GR5261 Statistical Methods in Finance

Final Project Report

Group Members:

Yixuan Liang	yl5390
Yuncong Liu	yl5440
Yiwei Lu	yl5515
Yingnan Shen	ys3750
Yui Yin Xi	yx2866
Muqing Yang	my2842

Instructor:

Prof. Zhiliang Ying

Department of Statistics

May 2, 2024

Contents

1 Introduction

Our project aims to enhance decision-making processes in the financial sector by contributing to the existing toolkit of investors through an analysis of various forecasting techniques. The datasets pertinent to this study are sourced from Yahoo Finance.

1.1 Motivation

In the rapidly evolving landscape of finance, the accurate forecasting of stock prices is both a crucial challenge and a substantial opportunity for investors. This report aims to address these challenges and apply the knowledge gained in GR5261 *Statistical Methods in Finance* by conducting a comprehensive analysis of stock price forecasting. The study focuses on e-commerce giants, with Amazon serving as the primary case study.

1.2 Framework

Our group has utilized two advanced predictive models, namely the Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks, to investigate their efficacy in forecasting stock price movements.

The ARIMA model is a well-established method in time-series analysis. We conucted the Augmented Dickey-Fuller (ADF) test to determine the presence of non-stationarity in the stock price data, followed by a meticulous fitting strategy and an evaluation of the model's forecasting capabilities. Additionally, we enhance our methodology through a rolling analysis to refine our predictions.

Subsequently, we introduced the LSTM model, diving into its implementation and discussing its intricacies and potential as a forecasting tool. This neural network technique, known for its prowess in handling sequential data, is analyzed to ascertain its suitability and performance in stock price prediction.

The highlight of our study is a comparative analysis of both models, providing valuable insights into their predictive accuracies and operational differences. Through this analysis, we aim to not only measure the effectiveness of each model but also propose potential improvements and envisage the future of stock price forecasting.

2 Stock Price Predicting

2.1 ARIMA Model

2.1.1 Introduction of ARIMA

ARIMA (AutoRegressive Integrated Moving Average) is a popular statistical method used for time series forecasting. It is a generalization of the simpler ARMA (AutoRegressive Moving Average) model and can capture more complex patterns in the data. In practice, It is effective for short-term forecasting and handling stationary data, often used for low-frequency datasets, such as GDP or yields. However, it often struggles with predicting turning points. For high frequency data, ARIMA's performance may be limited due to high volatility and unpredictable external factors. In the first part of our prediction analysis, we applied ARIMA to stock price forecasting, which underscored these limitations. The formulae are as follows:

$$AR(p) : Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + Z_t$$

$$MA(q) : Y_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad Z_t \sim WN(0, \sigma^2)$$

2.1.2 Fitting Strategies

Initially, we perform the Augmented Dickey-Fuller (ADF) test. If the p-value surpasses 0.05, indicating non-stationarity, we iteratively difference the series until it satisfies the ADF test. Typically, first-order differencing is adequate for this purpose. Subsequently, we utilize two nested loops to systematically explore potential values of p and q, ranging from 0 to 4, inclusively. The selection of model parameters is based on minimizing the Akaike Information Criterion (AIC).

2.1.3 Our Methodology

Algorithm 1 Implementation of ARIMA

Input: Close price of AMZN from 2012-01-01 to 2024-4-18;

Parameters: (p,d,q): (2,1,2); Training ratio: 0.95

Output: Stock price forecast visualization

1: Split the data into training and testing sets based on a specified training ratio, which we later use to illustrate the ARIMA model's capability for short-term forecasting.
2: Fit the Model
3: Prepare for Forecasting
4: Predict and Reverse Differencing

return Visualization and RMSE

Relevant Python code can be found in the attachments.

2.1.4 Results and Interpretation

According to figure 1, the model seems to capture the general trend, but an RMSE of 20 might be a little high for the average stock pricing being 150. The relationship between the overall forecasting performance, denoted as RMSE in Figure 2, and the proportion of the training set is observed to be positive, indicating stronger performance in short-term prediction scenarios. However, the ARIMA model exhibits limitations in capturing information amidst complex stock price movements, resulting in prediction outcomes that appear overly linear in contrast to the volatility of stock prices. To address this, we introduce two additional methods: rolling ARIMA and Long Short-Term Memory (LSTM) networks.

2.2 Rolling ARIMA Model

2.2.1 Introduction of Rolling ARIMA

The idea of employing Rolling ARIMA comes from extensive literature reviews and comparative studies between ARIMA and various machine learning methods, which demeritsistently use rolling ARIMA as the benchmark. Rolling analysis reflects its effectiveness in dealing with complex financial data.
Below are the Merits and Demerits of Rolling ARIMA compared with ARIMA.

Merits:

- Adaptation to New Data: align with the latest market trends.
- Handling Non-Stationarity: more effectively mitigate the impact of any residual non-stationarity that is not addressed by the I of ARIMA.
- Focused Predictions: allow models to concentrate on short-term forecasting—typically one day ahead.

Demerits:

- Lagging Reaction: may still react slower than desired to sudden changes in the market.
- Parameter Sensitivity: be highly sensitive to the choice of parameters (p, d, q) and the size of the rolling window.
- Overfitting Risk: becomes too closely tuned to recent 'noise'.

2.2.2 Our Methodology

The primary procedure is similar to the ARIMA algorithm(2.1.3), but do note that:
Window size = 30, and only the next prediction value is obtained with each iteration.
Parameter: ACF(for d) and AIC(for p, q); And (p, d, q) is recalculated at each iteration based on the data within the rolling window.

2.2.3 Results and Interpretation

According to figure 3, Rolling ARIMA performs much better than ARIMA, since the RMSE has been reduced from 20 to below 3.

2.3 LSTM

2.3.1 Introduction of LSTM

Long Short-Term Memory (LSTM) networks represent a category of recurrent neural networks. LSTMs are distinguished by their capability to retain information over prolonged intervals, rendering them highly suitable for tasks involving time-series prediction, including stock price forecasting. Figure 4 illustrates the fundamental framework of LSTM operation, which revolves around three essential gates: the input gate, the forget gate, and the output gate. These gates collaboratively determine the information that should be preserved or discarded as the network processes data sequentially over time.

Merits:

- Capture Long-Term Dependencies: overcome the vanishing gradient problem common in traditional recurrent neural networks.
- Robustness to Gaps in Data: handle cases where data might not be Demeritsistently available.
- Non-Linearity: model complex non-linear relationships that are often observed in financial markets.

Demerits:

- Prone to Overfitting: easily overfit on noise rather than underlying trends in stock price data.
- Sensitive to Hyperparameter Settings: heavily depends on the configuration of numerous hyperparameters, including the number of layers, the number of units per layer, learning rate, etc.
- Black Box Nature: the internal workings of a system or model are not visible.

2.3.2 Our Methodology

Algorithm 2 Implementation of LSTM

Input: Close price of several stocks from 2012-01-01 to 2024-4-18;

Output: Stock price forecast visualization

1: Random seed: Fix it for reproducibility
2: Data preprocess: Normalizing (MinMaxScale + Windowing by 60 days)
3: 2 hidden layers (LSTM, 128 nodes64 nodes); 2 dense layers
4: Adam optimizer; MSE loss function
5: Test set: last 60 days of the data
return Visualization and RMSE

Relevant Python code can be found in the attachments.

2.3.3 Results and Interpretation

Utilizing the LSTM model described previously, we conducted forecasting of the closing prices for 18 e-commerce companies. To exemplify the diverse market behaviors observed, we selected two companies for illustrative purposes. The first company is Amazon, representing a well-established entity with an extensive market history. The second company is a comparatively newer entity, characterized by a shorter market presence and more pronounced price fluctuations.

The model's predictive accuracy was evaluated in figure 5 using the Normalized Root Mean Square Error (NRMSE). The first graph in figure 6 is the sales volume graphs for selected e-commerce companies, revealing varying patterns of trading activity. For some companies like Chewy and Maplebear, it shows that there are sporadic spikes that could indicate irregular trading behavior or significant market events.

On the other hand, the risk-return profile figure 7 shows there is an intricate relationship between a company's risk/return profile and the LSTM model's forecasting accuracy. Companies with higher volatility and expected return, such as Sea and Revolve, are positioned further right on the graph, suggesting that the model might be capturing the inherent risk associated with their stock prices. Conversely, firms like Amazon and Ebay, which have lower risk and expected returns, lie closer to the origin, hinting at a more stable and less volatile stock performance. This stability could contribute to a more Demeritsistent historical data pattern, which the LSTM is potentially better equipped to learn from and predict.

The analyses collectively indicate that companies characterized by higher trading volumes do not consistently exhibit improved predictability, suggesting a nuanced relationship beyond surface-level observation. One plausible hypothesis is that erratic sales volumes introduce noise that complicates the LSTM model's ability to interpret data, consequently impacting its forecasting accuracy. Similarly, stocks displaying higher volatility, while potentially promising greater returns, also pose challenges for the LSTM model in accurately predicting price movements due to their unpredictable nature. These observations prompt further exploration into the distinct market dynamics and operational intricacies of these firms. Such an investigation could reveal factors contributing to the complexities of forecasting their stock prices with a high degree of precision.

3 Comparison of the Models

As depicted in Figure 8, the combined forecasting results of rolling ARIMA and LSTM reveal a close RMSE, with rolling ARIMA exhibiting a slightly lower value, as illustrated in Figure 9.
However, it is essential to consider that RMSE is not the sole criterion for evaluation, and other factors should also be taken into account. Both models demonstrate effective tracking of actual prices; nevertheless, they also display inevitable lag effects. In terms of prediction smoothness, LSTM outperforms rolling ARIMA, as the latter tends to be overly sensitive and can generate excessively volatile results. Additionally, a review of Table 1 reveals that the average execution time for rolling ARIMA is approximately 20 minutes. This duration may be attributed to our method of predicting one value at a time, underscoring a trade-off between accuracy and efficiency.

	Rolling ARIMA	**LSTM**
Average Execution time	20 mins	0.5 min
Tracking the Actual Prices	✓	✓
Predictive Accuracy (RMSE)	3.1916	3.9723
Smoothness of Predictions		✓
Lag effects	✓	✓
Reactivity to Changes	Too sensitive	General Trend

Table 1: Comparison of Rolling ARIMA and LSTM Models

4 Discussions

We've also engaged in a discussion on enhancing the forecasting capabilities of our models and have identified four potential avenues for improvement.

Possible Further Discussions:

- **Incorporating Exogenous Variables**: Include additional variables that could affect the forecasts, such as economic indicators, or other sector-specific drivers.
- **Hybrid Model**: Combining ARIMA with LSTM or other machine learning algorithms. ARIMA for the linear part and LSTM for residuals.
- **Seasonal Adjustment**: For strong seasonal patterns not captured by simple differencing, integrating seasonal differencing steps can improve model performance.
- **Error Correction and Intervention Analysis**: Adjust for past prediction errors or handle abrupt changes in the time series due to external shocks.
- **GARCH Model**: The involvement and analysis of GARCH model can be useful to forecast the volatility of financial time series data.

5 Bibliography

1. D. Ruppert and D. S. Matteson, *Statistics and Data Analysis for Financial Engineering*, 2nd edn, Springer, New York, 2015. ISBN 978-1-493-92613-8.

2. Modeling Financial Time Series with S-PLUS®, "Rolling Analysis of Time Series," Springer, New York, NY, 2006. https://doi.org/10.1007/978-0-387-32348-0_9.

3. R. Qiao, W. Chen, and Y. Qiao, "Prediction of stock return by LSTM neural network," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022. https://doi.org/10.1080/08839514.2022.2151159.

4. E. W. Saad, D. V. Prokhorov, and D. C. Wunsch, "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 6, pp. 1456–1470, 1998.

5. https://medium.com/making-sense-of-data/ . time-series-next-value-prediction-using-regression-over-a-rolling-window-228f0acae363

6 Appendix

Here are the aforementioned figures.

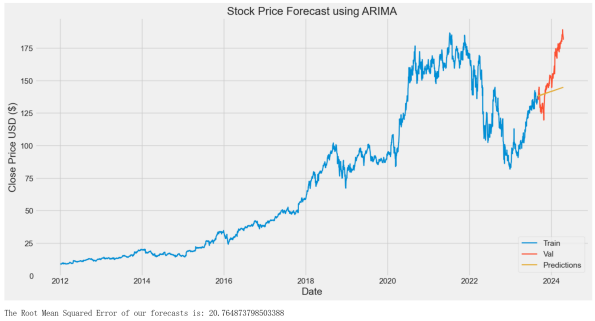


Figure 1: Stock Price Forecast with ARIMA

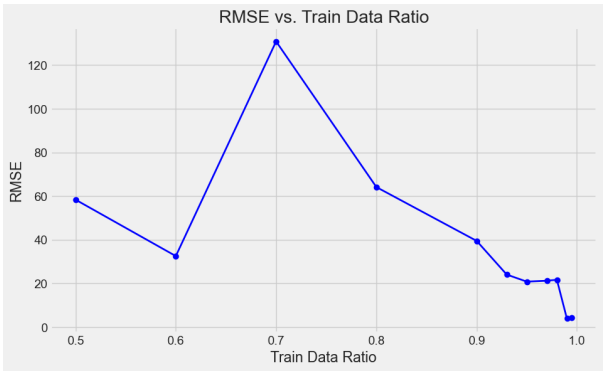


Figure 2: RMSE vs Train Data Ratio

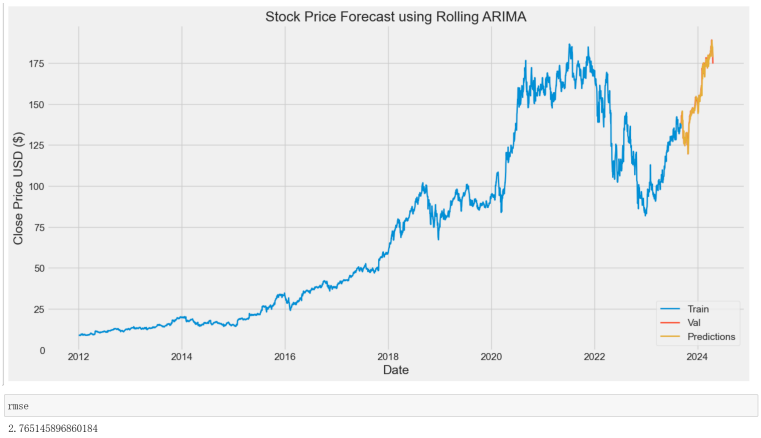


Figure 3: Stock Price Forecast with Rolling ARIMA

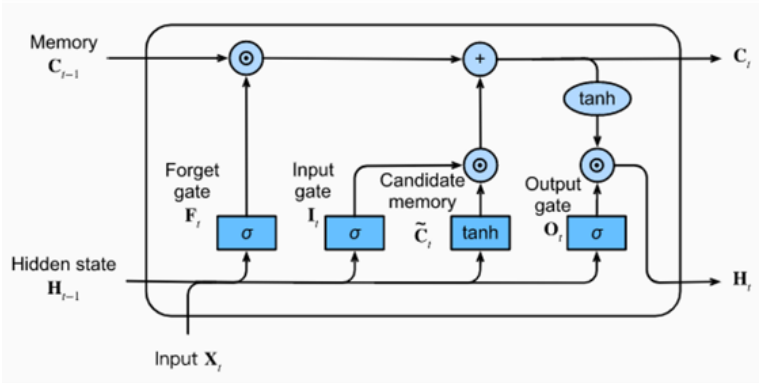
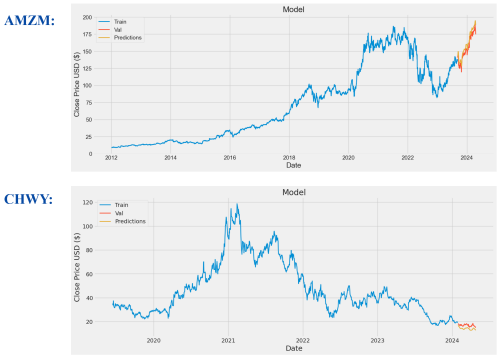


Figure 4: How LSTM works



Results from LSTM:

	ticker	NRMSE		ticker	NRMSE
8	EBAY	0.062595	16	BYON	0.066195
0	AMZN	0.073341	0	AMZN	0.067803
10	VIPS	0.102513	4	MELI	0.068662
1	BABA	0.102727	7	SE	0.088441
7	SE	0.116448	8	EBAY	0.094125
13	W	0.119727	6	JD	0.127425
16	BYON	0.125505	11	ETSY	0.144845
11	ETSY	0.125609	10	VIPS	0.146376
2	PDD	0.129581	3	MPNGY	0.148106
15	RVLV	0.154049	2	PDD	0.171358
4	MELI	0.174119	5	CPNG	0.185682
3	MPNGY	0.188454	1	BABA	0.205241
14	GLBE	0.344472	13	W	0.211634
5	CPNG	0.348806	15	RVLV	0.300330
6	JD	0.578275	14	GLBE	0.489098
9	CART	0.821662	12	CHWY	1.141740
12	CHWY	0.822837	9	CART	1.356627

Figure 5: LSTM Result Part 1

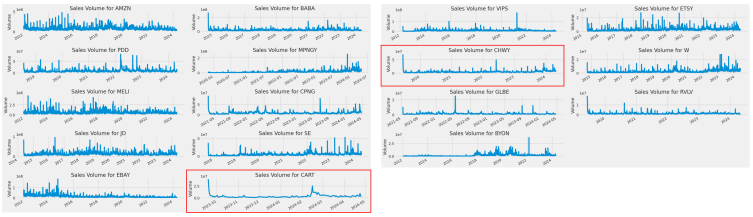
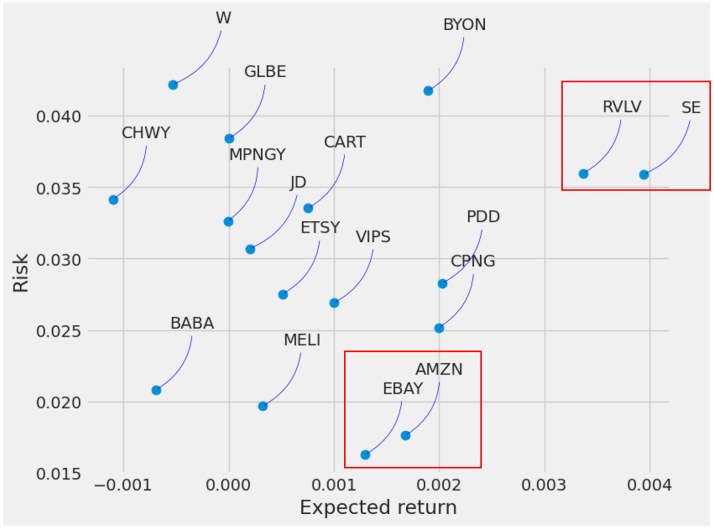


Figure 6: LSTM Result Part 2



NRMSE:
AMZN: 0.0733 RVLV: 0.300

Figure 7: LSTM Result Part 3

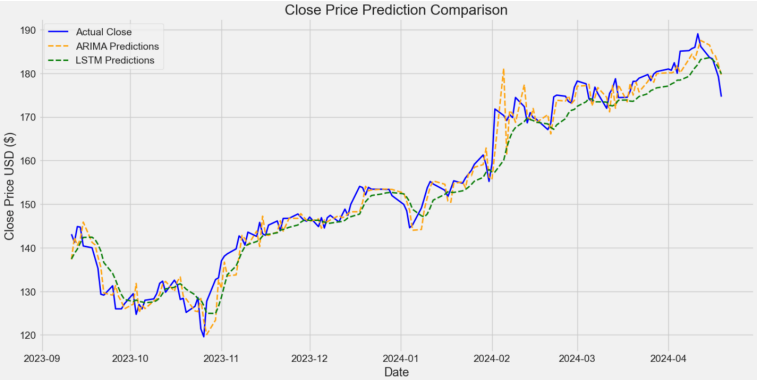


Figure 8: Close Price Prediction Comparison

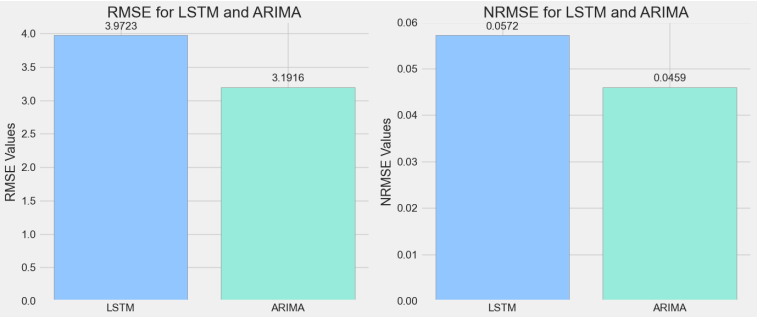


Figure 9: NRMSE