EDA of Yahoo Stock Price Prediction

Alexandra Guahnich
Georgia Institute of Technology
Atlanta, GA 30332
aguahnich3@gatech.edu

Raquel Ovadia
Georgia Institute of Technology
Atlanta, GA 30332
rbugin3@gatech.edu

Abstract

This study aims to analyze the maximum stock price in a given time period using financial data available through the Yahoo Finance API. The goal is to develop predictive models that can assist investors, analysts, and researchers in making informed decisions about stock investments. Specifically, we will focus on predicting stock price trends based on various financial metrics, including the lowest price in a given time period, the starting and ending of the trade in the same period, and the volume which refers to total amount of trading activity. Also, we will consider the date stamps on our trade entries in order to examine time dependency.

1 Introduction

Can predictive models examining financial metrics from the Yahoo Finance API accurately forecast the maximum stock price within a given time frame? How do seasonal patterns and key financial indicators, such as lowest price, opening/closing prices, and trading volume, influence the predictive accuracy, and what model(s) offer the most reliable predictions for assisting investors, analysts, and researchers in making informed decisions about stock investments? In order to dive into possible answers to these issues, we'll discuss forecasting methodologies in sections 3, 4, and 5, with concluding insights in section 6.

2 Description of Dataset

The dataset contains 1460 value sets (rows) and 6 variables pertaining to each trade input. From them, the 'High' column represents the dependent variable to be studied. Below are the explained headers of the dataset utilized in the analysis.

2.1 Dataset Headers

Each data sample corresponds to one trade entry and contains a total of 6 features that are described below:

1. **Date:** (Char) it contains the trading date.
2. **High:** (Numeric) it contains the maximum price of the stock in a given time period.
3. **Low:** (Numeric) it contains the minimum price of the stock in a given time period.
4. **Open:** (Numeric) it contains the stock price when the trading began.
5. **Close:** (Numeric) it contains the stock price when the trading ended.
6. **Volume:** (Numeric) it contains the total amount of trading activity.

Other Details: All prices are in US dollars.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

3 Multiple Linear Regression Analysis

3.1 Multiple Linear Regression Model

For the Multiple Linear Regression analysis, we initiated a regression to predict the ‘High’ value based on the variables ‘Low,’ ‘Open,’ ‘Close,’ and ‘Volume.’ All variables exhibited significance, with p-values below 0.05. However, we sought to assess multicollinearity to ensure the model’s appropriateness. We generated a Scatter Plot Matrix (Figure 1), revealing a strong correlation among ‘Low,’ ‘Open,’ and ‘Close.’ To delve deeper into this observation, we calculated the Variance Inflation Factor (VIF) for each regression variable. The VIF values for ‘Open,’ ‘Close,’ and ‘Volume’ were approximately 1369, 563, 827, and 1, respectively. Interpreting these findings, we confirmed our hypothesis of multicollinearity between ‘Low,’ ‘Open,’ and ‘Close,’ as their respective VIF values far exceeded 10. Notably, the ‘Volume’ variable demonstrated independence, with a VIF value significantly below 10.

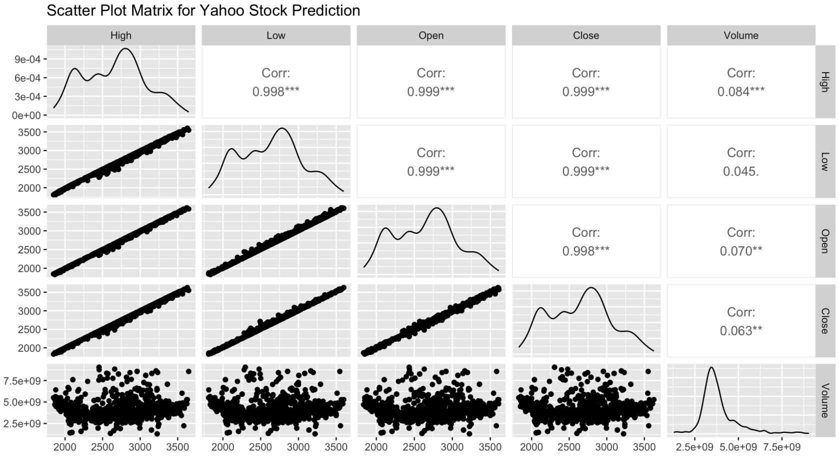


Figure 1: Scatter Plot Matrix

3.2 Mitigating Regression Pitfalls

To address multicollinearity, we considered three options. First, we explored removing one of the correlated variables. However, after systematically eliminating each correlated variable one at a time, the VIF values remained significantly greater than 10. Consequently, we concluded that this approach was not effective in resolving our multicollinearity issue.

For the second option, we experimented with combining the values of ‘Low,’ ‘Open,’ and ‘Close’ into a new variable called ‘Average.’ This successfully resolved our multicollinearity problem, as evidenced by the VIF value for ‘Average’ reducing to around 1. Despite this success, we hesitated to choose this approach, as it would reduce the number of variables from 4 to 2.

Finally, for the third option, we decided to implement Lasso Regression. To validate this decision, we compared the Mean Squared Error (MSE) of the second option (the one with the ‘Average’ variable) at 107.8994 with the MSE of the third option (Lasso Regression) at 103.7188. Since the MSE of the Lasso Regression model was lower than the MSE of the model including the ‘Average’ variable, we opted to retain the Lasso Model, as it minimized the MSE and included all 4 variables.

3.3 Model Accuracy Diagnostics

To assess the adequacy of our Lasso model, we conducted various diagnostics. Initially, we noted that the Percentage of Deviance Explained by the model, akin to R-squared, reached an impressive 99.95%, indicating robust model performance. Finally, we examined the Residual Plot for the Lasso Regression (Figure 2). In this figure, the residuals appear randomly scattered around zero, with no discernible trend or pattern. This observation supports the suitability of the Lasso Regression as a adequate model for our dataset.

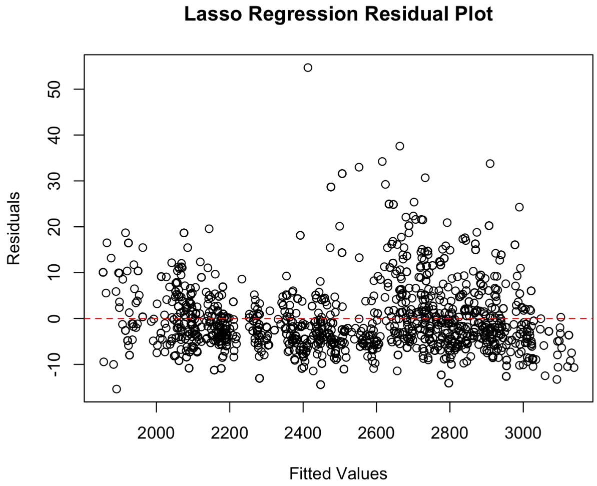


Figure 2: Lasso Regression Residual Plot

3.4 Remarks for Multiple Linear Regression

In summary, our final model is expressed as $\hat{y} = 2529.926 - 26.419x_1 + 236.670x_2 + 117.465x_3 + 2.736x_4$, where x_1 represents the 'Low' value, x_2 the 'Open,' x_3 the 'Close,' and x_4 the 'Volume.' This equation serves as a predictive tool for estimating the 'High' stock price in a given time period, providing an upper limit for Yahoo's stock price.

4 Time Series Regression Analysis

4.1 Time Series Regression Model

In this section, we analyzed the time series data, observing a consistent increasing trend in the 'High' values over time. The linear model line of $y_t = 1987 + 0.7436t$ indicates a clear linear trend within the data. Our fitted Time Series Regression model demonstrates a strong fit with R^2 value of 0.9186. The significance of the coefficients was confirmed through t-tests, rejecting the null hypothesis for each coefficient at $\alpha = 0.05$.

4.2 Model Accuracy Diagnostics

For model accuracy checks, we conducted diagnostic assessments. We investigated autocorrelation using the Durbin-Watson test, the sum of the squared difference of current and yesterday's residuals divided by the sum of todays squared residuals, where our data yielded a d-term of 2.002089. With a sample size of 1460 and $\alpha = 0.05$, the Durbin-Watson Table indications showed that values near 2 imply no significant autocorrelation, values significantly less than 2 (approaching 0) indicate positive autocorrelation, and values significantly greater than 2 (approaching 4) indicate negative autocorrelation, reinforcing the absence of autocorrelation in the residuals. Thus, we were successful in not rejecting the null hypothesis. In other words, no significant autocorrelation is present in the residuals of our regression model. Similarly to assessment tools used for stock forecasting studies, performing the Durbin-Watson Test allowed us to analyze the residual relationship between current and past price behavior.

4.3 Time Series Decomposition

We performed the decomposition of the Time Series to examine its components (Figure 3). The analysis displayed a consistent linear increase in trend and a yearly pattern in seasonality. Also, the remainder component displayed scattered data points of the residuals with minimal variability, suggesting that the residual portion captures little irregular fluctuations or random variations in the model. This second tool for autocorrelation proved the validity of the regression method and demonstrated that the independence assumption is not being violated.

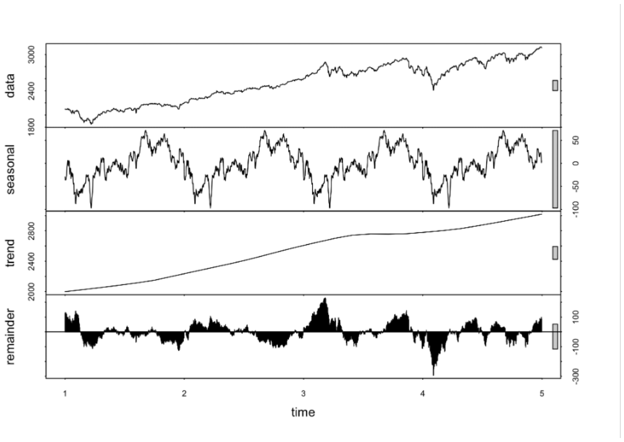


Figure 3: Time Series Decomposition Components

4.4 Remarks for Time Series Regression

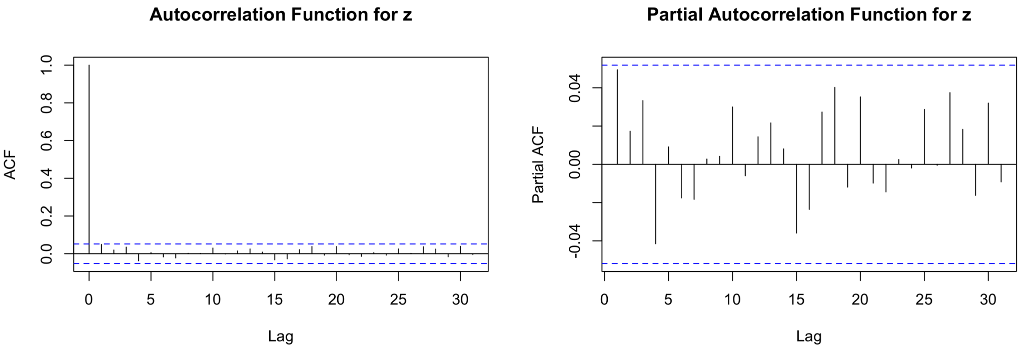
These analyses collectively indicate a well-fitted regression model that effectively captures the trends and patterns within the time series data, providing valuable insights to our overall EDA since the Time Series Regression presented data behavior over time and its dependency through the seasonal tendencies.

5 Nonseasonal Box-Jenkins Models Analysis

5.1 Nonseasonal Box-Jenkins Model and Transformation

For the Nonseasonal Box-Jenkins Models, we initiated the analysis by examining the plot of the time series y_t, as shown in Figure 3, in time vs. data. Observing that it was non-stationary, we applied a transformation to achieve stationarity. This transformation involved computing the returns using the formula $z_t = (y_t - y_{t-1})/y_{t-1}$ where $t = 2, ..., 1460$, which is considered as a first difference. Following this calculation, we successfully obtained a stationary time series, enabling us to proceed with our analysis.

5.2 Tentative Model Analysis



(a) Sample Autocorrelation Function

(b) Sample Partial Autocorrelation Function

Figure 4: Autocorrelation Functions

Subsequently, we analyzed the Sample Autocorrelation Function (SAC) plot (Figure 4(a)) and the Sample Partial Autocorrelation Function (PSAC) plot (Figure 4(b)). In Figure 4(a), we observe that the SAC cuts off after lag 1, while in Figure 4(b), the PSAC dies down. This suggests a tentative model of Moving Average of order 1, MA(1). To determine whether to include a constant term, δ, we examined the estimator for μ. We obtained that $\bar{z}$ was approximately 0.0003, very close to zero. However, we conducted a t-test to make a final decision. The null hypothesis was $H_0 : \mu = 0$, and the alternative hypothesis was $H_1 : \mu \neq 0$. We then calculated $t_0 = \bar{z}/(s_z/\sqrt{n-d+1}) = 0.000288/(0.00542/\sqrt{1460-2+1}) = 2.03$. Since t_0 is greater than 2, we rejected the null hypothesis, leading to the inclusion of the constant term. Therefore, our tentative model is expressed as: $z_t = 3 \times 10^{-4} + a_t - 0.0561a_{t-1}$.

5.3 Model Accuracy Diagnostics

To assess the accuracy of the model, we underwent several steps. First, as the model was already stationary, we needed to check for invertibility. The model satisfied invertibility conditions since $|\theta_1| = 0.0561 < 1$. Next, to determine the significance of the parameters in the tentative model, we conducted a t-test. The null hypothesis was $H_0 : \theta = 0$, and the alternative hypothesis was $H_1 : \theta \neq 0$. Given our earlier conclusion that the constant δ was significant, we focused the t-test on the θ_1 parameter. The calculated t-statistic was $t = \hat{\theta_1}/s_{\hat{\theta_1}} = 0.0561/0.0267 = 2.10$. With degrees of freedom well exceeding 30, we compared the t-statistic with $z_{0.025} = 1.96$. Since $|2.10| > 1.96$, we rejected H_0, indicating the significance of our θ_1 parameter. Finally, to evaluate the overall adequacy of the model, we employed the Ljung-Box Statistic. For this test, H_0 assumed that the model lacks autocorrelation, while H_1 suggested the presence of autocorrelation. The obtained Ljung-Box Statistic was $Q^* = 13.96$, and had a p-value of 0.1748. Given that the p-value exceeds 0.05, we concluded there is insufficient evidence to reject H_0, establishing the overall significance of the model.

5.4 Remarks for Nonseasonal Box-Jenkins Models

In conclusion, our finalized Nonseasonal Box-Jenkins Model, represented as $z_t = 3 \times 10^{-4} + a_t - 0.0561a_{t-1}$, effectively captures the dynamics of the time series. The model meets key assumptions, including stationarity and invertibility, and the significance of the estimated parameters, notably $\theta_1 = 0.0561$, was confirmed through t-tests. Overall, diagnostic tests, including the Ljung-Box Statistic, support the model's adequacy, making it a reliable tool for forecasting and understanding the time series dynamics.

6 Conclusion

6.1 Final Remarks

The methods we selected for forecasting –Multiple Linear Regression, Time Series Regression, and Nonseasonal Box-Jenkins Models– each revealed distinct insights.

Firstly, the initial Multiple Linear Regression encountered multicollinearity among 'Low,' 'Open,' and 'Close' variables, compromising the model. Various approaches were explored to mitigate this issue, eventually adopting Lasso Regression. This process ensured a more reliable model by reducing multicollinearity and minimizing the Mean Squared Error (MSE). The Lasso model, incorporating 'Low,' 'Open,' 'Close,' and 'Volume' proved effective in predicting the 'High' stock price, presenting a reliable solution for stock price forecasting.

Then, the Time Series Regression analysis showcased a strong model with an increasing trend in the 'High' stock price over time. Diagnostic assessments –including the Durbin-Watson test for autocorrelation, and the decomposition for trend and seasonality– affirmed a well-fitted model. These analyses provided comprehensive insights into the temporal behavior and dependencies within the data, highlighting the model's ability to capture trends and seasonal tendencies.

Finally, the Nonseasonal Box-Jenkins Models began with transforming a non-stationary time series into a stationary one. The analysis pointed to a tentative MA(1) model with an included constant term. Evaluating the model's accuracy through invertibility, parameter significance tests, and overall adequacy using the Ljung-Box Statistic, showcased its significance, affirming the model's suitability in examining autocorrelation and providing a final model representation.

Each forecasting method provided valuable insights and model improvements tailored to address specific challenges. Integrating the strengths of these methods could potentially enhance forecasting accuracy, leading to more reliable and comprehensive models in stock market prediction.

6.2 Future Work

Moving forward, avenues for further enhancement in forecasting models for stock prices emerge from these analyses. Integrating the strengths of multiple techniques –such as combining the predictive power of Multiple Linear Regression with the temporal understanding offered by Time Series Regression and Nonseasonal Box-Jenkins Models– could potentially yield more comprehensive and accurate predictions. Exploring ensemble methods that blend the outputs of various models or incorporating machine learning algorithms trained on a broader set of features could enrich forecasting capabilities. Additionally, refining the predictive models by incorporating external factors like market sentiments, economic indicators, or news sentiment analysis could offer a more holistic understanding of stock price fluctuations, thereby improving the accuracy of future forecasts.

Despite the strengths showcased in each approach, certain limitations should be acknowledged. For instance, in Multiple Linear Regression, the challenge of multicollinearity implied some concerns in model interpretation and required further work, leading to the selection of the Lasso Regression model. However, the reduction in variables might limit the model's ability to account all nuances in stock price prediction. Furthermore, while the Time Series Regression and Nonseasonal Box-Jenkins Models effectively accounted temporal dependencies and autocorrelation, they might overlook non-linear patterns or sudden market shifts that are frequent in stock markets, potentially affecting predictive accuracy. Acknowledging these limitations and pursuing potential improvements will be crucial in refining forecasting models, ensuring adaptability to the volatile market fluctuations.

References

[1] (2020) Time series forecasting with yahoo stock price (https://www.kaggle.com/datasets/arashnic/time-series-forecasting-with-yahoo-stock-price?select=yahoo_stock.csv). Accessed: 2023-10-30.

[2] (2023) A guide to regression analysis with time series data (https://www.influxdata.com/blog/guide-regression-analysis-time-series-data/). Accessed: 2023-10-30.

[3] (2022) Box-jenkins model: Definition, uses, timeframes, and forecasting (https://www.investopedia.com/terms/b/box-jenkins-model.asp). Accessed: 2023-10-30.