Tesla stock prediction with SVM, decision tree and random forest

Jierui Wang *

University of Toronto, 27 King's College Cir, Toronto, ON M5S 1A1, Canada

* Corresponding Author Email: jrui.wang@mail.utoronto.ca

Abstract. Stock price prediction is a critical task in financial markets, as it helps traders and investors make informed decisions. This paper compares 3 popular machine learning techniques, which are Support Vector Machines (SVM), Decision Trees, and Random Forests, in predicting stock movements. This paper will evaluate and compare the performance of these models for predicting Tesla's stock price movements by using historical stock data. The 20-day Simple Moving Average (SMA20) is used as a key technical indicator which is combined with the basic stock features such as Open, High, Low, and Volume. The results demonstrate that the SVM model achieved the highest accuracy and recall even when only using the SMA20. However, the SVM model provided a better balance between precision (51.5078%) and recall (93.6404%), achieving the highest F1 score which is 66.4591%. The three models have different focuses and advantages, which also lead to the difference between the test values. In this prediction, the SVM model performs the best, compared to the other two models. With the results of the test, the SVM model is perfect for recognizing price increases in a short time with a balance between precision and recall.

Keywords: stock prediction, machine learning, support vector machines.

1. Introduction

Stock price prediction plays an important role in financial markets, assisting investors in making decisions. However, predicting stock prices is challenging due to its volatility, non-linear patterns, and numerous external factors influencing market prices. In recent years, machine learning models have been applied to forecast stock price changes.

Lin et al. (2012) developed a Support Vector Machine (SVM)--based model with financial indexes and a quasi-linear SVM mode feature selection, which improved stock market trend prediction accuracy. Sai Reddy (2018) used SVM with a Radial Basis Function (RBF) kernel to analyze and predict stock prices with machine learning models using historical stock data which compares the performance of SVM and emphasizes the significance of machine learning in reducing uncertainty in the stock market predictions and noting that SVM effectively captured non-linear data patterns. Mahesh, B. (2019) mentioned that SVM is emphasized for its versatility in handling both linear and non-linear data, especially in high-dimensional spaces through kernel methods. The reason for its high ability to handle both linear and non-linear is also mentioned in the paper of Mahesh, B. which is an SVM model that works by identifying a hyperplane that can separate different classes in a dataset maximally.

This paper will focus on the 20-day Simple Moving Average (SMA20), a short-term moving average that is often used by traders to capture the trend of recent prices. In addition, it compares the predictive performance of three machine learning models: SVM, Decision Tree (DT), and Random Forest (RF) which are learned during the program, using Tesla stock data. Aiming to improve forecasting accuracy and find the best model for predicting stocks the next day.

2. Method

2.1. Dataset

This paper uses Tesla stock data from 2010 to 2023 from Kaggle, with features like daily Open, Close, High, Low, and Volume. The 20-day SMA20 and volatility are derived to improve predictive accuracy. Three machine learning models, SVM, Decision Tree, and RF, are used to predict stock price movements, with 80% of the data from the set used for training and 20% for testing.

2.2. Model

Somvanshi, & Chavan (2014) explain these two models in detail. SVM is a supervised learning model used for classification and regression. It aims to find the optimal hyperplane or set of hyperplanes that best separates the data into different classes. When meets the cases that the data is not linearly separable, this model will use kernel functions, like an RBF or polynomial kernels to transform the input space into a higher dimension feature space where data can be linearly separable. This model is highly effective in high-dimensional spaces, where the number of features is more complex. It is also robust against overfitting, especially when the data is clean and well-labeled. However, it requires careful tuning of hyperparameters, such as kernel type and regularization strength, which can be computationally expensive, especially for large data sets.

The DT is a tree-like model used for classification or regression tasks. It splits the dataset into subsets based on feature values, creating a structure of decisions and their possible consequences. Each split is determined by the attribute that provides the most meaningful data separation, using metrics such as information gain (ID3), gain ratio (C4.5), or Gini Index (CART). The result is a tree structure where decisions are made at the inner nodes and final predictions are made at the leaf nodes. DTs are easy to understand and interpret, which makes them particularly difficult in areas where interpretability is critical, such as healthcare and diagnostics. They are also versatile, able to handle both categorical and numerical data, and can withstand imported or noisy values. However, they have significant limitations. Without pruning, DT can become overtrained on data, making them less efficient on unseen data. In addition, small changes in the data set may cause significant changes in the tree structure, thus affecting consistency.

RF is an ensemble learning method that combines multiple DTs to improve predictive performance (Liu, Wang, and Zhang. 2012). This model builds many trees instead of relying on a single DT during training and combines their outputs. Each tree is trained on a random subset of features is considered. This randomness helps to reduce overfitting and can make the model more robust. RF can handle both classification and regression problems. The classification task works by having each tree "vote" for the most likely category, with the most votes becoming the final prediction. For the regression task, it takes the average of the predictions of all the trees. This integrated approach helps balance the model's tendency to overfit a single tree while still capturing the complexity of the data. However, this model can be computationally heavy, especially when working with large data sets or large numbers of trees. In addition, while RFs are excellent at accuracy, they sacrifice interpretability compared to individual DT. Understanding the individual contributions of features can be challenging because predictions are averaged across many trees.

Accuracy, recall, precision, and f1 scores are used to determine whether the models' predictions are accurate. Accuracy is measured by how often the model correctly predicts the target label across the data sets. Precision is measured by how many of the model's positive predictions were correct. Recall measures how many real positives were correctly identified by the model. F1 Score is the harmonic mean of Precision and Recall, balancing the 2 metrics.

3. Results

The models were evaluated by using accuracy, precision, recall, and f1 metrics. As shown below in Table 1: Results in the three models after tests, SVM has the highest Accuracy which is 51.13%, followed by RF (51.02%), and DT has the lowest Accuracy (49.43%). RF has the highest precision among the 3 models which is about 52.47%, and SVM is the second highest (51.51%), and the worst one is the DT which is about 51.06%, and the values of the 3 models of precision are close. Comparing the values of recall, SVM has the highest value which is around 93.64%. DT and RF models have significantly lower values of recall, their values are 52.85% and 55.92% respectively. For the F1

Score, the SVM model with 66.46% of it has the highest value among the 3 models. RF is about 54.14% and DT is the lowest, with 51.94% in F1 Score.

The results, present the performance of 3 models in predicting Tesla's stock, SVM has the best performance in accuracy, recall, and f1 score, and RF does best in precision. The DT performs worst in this test with the lowest value in accuracy, precision, recall, and f1 score.

Table 1. Results in the three models after tests

	Accuracy	Precision	Recall	F1 Score
SVM	0.5113	0.5151	0.9364	0.6646
DT	0.4943	0.5106	0.5285	0.5194
RF	0.5102	0.5247	0.5592	0.5414

4. Discussion

The results show that the SVM model is the most sensitive and effective, but RF also provides more consistent predictions with fewer errors. For further improvement in the future, using longer-period moving averages could enhance each model's performance more.

For SVM model improvement, Khan et al. (2023) propose an improved version called a genetic algorithm support vector machine (GASVM) combining SVM with genetic algorithms to fine-tune parameters and select better features. In their research, the method significantly improved the prediction accuracy of 10-day price movements to 93.7%, providing better results than traditional SVM. Similarly, later in 2024, Htun et al. Using SVM to predict the relative return of SandP 500 stocks. While their works show moderate results, combining SVM with other methods, such as rough set models, has been shown to improve its performance.

For RF model improvement, in the study of Khan and others (2023), a new method was used to improve RF performance. In their research, they improved the accuracy rate of the RF by optimizing the input characteristics and focusing on the stock price 15 minutes before the opening. As a result, the accuracy rate of the RF increased from 85.51% to 91.27%. This made it the best-performing model in their study.

For DT model improvement, Sharaf et al. (2021) point out that DT often overfits data and struggles to adapt to noisy stock market conditions. However, they perform better when they are used as part of an ensemble model such as RFs, which combine multiple DTs to improve results. Khan et al. (2023) Using a DT as a comparison baseline show that they perform worse than models such as RF and mixed methods. In addition, DT can still be used as building blocks for higher-level models.

Kara, Boyacioglu, and Baykan (2011) compared the accuracy of stock prediction of Istanbul, Artificial Neural Networks (ANN) and SVM used in their research. ANN model achieved an average prediction accuracy of 75.74% which is slightly higher than the SVM model (71.52%). They found although ANN and SVM are effective tools in predicting stock price direction, the accuracy and consistency of ANN outperformed SVM across the years. In addition, they thought the prediction accuracy could be improved by exploring further additional features, such as macroeconomic indicators, or refining the model parameters Later, Patel, Shah, and Thakkar, (2015) compared ANN, SVM, RF, and naive Bayes in their study. It explores two methods of inputting data, representing technical parameters in terms of continuous values and trended deterministic data. This paper concludes that RF performs better in the first approach, but all models improve in trend deterministic data.

Also, there are still some other machine learning models that can be used in stock prediction, Li and Pan (2021) used a deep learning ensemble model with Long Short-Term Memory (LSTM) and Grated recurrent Units (GRU) for stock prediction, demonstrating that the ensemble model outperforms standalone models in predicting stock trends. Demirel et al. (2021) compared Multilayer Perception (MLP), SVM, and LSTM models in stock price prediction for the Istanbul Stock Exchange, finding that LSTM outperformed SVM. The research of M., Patel, H., and Varma, S. (2017) also found that LSTM can provide more accurate predictions and explained how this model works, their work shows that LSTM uses advanced neural network techniques to understand sequential data and optimize predictions which can provide meaningful insights for the investors with an accurate prediction of future stock prices.

To improve the accuracy of machine learning, using a hybrid model can achieve this purpose. Lu et al. (2020) used the CNN-LSTM model, which is a hybrid model and daily stock prices of the Shanghai Composite Index from July 1st, 1991, to August 31st, 2020, to predict stock prices and highlight the potential of combining deep learning architectures to solve financial forecasting challenges. This hybrid model effectively captures the timing and feature-based complexity of stock price data and achieves high forecasting accuracy. Compared to other methods, such as MLP, CNN, RNN, LSTM, and CNN-RNN, this model achieved the lowest error rates which outperformed all other models in this test and the model also shows an excellent fit between predicted and actual values. As a result, this hybrid model provides a powerful tool for financial time series forecasting and practical implications for investors seeking to maximize returns. Though there are still some limitations in this combined deep learning machine, it can only consider the data that is price-related.

5. Conclusion

This study demonstrates that the 20-day moving average, when used with machine learning models, provides reasonable predictive power in capturing short-term stock price trends.

The SVM model achieved the highest accuracy (51.13%), recall (93.64%), and highest F1 score (66.46%) making it the most sensitive to identifying price movement and achieving a balance between precision and recall. And it is the most effective model overall. RF offered a good balance between precision and recall, leading to the highest F1 Score (54.14%), making it a good model for predicting Tesla stock prices and suitable for scenarios in which reducing false positives is important. DT performed the worst among the 3 models with the lowest accuracy (49.43%) and F1 score (51.94%), which shows it has limitations in dealing with complex datasets in the SMA20 feature.

Overall, SVM is the most well-rounded model among the 3 models, it is perfect for identifying price increases in a short time while maintaining a balance between precision and recall. And RF provides a relatively accurate prediction with fewer false positives which makes it fit for the critical prediction precision. For DT, in this test, it struggles with overfitting and lacks generalizability.

Future research could explore combining SMA20 with additional technical indicators such as RSI or MACD to improve model accuracy and predictive performance further. In addition, using longer moving averages, like 50 days, or combining SMA20 with some fundamental data could improve the ability of the models to predict long-term trends, and this might lead to a more precise prediction in the short term. Also, to improve the accuracy of forecasting the stock price change can use the hybrid model which combines several deep machine learning techniques, as mentioned before, the CNN-LSTM model, adding CNN to LSTM improves feature extraction, reduces prediction errors, and enhances the model's prediction capability.

References

[1] Li, Y., Pan, Y. 2021. A novel ensemble deep learning model for stock prediction based on stock prices and news. *Complex and Intelligent Systems*, 7, 217 - 230.

[2] Patel, J., Shah, S., Thakkar, P., Kotecha, K. 2015. Predicting stock and stock price index movement using machine learning techniques. *Expert Systems with Applications*, 42 (1), 259 - 268.

[3] Sai Reddy, V. 2018. Stock market prediction using machine learning. *International Journal of Computer Science and Information Technologies*, 9 (1), 10 - 14.

[4] Lin, G., Guo, Y., Hu, Y. 2012. An SVM-based approach for stock market trend prediction. *Procedia Computer Science*, 24, 267 - 274.

[5] Demirel, U., Cam, H., Unlu, R. 2021. Predicting stock prices using machine learning methods and deep learning algorithms: The sample of the Istanbul Stock Exchange. *GU Journal of Science*, 34 (1), 63 - 82.

[6] Somvanshi, M., Chavan, P. 2014. A review of machine learning techniques using DT and support vector machine. *IEEE*, 1 - 5.

[7] Mahesh, B. 2019. Machine learning algorithms - A review. *International Journal of Science and Research (IJSR)*, 9 (1), 381 - 386.

[8] Patel, M., Varma, S. 2017. Predicting stock prices using LSTM. *International Journal of Science and Research (IJSR)*, 6 (4), 1754 - 1756.

[9] Kara, Y., Boyacioglu, M. A., Baykan, Ö. K. 2011. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, 38 (5), 5311 - 5319.

[10] Lu, W., Li, J., Li, Y., Sun, A., Wang, J. 2020. A CNN-LSTM-based model to forecast stock prices. *Complexity*, 2020, Article ID 6622927.

[11] Liu, Y., Wang, Y., Zhang, J. 2012. New machine learning algorithm: RF. In B. Liu, M. Ma, J. Chang (Eds.), *Information Computing and Applications*, Third International Conference, ICICA 2012 (pp. 246-252). Springer, Berlin, Heidelberg.

[12] Sharaf, M., Hemdan, E. E., El-Sayed, A., El-Bahnasawy, N. A. 2021. StockPred: A framework for stock price prediction. *Multimedia Tools and Applications*, 80 (12), 17923 - 17954.

[13] Khan, A. H., Shah, A., Ali, A., Shahid, R., Zahid, Z. U., Sharif, M. U., Jan, T., Zafar, M. H. 2023. A performance comparison of machine learning models for stock market prediction with novel investment strategy. *PLoS ONE*, 18 (9), e0286362.

[14] Htun, H. H., Biehl, M., Petkov, N. 2024. Forecasting relative returns for S&P 500 stocks using machine learning. *Financial Innovation*, 10, 118.