Stock Market Prediction Using Different Timeseries Analysis Techniques Using R

Nidhi Shukla

Student No : 21071541

Supervised By: Ben Derric

Word Count: 12892

Independent project report submitted for the Masters of Science in Data Science

University of the West of England.

March 2023

**Acknowledgements**

This report is my master thesis, which I wrote and implemented during the final semester of my Master in Data Science at the University Of The West Of England. The work began on October 10th, 2022, and ended on March 10th, 2023. My supervisor, Mr Ben Derric, helped me conceptualise the problem description. His time, guidance and effort in supervising the project has been an inspiration, and his passion for giving advice, assistance, and criticism has substantially enhanced the quality and scope of the work

**Abstract**

Human curiosity to know the future leads to developing models to forecast which helps us prepare for adverse conditions or make the most out of an opportunity. The term stock market refers to several exchanges in which shares of publicly held companies are bought and sold. The stock market process is full of uncertainty; hence stock price forecasting is very important in finance and business. The entire idea of forecasting stock markets is to minimize risks and take profitable positions. Stock markets are notably volatile, dynamic, and non-linear. Market forecasts offer great profit opportunities and are a fundamental motivator for most researchers in the field, but Accurate forecasting of stock prices depends on several (macro and micro) factors like Politics, global economic conditions, unexpected events, company financial performance, etc. hence is difficult. Time series analysis has significance in financial analytics and forecasting and it can be utilized in any field. In finance, time series analysis is used for financial forecastings such as stock prices, assets, and commodities. In this paper, we used autoregressive integrated moving averages (ARIMA), neural networks (NN), Holt-Winter, Naïve, Prophet, KNN to forecast the closing price data of FTSE from 02/01/2009 to 11/11/2022. All models are evaluated using mean squared error (RMSE) and mean absolute percentage error (MAPE). **The results showed that ARIMA allowed to achieve over 80% accuracy in predicting stock index.**

**Contents**

[**Chapter 1. INTRODUCTION AND OBJECTIVES** 7](#_Toc130064859)

[**1.1 Introduction** 7](#_Toc130064860)

[**1.2 Scope** 7](#_Toc130064861)

[**1.3 Current Practices** 7](#_Toc130064862)

[**1.4 Objective** 9](#_Toc130064863)

[**1.4 Success Criteria** 10](#_Toc130064864)

[**1.5 Ethical, Legal and Professional Consideration** 10](#_Toc130064865)

[**1.6 Deliverables** 11](#_Toc130064866)

[**Chapter 2: Literature Review and Related Work** 12](#_Toc130064867)

[**A.** **Stock Market and Time series fundamentals** 12](#_Toc130064868)

[**B.** **Machine Learning Techniques** 12](#_Toc130064869)

[**C.** **Time Series Techniques** 13](#_Toc130064870)

[**D.** **AI and Deep Learning Techniques** 13](#_Toc130064871)

[**E.** **Accuracy matrix** 14](#_Toc130064872)

[**Chapter 3. Methodology** 15](#_Toc130064873)

[**Chapter 4. Baseline and proposed model** 17](#_Toc130064874)

[**4.1 Baseline models** 17](#_Toc130064875)

[**4.2 Proposed Models** 18](#_Toc130064876)

[**Chapter 5: Implementation** 25](#_Toc130064877)

[**5.1 Project Understanding** 25](#_Toc130064878)

[**5.2 Data Mining And Processing** 25](#_Toc130064879)

[**5.3. Data Exploration And Data preparation** 27](#_Toc130064880)

[**5.4 Model Training** 31](#_Toc130064881)

[**Chapter 6 Result and Conclusion** 43](#_Toc130064882)

[**6.1 EDA Results** 43](#_Toc130064883)

[**6.2 Model Performance** 43](#_Toc130064884)

[**6.3 Model Comparison Analysis** 47](#_Toc130064885)

[**6.4 Conclusion** 49](#_Toc130064886)

[**6.5 Future Work** 49](#_Toc130064887)

[**6.6 Learning Reflection** 49](#_Toc130064888)

[References 51](#_Toc130064889)

[Appendix 53](#_Toc130064890)

# **Chapter 1. INTRODUCTION AND OBJECTIVES**

This chapter introduces and focuses on the stock market forecasting problem that this study seeks to address. It includes an overview of specific goals and success criteria, consideration of ethical implications, and information required to access source code repositories.

## **1.1 Introduction**

A stock index represents a series of stock prices. The index is calculated from the prices of defined stocks and changes may reflect the overall performance of the stocks included in the index. In other words, **a stock index is a weighted average market value of a number of firms compared with the value on the base trading day (Bodie et al., 2013)**. Stock indices are calculated based on a few representative publicly traded stocks. To some extent, it can reflect price movements across financial markets and is therefore used as an important indicator of a country's future macroeconomic performance. **Accurate stock index forecasts are paramount to de-risking decisions by providing some key reference information.(Yan B., Aasma M., 2020)** Accurate forecasting of stock prices is very difficult. Despite the volatility, stock prices are not just randomly generated numbers. There is a lot of data to find patterns. So, financial analysts, researchers, and data scientists continue to search for analytical techniques to spot stock market trends. This gave rise to the concept of algorithmic trading, where orders are executed using automated pre-programmed trading strategies. It is also data generated on day to day basis, with fixed time intervals, hence they can be analyzed as a series of time-discrete data.

**What is time series data?**

**Time series forecasting is used to forecast future values ​​based on previously observed values ​​and is one of the best tools for trend analysis and future forecasting. This is recorded at regular time intervals and the order of these data points is important (Booth David.,2012).** Therefore, forecast models based on time series data include time as an independent variable. The output of the model will be the predicted value or classification at a given point in time. Time series forecasting (predicting future values ​​based on past values) is widely used in many real-world applications, such as weather forecasting and financial market forecasting. The stock market is a representative field that presents time-series data, and many researchers have studied it and proposed various models.

The general stock market trends in society are so dangerous for investing that most people are unable to make decisions based on general trends. For new investors, general research related to stocks or the stock market is not enough to make a decision. The seasonality and steady flow of each index help existing and new investors understand the stock market and make investment decisions.

## **1.2 Scope**

The scope of our study is to predict the stock index data using different TS algorithms and study their prediction efficiency. The proposed model will aid companies and individuals in take informed investment decisions. Currently, we are focusing on European Stock Indices but model can be used to predict any registered stock all over the world whose historical data is available for use on financial sites like Yahoo Finance or Google

## **1.3 Current Practices**

Fundamental analysis and technical analysis are the two most commonly used strategies to predict the stock market trend. Fundamental analysis focuses on company and market sentiments, policies, profits/loss, assets, takeovers, etc. whereas technical analysis uses a mathematical approach in attempts to identify opportunities by finding statistical trends based on historical value. (Chen, Yuh-Jen & Chen, Yuh-Min., 2013) Fundamental analysis is short-term and unpredictable so our focus is technical analysis.

Many algorithms by researchers were developed to predict prices with better accuracy. Machine learning, deep learning, neural network, time-series even hybrid.

The method we choose depends on many factors, including the context of our forecast, the relevance and availability of our historical data, the level of accuracy our need, the time period used for forecasting, and the cost-effectiveness of the forecast and time available for analysis. These factors should always be weighed on different levels.

Statistical models were first used to predict financial stock markets with some success. However, they have limitations as they assume linear and stationary series, which conflict with the dynamic and non-linear characteristics of real stock markets. Deep learning models can overcome the shortcomings of traditional statistical models in forecasting, but in some complex and dynamic financial systems, they are susceptible to noise and cannot discover hidden features of time series becomes difficult, learning ability declines, and limited predictions lead to accuracy.

**Advantages and disadvantages of various algorithms used by researchers.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Data | Purpose | Advantage | Disadvantage | Reference |
| Regression | Linear data, time-series | Analysis, classification, prediction | Easily scalable for linear data, easy to interpret result | Overfitting | Dr. P. K. Sahoo, Mr. Krishna charlapally  (2015) |
| FCM: Fuzzy c mean | Non-time series, Time-series | Clustering | Effective for small to medium datasets | Sensitive to noise - Has problems with handling high-dimensional datasets | Suganya and Shanthi (2012); |
| ARIMA: Autoregressive integrated moving average model | Time-series | Forecasting, clustering | Works well for linear time series and short-run forcasting | Ineffective for small data, time-consuming | Selvin et al. (2017) |
| LSTM: Long Short-Term Memory | Non-time series, Time-series | Analysis, classification, prediction | Makes good predictions as it analyses the interactions and hidden patterns within the data, self learning algo | Lacks a mechanism to index the memory while writing and reading the data The number of memory cells is linked to the size of the recurrent weight matrices | Selvin et al. (2017) |
| Holt winter | Non-time series, Time-series | Classification and prediction | Better forecast for close observations | Forecast lags behind actual trend, cant work with cyclic and seasonal data | Gelper, Sarah, Fried Roland & Croux, Christophe. (2010) |
| Neural network | Non-time series, Time-series | Classification and prediction | useful to show the time relationships between the inputs and outputs | Difficult to train,  The process is slow then ml | Papageorgiou, Markos, Kotsialos Apostolos & Poulimenos, Antonios. (2005). |
| RNN | Financial timeseries, non time series | Classification and prediction | Easy to implement, useful for real time predictiondoes not require iterative training | Requires memory to store model, huge size, added expense. | Al-Mahasneh et al. (2018) |
|  |  |  |  |  |  |

Though we studied the pros and cons of many algorithms which can be used for prediction our main focus was comparing different time-series model (due to time constraints only five models are created). There are proposals in the literature based on time series analysis and decomposition for predicting the future value of stocks, which yielded a high level of forecast accuracy.**[2]**

## **1.4 Objective**

The main aim and objective of this project is to help traders navigate the ever-changing stock market by providing a prediction investment tool using machine learning. The initiative intends to provide the newest machine learning technologies to retail investors who have not yet been exposed to these technologies. Our research's objective was to analyze FTSE stock prices from European Union stock markets over a lengthy 13-year period and create a reliable forecasting framework for estimating the value of the EU Index (FTSE). We have hypothesized that machine learning or deep learning models can successfully use these learnt features to effectively use aspects of previous movement patterns of daily index values to properly forecast future index values for a series. For the purposes of our current argument, we gave all the time series models we used a prediction horizon of two months, and we demonstrated using a number of metrics that these models are capable of reasonably accurate forecasting of the values of the FTSE index.

1. Which algorithms to use?

After researching various time-series models I have used

* **ARIMA**: A statistical model known as ARIMA is effective at forecasting time series, particularly for short-term predictions. In this study, utilizing historical stock market data and the ARIMA model, we offer a model for predicting stock market trends based on technical analysis. This model will automate the direction of future stock price indices and aid financial specialists in determining when is the best time to buy and/or sell stocks. R programming is used to create visualizations of the results. The results show ARIMA model has a strong potential for short-term prediction of stock market trend
* **Holt-Winter**: HW focuses on 3 major aspects for performing the predictions, average value, trend and seasonality. These aspects are types of exponential smoothing and hence the hold winter’s method is also known as triple exponential smoothing. In this study, with historical data and HW model, we conclude that HW gives the highest accuracy among all models generated still it has a major limitation: the multiplicative feature of the seasonality.
* **Naïve.:** In this study, utilizing historical stock market data and the Naïve model, we offer a model for predicting stock market trends based on technical analysis which was quick and fairly accurate.
* **Neural Network:** The ability of neural networks to detect nonlinear relationships in their input data makes them ideal for modeling nonlinear dynamic systems such as the stock market. However, neural networks are not perfect at predicting, It outperforms all other methods and helps us better understand dynamic and chaotic systems like the stock market.
* **KNN**. They show the similarity and time relationships between the inputs and outputs. But were difficult to train also process was slow then ml
* **Prophet**. Facebook created Function works with all types of seasonality, gives high accuracy ratio but is time consuming.

1. Which factors to consider while creating a model?

Stock indices depend on various external factors, like recession, government policies, company reputation, online sentiments, news, and more. Currently, we are not focusing on any of these factors except historical data.

1. How to measure their accuracy?

Matrices like rmse, mape, and mae are used to measure the performance of models.

1. What can be done to improve prediction?

As we have not taken external factors into consideration, we can use them like Twitter sentiment data. We can create two-layer deep learning model wherein the first layer processed data and Twitter data are merged, and processed and then we can implement our time series models, to improve the performance we can apply ARIMA + LSTM or Prophet + GARCH .

## **1.4 Success Criteria**

The success Criteria of any model depend upon its accuracy and timely prediction. Stock Market is highly volatile with heavy financial risks involved. Moreover, it changes each second hence time is an important constraint. Hence quickest model with the highest accuracy (85%) can be considered as a successful model.

## **1.5 Ethical, Legal and Professional Consideration**

The Dataset used in the study is available online on all finance websites like google finance and yahoo finance for research and analysis purposes. This work contains no personal data of any sort, so there are no ethical, legal, or professional considerations regarding its use.

## **1.6 Deliverables**

Model have been developed using R programming Language. Compiled and executed in Rstudio generating and executable rmd file. An Executable Jupyter Notebook(r-kernel) with proper documentation has also been created. All of these files are available on git ().

# **Chapter 2: Literature Review and Related Work**

In this Chapter, we will discuss about different research papers, blog, articles, refered to acquire relevant knowledge regarding the subject.

## **Stock Market and Time series fundamentals**

**What are stocks?**

Shares, also known as stocks, are securities that represent partial ownership of the issuing company. A unit of stock, called a "share", gives the owner a portion of the company's assets and a profit equal to the number of shares held. (investpedia.com). The determinants of stock market movements can be broadly categorized to: (i) macroeconomic determinants and (ii) institutional determinants. (Ho, Sin-Yu & Iyke, Bernard. 2017).

Nyasha, Sheilla & Odhiambo, Nicholas. (2013) in there paper showed the origins and development and reforms of the UK stock market. The UK stock market has evolved in terms of market capitalization, total shares traded and turnover. Though UK equity market has evolved over the years, but is subject to a wide range of challenges, including the uncertainty associated with new and regulatory changes that prevail domestically and internationally, and the sovereign debt crisis that has destabilized the UK equity market. still facing.

**What is Timeseries**

Time series forecasting is used to forecast future values ​​based on previously observed values ​​and is one of the best tools for trend analysis and future forecasting. This is recorded at regular time intervals and the order of these data points is important (Booth David.,2012). Time series data typically show patterns such as trends, seasonal variations, irregular cycles, sporadic changes in levels, and variability in economics, environmental, medicine, and other sciences. The purpose of investigating such series is often to detect unexpected interventions, assess the impact of known exogenous interventions, and extrapolate dynamic patterns in data to predict future observations. (Jose, Jonathan. 2022)

Over the past two decades, many crucial changes have taken place in the financial market environment. The development of powerful communication and trading capabilities has expanded investor options. Stock market forecasts involve predictions of the future performance of stocks on financial exchanges. Stock return forecasting is an important financial topic that has attracted the attention of researchers for many years. This includes the assumption that historically published fundamental information has some predictive relationship with future stock returns. Predicting stock prices and financial markets is one of the biggest challenges for the AI ​​community. Various technical, fundamental, and statistical indicators have been proposed and used with different results. Several methods have been proposed to forecast the market and assist in decision-making. In recent years, ML has been used in many research areas with great success. It changed the way investors use information, providing the best possible analysis for all types of investments. Many researchers have used ML algorithms to analyze historical financial data and other relevant information (such as economic conditions) to create tools to support investment decision-making.

## **Machine Learning Techniques**

Chou and Nguyen (2018) used nonlinear forecasting models to predict the stock prices of Taiwanese construction companies. The hybrid system exhibits excellent predictive performance and improves the overall return on investment performance. The proposed model showed a promising forecasting method for highly nonlinear time series whose patterns are difficult to capture with conventional models.

ML is an important tool for supporting financial investments. In general, statistical models assume that there is a linear correlation structure between time series values. However, the nature of time series in stock markets is nonlinear, unstable, chaotic, and highly noisy (Alves LM. 2018). Time series in the stock market analysis are time-series collections of observations such as daily sales totals and stock prices (Fu Jl. 2005). The popularity of ARIMA models stems from their statistical properties and superior Box-Jenkins model-building methodology. However, ARIMA models cannot capture nonlinear patterns, and it is not always practical to use linear models to simulate complex real-world problems (Zhang 2003). (Kumar D, Sarangi, P.K. and Verma, R.,2020) implemented the advent of machine learning and greater computer power, programmed techniques of prediction have shown to be more accurate in predicting stock values.

## **Time Series Techniques**

According to Zhong and Enke (2017), the groups of statistical approaches that fall into the category of univariate analysis for using time series as input variables are autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA), Generalized Autoregressive Conditional Heteroskedastic (GARCH) volatility, and Smooth Transition Autoregressive (STAR) models. ARMA combines autoregressive (AR) models, which attempt to explain momentum and mean-reverting effects commonly observed in trading markets, and moving average (MA) models, which attempt to capture observed shock effects over time increase. The main limitation of the ARMA model is that it does not take into account volatility clustering, an important empirical phenomenon in many financial time series. ARIMA is a natural extension of the class of ARMA models, allowing nonstationary series to be reduced to stationary series. ARIMA (Box et al. 2015) fits time series data to predict future points. Zhong and Enke (2017) describe another group of statistical approaches that typically use multiple input variables, including linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and regression algorithms.

Oluoch, P & Abade, Elisha. (2020). aimed to solve stock price data analysis problems using Holt-Winters Smoothening Stock Prediction Technology. This predictive technology makes it easier for stock investors on the Nairobi Stock Exchange to make sound financial decisions, allowing many to see a return on investment (ROI).

## **AI and Deep Learning Techniques**

Artificial intelligence techniques and metaheuristic optimization algorithms are powerful, but practitioners must be able to perform extensive manual operations. Users, such as financial market traders, are keen to get results that support their decision-making in a convenient way. The computational cost of such advanced algorithms is high due to their complexity. Authors Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, T. in 2007 conducted an empirical study to build a stock trading alert system using a backpropagation neural network (BPNN). Their NN was codename NN5. The system was trained and tested on historical price data of Hong Kong and Shanghai Banking Holdings from January 2004 to December 2005. Experimental results show that the implemented system can determine the direction of short-term price movements with an accuracy of about 74%. Authors Enke, D. and Thawornwong, S. in 2005 presented an approach to predict stock market returns using data mining techniques and neural networks. In this research, we tried to verify the predictive power of financial and economic variables by applying the method of variable relevance analysis in machine learning to data mining. The authors investigated the effectiveness of neural network models used for level estimation and classification. The results showed that trading strategies guided by neural network classification models produced higher profits than those proposed by other strategies with the same risks.

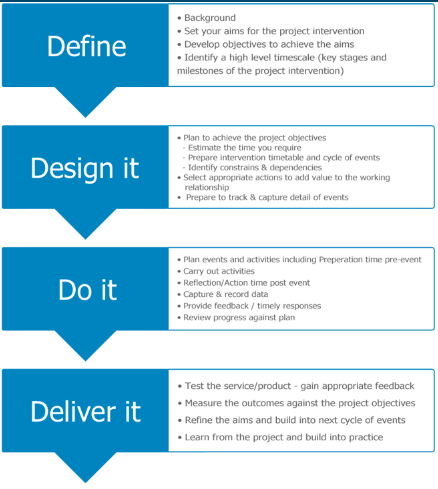
The Random Forest and Neural Network methods were used to forecast the following day's closing price for five businesses from various industries. (Shastry, January 2021) suggested method predicts share price utilising market data and machine learning techniques such as recurrent neural network called Long Short-Term Memory, and weights are adjusted for each data point using stochastic gradient. In compared to presently offered stock price prediction algorithms, our method will produce more accurate results. To influence the graphical outputs, the network can be trained and assessed using varying sizes of input data. (Chen, 2020) illustrate the CART and Autoregressive Series algorithms with a 35% return rate, a Highest Sharpe of 2.13, and a max drawdown of 0.34. The LSTM portfolio had a greater total return rate of 35% than the Australian market index, which had a return rate of 21%. The Long Short-term Memory predicted stock prices more accurately than the Autoregressive time - series data and regression (CART).

## **Accuracy matrix**

The evaluation metrics evaluate the generalization ability of the trained classifier. Evaluation metrics are used to measure and summarize the quality of a trained classifier when tested on unseen data. Accuracy, or error rate, is one of the most commonly used metrics in practice and is used by many researchers to assess a classifier's ability to generalize. Scoring metrics play an important role` in achieving the best classifier during classification training, Therefore, choosing an appropriate scoring metric is the key to distinguishing and obtaining the best classifiers. Hossin, Mohammad & M.N., Sulaiman. (2015).

# **Chapter 3. Methodology**

Project Development is an iterative process that is divided mainly into four phases. Two of these phases are theoretical which includes research and planning while the other two are practical which includes implementation and testing. To complete a given project appropriately on time it should be properly planned, for which there are project management methodology like Agile, Waterfall, 6Sigma KANBAN etc. I have used basic waterfall project management method to complete this project.

* research
* Theoretical framework

Phase1

* Scope and objective
* Proposed model
* Model baseline
* Scrums and timeline
* Dataset, constraints and ethical consideration
* Data Preprocessing
* Data analysis

Phase2

* Model preparation
* Model implementation
* Generate accuracy metrics
* Compare models
* Test results
* Conclusion and report

Figure 1: Project Methodology

Phase 1: A research strategy for a research project is an integral part. Researchers can use a variety of research techniques to examine the data and find answers to their questions. There are many types of data collection, analysis and collection techniques. Researchers must carefully coordinate their research procedures before starting their actual mission. It helps researchers conduct research properly and carefully. Literature review is conducted in the beginning of a research. However, it holds great importance throughout the research process. Systematic literature review is used in this dissertation to answer the research questions and achieve research objectives.

Our research starts with answering few basic question

1. What is stock market and what factors drive stock prices. What is timeseries?
2. Can we predict stock prices.
3. Current practices

Research articles were selected from the search engines like Google Scholar, researchgate and ieee.

The following research methods were used in this project.

• Search for research articles using Boolean search criteria

• Use truncation to find relevant articles.

• Separate student research projects with keywords.

The research strategies helped select few articles. The selected articles were then analysed and filtered according to the inclusion and exclusion criteria.

Inclusion Criteria

1) Articles published on authentic sites and journals were selected.

2) All the selected articles must directly address the research topic.

Exclusion Criteria

1) Unauthentic and unreliable sources were eliminated.

2) Research published in other languages was also excluded.

3) There were some redundant articles that were also excluded.

Once we are aware of current practices their strengths and weaknesses, we decide on models we will be using and plan steps and timelines for the implementation of models.

Phase 2: Implementation and conclusion.

The entire code is in the format of the jupyter Notebook using R language. R libraries like readr, quantmod load the dataset and perform the mathematical calculations on the dataset. Forecast, tidyquant are used to implement the different algorithms and forecast. ggplot and highcharter are needed to visualize the data in an interactive way. zoo, mts and xts are used to manage timeseries data. The historical data of the last 13 years was downloaded from the Yahoo finance website.

# **Chapter 4. Baseline and proposed model**

In this section we will review the methods used by researchers in literature review, to understand strengths and limitations of the models used. Moreover, this chapter illustrates models we propose and will discuss different aspects of proposed models along with its strengths and weaknesses.

## **4.1 Baseline models**

Chou and Nguyen (2018) proposed an intelligent time series system that used **sliding-window optimization** to predict stock prices of companies one step ahead. Project was divided in two stages.

Figure 2 Sliding Window optimization by Chon and Nguyen

This process repeats until all validations are performed.

**Limitations of model**

* Low computational speed, especially related to sliding window validation.
* High complexity of solving large mathematical loops in MATLAB programs. The number of computation increases with the number of validations.
* Many parameters of the "MetaFA and Time Series Parameters" system.
* This system does not produce excellent results for long-term investments.

Zhong and Enke (2017) used various models like **GARCH** (Generalized AutoRegressive Conditional Heteroskedasticity) which is a statistical model used to analyze time series data, where variance errors are assumed to be serially autocorrelated. The GARCH model assumes that the variance of the error term follows an autoregressive moving average process.

**Advantages**

* Simplicity
* Generate Volatility Clustering
* high kurtosis

**Disadvantage**

* Symmetrical positive and negative advance returns
* Restrictive
* Difficult to adapt prediction

Authors Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, used **BPNN**, Backpropagation is commonly used in training neural networks to compute a loss function in terms of network weights. Work with multilayer neural networks and observe the internal representation of the input-output mapping. The term backpropagation is used to describe backward propagation of errors. This is the heart of neural network training. In this concept, fine-tuning the weights of the neural network is based on error rates determined in previous iterations or runs. Properly setting the weights reduces the error rate and increases the confidence of the model through better generalization.

**Advantages**

* fast, simple, flexible, easy to analyze and implement
* does not contain parameters for tuning
* Efficient.

**Disadvantage:**

* The functionality or performance depends on the data input.
* Very sensitive to noisy data.
* Instead of mini-batches, a matrix-based approach is used.

Kumar, 2020 used **random forest** for prediction, where multiple decision tree are combined using bootstrap techniques and average of these DT help to make final decision.

**Advantages:**

* Reduced variance
* No scaling needed

**Disadvantages**:

* Increased time and complexity
* Additional resources required

Shastri January 2021 showed prediction results by **LSTM,** Long Short Term Memory. These are a special type of neural network called a recurrent neural network. RNNs address this issue by including feedback appearances that act as a kind of reminder. Past inputs to the model leave a trail. LSTM extends this idea to create both short-term and long-term memory components. LSTMs are therefore the perfect tool for anything with a sequence.

**Limitation:**

* LSTM training takes a long time
* Training LSTM requires more memory
* LSTM can easily overfit
* Dropout is much more difficult to implement with LSTMs.
* LSTM is sensitive to various random weight initializations

## **4.2 Proposed Models**

**4.2.1 ARIMA**

**A**uto**r**egressive **I**ntegrated **M**oving **A**verage, is collective process of AR, I, MA techniques each of which contributes to the final forecast. It considers past values (AR, MA) and predicts future values.

**Autoregressive (AR)**

An AR model predicts using a linear combination of past values ​​for that variable. The term autoregressive indicates that it is a regression on the variables itself. An autoregressive model of order p is

**mt = 0 + 1mt-1 + 2mt-2 + 3mt-3+…+ pmt-p.**

Where m is a linear function of its past p values while [0, p] are the regression coefficients that are determined after training.

**Integrated (I)**

Integrated stands for any differencing that must be applied to make the data stationary.

**Moving Average (MA)**

MA models use past forecast errors e to forecast future values in linear equation

**mt= 0 + 1et-1 + 2et-2 + 3et-3+…+ qet-q**

e also represents the random residual deviations between the model and the target variable.

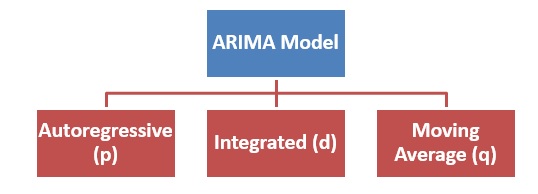


Figure 3: ARIMA Model

Standard notation ARIMA(p,d,q) is used where

p represents AR(number of lags included in model).

d represents I (number of times differencing is performed on raw data)

q represents MA (moving average window size)

**Advantage**

* ARIMA caters to a suite of standard structure in TS data
* It works well for short term prediction

**Disadvantage**

* Generate large trend errors
* Needs through training and tuning of variables as output is dependent on it hence becomes time consuming
* Ineffective for very large or very small data

**4.2.2 Holt Winter Model**

Holt-Winters forecasting is a time series method of modeling and forecasting values over time. It is a way of modeling his three dimensions of time series. Typical values ​​(average), slopes over time (trends), and periodically repeating patterns (seasonality). The Holt-Winters method uses exponential smoothing to encode past values ​​and use them to predict future values. The Holt-Winters model uses four forecasting techniques stacked one over the other.

AA

st=αxt+(1−α)st−1

st=αxt+(1−α)(st−1+bt−1)

bt=β(st−st−1)+(1−β)bt−1

Ft+m=st+mbt

Figure 4: Holt Winter Exponential Smoothing

Exponential smoothing refers to "smoothing" a time series using an exponentially weighted moving average (EWMA). For time series xt, st can be created using

***st*=*αxt*+(1−*α*)*st*−1 equation**

Charles Holt in the 1950s proposed Holt Exponential smoothing. It consists of two EWMAs: one for the smoothed values of *xt*, and another for its *slope*.

***st*=*αxt*+(1−*α*)(*st*−1+*bt*−1) , *bt*=*β*(*st*−*st*−1)+(1−*β*)*bt*−1**

To forecast the value in *m* time steps in the future. The formula for the *m*-step-ahead forecast, *Ft*+*m*, is:

***Ft*+*m*=*st*+*mbt***

**Advantages**

* Easy to learn and implement
* Gives betforecastcast
* Gives more significance to recent values

**Disadvantage**

* Forecast lags behind the actual trend
* Handles trends with difficulty

**4.2.3 Naïve Method**

The Naive Bayes Algorithm consists of two words, Naive and Bayes, and can be written as:

**Naive**: As it assumes that the occurrence of certain traits is independent of the occurrence of other traits.

**Bayes**: As it is based on the principles of Bayesian theorem.

Thus It is a classification technique based on Bayes' theorem that assumes independence between predictors. Simply put, naive Bayes classifiers assume that the presence of certain features in a class is independent of the presence of other features. NB models are easy to build and are especially useful for very large datasets. Bayes theorem provides a way of computing probability P(H|E) from P(E), P(H) and P(E|H).

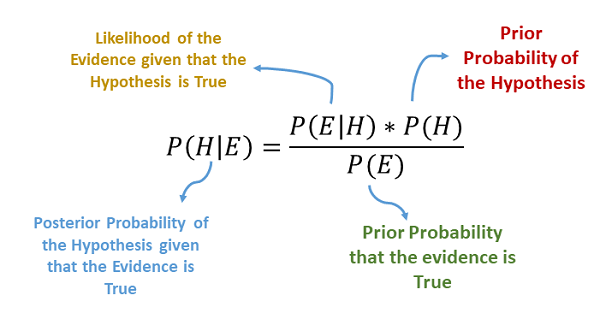


Figure 5: Naïve Bayes Equation

**There are three types of Naive Bayes models as shown below.**

**Gaussian**: Assume that the features follow a normal distribution. That is, if the predictors use continuous values ​​rather than discrete values, the model assumes those values ​​came from a Gaussian distribution.

**Polynomial**: The Multinomial Naive Bayes classifier is used when the data are multinomial. This is mainly used for document classification problems, i.e. which category a particular document belongs to, such as sports, politics, education, etc. The classifier uses word frequency as a predictor variable.

**Bernoulli**: A Bernoulli classifier works like a multinomial classifier, but the predictors are independent Boolean variables. For example, whether a particular word exists in a document. This model is also known for the document classification task.

**Advantages of naive Bayes:**

* Naive Bayes is one of the fast and simple ML algorithms for predicting classes of datasets.
* It can be used for both binary and multiclass classification.
* It has good performance in multi-class prediction compared to other algorithms.
* This is the most popular choice for text classification problems.

**Drawbacks of naive Bayes:**

* Naive Bayes cannot learn relationships between traits because it assumes that all traits are independent or unrelated.

**4.2.4 Neural Network**

Predictive neural networks are advanced data mining. A predictive neural network is conceptually a complex network of connected nodes that "learns" the structure of data. First, it analyzes historical data to determine how to predict known output values ​​using specified predictor variables. After this training phase, the neural network enters a testing phase, using new data to ensure that it has sufficient predictive power when presented with unprecedented information. When the prediction error of the network is small enough, it can accurately predict the future based on what it has "learned". An application that mimics the brain's ability to recognize patterns in data sets. Structure of Neural Network Prediction

**Structure of the neural network**

* **Input layer**: This will populate the next layer with past data values.
* **Hidden layer**: This is a key component of neural networks. I have a complex function that creates a predictor. Groups of nodes in the hidden layer called neurons represent mathematical functions that modify the input data.
* **Output layer**: This is where the predictions made in the hidden layers are collected to produce the final layer (the model's predictions).

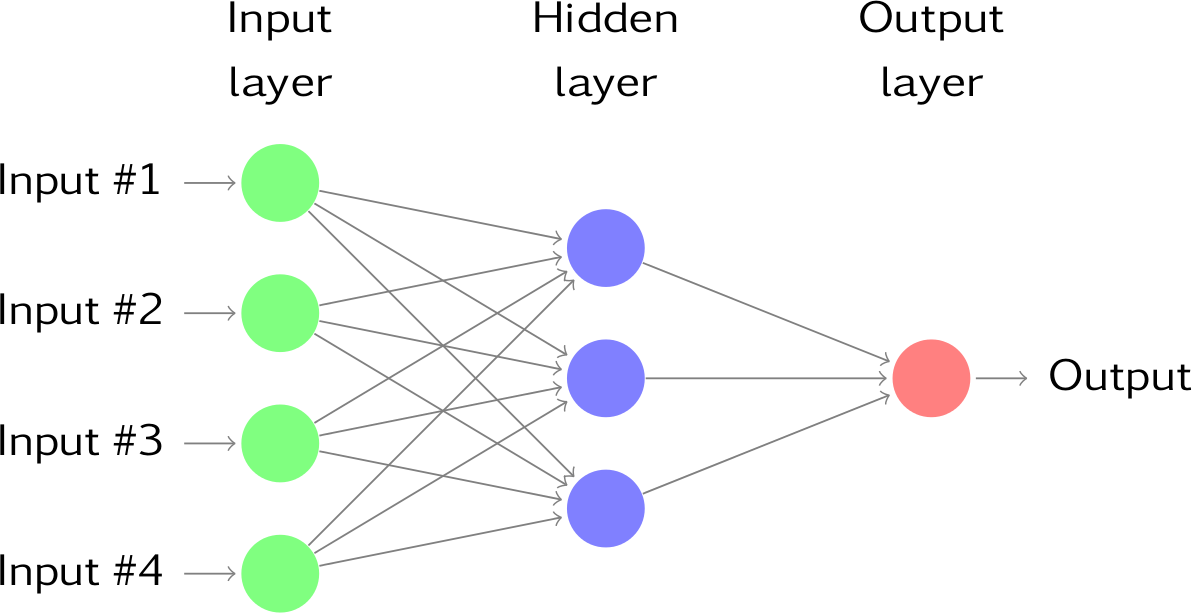


Figure 6: Working of Neural Network

**Advantage**

* Self learning ability
* State of the art results
* Strong fitting ability
* They can work with insufficient or irregular data
* They support parrellel processing

**Disadvantage**

* Uncertain prediction rate
* Limited data efficiency tends to loneger process
* Expensive to implement

**4.2.5 K Nearest Neighbor (KNN) Algorithm.**

K-Nearest Neighbor is one of the simplest machine learning algorithms based on supervised learning techniques. It assumes similarities between new cases/data and available cases and assigns new cases to the category that most closely resembles the available categories. The algorithm stores available data and classifies new data points based on similarity. It is also called a lazy learning algorithm because it saves the dataset and performs actions on the dataset during classification instead of learning from the training set immediately. The KNN algorithm stores only the training phase data set, and as it receives new data, it classifies the data into categories that are very similar to the new data.

KNN works as follows

1. Choose number of neighbors K
2. Compute the Euclidean distance of K neighbors
3. Get the K nearest neighbors according to the computed Euclidean distance.
4. Count the number of data points in each category in these k neighbors.
5. Assign new data points to the category with the largest number of neighbors.



Figure 7: KNN Algorithm Working

**Advantages:**

* Easy to implement.
* Robust to the noisy data
* More effective for the large training data.

**Disadvantages:**

Determine the value of K is most important and is user defined.

* High relative computational complexity

**4.2.6 Prophet**

Prophet is open source software released by Facebook. Prophet's key idea is that a better and much more flexible adjustment of the trend component will model seasonality more accurately, resulting in more accurate forecasts. Very flexible regression models like curve fitting were used instead of traditional time series models. That allows greater modeling flexibility, easier model fitting, and more elegant handling of missing data.

At its core, the Prophet method uses an additive regression model with four main components: The trend of linear or logistic growth curves.

* Daily seasonal components.
* Annual seasonal component.
* Weekly seasonal component.
* User-provided list of major holidays

The Prophet model can be conceptualized using the equation

****Y(t) is the forecast values

g(t) is trend variable that are nonperiodic

s(t) is seasonal variable(Annual, monthly, weekly)

h(t) is holiday is variables

∈ is variables not accommodated anywhere(noise variables)

Prophet provides uncertainty intervals for trend components by simulating future trend changes over time. The Prophet method uses Stan's probabilistic programming language. Stan performs MAP optimization of parameters in a fraction of a second, provides the ability to estimate parameter uncertainties using Hamilton-Monte Carlo algorithms, and fits in multiple interface languages making the steps reusable in open source.

**Advantage**

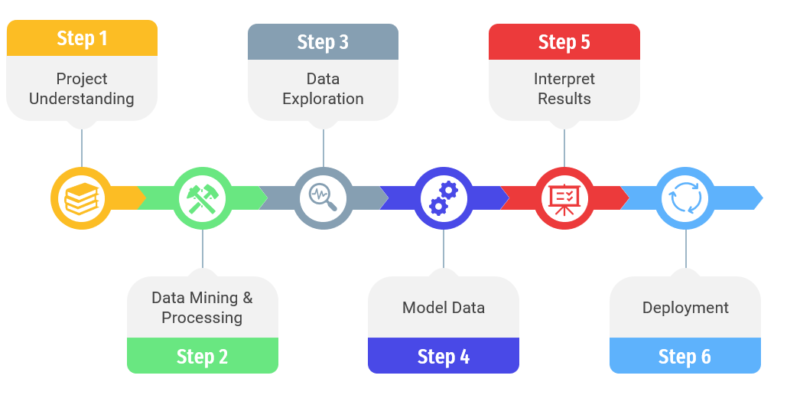
* Prophet accommodates seasonality with multiple periods allowing users to derive different assumptions about the trend
* Easily decomposes to time series where the additive model fits quickly with the ability to accommodate new components and hence can be tuned easily
* Robust to missing data

**Disadvantage**:

* Four seasonal trend dependencies can lead to overfitting.
* With irregular trends gives an uncertain prediction.
* Idiosyncratic changes (Noise ) are not accommodated and are assumed normally distributed
* Requires many iterations to produce an accurate result which makes it time consuming

# **Chapter 5: Implementation**

To deliver a project effectively we must follow a certain methodology composed of 6 different steps.



**Figure 8: project Implementation Steps**

## **5.1 Project Understanding**

In this step we determine the scope of the project and also determine the following:

1. The problem
2. The potential solution(s)
3. The necessary tools & techniques

Thorough research about stock indices, their importance in the economy, the need for accurate forecasts, pros and cons of current practices which are been used have been conducted and mentioned in chapter 1 and chapter 4. The research insights paved the way, to determine techniques we can use. As stock indices is a timed data we can analyze various timeseries algorithms to predict data. We have used basic, medium as well as advanced timeseries algorithms for forecast using R language and its libraries.

## **5.2 Data Mining And Processing**

Data which is raw information is crucial and most important asset of any organization. And this data serves as the core of any DS project. Accurate, reliable well organized and optimal data leads to statistically satisfying results. This step consists of three substeps.

Figure 8: Data Mining and Preprocessing Stages

**5.2.1. Data collection**

Data mining/acquisition is the first step in data processing. Data comes from available sources such as data lakes and data warehouses. It is important that the available data sources are reliable and well structured so that the data collected (and later used for information) is of the highest possible quality.

We are working on EU stock indices which include stock index for France (CAC 40), Germany(DAX), London(FTSE100) and Spain (IBEX 35). Data for EU Indices from 1st Jan 2009 to 11th Dec 2022.

In R library Quantmod (quantitative financial modeling framework) is used to download data, and. Charting. It supports functions like getSymbols(x) to get data from Yahoo(default) or Google and chartSeries() to draw charts in candle sticks format.

Our dataset is xts class object which has 6 features and a date index.

x.Open : open price of x on given date

x.High: Highest price of x on given date

x.Low: Lowest price of x on given date

x.Close: Closing price of x on Given date.

x.Volume: volume of x traded on given date

x.Adjusted: Price of x after paying off the dividend

**5.2.2 Data preprocessing**.

Once the data is collected, we enter the data preparation phase. Data preparation-"Data preprocessing", includes cleaning and organizing the raw data for the next phase of data processing. During processing, raw data is carefully checked for errors. The purpose of this step is to eliminate bad data (redundant, incomplete, or inaccurate data) and start creating quality data for the best business intelligence.

Data preprocessing techniques includes:

1. **Handling missing values**

Missing values ​​arise for various reasons during the process of data collection and observation. Time series models require that there are no data gaps along the time index, so observations with missing values ​​cannot be simply omitted (and reindexed as if there were no gaps). Instead, missing values ​​should be replaced with carefully chosen values ​​before fitting the model using "imputation". Median, spline interpolation, highest/lowest, previous/next value are methods used to replace missing values. In this project, we have replaced missing values with previous values.

1. **Duplicate data**

Time-series data have date time indexes that are unique (unless entered manually). We have downloaded datasets from reliable data sources so no such duplicity was found to process.

1. **Feature engineering**

There is no notion of time series input and output features. Instead, we need to select the variables to predict and use feature engineering to build all the inputs that will be used to predict future time steps. Ideally, we only want the input features that best help our learning method model the relationship between the input (X) and output (Y) we want to predict. The goal of feature engineering is to provide strong and ideally simple relationships between novel input and output features of the supervised learning algorithm being modeled. Downloaded dataset has just 6 features out of which our focus is on the closing price of indices per day. Hence we will be using only closing prices to generate a forecast model.

## **5.3. Data Exploration And Data preparation**

**1. Data Exploration** allows you to gain a deeper understanding of the datasets collected. A better understanding of data makes data analysts more efficient and streamlined. The EDA method used can be determined based on the type of variable (categorical or continuous). Data visualization tools help you explore your data and find important relationships and anomalies in your data sets.

There are several data exploration techniques based on type of variable

**Categorical variable**

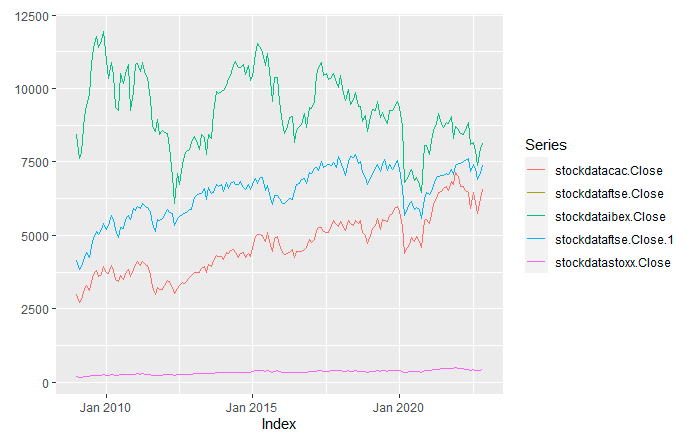
* number of unique values
* count multiple iterations

**Numeric variable**

* variance
* correlation
* Outlier detection
* trend analysis

Our dataset has numeric variables so we can use a summary() function in the R program to summarize the data frame into just one value or vector, showing min, max, quartile values, and median. For trend analysis, we have used visualization techniques relevant to financial data.

1. Visualizing Trend for all indices, Trend of FTSE50 and FTSE Covid data with moving average, Bollinger bands, CCI, ADX, MACD.



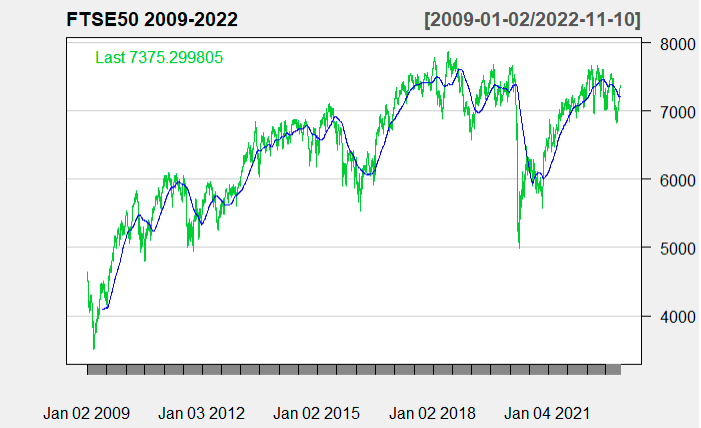
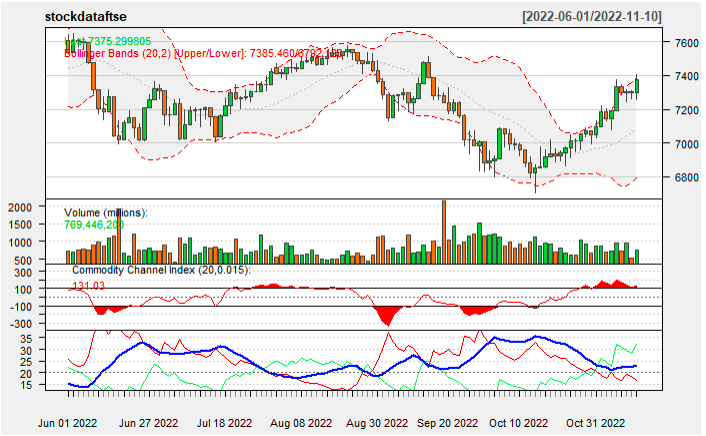


Figure 9 Trend analysis of all EU Stock Indices and FTSE 50 with 100b days MA

Figure 9 shows distribution of data of each indices from 2009 to 2022. Constant steady increase is observed with three notable depression 2011-12 effect of global depresseion on bank providing mortgages and companies associteed wit them. 2015-16 effet of brexit. 2019-20 effect of COVID 19.



**Bollinger bands**

**ADX**

**CCI**

Figure 10 FTSE 50 last 6 months with Bollinger bands, CCI and Directional Movement Index

Figure 10 shows last six months trend of FTSE 50. Bollinger Bands are a momentum indicator consisting of a simple moving average between two lines representing positive and negative standard deviations, measuring how close the price is to the mean. It predict potential market highs and lows relative to moving averages, helping traders visualize volatility and determine when trends will resume or reverse. Bands expand during volatile period and shrink during nonvolatile periods. We can determine stabilility overbought and over sold time frames. The Commodity Channel Index (CCI) is a technical indicator that measures current price levels relative to average price levels over a period of time. CCI is classified as a momentum oscillator. The basic assumption behind the CCI indicator is that commodities move in cycles, with highs and lows occurring at regular intervals.

**If the price is significantly above average, the CCI will be relatively high.**

**If the price is significantly below average, the CCI will be relatively low.**

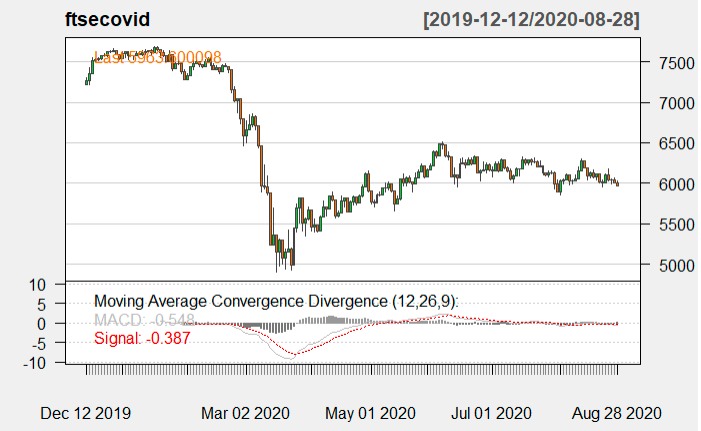


Figure 11 FTSE 50 covid depression data with MACD values

1. Interactive charts for detailed analysis for FTSE and comparison Chart of indice movement for FTSE and CAC

[FTSE PRICE CANDLE STICK CH 2009-2022](file:///C:\Users\L460\AppData\Roaming\Microsoft\Documents\stockmkt\lib\org.html)

[FTSE VS CAC PRICE 2009-2022](file:///C:\Users\L460\AppData\Roaming\Microsoft\Documents\stockmkt\lib\org1.html)

1. Stock return analysis

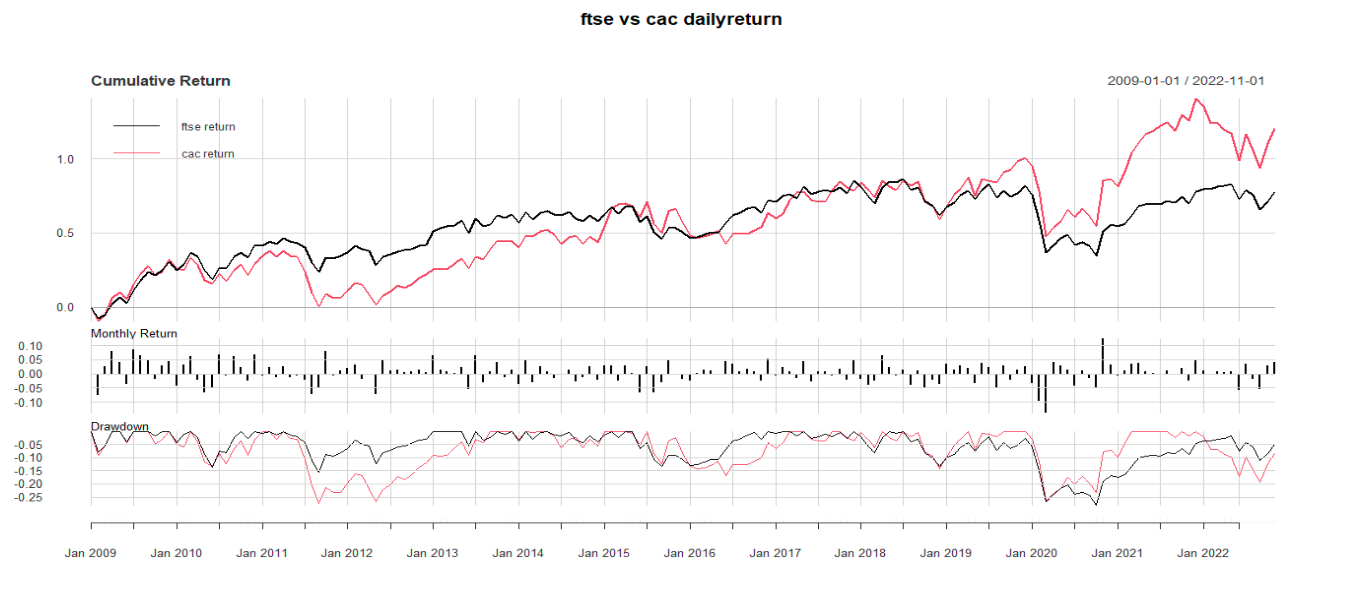


Figure 12 Cumulative monthly return of FTSE 50 and CAC

**2. Data Preparation**

Our dataset is an XTS (eXtensible Time Series ) object which can be easily used to visualize financial data in candlecticks form, *“The idea behind* ***xts****is to offer the user the ability to utilize a standard zoo object while providing a mechanism to customize the object's meta-data, as well as create custom methods to handle the object in a manner required by the user.”* But to prepare models for forecasting we need to convert them to Time Series Dataset by setting the relevant start and end of the dataset, along with its frequency. As the data is recorded on daily basis except for weekends frequency of the dataset is set as 260.

Our Time-Series have 3502 data for FTSE 50 from 2009 to 2022 while the Time series for Post covid FTSE data from 2019 to 2022 has 734 data with no Null Values and 260 frequency.

To model a time series we need to decompose it, remove its seasonality, convert it to stationary, and check its correlation.

* **Decomposition and removing seasonality**

A time series consists of three components: Trend cycle component, seasonal component, residual component. Decomposing helps for better understanding of time series, but can also be used to improve forecast accuracy. Decomposing a timeseries can be done by two methods, Addition or multiplication. Additive decomposition is best suited when the magnitude of seasonal variation or variation around a trend cycle does not vary at the series level. A multiplicative decomposition is more appropriate when the variation in the seasonal pattern or variation around the trend cycle appears to be proportional to the series level. Multiplicative decomposition is common in economic time series. Once the timeseries is decomposed its seasonality can be removed to make it more stationary.

* **Stationarity and Differencing**

Stationarity is an important characteristic of time series. A time series is said to be stationary if its statistical properties do not change over time. In other words, it has constant mean and variance, and covariance is independent of time. One way to test whether a time series is stationary is to perform an extended Dickie Fuller test. This test uses the following null and alternative hypotheses:

H0: Time series are not constant. In other words, it has a time-varying structure and not a constant variance over time.

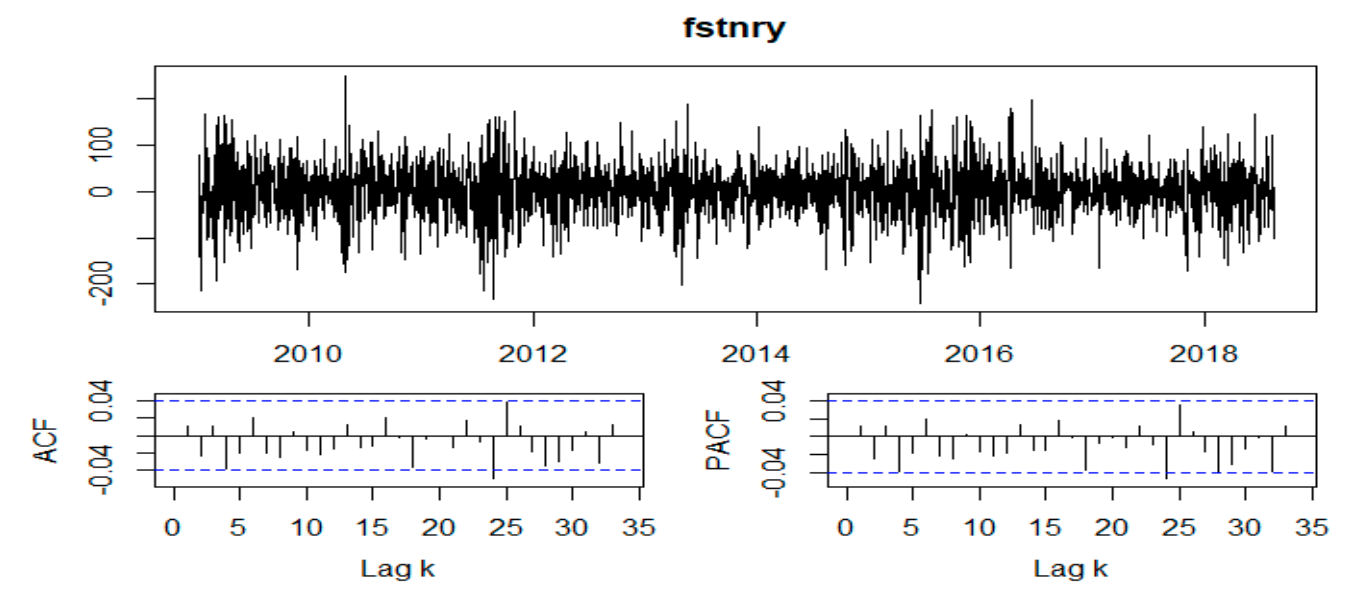
HA: The time series are stable.

**If the p-value of the test is below a certain significance level (for example, α = 0.05), you can reject the null hypothesis and conclude that the time series is stationary.**

The value of p for both FTSE 50 and post-covid is higher than 0.05 hence we need to make it stationary by **differencing** it. Differencing helps stabilize the mean of a time series by removing changes in the levels of the time series, thereby eliminating (or reducing) trends and seasonality.

* **AutoCorrelation**

When we are working with financial time series for forecasting or data simulation, decomposing the timeseries helped us to identify trends and seasonal components which when removed makes the series stationary. However, after all this we're left with a random elements. Sometimes one of these time series can be properly modelled via way of means of unbiased random variables. However, there are numerous situations, specially in finance, wherein consecutive factors of this random element time series will show correlation. That is, the behaviour of sequential factors affect each other or are correlated. To determine correlation ACF() and PACF() are used, it also helps us to determine p,d,q values which will be used for modelling.

****

**Figure 13 Stationary plot ACF and PACF plot of FTSE 50,**

From the plot we can determine ACF = p = 0, PACF = q = 0 whereas residual of the stationary series is evenly distributed.

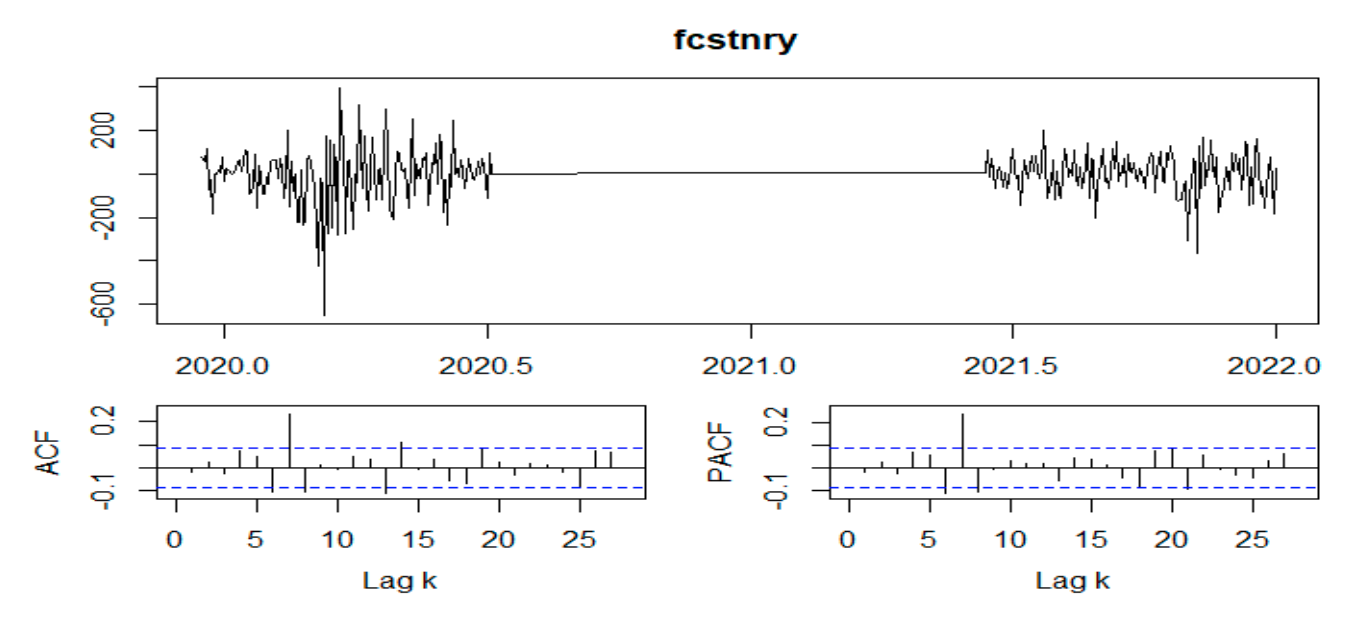


Figure 14 stationary residuals, ACF and PACF plot for FTSE post covid dataset.

As per the plot ACF as well as PACF is 0, but residuals of the stationary data shows null around 2020 june to 2021q june

## **5.4 Model Training**

As Discussed in chapter 4 we are creating Forecast model using Time Series techniques like ARIMA, Naïve, Holtwinters, Neural network, Prophet and KNN

Trained model

Test data

KNN

ARIMA

HW

Naive

Output Matrix

Accuracy comparision

NN

Prophet

Train Data 70 %

Test Data 30 %

* Analysis
* Data Preprocessing

Figure 15 Model Building

**5.4.1 ARIMA Model**

**A. DETERMINING value of pdq parameters**

ARIMA models can be represented in two forms:

1. Nonseasonal (ARIMA) model. The model is ordered in the form (p,d,q)

p = order of the autoregressive model

d = order of differentiational

q = order of moving average

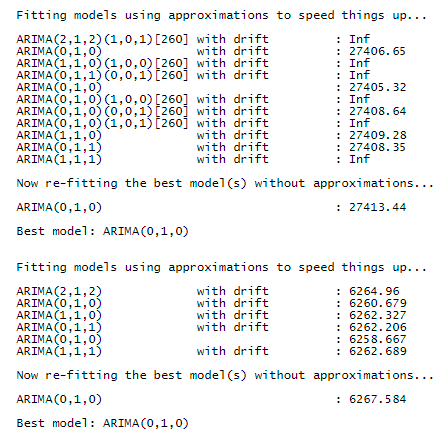
These values are determined using ACF and PACF values, which in our case is 0 (figure 13)

2. Seasonal models (SARIMA) These models take into account the seasonality of the data and perform the same ARIMA steps, but follow seasonal patterns. So if our data has a quarterly seasonal pattern, SARIMA will get the order of (p,d,q) for every point and (P,D,Q) for each quarter would be 1.

We have removed seasonality hence we are building Nonseasonal ARIMA value where we just need to determine p,d,q parameters only.

**B. Auto ARIMA:**

auto.ARIMA() heps in determining optimal values of p, d, q for ARIMA model. The optimum values for forecasting are those, whose AIC value is the lowest. Auto.ARIMA() confirmed values of p and q as 0 while that of d is 1. Hence ARIMA(0,1,0) is best for both datasets.



Auto.ARIMA() for post covid data is ARIMA(0,1,0) hence value of p=d=0 while I is 1. Model is using random walk

Auto.ARIMA() for FTSE 50 ARIMA(0,1,0) hence value of p=d=0 while I is 1. Model is using random walk

Figure 16 auto.ARIMA results

**C. ARIMA Model**

Training and testing the ARIMA model is next phase after finding the best p, d and q values ​​for ARIMA

model. The training data and test data are split by the following ratio: 70:30, where 70% of the data is trained while 30% of the data is used for model testing. The model is adjusted and a model prediction object is created for more analytical process. Apart from the p,d,q values generated by auto.ARIMA we have also used other values to create ARIMA(2,1,2), ARIMA(1,1,1),and ARIMA(1,1,0) to compare their accuracy and determine whether autoarima generated model is optimal or not.

**D. Ljung Box Test**

The Ljung box test is a method of testing for the lack of serial autocorrelation up to a particular lag k. This test determines if the error is iid (that is, white noise) or something more. Whether the autocorrelations of the errors or residuals are nonzero. Basically, this is a test of ineligibility.

H0, The data are independently distributed that and our model does not show lack of fit. (p<0.05 reject null hypothesis)

Ha, The data are not independently distributed. our model show a lack of fit. (p>0.05 cannot reject null hypothesis)

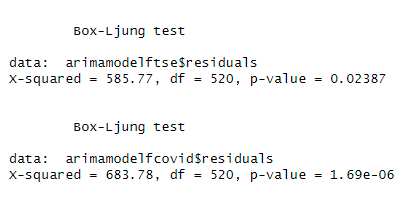


Figure 17 Box Ljung Test result (top FTSE 50 model, below post covid model)

**E. Predicted Model**

After confirming whether model found is optimal, ARIMA model with best fit is displayed using a visualization plot, the actual price (black), test price (red) projected price (in blue).

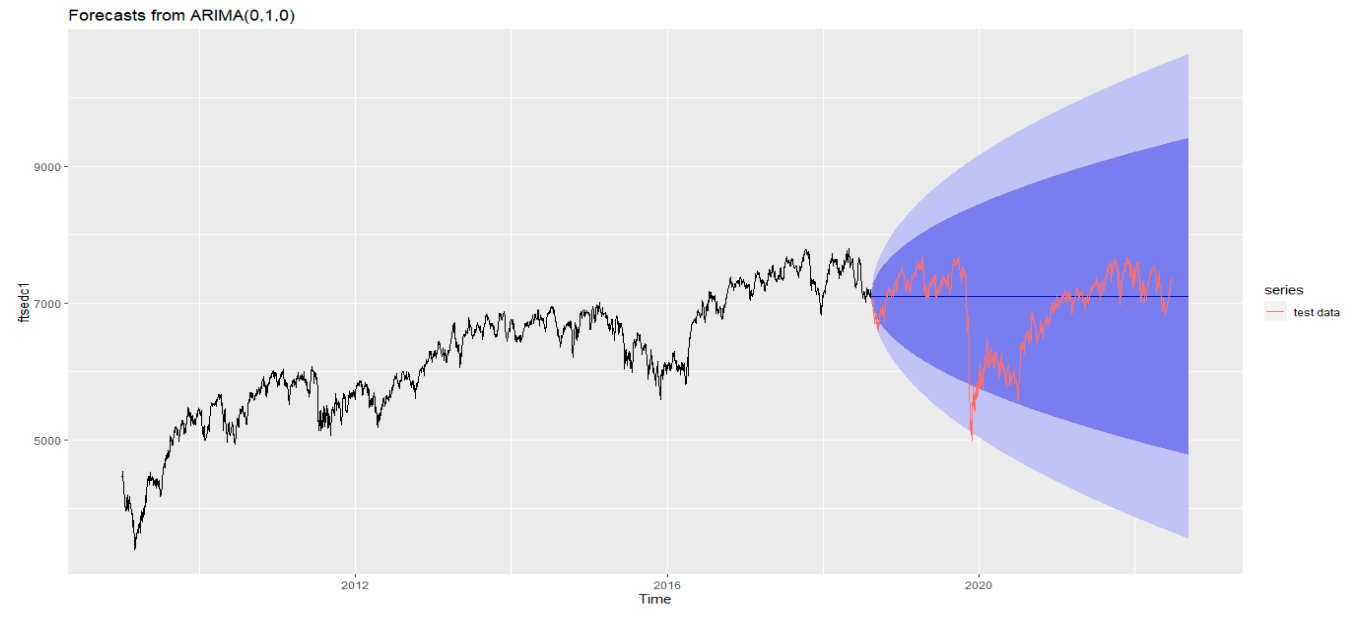
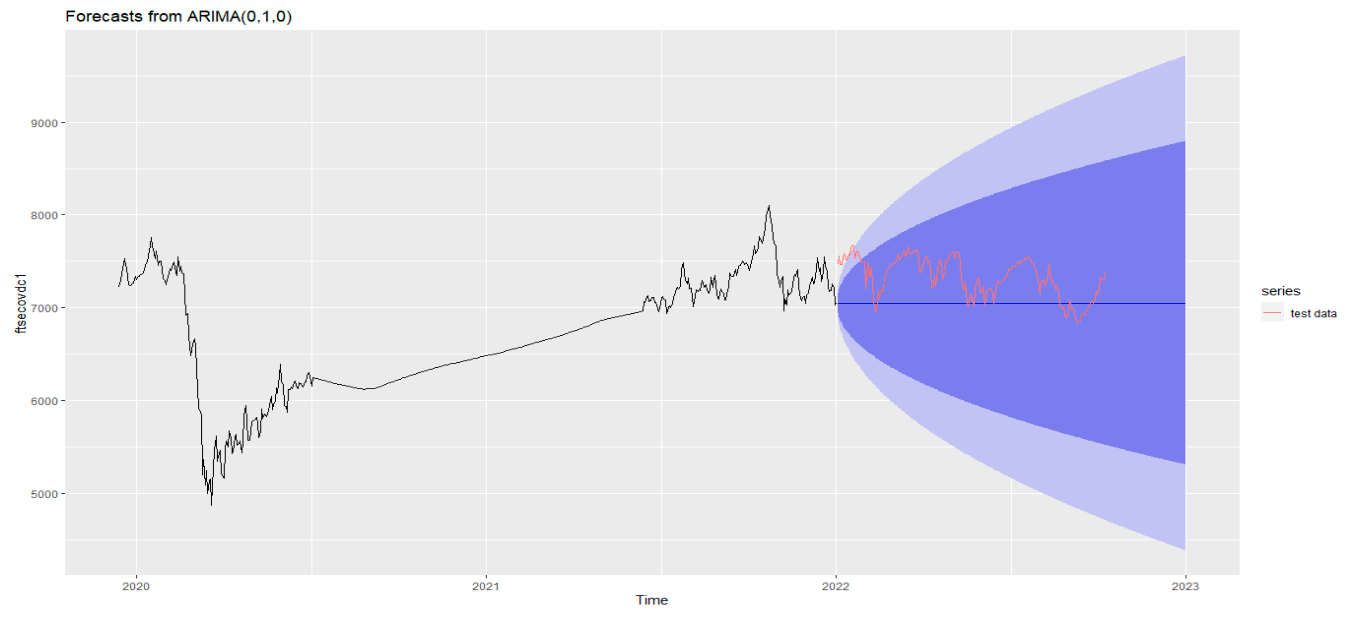


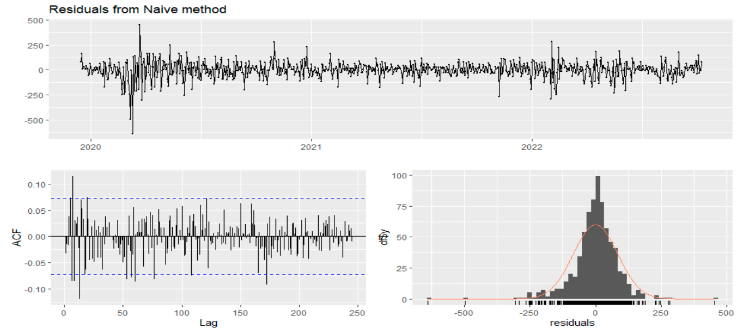
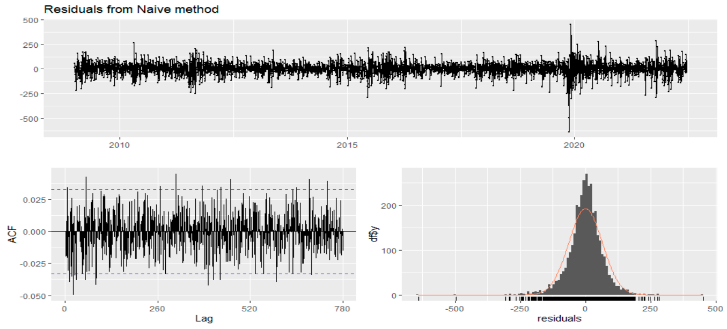
Figure 18 FTSE 50 ARIMA(010)

Figure 19 post covid ARIMA (0,1,0)

For both the models the predicted line is straight as it’s a random walk model. But if we compare it with test model it we can see that the test data is falling within the confidence level.

**5.4.2 Naïve Bayes Model**

Naïve is most basic forecasting technique, where it just forecast previous days value. Data set we have is already prepared for modelling. Hence we can implement model directly using Naïve() available in R.

Figure 20 a, b residuals ACF, and PACF values for FTSE and Post Covid

After confirming whether model found is optimal, ARIMA model with best fit is displayed using a visualization plot, the actual price (black), test price (red) projected price (in blue).

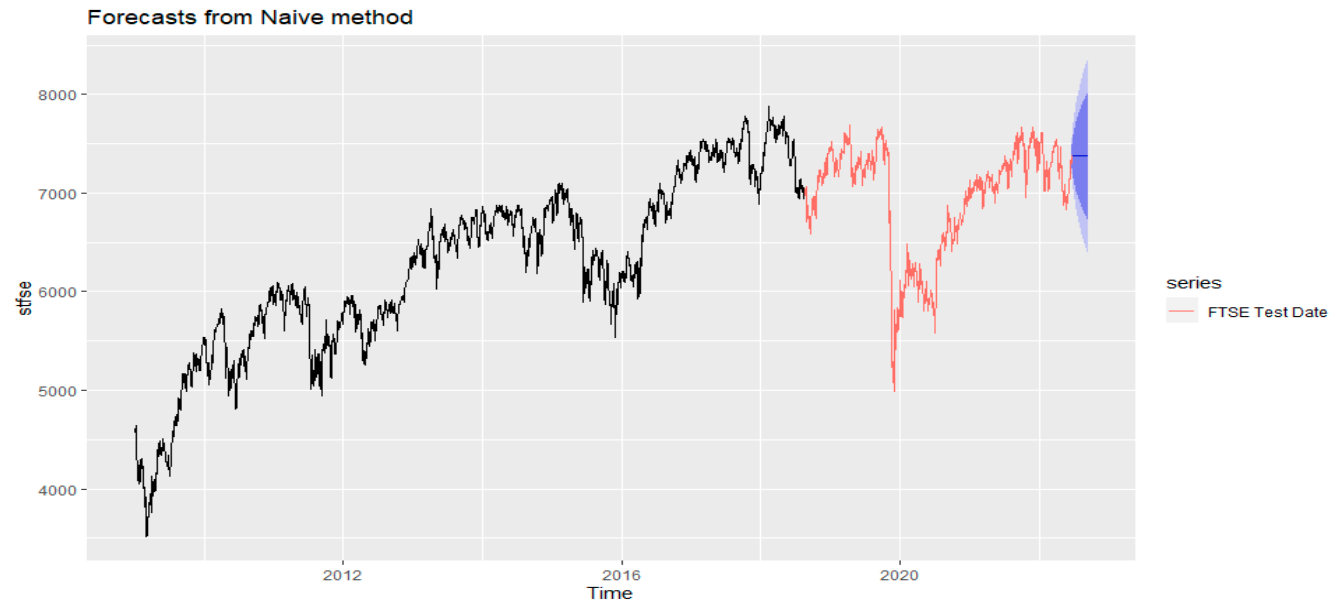


Figure 21 Naïve Bayes for FTSE 50

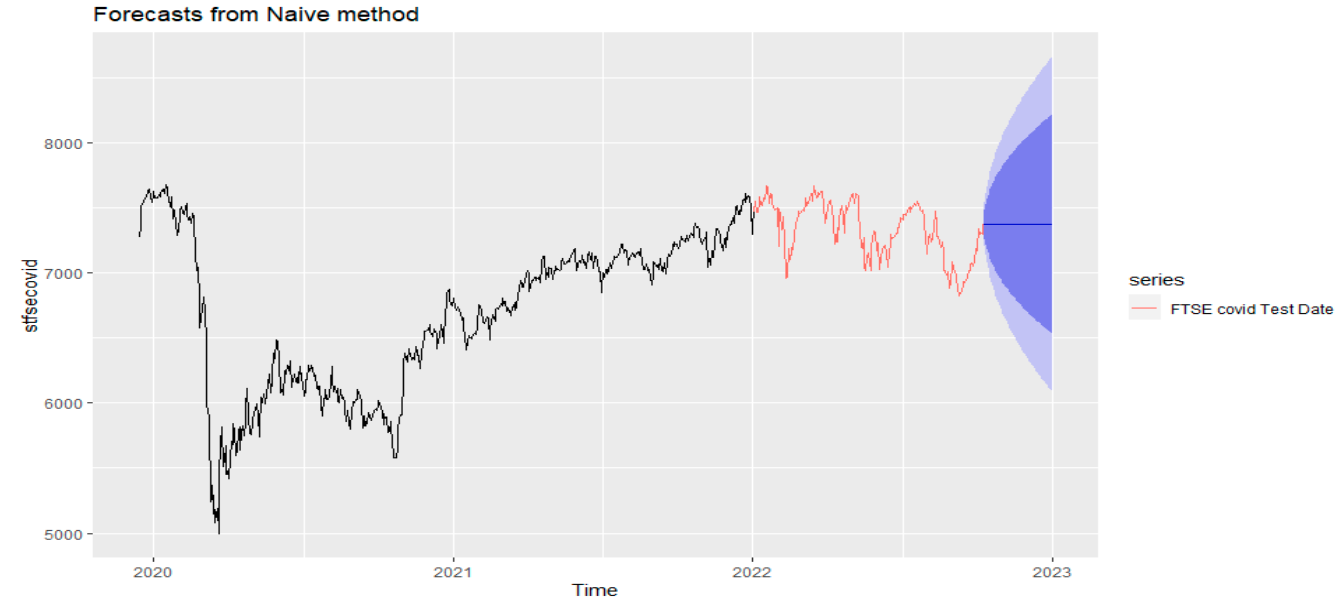


Figure 22 Naïve Bayes for Post Covid

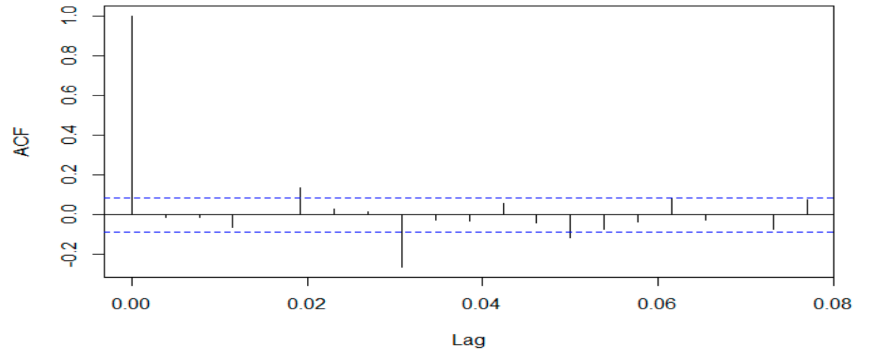
**5.4.3 Holt Winters Model**

Before we can predict future values, we need to create a fit model for the data. R provides HoltWinters() to automatically compute tuning parameters.

alpha: "level Values". The higher the alpha, the more recent observations are given more weight.

beta: "trend value". A high beta value means that the slope of the trend is highly dependent on the slope of the recent trend. If 0 does exponential smoothening

gamma: "Seasonal components". Higher gamma values ​​give more weight to recent seasonal cycles. If 0, non seasonal model is made

**Forecast Evaluation**

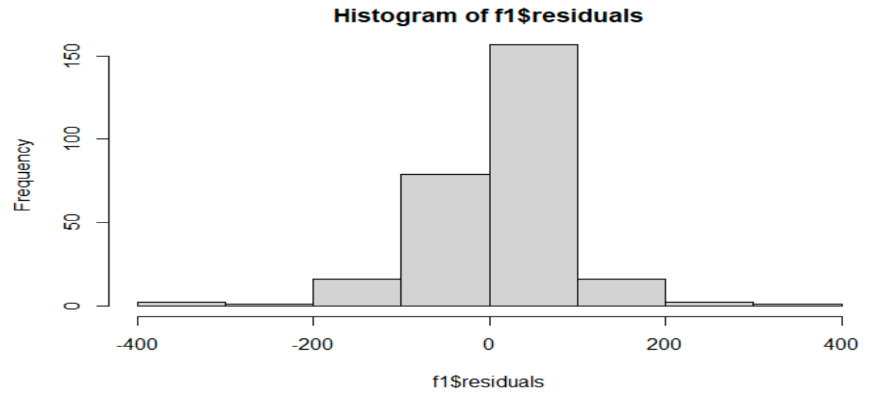
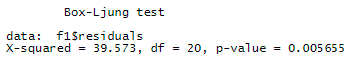
Forecast provides with residuals which can be used for evaluation. To best evaluate the smoothing function used in the model, we need to ensure that the prediction errors are uncorrelated. To capture this, we use the acf function to assess the correlation of the fitted residuals between points at different time intervals of the time series (lag). Ideally, when the delay is non-zero, the ACF bars are within the blue range bars shown below. The Ljung-Box test can also show the existence of these correlations. As long as the p-value exceeds 0.05, there is a 95% chance that the residuals are independent. Finally, it makes sense to check the histogram of the residuals to confirm normal distribution. If the residuals are highly skewed, the model can always overshoot in one direction.

Figure 23 Forecast evaluation of Holtwinter FTSE model

**Predicted Model**

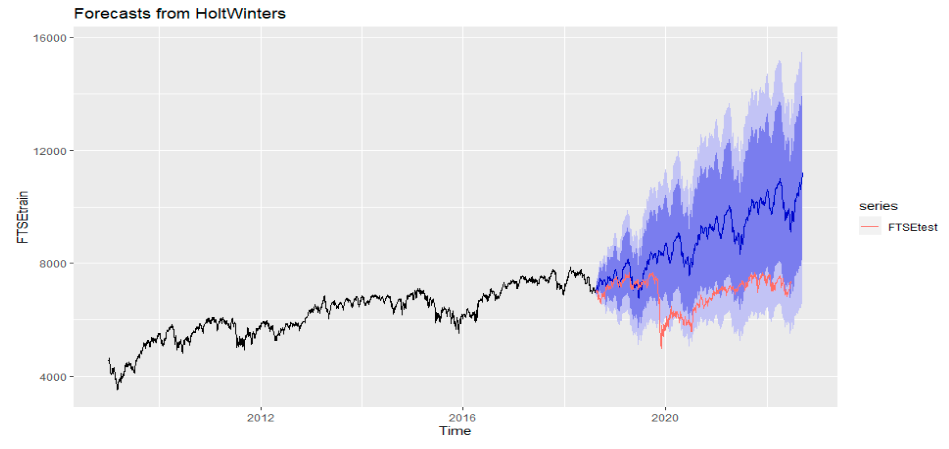


Figure 24 final model with prediction test and actual FTSE Data HW

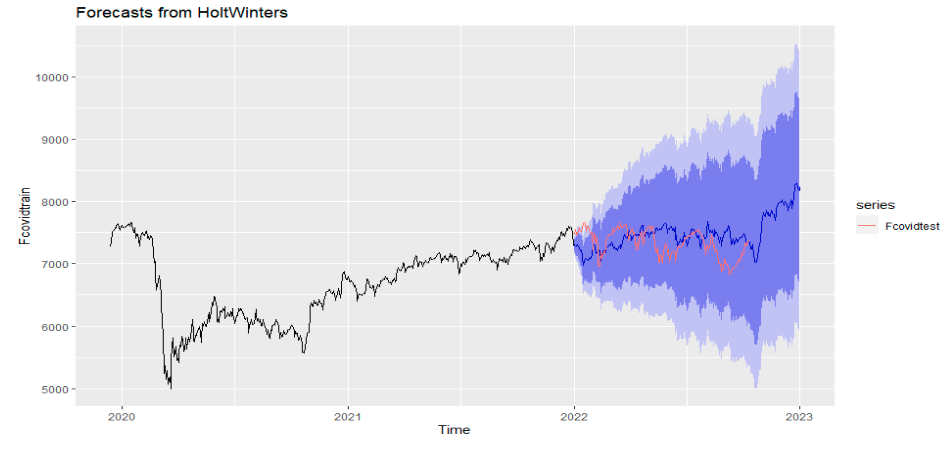


Figure 25 final model with prediction test and actual covid Data HW

**5.4.4 Neural Network**

Neural networks are a subset of machine learning inspired by the human brain. Neural networks are generally very data hungry means they need a lot of data to implement an accurate forecast. The nnetar function in the R forecasting package fits feedforward neural network models with delayed inputs at a single level to forecast univariate time series. A feedforward neural network is provided with the y delay values ​​as input and a single hidden layer containing the magnitude nodes. The inputs are for delays 1 to p and delays m to mP, where m = frequency (y). If xreg is specified, its columns are also used as input. If a value is missing in y or xreg, the corresponding rows (and other rows that depend on them as lags) are excluded from fitting. Overall, it incorporates a repeating network, each with random starting weights. These are averaged when calculating the forecast. The network is trained for one-step prediction. Multi-step predictions are computed recursively.

**Forecast Evaluation**

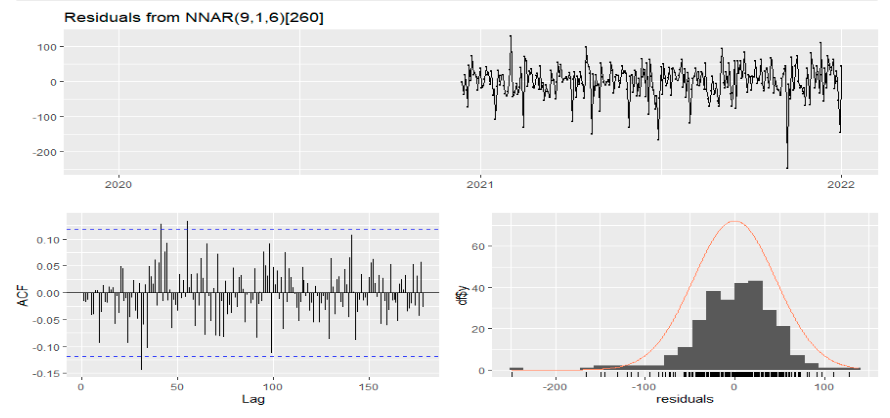
Forecast provides with residuals which can be used for evaluation. To best evaluate the smoothing function used in the model, we need to ensure that the prediction errors are uncorrelated. To capture this, we use the acf function to assess the correlation of the fitted residuals between points at different time intervals of the time series (lag). Ideally, when the delay is non-zero, the ACF bars are within the blue range bars shown below. Finally, histogram of the residuals to confirm normal distribution. If the residuals are highly skewed, the model can always overshoot in one direction.

Figure 26 neural nwtwork residual ceck FTSE Data

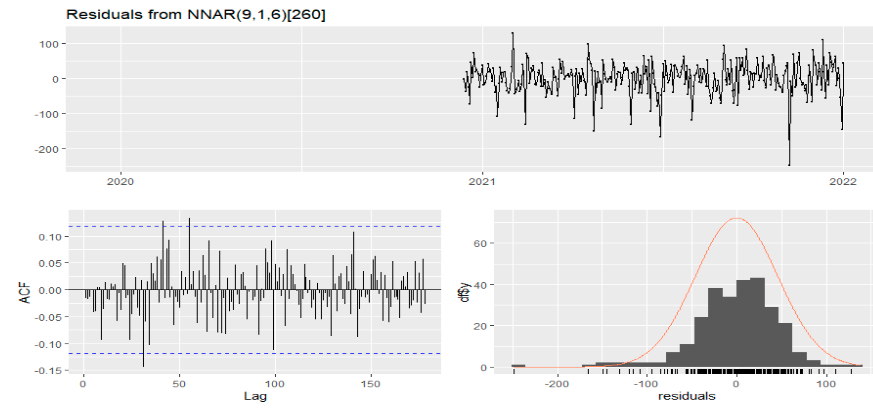


Figure 27 residual chart for FTSE cov data

**Test - Predicted Model**

For FTSE 50 predicted line is almost constant against FTSE Test Dataset

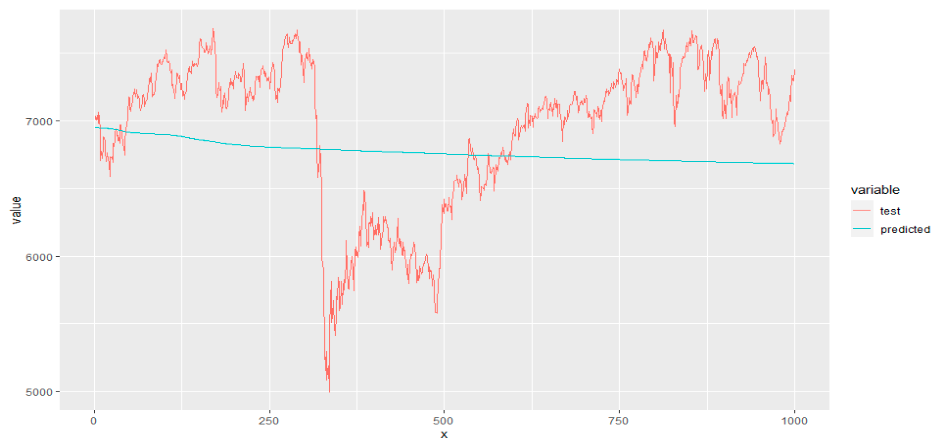


Figure 28 test – NN predicted plot for FTSE data

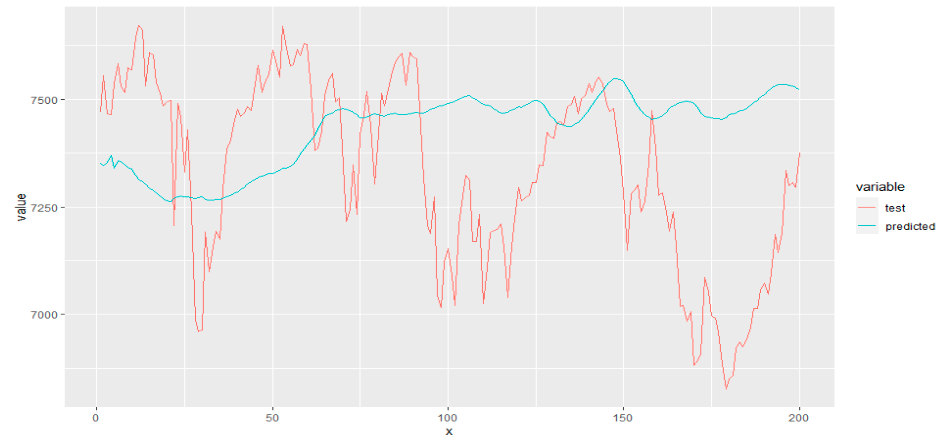


Figure 29 test – predicted plot for FTSE cov data

While in postcovid dataset, we can see uptrends and downtrends in the predicted values

**5.4.5 Prophet**

R is well known for its graphs, charts, and other analytical presentations that facilitate interpretation as R is a well-known tool for stock market analysis. The Prophet package is one of many other packages for R. Before using the Prophet package to predict the price of selected symbols, we transform the dataset so that we can analyze the data read by "Prophet". After transforming the dataset, we apply the model to the dataset to predict future values. Next, for better understanding, plot the data to find the trend of the FTSE index and divide the output by the trend component, weekly seasonality, and yearly seasonality.

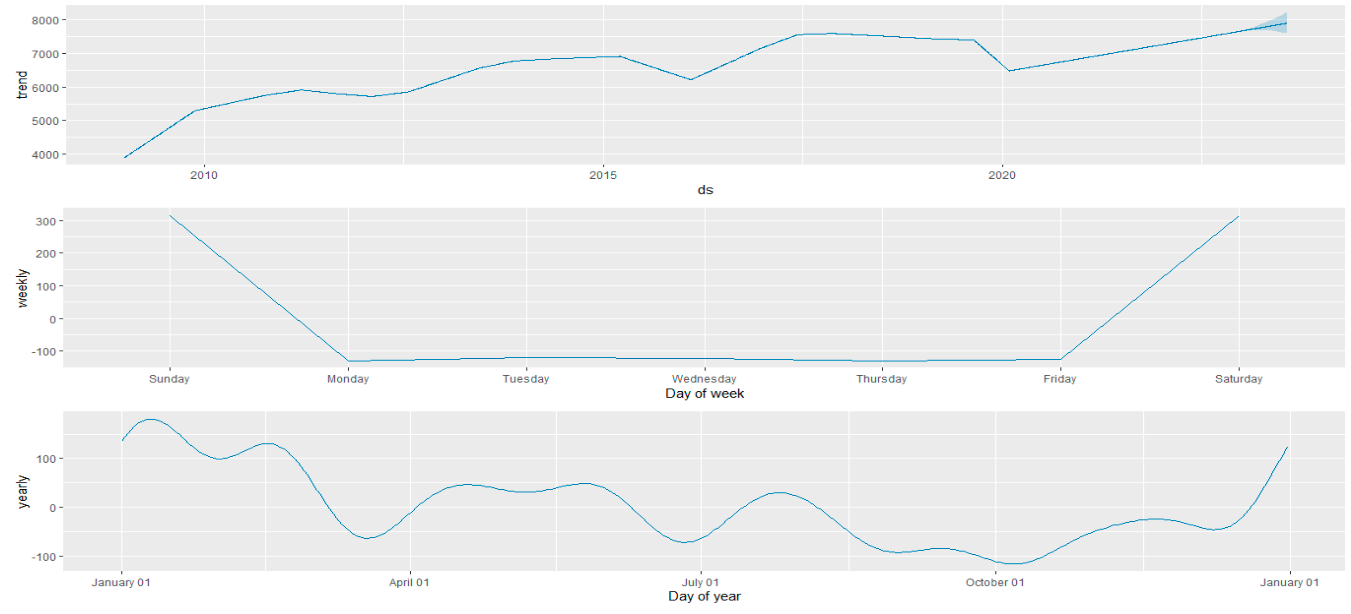


Figure 30 FTSE Prophet Trend Components

The Graph shows an upward trend for FTSE 50 index, operative from Monday to Friday, in yearly trend shows growth after market crash.

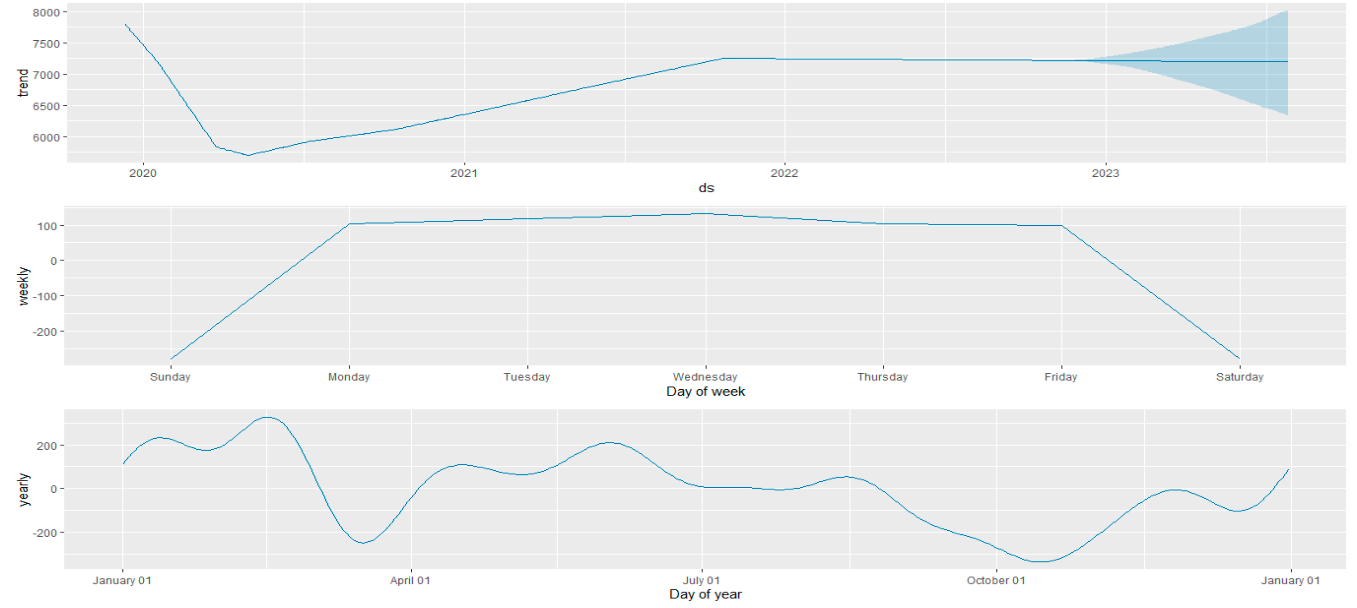
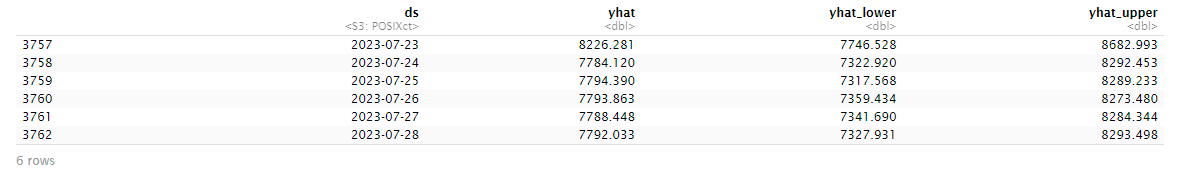


Figure 31 Post Covid Prophet Data Trend component

The entire forecasting process is wrapped in a single R function that takes an input data frame and gives number of sample predictions. Return and feed are calculated based on the adjusted closing price. This function returns a data frame containing the forecast and its error for each period as output.

The resultant Dataframe of forecast values can be realized as



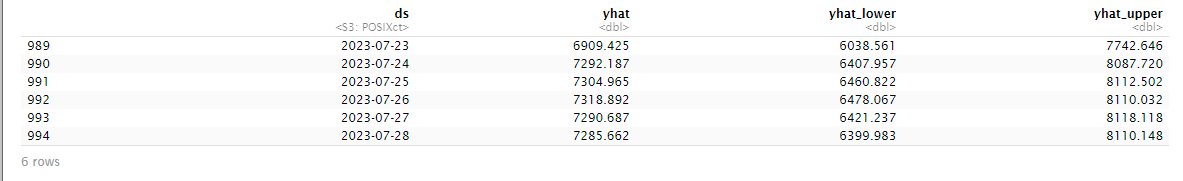


Figure 32 Prediction Dataframe for FTSE and postcovid data

We can observe different values for both the dataset which shows effect of data on training and testing of model.

**Final Model**

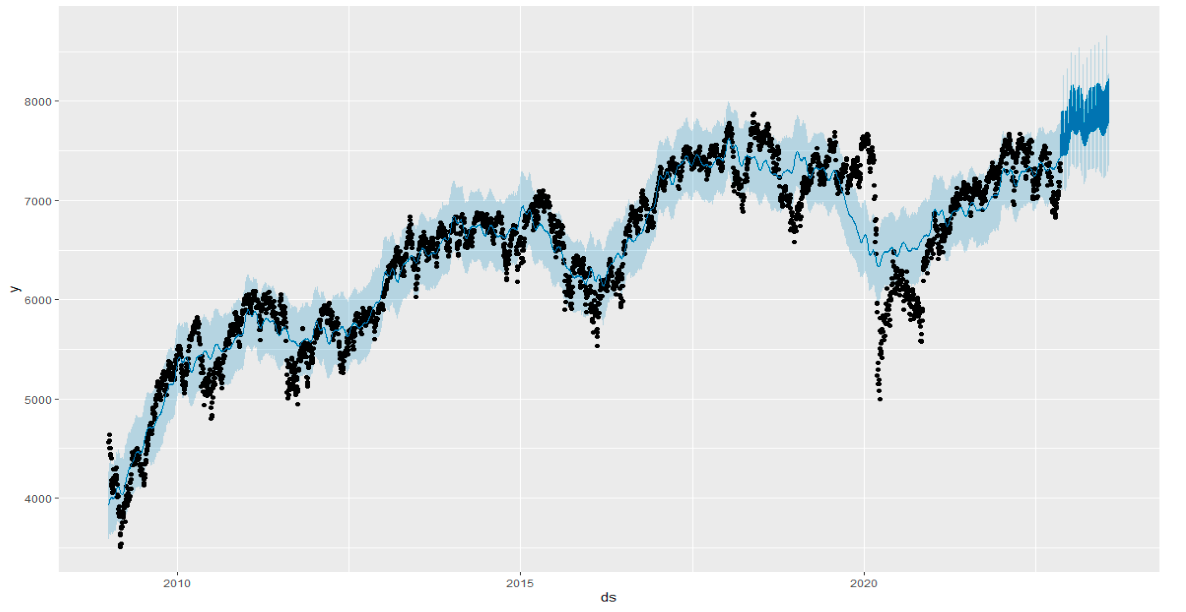


Figure 33 Prophet Model for FTSE 50

Upward movement in predicted values are shown in dark blue color, while confidence interval is in light blue tint. Only during covid actual values move away from confidence interval.

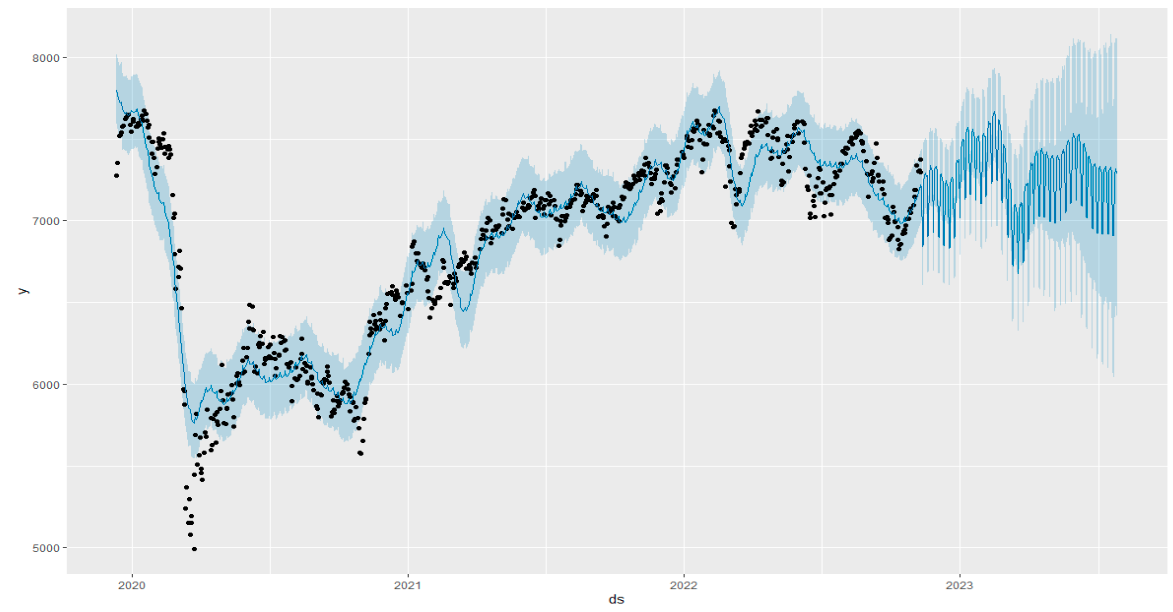


Figure 34 Prophet Model for post covid

Steady Movement of predicted values is visible with all values within confidence interval

**5.4.6 K NN Model**

For KNN Model we need to first deterimine optimal value of k usually is the square root of N, where N is the total number of samples. Our train data has 2500 and 550 rows/ data items. Hence value of K is taken as 50 and 22 respectively. While lags is taken as 1:260(frequency of Data set. The forecasting technique used by KNN\_forecasting function is MIMO – multiple input multiple output. All other default parameters are considered while fitting the model

nearest\_neighbors () shows the instance, kn neighbors, and its targets. These targets are averaged to forecast the future h periods.

**Final model**

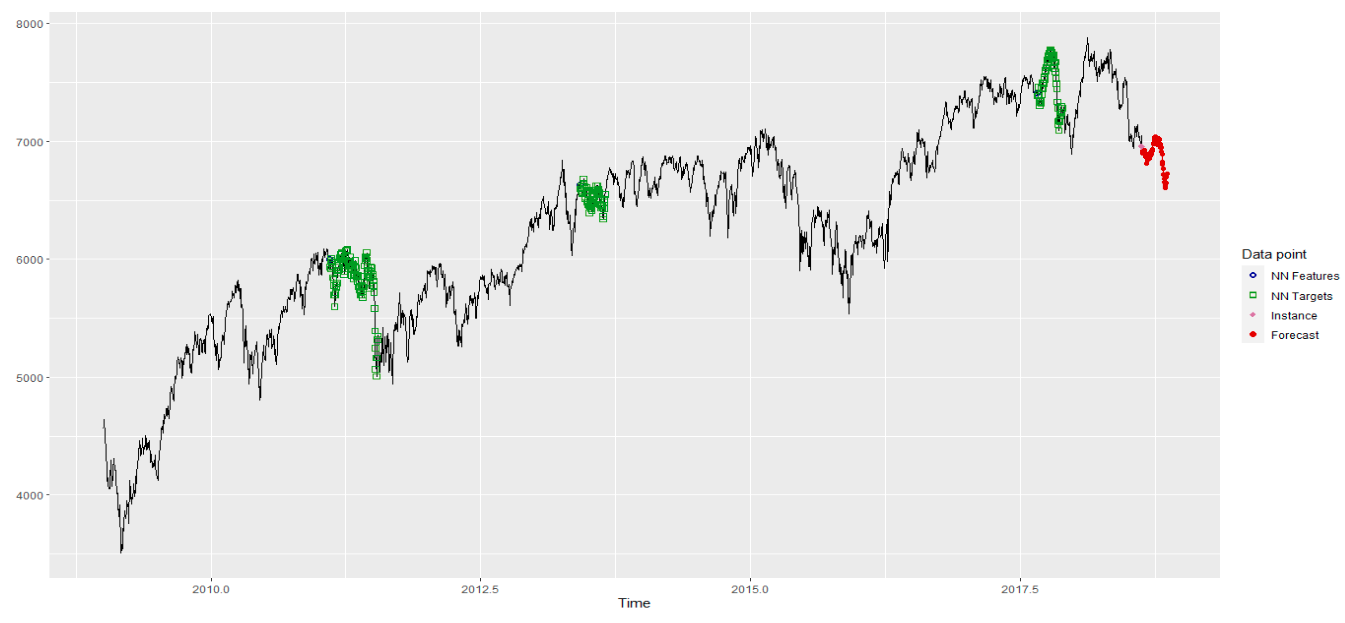


Figure 35 KNN model for FTSE data

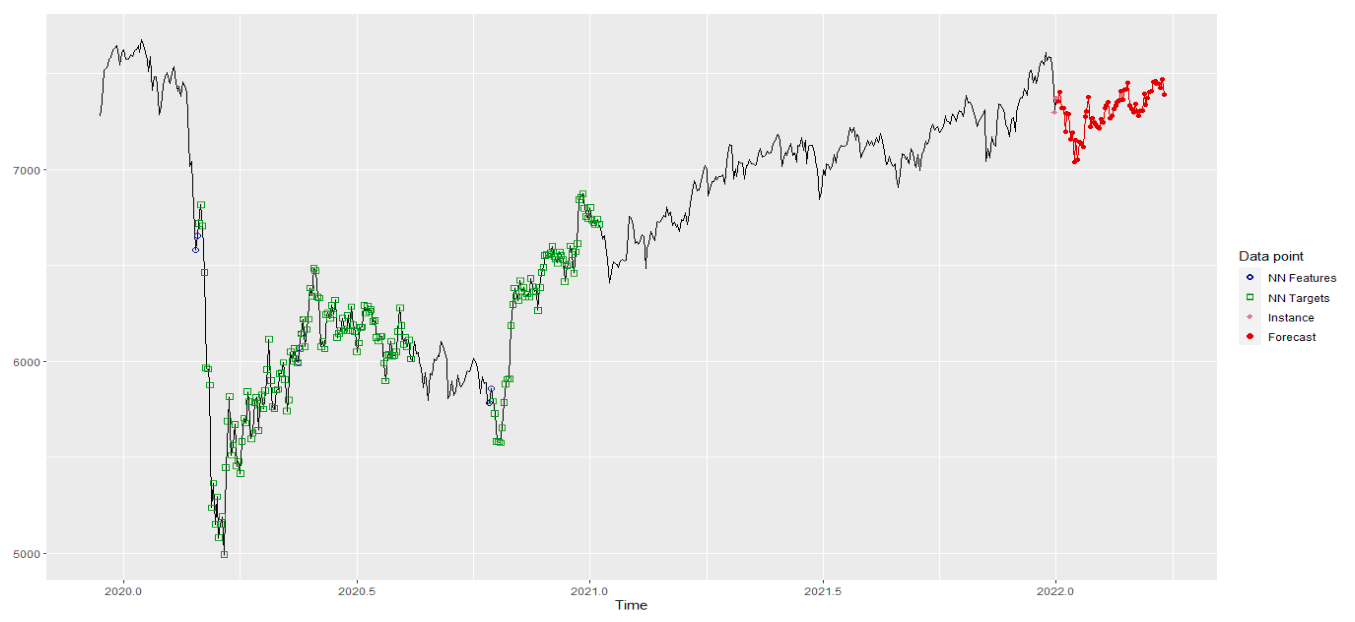


Figure 36 KNN model for post covid

# **Chapter 6 Result and Conclusion**

## **6.1 EDA Results**

**Figure 9** shows distribution of data of each indices from 2009 to 2022. Constant steady increase is observed with three notable depression 2011-12 effect of global depresseion on bank providing mortgages and companies associteed wit them. 2015-16 effet of brexit. 2019-20 effect of COVID 19.

**Figure 10** shows last six months trend of FTSE 50. Bollinger Bands are a momentum indicator consisting of a simple moving average between two lines representing positive and negative standard deviations, measuring how close the price is to the mean. It predict potential market highs and lows relative to moving averages, helping traders visualize volatility and determine when trends will resume or reverse. Bands expand during volatile period and shrink during nonvolatile periods. We can determine stabilility overbought and over sold time frames. **Our model showed high voltality arounf jun 27, 2022 and October 10, 2022. We also visualized few oversold and over bought stocks**. The Commodity Channel Index (CCI) is a technical indicator that measures current price levels relative to average price levels over a period of time. CCI is classified as a momentum oscillator. The basic assumption behind the CCI indicator is that commodities move in cycles, with highs and lows occurring at regular intervals.

**The CCI will be relatively high if the price is significantly above average.**

**The CCI will be relatively low if the price is significantly below average.**

**Results from Bollinger band visualization and CCI Charts were same.**

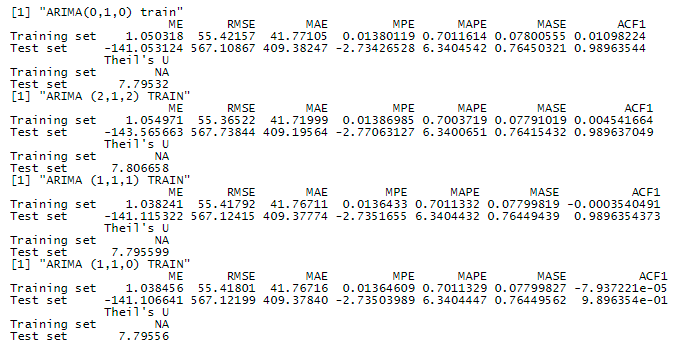
**Figure 12** Stock return analysis

FTSE showed higher returns in past but in last 3 years (post covid ) returns by CAC is higher compared to that of FTSE

## **6.2 Model Performance**

* + 1. **ARIMA Performance**

**Accuracy comparision of different ARIMA models**

Figure 37 different ARIMA model assessment

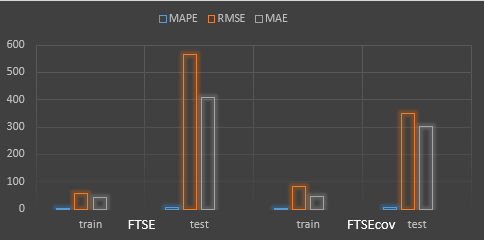
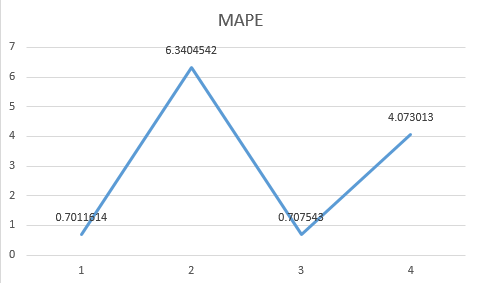
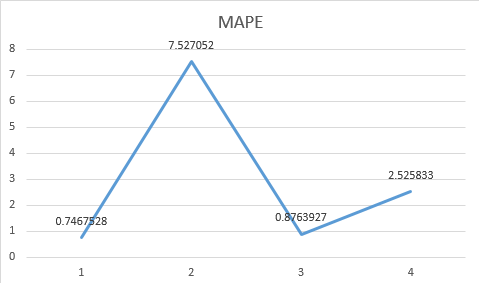
 

Figure 38 Train test accuracy and MAPE values for FTSE and FTSE post covid data

**Interpretation**

According to accuracy matrices in fig 37 for all the four ARIMA models, theirs negligible difference between either of them. But RMSE for the test data is least for ARIMA (0,1,0) which is also auto.ARIMA() suggested model hence we can conclude that auto.ARIMA () does give us the optimum value for p,d,q for best fit model. Now coming to the main model we have taken two datasets FTSE 50(huge historical data) and FTSE 50 post covid (recent data) by modelling dataset of different size we can interpret whether given model performs better with limited data or needs more data for training.

In the fig 38 above important accuracy measures for FTSE and FTSE post covid data are shown. MAPE values for train model for both datasets is near to 0, which suggest model is well fitted, in test data also its <20% means our model is performing with accuracy over 80% ,RMSE and MAE values is relatively high, From which we can see that ftsecov model has outperformed FTSE model clearly for test data, but these values depend upon the size of the dataset. Percentage wise FTSE has performed better than the postcovid model.

* + 1. **Naïve Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **train/test** | **MAPE** | **RMSE** | **MAE** | **data** |
| **Train** | 0.7467528 | 64.36734 | 46.223 | FTSE |
| **Test** | 7.527052 | 690.5633 | 477.1628 | FTSE |
| **Train** | 0.8763927 | 84.61951 | 57.96426 | FTSE cov |
| **Test** | 2.525833 | 219.0542 | 182.7315 | FTSE cov |

Figure 39 Accuracy matrix and MAPE Value

**Interpretation**

In the figure above important accuracy measures for FTSE and FTSE post covid data are shown. MAPE values for train model for both datasets is near to 0, which suggest model is well fitted, in test data also its <20% means our model is performing with accuracy over 80% ,RMSE and MAE values is relatively high, From which we can see that ftsecov model has outperformed FTSE model clearly for test data, but these values depend upon the size of the dataset. Percentage wise Postcovid model has performed better than the FTSE model.

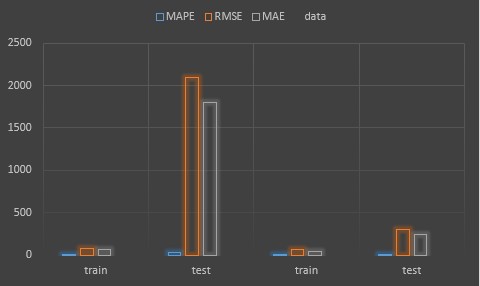
Pre-Covid Test:

As per the performance Matrix of naïve-bayes model, pre covid data set gives a good accuracy where all the matrices is well within the limits of accuracy. RMSE value of 690.56, MAE of 477.16, MAPE of 7.52 shows high accuracy.

Post Covid:

As per the performance Matrix of naïve-bayes model, pre covid data set gives a good accuracy where all the matrices is well within the limits of accuracy. RMSE value of 219.05, MAE of 182.7316, MAPE of 2.52 shows high accuracy and also suggests that the training data set worked small with large data and accuracy has fell down while we used a Big set of data, with. While evaluating the data with respect to MAPE values, Post covid data still showed lesser error count. Then that by pre covid data.

* + 1. **HoltWinters Model**



ftsecov

FTSE

**Figure 40 Acuuracy matrix graph HW model**

**Interpretation**

Holt Winters Model performed well in postcovid dataset, means it worked better with less data. Time consumed was not huge. Moreover the prediction data didn’t moved in straight line like previous models instead showed exponential smoothing curves across data. MAPE value for train as well as test is less than 20% hence Models are pretty good. But forblage datset where it performed good in training performed extremely poor in testing.

* + 1. **Neural Network**

**Accuracy**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **train/test** | **MAPE** | **RMSE** | **MAE** | **data** |
| **train** | 0.6818334 | 57.13394 | 42.47769 | FTSE |
| **test** | 7.7620843 | 602.59468 | 532.16035 | FTSE |
| **train** | 0.5524685 | 71.12106 | 39.61945 | FTSE cov |
| **test** | 3.3572268 | 296.95511 | 243.56448 | FTSE cov |

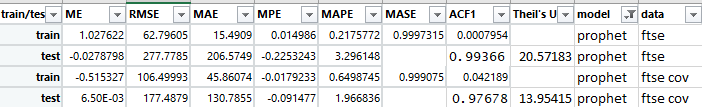
Figure 41 accuracy matrix NN

**Interpretation**

Value of MAPE is just 3.35 for Postcovid Data which is really low percentage error. While value of RMSE is also less percentagewise with entire dataset. Value of MAPE is not bad in FTSE Data set too. But major concern with this model is time consumption.

* + 1. **Prophet Model**

**Accuracy**

****

**Figure 42 Accuracy matrix Prophet**

**Interpretation**

Post-Covid:

As per the results on Prophet model, post covid data set gives very high accuracy where all the matrices is well above the consideration. This explains that the test model is accurate with RMSE value of 177.48, but very less value of MAPE suggest Overfitting

Pre-Covid:

Through this this is the best model so far for high volume of data, when compared to all the above models but it consumed ample time for execution,

* + 1. **KNN**

Accuracy of a KNN prediction object is measured using rolling\_origin() as it creates global\_accuracy values for test data

**Interpretation**

Pre-Covid:

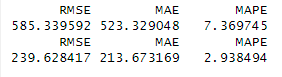
As per the results on Prophet model, pre covid data set gives very less accuracy where all the matrices is well above the consideration. Which explains that the test model is less accurate than both the above models with RMSE value of 585.3, MAE of 523.3, MAPE of 7.4.

Figure 43 accuracy measure KNN

Post Covid:

As per the matrices here, the training data set worked more accurate when compared to all the above models, with RMSE value of 239.6, MAE of 213.67, MAPE of 2.9. While evaluating the data, like the pre-covid testing there is a slight decrease in RMSE,MAE and MAPE value hence we are getting better results for smaller dataset. However, this model is not accurate as the previous model.

## **6.3 Model Comparison Analysis**

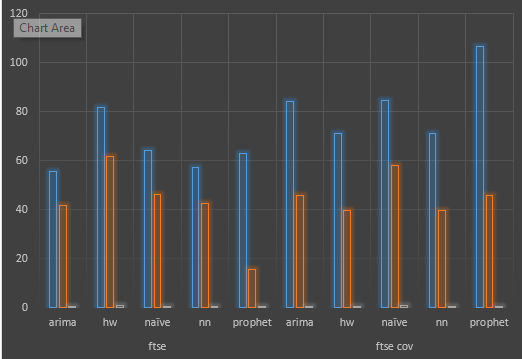


Figure 44 Performance Matrix for train data FTSE and post covid data for all models

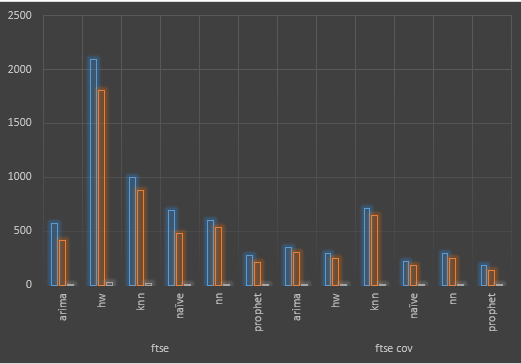


Figure 45 Performance Matrix for test data FTSE and post covid data for all models

We have six Models working on small as well as big data set. Where ARIMA, Naïve worked better for huge data while Holtwinters KNN worked better with small data. Neural Networks and Prophet gave almost same performance irrespective of size of data they took more time in execution.

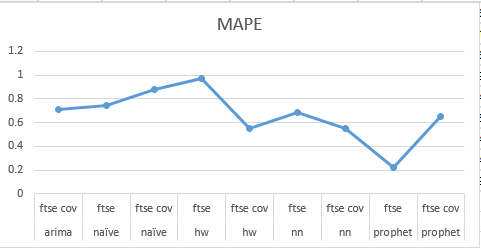
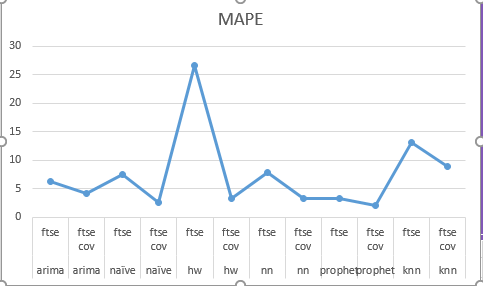
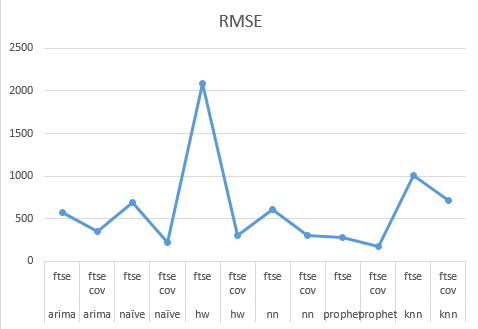


Figure 46 Train(left) test (Right) MAPE Values

MAPE values for train data set is below 1 indicating closely fit models. According to MAPE Values Holt winters model performed worst for FTSE dataset with more than 20% error. Even KNN model performance is not upto mark. Prophet showed highest accuracy with 0.2 % error in train and around 2% error in test set which looks like an overfit model.

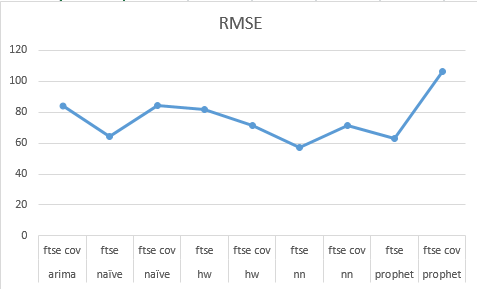


Figure 47 Train(left) test (Right) RMSE Values

RMSE is an important indicator to measure the performance of a model. Here though the value of MAPE was least in Prophet it has highest RMSE for train data set which shows our assumption of overfitting is true. Holtwinters and KNN model are performing poorly, while ARIMA, Naïve, and NN still looks good enough as predition model.

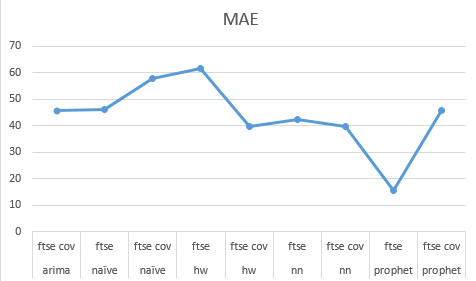
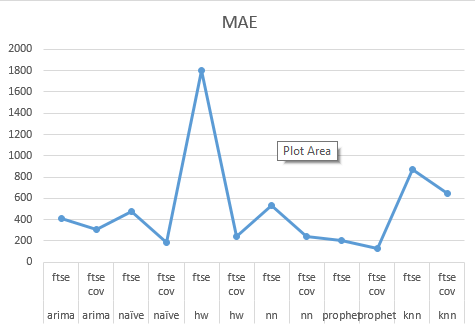
 

Figure 48 Train(left) test (Right) MAE Values

Results from MAE values are almost same as we saw with MAPE and RMSE.

## **6.4 Conclusion**

Looking at matrices we can conclude that ARIMA (for huge data) while holtwinter (for small data) are better with respect to MAPE RMSE, MAE and time constraint. While Prophet seems overfit, and KNN was unpredictable for our dataset. Overall accuracy of all the model was between 75-80 % which shows that success criteria set at the beginning of the project has been met.

## **6.5 Future Work**

Although our research has been successful, stock markets which is highly volatile includes huge risk of investment, Hence we need to improve our accuracy for minimizing loss and maximizing profits by taking better informed decisions. In this research our main focus was to support technical analysis of stock indices using historical data we started with an assumption that other non-measurable sentimental factors like company values, policies, profit loss statements, public sentiments are already reflected in the historical data, which is partially true. Now with the advancement of technology and digital media we can fetch such sentiments data from twitter by incorporating such data in our prediction model we can make our model do fundamental as well as technical analysis, which would generate more accurate results. Moreover in this paper we have used one model at a time where each model have its own pros and cons we can create a three level prediction model.

Stage 1 merge historic and sentiment data.

Stage 2 Create quick and basic model like Naïve which gives over 70% accuracy

Stage 3 move the result too final model like Prophet or LST.M or GARCH to enhance our model performance.

## **6.6 Learning Reflection**

Learning is a continuous process that either is acquired by gaining knowledge or by gaining experience.

For this project various blogs, articles, and research papers I acquired domain knowledge to start with the project. I can’t say that stock market was a new thing to learn but I lacked in-depth knowledge about it. Through the literature, I came to know about fundamental analysis, technical analysis, Bollinger bands, CCI index how they help in prediction, and Factors affecting stock prices. To create prediction models I used Time series modeling, instead of ML as stock market is a regular recorded timed data. Literature enlightened me about different types of Time series algorithms when and where to use a particular algorithm.

Gaining theoretical knowledge is easy but while implementing it practically we encounter various errors finding solutions to them gives us experience. I encountered such errors almost in every stage of development phase.

First problem was getting dataset in timeseries format. Dataset was easily available online and my project objective was to create mts class object for EU Stock Indices. Which required heavy processing and still I was getting erronatic values in date column. Fortunately in R there is a function getSymbols() which helped me in an instant and the downloaded dataset was an xts class object which supports advanced charting function.

Making the time series stationary by identifying and removing seasonal factors was again important. For any timeseries model we need to determine certain parameters like p,d,q for ARIMA and alpha, beta, gamma for HoltWinters. For p,d,q values we need to find ACF and PACF values with an integral lag.max my ACF and PACF chart repeatedly was showing decimal numbers in x axis which actually shows lags to get it in required format I kept changing my lag.max values still it showed decimals, when finally I realized that acf() I have to pass coredata(ts) as parameter whereas I was just using ts object.

Each day during project development phase was a learning experience. Practically I learned about seasonality, stationarity, ACF, PACF, different parameters used for developing a ts model, creating interactive candle stick chart, zoo charts, ggplots, different accuracy measures and how they are calculated. Whereas theoretically I learned about benefits and limitations of different time series model.

# References

1. Yan B., Aasma M. A novel deep learning framework: Prediction and analysis of financial time series. Expert Syst. Appl. 2020;159:113609.
2. Decomposition of Time Series Data of Stock Markets and its Implications for Prediction – An Application for the Indian Auto Sector. Sen & Datta Chaudhuri ( 2016a; 2016b;2016c;2016 d). (Sen & Datta Chaudhuri, 2016; Sen & Datta Chaudhuri, 2017a; Sen & Datta Chaudhuri 2017b; Sen & Data Chaudhuri, 2017c; Sen & Datta Chaudhuri, 2017d; Sen & Datta Chaudhuri, 2018a; Sen & Datta Chaudhuri, 2018b).
3. J. -S. Chou and T. -K. Nguyen, "Forward Forecast of Stock Price Using Sliding-Window Metaheuristic-Optimized Machine-Learning Regression," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3132-3142, July 2018, doi: 10.1109/TII.2018.2794389.
4. Fu, J.L., 2005. Agency conflicts, expropriation and firm value: Evidence from securities-market regulation in China. *Expropriation and Firm Value: Evidence from Securities-Market Regulation in China (March 1, 2005)*.
5. Alves, L.M., 2018. Price Clustering in Bank Stocks during the Global Financial Crisis.
6. Zhong, X. and Enke, D., 2017. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, *67*, pp.126-139.
7. Box, G.E., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M., 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.
8. Obuolo, E.P., 2020. *Prediction of Stock Prices Using Predictive Data Analytics. a Case of Nairobi Securities Exchange* (Doctoral dissertation, University of Nairobi).
9. Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K. and Wong, T.L., 2007. Design and implementation of NN5 for Hong Kong stock price forecasting. *Engineering Applications of Artificial Intelligence*, *20*(4), pp.453-461.
10. Enke, D. and Thawornwong, S., 2005. The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, *29*(4), pp.927-940.
11. Kumar, D., Sarangi, P.K. and Verma, R., 2020. A systematic review of stock market prediction using machine learning and statistical techniques. *Materials Science*, *2214*, p.7853.
12. Shastry, M., January 2021. Stock Price Prediction Using LSTM. Test Engineering and Management, 83(May-June 2020), pp. 5246-5251
13. Chen, C. H. &. L., 2020. Stock Prediction Using Deep Learning with Long-Short-TermMemory Networks. Issue December 20
14. Al-Mahasneh, A.J., Anavatti, S., Garratt, M. and Pratama, M., 2018. Applications of general regression neural networks in dynamic systems. *Digital Systems*, *10*.
15. Kotsialos, A., Papageorgiou, M. and Poulimenos, A., 2005. Long‐term sales forecasting using holt–winters and neural network methods. *Journal of Forecasting*, *24*(5), pp.353-368.
16. Gelper, S., Fried, R. and Croux, C., 2010. Robust forecasting with exponential and Holt–Winters smoothing. *Journal of forecasting*, *29*(3), pp.285-300.
17. Selvin, S., Vinayakumar, R., Gopalakrishnan, E.A., Menon, V.K. and Soman, K.P., 2017, September. Stock price prediction using LSTM, RNN and CNN-sliding window model. In *2017 international conference on advances in computing, communications and informatics (icacci)* (pp. 1643-1647). IEEE.
18. Suganya, R. and Shanthi, R., 2012. Fuzzy c-means algorithm-a review. *International Journal of Scientific and Research Publications*, *2*(11), p.1.
19. Suganya, R. and Sathya, S., 2022. Stock Price Prediction Using Data Mining with Soft Computing Technique. In *Expert Clouds and Applications: Proceedings of ICOECA 2022* (pp. 199-209). Singapore: Springer Nature Singapore.
20. Booth, David. (2012). Time Series Analysis. Technometrics. 39. 102-103. 10.1080/00401706.1997.10485448.
21. Chen, Yuh-Jen & Chen, Yuh-Min. (2013). A fundamental analysis-based method for stock market forecasting. Proceedings of the 2013 International Conference on Intelligent Control and Information Processing, ICICIP 2013. 354-359. 10.1109/ICICIP.2013.6568097.
22. Nyasha, Sheilla & Odhiambo, Nicholas. (2013). Stock Market Development In The United Kingdom: Prospects And Challenges. International Business & Economics Research Journal (IBER). 12. 725. 10.19030/iber.v12i7.7963.
23. Ho, Sin-Yu & Iyke, Bernard. (2017). Determinants of stock market development: a review of the literature. Studies in Economics and Finance. 34. 143-164. 10.1108/SEF-05-2016-0111.
24. Jose, Jonath. (2022). INTRODUCTION TO TIME SERIES ANALYSIS AND ITS APPLICATIONS.
25. Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.

Additional Articles and papers to gather subject knowledge

1. Singh, T., Kalra, R., Mishra, S. et al. An efficient real-time stock prediction exploiting incremental learning and deep learning. Evolving Systems (2022).
2. Arie Harel, Giora Harpaz, Forecasting stock prices. International Review of Economics & Finance Volume 73, 2021, Pages 249-256, ISSN 1059-0560.
3. RakhiMahant, TrilokNathPandey, Alok Kumar Jagadev, and SatchidanandaDehuri ―Optimized Radial Basis Functional Neural Network for Stock Index Prediction,‖ International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016
4. Roondiwala, Murtaza & Patel, Harshal & Varma, Shraddha. (2017). Predicting Stock Prices Using LSTM. International Journal of Science and Research (IJSR). 6. 10.21275/ART20172755.

# Appendix

Appendix 1 : Table of Figures

|  |  |
| --- | --- |
| Figure 1: Project Methodology | 14 |
| Figure 2 Sliding Window optimization by Chon and Nguyen | 16 |
| Figure 3: ARIMA Model | 18 |
| Figure 4: Holt Winter Exponential Smoothing | 18 |
| Figure 5: Naïve Bayes Equation | 19 |
| Figure 6: Working of Neural Network | 20 |
| Figure 7: KNN Algorithm Working | 21 |
| Figure 8: project Implementation Steps | 23 |
| Figure 8: Data Mining and Preprocessing Stages | 23 |
| Figure 9 Trend analysis of all EU Stock Indices and FTSE 50 with 100b days MA | 25-26 |
| Figure 10 FTSE 50 last 6 months with Bollinger bands, CCI and Directional Movement Index | 26 |
| Figure 11 FTSE 50 covid depression data with MACD values | 27 |
| Figure 12 Cumulative monthly return of FTSE 50 and CAC | 27 |
| Figure 13 Stationary plot ACF and PACF plot of FTSE 50, | 29 |
| Figure 14 stationary residuals, ACF and PACF plot for FTSE post covid dataset for ARIMA. | 29 |
| Figure 15 Model Building | 30 |
| Figure 16 auto.ARIMA results | 31 |
| Figure 17 Box Ljung Test result (top FTSE 50 model, below post covid model) | 31 |
| Figure 18 FTSE 50 ARIMA(010) | 32 |
| Figure 19 post covid ARIMA (0,1,0) | 32 |
| Figure 20 a, b residuals ACF, and PACF values for FTSE and Post Covid for naives | 33 |
| Figure 21 Naïve Bayes for FTSE 50 | 33 |
| Figure 22 Naïve Bayes for Post Covid | 33 |
| Figure 23 Forecast evaluation of Holtwinter FTSE model | 34 |
| Figure 24 final model with prediction test and actual FTSE Data HW | 34 |
| Figure 25 final model with prediction test and actual FTSE Data HW | 35 |
| Figure 26 neural networks residual check FTSE Data | 35 |
| Figure 27 residual chart for FTSE cov data | 36 |
| Figure 28 test – NN predicted plot for FTSE data | 36 |
| Figure 29 test – predicted plot for FTSE cov data | 36 |
| Figure 30 FTSE Prophet Trend Components | 37 |
| Figure 31 Post Covid Prophet Data Trend component | 37 |
| Figure32 Prediction Dataframe for FTSE and postcovid data | 38 |
| Figure 33 Prophet Model for FTSE 50 | 38 |
| Figure 34 Prophet Model for post covid | 39 |
| Figure 35 KNN model for FTSE data | 39 |
| Figure 36 KNN model for post covid data | 40 |
| Figure 37 different ARIMA model assessment | 41 |
| Figure 38 Train test accuracy and MAPE values for FTSE and FTSE post covid data | 42 |
| Figure 39 Accuracy matrix and MAPE Value | 42 |
| Figure 40 Accuracy matrix graph HW model | 43 |
| Figure 41 accuracy matrix NN | 43 |
| Figure 42 Accuracy matrix Prophet | 44 |
| Figure 43 accuracy measure KNN | 44 |
| Figure 44 Performance Matrix for train data FTSE and post covid data for all models | 45 |
| Figure 45 Performance Matrix for test data FTSE and post covid data for all models | 45 |
| Figure 46 Train(left) test (Right) MAPE Values | 46 |
| Figure 47 Train(left) test (Right) RMSE Values | 46 |
| Figure 48 Train(left) test (Right) MAE Values | 46 |