Science and Technology and Social Development Proceedings Series

Article

2024 International Conference on Business Economics, Education, Arts and Social Sciences (EASS 2024)

Forecasting Models for Apple Inc. Stock Price Using Regression Smoothing and Box Jenkins Time Series Analysis

Chenhao Jin [1,*]

[1] University of Waterloo, 200 University Ave W, Waterloo, ON, Canada
* Correspondence: Chenhao Jin, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada

Abstract: In financial markets, stock price forecasting plays a critical role in investment decision-making, especially for globally influential companies like Apple Inc. This study aims to develop and assess models for predicting Apple Inc.'s stock price using various approaches, including regression analysis, smoothing methods, and the Box-Jenkins methodology. We analyzed ten years of Apple Inc.'s historical adjusted closing price data to construct models such as unregularized regression, regularized regression (Ridge and Lasso), smoothing methods (including exponential smoothing and moving averages), and the Box-Jenkins (SARIMA) model. The dataset was divided into training and test sets, and the predictive performance of each model was evaluated using the Average Prediction Squared Error (APSE). The findings indicate that the Simple Exponential Smoothing model performed best for short-term predictions, with an APSE of 0.01455. The Box-Jenkins model achieved an APSE of 0.08270, unregularized regression 0.021, while Ridge and Lasso models yielded APSEs of 0.6 and 4.97, respectively. In summary, smoothing methods are well-suited for short-term forecasting, while the Box-Jenkins method offers greater stability but comes with added complexity. Investors should choose forecasting models based on their specific requirements. This study provides empirical evidence for stock price forecasting and contributes new insights into the application of financial analysis and data science techniques.

Keywords: stock price forecasting; time series analysis; regression models; smoothing techniques; Box-Jenkins (SARIMA) Model

Published: 09 November 2024

1. Introduction

Stock Market is a system where currency, stocks, equities, and other financial commodities are bought and sold by individuals [1]. Investors tend to buy stocks with the potential to rise in value, rather than stocks expected to decline in the future. Therefore, in the financial market, there is a strong demand to forecast stock prices using statistical analysis [2], deep learning, and machine learning techniques to maximize capital gains and minimize losses.

In this study, we analyze ten years of adjusted price data for Apple Inc. and propose appropriate models to forecast the future price movements of Apple's stock. This analysis can assist in portfolio management decisions, helping investors optimize their investment strategies for risk and return. Forecasting the stock price of a company as influential as Apple allows us to apply theoretical knowledge to real-world data, addressing the challenges of non-stationarity, market sentiment, and external economic factors that influence stock prices.

The primary objective of this study is to develop a reliable model or set of models that can accurately forecast the future stock price of Apple Inc. This study offers a valuable opportunity to apply statistical and analytical skills to tackle real-world problems, preparing students for careers in data science, financial analysis, and related fields. Time series of stock prices are crucial for the prediction of stock prices, and time series analysis remains the most widely used method for such forecasts [3]. In this study, we employ four models: unregularized regression, regularization methods, smoothing methods, and the Box-Jenkins approach.

Firstly, for the unregularized regression model, polynomial regression will be used to capture trend behavior, with mean square error (MSE) as the primary criterion. The assumptions of the selected model will be checked, and the next 12 monthly stock prices of Apple Inc. will be forecasted. Secondly, for the regularization methods, Ridge and Lasso models will be evaluated using cross-validation to find the model with the optimal degree, which minimizes the cross-validation (cv) value. APSE (Adjusted Prediction Squared Error) will be applied on the testing set to assess prediction performance. Thirdly, for the smoothing methods, various techniques such as moving averages, simple exponential smoothing, double exponential smoothing, and differencing will be applied to estimate or eliminate trend and seasonality. The best prediction model will be chosen based on APSE for the testing set. If necessary, residuals will be examined after detrending and deseasonalizing the data. Finally, for the Box-Jenkins method, since the data is retrieved from a financial institution [4], it is reasonable to assume that the data exhibits both trend and seasonality. The SARIMA model, based on the Box-Jenkins methodology, will be employed. First, the stationarity of the data will be determined. If stationary, ACF and PACF plots will be used to identify potential coefficients for the SARIMA model. If not, necessary transformations and differencing will be applied to achieve stationarity, after which ACF and PACF plots will be used to determine the coefficients. APSE will be applied to assess the prediction accuracy of the chosen models, and the final model will be selected based on its performance.

2. Exploratory Data Analysis

In this section, a preliminary data analysis will be conducted before building models. This will involve utilizing different plots, such as the autocorrelation function (ACF) plot, time series (ts) plot, and decompose plot, to analyze the data.

2.1. Data Description & Summary

The data for APPL's stock price can be accessed from Yahoo Finance. The dataset consists of seven columns: "Date", "Open", "High", "Low", "Close", "Adjusted Close Prices", and "Volume", with 123 observations in total. This project will primarily focus on the adjusted closing price (in dollars per share) (2014-02-01 to 2023-03-08).

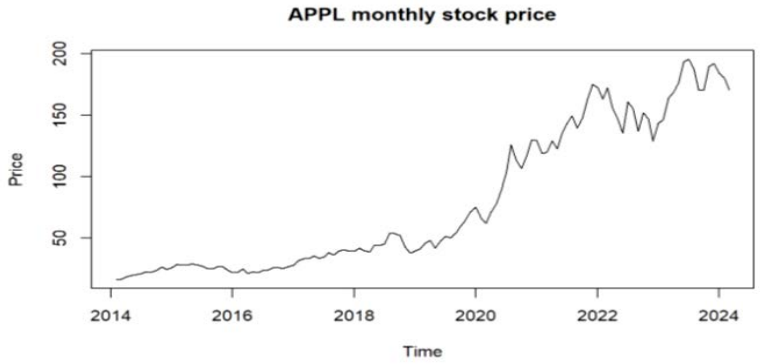

Figure 1. Time Series Plot (Original Data).

Based on Figure 1, it could be noticed that the variance is not constant over time. This can be seen from the sharp increase between the years 2020 to 2022 as compared to the more static data value between the years 2014 to 2020. Consequently, a log transformation will be applied to stabilize the variance to a certain extent before modelling. In addition, an overall increasing trend can be witnessed. Notably, there is a significant surge in stock price around 2020, which could be associated with market reactions to significant events, such as the COVID-19 pandemic.

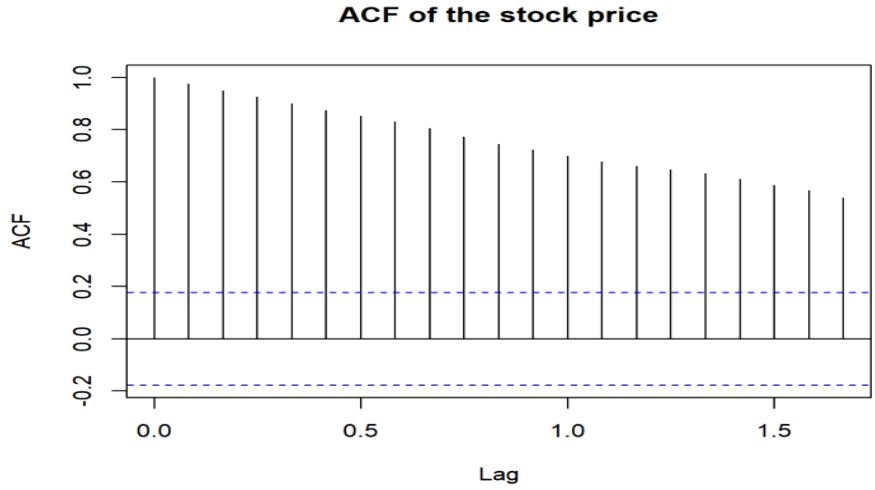

Figure 2. ACF Plot (Original).

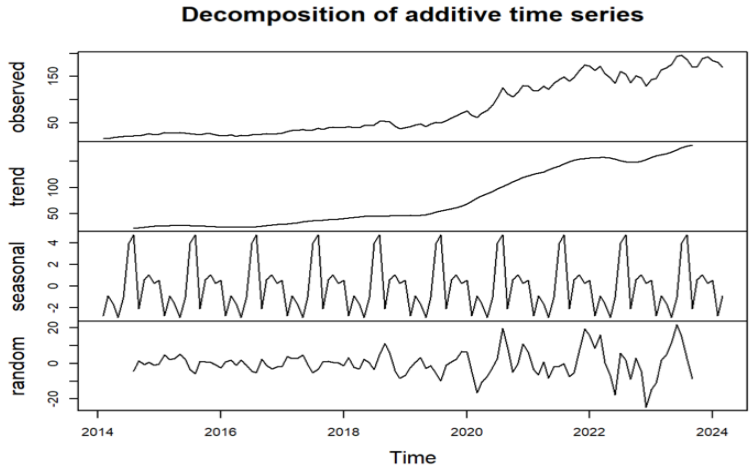

Figure 3. Decomposed Plot (Original).

Figure 2 displays a slow and linear decay, indicating a trend. However, no significant seasonal patterns can be noticed from the ACF plot. Figure 3 shows an upward trend in the trend component. Also, it is noticeable that the range of the seasonal component is relatively negligible compared to the observed and trend components ([-3,5] vs [10,160]). Therefore, it may be feasible to model the data without considering a seasonal pattern.

3. Modeling

In this section, models based on 4 different methods will be built and the prediction power of these 4 models will be assessed. The whole dataset will be divided into two sets, the training set consists of the first 8 years of original data while the test set consists of the last 2 years of original data. The model building will make use of the training set, while model validation is based on the test set.

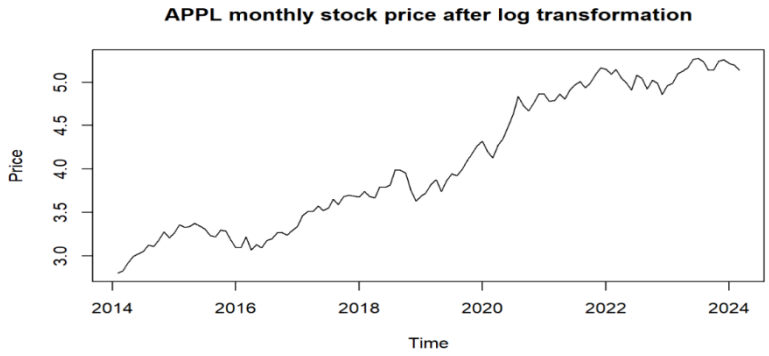

Figure 4. Time Series Plot (log-Transformed Data).

Figure 4 is the time series plot of log-transformed data.

3.1. Regression(Un-regularized)

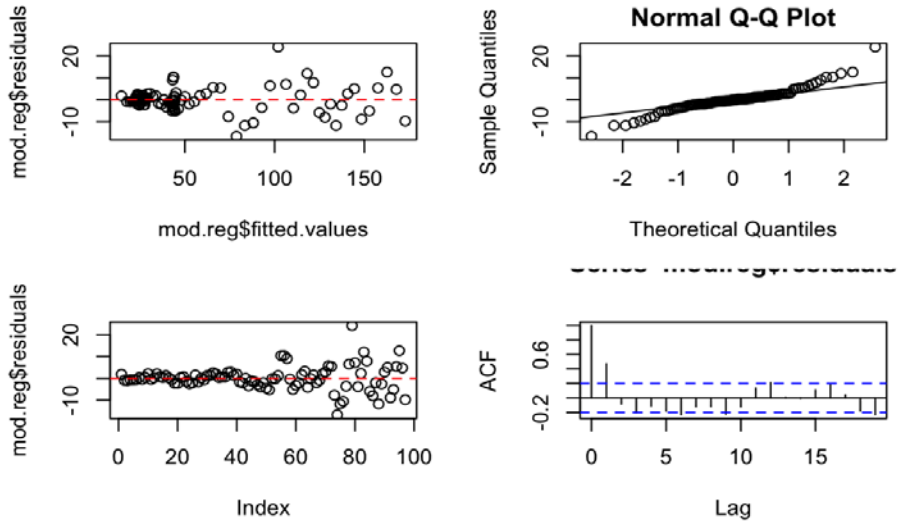

Figure 5. Assumptions of Un-regularized Regression.

A polynomial analysis with degree from 1 to 10 is conducted to find the best degree based on the criteria of mean square error (MSE). The result shows (see Appendix) that a poly- nomial with degree of 1 is the best among the given range. The seasonality does not occur in this regression, since the APSE of the trend only is smaller than the APSE of seasonality.

Figure 5 presents the verification of assumptions of the final regression model with degree of 1. Both graphs on the left indicate that the residuals are distributed around zero y-axis. It suggests that the zero-mean assumption is satisfied. However, the graph on the bottom left shows that the variation of the residuals increases as the index increases. Thus, the constant- variance assumption does not hold for the final model. As for the normally distributed assumption, the graph on the top right indicates that data at the middle lies around on the straight line with some off at the right tail. At last, we can discover that there are some correlations between variates on the ACF plot.To sum up, only the assumption of zero-mean is satisfied, while independence, normally distributed, and constant variance are violated.

3.2. Regression(regularized)

Regularization is an advanced method of regression. It adds a penalty term into the loss function of regression. Different from non regularized regression, multicollinearity will be solved under regularization. Regularization also helps reduce variance and could result in variable selection by the choice of penalty function. For the following operation, log-transformed data will be used to ensure data consistency and avoid non-constant variance. For the purpose of prediction, a cross-validation method will be used to find the model with the best degree, which has the smallest cross-validation value, for both Ridge and Lasso Model. Then, APSE on the testing set will be applied to test prediction power. For the regularization method, degree 1 to 15 was chosen.

The first step for finding the best degree is to find the optimal cross-validation value. The fitted ridge model will be built based on degree 11 and its associated lambda. The fitted lasso model will be built based on degree 9 and its associated lambda.

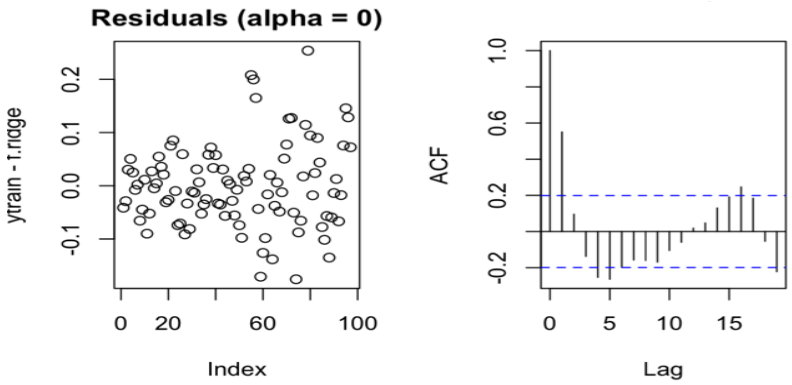

Figure 6. :Assumptions of Ridge.

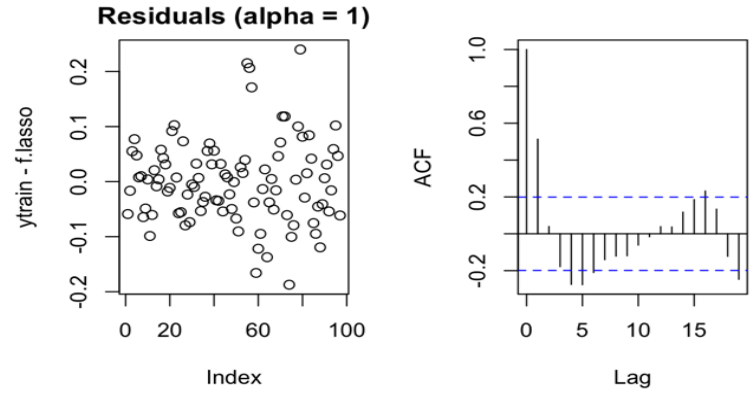

Figure 7. Assumptions of Lasso.

Next step is to perform a residual diagnosis. Figure 6 and 7 provides four graphs which could be used to assess the model assumption. The Residual vs Time indicates that it satisfies zero-mean assumption. However, there exists a damp-sine wave pattern in the ACF plot, showing the residuals are correlated. Therefore, not all model assumptions are satisfied with respect to the Ridge and Lasso model.

Compare Lasso and ridge regression:

As Table 1 presents, the ridge model with degree of 9 preserves a smaller APSE compared to lasso model with degree of 11. It suggests that under regularization, the ridge model with degree 9 has a stronger prediction power.

Table 1. APSEs of Ridge and Lasso models.

Model	Degree	APSE
Ridge	11	0.6
Lasso	9	4.97

3.3. Smoothing

3.3.1. The Best Smoothing Model for Prediction

Based on Figure 8, the simple exponential smoothing model has the smallest APSE, which is 0.01455135, thus, the model would be chosen as the best model for prediction in this section.

Method <chr>	APSE <dbl>
simple exponential smoothing	0.01455135
double exponential smoothing	0.11834105
additive HW method	0.07236182
multiplicative HW method	0.08346342

Figure 8. APSEs of all models.

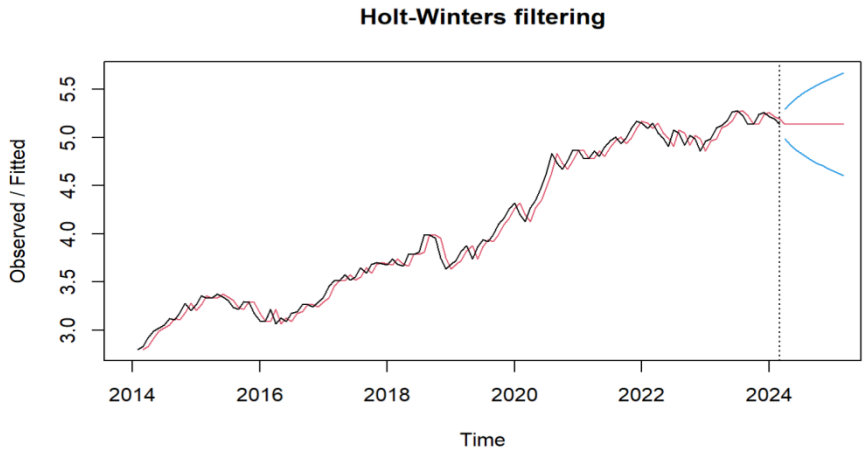

Figure 9. Simple Exponential Smoothing with Prediction.

In Figure 9, the black line is the data, the red line is the prediction, and the blue lines are the prediction interval. Under the simple exponential smoothing model, the prediction after 2024.03 would be stable.

3.3.2. Residuals of the Model Fitted to the Whole Data

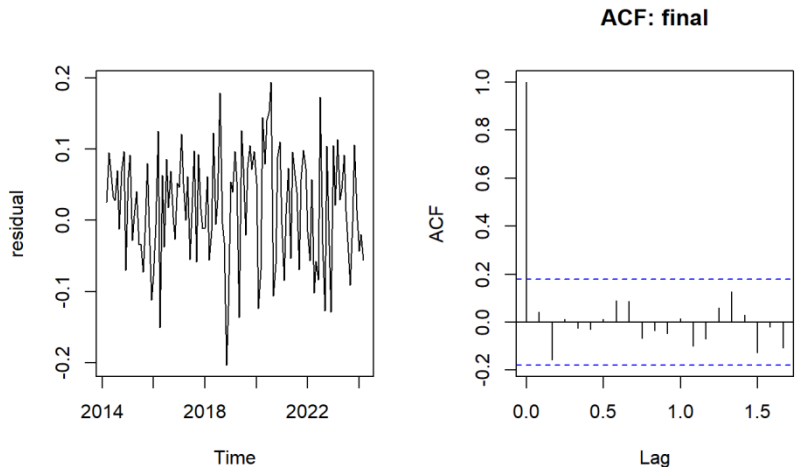

Figure 10. Residual Analysis.

The Time and Residual plot in Figure 10 shows that these residuals appear to fluctuate around a mean of zero without any clear trend or seasonality, and from the ACF there are no significant spikes and there is no slow decay, therefore, we can say it is stationary, and it is also white noise because all of the lags are inside of the blue section (except for lag0 = 1).

3.4. Box-Jenkins:

From the plot of the original data(Figure 1), we see the data clearly does not appear in constant variance, and since our ultimate goal is to have our data stationary via transformations and/or differencing, we first want to fix the non-constant variance before applying the Box-Jenkins Methodology. As a result, we will apply the power transformation on our data, the following graphs are the power transformation on our data using different values of alpha, ranging from -2 to 2:

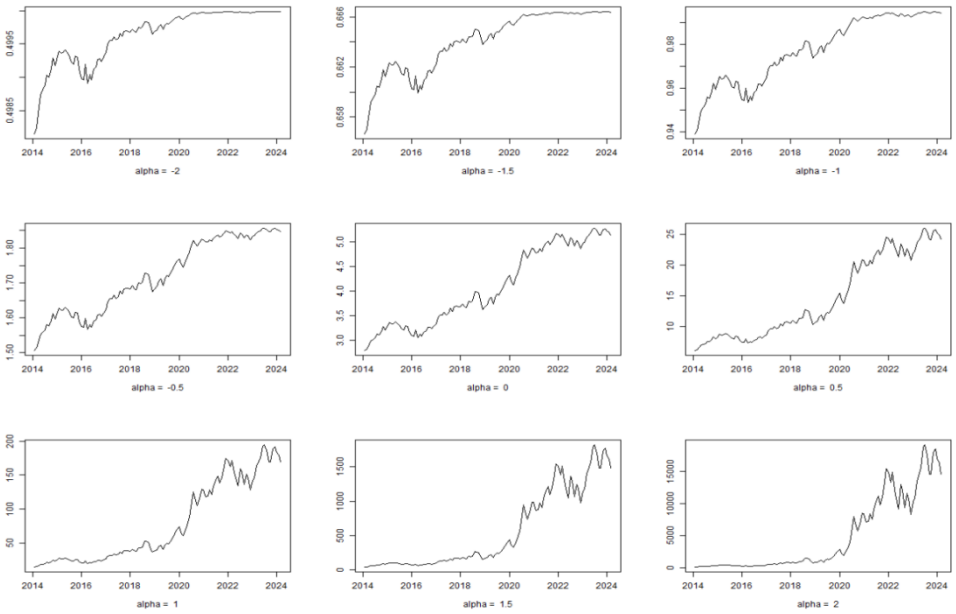

Figure 11. Power transformation of data with various alpha ranging from -2 to 2.

From the above Figure 11, we see one good choice of alpha would be alpha = 0, i.e., log transformation. Thus log transformation is applied to our data. A clear trend can still be noticed from the transformed data, which suggests the use of a regular differencing. Applying the regular differencing yields:

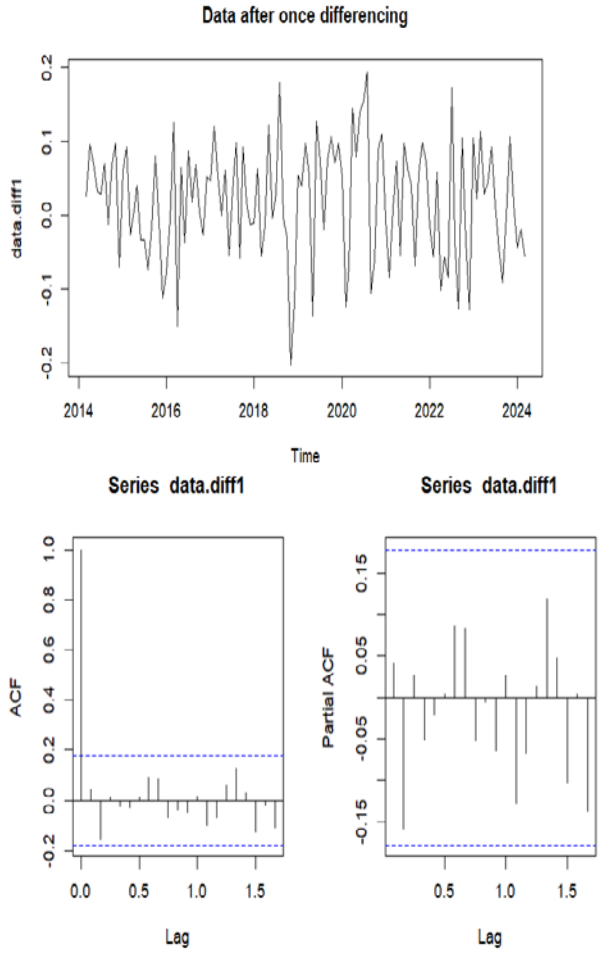

Figure 12. Plot, ACF and PACF plot of transformed data after once differencing.

As can be seen in the differenced plot, after one time differencing, the data is stationary with no clear trend or seasonality. From the ACF plot we can also see the data cuts off after lag 1, and there are no patterns of seasonality. Similarly, from the PACF plot, we see all the values are within the range of the dashed blue lines. The data now seemed stationary and is close to white noise after the first differencing, this means ∇X_t isstationary,and-wehavep=0,d=1,q=0,s=0forourARIMAmodel.

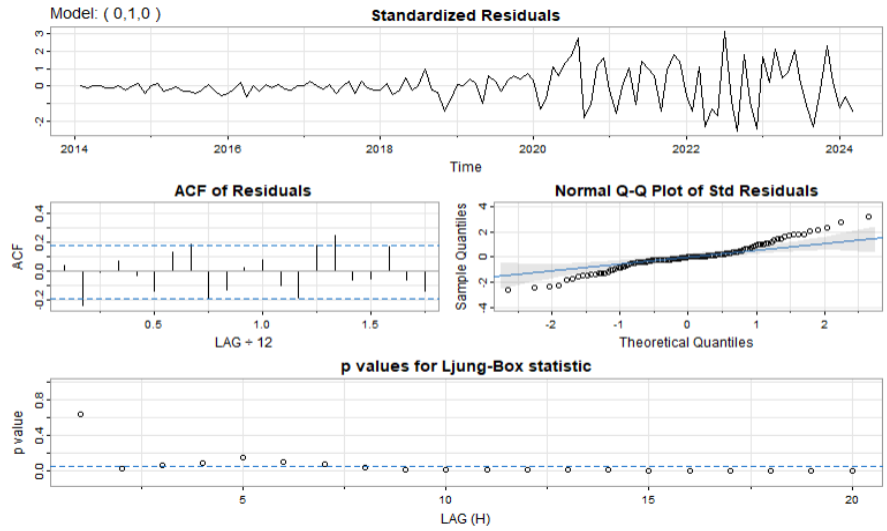

Figure 13. Model Diagnostic for ARIMA.

Then we predict using ARIMA model, and the next year's adjusted price for AAPL is predicted by the following:

And we have the plot for the predicted value from our ARIMA model, along with its prediction interval:

```
$pred
         Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep
2024                           172.0050 173.2800 174.5550 175.8300 177.1050 178.3800
2025 183.4799 184.7549 186.0299
         Oct      Nov      Dec
2024 179.6550 180.9299 182.2049
2025
```

Figure 14. Predicted value of ARIMA model.

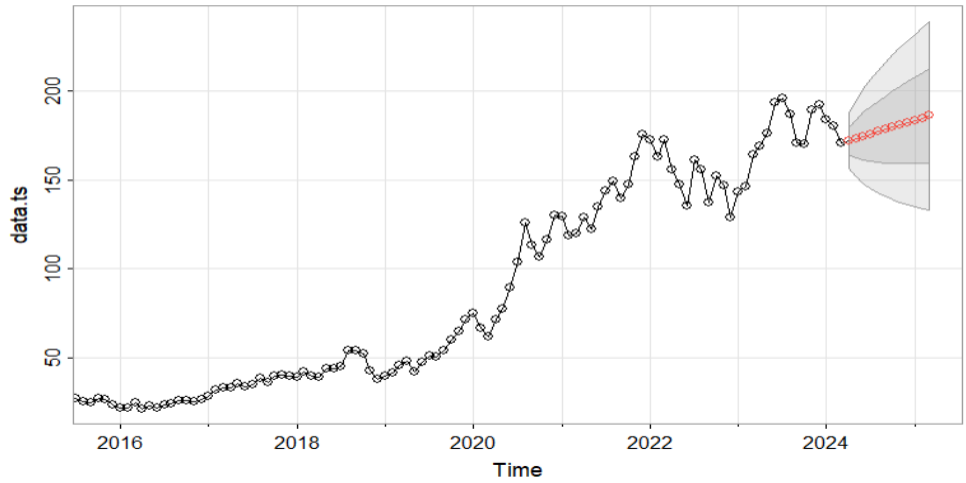

Figure 15. Predicted value and interval for AAPL data.

Our APSE for ARIMA model is 0.08269791.

4. Result

In this study, models constructed using four different methods were applied to predict the stock price of Apple Inc., and their predictive performance was evaluated using the Average Prediction Squared Error (APSE). The results are shown in Table 2:

Table 2. APSE Comparison of Different Models for Predicting Apple Inc.'s Stock Price.

Method	APSE
Unregularized Regression	0.021
Ridge	0.6
Lasso	4.97
Simple Exponential Smoothing	0.01455
Box-Jenkins	0.08270

From Table 2, it can be seen that the Simple Exponential Smoothing (SES) model performs best for short-term predictions, with an APSE of 0.01455, outperforming the other methods. The Box-Jenkins method has an APSE of 0.08270, demonstrating strong stability and being particularly suitable for data with trends and seasonal components. In contrast, the unregularized regression model has an APSE of 0.021, showing moderate performance, while the Ridge and Lasso regression models have APSEs of 0.6 and 4.97, respectively, indicating weaker predictive power.

Based on the model predictions, we estimate that Apple Inc.'s adjusted closing price will stabilize around $170.7304 over the next 12 months (see Figures 16 and 17), with a 95% confidence prediction interval. Although the SES model performs best for short-term forecasts, different models may be required for long-term predictions to improve accuracy.

Overall, the Simple Exponential Smoothing method is well-suited for short-term stock price forecasting, while the Box-Jenkins model is more effective when the data exhibits longer-term trends and seasonal variations. The comparison of APSE across different methods further validates the strengths and weaknesses of these models in various scenarios.

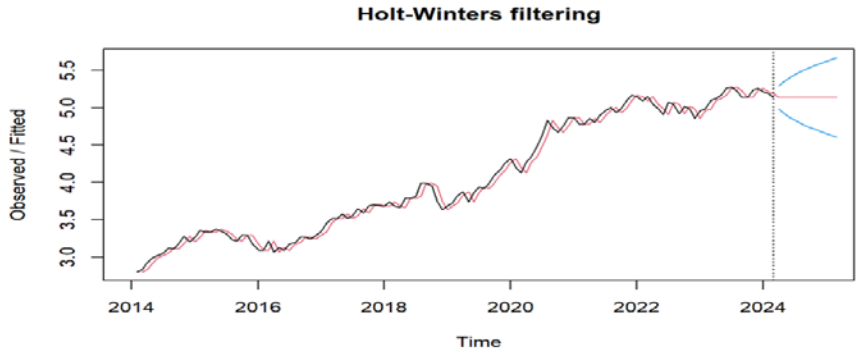

Figure 16. SES for Entire Data Figure.

	fit	upr	lwr
Apr 2024	170.7304	199.1057	146.3990
May 2024	170.7304	212.1972	137.3669
Jun 2024	170.7304	222.8236	130.8159
Jul 2024	170.7304	232.1945	125.5364
Aug 2024	170.7304	240.7765	121.0620
Sep 2024	170.7304	248.8078	117.1542
Oct 2024	170.7304	256.4297	113.6720
Nov 2024	170.7304	263.7336	110.5239
Dec 2024	170.7304	270.7830	107.6466
Jan 2025	170.7304	277.6237	104.9942
Feb 2025	170.7304	284.2904	102.5320
Mar 2025	170.7304	290.8099	100.2334

Figure 17. Pred. Values & Intervals.

5. Conclusion

This study developed and evaluated various models for predicting Apple Inc.'s stock price, including unregularized and regularized regression, smoothing techniques, and the Box-Jenkins (SARIMA) method. Among these, the Simple Exponential Smoothing (SES)

model exhibited the best short-term predictive performance with the lowest APSE of 0.01455. The Box-Jenkins model also showed strong stability with an APSE of 0.08270, making it particularly suitable for data with identifiable trends and seasonality. In contrast, the unregularized regression model yielded moderate results, while the Ridge and Lasso models demonstrated weaker predictive power.

Despite these promising findings, the study has certain limitations. While the SES model performs well for short-term forecasting, its accuracy may diminish for long-term predictions due to shifting market dynamics. Additionally, the complexity of the Box-Jenkins method, which requires substantial data preprocessing, could limit its practicality in time-sensitive applications. Furthermore, the exclusive reliance on historical data may overlook sudden market changes or external economic factors that could impact stock prices.

Future research should aim to improve long-term forecasting accuracy by incorporating more advanced machine learning techniques, such as neural networks or ensemble approaches. Additionally, integrating real-time data, including sentiment analysis and macroeconomic indicators, could enhance the robustness of predictive models. Such advancements would lead to more versatile and accurate forecasting tools, better equipped to address the complexities of modern financial markets.

References

1. Himanshu Gupta, Aditya Jaiswal; A study on Stock Forecasting Using Deep Learning and Statistical Models(2024)
2. R. Dias, P. Alexandre, P. Heliodoro, Contagion in the LAC financial markets: The impact of stock crises of 2008 and 2010. Littera Scripta, 13(1), 32-45 (2020)
3. Forecasting of Indian stock market using time-series ARIMA model 2014 2nd International Conference on Business and Information Management, ICBIM 2014 (2014), pp. 131-135, 10.1109/ICBIM.2014.6970973
4. "Apple Inc. (AAPL) Stock Historical Prices & Data." Yahoo! Finance, Yahoo!, 6 Apr. 2024, ca.finance.yahoo.com/quote/AAPL/history.
5. WARREN L. YOUNG; The Box-Jenkins approach to time series analysis and forecasting : principles and applications(1977)