Vector Machines (SVM) and Random Forest are two different machine learning techniques that were used in this paper's research to predict stock prices. In the field of machine learning, these are both potent and extensively utilized methods, each with special advantages and skills. Furthermore,

DOI: 10.54254/2755-2721/46/20241571

this paper delves into a comparative analysis of the outcomes obtained with and without the application of Principal Component Analysis (PCA). This comparison is part of the study's effort to clarify how dimensionality reduction affects this machine learning models' capacity for prediction. This exploration serves as a valuable addition to the study, providing insights into the potential benefits of PCA in the context of stock price prediction and enabling a more comprehensive evaluation of the methodologies [5].

5.1. Support Vector Machines (SVM)

Support Vector Machines (SVM) are a powerful machine learning method that are well-known for their effectiveness in regression and classification applications. SVMs are unique in that they perform well in situations when the data cannot be separated linearly [6]. They achieve this by ingeniously mapping the data into higher-dimensional spaces, thereby unveiling optimal decision boundaries. SVMs come with several notable advantages. They are exceptionally effective in high-dimensional spaces, rendering them versatile tools for various data domains. Their robustness against overfitting, when appropriately regularized, is a significant asset. Moreover, SVMs can gracefully handle complex, non-linear relationships, thanks to the flexibility offered by kernel functions.

Additionally, they offer feature importance scores, simplifying feature selection [7]. Nonetheless, SVMs do come with certain drawbacks. They can be computationally intensive, particularly with large datasets. The choice of kernel and hyperparameters can substantially impact performance, demanding meticulous tuning. Therefore, this study will employ grid search as a critical step in the methodology. By sweeping through a predefined grid of hyperparameters and evaluating the model's performance using cross-validation, this study aims to pinpoint the hyperparameter settings that yield the best results.

In a binary classification scenario, the linear SVM formula is succinctly expressed as follows.

$$f(x)= \text{sign}(w*x+b) \tag{1}$$

Where $f(x)$ signifies the predicted class label, x denotes the input data, and w and b are the weight vector and bias term, respectively. The overarching objective of training a linear SVM is to uncover the optimal w and b that maximize the margin while minimizing classification errors, making SVMs an invaluable asset in machine learning and data analysis.

5.2. Random Forest

Random Forest is a potent ensemble learning method in machine learning that may be applied to both regression and classification issues. It combines the strengths of many decision trees to produce forecasts that are incredibly accurate. A Random Forest is a collection of decision trees constructed mathematically, with each tree having a random sample of data and a random subset of attributes. The aggregate of the predictions from each individual tree is the final forecast, which is usually determined by average (for regression) or by a majority vote (for classification) [8]. Random Forests offer remarkable advantages including exceptional predictive accuracy, reduced risk of overfitting through aggregation, adeptness with large datasets and feature-rich data, feature importance insights, and resilience to noisy data. However, they come with computational intensity, interpretability challenges due to ensemble nature, and potential performance gaps in specialized domains like image recognition where deep learning shines [9].

6. Experiment Results

This study uses grid research for hyperparameter optimization. In the SVM model optimization, this study traverses a parameter grid that encompasses 'C' and 'kernel' settings. 'C' represents the regularization parameter, with values in [0.1, 1,3,5, 7,10]. It controls how much margin is maximized and how much is lost due to misclassification. In addition, 'kernel' provides options for 'linear,' 'rbf' (Radial Basis Function), and 'poly' (Polynomial) kernels, each of which affects how well the model can represent intricate relationships in the data. For random forest, this study carefully navigates the

DOI: 10.54254/2755-2721/46/20241571

'n_estimators' parameter, considering three different quantities of trees: 50, 100, and 150. This exploration strikes a delicate balance between achieving robust and accurate predictions while managing computational demands effectively. This study also ventures into the depths of decision trees through the 'max_depth' parameter. By choosing from options like 'None,' 10, 20, and 30, this study control how deep these trees can grow. Allowing unrestricted depth can capture intricate data patterns but risks overfitting, where the model learns the training data too intimately, potentially compromising its ability to generalize to unseen data.

This study has revealed the ideal settings for the machine learning models through thorough grid search. Whether or not this study uses Principal Component Analysis (PCA), the Support Vector Machine (SVM)'s optimal configuration is 'C=5' with a 'linear' kernel [10]. In the realm of Random Forest, the parameter choices differ based on the presence of PCA. Without PCA, this study has determined that the best settings are 'max_depth=20' and 'n_estimators=100,' optimizing model performance [10]. However, when PCA is in play, a slightly different combination emerges as the champion— 'max_depth=10' paired with 'n_estimators=150.' These meticulously selected parameters ensure that the models are finely tuned to deliver accurate predictions in the world of stock price direction forecasting.

The final results (Table 2) of the analysis demonstrate the performance of two machine learning models in stock price prediction. Support Vector Machine (SVM) excels with an impressive accuracy of 89.11%. On the other hand, Random Forest also demonstrates strong predictive power, achieving an accuracy of 75.25% without PCA.

Remarkably, when PCA techniques are applied to the dataset, the accuracy of Random Forest significantly improves to 87.13%. This observation suggests that PCA plays a valuable role in enhancing the predictive accuracy of Random Forest when applied to the specific dataset under consideration. It underscores the importance of dimensionality reduction techniques in optimizing the performance of machine learning models in financial prediction tasks.

Table 2. Final results

ACCURACY (%)	SVM	Random Forest
With PCA	89.11%	87.13%
Without PCA	89.11%	75.25%

7. Conclusion

The main goal of this extensive paper was to forecast Amazon stock price movements using two potent machine learning models: Random Forest and Support Vector Machine (SVM). This study also aimed to investigate the effects of Principal Component Analysis (PCA). The results shows that SVM emerged as the star performer, achieving remarkable accuracy with a prediction rate of 89.11%. Equally intriguing is the role of PCA in enhancing the predictive power of random forest. Without PCA, Random Forest achieved a commendable accuracy of 75.25%. However, with the integration of PCA techniques, this paper witnessed a substantial increase in accuracy, elevating it to an impressive 87.13%. This underscores the valuable contribution of dimensionality reduction in optimizing Random Forest's performance on the specific dataset. The study concludes by stressing the importance of PCA as a tool for enhancing Random Forest's accuracy as well as the supremacy of SVM in stock price prediction. These findings offer valuable insights into the dynamics of machine learning models in financial forecasting, empowering us to make more informed decisions in the everchanging landscape of stock markets.

References

[1] Patel, R., Choudhary, V., Saxena, D., & Singh, A. K. (2021, June). Review of stock prediction using machine learning techniques. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 840-846). IEEE.

DOI: 10.54254/2755-2721/46/20241571

[2] Lawal, Z. K., Yassin, H., & Zakari, R. Y. (2020, December). Stock market prediction using supervised machine learning techniques: An overview. In 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). IEEE.
[3] Yahoo Finance. Available at: https://finance.yahoo.com (Accessed on 05 November 2023).
[4] Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, 197, 116659.
[5] Zhong, X., & Enke, D. (2019). Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Financial Innovation, 5(1), 1-20.
[6] Kumar, L., Pandey, A., Srivastava, S., & Darbari, M. (2011). A hybrid machine learning system for stock market forecasting. Journal of International Technology and Information Management, 20(1), 3.
[7] Yu, H., Chen, R., & Zhang, G. (2014). A SVM stock selection model within PCA. Procedia computer science, 31, 406-412.
[8] Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.
[9] Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. Int. J. Recent Technol. Eng, 8(1), 1224-1228.
[10] Wang, H., & Hu, D. (2005, October). Comparison of SVM and LS-SVM for regression. In 2005 International conference on neural networks and brain (Vol. 1, pp. 279-283). IEEE.