Comparison of Stock Price Prediction Models using Machine Learning

Felipe de Jesús Liévano
William Andrés Bayona
Sara Alejandra Gómez
Juan Martin Vasquez

May 30, 2024

Abstract

This study evaluates the efficacy of four different machine learning models in predicting stock market dynamics using the S&P 500 index. Specifically, the Random Forest (RF) and Support Vector Machine (SVM) models are utilized for predicting market movement, whereas the Hidden Markov Model (HMM) and Long Short-Term Memory (LSTM) focus on price predictions. Performance is assessed based on metrics such as mean absolute error, Sharpe ratio, profitability, and computing time. Notably, the HMM achieves the highest accuracy in price prediction, with a mean absolute error of 0.89% and an R^2 of 0.91. In terms of movement prediction, the RF model excels, recording an accuracy score of 0.5474, an AUC of 0.5404 on the ROC curve, and a Sharpe ratio of 2.1498. Both the LSTM and RF models return around 10%. SVM and RF are highlighted for their computational efficiency and consistent execution. These results provide evidence against the efficient market hypothesis by demonstrating that machine learning models can effectively approximate stock behavior and sometimes match or exceed market returns. However, these findings are preliminary, and further investigation with varied portfolios and different time frames is advised.

Keywords: Market Efficiency Hypothesis, Machine Learning Algorithms, Stock Behavior Forecast, Model Benchmarking.

1 Introduction

It has always been believed that predicting future stock prices is impossible. It is taken for granted that the market is efficient, and that the distribution of stock prices follows a random walk process. It is assumed that the market has already considered all available information and quantified it in the current stock price, and therefore it is impossible to make future predictions based on historical information. However, the existence of the Medallion Fund is one of the best challengers to the efficient market hypothesis, given its astonishing returns of an annualized 66% return (before commissions). Their approach to trading is based purely on a secret predictive mathematical model. "The returns are large enough to stretch that explanation to the limit. Whatever the source of the performance, Medallion is a Michelson–Morley-level challenge to the hypothesis of market efficiency" (Cornell, 2020).

The hypothesis of this project proposes that by using advanced machine learning models, such as decision trees, neural networks, and hidden Markov models, it is possible to predict the price movements of the S&P 500, challenging the validity of the efficient market hypothesis. This approach will be evaluated by comparing the predictive performance of these models using fit metrics such as mean absolute error (MAE) and Receiver Operating Characteristic (ROC) Curve Analysis, using historical data

from the S&P 500 to train and validate each model in a standardized manner using several benchmarks to compare the models. This approach is based on the research done by Kevin M. Harper, 2016, in his thesis *Challenging The Efficient Market Hypothesis With Dynamically Trained Artificial Neural Networks*. In his thesis, the author trains and tests a model very extensively, sadly having lackluster results. "Research indicates that statically trained models, while they may produce good returns for finite test periods, are subject to performance degradation over time" (Harper, 2016). In our research, we arrive at similar conclusions.

2 Objectives

2.1 General Objective

Evaluating the Efficient Market Hypothesis through the Comparison of Machine Learning Models to Analyze Predictability in Financial Market Prices.

2.2 Specific Objectives

1. Analyze different predictive models: Random Forest, Hidden Markov Model (HMM), Long Short-Term Memory (LSTM), and Support Vector Machine (SVM).

2. Simulate the behavior of the models on the S&P 500 index, over the same period of time.

3. Compare the performance of each model, identifying their strengths and weaknesses.

4. Evaluate the profitability of each model, using the same trading strategy (Backtracking).

3 Literature Research

This section reviews the literature on various predictive models used for stock price prediction. Each model's effectiveness, methodologies, and the specifics of its application in financial markets are detailed based on recent studies.

3.1 SVM (Movements)

The Support Vector Machine (SVM) model for stock price prediction is based on minimizing an upper bound of the generalization error, rather than merely reducing the training error. This is achieved by fitting a hyperplane in a higher-dimensional space that maximizes the distance to the nearest support vectors, thus reducing structural risk. SVM regression balances the minimization of the squared training error with a regularization term that penalizes model complexity, regulated by the C parameter. Additionally, slack variables and the epsilon parameter allow flexibility and tolerance to certain prediction errors, making SVM a robust and effective model for prediction in financial markets (Tripathy, 2019).

3.2 LSTM (Price)

The Long Short-Term Memory (LSTM) algorithm is effective in predicting stock prices due to its ability to handle long-term data sequences, essential for time series analysis. LSTM, a type of recurrent neural network, utilizes memory cells that regulate the flow of information through input, forget, and output gates, mitigating the problem of gradient vanishing. This allows learning from long-term dependencies, making LSTMs suitable for analyzing complex patterns in stock prices. Their ability to exploit temporal patterns in financial data makes them a powerful tool for investors and analysts in volatile markets (Phuoc et al., 2024).

3.3 Random Forest (Movement)

The Random Forest model for stock price prediction uses multiple decision trees to enhance the accuracy and robustness of predictions. This supervised learning approach classifies whether a stock price will increase or decrease relative to previous days' prices, and is part of ensemble learning methods that outperform individual algorithms. Decision trees identify the best threshold to divide the feature space, and randomization in tree construction reduces the sensitivity to overfitting. Random Forest handles high-dimensional data well and evaluates the impor-

tance of each feature, aiding analysts in understanding which variables influence the direction of stock prices, thus improving the accuracy of predictions for investment strategies and risk management in financial markets (Suryoday, B. et al., 2018).

3.4 HMM (Price)

Hidden Markov Models (HMM) are effective in predicting stock prices due to their ability to model stochastic and non-stationary processes. HMMs rely on hidden states that influence observations, such as stock prices, using state transition and emission probabilities. Transition probabilities describe the market state change from one day to another, while emission probabilities indicate the likelihood of observing a specific price change. This model captures how internal and external events affect prices, allowing handling of market dynamics and volatility. By adjusting parameters with historical data, HMMs can predict future price changes, providing a robust tool for analysts and traders in anticipating market movements (Verma, A. et al., 2021).

4 Methodology

In our project, we evaluated machine learning models that predict stock prices and market behaviors, using a comprehensive and consistent methodology. Some models, such as the Hidden Markov Model (HMM) and Long Short-Term Memory (LSTM), are designed to forecast actual prices, while others, like the Support Vector Machine (SVM) and Random Forest, focus on predicting the direction of price movements.

To ensure a fair and thorough comparison across these models, we have selected common metrics that evaluate both types of predictions:

1. **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in the predictions, providing a straightforward metric for error magnitude across all models.

2. **Coefficient of Determination ($\mathbf{R^2}$)**: Used only for the price forecasting models (HMM and LSTM), this metric quantifies how well variations in the actual prices are replicated by the models' predictions.

3. **Accuracy Score**: Assesses the overall accuracy in predicting the correct direction of price movement for all models.

4. **Receiver Operating Characteristic (ROC) Curve Analysis**: Evaluates the models' ability to discriminate between classes effectively, with the area under the curve (AUC) indicating overall discriminatory power.

5. **Sharpe Ratio**: Measures the risk-adjusted return, providing insights into the models' effectiveness under practical trading conditions.

6. **Performance Over 100 Days**: Reflects the models' predictive performance over the most recent market conditions.

7. **Profitability**: Calculates the hypothetical profitability based on the models' predictions against actual market movements.

8. **Computing Time**: Tracks the computational efficiency required for training and prediction.

9. **Execution Consistency**: Assesses the reliability of each model in operation, focusing on stability and frequency of errors.

The models were trained using a total of 3605 daily S&P 500 index close prices, starting from January 1, 2010, up to the 101st most recent day as of May 1, 2024. The evaluation period will utilize the data from the 100th to the most recent day to assess each model's accuracy and applicability in current market conditions.

Additionally, the *Back Testing* methodology is a way of analyzing the potential performance of a trading strategy by applying a set of real-life historical data. This process allows the evaluation of how a strategy would have performed in the past, thus providing a basis for predicting its possible success in the future. Each of the models uses specific input parameters, such as a 1% Stop Loss and a 3% Take Profit. These parameters are crucial for managing risk and

ensuring that potential losses are kept under control while maximizing profits.

The financial results and the behavior of each investment policy are determined by the models analyzing a period of 100 days. This analysis includes evaluating market fluctuations and how each strategy responds to these variations. Regarding the purchase of an asset, the strategy considers tolerance to price variation to avoid losses and ensure profits. If the asset's price exceeds a 3% growth or decreases by 1%, the asset is sold. This selling threshold ensures that profits are capitalized on and losses are minimized, thus maintaining the value of the equity updated and optimized.

5 Results

After running the models under standardized parameters, the following results were obtained.

Modelo	HMM	LSTM	SVM	Random Forest
MAE	0.89%	8.75%	-	-
R^2	0.9128	0.89	-	-
Accuracy Score	0.535353	0.519	0.516	0.5474
ROC	0.5218	0.511	0.4934	0.5404
Sharpe Ratio	0.68	1.427	1.47	2.1498
Rendimiento 100 dias	2.89	10.1153	8.1159	10.24
Rentabilidad	3%	10%	8%	10%
Eficiencia Computacional	2 min	40 min	1 Seg	1 Seg
Consistencia en Ejecucion	NO	SI	SI	SI

Figure 1: Benchmarking among Machine Learning Models.

Starting with the Mean Absolute Error (MAE), the HMM exhibits the highest accuracy with an MAE of only 0.89%, considerably lower than the 8.75% of the LSTM, indicating that the HMM has higher accuracy in predictions. In terms of the coefficient of determination (R^2), the HMM is higher at 0.91, demonstrating that almost 91% of the variability in stock prices is explained by this model, closely followed by the LSTM with 0.89. This can also be observed in the comparison between predicted prices and actual prices with respect to the closing price of the S&P over a period of 100 days (Figure 2 and Figure 3).

In terms of Accuracy Score and the area under the ROC curve, Random Forest outperforms the other models with scores of 0.5474 and 0.5404, respectively.

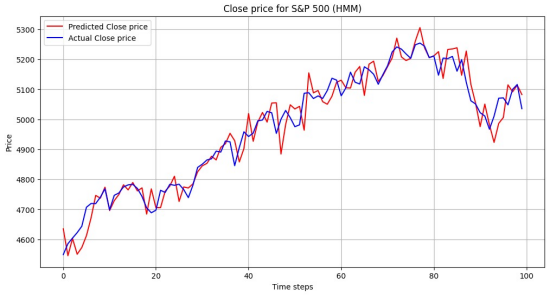

Figure 2: Comparison between price prediction (HMM) and the actual price of the S&P 500 closing price.

Figure 3: Comparison between price prediction (LSTM) and the actual price of the S&P 500 closing price.

This indicates a better ability of the Random Forest to correctly classify the movement of stock prices compared to the other models. The Sharpe Ratio, which measures risk-adjusted return, is highest for Random Forest (2.1498), suggesting that this model provides a superior return for every unit of risk assumed compared to LSTM (1.427) and SVM (1.47).

Furthermore, both Random Forest and LSTM demonstrate remarkable performance at 100 days, with very close values of 10.24 and 10.11, respectively, highlighting them in terms of potential gains over a fixed period. Regarding profitability, both LSTM and Random Forest offer 10%, surpassing HMM and SVM, which show 3% and 8%, respectively. (Figures 3, 4, 5, 6)

From the perspective of computational efficiency

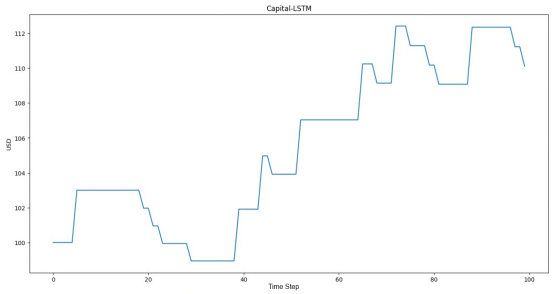

Figure 4: Gains obtained by a Back testing test for an LSTM model.

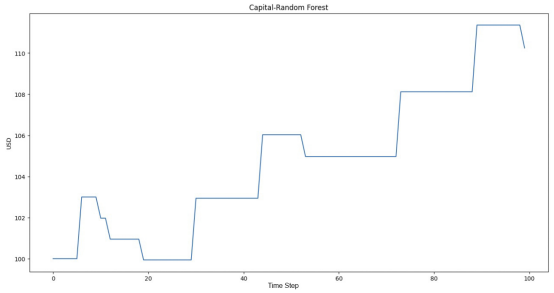

Figure 5: Gains obtained by a Back testing test for an RF model.

and consistency in execution, SVM and Random Forest stand out, as both require only one second to run and have proven to be consistent in their executions. In contrast, the HMM, although fast, fails in execution consistency (as sometimes when it runs, it presents errors). This could be a limiting factor for its use in real-time trading applications where consistency and speed are crucial.

In conclusion, Random Forest seems to be the most balanced model, offering a good combination of accuracy, profitability, and efficiency. Each model has specific advantages that could make them more suitable for different applications within stock price prediction, depending on the specific needs of accuracy, long-term interpretation, and risk management capabilities.

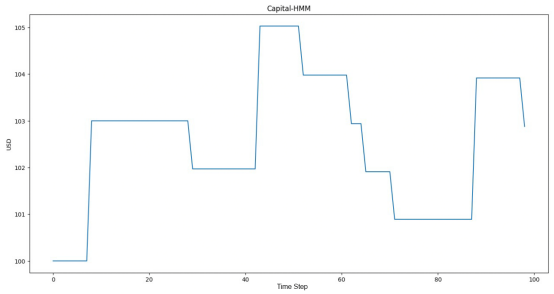

Figure 6: Gains obtained by a Back testing test for an HMM model.

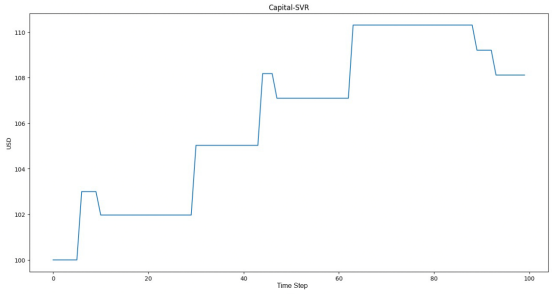

Figure 7: Gains obtained by a Back testing test for an SVR model.

6 Discussion

According to the results, it is observed how the models were able to question the efficient market hypothesis. Models such as the HMM and LSTM achieve a high fit of the data, demonstrating that it is possible to approximate the behaviour of a stock and thus understand its future behaviour beforehand. Likewise, when predicting the movement or classifying price patterns, it is possible in some cases to obtain returns like the S&P500, as models such as Random Forest achieve a return of 10.24% when in the same interval the actual market return is around 10.69%. However, the evidence presented is anecdotal with respect to the size and diversity of the market. We also consider that these cases, although successful, are not generalizable to all assets and all prediction time windows. Given this, we only limit ourselves to questioning whether the market is 100% efficient

since more evidence would be needed to have a complete counterexample that would allow us to attack the efficient market hypothesis. With this, it can only be evidenced that the market is possibly not completely efficient without denying that the hypothesis is true.

7 Conclusion

Based on the results obtained, and the discussion presented, it can be observed that:

- No model was able to outperform the return of the S&P500 given the back testing strategy proposed during the defined time window.
- Among the classification models, the one closest to the market results was Random Forest. This may be due to each tree performing a prediction (price rise or fall) and the final decision being made by "majority vote" of all trees, which reduces the risk of overfitting and improves the generalization capability of the model. This "ensemble learning" technique allows for capturing complex and nonlinear relationships between input features and price direction.
- The results are conditioned by the training and testing data in addition to only having been tested with the S&P500. Therefore, it is recommended that in the future, the models be evaluated on different portfolios with variations in the time windows to be used.
- It is possible to empirically question the efficient market hypothesis. However, more testing and model adjustments are needed to obtain data that allow us to conclude whether there is a way to model and predict prices. In this way, the efficient market hypothesis is not denied, but rather the possibility of seeking arguments that show that it is not 100% efficient is opened.

8 References

[1] Cornell, B. (2020). Medallion Fund: The Ultimate Counterexample? The Journal of Portfolio Management (46).

[2] Harper, K. (2016). Challenging The Efficient Market Hypothesis With Dynamically Trained Artificial Neural Networks. UNF Graduate Theses and Dissertations. (718)

[3] Phuoc, T. et al. (2024) Applying machine learning algorithms to predict the stock price trend in the stock market – the case of Vietnam, Humanit Soc Sci Commun (11), 393.

[4] Suryoday, B. et al. (2018) Predicting the direction of stock market prices using tree-based classifiers, The North American Journal of Economics and Finance (47), 552-567.

[5] Tripathy, N. (2019). Stock Price Prediction Using Support Vector Machine Approach. Ocford United Kingdom; Indian Institute of Management.

[6] Verma, A. et al. (2021) stock price prediction using Hidden Markov models and understanding the nature of underlying hidden states.

8.1 Code

[1] Cabrales, S. (n.d.). Algoritmos de negociación basados en machine learning. [MOOC]. Coursera. https://www.coursera.org/learn/algoritmos-de-negociacion-basados-en-machine-learning

[2] Jain, A. (2018). Stock Forecasting using Hidden Markov Models. https://github.com/ayushjain1594/Stock-Forecasting

[3] Kumar, A. (2023).Stock Price Prediction Using LSTM. GitHub. https://github.com/034adarsh/Stock-Price-Prediction-Using-LSTM

9 Project Repository

https://github.com/WilliamBayona/Comparison-of-Stock-Price-Prediction-Models-using-ML

Dedicated to Jim Simons (1938 -2024)