Round 1 Report and Reflection (Research)

Haechan Oh

Minerva University

CP191

Prof. Mathur

October 14, 2024

The research explores the potential of using technical indicators in a combination with machine learning models for stock price movement forecasting. The major goal of this project is to check how time series forecasting, using ARIMA and integrated into a RandomForest model, would influence the predictive performance of this model. This project was structured into a series of steps that included data processing, feature engineering, integration of ARIMA, model development, and then performance evaluation using visualization.

Collection and processing of data was the initial step for the research. We obtained historical stock market data on the S&P 500 index from Yahoo Finance, ranging from January 2020 to January 2023. Our analysis is based on daily observations of Open, High, Low, Close price, and Volume. We decided to use an exponential smoothing of the closing prices as a method of smoothing short-run fluctuations and giving more focus to the longer-term trend in stock prices. This yielded a "Smoothed_Close" used for further calculations and as input for both machine learning models.

Our feature engineering was supported by a set of well-established technical indicators. These are indicators from historical data of stock prices and volume, which are commonly used by traders and financial analysts in predicting future movements of the market. Various indicators, as identified for this study, include RSI, Stochastic Oscillator, Williams %R, Moving Average Convergence Divergence, Rate of Change, and On-Balance Volume. Each of these indicators gauges a different dimension of market behavior, ranging from momentum to trend strength and further to buying or selling pressure. In this respect, RSI and Stochastic Oscillator are classified as momentum indicators because they exhibit the speed and amplitude of changes in prices, while MACD is an indicator that reflects the convergence or divergence between

moving averages with the purpose of portraying trend reversals. These indicators, computed from the smoothed closing price, became our feature set for the machine learning models.

Besides the technical indicators, we also included ARIMA in the feature set. The ARIMA is a statistical model of time series forecasting; it picks up patterns in past data and uses them to predict future values. ARIMA mainly consists of three components: AutoRegression (AR), Integrated (I), and Moving Average (MA). The AR part focuses on the relationship between a value at a point in time and previous values in time-lag values. The MA part models the relationship between a value and past forecast errors. Integrated puts the time series into stationary by applying differencing to remove either trends or seasonality. Using that for stock price forecasting, we've added the predicted values as a new feature in our RandomForest model. This allowed us to marry the strength of ARIMA in sequential forecasting with pattern recognition capabilities supported by machine learning.

We developed two different machine learning models in order to compare them. First, we created a baseline RandomForest classifier using only the technical indicators as features. RandomForest is one of many ensemble learning methods that generate a great number of decision trees and combine the outputs in order to predict with greater accuracy. That makes it very apt for classification problems such as stock movement prediction because it helps to reduce overfitting and make the model more robust by averaging out across trees. The second model added the ARIMA forecast to the feature set with the technical indicators. By doing this, we tried to assess if adding ARIMA's sequential forecasting data would increase the predictive performance of the RandomForest model.

We further compare the performances of both models using four major metrics: accuracy, precision, recall, and F1 score. Accuracy is the proportion of correctly predicted price movements. Precision evaluates a model's performance for correct identifications of positive price movements. Recall reflects the model's ability to detect all actual positive movements, whereas the F1 score is a harmonic mean of precision and recall; hence, it gives a balanced view of model performance. These metrics have been evaluated for both models on a range of prediction periods from 1 to 30 days in order to understand how the models performed over different time horizons.

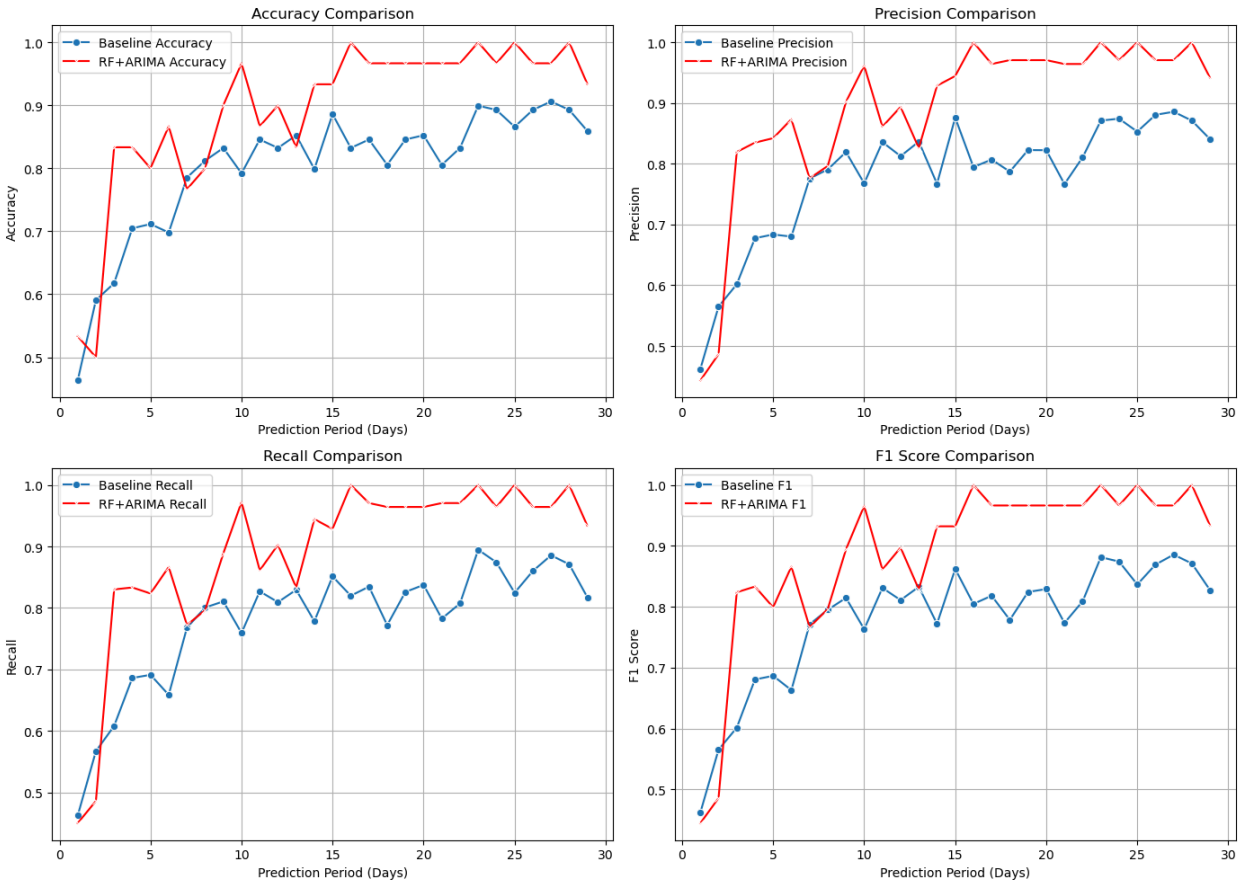

Figure 1. Two models compared on four different metrics: accuracy, precision, recall, and F1.

The visualization above shows the clear difference between the two models. Overall, speaking about accuracy, the model with ARIMA integration is always higher than that of the baseline model. Since most of the prediction periods tend to increase, the ARIMA-enhanced model attained higher accuracy, managing to stabilize at close to perfect accuracy. Precision still shows similar results, where the integrated model outperforms the baseline model, especially in identifying the model performance of positive price movements. Both Recall and F1 score showed this trend, which means the ARIMA-enhanced model was better at correctly classifying the stock price movements with better balancing between precision and recall. On the other hand, the baseline model was more volatile and had scores on all metrics that were lower, especially for increasing periods of the prediction period.

In conclusion, incorporating ARIMA and technical indicators into the model significantly enhanced the capability of the RandomForest model in predicting stock price movements. With machine learning, we were able to further leverage the power of ARIMA in time series forecasting, detecting patterns and thus improving accuracy, precision, recall, and F1 score from the baseline model. The results of this study indicated that the integration of machine learning with statistical models, such as ARIMA, is a more holistic and robust approach toward stock market prediction by providing insights from both the sequential nature of stock prices and market trends captured by technical indicators. Based on this approach, further research and practical applications in the analysis of financial markets are highly promising.

Analysis of deliverable

The research has shown clearly how the integration of ARIMA forecasts with a RandomForest model could lead to a better accuracy in the stock price forecast. Regarding the technical setup, everything was done efficiently, right from data processing to feature engineering, thus presenting a robust dataset for analysis. The chosen combination of technical indicators with ARIMA managed to grasp both the market trends and sequential patterns in the stock data.

The strengths are that in this project, there was a very methodical comparison of two different models: a baseline RandomForest model versus one with ARIMA forecasts added. Indeed, the addition of ARIMA predictions really improved the performance on virtually all key metrics, which has shown the value in such hybrid approaches that meld time series forecasting with machine learning. More importantly, the capability to visualize these results made it very clear where the improvements came from and also how consistent over time the ARIMA-enhanced model was.

However, there are a couple of areas for improvement: first, both the hyperparameter tuning of the ARIMA and RandomForest models may lead to even better performances. As an example, the parameters p, d, and q of the ARIMA model were only superficially optimized, and likewise, the RandomForest model can be tuned even further by using a grid search or cross-validation.

If I were to proceed further with this project, I would focus on a number of extensions which could make the project at least comprehensive and possibly turn the project into a Capstone project. A very valid next step could be the introduction of more complex models, such

as neural networks-especially LSTMs-designed for sequential data. It could be mixed with even more sophisticated feature engineering techniques in order to grasp even more peculiar market behaviors. One can go further by expanding the data to feature more financial indicators or macroeconomic factors such as interest rates, inflation data, geopolitical events, and so on that would better analyze the results of modeling. Another possible extension may be in developing a real-time stock prediction system with inputs from live data streams and the construction of a web app allowing the interaction of the end user with the system. That would be beyond mere predictive modeling into deployment techniques such as Flask, Docker, or cloud services such as AWS or Azure. This would make the project a full-stack data science work applicable to real-world practice. Constructing such a system could be a Capstone project because it would bring together not only the advance machine learning methods but also deployment and scalability concerns central to the data science job.

One was the trade-off between model complexity and interpretability: while more complex models such as ARIMA and RandomForest may yield better results, they required a lot of tuning, and interpretation of results might also become challenging. That taught me the importance of simplification wherever possible without sacrificing the accuracy of the models. In future projects, I will always try to allow more time for model tuning and look first into simpler models before going to advanced techniques.

In my concentration courses this semester, I learn more about advanced machine learning algorithms and cloud computing highly related to my potential Capstone. What's more, I plan to conduct several informational interviews with professionals working in the financial technology sector to further understand how predictive modeling is applied in real-world finance.

Appendix

HCs to be scored:

#dataviz: I applied this HC effectively by generating detailed visual comparisons between the baseline RandomForest model and the ARIMA-augmented model. Through line graphs displaying accuracy, precision, recall, and F1 scores, I provided a clear and interpretable way to analyze the performance differences between the models. This allowed for easy understanding of which model performed better over different time periods and why. Choice of line graphs for comparing changes across time horizons shows an understanding of how best to visualize trends in your data.

#modeling: The core of this project involved developing two different predictive models: the baseline RandomForest model and the enhanced model incorporating ARIMA. I applied this HC well by not only building these models but also carefully evaluating their performance using various metrics. Furthermore, I interpreted the results to conclude that the ARIMA-enhanced model outperformed the baseline model. This process of building, testing, and comparing models demonstrates proficiency in model evaluation and interpretation.

Additional HCs applied:

#selfawareness: I applied this HC in this project by selecting a topic that aligned with my strengths in machine learning and stock market prediction, while acknowledging my time constraints and capabilities. Knowing that I had around 8 hours to complete the task, I chose a manageable scope by integrating ARIMA with RandomForest, two models I was already familiar with, rather than overextending myself into unfamiliar territory. This self-awareness

helped me avoid the common pitfall of overestimating my abilities and ensured that I could complete the project within the set time frame, focusing on quality while still challenging myself to learn and improve.

#sampling: In this project, I applied **#sampling** by selecting stock data from the S&P 500 (SPY) as a representative sample of the broader financial market. The dataset, covering the period from January 2020 to January 2023, was chosen because it captures significant market fluctuations, allowing the models to learn from various market conditions. By focusing on this time range, I aimed to ensure that the sample included both bullish and bearish trends, making the predictions more generalizable. Although the sample is limited to one index, the use of technical indicators and ARIMA forecasting methods allowed for a reasonable approximation of stock movement patterns, providing insights that could potentially be applied to similar markets or stocks with analogous behaviors. The sampling process thus aimed to balance specific market dynamics with broader applicability, ensuring the results were interpretable and generalizable to other contexts.

#composition: In this project, I applied **#composition** by ensuring that my written communication was clear, concise, and tailored to the intended audience. I avoided unnecessary complexity and focused on explaining key concepts—such as technical indicators and the integration of ARIMA—using straightforward language. By carefully structuring the essay, I ensured smooth transitions between data processing, model development, and performance evaluation, allowing the reader to follow the progression of ideas easily. Additionally, I used visualization to complement the written explanations, reinforcing the analysis in a direct and accessible way. This attention to clarity and precision helped convey the technical aspects of the project without overwhelming the reader, demonstrating effective composition throughout.

Link to github repository for code reference:

https://github.com/haechan01/ML-stock-market-prediction

References

Hardikkumar. (2024, May 31). *Stock market forecasting using time series analysis with Arima Model*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analysis-with-arima-model/

Khaidem, L., Saha, S., & Dey, S. R. (2016, April 29). *Predicting the direction of stock market prices using random forest*. arXiv.org. https://arxiv.org/abs/1605.00003