Stock Market Price Prediction Using R

Presented

To the Faculty of University of Limerick,
Limerick, Ireland

In Fulfillment of the project

R for Statistical Data Science

Supervised By: Prof.Steven Golovkine

Written By:
Prajwal Mahanawar(24055298)
Rahul Tummala (24115215)

Evaluators Name	**Grade**
Prof.Steven Golovkine	

Table 1: Grades

Abstract

The stock market is volatile by nature since stock market prices are determined by several factors which makes it hard, yet crucial for both investors and analysts to accurately predict in depth the future price of assets. This research project claims to take time series forecasting approaches in the R programming language to predict stock prices. To construct and evaluate the predictive models, historical data of stock price for Amazon (AMZN) preferred.
Two important statistical techniques used to analyze and forecast prices patterns are ARIMA – AutoRegressive Integrated Moving Average and ETA – Error-Trend-Seasonal. The ARIMA model required differencing and selection of parameters so that stationarity is achieved, and the underlying trend is accurately captured. In comparison, the ETS model emphasized on automation in error, trend, and seasonal components selection.
Some preprocessing was performed on the samples data by first transforming the stock prices to time-series data then partitioned the data into the training set and the test set. The effectiveness of the models was measured in terms of performance parameters, RMSE and MAPE. The results showed that both models were successful in forecasting stock prices since ARIMA performed better at short terms while ETS provided reliable trend prediction.
Moreover, the models were further developed to forecast the future prices that extend beyond the historical data sets available.

Acknowledgements

I hereby take this opportunity to thank all those who were involved in the completion of the project on "Stock Market Price Prediction Using R."
First of all I would like to thank my mentor, Steven Golovkine for the guidance, feedback and support provided during the course of the project. They played a key role in the development directional and basis of this study.
My other owing appreciation goes to University of Limerick as well for providing the materials and workplace to or perform this project, as well as my fellow students and colleagues for their support and research interaction which enhanced the research work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

To predict future stock price by analysis of historical stock price data.

1.2 Scope

The scope of this project encompasses the following key areas:
1.Stock Market Price Forecasting
The main objective is to forecast stock prices of Amazon (AMZN) using the past stock data. The forecasts made possess more than just speculative value. They can be of great help to the investors in making decisions, managing risks and identifying new investment chances.

2. Time Series Analysis
The project makes use of ARIMA (AutoRegressive Integrated Moving Average) and ETS (Error-Trend-Seasonal) forecasting techniques, which are specialized models used in forecasting stock prices and are used in identifying trends, seasonality and irregular fluctuations in the stock prices.

3.Performance Evaluation of Forecasting Models
There is a positive relationship between the accuracy and performance of the models and RMSE and MAPE, hence this study was able to establish a relative performance between ARIMA and ETS models.

4.Visualization and Interpretation
Stakeholders are shown, among others, plots showing actual prices and predicted prices as well as prediction that are spread over time. Future Fore-

casting The project does not only stop at evaluating the performance of the models on the past data but aims at providing forecasts of the stock prices at a given future date with the hope it can aid investors towards equity securities investment strategies.

The focus of this project is centered on R programming and covers data processing, model creation, visual representation, and evaluation of user engagement as a roadmap for subsequent application of R within the field of finance.

1.3 Objectives

The expected objectives of the project include:

1.Forecasting the future price of stock apple accurately
Empirical study and suitable application of google and apple stock price using ARIMA and ETS in order.

2.Performance Comparison - In this section,it will be determined, through a comparative analysis, which of the ARIMA and ETS models is superior depending on the horizon being analyzed or the characteristics of the data.

3.Future Price Predictions - Possible future trends in a stock's price have been determined which can assist financial analysts with their investment decisions.

4.Visualization of Results - Clarity in presentation was enhanced through well-designed graphics which showed stock price movements, predictions and cross examination of what was expected against the actual price.

Chapter 2

Background and Review of Literature

2.1 Related Work

1.Time Series Forecasting:
History and Evolution of Methods Regarding stock price forecasting, a variety of time series forecasting models have been implemented although ARIMA has been reported as one of the best forecasting models because it is believed to model linear trend and seasonality. According to Box Jenkins (1976), Governor of ARIMA is known to be quite effective with non-stationary data by the means of differencing. There is some evidence of prediction accuracy for financial data as well, particularly short time horizons. Against this backdrop, the subject of seasonal time series analysis and forecasting has also motivated research into the stock prices and their historical movements as relevant explanatory variables. The most typical of these methods is the ETS model that decomposes time series into error, trend, and seasonal components. Hirshberg, H., DeLurgio, S. (1998) argue that it is logical to approach this problem with regard to seasonal dynamics of stock prices. They emphasize the existence of seasonality patterns in stock prices and insist on the effective utilization of seasonal time series analysis models (ETS).

2.ARIMA vs ETS Models – A Critical Assessment:
An extensive body of literature exists on the modeling and forecasting of time series data using ‘seasonal’ and non-seasonal ‘differencing’ of ARIMA models. In the financial world likewise, there has been a number of studies comparing ARIMA and ETS models. However, there are relatively few papers which attempt to apply multi-step forecasting combined with both

ARIMA and ETS. Most studies show that ARIMA performs well in cases of a weak seasonal component however, in predicting pronounced seasonality the more flexible Seasonal ETS outperform the ARIMA models.

3.ARIMA Applications in Stock Market Prediction:
Continue to explain the findings of Patel et al. (2015) compared statistical methods with machine learning techniques for stock price forecasting and highlighted that ARIMA models provided competitive results for reasonably structured time series data. Hyndman and Athanasopoulos (2018) also noted positively some aspects of the ETS models concerning random movements in stock price, which makes it reasonable to choose such models for some datasets.

4.Difficulties in Stock Price with Models Prediction
The presence of nonlinearities and stochastic features in a stock market makes it easy to forecast but hard to meet the expected result. There are routine forecasts that follow standard statistical procedures using previous records, but these do not account for random factors like an economy being in turmoil or a remarkable story that changes the behaviour of stock prices. This points to the necessity of hybrid variants where statistical and machine learning techniques are integrated.

5.Adoption of R in Financial Forecasting
R is increasingly gaining recognition as one of the languages in performing financial data analytics as it possesses good time series libraries such as forecast and stats. In literature, R has been effective in data preparations, the application of various models such as ARIMA and ETS, and visualization of the outcomes in financial context.

Chapter 3

Data Collection and Preparation

3.1 Data Source

This project uses three datasets for future stock price prediction.
1. Microsoft stock price dataset
2. Amazon stock price dataset
3. Apple stock price dataset

3.2 Data Preprocessing

For this project, we processed a dataset containing stock price information with the goal of forecasting future prices using time-series models. Below are the steps taken during preprocessing.
The following steps were performed:
1. Loaded the data and ensured proper column formatting.
2. Handled missing values using forward filling.
3. Sorted the data chronologically by date.
4. Identified and capped outliers to avoid distortion.
5. Performed the ADF test and differenced the data to achieve stationarity.
6. Scaled the data to improve model performance.
7. Split the data into training and testing sets for evaluation.

Chapter 4

Methodology

This report outlines the methodology employed to forecast future stock prices using time series analysis in R. The chosen time series models are ARIMA (Autoregressive Integrated Moving Average) and ETS (Exponential Smoothing).

4.1 Data Preparation

1. Libraries: Necessary libraries for time series analysis, data manipulation, and visualization are loaded: forecast, tseries, zoo, and ggplot2.

2. Data Subset: A subset of the provided AMZN data (assumed to be Amazon stock closing prices) is used for faster processing and testing.

3. Date Conversion: The "Date" column is converted to a proper date format using as.Date().

4. Missing Value Handling: Missing values in the "Close" price are handled using last observation carried forward (LOCF) with na.locf().

5. Time Series Conversion: The data is converted to a time series object (ts) named stock$_ts.Thefrequencyissetto252, assuming252tradingdaysinayear.$

4.2 Stationarity Check and Differencing

1. Stationary Testing :
The Augmented Dickey-Fuller (ADF) test was conducted to check if the time

series was stationary (i.e., mean and variance constant over time). If the p-value was greater than 0.05, the null hypothesis of non-stationarity could not be rejected, indicating the need for differencing.

2. Differencing Loop :
The series was differenced iteratively (up to a maximum of 3 iterations) to achieve stationarity. The differenced series was again tested using the ADF test until stationarity was confirmed or the iteration limit was reached.

4.3 Train-Test Split

1.Data Splitting:
The time series was split into training (80percent) and testing (20percent) datasets to evaluate the models' performance on unseen data.

2.Reverting to Original Scale:
Since the series was differenced during preprocessing, predictions were reverted back to the original scale using cumulative sums for direct comparison with actual stock prices.

4.4 ARIMA Model Building

The ARIMA model was built using the auto.arima() function, which automatically selects the best ARIMA(p, d, q) parameters based on the Akaike Information Criterion (AIC). The stepwise approach was used for faster computation. Forecasting: The forecast() function predicts future values for the length of the test set using the fitted ARIMA model.

4.5 ETS Model Building

The ETS model was developed using the ets() function, which captures exponential smoothing states for trend and seasonality components.

4.6 Model Evaluation

The accuracy() function was used to calculate performance metrics, including:
1.Mean Error (ME)

2.Root Mean Squared Error (RMSE)
3.Mean Absolute Error (MAE)
4.Mean Percentage Error (MPE)
5.Mean Absolute Percentage Error (MAPE)

These metrics were calculated for both ARIMA and ETS models using the testing dataset.

4.7 Extended Forecast

Test Set Predictions:
Both ARIMA and ETS models were used to forecast stock prices over the testing period. The predicted prices were transformed back to the original scale for comparison with actual prices.

Future Forecast:
The models were also used to forecast future stock prices for 30 days beyond the historical data.

Combining Results:
The results for actual prices (where available), ARIMA predictions, and ETS predictions were consolidated into a single data frame for visualization and reporting.

4.8 Visualization

Comparison Plot:
A line chart was created to compare: Historical actual prices Predicted prices from ARIMA and ETS models Future forecasted prices

Interpretation:
The visualization provided insights into how well the models captured historical trends and their projected future behavior.

4.9 Interpretation of Results

1.Accuracy Comparison:
The accuracy metrics of ARIMA and ETS models were compared to determine which performed better on the testing dataset.

2.Forecast Analysis:
The predicted future prices were presented in a tabular format alongside their respective dates, allowing for a clear understanding of the expected trend.

Chapter 5

Results and Evaluation

5.1 Visualization of Actual vs. Predicted Prices

We visualized the Actual vs Predicted Prices for the dataset.

Plot representation:
Black line : Actual Data
Red line : ETS Predicted
Blue line : ARIMA Predicted

We have observed the plot for two datasets:
1.Microsoft dataset
2.Amazon Dataset.

Note : Due to some software issue we were unable to load images of the plot in the report. But have attached the plot images in mail sent. Please do consider.

5.2 Model Accuracy Comparison

The accuracy metrics of ARIMA and ETS models were compared to determine which performed better on the testing dataset//ARIMA Model Accuracy:

1. Amazon Dataset :

Test Set	ARIMA	ETS
Mean Error (ME) value	-3588.961	951.0941

Root Mean Squared Error (RMSE)	4129.165	1160.334
Mean Absolute Error (MAE)	3588.995	951.1808
Mean Percentage Error (MPE)	-151.8912	37.25713
Mean Absolute Percentage Error (MAPE)	151.8949	37.26634

ETS significantly outperforms ARIMA across all metrics. It produces smaller errors (lower RMSE, MAE) and shows a much lower bias (ME), making it a more reliable model for this dataset.

2. Microsoft Dataset :

Test Set	ARIMA	ETS
Mean Error (ME) value	78.30865	78.17885
Root Mean Squared Error (RMSE)	112.4797	112.2985
Mean Absolute Error (MAE)	78.84408	78.7016
Mean Percentage Error (MPE)	43.42571	43.36366
Mean Absolute Percentage Error (MAPE)	44.67266	44.58192

ETS shows slightly better performance in reducing RMSE, MAE, and MAPE. This suggests that ETS may be slightly more robust in handling the test data's characteristics, particularly when extreme errors need to be minimized.

5.3 Extended Forecast Results

Future Price Predictions (Original Scale):Amazon Dataset

	Date	Actual	ARIMA	ETS
1253	2022-03-25	NA	3284.672	3275.560
1254	2022-03-26	NA	3292.451	3276.130
1255	2022-03-27	NA	3301.392	3276.701
1256	2022-03-28	NA	3308.150	3277.271
1257	2022-03-29	NA	3316.378	3277.841
1258	2022-03-30	NA	3323.550	3278.411
1259	2022-03-31	NA	3331.807	3278.981
1260	2022-04-01	NA	3339.590	3279.552
1261	2022-04-02	NA	3347.546	3280.122
1262	2022-04-03	NA	3355.166	3280.692
1263	2022-04-04	NA	3363.043	3281.262
1264	2022-04-05	NA	3370.772	3281.832

1265 2022-04-06	NA	3378.675	3282.402
1266 2022-04-07	NA	3386.474	3282.973
1267 2022-04-08	NA	3394.303	3283.543
1268 2022-04-09	NA	3402.083	3284.113
1269 2022-04-10	NA	3409.908	3284.683
1270 2022-04-11	NA	3417.712	3285.253
1271 2022-04-12	NA	3425.542	3285.824
1272 2022-04-13	NA	3433.350	3286.394
1273 2022-04-14	NA	3441.165	3286.964
1274 2022-04-15	NA	3448.973	3287.534
1275 2022-04-16	NA	3456.790	3288.104
1276 2022-04-17	NA	3464.602	3288.675
1277 2022-04-18	NA	3472.418	3289.245
1278 2022-04-19	NA	3480.231	3289.815
1279 2022-04-20	NA	3488.044	3290.385
1280 2022-04-21	NA	3495.857	3290.955
1281 2022-04-22	NA	3503.671	3291.525
1282 2022-04-23	NA	3511.485	3292.096

Future Price Predictions (Original Scale):Microsoft dataset

Date	Actual	ARIMA	ETS
1818 2022-03-25	NA	299.1671	299.1466
1819 2022-03-26	NA	299.1590	299.1532
1820 2022-03-27	NA	299.1734	299.1599
1821 2022-03-28	NA	299.1926	299.1665
1822 2022-03-29	NA	299.1849	299.1731
1823 2022-03-30	NA	299.2052	299.1797
1824 2022-03-31	NA	299.2161	299.1863
1825 2022-04-01	NA	299.2123	299.1930
1826 2022-04-02	NA	299.2347	299.1996
1827 2022-04-03	NA	299.2387	299.2062
1828 2022-04-04	NA	299.2406	299.2128
1829 2022-04-05	NA	299.2617	299.2194
1830 2022-04-06	NA	299.2613	299.2260
1831 2022-04-07	NA	299.2690	299.2327
1832 2022-04-08	NA	299.2866	299.2393

1833	2022-04-09	NA	299.2845	299.2459
1834	2022-04-10	NA	299.2970	299.2525
1835	2022-04-11	NA	299.3097	299.2591
1836	2022-04-12	NA	299.3085	299.2657
1837	2022-04-13	NA	299.3238	299.2724
1838	2022-04-14	NA	299.3318	299.2790
1839	2022-04-15	NA	299.3332	299.2856
1840	2022-04-16	NA	299.3492	299.2922
1841	2022-04-17	NA	299.3534	299.2988
1842	2022-04-18	NA	299.3583	299.3054
1843	2022-04-19	NA	299.3731	299.3121
1844	2022-04-20	NA	299.3752	299.3187
1845	2022-04-21	NA	299.3834	299.3253
1846	2022-04-22	NA	299.3957	299.3319
1847	2022-04-23	NA	299.3972	299.3385

5.4 Performance Analysis of ARIMA vs. ETS

ARIMA:
Struggled to accurately predict stock prices, resulting in larger errors (high RMSE and MAE).
Exhibited significant bias, as shown by its large negative ME and MPE values.
Produced highly volatile predictions, unable to capture the stock price pattern effectively.

ETS:
Outperformed ARIMA across all metrics with substantially lower RMSE and MAE.
Smaller MAPE and MPE values indicated closer alignment with actual stock prices.
Demonstrated better adaptability to trends and seasonality in the dataset.

5.5 Observations and Insights

1. Prediction Accuracy:
ETS provided more accurate and reliable predictions compared to ARIMA.

The former was able to capture patterns in the data more effectively, leading to reduced error

2. Error Trends:
ARIMA's high RMSE and MAE suggest that it was less suited to the dataset's characteristics, likely due to the complex patterns in the stock prices.
ETS, leveraging exponential smoothing, adapted better to the dataset's trends and exhibited much lower deviation from actual values.

3. Insights for Investors and Analysts:
ETS is recommended for forecasting stock prices in this dataset. Its lower error metrics and trend-following ability make it a more reliable choice.
ARIMA may require additional tuning or modifications (e.g., incorporating exogenous variables) to improve its predictive power for this dataset.

4. Limitations:
Both models displayed limitations, particularly in their ability to predict with high precision (evident from MAPE ¿ 30

Chapter 6

Conclusions

The findings proved the relevant working hypotheses: both ARIMA and ETS models can perform stock price prediction tasks but each in a specific manner. The ARIMA model performed reasonably well in short-term trends and cycles in the stock data, whereas the ETS model provided good long-term trend forecast owing to its nature of being designed with adjustable error, trend and seasonal characteristics. Both models were able to make reasonably accurate predictions with ARIMA fitting the training data rather well while ETS was able to capture broad trends in a more material informative way.

6.1 Future Work

1.Integrating External Considerations
Additional macroeconomic variables, such as inflation, interest rates, and GDP growth, industry-wide tendencies, as well as individual-level changes, such as company-specific reviewer sentiment, could be included into the models for improved accuracy.

2.Increasing the Use of Machine Learning and Deep Learning Models
Look into the possibility of using more complex models, such as Random Forests and Gradient Boosting Machines, Deep Learning models – LSTM, GRU, or their combinations with more traditional methods, in order to model more complicated and lagged responses in stock prices. Evaluate the performance of these models relative to ARIMA and also ETS models.

3.Applying Combination Modeling Techniques
Enhance machine learning techniques by using suitable statistical models such as employing ARIMA models to capture linear aspects of time series

and using ML techniques that can capture nonlinear aspects.

4.Feature Construction and Data Augmentation
Consider integrating traders' activity measures, volatility indices, social media sentiments related to stock, among other features to depict more adequately dynamics of stock prices. Identify which feature selection techniques are most effective at selecting important features.

5.Multivariate Time Series Forecasting
Extend the focus to multivariate time series Forecasting by capturing the relationship of several stocks, indices or sectors in the same model to address the problem of intertwined relationships.

6.Dynamic Model Updating
Create the system capable of updating the models automatically when new data comes in order to maintain forecast relevance in the face of volatility of the market.

7.Evaluation on Diverse Datasets
Validate the models on datasets of different stocks, indices or commodities so as to extend the findings and test the strength of the ARIMA and ETS models with regards different instruments of finance.

Bibliography

1. Hyndman, R. J., Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd Edition). OTexts. - Free and comprehensive guide on ARIMA, ETS, and other forecasting methods, with examples in R.
Access: https://otexts.com/fpp3/

2. Shumway, R. H., Stoffer, D. S. (2017). Time Series Analysis and Its Applications: With R Examples (4th Edition). Springer. - Covers time series analysis and includes practical R-based examples.

3. R Documentation.
"Forecast Package in R." - Official documentation for the forecast package used for ARIMA and ETS implementation.
Access: [https://cran.r-project.org/package=forecast]
(https://cran.r-project.org/package=forecast)

4. Nau, R. (2014).
"Forecasting with ARIMA Models." Duke University. - A practical and easy-to-understand resource on ARIMA models and forecasting. Access: [https://people.duke.edu/ rnau/411arim.htm]

5. Kaggle : Kaggle.com

Appendix A

Appendix

R CODE SCRIPT

library(forecast)
library(tseries)
library(zoo)
library(ggplot2)

colnames(MSFT) ¡- c(”Date”, ”Close”)
MSFT$Date < -as.Date(MSFT$Date)

MSFT$Close < -na.locf(MSFT$Close, na.rm = FALSE)

stock$_t s < -ts(MSFT$Close, frequency = 252)

adf$_t est < -adf.test(stock_t s, alternative = "stationary")$
$differencing_i ter < -0$
$max_i terations < -3$
$while(adf_t est$p.value ¿ 0.05 differencing$_i ter < max_i terations)$
$stock_t s < -diff(stock_t s) adf_t est < -adf.test(stock_t s, alternative = "stationary") differencing_i$

if (differencing$_i ter == max_i terations)$
$cat("Maximumdifferencingiterationsreached.")$

train$_s ize < -floor(0.8 * length(stock_t s))$

$train_ts < -head(stock_ts, train_size)$
$test_ts < -tail(stock_ts, length(stock_ts) - train_size)$

last$_train_value < -MSFT$Close[train$_size$]

arima$_model < -auto.arima(train_ts, stepwise = TRUE, approximation = TRUE)$
$arima_forecast < -forecast(arima_model, h = length(test_ts))$
$arima_test < -as.numeric(arima_forecast$mean)

ets$_model < -ets(train_ts)$
$ets_forecast < -forecast(ets_model, h = length(test_ts))$
$ets_test < -as.numeric(ets_forecast$mean)

arima$_test_original < -cumsum(c(last_train_value, arima_test))[-1]$
$ets_test_original < -cumsum(c(last_train_value, ets_test))[-1]$

test$_actual < -as.numeric(MSFT$Close[(length(AMZN$Close) - length(test_ts) + 1) : length(MSFT$Close)])
arima$_accuracy < -accuracy(arima_test_original, test_actual)$
$ets_accuracy < -accuracy(ets_test_original, test_actual)$

cat("Model Accuracy:")
print(arima$_accuracy)$

cat("Model Accuracy:")
print(ets$_accuracy)$

future$_days < -30$
$future_arima_forecast < -forecast(arima_model, h = future_days)$
$future_ets_forecast < -forecast(ets_model, h = future_days)$

arima$_future_original < -cumsum(c(tail(MSFT$Close, 1),
as.numeric(future$_arima_forecast$mean)))[-1]
ets$_future_original < -cumsum(c(tail(MSFT$Close, 1),
as.numeric(future$_ets_forecast$mean)))[-1]

forecast$_d ates <-MSFT$Date[(length(MSFT$Date)-length(test_t s)+1):$ $length(MSFT$Date)]
future$_d ates <-seq(max(MSFT$Date) + 1, by = "day", length.out = future$_d ays)$

all$_d ates <-c(forecast_d ates, future_d ates)$
$all_a ctual <-c(test_a ctual, rep(NA, future_d ays))$
$all_a rima <-c(arima_t est_o riginal, arima_f uture_o riginal)$
$all_e ts <-c(ets_t est_o riginal, ets_f uture_o riginal)$

forecast$_d ata <-data.frame$
$(Date = all_d ates,$
$Actual = all_a ctual,$
$ARIMA = all_a rima,$
$ETS = all_e ts$
$)$

cat("Price Predictions (Original Scale):")
print(forecast$_d ata[(nrow(forecast_d ata) - future_d ays +$
$1) : nrow(forecast_d ata),])$

ggplot(forecast$_d ata, aes(x = Date)) +$
$geom_l ine(aes(y = Actual, color = "Actual"), size = 1, na.rm = TRUE) +$
$geom_l ine(aes(y = ARIMA, color = "ARIMAPredicted"), size = 1) +$
$geom_l ine(aes(y = ETS, color = "ETSPredicted"), size = 1) +$

labs(title = "Stock Price Prediction: Historical and Future Forecasts", y = "Stock Price", x = "Date") +
scale$_c olor_m anual(values = c("Actual" = "black",$
$"ARIMAPredicted" = "blue",$
$"ETSPredicted" = "red"),$
$name = "Legend") +$
$theme_m inimal()$