APPLICATIONS OF MACHINE LEARNING IN STOCK MARKET

Thesis Submitted for the Partial Fulfilment of the Award of the Degree of
Master of Technology
in
Industrial Mathematics and Scientific Computing
by

Doyel Bag
MA21M011

Under the Supervision of

Prof. S. Sundar
Dr. Sri vallabha Deevi (Tiger Analytics)

Department of Mathematics,
Indian Institute of Technology Madras,
Tamil Nadu, 600036, India.
May 2023

INDIAN INSTITUTE OF TECHNOLOGY MADRAS
DEPARTMENT OF MATHEMATICS

NAME: Doyel Bag
Roll No: MA21M011
PROGRAMME: M.Tech.
DATE OF JOINING: 28.07.2021
Guide : Prof. Dr. S. Sundar
Co-guide Dr. Sri Vallabha Deevi

Certificate

This is to certify that the report **”Applications of Machine Learning In Stock Market”** , submitted by **Doyel Bag (MA21M011)**, in partial fulfillment of requirement for the award of the degree of Master of Technology in Industrial Mathematics and Scientific Computing, Indian Institute of Technology Madras, is the record of the work done by her during the academic year 2022-2023 in the Department of Mathematics, IIT Madras, India, under my supervision.

Prof. Dr. S. Sundar
Department of Mathematics
IIT Madras

Dr. Sri Vallabha Deevi
Tiger Analytics
Chennai, India

Acknowledgement

I would like to thank Prof. Dr. S. Sundar and Dr. Sri Vallabha Deevi for their constant help and support, without which the completion of my project for this academic year would be impossible. They have always been there to guide me throughout the course and have patiently solved all the queries associated with the project. Lastly, I would like to thank the Department of Mathematics, IIT Madras and Tiger Analytics for providing me the opportunity to work in this exciting field.

Doyel Bag
MA21M011
Department of Mathematics
IIT Madras

Abstract

The project aims to develop and compare various time series models and deep learning models for forecasting future stock prices and perform a comprehensive analysis of a portfolio of financial assets, such as stocks, bonds, and commodities, to evaluate its performance and identify opportunities for optimization. The models will be trained on historical stock price data and evaluated based on their accuracy in predicting future prices. The project will involve data preprocessing, feature selection, and model training and validation. Different types of time series models will be implemented, such as AR, MA, ARMA, ARIMA, Seasonal ARIMA (SARIMA), LSTM and their performance will be compared using standard performance metrics such as Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error. The project will also involve applying the best-performing model to forecast future stock prices for a given time horizon. The project will provide insights into the relative strengths and weaknesses of different time series models for forecasting stock prices and demonstrate the feasibility of using such models for investment decision-making. Furthermore, the project will employ portfolio optimization techniques to construct an optimal portfolio that maximizes returns for a given level of risk.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Time series forecasting and portfolio analysis are two critical steps in finance and investment management. Predicting the future movement of stock prices is a difficult task for researchers. Although proponents of the efficient market hypothesis believe that it is impossible to design a forecasting framework that can accurately predict stock price movement , significant literature has shown that the seemingly random movement of the time series can be accurately predicted[44]. Designing such predictive models requires the selection of appropriate variables, appropriate variable transformation methods, and tuning of model parameters. In this work, we propose a very robust and accurate stock price forecasting framework that consists of a set of statistical, machine learning and deep learning models. We use daily stock price data collected at five-minute intervals for well-known companies listed on the National Stock Exchange of India. The granular data is aggregated into three time periods of the day, and the aggregated data is used to build and train the forecast model. We try to build models to predict future stock price using a combination of statistical, machine learning, and deep learning methods in data on stock prices and perform portfolio analysis. This efficient learning leads to a very robust training of models for short-term stock price prediction and stock movement model predictions. We built seven models based on statistical and machine learning methods. In addition to these models, a deep learning regression model using a long-short-term memory (LSTM) network was constructed. Detailed results are provided for the performance of these models, and a rigorous analysis of the results is performed. The goal of a portfolio analysis is to provide actionable insights and recommendations to help investors optimize their portfolio and achieve their financial goals.

Chapter 2

Time Series

2.1 Time Series Data

Data collected at regular intervals is called time series data. It consists of a series of observations or measurements taken at successive points in time. Examples of time series data include daily temperature measurements, stock prices, weekly or monthly sales figures, and quarterly economic indicators[27].

Analyzing time series data provides insight into trends, patterns, and relationships that may not be evident in cross-sectional data. Time series data can be analyzed using a variety of techniques, including statistical models, machine learning algorithms, and data visualization tools.

Some common characteristics of time series data include seasonality, trending, and auto-correction. Seasonality is a recurring pattern that occurs regularly, such as an increase in sales of during the holidays. Generally a trend is the direction in which a time series moves over time, such as the upward trend in a stock price over several years. The Auto-correlation refers to the correlation between observations at different points in time, for example a positive correlation between a company's sales figures for several consecutive quarters.

Time series analysis[31] is a statistical technique used to analyze and interpret data that varies over time. It is a study of a series of data points ordered sequentially and identifies patterns, trends, and seasonality in the data.

There are many methods of time series analysis, including statistical methods such as autoregressive integrated moving average (ARIMA)[62] models, exponential smoothing models[26], and spectral analysis. These methods can be used to forecast future values, identify outliers and anomalies, and model the underlying relationships between variables. Time series analysis has applications in a wide range of fields, including finance, economics, engineering, and environmental science.

2.2 Types of Time Series Data

A time series can be Univariate, Bivariate, or Multivariate as shown in Figure2.1.

Univariate Time Series: Univariate time series analysis[13] is a statistical method that is used to analyze and understand the behavior of a single time series variable over time. This type of analysis is commonly used in fields such as finance, economics, and meteorology, where time series data is frequently collected and studied.

The analysis of univariate time series involves studying the patterns and trends in the data, identifying any seasonality or cyclicality, and modeling the data to forecast future values. This is typically done through techniques such as time series decomposition, autocorrelation analysis, and ARIMA modeling.

In general, the goal of univariate time series analysis is to identify and understand the underlying factors that contribute to the behavior of the variable over time. This can help researchers and analysts make predictions about future values of the variable and inform decisions related to investment, planning, and risk management.

Examples of univariate time series include stock prices, exchange rates, temperature readings, and sales data.

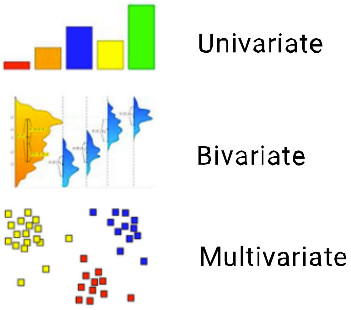

Figure 2.1: Univariate-Bivariate-Multivariate[30]

Bivariate Time Series: Bivariate time series analysis[52] is a statistical method used to analyze and understand the relationship between two time series variables over time. This type of analysis is commonly used in fields such as economics, finance, and marketing, where the behavior of two related variables is of interest.

The analysis of bivariate time series involves examining the correlation and causality between the two variables, identifying any seasonality or cyclicality in each variable, and modeling the data to forecast future values. This is typically done through techniques such as time series regression, Granger causality test, and vector autoregression (VAR) modeling.

In general, the goal of bivariate time series analysis is to identify and understand the relationship between the two variables, which can help researchers and analysts make predictions about future values of one variable based on the behavior of the other variable. Examples of bivariate time series include the relationship between stock prices and interest rates, the relationship between oil prices and inflation, and the relationship between advertising spending and sales.

Multivariate Time Series: Multivariate time series analysis[34] is a statistical method used to analyze and understand the relationship between multiple time series variables over time. This type of analysis is commonly used in fields such as economics, finance, engineering, and environmental science, where multiple variables are interdependent and can influence each other.

The analysis of multivariate time series involves examining the correlations and causality between the variables, identifying any seasonality or cyclicality in each variable, and modeling the data to forecast future values. This is typically done through techniques such as multivariate regression, VAR modeling, and dynamic factor models.

In general, the goal of multivariate time series analysis is to identify and understand the complex relationships between multiple variables. Examples of multivariate time series include the relationship between economic indicators such as GDP, inflation, and unemployment rates, the relationship between weather variables such as temperature, precipitation, and wind speed, and the relationship between traffic flow, air quality, and weather conditions.

2.3 Forecasting Time Series Data

Forecasting time series data[46] involves using statistical models and techniques to predict future values of a time series based on historical data. There are several methods can be used for time series forecasting, including exponential smoothing, ARIMA (autoregressive integrated moving average), and machine learning models. Types of forecasting methods are shown in Figure2.2.

Exponential smoothing is a popular method for time series forecasting, as it can be applied to a wide range of time series data, including data with trend and/or seasonal patterns. It works by assigning exponentially decreasing weights to past observations, with more weight given to recent observations.

ARIMA models are another commonly used method for time series forecasting. They are based on the idea that a time series can be modeled as a combination of autoregressive (AR) and moving average (MA) components. The model can be used to identify patterns in the data, such as trends and seasonal patterns, and make predictions based on these patterns.

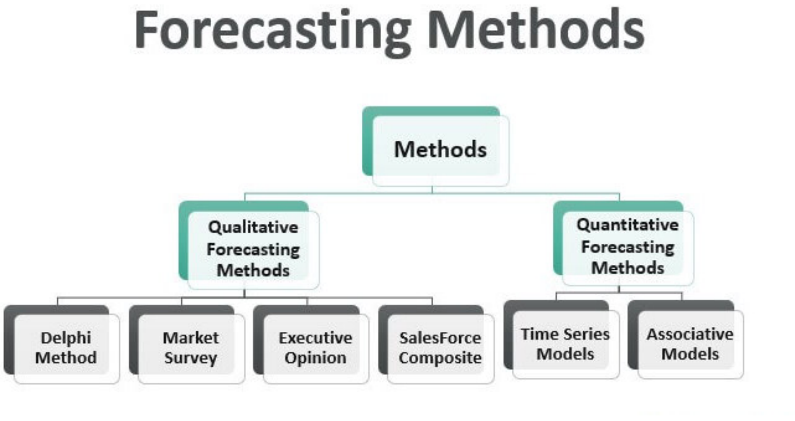

Figure 2.2: Forecasting Methods[57]

Machine learning models, such as random forests[20] and neural networks[28], can also be used for time series forecasting. These models can capture complex patterns in the data and are often used when there are non-linear relationships between variables.

When forecasting time series data, it is important to evaluate the accuracy of the model. This can be done by comparing the predicted values to actual values in the test data set. Some common metrics to measures the forecast accuracy are root mean squared error (RMSE), mean absolute error (MAE) and mean squared error (MSE).

Overall, forecasting time series data is an important task in many industries, as it can provide valuable insights and help organizations make informed decisions about future planning and resource allocation. In Figure2.3, it is shown that the Historical data is used to forecast the future value.

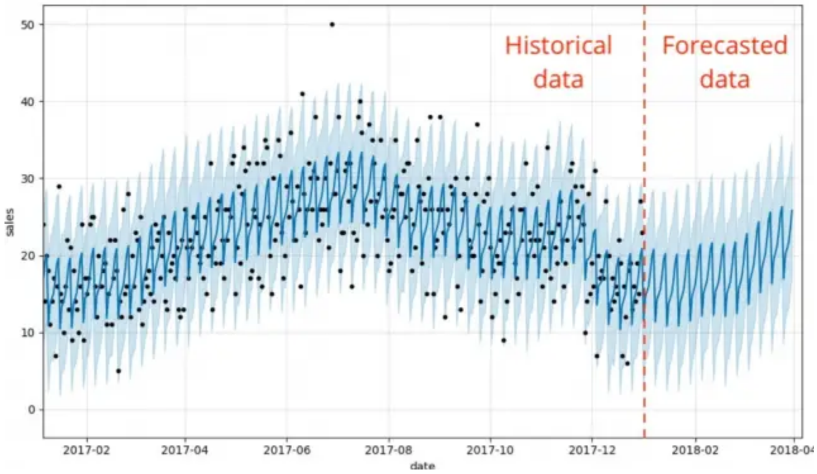

Figure 2.3: Data Forecasting[38]

2.4 Components

In time series analysis, a time series is a sequence of observations recorded over time. There are several components in a time series, as illustrated in Figure 2.4

Trend: The long-term movement of a time series is called Trend. It captures the direction in which the time series is moving over time. Trends can be either upward (positive trend), downward (negative trend), or horizontal (no trend)[48].

Seasonality: Seasonality[55] is the variation that occur at regular intervals in a time series. Seasonal patterns can be less than one year, such as daily, weekly, monthly and can be caused by factors such as weather, holidays, or business cycles.

Cyclical: Cyclical components[3] are irregular fluctuations in a time series that can occur over varying time periods. They can be caused by economic factors such as business cycles, and can have a periodicity of several years.

Irregular: Irregular components[12] are the random fluctuations that occur in a time series. It can be caused by sudden changes in the market, natural adversity, or other unforeseen events.

By understanding the different components of a time series, we analyse the patterns of the time series and predict future behavior. In Figure2.4, it is showing the pattern of the components in graph of time series data.

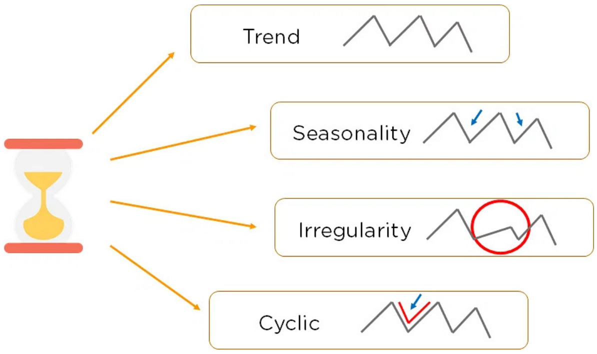

Figure 2.4: Time Series Components[63]

2.5 Stock Market

The stock market is a place where investors buy and sell shares of traded companies. It is a place where companies raise capital by selling ownership shares to the public, and where investors can potentially earn a return on their investment by buying low and selling high.

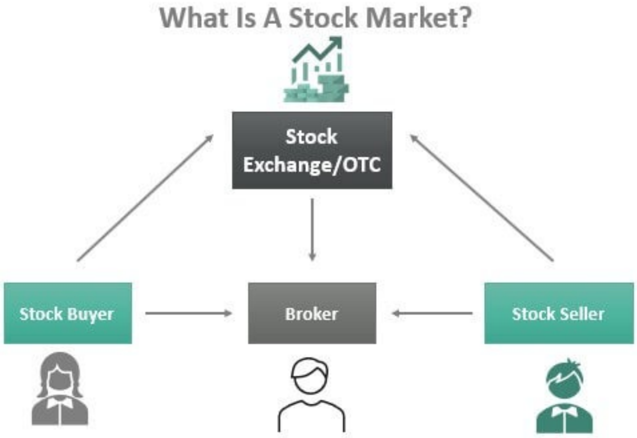

Figure 2.5: Stock Market[67]

Data plays a critical role in the stock market. Investors use various types of data to make decisions about buying and selling stocks. This data can include financial statements, earnings reports, news articles, economic indicators, and market data such as stock prices and trading volumes.

In recent years, the use of data and technology in the stock market has increased dramatically, with the rise of algorithms, high-frequency trading, and big data analytics. These advancements have led to increased efficiency in the market, but also raised concerns about the impact on human decision-making and the potential for market manipulation.

Investors can use this data to perform various types of analysis, such as fundamental analysis, which involves analyzing a company's financial data and industry trends to determine its overall health and growth potential. Technical analysis, on the other hand, involves looking at price and volume data to identify patterns and trends that can help predict future price movements.

Chapter 3

Data Preprocessing

3.1 Preprocessing Time Series Data

Preprocessing time series data is an essential step before using it for analysis or modeling. Here are some common preprocessing steps for time series data[9]:

Data Cleaning: This step involves identifying and correcting any missing, erroneous, or inconsistent data points in the time series. This can include filling in missing values, removing outliers, and smoothing noisy data.

Resampling: Depending on the nature of the data, it may be necessary to resample the time series to a lower or higher frequency. For example, if the data is collected at irregular intervals, it may need to be resampled to a regular interval to make it easier to work with.

Normalization: This step involves scaling the time series data to a common range, such as between 0 and 1 or -1 and 1. Normalization can help to eliminate the effects of scale and make it easier to compare different time series.

Differencing: Differencing involves calculating the difference between consecutive data points in the time series. This can help to remove trends and seasonality in the data, making it easier to model and analyze.

Transformations: Some time series data may benefit from transformations such as logarithmic or exponential transformations, which can help to stabilize variance and make the data more normally distributed.

Feature Engineering: Depending on the problem at hand, it may be necessary to create new features from the time series data. For example, creating lag features, rolling averages, or moving averages can help to capture important patterns in the data.

Splitting Data: Lastly, the time series data is split into training, validation, and testing sets. The training set is used to train the model, the validation set is used to tune hyperparameters, and the testing set is used to evaluate the model's performance on unseen data.

3.2 About Data

In this study, five years (2017-01-01 to 2021-12-31) of daily stock data from 10 companies are taken. Data is downloaded from Yahoo finance website[5]. The name of the companies are Reliance, Tesla, Apple, Amazon, Ford Motor, Microsoft, FMC Corporation, Meta Platforms, Ecolab Inc and LyondellBasell.

Table 3.1: Reliance Data

Date	High	Low	Open	Close	Volume	Adj Close
2021-12-27	2378.00	2348.1	2361.55	2370.25	1853948.0	2363.13
2021-12-28	2404.85	2373.05	2375.6	2398.4	2941883.0	2363.13
2021-12-29	2419.00	2391.00	2391.00	2402.5	7118779.0	2395.28
2021-12-30	2404.94	2400.00	2400.00	2359.1	13537254.0	2352.01
2021-12-31	2383.89	2373.00	2373.00	2368.14	4373768.0	2361.04

Table 3.2: Tesla Data

Date	Open	High	Low	Close	Adj Close	Volume
2021-12-23	335.6	357.66	332.51	355.66	355.66	92713200
2021-12-27	357.89	372.33	356.906	364.64	364.64	71145900
2021-12-28	369.82	373.00	359.47	362.82	362.82	60324000
2021-12-29	366.21	368.00	354.71	362.06	362.063	56154000
2021-12-30	353.77	365.18	351.049	356.77	356.77	47040900

Table3.1 is Reliance data set and Table3.2 is Tesla Data Set. The Reliance data set has 1235 entries and 7 columns and the Tesla Data Set has 1258 entries and 7 columns. Columns of the data set are

- **Date:** This column represents the date of a particular stock.
- **High:** High represents the highest price of the stock on that particular date.
- **Low:** Low represents the lowest price of the stock.
- **Open:** It is the opening price of stock.
- **Close:** It is the closing price of stock.
- **Volume:** Volume is the total number of shares that have been bought or sold in a specific period of time or during the trading day.
- **Adj Close:** Adjusted close is the closing price after adjustments for all applicable splits and dividend distributions.

Table3.3 and Table3.4 represent the numerical behaviors of the Reliance data set and Tesla data set.

Table 3.3: Reliance Data Description

	Open	High	Low	Close	Volume	Adj Close
count	1235.00	1235.00	1235.00	1235.00	1235.00	1.23e+03
mean	1405.35	1421.51	1387.61	1403.73	1388.45	1.03e+07
std	564.26	569.99	556.48	562.53	565.69	7.68e+06
min	503.72	509.56	501.64	503.18	488.82	7.87e+05
25%	931.17	938.85	921.53	931.54	911.62	5.82e+06
50%	1269.02	1279.67	1254.67	1266.49	1248.98	8.07e+06
75%	1964.95	1983.75	1938.65	1960.47	1948.32	1.21e+07
max	2742.75	2751.35	2708.00	2731.85	2723.64	6.58e+07

Table 3.4: Tesla Data Description

	Open	High	Low	Close	Adj Close	Volume
count	1258.00	1258.00	1258.00	1258.00	1258.00	1.258e+03
mean	83.21	85.025	81.303	83.29	83.29	1.34e+08
std	99.14	101.31	96.83	9.25	99.25	9.052e+07
min	12.073	12.44	11.79	11.93	11.93	2.94e+07
25%	19.97	20.32	19.57	19.95	19.95	7.55e+07
50%	23.17	23.49	22.81	23.16	23.16	1.051e+08
75%	140.45	143.45	136.81	140.38	140.38	1.57e+08
max	411.47	414.49	405.66	409.97	409.97	9.14e+08

where,
- **count:** Frequency of the data,
- **mean:** The Average of the values column wise,
- **std:** The standard deviation of the values column wise,
- **min:** The min values the column,
- **max:** The min values the column,
- **25%:**, **50%:** and **75%:** represent percentiles.

Figure 3.1 gives an illustration of quantiles.

3.3 Data

Data visualization[66] refers to the use of graphical and visual elements to represent and communicate data effectively. It is a way of presenting data in a more understandable and appealing manner, using charts, graphs, maps, and other visual aids.

Primary goal of data visualization is to present data in a way that is easy to understand, analyze, and draw insights from. It is a crucial tool in various fields such as business, science, engineering, finance, and many others. Data visualization can help identify

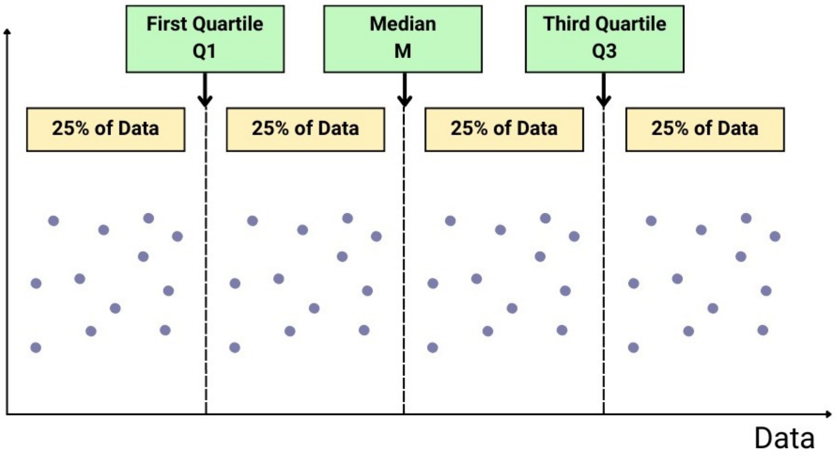

Figure 3.1: Quartile Deviation

patterns, trends, and relationships in data that may be difficult to discern from raw data. It can also help communicate complex data to a broader audience in a way that is both engaging and accessible.

There are various types of data visualization techniques, including:

1. **Line charts:** Line charts are commonly used to represent data that changes continuously, such as stock prices, temperature, or population growth. They can also be used to compare multiple data sets by displaying them on the same graph with different colors or line styles.

   To create a line chart, the data is typically plotted on a coordinate grid with time on the x-axis and the variable being measured on the y-axis. Each observation point is represented by a dot, and the dots are connected by a line to show the overall trend.

2. **Bar charts:** Bar charts are commonly used to represent categorical data, such as survey results, sales figures, or population statistics. They can be used to compare the values of different categories by displaying the bars next to each other or to show changes in a single category over time by displaying the bars sequentially.

   To create a bar chart, the data is typically plotted on a vertical or horizontal axis, with the length of the bar representing the value of the data point. Each bar is typically separated by a small gap to visually distinguish each category or time period being displayed.

3. **Scatter plots:** A scatter plot, also known as a scatter diagram, is a type of data visualization that displays the relationship between two variables. It is used

to identify patterns or trends in the data and to identify potential correlations between the variables.

Scatter plots are commonly used in scientific and statistical research to analyze the relationship between two quantitative variables, such as age and weight or temperature and humidity. Each point is represented by a dot, and the dots are plotted on a coordinate grid with one variable on the x-axis and the other variable on the y-axis.

4. **Pie charts:** A pie chart is a type of data visualization that displays data as a circular graph divided into slices, where each slice represents a category or proportion of the data. It is commonly used to show the distribution of categorical data, such as the percentage of a population with a particular characteristic.

   To create a pie chart, the data is typically represented as a set of percentages or proportions. The circle is divided into slices, with each slice representing a category and its size proportional to its percentage or proportion of the data.

5. **Heat maps:** A heat map is a type of data visualization that displays data values as colors in a two-dimensional matrix. It is commonly used to show the distribution or intensity of a phenomenon across different variables or time periods.

   To create a heat map, the data is represented as a matrix in which each row represents a variable and each column represents a time period or another variable. Each data value is represented by a color where darker colors indicating higher values and lighter colors indicating lower values.

The choice of the visualization technique and design should be based on the specific needs of the project and should help the audience gain insight and understanding of the data. Some popular tools for creating data visualizations include Tableau, Excel, Python libraries like Matplotlib and Seaborn, and R libraries like ggplot2.

Plot Open vs Close
Figure 3.2 is plot of Open Vs Close columns of Reliance data.

Plot High vs Low
Figure 3.3 is plot of High Vs Low columns of Reliance data.

3.4 Feature Engineering

To use year, month, days from column 'Date', extract year, month, day separately[50].

In Table3.5 **year**, **month** and **day** are the extracted columns from **Date** column.

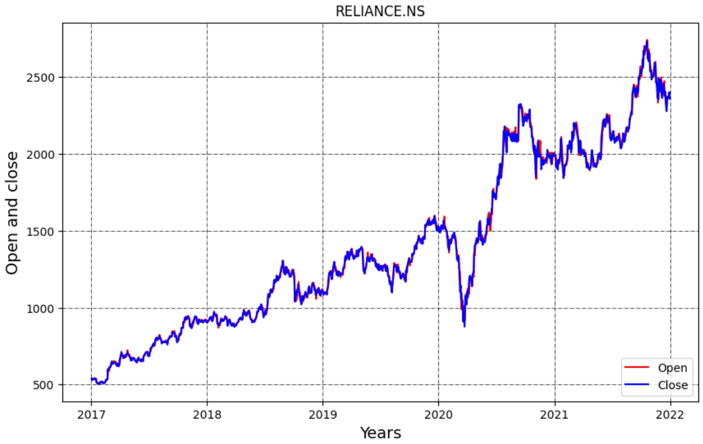

Figure 3.2: Open vs Close

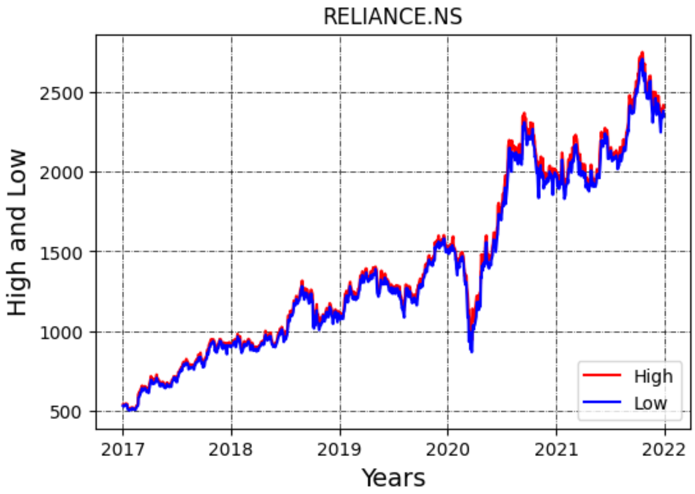

Figure 3.3: Height vs Low

Table 3.5: Extract y-m-d

	Date	Open	High	Low	Close	Adj Close	Volume	year	month	day
1230	2021-12-24	2370.00	2392.00	2337.55	2372.8	2365.67	3639616	2021	12	24
1231	2021-12-27	2361.55	2378.00	2348.1	2370.25	2363.13	1853948	2021	12	27
1232	2021-12-28	2375.6	2404.85	2373.05	2398.39	2391.2	2941883	2021	12	28
1233	2021-12-29	2391.00	2419.00	2382.1	2402.5	2395.28	7118779	2021	12	29
1234	2021-12-30	2400.00	2404.94	2345.6	2359.1	2352.01	13537254	2021	12	30

Quarter is added based on month which is shown in Table3.6.

Splitting Data:
Time series data cannot be split into training and testing randomly. While dividing

Table 3.6: Extract quarter

	Date	Open	High	Low	Close	Adj Close	Volume	year	month	day	quarter
276	2018-02-09	880.65	893.53	877.68	889.41	870.38	5344463	2018	2	9	1
563	2019-04-12	1337.32	1344.15	1324.09	1330.48	1310.04	5975855	2019	4	12	2
945	2020-10-30	2033.5	2065.1	2021.8	2054.5	2041.76	15700946	2020	10	30	4
436	2018-10-05	1089.17	1104.82	1028.4	1039.001	1023.032	22151189	2018	10	5	4
1138	2021-08-11	2096.94	2120.00	2083.39	2117.3	2110.94	4238859	2021	8	11	3

time series data in train and test, it is important to maintain continuity of the data. Five years of data is used in this study. Training testing split is done in the following manner. One year of data is used for training and the next observation is considered as testing. The training window moves forward with time, and the next observation is considered as testing. Next consider two years data as training data and next observation is considered as test data and so on. Table3.7 illustrates the process of training and testing split.

Table 3.7: Splitting Data

training data	test data
364 observations	365-th observation
730 observations	731-th observation
1000 observations	1001-th observation
1234 observations	1235-th observation

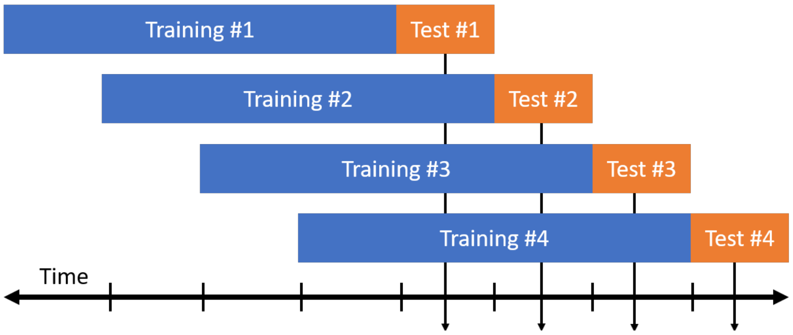

Figure 3.4: Splitting Time Series Data[1]

Figure3.4 gives an illustration of how to split time series data into train and test data.

Chapter 4

Statistical Models

Statistical models are mathematical models that use statistical techniques to analyze and explain data. They are used in data science, machine learning, engineering, or operations research.

Statistical models are used to explain the relationships between variables, make predictions, and test hypotheses. They can take various forms, from simple linear regression models to more complex models such as multivariate regression, time-series models, and machine learning models.

In statistical modeling, data is analyzed to identify patterns, relationships, and trends, and to develop mathematical equations that describe the underlying processes. The model parameters are estimated from the data, and the model is validated by comparing its predictions to new data. The use of statistical models has revolutionized many fields by enabling more accurate predictions and better understanding of complex systems. However, statistical models are not perfect and can have limitations. it's miles essential to carefully consider the assumptions and limitations of a model earlier than the use of it to make decisions or draw conclusions.

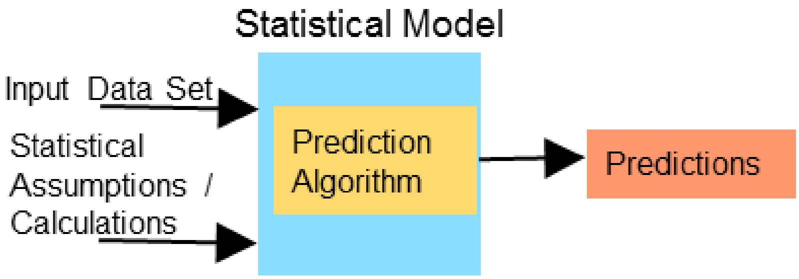

Figure 4.1: Statistical models[33]

4.1 Decomposition

Time series decomposition is the process of breaking down a time series into its component parts, typically a trend, seasonality, and residual (or noise) component. The decomposition of time series helps to understand the underlying structure of the data, identify any trends or seasonal patterns, and make more accurate forecasts by modeling each component separately. There are various methods for decomposing time series data, including classical decomposition, seasonal decomposition of time series, and wavelet decomposition.

Additive Decomposition: Additive decomposition is one of the methods for decomposing a time series into its component parts, including trend, seasonal, and residual components. In the additive method, the components are added together to form the time series.

Mathematically, an additive decomposition of a time series Yt can be represented as:

Y_t = Trend Component + Seasonal Component + Residual Component

where the trend component is the long-term pattern in the data, the seasonal component represents the seasonal variation, and the residual component is the random variation that cannot be explained by the trend or seasonal components.

The additive decomposition method is useful when the seasonal pattern of the time series is relatively constant over time, and the magnitude of the seasonal variation is independent of time. However, if the seasonal pattern changes over time or the magnitude of the seasonal variation is proportional to the level of the time series, the multiplicative decomposition method may be more appropriate.

Multiplicative Decomposition: Multiplicative decomposition is another method for decomposing a time series into its components, that is trend, seasonal, and residual components. In the multiplicative method, the components are multiplied together to form the time series.

Mathematically, a multiplicative decomposition of a time series Y_t can be represented as:

Y_t = Trend Component x Seasonal Component x Residual Component

where the trend component is the long-term pattern in the data, the seasonal component represents the seasonal variation, and the residual component is the random variation that cannot be explained by the trend or seasonal components.

The multiplicative decomposition assumes that the seasonal variation is proportional to the level of the time series, meaning that the magnitude of the seasonal variation increases or decreases with time. This means that the seasonal component is multiplied by the time series at each point in time to capture the seasonal variation. However, if the seasonal pattern is relatively constant over time and the magnitude of the seasonal variation is independent of the level of the time series, the additive decomposition method may be more appropriate.

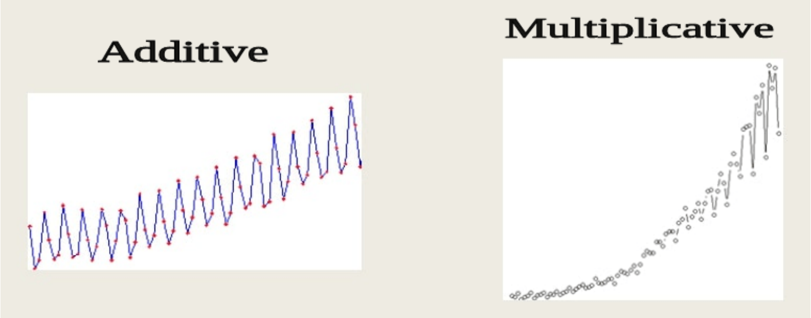

Figure 4.2: Additive and Multiplicative Decomposition[70]

4.2 Normality Check

Normality check of data is an important step in statistical analysis as many statistical tests, such as t-test, ANOVA, and correlation, assume that the data are normally distributed.

There are several methods to check the normality of data:

Histogram: A histogram is a plot that is a graphical representation of the frequency distribution of a dataset. It can be used to visually inspect the shape of the distribution. A normal distribution will have a bell-shaped curve, whereas skewed distributions will have a curve that is skewed to the left or right.

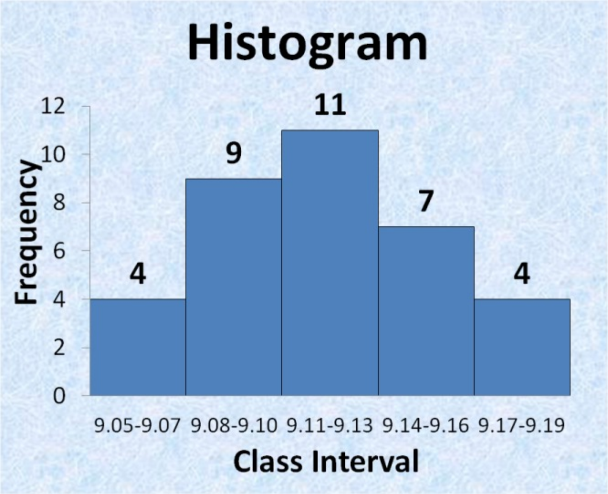

Figure 4.3: Histogram[69]

Q-Q plot: A Q-Q plot, or quantile-quantile plot, is a graphical method for comparing the distribution of a sample to a normal distribution. If the sample distribution is normal, the points on the Q-Q plot will fall on a straight line.

Shapiro-Wilk test: The Shapiro-Wilk test is a statistical test that can be used to determine whether a dataset is normally distributed. The null hypothesis of the test is that the data is normally distributed. If the p-value is less than the significance level (usually 0.05), then the null hypothesis is rejected, and the data is considered to be non-normal.

Kolmogorov-Smirnov test: The Kolmogorov-Smirnov test is another statistical test to test the normality of data. It compares the empirical cumulative distribution function of the sample to the theoretical cumulative distribution function of a normal distribution. If the p-value is less than the significance level, then the null hypothesis is rejected, and the data is considered to be non-normal.

It is important to know that none of these methods can definitively prove that a dataset is normal or non-normal. They can only provide evidence for or against normality. It is sufficient to use a combination of these methods to assess the normality of data.

4.3 Heteroscedasticity Test

Heteroscedasticity refers to the situation where the variance of errors in a regression model is not constant across the range of the independent variable(s). This violates one of the assumptions of the ordinary least squares (OLS) regression model, which assumes that the variance of errors is constant (homoscedasticity).

To test for heteroscedasticity in a regression model, a variety of diagnostic tests can be performed.

Graphical analysis: Plot the residuals (i.e., the differences between the actual and predicted values) against the predicted values or the independent variable(s). If the variance of the residuals changes systematically with the predicted values or the independent variable(s), this indicates heteroscedasticity.

Breusch-Pagan test: The Breusch-Pagan test is a statistical test used in econometrics to test heteroscedasticity which represents non-constant variance in the error terms of a regression model.

The Breusch-Pagan test involves a regression of the squared residuals from the original regression model on the independent variables in the model. The test statistic is then calculated as the number of observations times the R-squared from this auxiliary regression. Under the null hypothesis of homoscedasticity, the test statistic follows a chi-squared distribution with degrees of freedom, the number of independent variables in the original regression model.

If the calculated test statistic is greater than the critical value of the chi-squared distribution at a given significance level and the p-value is less than 0.05, then the null hypothesis of homoscedasticity is rejected and the alternative hypothesis of heteroscedasticity is accepted.

The Breusch-Pagan test is commonly used in applied econometric analysis to check for the presence of heteroscedasticity in regression models and to decide whether to use robust standard errors or to transform the data to correct for heteroscedasticity.

White test [72]: This test is similar to the Breusch-Pagan test, but it also includes higher-order terms of the independent variable(s) and their cross-products. The null hypothesis is again homoscedasticity, and the test statistic follows a chi-square distribution. If the p-value is less than the significance level, the null hypothesis will be rejected with a conclusion that heteroscedasticity is present.

Overall, there is no single ”best” test for heteroscedasticity, and the choice of test depends on the specific characteristics of the data and the regression model.

4.4 Stationarity Check

Stationarity refers to the property of a time series data where its statistical properties remain constant over time, such as mean, variance, and auto-correlation. A stationary time series is easier to model and forecast than a non-stationary one[73].

To check for stationarity in a time series, following tests can be performed:

Visual inspection: Plot the time series data and check if the statistical properties remain constant over time. Specifically, check if there is a trend, seasonality, or cycles that could indicate non-stationarity.

Summary statistics: Calculate the mean and variance of the time series data over different time periods. If the mean and variance are not depend on time, then the series is stationarity.

Augmented Dickey-Fuller (ADF) test: This is a hypothesis test that tests whether a time series has a unit root (i.e., a root that equals 1) and is therefore non-stationary. The null hypothesis of the test is the series has a unit root and the alternative hypothesis is the series does not has a unit root. If the p-value is less than the significance level (e.g., 0.05), the null hypothesis will be rejected and conclude that the time series is stationary.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test: This is another hypothesis test that tests for stationarity in a time series. The null hypothesis is that the time series is stationary, and the alternative hypothesis is that it is not. If the p-value is less than the significance level, the null hypothesis will be rejected and conclude that the time series is non-stationary.

Phillips-Ouliaris test with Trend Breaks: The PSR test is a type of unit root test[15], which test the presence of a unit root in the data. A unit root indicates that the series is non-stationary and that the statistical properties of the series are changing over time.

The PSR test is an extension of the Phillips-Perron test that allows for multiple structural breaks in the time series trend. The presence of structural breaks can affect the results of unit root tests, as the trend may be different in different parts of the time series. The PSR test accounts for this possibility by allowing for trend breaks at unknown points in time.

In general, the PSR test involves estimating a regression model that includes a time trend and any additional covariates that may be relevant. The residuals from this model are then used to test for the presence of a unit root. The PSR test has been shown to have good power in detecting unit roots in time series data with structural breaks.

Overall, there is no single ”best” test for stationarity, and the choice of test depends on the specific characteristics of the data and the time series model.

Chapter 5

Models

5.1 Linear Regression

Linear regression[6] is a statistical method used to model the relationship between two or more variables by fitting a linear equation to the given data. The goal of linear regression is to find the line of best fit that summarizes the relationship between the dependent variable (also known as the response variable or outcome variable) and the independent variable(s) (also known as predictor variable(s) or explanatory variable(s))[47]. The best fit line is represented by a linear equation of the form

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n,$$

where Y is the dependent variable, X_1, X_2, …, X_n are the independent variables, and b_0, b_1, b_2, …, b_n are the coefficients that determine the slope and intercept of the line. The coefficients are estimated from the observed data using a method called least squares fit.

Linear regression is widely used in many fields, including economics, finance, biology, engineering, and social sciences. It is widely used in prediction, forecasting, and modeling the relationship between variables. There are also different types of linear regression models[59], including simple linear regression (one independent variable), multiple linear regression (more than one independent variable), and polynomial regression (non-linear relationships between variables).

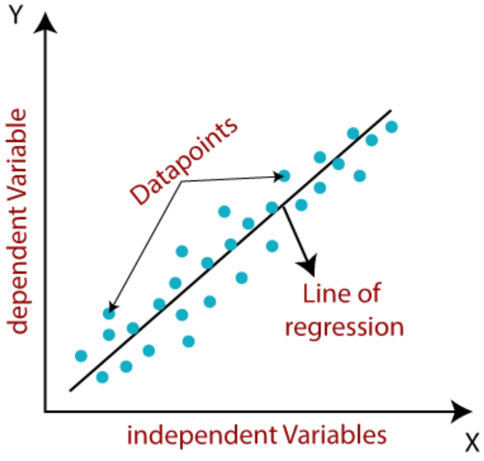

Figure 5.1: Linear Regression[7]

5.1.1 Cost Function

The cost function of linear regression is a mathematical expression that measures the error between the actual values and the predicted values of the dependent variable in a linear regression model. It is also known as the objective function, loss function, or mean squared error (MSE) function[22].

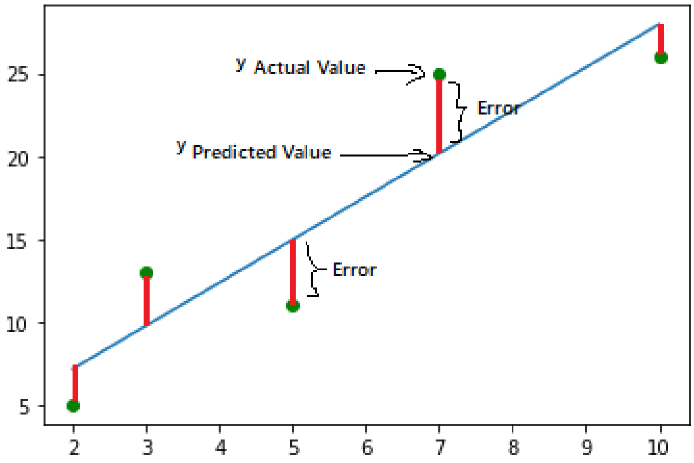

Figure 5.2: Error of Linear Regression[54]

The cost function for linear regression is defined as the sum of the squared differences between the predicted values $(\hat{y})$ and the actual values (y) of the dependent variable, divided by the number of observations (N):

$$J = \frac{1}{2N} \sum (\hat{y} - y)^2$$

where $\hat{y}$ represents the predicted values, y represents the actual values, and N represents the number of observations in the dataset.

The goal of linear regression is to minimize this cost function by finding the optimal values of the parameters, which results in the best possible fit between the predicted and actual values of the dependent variable. The optimization of the cost function can be achieved using various methods such as gradient descent, normal equations, or other optimization algorithms.

There are some most used Regression cost functions are discussed below,

Mean Absolute Percentage Error: MAPE measures the percentage difference between the predicted and actual values.
The MAPE is calculated using the following formula,

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i|} * 100\%$$

where, n is the number of data points
y_i is the actual value of the i-th data point
$\hat{y}_i$ is the predicted value of the i-th data point.

Mean Absolute Error: MAE This cost function measures the average absolute difference between the predicted and actual values. The formula for calculating the mean absolute error is

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

where, n is the total number of observations,
y_i is the actual value of the i-th observation,
$\hat{y}_i$ is the predicted value of the i-th observation.

Root Mean Squared Error: RMSE is the square root of the average squared difference between the predicted and actual values.
The formula for calculating the root mean square error is

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

where, n is the total number of observations,
y_i is the actual value of the i-th observation,
$\hat{y}_i$ is the predicted value of the i-th observation.

5.1.2 Optimization

Optimization of linear regression[29] involves finding the values of the regression coefficients that minimize the sum of squared residuals between the predicted values and the actual values of the dependent variable.

There are several methods to optimize linear regression:

Ordinary least squares: Ordinary least squares, or OLS is the most commonly used method for linear regression optimization. It involves minimizing the sum of squared residuals by setting the partial derivative of the sum of squared residuals with respect to each regression coefficient to zero. The resulting equations can be solved using matrix algebra.

Gradient descent: Gradient descent is an iterative optimization algorithm that can be used to optimize linear regression. It involves updating the regression coefficients by taking steps in the direction of steepest descent of the cost function. The learning rate, which determines the step size, needs to be carefully chosen to ensure convergence. Equation of Gradient descent is

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} [(h_\theta(x^i) - y^i) x_j^i]$$

where,

• θ_j is the j^{th} model parameter,

• α is the learning rate, which controls the step size of the parameter updates,

• m is the number of training examples,

• $h_\theta(x^i)$ is the predicted output for the i^{th} input using the current values of the model parameters θ,

• y^i is the actual output for the i^{th} input,

• x_j^i is the j^{th} feature of the i^{th} input.

Stochastic gradient descent: SGD is similar to gradient descent, but it updates the regression coefficients using a randomly selected subset of the training data at each iteration. This can speed up convergence and reduce the computational cost.

Ridge regression: Ridge regression is a regularized version of linear regression that adds a penalty term to the sum of squared residuals. This penalty term is proportional to the square of the magnitude of the regression coefficients. Ridge regression can help prevent overfitting and improve the generalization performance of the model.

Lasso regression: Lasso regression is another regularized version of linear regression that adds a penalty term which is the sum of absolute values of the regression coefficients. This penalty term can cause some of the coefficients to become exactly zero, effectively performing feature selection. Lasso regression can also help prevent overfitting and improve the interpretability of the model.

Overall, the choice of optimization method depends on the size and complexity of the data, as well as the specific goals of the analysis.

5.2 Time Series models

Time series models build relation between successive observations of a variable. Depending on the types of relations, they are classified into the following[21].

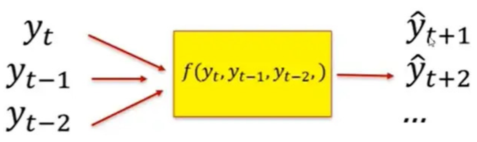

Figure 5.3: Time Series Model

5.2.1 Autoregressive Model

An autoregressive, or AR model[37] is a type of time-series model that predicts future values of a variable based on its own past values. In an AR model,at any given time step the value of the variable is a linear combination of its past values, with coefficients that are learned from the data.

The order of an AR model refers to the number of past values used to predict the current value. For example, an AR(1) model uses only the previous value to predict the current value, while an AR(2) model uses the two previous values.

Mathematically, an AR(p) model can be written as:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

where y_t is the value of the variable at time t, c is a constant term, $\phi_1, \phi_2, \ldots, \phi_p$ are the coefficients for the past values, ϵ_t is a random error term, and p is the order of the model.

AR models can be useful for predicting future values of a time series, identifying trends and patterns in the data, and estimating the impact of past events on the variable of interest. They are commonly used in fields such as finance, economics, and engineering.

5.2.2 Moving Average Model

Moving average or MA[8] is a time series model that is used to explain the dependencies between observations in a time series. In this model, the current observation is assumed to be a linear combination of the past 'q' observations of the series, and an error term or residual.

The mathematical equation for an MA(q) model can be represented as:

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

where Y_t is the observation at time 't', μ is the mean over all observations, at time t ϵ_t is the white noise error term, and $\theta_1, \theta_2, \ldots, \theta_q$ are the coefficients of the past 'q' error terms. The order of the MA model is determined by the value of 'q'.

In this model, the coefficients of the past error terms $(\theta_1, \theta_2, \ldots, \theta_q)$ determine the magnitude and duration of the effect of the past error terms on the current observation. A positive value of the coefficient indicates that a positive error in the past will increase the current observation, while a negative value indicates the opposite.

MA models are used to capture the random shocks or fluctuations in the time series data that cannot be explained by the past values of the series. They are often used in combination with AR (Autoregressive) models to build ARMA models, which can capture both the short-term and long-term dependencies in a time series.

5.2.3 ARMA Model

ARMA, or Autoregressive Moving Average , is a statistical model used to analyze time-series data. It combines two simpler models, the autoregressive (AR) model and the moving average (MA) model, to create a more sophisticated model that captures both trend and seasonality.

The ARMA model is specified by two parameters, p and q. The p parameter represents the order of the autoregressive (AR) component, and the q parameter represents the order of the moving average (MA) component. The AR component captures the linear dependence of the current value on the previous p values, while the MA component captures the linear dependence of the current value on the previous q errors.

Mathematically, an ARMA(p,q) model can be represented as:

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

where Y_t represents the time series at time t, c is a constant term, $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are the AR and MA parameters, respectively, and $\epsilon_t, \ldots, \epsilon_{t-q}$ are the error terms.

The ARMA model is commonly used in fields such as economics, finance, and engineering to model and forecast time-series data. It can be estimated using various statistical techniques such as maximum likelihood estimation or Bayesian estimation.

5.2.4 ARIMA Model

ARIMA, or Autoregressive Integrated Moving Average is a time series forecasting model that is generally used in statistics and econometrics. It is a combination of three different methods: autoregression, integration, and moving average.

Autoregression involves predicting a variable based on its own past values, while moving average involves predicting a variable based on its past errors. Integration involves transforming a non-stationary time series into a stationary one by differencing the series.

The ARIMA model can be represented as ARIMA(p, d, q), where p is the order of AR terms, d is the degree of differencing, and q is the order of MA terms. The model is used to forecast future values of a time series using its past behavior.

ARIMA models are widely used in various fields, such as economics, finance, and engineering, to predict future values of a time series. The model can be used to forecast stock prices, weather patterns, and many other variables that are dependent on time. The accuracy of the ARIMA model depends on the quality of the data used and the appropriateness of the model's assumptions for the specific time series being analyzed.

Integrating Factor

In time series analysis, the term ”integrating factor” is not commonly used, but a similar concept of a transformation factor is often used to transform a non-stationary time series into a stationary one[65].

A stationary time series has statistical properties that remain constant over time, such as constant mean and variance. On the other hand, a non-stationary time series may exhibit trends, seasonality, or other time-varying characteristics.

The ARIMA model is a popular method for modeling time series data, and it involves taking differences of the time series data to achieve stationarity. The order of differencing is determined by the parameter d in the ARIMA model.

The transformation factor used in this case can be written as:

$$\Delta^d = (1 - L)^d$$

where L is the lag operator, and d is the order of differencing. This transformation factor is equivalent to taking the differences of the data d times.

Multiplying the time series data by this transformation factor can help convert a non-stationary time series into a stationary one, allowing for modeling using the ARMA part of the ARIMA model.

Therefore, while the term ”integrating factor” is not commonly used in time series analysis, a similar concept of a transformation factor is used in some models to transform non-stationary time series into stationary ones.

5.2.5 SARIMA Model

SARIMA, or Seasonal Autoregressive Integrated Moving Average models are a class of time series models used for forecasting. They are an extension of the ARIMA (Autoregressive Integrated Moving Average) model. SARIMA models are used when the time series exhibits seasonal patterns, meaning that there is a repeating pattern over time.

SARIMA models are characterized by four components: the seasonal component, the autoregressive component, the moving average component and the integrated component. The seasonal component models the seasonal pattern in the data, while the autoregressive and moving average components model the non-seasonal patterns. The ”I” in SARIMA stands for ”integrated,” which means that the model includes differencing to make the time series stationary.

The parameters of a SARIMA model include the order of the autoregressive, differencing, and moving average components (ARIMA order), as well as the seasonal order, which includes the seasonal autoregressive, differencing, and moving average components (SAR order).

The full notation of a SARIMA model is representated as in Figure5.4.

$$SARIMA\,\underbrace{(p,d,q)}_{non-seasonal}\,\underbrace{(P,D,Q)_m}_{seasonal}$$

Figure 5.4: SARIMA notation[53]

SARIMA models are widely used in various fields, including economics, finance, and meteorology, among others. They have proven to be effective in modeling and forecasting seasonal time series data. However, choosing the optimal SARIMA model can be a challenging task, and various methods are used to determine the best model, such as BIC(the Bayesian Information Criterion) and AIC(the Akaike Information Criterion).

5.2.6 FB Prophet Model

FB Prophet is an open-source time-series forecasting tool developed by Facebook's Core Data Science team. It is designed to make it easier for analysts and developers to forecast time-series data, even if they have limited experience in time-series analysis.

Prophet works by decomposing time-series data into its underlying components such as trend, seasonality, holidays, and outliers. It then fits separate models to each of these components and combines them to make a forecast. Figure?? is the formula of FB Prophet model.

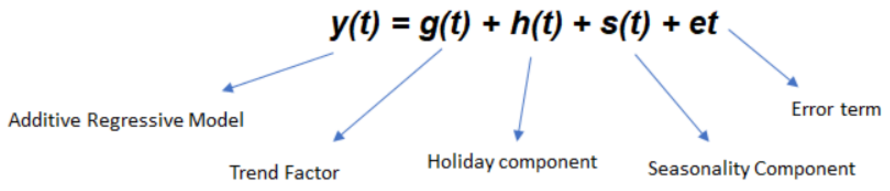

Figure 5.5: FB Prophet [10]

Prophet is easy to use and requires minimal parameter tuning, which makes it an attractive tool for forecasting tasks in various industries. However, it is important to note that Prophet may not always be the best choice for every time-series forecasting problem, and it is always important to consider other models and methods to ensure the best results.

Prophet is designed to be user-friendly and highly customizable. While it has gained popularity for its ability to handle seasonality and holidays, it is true that Prophet can overfit to the training data, which can lead to poor generalization and inaccurate forecasts.

There are several reasons why Prophet can overfit:

Lack of regularization: Prophet allows users to specify a variety of hyperparameters that can affect the model's ability to generalize. If these hyperparameters are not carefully tuned, the model may overfit to the training data.

Lack of cross-validation: Prophet does not have built-in support for cross-validation, which is an important technique for preventing overfitting. Without cross-validation, it can be difficult to assess whether the model is overfitting or not.

Limited flexibility: While Prophet is highly customizable, it has some limitations in terms of the types of time series it can handle. This can lead to overfitting if the data does not conform to Prophet's assumptions.

To address these issues, it is important to carefully tune the hyperparameters of the model and to use cross-validation to assess its performance. Additionally, it may be helpful to explore alternative modeling approaches or to preprocess the data in ways that are more amenable to Prophet's assumptions.

5.3 Deep learning models

Deep learning models are a type of machine learning algorithms that use artificial neural networks with multiple layers to learn from data. Deep learning models have revolutionized the field of machine learning and have achieved state-of-the-art results in various tasks such as image recognition, speech recognition, natural language processing, and many more.

5.3.1 Recurrent Neural Networks

Recurrent Neural Networks, or RNNs are a type of neural networks that are commonly used for time series analysis. Unlike feedforward neural networks, RNNs can maintain a hidden state that depends on previous inputs, which makes them well-suited for analyzing sequences of data, such as time series.

In an RNN, each neuron has an additional input that represents the previous hidden state. This hidden state is updated at each time step, and can be thought of as a memory of the past inputs. The current input and the previous hidden state are combined using a set of weights, and the resulting activation is passed through a non-linear activation function. This output is then used as the hidden state for the next time step. Figure5.6 represents the architecture of RNN.

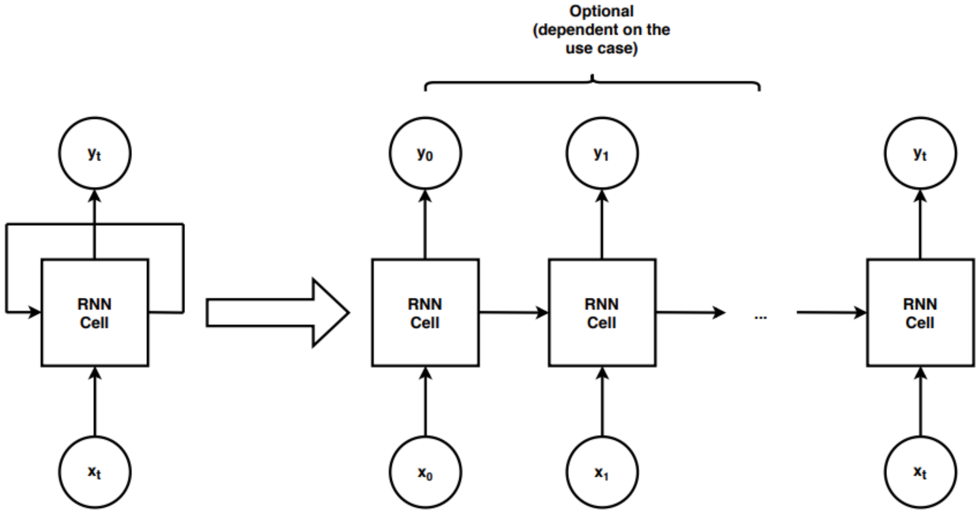

Figure 5.6: RNN[4]

There are different types of RNNs, such as the basic RNN, Long Short-Term Memory (LSTM) networks, and Gated Recurrent Unit (GRU) networks. These networks have different architectures and mechanisms for updating the hidden state, but they all share the ability to capture temporal dependencies in data.

To train an RNN for time series analysis, the network is typically fed a sequence of inputs and asked to predict the next value in the sequence. The weights of the network are updated based on the prediction error, and the process is repeated for multiple time steps.

RNNs have been used for a variety of time series analysis tasks, including stock price prediction, weather forecasting, and speech recognition. They have additionally been utilized in mixture with other neural network architectures, inclusive of Convolutional Neural Networks (CNNs), to analyze time series statistics in multiple dimensions.

5.3.2 LSTM Model

LSTM stands for Long Short-Term Memory. It is a type of artificial neural network architecture that is particularly well-suited for processing sequential data such as speech, text, or time-series data.

The architecture was designed to address the vanishing gradient problem that occurs in traditional recurrent neural networks (RNNs), which can make it difficult for these networks to learn long-term dependencies.

LSTMs use a series of gates to control the flow of information through the network. These gates include an input gate, an output gate, and a forget gate. Figure5.7 represents the architecture of LSTM.

Forget Gate: The forget gate decides which information from the previous hidden state should be discarded or forgotten, and which new information should be added to the current hidden state.

Input Gate: The input gate controls the flow of new information into the current hidden state. It decides which new information is important to add to the current hidden state.

Output Gate: The output gate controls the output of the current hidden state. It decides which parts of the current hidden state should be used as output and which parts should be discarded.

The ability of LSTMs to selectively retain or forget information over long periods of time makes them particularly useful for processing sequential data. LSTMs have been applied successfully in a wide range of applications, including speech recognition, machine translation, and sentiment analysis.

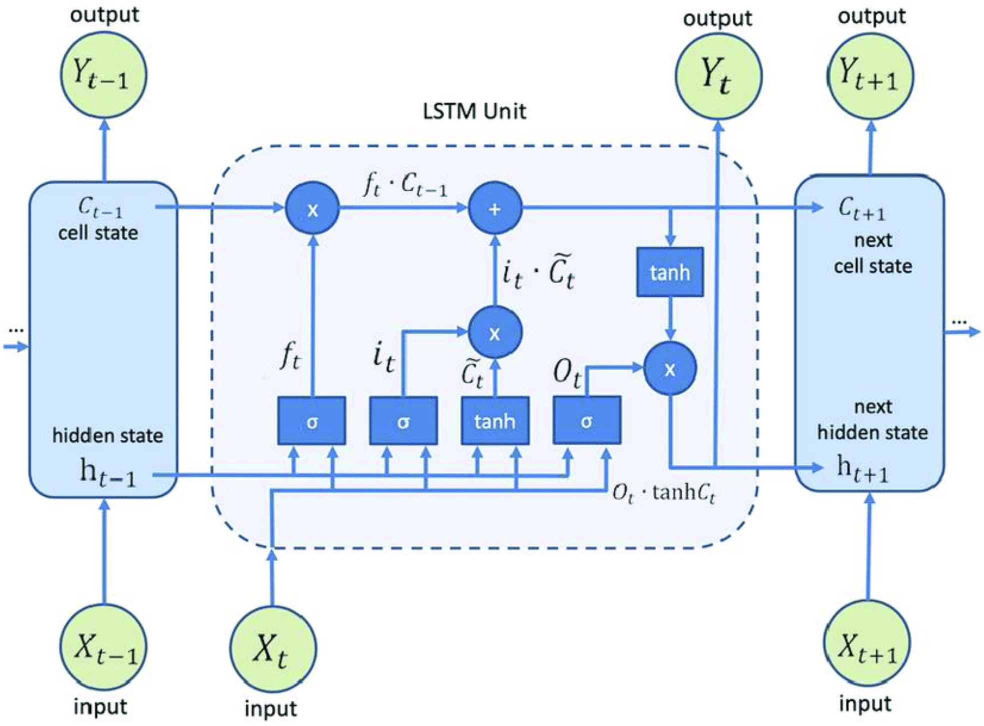

Figure 5.7: LSTM[11]

Chapter 6

Results

The result of time series models depends on the specific model being used and the quality of the data being analyzed. In general, the goal of time series models is to make accurate forecasts of future values based on past observations. The accuracy of the forecasts can be evaluated using various metrics such as MSE(mean squared error), MAE(mean absolute error) and RMSE(root mean squared error).

6.1 Model evaluation methods

When comparing the outputs of several models, it is important to use appropriate metrics that capture the performance of the models in different aspects. Here are some methods to compare outputs of several models:

Accuracy metrics: Accuracy metrics are some of the most common methods to compare model outputs. MSE(mean squared error), MAE(mean absolute error), RMSE(root mean squared error) are examples of accuracy metrics. These metrics measure the difference between the predicted values and the actual values in the dataset. The model with the lowest value of the accuracy metric is considered to be the best.

R-squared: R-squared is a metric that measures how well the model fits the data. It provides a measure of how much of the variance in the dependent variable is explained by the independent variables. The model with the highest R-squared value is considered to be the best.

Cross-validation: Cross-validation is a method to evaluate the performance of models by testing them on data that was not used during training. This helps to ensure that the model is not overfitting the training data. Cross-validation can be used to compare the performance of different models.

Visual inspection: Sometimes, it can be helpful to visually compare the outputs of different models by plotting the predicted values against the actual values. This can help to identify patterns and differences between the models.

Information criteria: Information criteria such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) can also be used to compare the outputs of several models. These criteria balance the goodness of fit of the model with the complexity of the model. The model with the lowest information criterion value is considered to be the best.

Overall, it is important to use a combination of these methods to compare the outputs of several models, as each method provides a different perspective on model performance.

6.2 Result of Time Series Models

The error output from the models that have been applied to data set 1 and data set 2 are plotted in Figure 6.1. Data set 1 is the Reliance Data set and data set 2 is the Tesla data set as defined in chapter-3. These two plots illustrated that LSTM performs better than other models in these kind of predictive tasks.

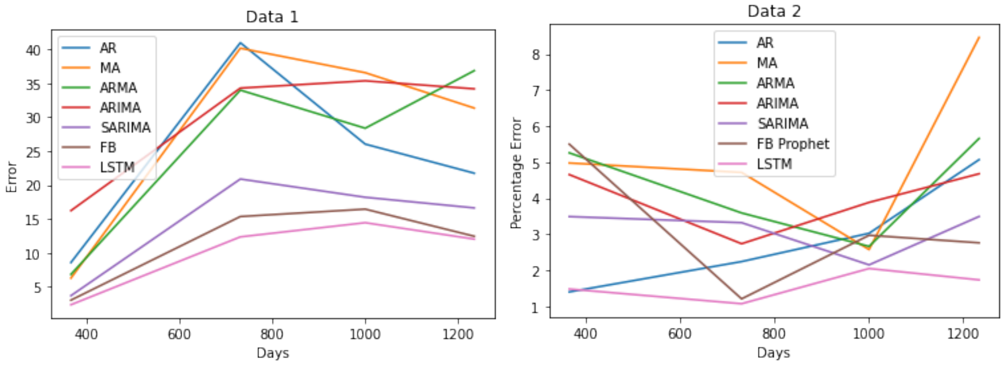

Figure 6.1: Error plot of time series models

To compare all the error of all time series models, box plot is used. In Figure 6.2, the box plot is the combine error plot of all data set. From the plot, it is clear that error for LSTM model is less compared to other models.

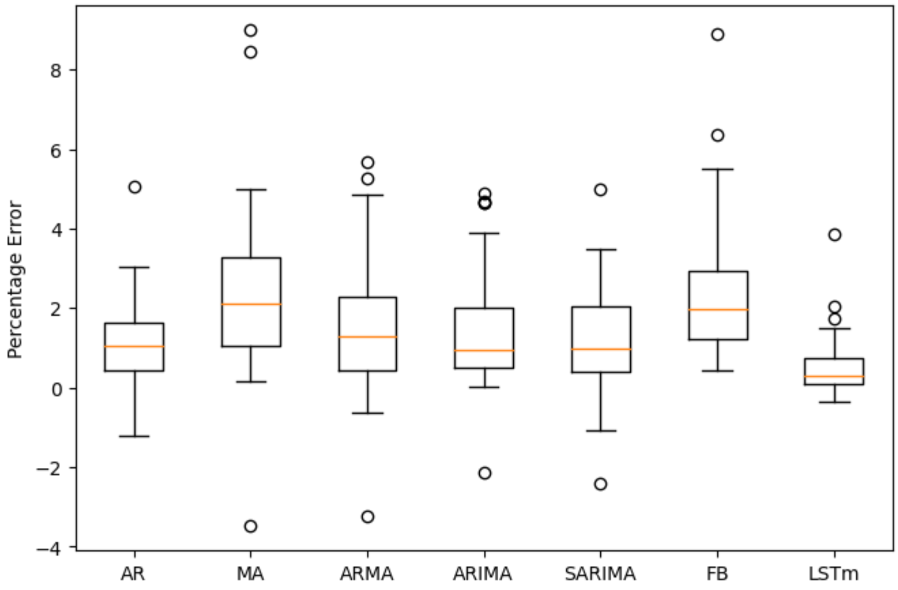

Figure 6.2: Combined error plot of Data

6.3 Result of Statistical Models

Statistical properties of the data sets, like decomposition, normality, heteroscedasticity, stationarity are checked for the data sets. Results of statistiscal properties of data set 1 are bellow.

Decompositions: Figure 6.3 shows that data has trend and seasonality.

Q-Q Plot to check normality: In Figure 6.4 Q-Q plot is used for check stationarity. Since the blue line is not roughly equal to the straigh line, the data seem to not follow normality.

KPSS test to check stationarity:

Since test statistic value is greater that critical value[2] and p-value[6] is less than 0.05, reject the null hypothesis. Hence the data is not stationary.

Breusch-Pagan test to check heteroscedasticity:

The result of Breusch-Pagan test is p-value: 6.89e-20. Since p-value is less than 0.05, reject the null hypothesis. Hence heteroscedasticity is present in the data. Thus the data is not stationary, not normal and heteroscedasticity. The time series models are applied on a simple data set, that is weather dataset[45] of Jena, a place in Germany.

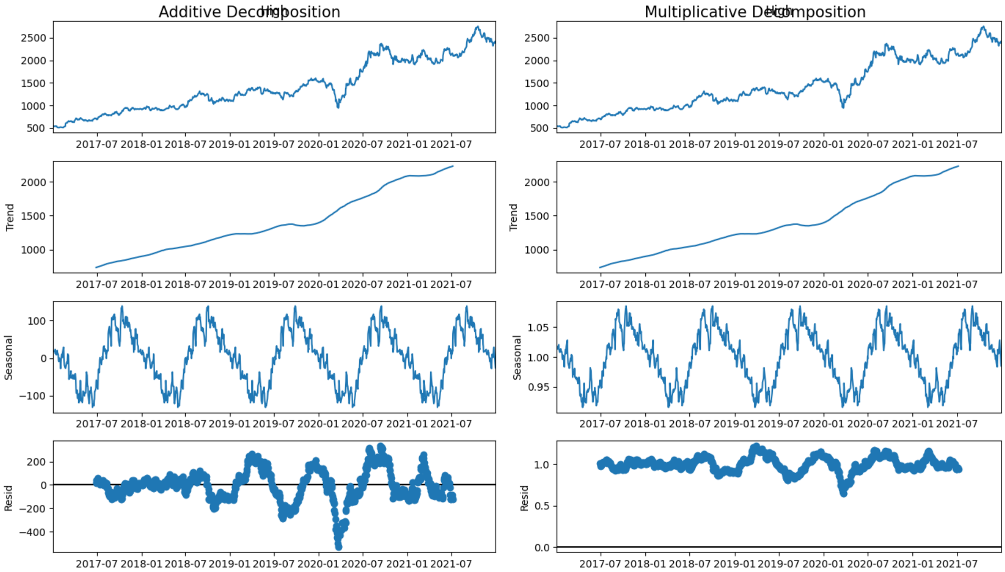

Figure 6.3: Decomposition

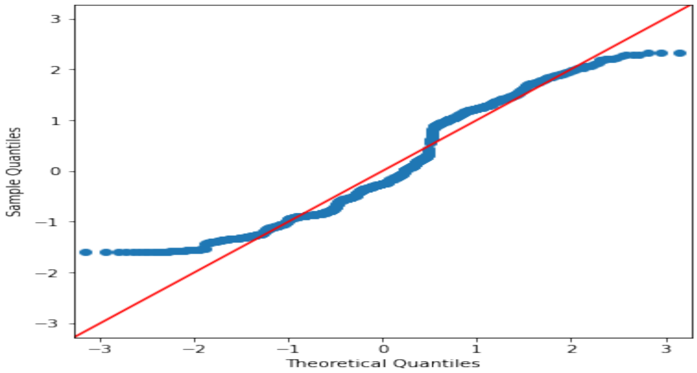

Figure 6.4: q-q plot

Table 6.1: KPSS result of Data set 1

KPSS Statistic	5.48
p-value	0.01
num lags	20
Critical Values:	
10%	0.347
5%	0.463
2.5%	0.574
1%	0.739

Table 6.2: Temperature Data set

Date	Temp
2011-11-01	2.95
2009-08-08	25.14
2016-07-20	27.27
2014-12-26	0.46
2014-04-30	20.02
2015-10-10	5.23
2009-10-08	15.41
2014-08-21	16.81
2010-10-13	4.02
2014-06-03	3.65

Few columns of the data set is:
- **Date:** Particular date on which the temperature reading was taken.
- **Temp:** Temperature reading on a particular date.

KPSS test to check stationarity:

Table 6.3: KPSS Test Values

KPSS Statistic	2.054
p-value	0.01
num lags	362
Critical Values:	
10%	0.347
5%	0.463
2.5%	0.574
1%	0.739

Since test statistic value is grater than critical value and p-value is less than 0.05, reject the null hypothesis and accept alternative hypothesis. Hence data is not stationary.

Error plot of the time series models(AR, MA, ARIMA, SARIMA) of temperature data in Fig 6.5. For temperature data the models error are less, all models perform equally well on temperature data. So the conclusion is that when the data is simple(st temperature data), then the models perform well on the data set but when complexity of data increases(st Stock Data), different models have varying accuracies.

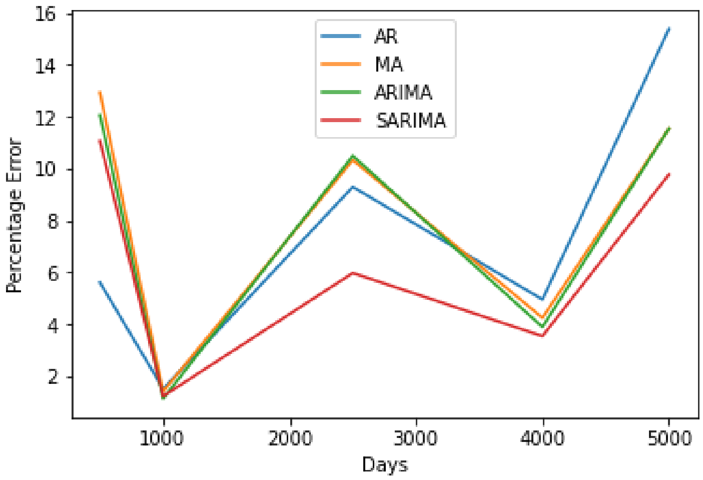

Figure 6.5: Error Plot of Temperature Data

Chapter 7

Stationarity of Time Series

In time series analysis, a stationary series is a sequence of observations that has certain statistical properties, such as mean, variance, auto-covariance remaining constant over time[42]. In contrast, a non-stationary time series is one where the statistical properties change over time, making it difficult to analyze and make accurate predictions. Figure7.1 and Figure7.2 are the plots of High values of Reliance data before and after making stationary.

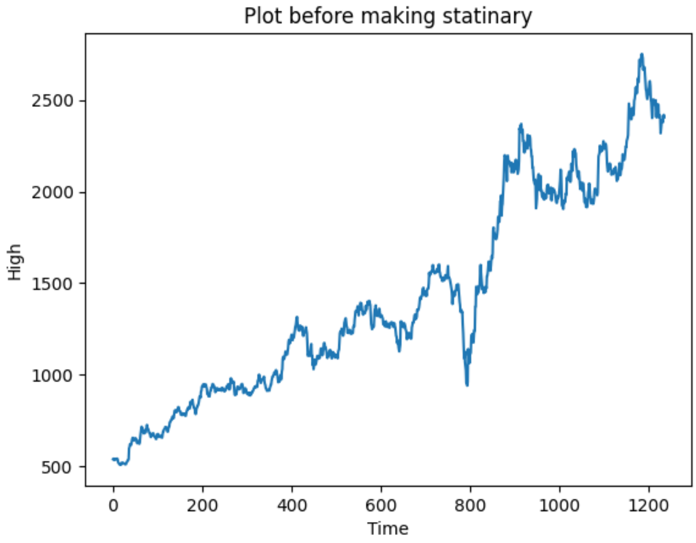

Figure 7.1: Before making stationary

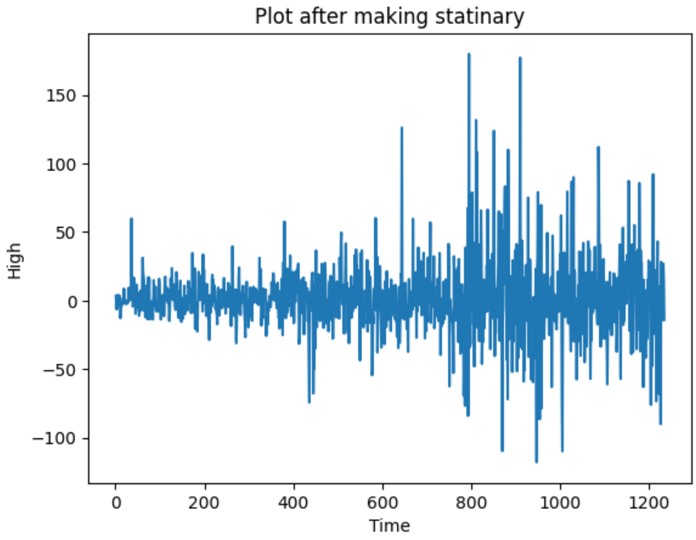

Figure 7.2: After making stationary

Here is some reasons why stationarity is important:

Prediction: Stationarity is crucial for time series prediction. Stationary time series are predictable and easier to model accurately. This is because the underlying statistical properties remain the same over time, allowing us to use past observations to forecast future values.

Statistical Inference: Stationarity is a necessary assumption for many statistical tests and models such as regression, hypothesis testing[56], and ANOVA[64]. This is because these methods rely on the assumption that the data are independent and identically distributed(IID)[6], which is only valid if the data are stationary.

Data Exploration[32]: Stationarity helps in identifying patterns in the data. By removing non-stationarity, it becomes easier to identify trends, seasonality, and other underlying patterns in the data that might be missed otherwise.

Signal Processing[51]: Stationarity is important in signal processing applications where the goal is to extract information from noisy signals. By removing non-stationarity, it becomes easier to identify the underlying signal from the noise.

7.1 Making Time Series Stationary

Stationarity is important because it provides a stable and predictable framework for analyzing data, making it easier to model and predict future outcomes, and to extract meaningful information from the data. To make a non-stationary time series stationary, following techniques are used.

Differencing: Take the first difference of the time series by subtracting each observation from the previous observation. If the time series has a trend, take higher order differences[24].
The first-order differencing of a time series can be defined as the difference between consecutive observations. Mathematically, if Y_t is the value of the time series at time t, then the first-order difference can be defined as:

$$\Delta Y_t = Y_t - Y_{t-1}$$

where ΔY_t represents the first-order difference of Y at time t.

The second-order difference can be defined as the difference between the first-order differences. Mathematically, if ΔY_t is the first-order difference of Y at time t, then the second-order difference can be defined as:

$$\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2}$$

where $\Delta^2 Y_t$ represents the second-order difference of Y at time t.

Higher order differences can also be defined in a similar manner. By applying differencing to a time series, we can often remove trends and seasonality, making the series stationary and enabling the use of statistical models for forecasting and analysis.

Seasonal Differencing: If the time series has a seasonal component, take the seasonal difference by subtracting each observation from the observation at the same season in the previous year. Mathematically, the first-order seasonal difference of a time series Y at time t with a seasonal period of s can be defined as:

$$\Delta_s Y_t = Y_t - Y_{t-s}$$

where $\Delta_s Y_t$ represents the first-order seasonal difference of Y at time t with a seasonal period of s. The second-order seasonal difference of a time series can be defined as the difference between the first-order seasonal differences:

$$\Delta_s^2 Y_t = \Delta_s Y_t - \Delta_s Y_{t-s}$$

where $\Delta_s^2 Y_t$ represents the second-order seasonal difference of Y at time t with a seasonal period of s.

Higher order seasonal differences can also be defined in a similar manner. By applying seasonal differencing to a time series, we can remove the seasonal patterns and make the series stationary, which can be useful for modeling and forecasting.

Logarithmic Transformation: If the time series has a varying variance over time, a logarithmic transformation can be apply to stabilize the variance. Mathematically, the logarithmic transformation of a time series Y can be defined as:

$$\log(Y)$$

where log denotes the natural logarithm.

Once the time series is transformed into a stationary series, you can use statistical models like ARIMA or other machine learning techniques for time series forecasting or analysis.

7.2 Transforming to Normal Distribution

Transforming non-normal data into normal data can be useful in some statistical analyses or machine learning algorithms that assume normally distributed data. There are several methods that can be used to transform non-normal data into normal data, including:

Box-Cox transformation: The Box-Cox transformation is a widely used method for transforming non-normal data into normal data. It involves taking the logarithm or power of the data to normalize it [58].

Log transformation: Taking the logarithm of non-normal data can help normalize it if the data has a positive skew [35].

Square root transformation: This transformation is useful when the data is skewed to the right (positive skew) and has a minimum value of zero [18].

Inverse transformation: The inverse transformation is useful for data that has a minimum value of zero and is skewed to the right.

It's important to keep in mind that transforming data can change its interpretation and make it difficult to communicate results to non-technical audiences. Therefore, it's essential to carefully consider the implications of transforming data before doing so.

7.3 Remove Heteroscedasticity

Heteroscedasticity refers to the situation where the variability of the errors in a statistical model is not constant across the range of predictor variables. This can lead to biased or inefficient estimates of the regression coefficients and standard errors, making it difficult to draw accurate conclusions from the analysis. To remove heteroscedasticity, there are several methods can be used.

Transforming the data: One way to remove heteroscedasticity is to transform the data, such as taking the logarithm or square root of the response variable or predictor variables. This can help to stabilize the variance of the errors[71].

Weighted regression: Another method is to use weighted regression, where observations with higher variance are given less weight in the analysis. This can help to reduce the impact of the heteroscedasticity on the estimates of the regression coefficients[49].

Robust standard errors: Using robust standard errors can help to adjust for the presence of heteroscedasticity. Robust standard errors are less sensitive to the distribution of the errors and can provide more accurate estimates of the standard errors[41].

It's important to diagnose and address heteroscedasticity before interpreting the results of a statistical analysis. Ignoring heteroscedasticity can lead to biased or inefficient estimates and incorrect conclusions.

7.4 Results

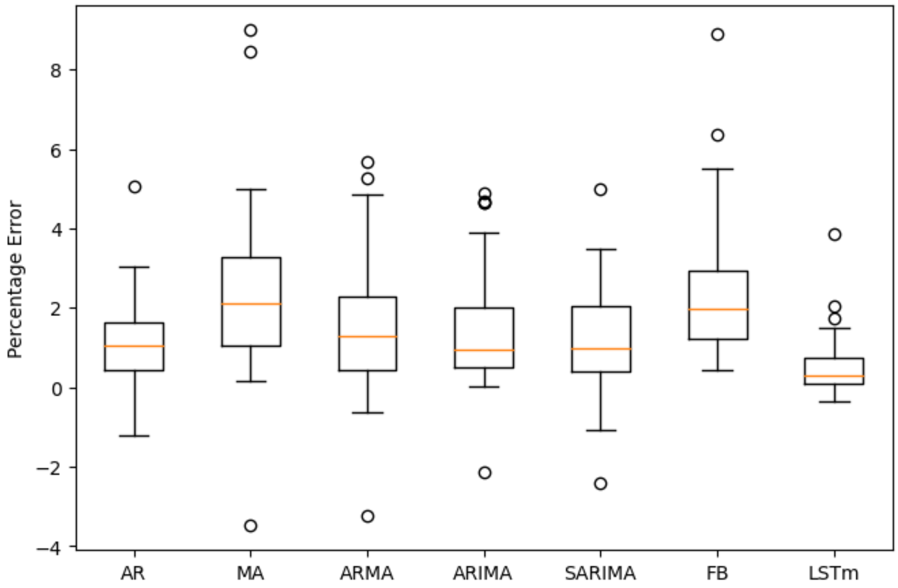

Figure 7.3: Combined error plot of Reliance data before making the data stationary

Figure 7.3 is the box plot of the errors of Reliance data before making the data stationary. Figure 7.4 is the box plot of the errors of Reliance data after making the data stationary. Comparing plots, in Figure 7.3, before making the data stationary LSTM Model performs well than other models but in Figure7.4, after making the data stationary all the models perform equally well. The conclusion is when the data has simple behaviour, that is there is no trend or seasonality in the data then all the time series

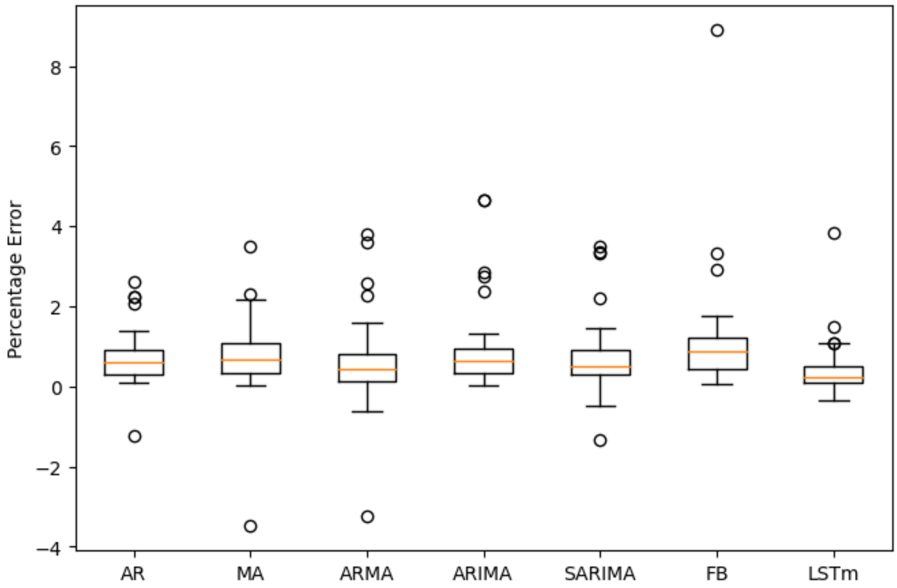

Figure 7.4: Combined error plot of Reliance data after making the data stationary

models perform equally well. But when the complexity increases then the complex models like LSTM perform better than simple models like AR, MA, ARMA, ARIMA, Prophet.

Chapter 8

Portfolio Analysis

A portfolio[60] is a collection of financial assets, such as mutual funds, bonds, stocks, and other investments, owned by an individual or institution. The objective of a portfolio is to diversify investments and spread risk, as well as to maximize returns primarily based on an individual's or institution's financial goals, risk tolerance, and investment horizon.

Portfolios can be managed actively or passively, depending on the investment strategy and goals. Active management involves making investment decisions based on market research and analysis, while passive management involves investing in a pre-selected basket of assets that track a market index.

There are several types of portfolios, including equity portfolios, fixed income portfolios, balanced portfolios, and alternative investment portfolios. Equity portfolios consist mainly of stocks, while fixed income portfolios consist mainly of bonds. Balanced portfolios are a mix of both stocks and bonds, while alternative investment portfolios may include assets such as real estate, private capital, and investment fund.

Portfolios are typically evaluated based on performance metrics such as returns, volatility, and risk-adjusted returns. It's important for investors to regularly review and adjust their portfolios to ensure that they meet their financial goals and risk tolerance.

8.1 Portfolio Optimization

Portfolio optimization[19] is the process of selecting a group of assets such that the expected return of a portfolio will be maximum for a given level of risk. An investor's mission is to build a portfolio that balances risk and return while taking into consideration the investor's goals, time horizon, and risk tolerance.

There are various techniques for optimising a portfolio, such as maximum diversification, lowest variance, and mean-variance optimisation. The most popular technique is mean-variance optimisation, which entails figuring out the expected return, variance, and covariance between each asset in the portfolio. The asset combination that yields the maximum predicted return for a particular level of risk is then found using the optimisation algorithm.

An investor would often utilise statistical analysis and optimisation software to establish an efficient frontier, which represents the set of portfolios with the best predicted return for a given amount of risk, in order to conduct portfolio optimisation. The investor can then select the portfolio that best meets their objectives and risk tolerance, based on the efficient frontier.

It's critical to note that portfolio optimization isn't always a one-time event, however an ongoing exercise that requires continuous tracking and adjustment. As market conditions change, the optimal mix of assets in a portfolio may also change, and adjustments may need to be made to maintain the desired risk-return profile. Figure8.1 is a outline of portfolio analysis.

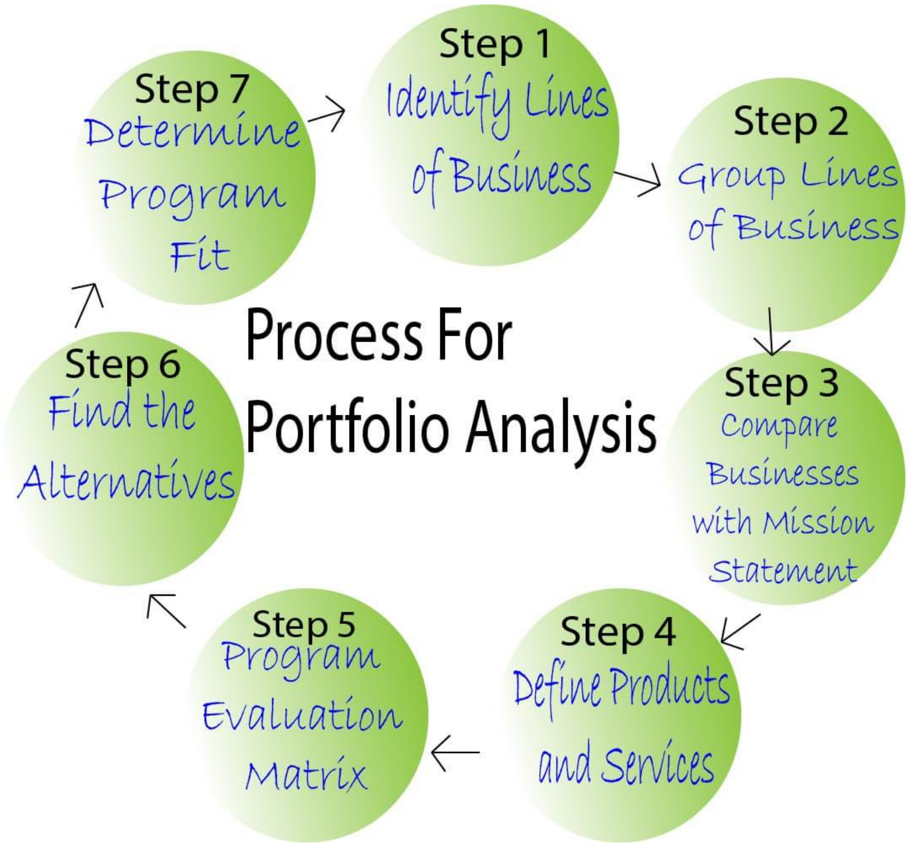
Process For Portfolio Analysis
Step 1 Identify Lines of Business
Step 2 Group Lines of Business
Step 3 Compare Businesses with Mission Statement
Step 4 Define Products and Services
Step 5 Program Evaluation Matrix
Step 6 Find the Alternatives
Step 7 Determine Program Fit

Figure 8.1: Steps of Portfolio Analysis[39]

Fundamental terms in portfolio optimization

Portfolio optimization is the process of selecting a portfolio of assets that maximizes expected returns for a given degree of risk. There are several fundamental terms that are commonly used in portfolio optimization, including:

Expected return: The expected return of an asset is the average return that an investor can expect to earn on that asset over a given period of time.

Risk: Risk is the potential for loss that an investor faces when investing in an asset. It is often measured by the volatility of the asset's returns.

Standard deviation: The standard deviation is a statistical measure that expresses how far a set of data deviates from the mean. The standard deviation is used to assess an asset's risk when optimising a portfolio.

Correlation: A statistical measure known as correlation measures how much two assets move in relation to one another. The correlation between two assets is significant in the context of portfolio optimisation because it determines the degree of diversity that may be obtained by making investments in those assets.

Efficient frontier: The collection of all portfolios that provide the best expected return for a specific amount of risk is represented by the efficient frontier, which is a curve.

Portfolio weight: The portfolio weight is the percentage of a portfolio that is allocated to a particular asset.

Asset allocation: The practise of distributing a portfolio among several asset classes, such as cash,stocks and bonds is known as asset allocation.

Diversification: Diversification is the practice of making an investment in an expansion of assets as a way to reduce the risk of a portfolio.

These fundamental terms are essential for understanding the principles of portfolio optimization and for constructing investment portfolios that meet an investor's risk and return objectives.

8.2 Assets

Assets[17] are economic resources that have value and are owned or controlled by an individual, company, or institution. Assets can be tangible or intangible and can be used to generate revenue or provide future economic benefits.

Tangible assets are physical assets that have a finite lifespan and can be seen, touched, or felt. Property, plant, inventory, and machinery are examples of assets. On the other hand, intangible assets are neither visible or touchable and do not have a physical presence. Patents, trademarks, copyrights, brand names, and goodwill are examples of intangible assets.

Assets are typically classified as current assets or non-current assets. Current assets are belongings that are predicted to be converted into cash within one year or one running cycle, while non-current day belongings are belongings that aren't predicted to be converted into cash within one year or one operating cycle.

Moreover, assets can be categorised as either financial or non-financial. Non-financial assets include property, plant, and equipment, inventory, and other tangible assets, whereas financial assets include investments like stocks, bonds, and mutual funds.

Assets are an important component of a company's financial statements and are used to calculate important financial ratios such as return on assets (ROA) and asset turnover ratio. They also play a critical role in financial planning and investment decision-making for individuals and institutions.

8.3 Returns

Returns[36] refer to the profit or loss generated by an investment over a predetermined amount of time, expressed as a percentage of the initial investment. Returns may be positive or negative, relying on whether the investment generated a profit or incurred a loss. There are several types of returns, including:

Absolute returns: Without taking into consideration the initial investment amount, these are the total returns produced by an investment during a given time period.

Annualized returns: These are the average returns generated by an investment over a specific period of time, annualized to provide a standard measure of performance.

Total returns: These are the sum of all returns generated by an investment over a specific period of time, including capital gains, dividends, and interest payments.

Risk-adjusted returns: These are returns that take into account the level of risk associated with an investment. The most commonly used risk-adjusted return measure is the Sharpe ratio, which compares the excess return of an investment to its volatility.

Returns are an important performance measure for investors, as they provide an indication of how well an investment has performed over a certain period of time. They are used to evaluate the effectiveness of investment strategies and to compare the performance of different investments. Investors also use returns to assess the risk-return trade-off of an investment, as investments with higher returns generally come with higher levels of risk.

8.4 Risk

Risk[14] refers to the uncertainty or variability associated with an investment, and the possibility that the actual outcome may differ from the expected outcome. Risk is a fundamental part of investing and is present in all types of investments, including stocks, bonds, and mutual funds. There are several types of risk, including:

Market risk: The risk that an investment may decline in value due to factors such as changes in economic conditions, interest rates, or geopolitical events.

Credit risk: The risk that an issuer of debt securities may default on their payments.

Liquidity risk: The risk that an investment will be difficult to sell or convert to cash without suffering significant losses.

Inflation risk: The risk that the value of an investment may decline due to inflation eroding its purchasing power.

Currency risk: The risk that changes in exchange rates may have a bad effect on the return on a foreign currency investment.

Political risk: The risk that changes in government policies or instability in a country may negatively impact an investment.

Investors manage risk by diversifying their portfolios across different asset classes, industries, and geographic regions. They also use risk management techniques such as hedging, stop-loss orders, and other risk reduction strategies to limit potential losses.

It's important to note that higher levels of risk are typically associated with higher potential returns, and that investors should carefully consider their risk tolerance and investment goals before making investment decisions.

8.5 Modern Portfolio Theory (MPT)

Modern Portfolio Theory (MPT)[25] is a financial theory developed by Harry Markowitz in the 1950s. MPT is based on the idea that an investor can construct an optimal portfolio by combining different investments in a way that maximizes expected returns for a given level of risk. MPT assumes that traders deliberately try to minimize risks and maximize returns of the portfolio.

MPT starts by defining risk as the variance of the returns of an investment. It then suggests that an investor can minimize risk by diversifying their investments across different asset classes, such as stocks, bonds, and cash. By diversifying, an investor can reduce the overall risk of their portfolio because the performance of one asset class is unlikely to be correlated with the performance of another.

MPT also introduces the concept of the efficient frontier, which is the set of all portfolios that offer the high return for a given level of risk. The efficient frontier is plotted on a graph, with expected return on the y-axis and risk on the x-axis. The investor can then choose a portfolio along the efficient frontier that meets their desired level of risk.

MPT has been widely used by financial professionals to construct investment portfolios for clients. However, MPT has been criticized for its assumptions of rationality and efficient markets, which won't always hold genuine inside the real global. Additionally, MPT does not take into account factors such as market volatility, liquidity, and transaction costs, which can impact the performance of a portfolio.

8.6 Efficient Frontier

The efficient frontier is an idea of Modern Portfolio Theory (MPT). MPT refers to the set of all portfolios, offered the highest expected return of the given level of risk. The efficient frontier is plotted on a graph, with expected return on the y-axis and risk on the x-axis.

The efficient frontier is determined by analyzing the expected returns and risks of various investment portfolios that are constructed by combining classes of different assets, such as stocks, bonds, and cash. Portfolios that lie on the efficient frontier are considered to be the most efficient, as they provide the highest expected return for a given level of risk.

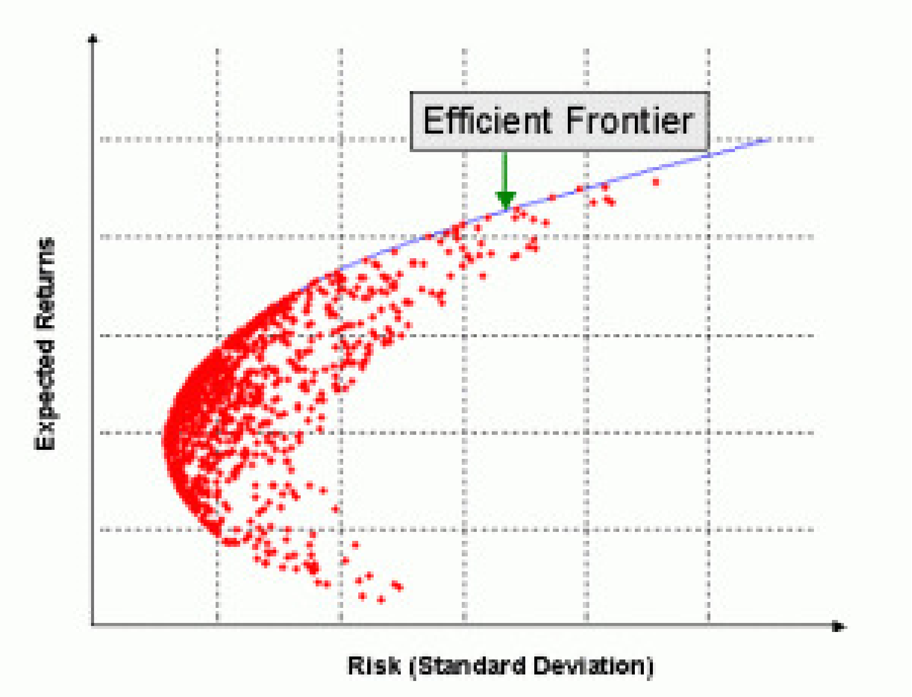

Figure 8.2: Efficient Frontier[40]

Investors can use the efficient frontier to construct portfolios that meet their specific risk and return objectives. For example, an investor who is willing to take on more risk may choose a portfolio that lies further to the right on the efficient frontier, while an investor who is more risk-averse may choose a portfolio that lies closer to the left.

It is important to note that the efficient frontier is based on assumptions of rationality and efficient markets, which may not always hold true in the real world. Therefore, investors should use the efficient frontier as a tool for constructing portfolios, but should also consider other factors when making investment decisions.

8.7 Covariance matrix

A covariance matrix is a square matrix that contains the variances and covariances of a set of variables. In particular, the diagonal elements of a covariance matrix represent the variances of the individual variables, while the off-diagonal elements represent the covariances between pairs of variables.

The covariance matrix is symmetric, since the covariance between X and Y is the same as the covariance between Y and X. Moreover, the diagonal elements are always non-negative, since the variance of a variable is always non-negative.

If we have k variables, then the covariance matrix will be an k x k matrix. The (i,j)-th element of the covariance matrix will be the covariance between the i-th and j-th variables.

The formula for the covariance matrix is as follows:

Let X be an nxp matrix, where each row corresponds to an observation, and each column corresponds to a variable. Let μ be the px1 vector of column means, where the i-th element of μ is the mean of the i-th column of X[43]. Then, the covariance matrix Σ of X is given by:

$$\Sigma = \frac{1}{n}(X - \mu)^T(X - \mu)$$

where $(X - \mu)$ is the centered data matrix obtained by subtracting the column means μ from each column of X, and "T" denotes the transpose operator.

The (i,j)-th element of the covariance matrix Σ is the covariance between the i-th and j-th variables of X, and is given by:

$$\Sigma_{i,j} = \frac{1}{n}\sum_{k=1}^{n}(X_{k,i} - \mu_i)(X_{k,j} - \mu_j)$$

where $X_{k,i}$ denotes the i-th element of the k-th row of X.

Covariance matrices are often used in multivariate analysis, where we want to understand the relationship between multiple variables. They can be used to identify patterns and relationships among the variables, and to construct linear combinations of the variables that capture the maximum amount of variation in the data. Covariance matrices are also used in machine learning algorithms such as principal component analysis (PCA)[16] and linear discriminant analysis (LDA)[68].

8.8 Correlation matrix

A correlation matrix is a square matrix that contains the correlation coefficients between pairs of variables. In contrast to a covariance matrix, a correlation matrix is standardized, meaning that each correlation coefficient is scaled to lie between -1 and 1.

The diagonal elements in a correlation matrix are always equal to 1, since the correlation between a variable and itself is always equal to 1. Moreover, the correlation matrix is symmetric, since the correlation between X and Y is the same as the correlation between Y and X.

The correlation coefficient measures the direction and strength of the linear relationship between two variables. A correlation coefficient is +1 means a perfect positive relationship, a correlation coefficient is -1 means a perfect negative relationship, and a correlation coefficient is 0 means no linear relationship between the variables.

The formula for the correlation coefficient(r) between two variables X and Y[23] can be expressed as:

$$r = \frac{(n\sum XY - \sum X \sum Y)}{\sqrt{\Big((n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)\Big)}}$$

where n is the number of observations, $\sum$ represents the sum of the variable, $\sum XY$ represents the sum of the product of X and Y, $\sum X$ and $\sum Y$ represent the sum of X and Y respectively, $\sum X^2$ represents the sum of the square of X and $\sum Y^2$ represents the sum of the square of Y.

If we have n variables, then the correlation matrix will be an nxn matrix. The (i,j) element of the correlation matrix will be the correlation coefficient between the i-th and j-th variables.

Correlation matrices are often used to identify patterns and relationships among multiple variables. They can be used to understand the degree of association between different variables and to help identify which variables are most strongly correlated with each other. Correlation matrices are also used in statistical inference, where we want to estimate the parameters of a multivariate distribution.

8.9 Sharpe Ratio

The Sharpe ratio is a measure of risk-adjusted performance developed by William F. Sharpe. It is widely used in finance to evaluate the return of an investment compared to its risk[61].

The formula for the Sharpe ratio is:

$$\text{Sharpe Ratio} = \frac{R_p - R_f}{\sigma_p}$$

Where:
- R_p is the average return of the portfolio,
- R_f is the risk-free rate of return,
- σ_p is the standard deviation of the investment's returns.

The Sharpe ratio measures the excess return per unit of risk. A higher Sharpe ratio indicates a better risk-adjusted performance of the investment.

For example, if an investment has an average return of 12% and a standard deviation of 8%, and the risk-free rate of return is 4%, then the Sharpe ratio would be:

Sharpe Ratio = (12% - 4%) / 8% = 1.00

This means that for each unit of risk taken on by the investment, the investor is receiving a return that is 1.00 times greater than the risk-free rate. A higher Sharpe ratio would indicate a better risk-adjusted return.

8.10 Data for Portfolio Analysis

The data set is downloaded from Yahoo finance website. The data set is a group of stock data of ten companies, Reliance, Tesla, Apple, Amazon, Ford Motor, Microsoft, FMC Corporation, Meta Platforms, Ecolab Inc and LyondellBasell.

Table 8.1: Portfolio Data

Date	AAPL	AMZN	ECL	F	FMC	LYB	META	MSFT	RELIANCE	TSLA
2018-01-02	40.88	59.45	127.11	9.8	75.79	83.56	181.41	80.73	883.27	21.36
2018-01-03	40.88	60.2	127.93	9.94	76.76	83.58	184.66	81.11	886.81	21.15
2018-01-04	41.07	60.47	128.97	10.12	77.15	83.99	184.33	81.82	892.14	20.97
2018-01-05	41.53	61.45	129.64	10.29	77.65	85.15	186.85	82.84	895.00	21.1
2018-01-08	41.38	62.34	129.19	10.25	77.72	84.96	188.27	82.92	900.14	22.42

where

Table 8.2: Portfolio Data

Company Code	Company name
AAPL	Apple
AMZN	Amazon
ECL	Ecolab Inc
F	Ford Motor
FMC	FMC Corporation
LYB	LyondellBasell
META	Meta Platforms
MSFT	Microsoft
RELIANCE	Reliance
TSLA	Tesla

Correlation Matrix of the data set:

Table 8.3: Correlation Matrix

	AAPL	AMZN	ECL	F	FMC	LYB	META	MSFT	RELIANCE	TSLA
AAPL	1.00	0.653	0.523	0.393	0.478	0.405	0.574	0.774	0.128	0.473
AMZN	0.653	1.00	0.407	0.304	0.342	0.274	0.605	0.716	0.080	0.437
ECL	0.523	0.407	1.00	0.523	0.525	0.531	0.396	0.574	0.189	0.327
F	0.393	0.304	0.523	1.00	0.463	0.545	0.324	0.395	0.208	0.334
FMC	0.478	0.342	0.525	0.463	1.00	0.616	0.368	0.512	0.213	0.333
LYB	0.405	0.274	0.531	0.545	0.616	1.00	0.295	0.409	0.246	0.265
META	0.574	0.605	0.396	0.324	0.368	0.295	1.00	0.614	0.107	0.337
MSFT	0.774	0.716	0.574	0.395	0.500	0.409	0.614	1.00	0.132	0.469
RELIANCE	0.128	0.082	0.189	0.208	0.211	0.246	0.107	0.134	1.00	0.109
TSLA	0.473	0.437	0.327	0.334	0.333	0.265	0.337	0.469	0.109	1.00

In Correlation Matrix(Table 8.3), the correlation of an asset with itself always equal to 1.

Portfolio Expected Returns:
To Plot the efficient frontier, we calculate the Volatility and returns of assets.

Return and Volatility of the companies:

In Table8.4 Tesla has the maximum risk attached but it also offers the maximum returns. Ford Motor lies approximately in the middle having average risk and return rates. Next, assign different weights for assets to plot the graph of efficient frontier and calculate the return and volatility of that particular portfolio. Since the sum of weights must be 1, the weights of individual assets are divided by their cumulative sum.

Return and Volatility with weights of the companies:

where Returns are the product of individual expected returns of asset and its weights, Volatility is the annual standard deviation.

Table 8.4: Return and Volatility

Company	Returns	Volatility
AAPL	0.448771	0.328925
AMZN	0.130113	0.351677
ECL	0.044299	0.296493
F	0.302660	0.411333
FMC	0.222586	0.331976
LYB	0.065341	0.410244
META	0.121424	0.435379
MSFT	0.311361	0.304900
RELIANCE	0.240494	0.307538
TSLA	1.884653	0.642530

Table 8.5: Return and Volatility with weights

	Returns	Volatility	AAPL	AMZN	ECL	F	FMC	LYB	META	MSFT	RELIANCE	TSLA
0	0.45	0.25	0.016	0.05	0.08	0.209	0.091	0.029	0.062	0.071	0.23	0.14
1	0.32	0.261	0.16	0.15	0.03	0.15	0.127	0.075	0.15	0.0001	0.08	0.057
2	0.31	0.26	0.062	0.17	0.09	0.11	0.093	0.12	0.16	0.034	0.07	0.079
3	0.35	0.27	0.06	0.034	0.08	0.13	0.13	0.12	0.26	0.08	0.03	0.06
4	0.39	0.27	0.14	0.02	0.04	0.072	0.13	0.23	0.08	0.13	0.026	0.106

In Table 8.5 ,a number of portfolios with different weights, returns and volatility. Plotting the returns and volatility from this dataframe will give the efficient frontier for portfolio.

The Efficient Frontier Plot

On the graph Figure 8.3, each point on the line (left edge) represents an optimal portfolio of stocks that maximises the returns for any given level of risk. The point which are the portfolios are sub-optimal for a given risk level. For every point offers higher returns for the same risk.

1. Minimum risk (left most point)
2. Maximum returns (top most point)

From Table8.6 the minimum volatility is in a portfolio where the weights of Apple, Amazon, Ecolab Inc, Ford Motor, FMC Corporation, LyondellBasell, Meta Platforms, Microsoft, Reliance, and Tesla are 4%, 27%, 26%, 1%, 5%, 23%, 1%, 6%, 27% and 8% respectively.

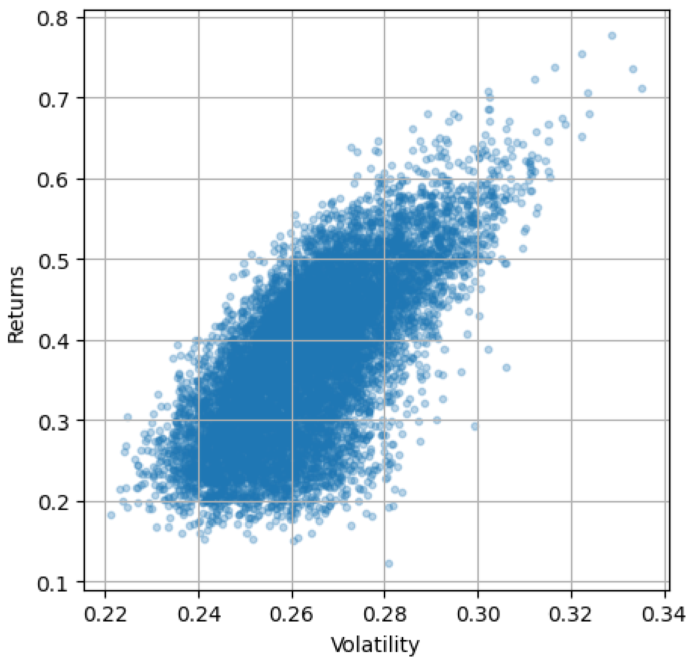

Figure 8.3: Efficient Frontier plot

Table 8.6: Minimum volatility

Returns	0.181
Volatility	0.221
AAPL weight	0.043
AMZN weight	0.27
ECL weight	0.258
F weight	0.0085
FMC weight	0.046
LYB weight	0.023
META weight	0.0084
MSFT weight	0.061
RELIANCE weight	0.271
TSLA weight	0.0084

Plot the minimum volatility portfolio

In Figure 8.4, the red star denotes the most efficient portfolio with minimum volatility. Any point to the right of efficient frontier boundary is a sup-optimal portfolio. The plot gave portfolio with minimum volatility, but the return on this portfolio is pretty low. Any investor wants to maximize his return, even if it is a trade-off with some level of risk. Use Sharpe Ratio to find the optimal risky portfolio and optimize portfolio to the maximum.

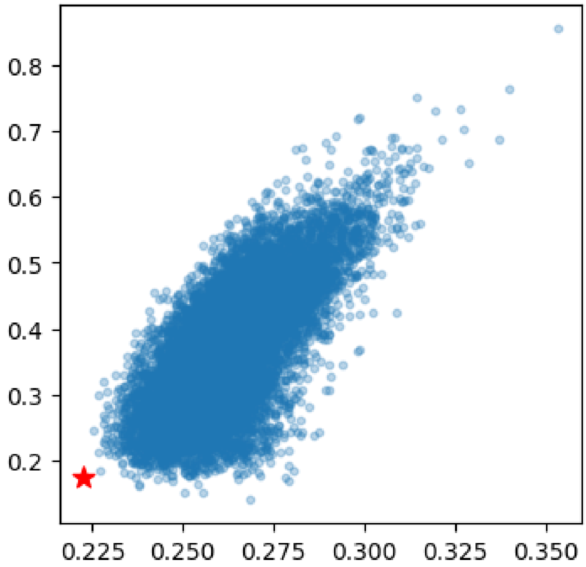

Figure 8.4: Efficient Frontier with minimum volatility

Table 8.7: Optimal Portfolio

Returns	0.78
Volatility	0.33
AAPL weight	0.23
AMZN weight	0.07
ECL weight	0.01
F weight	0.05
FMC weight	0.10
LYB weight	0.013
META weight	0.093
MSFT weight	0.011
RELIANCE.NS weight	0.053
TSLA weight	0.3

From Table8.6 and Table8.7, while the difference in risk between minimum volatility portfolio and optimal risky portfolio is just 11%, the difference in returns is a whopping 60%.

The graph of efficient frontier for Optimal Portfolio

Figure8.5 is the efficient frontier of optimal portfolio. The green star represents the optimal risky portfolio.

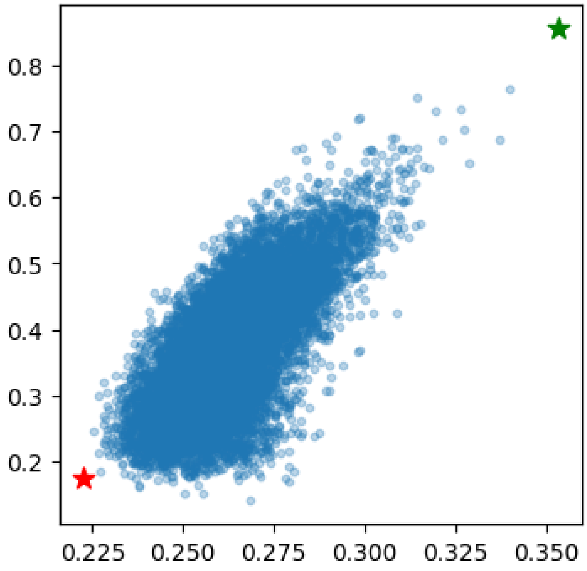

Figure 8.5: Efficient Frontier of Optimal Portfolio

Chapter 9

Conclusion

The objective of the project was apply some time series models and deep learning models like AR, MA, ARMA, ARIMA, SARIMA, FB Prophet, LSTM are applied on stock data. Compare the errors of the models and predict future stock price based on the best model obtained. Also perform a portfolio analysis on a group of stock data and find the optimal portfolio, which will help an investor to built a better portfolio. The conclusion is the traditional time series models such as AR, MA, ARMA, ARIMA, and SARIMA are based on linear assumptions and may not be able to capture the complexity of the data. FB Prophet is a more advanced model than traditional time series models but is still limited in its ability to capture complex patterns in the data. LSTM is a suitable choice for time series analysis and forecasting when dealing with complex data. But all models perform equally well on a simple data.

The conclusion of portfolio analysis depends on the specific analysis conducted and the goals of the investor. However, portfolio analysis can provide insights into the risk and return characteristics of a portfolio, as well as its diversification and allocation of assets. By analyzing the historical performance and volatility of the assets in a portfolio, investors can identify opportunities to improve its risk-adjusted return. By assessing the correlation between different assets, investors can identify opportunities to diversify the portfolio and reduce its overall risk. Portfolio optimization techniques identifies the optimal allocation of assets that maximizes the return for a given level of risk. Overall, portfolio analysis can be a valuable tool for investors to manage risk and maximize returns.

Bibliography

[1] Avoiding data leakage machine learning, April 2023. URL.

[2] Critical value, April 2023. URL.

[3] Cyclic and seasonal time series, April 2023. URL.

[4] Efficient frontier, April 2023. URL.

[5] finance.yahoo.com, April 2023. URL.

[6] Linear regression, April 2023. URL.

[7] Linear regression in machine learning, April 2023. URL.

[8] Moving average models, April 2023. URL.

[9] Pre process time series data, April 2023. URL.

[10] Preprocessing time series data, April 2023. URL.

[11] Structure of the lstm model, April 2023. URL.

[12] Time series analysis the basics, April 2023. URL.

[13] Time series characteristics, April 2023. URL.

[14] Types of risks in stock market, April 2023. URL.

[15] Unit root test, April 2023. URL.

[16] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[17] Adam Barone. Asset, April 2023. URL.

[18] Maurice S Bartlett. The square root transformation in analysis of variance. *Supplement to the Journal of the Royal Statistical Society*, 3(1):68–78, 1936.

[19] Michael J Best. *Portfolio optimization*. CRC Press, 2010.

[20] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[21] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

[22] Robert Costrell, Eric Hanushek, and Susanna Loeb. What do cost functions tell us about the cost of an adequate education? *Peabody Journal of Education*, 83(2):198–223, 2008.

[23] Joan Daemen, Vincent Rijmen, Joan Daemen, and Vincent Rijmen. Correlation matrices. *The Design of Rijndael: The Advanced Encryption Standard (AES)*, pages 91–113, 2020.

[24] David A Dickey. Stationarity issues in time series models. *SAS Users Group International*, 30, 2015.

[25] Edwin J Elton and Martin J Gruber. Modern portfolio theory, 1950 to date. *Journal of banking & finance*, 21(11-12):1743–1759, 1997.

[26] Thomas B Fomby. Exponential smoothing models. *Mannual sas/ets software: time series forecasting system*, 6:225–235, 2008.

[27] Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[28] Alexander I Galushkin. *Neural networks theory.* Springer Science & Business Media, 2007.

[29] Abdiel Goni. Linear model for optimization, April 2023. URL.

[30] Shashank Gupta. Univariate bivariate multivariate analysis, April 2023. URL.

[31] James Douglas Hamilton. *Time series analysis.* Princeton university press, 2020.

[32] Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. Overview of data exploration techniques. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 277–281, 2015.

[33] Gartner Inc. Prediction models traditional versus machine learning, April 2023. URL.

[34] Gwilym M Jenkins and Athar S Alavi. Some aspects of modelling and forecasting multivariate time series. *Journal of time series analysis*, 2(1):1–47, 1981.

[35] Oliver N Keene. The log transformation is special. *Statistics in medicine*, 14(8):811–819, 1995.

[36] Will Kenton. Rate of return, April 2023. URL.

[37] Jae H Kim. Forecasting autoregressive time series with bias-corrected parameter estimators. *International Journal of Forecasting*, 19(3):493–502, 2003.

[38] Ajitesh Kumar. Different types of time series forecasting models, April 2023. URL.

[39] Saakshi Malhotra. Portfolio analysis in strategic management, April 2023. URL.

[40] Manish. Auto-regressive model, April 2023. URL.

[41] Mohammad Ali Mansournia, Maryam Nazemipour, Ashley I Naimi, Gary S Collins, and Michael J Campbell. Reflection on modern methods: demystifying robust standard errors for epidemiologists. *International Journal of Epidemiology*, 50(1):346–351, 2021.

[42] Radu Manuca and Robert Savit. Stationarity and nonstationarity in time series analysis. *Physica D: Nonlinear Phenomena*, 99(2-3):134–161, 1996.

[43] Covariance Matrix, Marion R Reynolds Jr, and Gyo-Young Cho. Multivariate control charts for monitoring the mean vector and covariance matrix. *Journal of Quality Technology*, 38(3):230–253, 2006.

[44] Sidra Mehtab and Jaydip Sen. A time series analysis-based stock price prediction using machine learning and deep learning models. *International Journal of Business Forecasting and Marketing Intelligence*, 6(4):272–335, 2020.

[45] Baligh Mnassri. Jena climate dataset, April 2023. URL.

[46] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[47] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[48] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

[49] Allan Aasbjerg Nielsen. Least squares adjustment: Linear and nonlinear weighted regression analysis. *Danish National Space Center/Informatics and mathematical modelling, Technical Univ. of Denmark*, 2007.

[50] Naveen (NNK). pandas get day-month and year from datetime, April 2023. URL.

[51] Sophocles J Orfanidis. *Introduction to signal processing*. Prentice-Hall, Inc., 1995.

[52] Milan Paluš and Martin Vejmelka. Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections. *Physical Review E*, 75(5):056211, 2007.

[53] Marco Peixeiro. Time series forecasting with sarima, April 2023. URL.

[54] Thirasha Praween. Basic understanding of cost function formula, April 2023. URL.

[55] Tommaso Proietti and Diego J Pedregal. Seasonality in high frequency time series. *Econometrics and Statistics*, 2022.

[56] Adrian E Raftery, WR Gilks, S Richardson, and D Spiegelhalter. Hypothesis testing and model. *Markov chain Monte Carlo in practice*, pages 165–187, 1995.

[57] Abhilash Ramachandran. Forecasting methods, April 2023. URL.

[58] Remi M Sakia. The box-cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(2):169–178, 1992.

[59] Prashant Sharma. Different types of regression models, April 2023. URL.

[60] William F Sharpe. Portfolio analysis. *Journal of Financial and Quantitative Analysis*, 2(2):76–84, 1967.

[61] William F Sharpe. The sharpe ratio. *Streetwise–the Best of the Journal of Portfolio Management*, 3:169–185, 1998.

[62] Robert H Shumway, David S Stoffer, Robert H Shumway, and David S Stoffer. Arima models. *Time Series Analysis and Its Applications: With R Examples*, pages 75–163, 2017.

[63] Simplilearn. Time series analysis, April 2023. URL.

[64] Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

[65] Stephanie. Order of integration, April 2023. URL.

[66] Emily Stevens and England. Data visualization, April 2023. URL.

[67] Wallstreetmojo Team. Stock market, April 2023. URL.

[68] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. Linear discriminant analysis: A detailed tutorial. *AI communications*, 30(2):169–190, 2017.

[69] Tqp. Histogram, April 2023. URL.

[70] Yugesh Verma. Decompose time series, April 2023. URL.

[71] George CS Wang and Charles K Akabay. Heteroscedasticity: How to handle in regression modeling. *The journal of business forecasting*, 13(2):11, 1994.

[72] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

[73] Annette Witt, Jürgen Kurths, and A Pikovsky. Testing stationarity in time series. *physical Review E*, 58(2):1800, 1998.