WROCŁAW UNIVERSITY OF
SCIENCE AND TECHNOLOGY
FACULTY OF ELECTRONICS

DEGREE PROGRAMME: INFORMATIC

SPECIALISATION: INTERNET ENGINEERING

**MASTER'S THESIS**

# Modele do prognozowania w oparciu o szeregi czasowe

# Models for time series forecasting

Author: RAJKUMAR BATHANI

Thesis advisor:

Dr hab. inż. Henryk Maciejewski

THESIS GRADED:

**WROCŁAW 2020**

# Acknowledgement

I would like to express my sincere gratitude to my supervisor Dr. Henryk Maciejewski for his consistent support and recommendations on this dissertation. He has spent a lot of his time in advising and supporting me, starting from formulating a problem to write this dissertation and to motivate me to choose a topic that motivates me to work hard, and hereby, I deliver my gratefulness and appreciation to him.

I would also like to thank my family, especially Mom and Dad, for the constant support they have given me throughout my time at PWr. Secondly, I would like to thank my friends for all the support, courage and guidance they have given me during this intense academic year.

At the end, I thank everyone who believed in me and supported me for continuing my studies abroad, without you all, this would not have been possible.

# Abstract

This study proposes a comparative study of chosen time series forecasting techniques. This dissertation intends to present time series forecasting techniques such as Autoregression, Moving Average, Smoothing Techniques and Prophet System. This thesis is focused on providing empirical comparison between all selected models in a selected business application context. In order to perform forecasting, time series datasets have been gathered. The chosen datasets are as follows: 1) Stock Data of Amgen and 2) Daily Minimum Temperature Data. Datasets have been chosen based upon their statistical properties. This thesis describes the forecasting process and necessary statistical tests to be performed prior fitting a forecasting model. This thesis talks about Literature Analysis on Data Mining Techniques to forecast Stock prices.

Thus, this research aims at making comparison between all models by fitting them on both datasets. At the end, it also aims at forecasting future values and compare the performance of the models in selected business context.

# TABLE OF CONTENTS

# Table of Figures

# List of Tables

# Chapter 1 Introduction

## 1.1 Motivation

Over the years, technology has changed our lives drastically, we became dependent on technology and digitalization. With the increasing number of users, a large amount of data is being generated rapidly. Most of the generated data are time-stamped data, it means time was added as an independent variable when the data was recorded. It is very useful when it comes to predicting future values by analysing past observations.

Any parameter in the real life that changes over time, time series can be generated on that parameter. Also, in almost every scientific field, a large volume of data being gathered. Such as finance, economics, banking, telecommunication, daily fluctuations of the stock market, inventory planning, production, weather forecasting, and so on. Additionally, for every small or huge business and supermarket, time series forecasting is common to track the price movement over time. To conclude, business owners are usually interested in performing forecasting, for economic reasons. Nowadays, time series covers a wide range of real-life problems. Thus, it is becoming a very popular topic in the field of data science and machine learning.

Notably, we cannot ignore the fact that the availability of these time series data has brought huge advantages in many areas and helped in solving complex problems. On the other hand, it is surely advisable to use right models to forecast datasets. Otherwise, the wrong forecasted values could lead to financial crisis and business would start downgrading. Despite the availability of many statistical models and machine learning methods, the easy management and analysis of these data still pose a great challenge.

My personal motivation behind writing this thesis is, I like to solve real life problems which are dependent on time. Finding a correlation between two events and determining future forecasting helps in finding solutions for many problems. There is also a lot to learn when we observe a dataset which changes according to time, getting into such datasets and observing insights is an inspiration to solve problems.

This thesis derives from such problems of choosing a right forecasting model for a given problem. This research will provide an empirical comparison of models in selected business content and provide a generalized approach for forecasting with a chosen model. By learning mathematics behind these models help in predicting accurate future values. I have chosen two datasets to derive conclusions for this thesis: Stock data of a Biotechnology company Amgen and 10 years minimum daily temperature data of Melbourne, Australia starting from 1981 to 1990.

## 1.2 Problem Statement

Time series datasets always have time as an independent variable, which adds additional information and makes time series problems more difficult to handle compared to other prediction tasks. Forecasting has been always considered as a difficult task to do as without knowledge of forecasting model it is very difficult to get correct future forecasting. Often it is observed that poorly fitted forecasting model leads to inaccurate results. Where forecasting stock market prices is a very difficult and challenging task because the datasets are often visualized with non-linear and non-stationary trends. For stocks price, time series is a widely used technique to track the fluctuations in stock prices over time. Time series forecasting experts habitually required to work with different types of dataset. So, identifying right patterns in the time series plays an important role. This research aims to provide an approach to perform statistical tests to identify the important patterns and nature of the datasets.

However, such forecasting can be difficult because of the involvement of many steps in the complete forecasting process. Also, there is a proper need of applying right model according to the nature of the datasets. By applying visualization techniques and some statistical methods, this can be achieved.

This thesis will try to study behaviour of the models by understanding the intuition and mathematics behind them on different datasets. Such associated problems with model selection and analysis are:

- The proper need of selecting and understanding the appropriate statistical models for forecasting. This dissertation will try to perform different tests on chosen datasets and will provide a generalized analysis of all models.
- Sometimes it is hard to identify the clear pattern in the dataset (trend, seasonality, cyclic changes and irregularity). Without performing important statistical tests, it is difficult to have correct forecast. This research also aims at providing all tests must be performed prior to fit a forecasting model.
- Finding the optimal parameters for model can be a challenging task to perform. Even though, availability of auto generated functions, sometimes, it is hard to find optimal parameters. There are some performance evaluation metrices are available to help with finding optimal parameters.
- There are a few experts available for forecasting with time series. Just fit a random model to any dataset would not give correct forecast.
- It is very challenging to understand the working procedure of models such as Moving Average and ARIMA based models. Therefore, analysts always get problem while forecasting. This study aims at describing model fitting scripts on different sizes of datasets. Which allows to study the model.
- Not always forecasted future values are trustworthy, this research will provide a study on when to rely on future forecasted values in a selected business context. By making empirical comparison between models this issue can be addressed.

## 1.3 Goal of the Thesis

The principal objective of this research is to study the several models based on auto regression, moving average and smoothing techniques, selected models such as (AR, MA, ARIMA, Exponential Smoothing techniques and Prophet System) for forecasting in a selected business context. To achieve the above defined goals, two datasets are derived based on their statistical time series properties ((1) Stock Data (Non-stationary) and (2) Temperature Data (Stationary)). In result, it would give clear understanding of the performance of these models on different datasets. Moreover, this research will also try to derive optimal solution for mentioned problems in the previous section. The detailed study of the mathematical behaviour of the models would help in getting accurate forecast values.

The focus of this research is not only on forecasting future prices but also to study the behaviour of the models. This research will also make an experimental comparison between all models from the obtained results, which will aim at providing when to fit which specific forecasting model. It is always advisable to choose a right forecasting model for the given problem. Sometimes, wrong forecasting model might lead to inaccurate forecast. This research also aims at providing best possible forecasting in selected business context (stock data) using the daily closing price and identify the trends and patterns in the dataset by applying different models and adjust the parameters of the models using optimization techniques.

This dissertation will discuss forecasting procedure using three different techniques such as: using auto regression, using moving average and using smoothing techniques. Also, it will fit a prophet forecasting model to achieve future forecast. The research is not only limited for the chosen datasets but will also provide a general approach on how to do forecasting with all chosen models. The obtained results and conclusions will help in determining a correct model for forecasting.

## 1.4 Thesis Outline

In order to attain the above-mentioned goals, this research document is structured as thus.

The thesis contents are mainly divided into three sections. The first section contains three chapters: Chapter 1, Chapter 2 and Chapter 3. The first chapter discusses the motivation behind choosing the research topic in time series forecasting. It emphasizes the problems and difficulties while forecasting and choosing a right forecasting model. In order to attain the conclusion, the set goals are also mentioned in the same chapter. The next chapter talks about the other possible forecasting approaches and what should be the forecasting approach. The third chapter is about introduction to time series. And will also explain the necessary statistical tests to perform before applying a forecasting model.

The second section comprises of three chapters: Chapter 4, Chapter 5 and Chapter 6. These chapters are mainly about the forecasting approach with chosen models based on Autoregression, Moving Average and Smoothing techniques. Moreover, the datasets were set

to different train-test sizes to analyze the performance of the models, where a model was applied on train data and performance was evaluated on test data. The fourth chapter illustrates details on forecasting with AR (Autoregression) model, MA (Moving Average) model and ARIMA (Autoregression Integrated Moving Average) model. The generalized approach is discussed starting from applying a model until it forecasts the results.

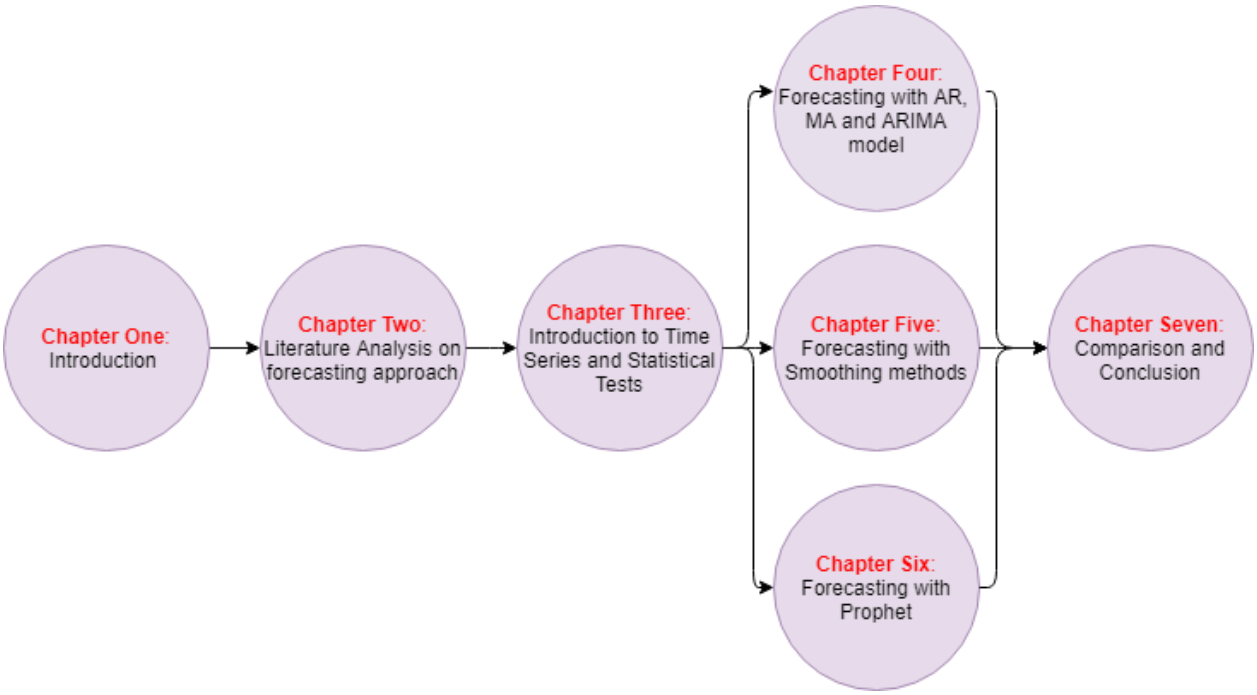



*Figure 1.1 Diagrammatic representation of thesis structure*

The next chapter is about forecasting with Smoothing methods such as: Simple Exponential Smoothing, Double Exponential Smoothing and comprehensive Exponential Smoothing. Correspondingly, the results from smoothing techniques are compared to evaluate the differences between them. The sixth chapter is about applying a forecasting tool called Prophet to the datasets. Which is developed by Facebook's Data Science team. Prophet is considered as one of the best tools to make future forecasting. The Figure 1.1 shows how the whole dissertation is structured.

The third section contains only one chapter, which is Chapter 7, It compares the forecasted future values by all the forecasting models studied in this research. It describes how accurate are the future forecasted values. This chapter also discusses about generalized analysis of forecasting models.

# Chapter 2 Literature Analysis

## 2.1 The Forecasting Process

According to (Montgomery, Jennings, & Kulahci, 2008) [1] forecasting is a process of several activities. A forecasting process includes activities such as:

1. Defining a problem definition
2. Collecting data in a selected business application
3. Analyzing data
4. Selecting appropriate model and fit the model to the dataset
5. Validation of selected model
6. Model deployment for forecasting
7. Evaluating model performance and monitor performance

Problem definition involves setting boundaries of forecast and managing expectations of the customer. Also, it aims on deciding what accuracy matrices should be used based on the requirements. In addition, it adds business decision approach to identify associated risks with forecast. Data collection entails of collecting historical information and identify if all data is useful for the current problem. Often it is required to handle missing values or solving other data-related problems. During this phase it is suggested to initiate planning how the data collection in the future will be handled so that the reliability and integrity of the data will be preserved.

The most important step is to analyze the data by constructing various time series plots, it is possible to recognize patterns, such as trends and seasonality or other cyclical components. These results can be obtained by performing statistical tests that help in building a strong foundation for choosing a model. The purpose of this step is to obtain some understanding of data and a sense of how strong the underlying patterns are.

Once the patters are identified, the next step is to choose one or more forecasting models and fitting the model to the data. This step also aims in identifying the parameters which give optimal result. Model validation handles the evaluation procedure of a model which is to determine how likely it is to perform in the intended application. A widely used method for validating a forecasting model is to divide a data into two sets - a training set and a testing set, which is also called as train-test split. The model is fit to only the training data segment, and then forecasting is done on test data segment. This can provide useful guidance on how the model will perform when introduced to new data.

Model deployment involves getting the model and resulting forecasts. Also, the comparison of forecasting models can be carried out in this step. And the last step is to monitor the performance, it should be done once the model has been deployed to ensure it is performing satisfactorily. It is also assumed most of the time that according to the nature of forecasting it

changes over time, a well performed model in past may deteriorate in performance. Usually performance worsening will result in large forecasting errors.

The above-mentioned procedures are a stepwise approach to do forecasting with each individual process has an equal contribution. For a given dataset, if one fails to fulfill the requirements of any of the above-mentioned procedures then the probability of getting incorrect or biased predictions is very high.

## 2.2 Data Mining Techniques for Financial Stock Market Forecast

The research study of (Senthamarai, Sailapathi, Mohamed, & Arumugam, 2010) [2] has developed an automated computer program using data mining and predictive technologies to do trading. Where the focus was on identifying hidden patterns from historical data which helps investors to make wise investment decisions. Five methods were combined to predict the movements of the closing prices in the future. The used methods were Typical Price (TP), Bollinger Bands, Relative Strength Index (RSI), Chaikin Money Flow Indicator (CMI), and Moving Average (MA). Also, it examined numerous global events and their issues predicting on stock markets.

This is how all the different methods were calculated. According to (Senthamarai, Sailapathi, Mohamed, & Arumugam, 2010) [2] Typical Price was calculated by adding the high, close, and low prices together, and then further the combined prices were divided by three. Chaikin Money Flow Indicator is a technical analysis indicator used to calculate the flow of all the investments. This condition would be indicative of security risks associated with investments. If CMI flow value is below zero indicates that security is under selling pressure. And the reading of the above +0.25 is considered to be strong buying pressure. Relative Strength Index compares the total number of days a stock stays up and stays down. It is calculated between various time spans usually between 9 and 15 days. It generally calculates if the stock is overbought or is oversold. Moving Average method were used to keep track of the average price over a period of time. When prices fall below the moving average it tends to keep falling over time. Contrariwise, if prices were accounted above the moving average then it is considered to have a rising tendency [2].

Consistent with (Senthamarai, Sailapathi, Mohamed, & Arumugam, 2010) [2], the Bollinger Signal is decided based on drawing three lines, which mostly based on market volatility. The lines were used to make a smarter decision. The upper and lower lines suggest the volatility, where widen space between lines considered as high volatility and closer lines indicated as less volatility. And the middle line was moving average line. So, these three lines used to capture the stock movement.

The result (Senthamarai, Sailapathi, Mohamed, & Arumugam, 2010) [2] proved that the combined techniques algorithm has (50%) a better chance of predicting the following day’s closing price would increase or decrease. Where it was observed that the algorithm performed

well only on half of the stocks. So, the prediction was correct at least 50% of the time. Hence, it was believed that the winning probability is only 50%.

## 2.3 Neural Networks and Forecasting

According to (Montgomery, Jennings, & Kulahci, 2008) [1] an artificial neural network can be used to solve different complex problems. In the forecasting procedure, neural networks are used to solve high-dimensional, nonlinear data prediction problems. There are neural networks such as; ANN (Artificial Neural Network), LSTM (Long short-term memory) and Multilayer feedforward artificial neural network are most popular neural networks to solve prediction problems.

## 2.4 Forecasting Stocks with Multivariate Time Series Models

Basically, a time series is divided into two different types: Univariate and Multivariate. This thesis aims on forecasting with univariate time series where a time series has only one-time dependent variable. A multivariate time series consists of more than one-time dependent variable. Each time dependent variables in a multivariate time series has some dependency on other variables. VAR (Vector Auto Regression) is one of the most commonly used techniques for multivariate time series forecasting. In data analysis, sometimes it is reasonable to use multivariate time series because flexibility offered by univariate time series could be limiting.

(Iwok & Okoro, 2016) [3] developed a VAR model to forecast six different Nigerian banks stock prices that were found to be analytically interrelated. VAR model gives possibility to build a model using set of time dependent variables. Vector Auto Regression model constructs equation that makes each endogenous variable a function of their own past values and also past values of all other endogenous variables. The best fit model is chosen by selecting all model with least AIC (Information Criterion) model. Where, model diagnostic check was performed after a model was fitted to the dataset. The further diagnostic process was carried out by examining the behaviour of the residual matrices. It was believed that residuals are expected to follow a white noise pattern, which basically means that there is no underlying patter in the residuals. The best accounted model was VAR (1). It means the least AIC value was observed at lag 1.

In result, it can be said that multivariate time series models are useful only when the correlation between various time dependent variables is established. Also, the domain knowledge is required to perform such analysis. Sometimes, it may happen that wrongly identified interrelation between variables leads to incorrect forecast. The mathematical complexity of multivariate models is comparatively higher than univariate models. Hence, it is advisable to use right strategy to choose a model for forecasting.

# Chapter 3 Theoretical Background on Time Series and Statistical Tests

## 3.1 What is a Time Series?

A time series is a sequence of data points which can be generated by observations or successive measurements. The data points are usually taken on regular intervals (weeks, days, months, years), but the observations could also be irregular. To build successful forecast, it is mandatory to first perform the statistical tests and identify the nature of the time series. The results of these tests help in building key foundation for any type of forecasting model.

Each time series contains trend and seasonality. By doing an ETS decomposing on a time series divides a time series into three parts: Error, Trend and Seasonality. Sometimes, it is very difficult to say whether a time series has clear trend or seasonality. But, by decomposing a time series helps in easily determining these factors.

Where, trend represents, an upward or downward growth of the observations. And seasonality is a term which refers to the cyclic pattern in the dataset. It could be daily, weekly, monthly or yearly. And error term referred to a set of observations which are left in the time series after decomposing it to error, trend and seasonality. The next section gives introduction to the selected time series datasets.

## 3.2 The Datasets

In order to achieve goals and study the detailed statistical methods and models two datasets have been chosen with different characteristics such as: stationarity, seasonality and trend.

### 3.2.1 The AMGEN stock dataset

The stock prices (values such as: High, Low, Open, Close, Total traded volume) of the Biotechnology company named “AMGEN” headquartered in California, United States have been used as one of the datasets used in this research. With solely aim on forecasting future values and some advanced analytics graphs have been plotted to visualize the general pattern being followed throughout the time series.

The stock market is a general term alluding to the exchanging of securities through different physical and electronic trades over the counter market. The securities exchange is one of the most essential zones of a market economy since it gives organizations access to capital by permitting speculators to purchase portions of ownership in an organization. Which often known as percentage holders of the companies. By purchasing some stocks of possession, investors have possibilities to perhaps pick up money by benefitting from organizations' future thriving. Well, there are a few numbers of investors who make profit, because the stock prices keep fluctuating over time and it is very difficult to say about the future values.

| Date | Open | High | Low | Close | Adj Close | Volume |
|---|---|---|---|---|---|---|
| **2013-01-02** | 87.360001 | 89.239998 | 87.290001 | 89.150002 | 74.684219 | 5772800 |
| **2013-01-03** | 89.529999 | 89.580002 | 88.339996 | 88.589996 | 74.215080 | 3871300 |
| **2013-01-04** | 88.589996 | 89.279999 | 88.419998 | 88.980003 | 74.541794 | 3280500 |
| **2013-01-07** | 88.489998 | 88.800003 | 87.760002 | 88.529999 | 74.164810 | 2573300 |
| **2013-01-08** | 88.440002 | 88.680000 | 87.440002 | 88.150002 | 73.846474 | 5170700 |
| **2013-01-09** | 88.339996 | 88.919998 | 88.209999 | 88.669998 | 74.282089 | 3372500 |
| **2013-01-10** | 88.320000 | 89.000000 | 87.269997 | 87.809998 | 73.561653 | 5861200 |
| **2013-01-11** | 87.949997 | 88.050003 | 86.870003 | 86.959999 | 72.849571 | 4681700 |
| **2013-01-14** | 87.169998 | 87.250000 | 86.500000 | 86.919998 | 72.816048 | 4401000 |
| **2013-01-15** | 86.559998 | 86.860001 | 85.000000 | 85.080002 | 71.274628 | 6277900 |

*Figure 3.1 The AMGEN stock data for 6 years starting from 2013 until 2019.*

The Figure 3.1 (above) illustrates that how the time series of stock data looks like, where it can be seen that values such as: Open, High, Low, Close, Adjusted Close and Traded volume have been captured.

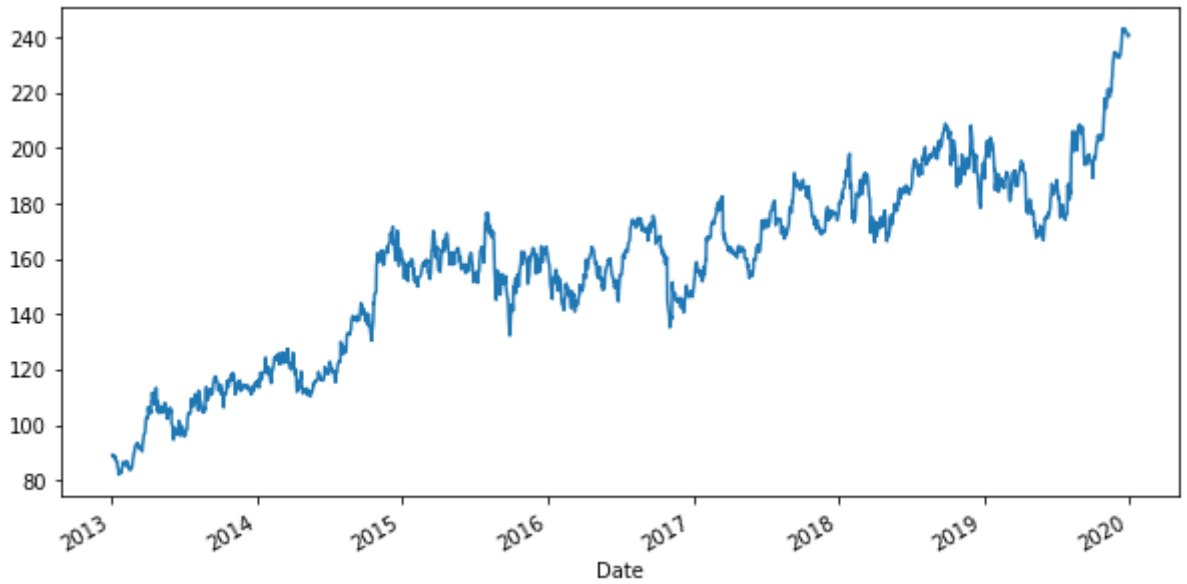



*Figure 3.2 The fluctuations of Close stock prices throughout the time series.*

The values of daily closing prices are used for model fitting and forecasting as it captures all the fluctuations happened during a day. The frequency for this time series was set to business days which basically means that values are gathered only on business days and holidays are

not taken into consideration. Values such as highest and lowest prices are used to plot the graphs in order to analyze the patterns over time.

### 3.2.2 The Temperature dataset

The second dataset is a time series captured on daily minimum temperature data recorded in Melbourne, Australia for a period of 10 years, starting from 1981 until 1990. The frequency for this time series was set to daily data, for details see Figure 3.3 (below). The main purpose of using this dataset was not to forecast but instead study model's behaviour as this time series has no trend nor does it have clear seasonality, as it can be said it is stationary dataset.

| Date | tempmin |
|---|---|
| 1981-01-01 | 20.7 |
| 1981-01-02 | 17.9 |
| 1981-01-03 | 18.8 |
| 1981-01-04 | 14.6 |
| 1981-01-05 | 15.8 |
| 1981-01-06 | 15.8 |
| 1981-01-07 | 15.8 |
| 1981-01-08 | 17.4 |
| 1981-01-09 | 21.8 |
| 1981-01-10 | 20.0 |

*Figure 3.3 The daily minimum temperature data recorded in Melbourne, Australia, from 1981 to 1990.*

The temperature dataset reveals patterns over time. It is common that temperature dataset follows kind of similar pattern over the year, the time period during winter observes lower temperature than summer and during rainy season, temperature remains within a moderate range which is always in between the highest temperature recorded during summer and lowest measured during winter.

Such patterns are not easily identifiable in all time series datasets. For example, Stock prices datasets contain frequent fluctuations and not often have a clear trend either upward or downward.

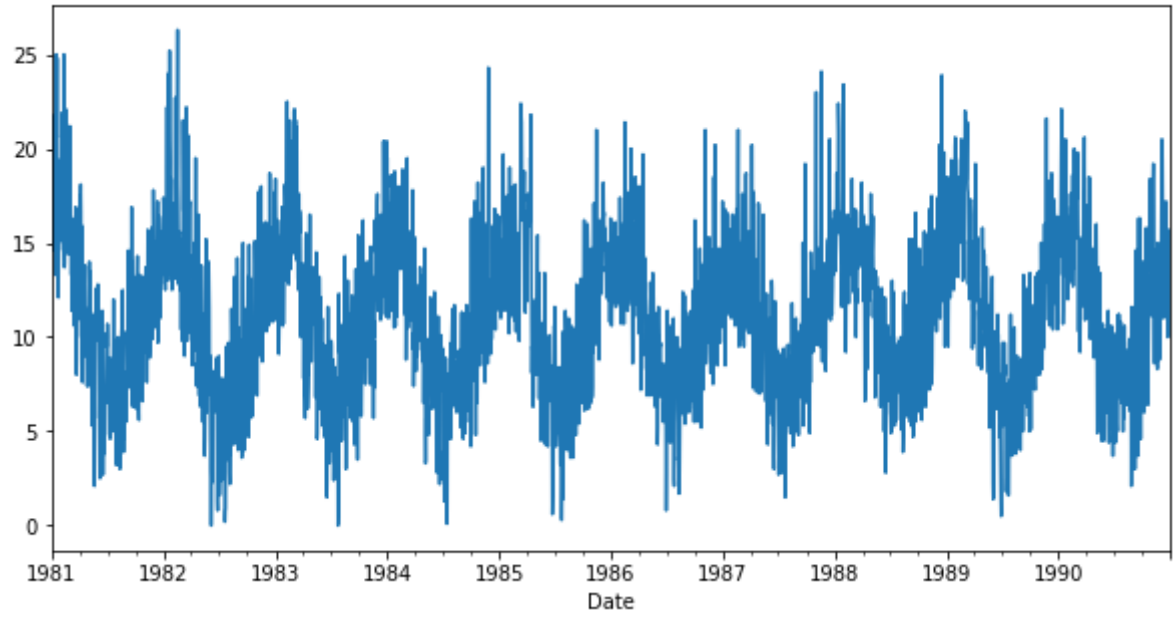



*Figure 3.4 The daily minimum temperature values have been plotted to visualize patterns.*

## 3.3 Stationarity

Stationarity is a key statistic when it comes to time series analyses and working with time data. The question behind it is, has the data contains same statistical properties throughout the time series?

- Variance
- Mean
- Autocorrelation

Most analytics procedures in time series analysis require the data to be stationary, or the model function will make the data stationary during the model set up if data is not yet stationary. There are techniques such as transformations and differencing can be used in case of non-stationary data. Differencing basically adjusts data according to the time span that differ in variance or mean. The goal of differencing is to make the statistics the same throughout the whole dataset.

Generally, time series often have a trend in them. It can be observed by calculating mean that it changes over time as a result of trend. If the de-trading procedure was applied, then it would be possible to check the stationarity in the dataset.

However, it is very difficult to decide whether a given time series is stationary or not; sometimes it is very confusing such as: a time series with cyclic pattern over time but have no trend and seasonal pattern in it- it is stationary time series then. To conclude, we have a statistical test called augmented Dickey-Fuller test (also called Unit Root Test) can be used to perceive if the dataset has stationarity.

### 3.3.1 Determining Stationarity of the datasets (By performing Dickey-Fuller test and plotting graphs)

The Dickey-Fuller test is a statistical test in the form a null hypothesis test which was developed with some assumptions and calculates the p value.

- The test presents null hypothesis which states, that the value of Φ = 1 (this is also called a unit test).
- If the returned p value is low (<0.05), then the null hypothesis will be rejected, and it can be concluded that dataset is stationary.
- If the returned p value is high (>0.05), then the null hypothesis will not be rejected.

```
Augmented Dickey-Fuller Test:
ADF test statistic        -1.174090
p-value                    0.684641
# lags used                0.000000
# observations          1761.000000
critical value (1%)       -3.434069
critical value (5%)       -2.863183
critical value (10%)      -2.567645
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

*Figure 3.5 Dickey-Fuller test on Stock dataset*

The most important part here is to analyze the p value and based on that to determine the result. From the Figure 3.5(above), it can be observed that it is (>0.05). Hence, the time series is not stationary, and we do not reject the null hypothesis. While, the test result is totally different for second dataset, where the p value was calculated as 0.00. Thus, the null hypothesis can be rejected, and it can be said that it is stationary, for more details see Figure 3.6 (below).

```
Augmented Dickey-Fuller Test:
ADF test statistic        -4.441196
p-value                    0.000251
# lags used               20.000000
# observations          3631.000000
critical value (1%)       -3.432152
critical value (5%)       -2.862336
critical value (10%)      -2.567194
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

*Figure 3.6 Dickey-Fuller test on Temperature dataset*

The same results can also be determined by plotting the plots of mean and standard deviation on both the time series. If values of mean and standard deviation do not fluctuate over time

and remain constant, then it can be said that given time series is stationary or it is not stationary.

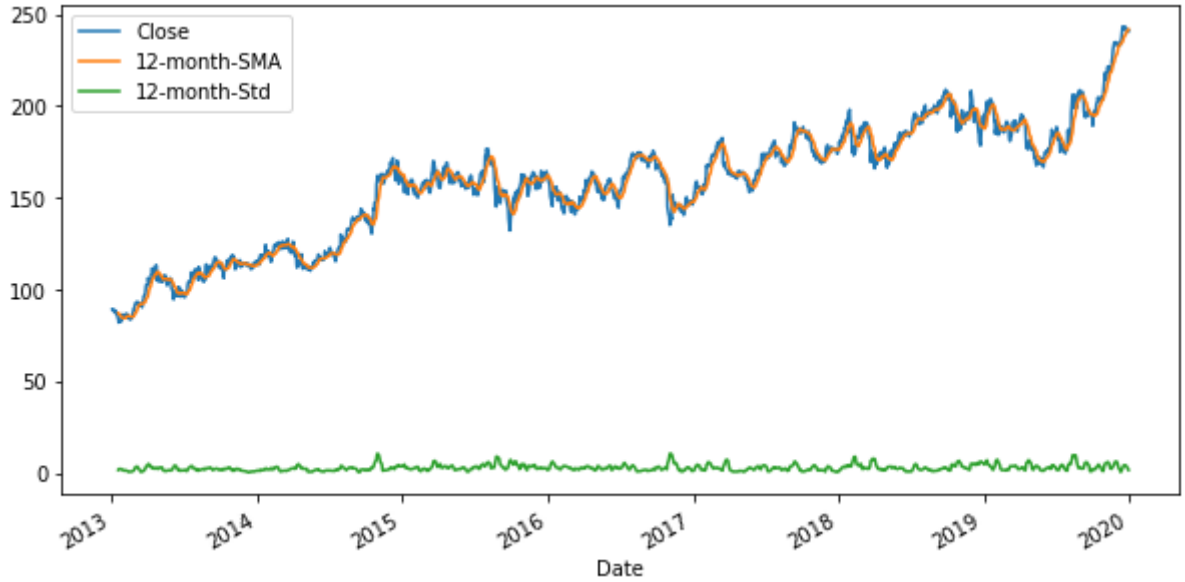



*Figure 3.7 The graph of mean and standard deviation to determine the stationarity(Stock data)*

From Figure 3.7, it can be said clearly, that the graph of mean values changes over the time as time series progresses but values of standard deviation remain constant. In result, it can be concluded that given time series of stock prices is not stationary, as mean values keep changing and that was also proved by performing Dickey-Fuller test. Well, it is also clear to notice upward trend in the dataset.

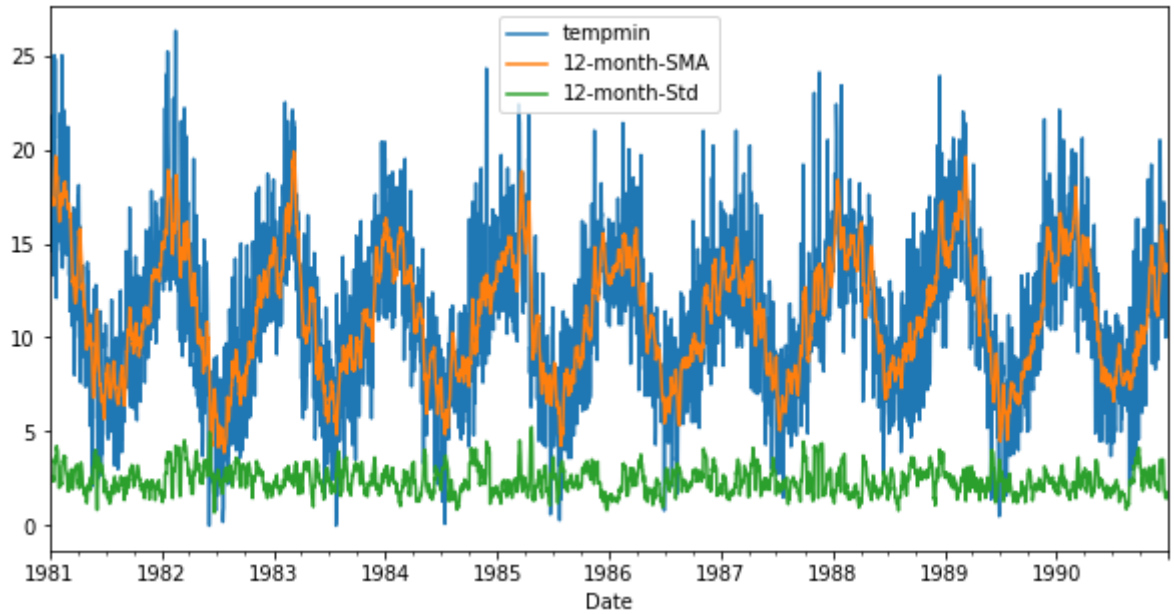



*Figure 3.8 The graph of mean and standard deviation to determine the stationarity(Temperature data)*

The temperature dataset was proven to be stationary from the results of the Dickey-Fuller test and the plot in Figure 3.8 also adds more information into the fact by showing constant values of mean and standard deviation over time. Likewise, the mean values were changing but

never went beyond threshold and stayed within a range, which mean the time series does not have clear trend.

## 3.4 ACF and PACF theory

ACF and PACF are important plots used to identify a relationship between an observation in time series with observations at prior time stamp. ACF stands for autocorrelation and PACF stands for partial autocorrelation. They can also be used to determine the parameters for AR and MA components in ARIMA model. Correlation is a measure of how strongly the two variables are linearly related. It goes between a range of (-1, +1), if the correlation is closer to -1, it means the stronger the negative linear relationship. And if the correlation is closer to +1 then it means the positive linear relationship is stronger.

ACF plot is used to find the correlation between the observation at the current time stamp and the observations at previous time stamps. For example, if we believe that today's stock price is related to yesterday's stock price then we can calculate ACF value to measure how strong two day's prices are related. Autocorrelation is basically applied on univariate time series. So, the values are compared with itself, lagged by x time units. So, the y axis is the correlation and the x axis are the number of time unit lags. It is essentially possible that in general there is a decline of some sort, the further away you get with the shift, the less likely the time series would be correlated with itself.

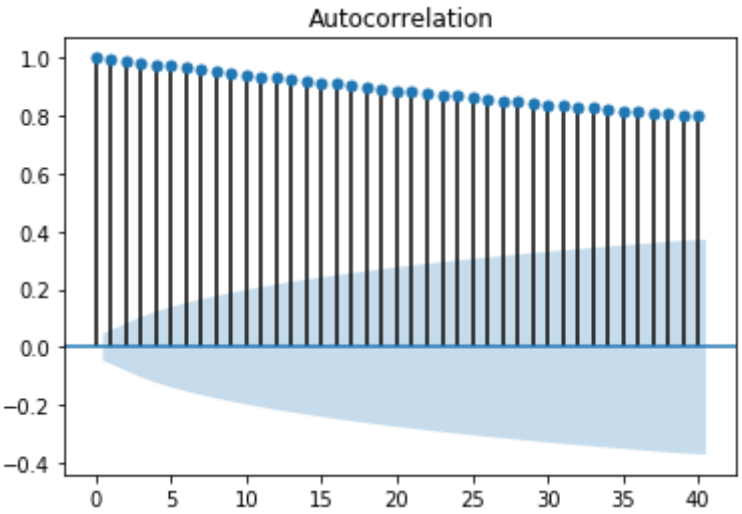



*Figure 3.9 ACF plot on Stock data*

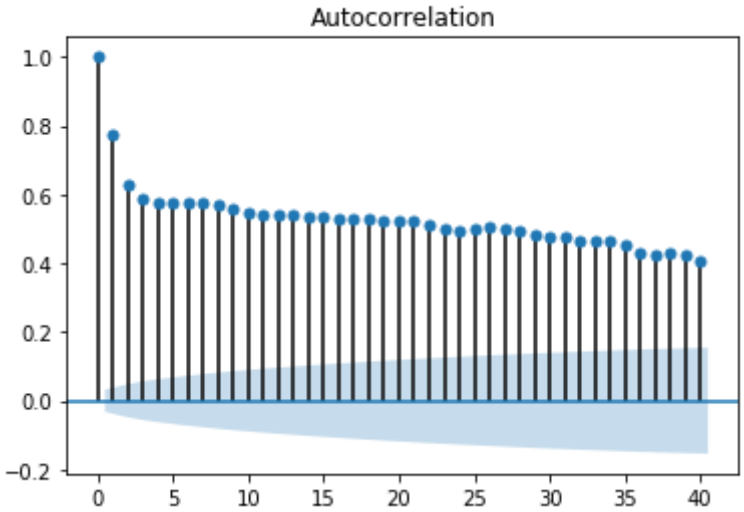



*Figure 3.10 ACF plot on Temperature data*

By observing ACF plots for both the datasets, it can be clearly seen that ACF plot for non-stationary data (Figure 3.9) does not show any sharp drop off, because of presence of either trend or seasonality component. While, on the other hand ACF plot for stationary data (Figure 3.10) has shown sharp drop off, which is visible after first lag observation.

On the other hand, PACF plot measures correlation between the today's values and values in the past, but also considers influence of the other days. For example, today's stock price can be correlated to the day before yesterday. So, PACF of yesterday is the actual correlation between today and yesterday after taking out the influence of the day before yesterday.

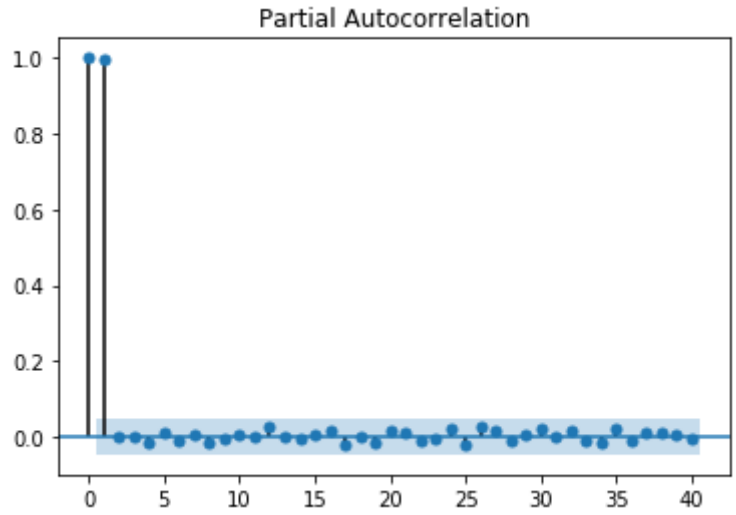



*Figure 3.11 PACF plot on Stock data*

The careful observation of PACF plots for both datasets shows that it is easy to analyze PACF plot for the data that is already stationary.

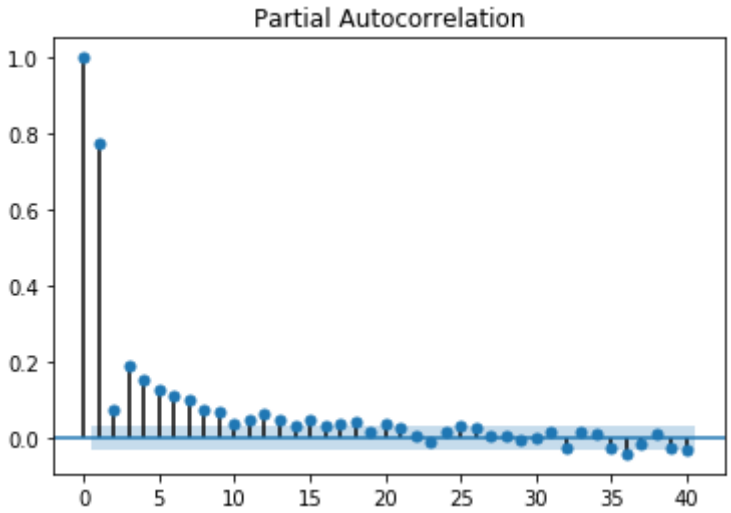



*Figure 3.12 PACF plot on Temperature data*

## 3.5 Data Pre-processing and Preparation

Data Pre-processing is the most important step in any forecasting or data analysis process. In order to build a proper forecasting model, the data must be cleaned. Pre-processing tasks include; filling missing values, setting the frequency which is the most important part in time series forecasting.

In this dissertation, both datasets have gathered from the authentic sources. Thus, they do not contain any missing value. Even, time series datasets do not contain missing values, as chronological order is important in time series. But I had to set time series frequency according to the cycles time series contain. For example, if a time series has daily data then frequency will be “D”. If a time series has monthly data, then the frequency will be “MS”. Some time series might have frequency of business days such as Stock Market data. Then the frequency must be set to “B”.

Sometimes, after adding frequency to a time series, there might be possibility of a few number of additional rows, this is the most common problem with business data frequency. In that case, according to the business requirements, those missing values must be filled logically. Such operations of filling missing values incudes; filling previous observations value, filing one-time ahead observations value, or filling with the average of all the observations. This is how both datasets were prepared and pre-processed.

# Chapter 4 Forecasting with AR (Autoregression), MA (Moving Average) and ARIMA (Autoregressive Integrated Moving Average) Models

## 4.1 AR (Autoregression) Model

Autoregression model predicts future values based on its own past values. Attesting the previous sentence, it is even advisable while working with time series data sets to observe the past values and correlation between the values. AR model forecasts a series based solely on the past values in the series-called lags. In addition, it is natural to build a correlation between current values and past values, however, there could be some external factors that affecting predictions. But we should give more priority to the past values, for example, a production company produces goods by considering the total consumption happened in the past month.

Generally, an AR model can be denoted as AR(p), where p denotes the number of lags observations. Where, a first order AR model is written as AR (1). And mathematically the AR model can be formulated as according to Eq. (4.1) where $y_t$ is the current time stamp where we are predicting the values and $y_{t-1}$ is representing the values of y during the previous period. For instance, if $t$ is the current period and we have weekly values, then $t-1$ would represent last week's values. The coefficient Φ is any numeric constant by which we multiply the lagged variable. And the value of Φ will always be between -1 and 1. A residual $\varepsilon_t$ gives the difference between our prediction for period $t$ and the correct value.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{4.1}$$

The AR (1) model with lag = 1 would mathematically look like Eq. (4.2) where current values of $y_1$ depends on only one past value. Likewise, A time series with AR (2) model would be written as Eq. (4.3).

$$y_1 = c + \phi_1 y_{t-1} + \varepsilon_t \tag{4.2}$$

$$y_2 = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \tag{4.3}$$

The Autoregression basically build linear model between current values and values recorded during previous days. However, we can think of some external factors that might affect the forecast significantly. AR model often gives correct forecast with a time series with no trend or seasonality, it means stationary. In normal terms, an AR model is a linear model, which assumes that values at given period are equal to some portions of values in the last period with some constant benchmark $c$ and some unpredictable shocks which generally describes as error terms. It is vital to understand that any AR model cannot be used on given time series, but it is essential to determine how many lags to include in the analysis. The complexity of the model proportional to the number of lags used, a greater number of lags requires to

determine more coefficients and thus there is a possibility that some of them would not be significant.

### 4.1.1 Forecasting Using AR model on Temp data

Autoregressive models are basically a regression technique. Instead of looking at value at time period t, it considers value at time period t-1. The Temperature time series is stationary; it does not have trend and it also does not radiate clear seasonality. Hence, AR model might be the best fit, as it often neglects to predict seasonality from time series.

- **Steps for fitting AR model on time series**:

**Step 1**: Start fitting AR (1) model on time series and calculate AIC (Information Criterion); which is used to compare the performance of ARIMA based models.

**Step 2**: Then after fit another AR model with higher order of parameters and check the AIC, a model with least AIC value will be chosen as the best model; this procedure called model estimation.

**Step 3**: Once the model has been estimated, then we need to apply the model on the dataset and forecast. For, forecast accuracy evaluation RMSE (Root Mean Squared Error) metrics has been used, which generally calculate the difference between predicted value and actual value.

**Step 4**: The final step is determining a train-test dataset which gives least RMSE value.

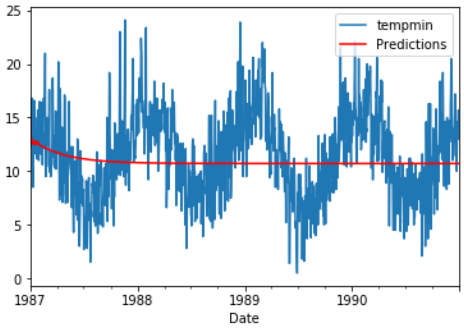
*Table 4.1 choosing the correct AR model*

| AR (p) | AIC (Information Criterion) |
|---|---|
| **AR (1)** | 1.8916 |
| **AR (2)** | 1.8916 |
| **AR (3)** | 1.8499 |
| **AR (4)** | 1.8273 |
| **AR (5)** | 1.8111 |
| **AR (10)** | 1.7790 |
| **AR (15)** | 1.7678 |
| **AR (20)** | 1.7665 |
| **AR (23)** | 1.7659 |
| **AR (27)** | 1.7647 |
| **AR (29)** | 1.7642 |

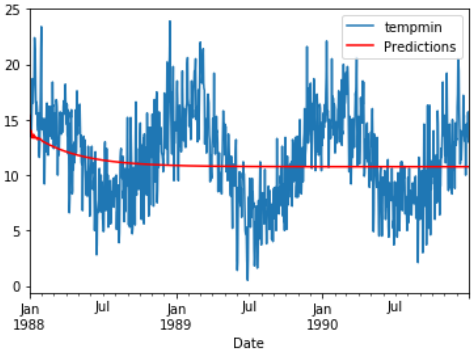
According to the Table 4.1, AR model with 29 lagged variables was best fit; other models with higher lag order would be overfitting the model. And the same order can be estimated by using AR () function from Stats models in Python. The next step would be to divide a dataset into different train and test sizes, the below Table 4.2, illustrates the calculated RMSE (Root Mean Squared Error) values with AR (29) model on different train and test sizes. The least AIC was calculated for AR (29) model. The use of more lagged variables states the more accurate results, meaning showing strongly correlation between past observations.

*Table 4.2 Accuracy (RMSE) values according to the train-test split sizes*

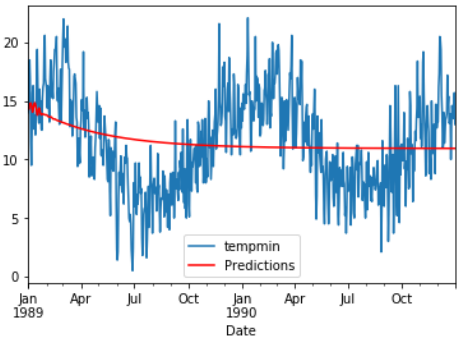
| Train | Test | RMSE |
|---|---|---|
| 2191 (60%) | 1461 (40%) | 3.958 |
| 2556 (70%) | 1096 (30%) | 3.940 |
| 2921 (80%) | 731 (20%) | 3.955 |
| 3286 (90%) | 366 (10%) | 3.611 |
| 3622 (99%) | 30 (1%) | 3.281 |
| 3644 (99%) | 7 (1%) | 1.229 |



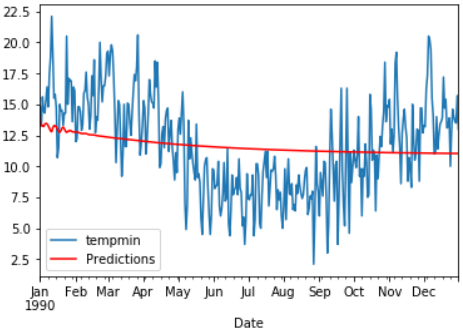

*(A) 60% train and 40% test*



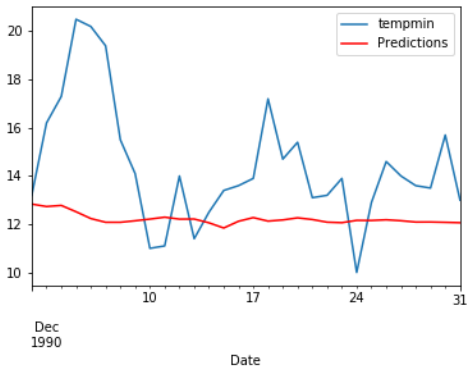

*(B) 70% train and 30% test*



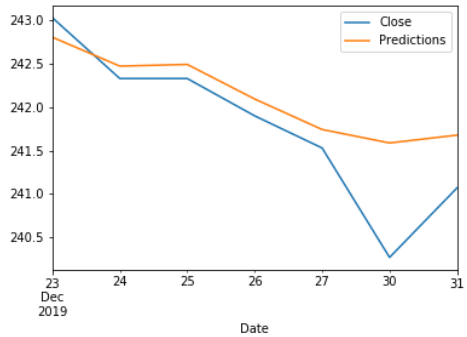

*(C) 80% train and 20% test*




*(D) 90% train and 10% test*




*(E) 1-month forecasting period*




*(F) 1-week forecasting period*

*Figure 4.1 Forecasting with different train-test splits, (A)60% -40%, (B)70%-30%, (C)80%-20% , (D)90%-10%, (E) 1 month and (F) 1 week.*

The AR model is very sensitive towards the length of the forecasting period. Thus, Table 4.2, shows how the value of RMSE is decreasing according to the decreasing test length. The lowest RMSE value was calculated when the length of the test set was at minimum. And highest RMSE value was recorded when train and test sizes were divided into 60% and 40% in length each.

Likewise, forecasting comparison plots are also plotted for each train and test size split. From Figure 4.1, it is observed that compared to the fluctuations in actual data and what is predicted by the model is not similar, but when the forecasting period was only a week, the predicted values are similar to actual values. Otherwise, the predicted values were range within a straight line which was assumed to be predicted equal to the average value of a period. The probability of how much we can rely on future predictions made by Autoregression model does not evolve strong desirability.

### 4.1.2 Forecasting Using AR model on Stock Data

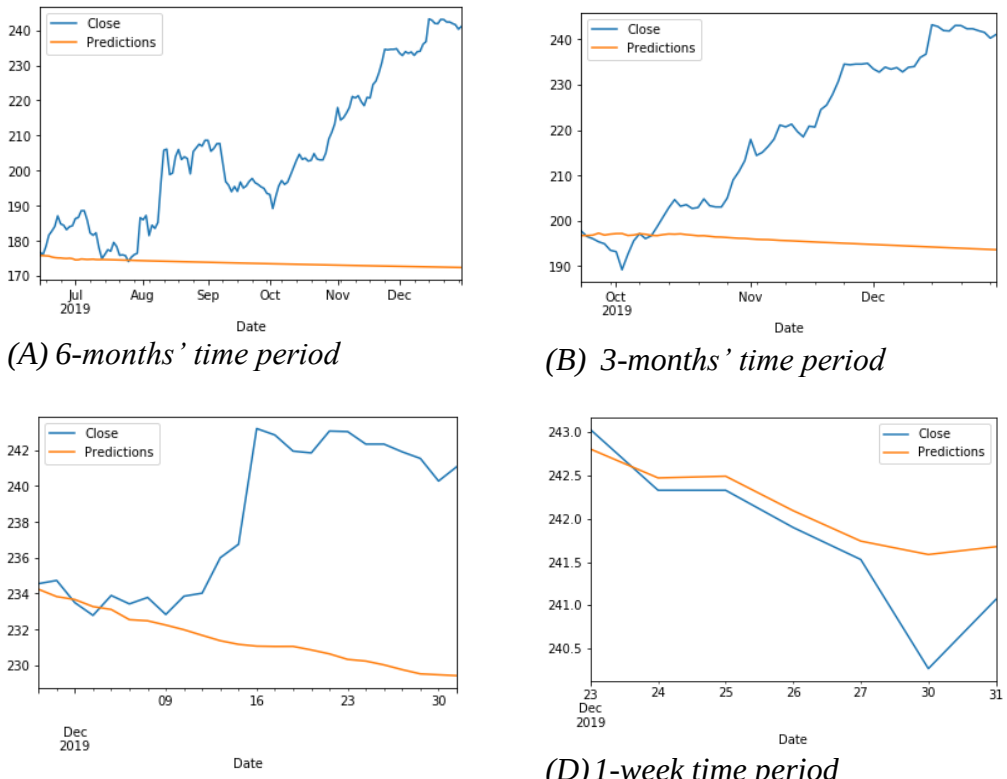
The stock time series is not stationary; hence, there is either trend or seasonality present in it. The dataset is split into train and test. Where, train and test sizes kept to different ratio. It is often suggested that the length of the test dataset must be equal to the length of the forecasting time period. The AR model has been applied on each dataset and it was calculated as AR (25), it means AR model with 25 lagged variables was best fit and the least AIC was calculated for AR (25) model.

*Table 4.3 Accuracy (RMSE) values according to the length of forecasting period*

| Forecasting Period | Train | Test | RMSE |
|---|---|---|---|
| **6 months** | 1681 | 144 | 37.415 |
| **3 months** | 1753 | 72 | 29.639 |
| **1 month** | 1801 | 24 | 8.548 |
| **1 week** | 1818 | 7 | 0.571 |

For Stock data, the RMSE values gave contradictory results, that can be proven from Table 4.3. Where, the highest RMSE value was observed for 6 month forecasting period. And that is 37.41. It can also be seen that RMSE value is decreasing according to the length of test data set. The lowest RMSE value was derived when forecasting period was set to 1 week and that was considerable lowest error value ever recorded, that was 0.571. All these RMSE values obtained by fitting AR (25) model. To conclude, results observed for both datasets, clearly demonstrates that AR model does not capture trend nor seasonality in the dataset.

Forecasting comparison plots (Figure 4.2) also shows that AR model does not detect any patterns in the dataset, except when forecasting period is too short, it detects nearly the same values. While modelling with AR model, it is necessary to select a time series with either linear trend or without any patterns. Such strange behaviour of AR model on Stock data, is may be because of clear trend and frequently fluctuating values. Where, such behaviour was not observed for stationary Temp dataset.




*(A) 6-months' time period*

*(B) 3-months' time period*

*(C) 1-month time period*

*(D) 1-week time period*

*Figure 4.2 Forecasting with different forecasting time periods, (A)6 months, (B)3 months, (C)1 month and (D)1 week.*

The AR model does not able to capture a lot of movements throughout the time series. Instead, it predicts the values within the range of average values.

### 4.1.3 Forecast Improvement by Implementing Walk-Forward Validation

Walk Forward validation technique helps in improving the accuracy of forecasting as at each iteration it adds previously forecasted values to the dataset; by which every time model would know the previously forecasted values and error. Thus, the forecast can be improved. At each step, we will create an individual model for each of the test value prediction. And in the next step, we will increase our training dataset to include the new information, then we will again create a new model and forecast for one time period ahead.

- **Steps to implement Walk Forward Validation**:

**Step 1**: Implement a for loop as the same length of test size. Then at each step a new model will be created, and it will predict the future.

**Step 2**: In the next step, the previously forecasted information will be added to the train dataset; by doing that all the time train dataset will have incorporated new information. And the new model will have past performance knowledge.

**Step 3**: Thus, the forecast can be improved.

*Table 4.4 Forecast improvement using Walk Forward validation and decreasing RMSE value.*

| Dataset | Previous RMSE value | New RMSE value | Forecasting period |
|---|---|---|---|
| **Stock Data** | 0.571 | 0.520 | 1-week |
| **Temp Data** | 1.220 | 1.200 | 1-week |

The idea behind implementing Walk Forward validation was to improve the forecast accuracy. The new RMSE values in table 4.4, illustrates that error rate is decreased once the Walk Forward validation has been implemented. Well, the difference is not that huge, but in stock market sometimes it can make a huge difference.

## 4.2 Moving Average (MA) Model

The MA model uses past forecast errors in a regression type model. Sometimes, regression like models do not fully manage to capture the patterns in the time series, there might have some information left which can be beneficial; these errors or values called residuals. So, in order to have correct future values, we will apply model on residuals and forecast them. Later, they will be added to actual forecast.

Generally, an MA model can be denoted as MA (q), a moving average model of order q. mathematically, an MA model can be formulated as (4.4). Where, $\theta$ is the parameter, changing $\theta$ results in different time series patterns. $\varepsilon$ represents the error term for used past observations.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{4.4}$$

- **Steps for fitting MA model on time series**:

**Step 1**: The first steps to forecast with an MA model is to first predict the values using AR model. Start fitting AR () model on time series and calculate AIC (Information Criterion); which can be used to compare the performance of ARIMA based models.

**Step 2**: The prediction made using AR model can be denoted as ŷ, while the actual values can be denoted as y. The difference between ŷ and y is called the residuals or forecasting error.

**Step 3**: The residual might have some information left which can be predicted by applying forecasting model on residuals. In result, the future residual will be predicted. In order to predict future residuals, an AR model should also be applied on residuals.

**Step 4**: The forecasted residuals then will be added to original forecast. Hence, the final forecast can be better predicted by analyzing the error term.

The overall procedure of fitting an MA model can be divided in to two parts; the initial steps include to build a prediction model and to generate a forecast, and the next steps are about finding residuals from the forecast and then to add them to the original forecast to improve accuracy of the forecast. The Moving Average method is often used along with Auto

Regression method or with Naïve Persistence model, where AR method is used as the first level prediction model and then MA is used to update the initial predictions.

### 4.2.1 Fitting an MA model on Stock data

Stock dataset is nonstationary dataset. And as discussed above, fitting an MA model and forecasting with MA model is a two-step approach. Where first an AR model is fitted on the actual value and an AR model is applied on residuals. While forecasting with MA model, it is assumed that residual (error) is a linear combination of error terms. For Stock dataset, at first level the Naïve Persistence model was used, which assumes that future values are calculated on the assumption that conditions remain unchanged between current time and future time.

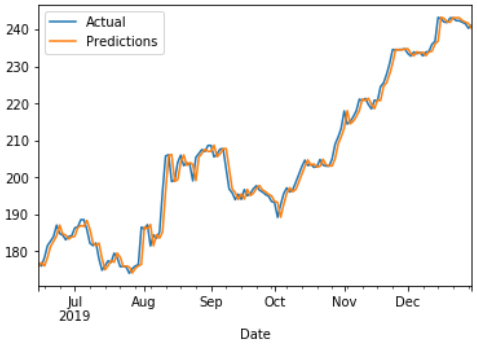
Once the Naïve Persistence model was fitted, residuals are calculated. To calculate the residuals, we need to shift the actual values by 1 time period ahead. Later, the difference between shifted values and actual values are called residuals. The next followed step was to apply model on residuals and calculate predictions for residuals. At the end, predicted residuals are added to the actual forecasted values. Hence, the forecast can be improved.

*Table 4.5 RMSE value based on different forecasting period*

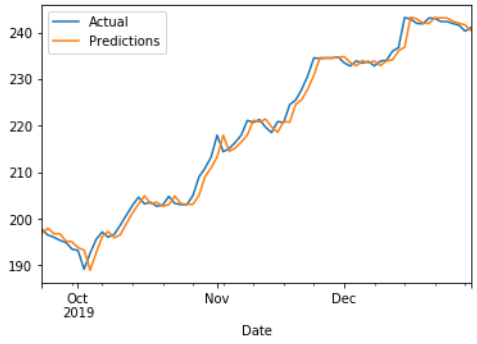
| Forecasting Period | Train | Test | RMSE |
|---|---|---|---|
| **6 months** | 1681 | 144 | 2.609 |
| **3 months** | 1753 | 72 | 1.893 |
| **1 months** | 1801 | 24 | 1.575 |
| **1 week** | 1818 | 7 | 0.641 |

From Table 4.5, the obtained RMSE values can be seen for each different forecasting period. It is necessary to assume that the length of test dataset is equal to the defined forecasting period. The forecast accuracy is improved than what was observed with AR model. Because one more step of predicting residuals is implemented. Even for a longer duration forecasting period, the obtained RMSE value was 2.6. Which means Moving Average model has performed better than Autoregression model.

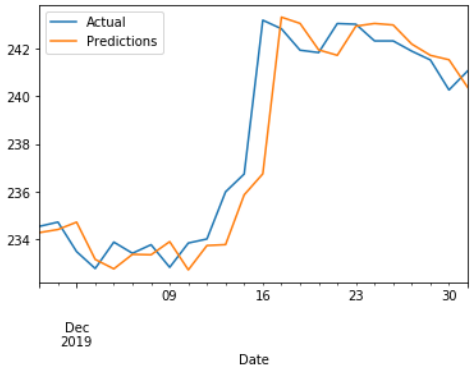
When observing difference between actual values and predicted values from plotted graphs in Figure 4.3, the MA model has predicted values nearly similar to the actual values. In comparison with Autoregression model, the MA model has performed better. Sometimes, it is necessary to observe the residuals because they might contain some important information which could change the forecast drastically. As we already discussed that Autoregression model does not do good job when a time series has trend or seasonality. So, very often after fitting an AR model, there are residuals left. So, applying this moving average model would help in identify patterns in time series.



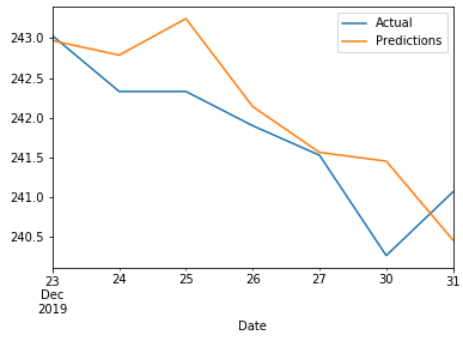

*(A) 6 months forecasting period*




*(B) 3 months forecasting period*




*(C) 1 month forecasting period*




*(D) 1 week forecasting period*

*Figure 4.3 Forecasting with different forecasting time periods*

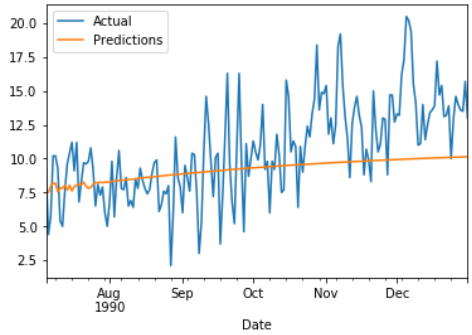
### 4.2.2 Fitting an MA model on Temp data

The Temp dataset is a stationary dataset. In order to forecast or predict the performance, the dataset was divided into different train and test sizes that too in accordance with the forecasting period. For Temperature dataset, at first level an AR model was used, then after calculating residuals, an AR model was applied on residuals as well. The according to the above-mentioned steps, predicted residuals by AR model was added to the actual forecast.

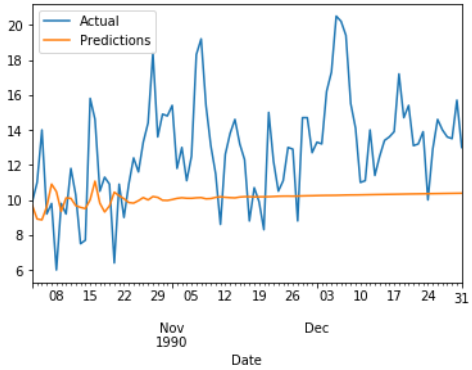
*Table 4.6 RMSE values based on forecasting period*

| Forecasting Period | Train | Test | RMSE |
|---|---|---|---|
| **6 months** | 3472 | 180 | 3.406 |
| **3 months** | 3562 | 90 | 3.925 |
| **1 months** | 3622 | 30 | 3.340 |
| **1 week** | 3645 | 7 | 1.049 |

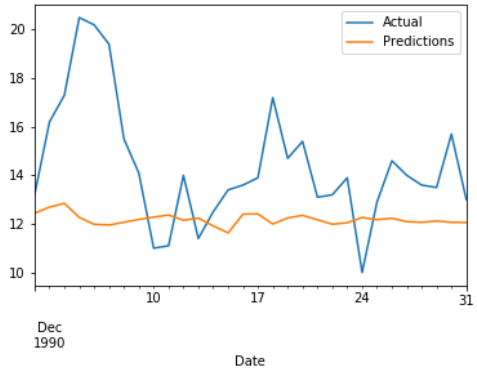
To evaluate the performance of an MA model, RMSE value was calculated for each train and test size. The lowest value was calculated 1.049 for 1 week forecasting period. While, on the other hand, the highest value was calculated 3.925 for 3 months forecasting period. Generally, the performance of AR and MA is same for stationary dataset. As it can be seen from the Table 4.6 that RMSE values lie around 3 and for AR model the same results were observed.



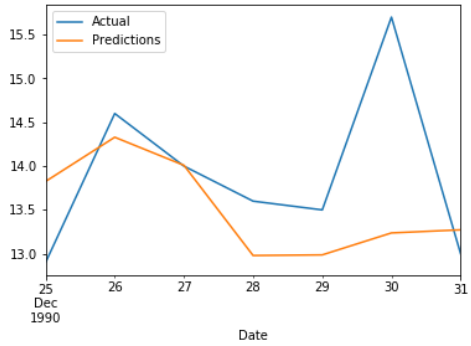

*(A) 6 months forecasting period*




*(B) 3 months forecasting period*




*(C) 1 month forecasting period*




*(D) 1 week forecasting period*

*Figure 4.4 Forecasting with different forecasting time period*

From the Figure 4.4, it is visible that MA model does not capture the cyclic behaviour of Temp data but it predicts the values between the thresholds set by the actual values. The blue line represents the Actual test data values, whereas the orange line depicts the predicted values. Hence, it can be said that an MA model is a good model when a time series does not have frequent fluctuations.

## 4.3 Forecasting using Autoregressive Integrated Moving Average (ARIMA)

### 4.3.1 Introduction to ARIMA models

ARIMA (Auto-regressive Integrated Moving Average) models are, in theory, the most general class for forecasting a time series. ARIMA is a combination of AR (Auto regression) model and MA (Moving-Average) with Integration (Differencing) which basically means how many times we had to difference the data to get it stationary. An ARIMA model has three orders: p, d and q. And it can be written as ARIMA (p, d, q) where p represents the AR model and q represents MA model.

$$y_t = c + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (4.5)$$

In general, an ARIMA model mathematically can be written according to Eq. (4.5), $y_t$ and $y_{t-1}$ represents values at current period and one period ago respectively. Similarly, $\varepsilon_t$ and $\varepsilon_{t-1}$ are the error terms for the same two period and $c$ is just a baseline constant factor. The

parameter $\phi_1$ expresses what part of the value last period is relevant in explaining the current one. And $\theta_1$ expresses what part of the error last period is relevant in explaining the current value.

There are many available variants of ARIMA model which can be used for different purpose, such as: Non-seasonal ARIMA, which is generally denoted as ARIMA (p, d, q) where p represents the order of autoregressive model, d is the degree of differencing and q is the order of moving average model. Seasonal ARIMA (SARIMA) model is usually denoted with ARIMA (p, d, q) (P, D, Q) m, where m is the period of seasonality. P, D, Q will be telling the impact of seasonality on auto regression, integration and moving average.

- **Steps for fitting an ARIMA model on time series**:

**Step 1**: First it is necessary to visualize the time series data, we also need to make sure that given data set is stationary.

**Step 2**: If the dataset is not stationary, then it must be converted to stationary time series data. We can perform operations such as; taking Log, Detrending and Differencing to remove trend and seasonality.

**Step 3**: There are two ways to choose the optimal value for p, d and q. Which can be done by applying AUTO_ARIMA function and either it can be done by analyzing ACF and PACF plots.

**Step 4**: The remaining steps are same as models such as AR and MA. We need to estimate ARIMA model which gives lowest AIC value and further the forecasting must be built on that model.

**Step 5**: The model with least AIC value would calculate least RMSE value and that is the optimal model.

### 4.3.2 Fitting an ARIMA model on Stock data

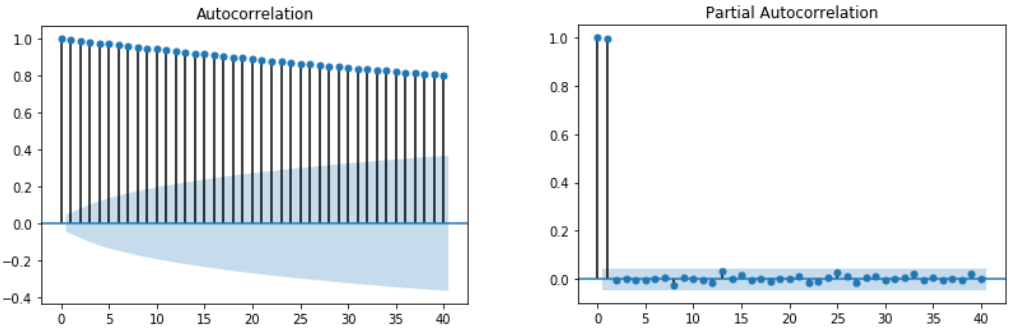
The Stock data has no clear seasonality, but it has increasing trend over time. In order to have correct forecasting, a non-stationary time series must be converted to stationary before applying ARIMA model. In order to have stationary dataset, the differencing has been performed on the dataset and after that Dickey-Fuller test has been applied to check whether if there is still any seasonality or trend component remained in the dataset. In order to convert nonstationary time series into stationary differencing procedure is followed.

```
Augmented Dickey-Fuller Test:
ADF test statistic       -42.453975
p-value                    0.000000
# lags used                0.000000
# observations          1823.000000
critical value (1%)       -3.433942
critical value (5%)       -2.863127
critical value (10%)      -2.567615
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

*Figure 4.5 Confirming stationarity after differencing time series*

The differencing procedure shifts the values in the future by one step. Hence, if there is any sort of seasonality or trend then it will be removed. After applying differencing, the Dickey-Fuller test has been applied and p-value is calculated as 0.00, which confirms the stationarity, see Figure 4.5, which also confirms the order of differencing (d) = 1. If the dataset did not confirm the stationarity, then the differencing procedure should have been performed with increasing order of differencing until the dataset confirms the stationary characteristics.

The next step is to choose p and q parameters, here, the order of AR (p) and MA (q) were determined by visualizing ACF and PACF plots.




*(A) Autocorrelation Plot* *(B) Partial Autocorrelation Plot*

*Figure 4.6 Determining value of p and q by analyzing ACF and PACF plot*

According to (Nau, 2019) [4], If the autocorrelation plot shows positive autocorrelation at the first lag (lag-1), then it suggests to include the AR terms in relation to the lag. If the autocorrelation plot shows negative autocorrelation at the first lag, then it suggests including MA terms. Identification of an AR model is often best done with the PACF. Whereas, the order of an MA model is often best done with the ACF plot.

- To determine an AR term, it can be determined from PACF plot, if PACF plot produces sharp cutoff then the number of AR term would be the same where PACF cuts off. See Figure 4.6 (B), it can be easily seen that PACF cutoff sharply at lag = 2. Thus, the value of AR term (p) = 2 [4].
- For calculating an MA term, if ACF plot gradually decline then it can be said that a time series contains an AR signature. Hence an MA term would be between 1 and 2. We need

to apply model with changing values of an MA term and observe the AIC value, that which one gives least AIC value.

However, it can be very difficult to read these plots, so it is often more effective to perform a grid search across various combinations of p, d, q values. Because this is a very realistic approach. There is no point in adding in the human error aspect by trying to read these very complex plots and deciding what the p values and q values.

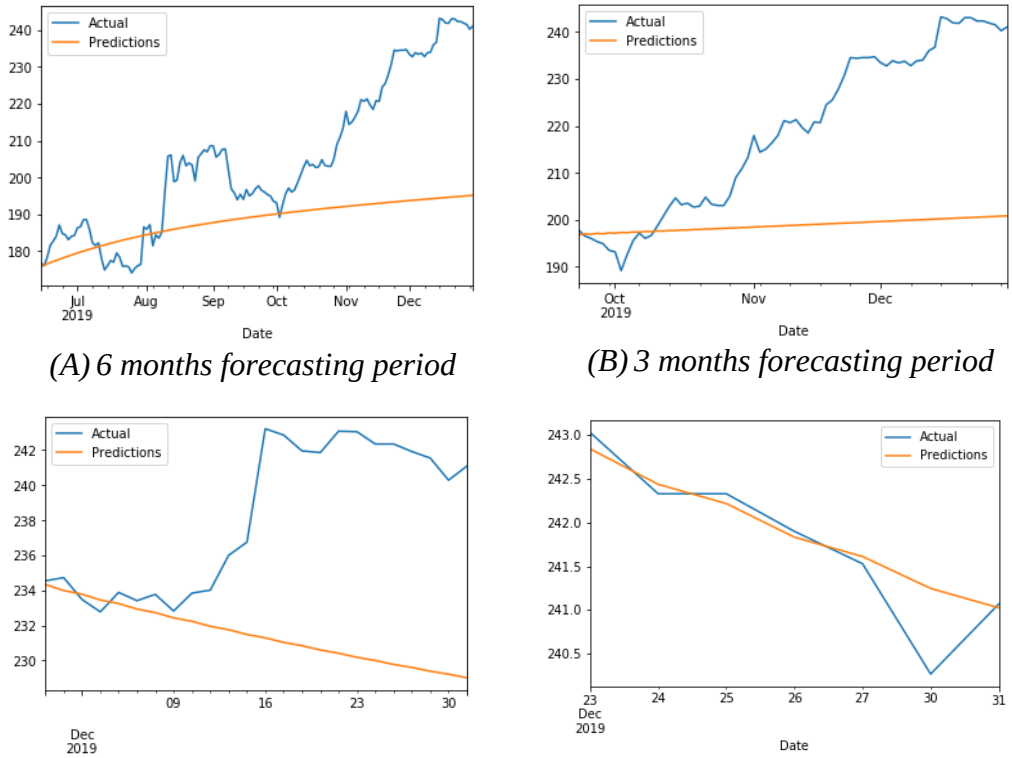
*Table 4.7 ARIMA model performance evaluation for different forecasting period on Stock data*

| Forecasting Period | Train | Test | ARIMA (p, d, q) | AIC | RMSE |
|---|---|---|---|---|---|
| **6 months** | 1681 | 144 | (1,1,1) | 7535.18 | 22.974 |
| **3 months** | 1753 | 72 | (2,1,2) | 7921.82 | 25.700 |
| **1 months** | 1801 | 24 | (3,1,1) | 8132.62 | 8.625 |
| **1 week** | 1818 | 7 | (2,1,2) | 8200.54 | 0.383 |

The Stock data having a clear upward trend and that depicts that it required a differencing before applying ARIMA model. The above Table 4.7 describes the results obtained after applying AUTO_ARIMA () function to the train dataset and the train dataset was trained to fit ARIMA model for different forecasting period. When the train dataset was longer in length, the greater number of lags are found to be correlated with past values. For example, for 1 month forecasting period, the train dataset length was 1801 observations and best filled ARIMA model was (3,1,1), meaning 3 terms of AR, and 1 term of MA. Whereas, the lowest RMSE was calculated for 1 week forecasting period and that was 0.383. While, the least value of AIC was calculated for 6 months forecasting period and that was around 7535, because that train and test size split uses ARIMA model with least number of AR and MA term. At the same time, the highest RMSE value calculated was 25.700 which is better than what was calculated by AR model.

The ARIMA model is a combination of AR and MA terms with d is for differencing. It can be observed from the table that after combining AR and MA, the number of lag observations in the past are decreased than what was considered in AR model. For the same length of dataset, in AR model, total number of lags in the past was AR (25), which is reduced to AR (3) in ARIMA model.

The forecasted values are plotted against the actual value and that can be seen in Figure 4.7, where ARIMA model does better work than an AR model. It has reduced RMSE values than RMSE values calculated by AR model. Still, the trend and seasonality are not clearly predicted by an ARIMA model. Hence, the models which are based on regression, moving average or any ARIMA based models are better at predicting values within the threshold of actual values. While they do not forecast according to the clear seasonality affects.




*(A) 6 months forecasting period*

*(B) 3 months forecasting period*

*(C) 1 month forecasting period*

*(D) 1 week forecasting period*

*Figure 4.7 Forecasting with different forecasting time period on Stock data*

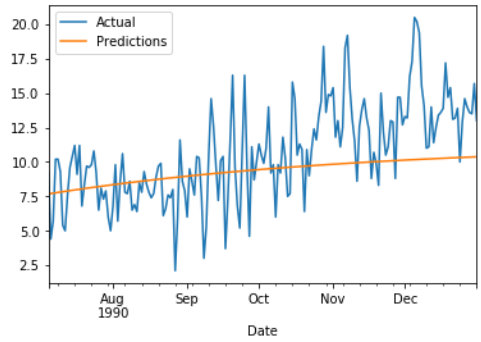
### 4.3.3 Fitting an ARIMA model on Temp data

ARIMA model is best fitted with stationary data. The Temperature dataset is stationary and an ARIMA model also does a good job predicting values. Which is giving maximum value of 3 for RMSE. To evaluate the forecasting performance the same procedure has been followed throughout the dissertation. Which is to divide a dataset into train and test split. For Temp dataset, the dataset was divided according to four different forecasting period. The lowest RMSE value observed for 1 week forecasting period with best fitted ARIMA model was (3,0,1) and value was 1.302. To notice, with stationary dataset, ARIMA model estimates almost same parameters for AR and MA term regardless of length of datasets.

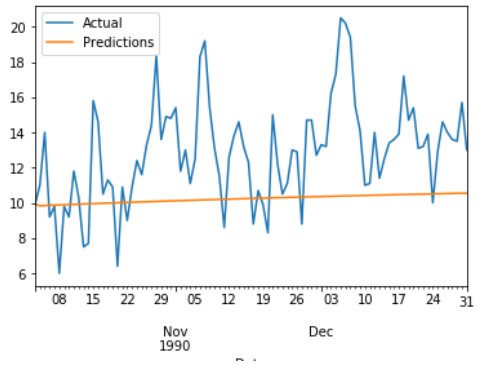
*Table 4.8 ARIMA model performance evaluation on different forecasting period on Temp data*

| Forecasting Period | Train | Test | ARIMA (p, d, q) | AIC | RMSE |
|---|---|---|---|---|---|
| **6 months** | 3472 | 180 | (3,0,1) | 15961.93 | 3.322 |
| **3 months** | 3562 | 90 | (1,0,3) | 16372.96 | 3.840 |
| **1 months** | 3622 | 30 | (3,0,1) | 16652.61 | 3.212 |
| **1 week** | 3645 | 7 | (3,0,1) | 16757.37 | 1.302 |

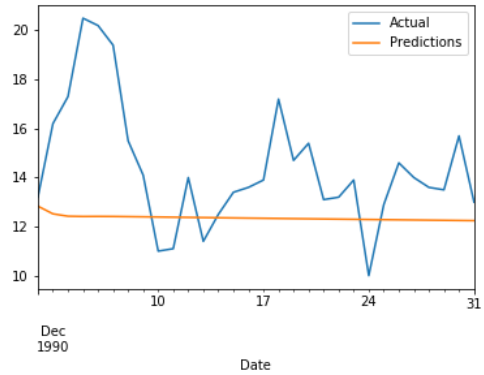
From figure 4.8, it can be seen how the predicted values are different than the actual values. The results are accurate than what was attained for Stock data. Temp data is stationary and compared to Stock data it has less variations.



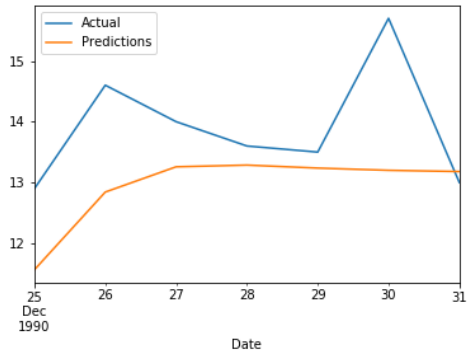

*(A) 6 months forecasting period*




*(B) 3 months forecasting period*




*(C) 1 month forecasting period*




*(D) 1 week forecasting period*

*Figure 4.8 Forecasting using ARIMA model with different forecasting period on Temp data*

## 4.4 Derived Conclusions after Fitting Autoregression and Moving Average Models

Autoregression model forecasts values based on the relationship between an observation and some number of lagged observations. Sometimes, it may happen that there is no strong relationship exists between past observations and current observations. Then, to fit an AR model might lead to inaccurate forecasts. AR model is a good choice when the correct forecasting decisions are made in accordance of the only past observations. But while predicting future forecasting with AR model would not give accurate forecasts-which is discussed further in Chapter 7.

On the other hand, MA model tries to predict forecasting based on errors and predict the errors which is so called ‘residuals’, which are left unattended after fitting a first level AR model. As this thesis discussed so far about the disadvantages of using an AR model, is that not efficiently capturing trend and seasonality. An MA model would consider as an extension of an AR model.

Combination of AR and MA model with differencing term makes a very powerful ARIMA model. Which is considered to give robust predictions irrespective of stationarity of a time series. Well, there are techniques available to convert nonstationary data into stationary data. Combining AR and MA term would reduce the mathematical complexity of individual model such as; AR and MA. Merging both together would reduce the number of lagged observations

Used in each model. For example, for Stock data, the best fitted AR model was with 25 lagged observations which can also be denoted as AR (25). Even though, approaching an MA model would still require using 25 lagged observations. But right after implementing ARIMA model which calculates both term AR and MA together. In result, the lagged value was decreased to 2.

ARIMA model forecasts future values based on past observations and error terms. The probability that it would perform better than other models in the future are higher. But it is understandable to wisely choose a right forecasting model. If a time series has a linear trend and no seasonality, then any of the ARIMA based models would perform better. Otherwise, it would require to some procedures such as; de-trending or differencing.

The nature of the time series is the most important factors in building the forecast. In our example, all three models have performed equally for Temp data, which is a stationary time series. On the other hand, we observed some biased values while forecasting with Stock data, because of its frequent changes throughout the time series and a clear upward trend. That is why to discuss important of Statistical tests, that is why this dissertation structured Chapter 3 prior to other chapters in the thesis. Because the importance of understanding a time series and results obtained from the statistical tests are a key foundation of accurate and correct forecast. The efficiency of forecasting in future using ARIMA based models are discussed in Chapter 7.

# Chapter 5 Forecasting using Smoothing Methods

## 5.1 Introduction to Exponential Smoothing Methods

Exponential Smoothing is a generalized forecasting technique which belong to ETS (Error-Trend-Seasonality) models. These models primarily focus on main components such as trend and seasonality. Basically, these models take each of these terms for smoothing and may add them, multiply them, or even just leave some of them out. Smoothing methods often required to have identified trend and seasonality in the dataset. Therefore, moving average smoothing methods such as: simple moving average and weighted moving method are used to smooth a time series by averaging (with or without weights) for a fixed number of continual terms. Over time, in a simple moving average all the observations are allotted with the same weight in averaging. While in a weighted moving average method it is possible that all the observations are assigned with different weights in averaging. The most recent observations are often given higher weights in comparison with older observations where weights lower.

The assigning of different weights can be explained by the fact that recent data are more important for a forecasting new data than old observations. Exponential Smoothing is an important quantitative forecasting technique. The weights are decreasing exponentially as the observations gets older. Meaning that recent observations are given higher weights as they contribute significantly in the future forecasting. The smoothing methods models three aspects of time series: the trend, seasonality and trend slope. These three ways to model a time series give rise to three types of exponential smoothing: single or simple exponential smoothing, double exponential smoothing, and tripe exponential smoothing also called as the Holt-Winters method.

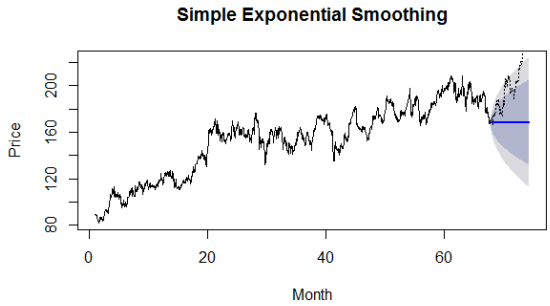
The smoothing techniques generally comprises of three smoothing equations. One for the level ℓ_t, one for the trend b_t, and one for the seasonal component s_t, with corresponding smoothing parameters α, β and γ. The nature of the seasonal component depends on two different variations. The additive method is used when the seasonal variations are nearly constant through time series, while the multiplicative method is applied when the seasonal variations are changing proportional to the level of the series.

## 5.2 Simple Exponential Smoothing (SES)

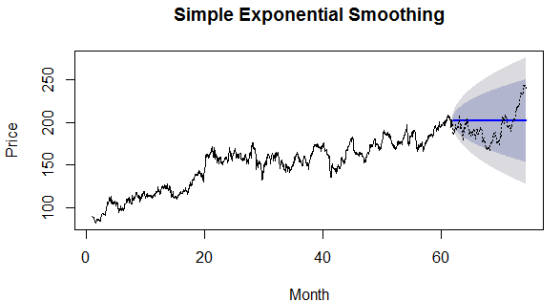
Simple exponential smoothing forecasts future values by using a weighted average of all the previous values in the series. It is best suited for a short-term forecasting, and usually best fit when the forecast horizon is shorter in length like a week or a month. Simple exponential smoothing is also known as exponential weighted moving average (EWMA) model. Simple exponential smoothing assumes that a time series does not contain trend or seasonality. The most important thing about simple exponential smoothing is that there is only one smoothing parameter α.

$$y_t = \alpha x_t + (1-\alpha) y_{t-1} \qquad 0 \leq \alpha \leq 1 \tag{5.1}$$

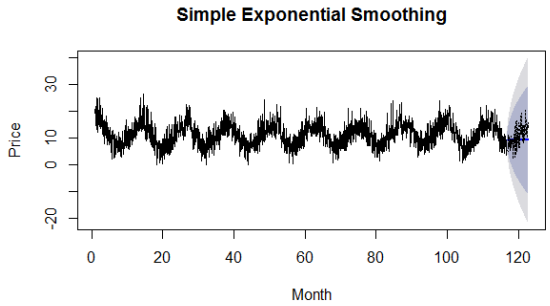
Simple Exponential Smoothing only requires a single smoothing parameter, called *alpha* (α). The α parameter controls the rate at which the observations influence the forecast. SES assumes that a time series only contains level ($L_t$), no trend or seasonality. The assumptions of the Exponential smoother is that this level will stay put and not move. Therefore, the k step ahead SES forecast is simply the most recent estimate of level at time t. The (5.1) illustrates how to estimate the level, which is also called the Level updating formula. The level at time t ($y_t$) and updating the previous level at time t-1 by integrating information from our most recent data point $x_t$. It is also like weighted average where α and (1- α) are considered as weights. In order to start the forecasting procedure, it is mandatory to initialize $L_1 = Y_1$. So, it starts by setting $L_1$ equals to the first record.



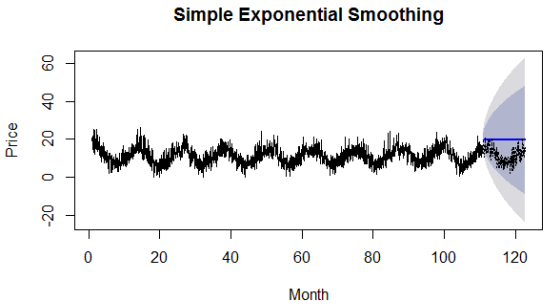

*(A) 6 months forecasting period Stock data*




*(B) 1-year forecasting period Stock data*




*(C) 6 months forecasting period Temp data*




*(D)1-year forecasting period Temp data*

*Figure 5.1 Simple Exponential Smoothing forecasting on both datasets with different time period*

The value of α lies between 0 and 1. The closer the smoothing parameter α is to 1, the more the model relies on the latest observations. A value close to 0, will produce a smoother model even the old data being taken into consideration. In order to perform Simple Exponential Smoothing, the both datasets are divided into train and test sizes. The forecasting was obtained for 6 months and 1-year time period for both the datasets. From Figure 5.1, it can be clearly seen how the model has predicted values compare to actual test values. As the model does not consider trend and seasonality, it predicts the almost same values for all the observations. The forecasting values stay within same range and do not replicate more changes. When α=1, the past values have no influence on the algorithm and the level just

remains the way it started out. Whereas, α=0, meaning past values have influence on the algorithm.

The R studio has in built function called SES (), which is used to forecast using Simple Exponential Smoothing. Where the value of α is auto generated, and it could also be assigned manually. The function calculates RMSE and AIC value as well.

*Table 5.1 Simple Exponential Smoothing forecasting performance evaluation*

| **Dataset** | **Forecasting Period** | **Smoothing Parameter: $\alpha$** | **AIC** | **RMSE** |
|---|---|---|---|---|
| **Stock Data** | 6 months | 0.979 | 14569 | 2.323 |
| | 1 year | 0.968 | 13038 | 2.255 |
| **Temperature Data** | 6 months | 0.452 | 35040 | 2.610 |
| | 1 year | 0.447 | 33082 | 2.622 |

For the first Stock data, α is calculated nearly equal to 1. Which means that recent observations have the strong influence on the algorithm and strong effects on forecast values. While, for the second dataset of Temperature data α is calculated equal to 0.4, meaning it assigns equal weights to both past and recent observations. Simple Exponential Smoothing is not affected by the length of the dataset and calculates the same RMSE values for each size of datasets. Simple Exponential Smoothing fails to account for other contributing factors like trend and seasonality. Hence, predicting future forecasting with Simple Exponential Smoothing is not efficient. On the other hand, it can be used to identify trend and seasonality in the datasets.

## 5.3 Double Exponential Smoothing (Holt's Method)

Double Exponential Smoothing also known as Holt's Method, is an extension of Simple Exponential Smoothing by allowing a new parameter $\beta$ which is responsible for identifying trend in a time series. It predicts trend very well. The equation (5.2) calculates the level $l_t$ at time t, and $b_t$ which is used to define general trend. Later, both can be combined to (5.4) predict some period ahead, where h denoted the number of periods into the future.

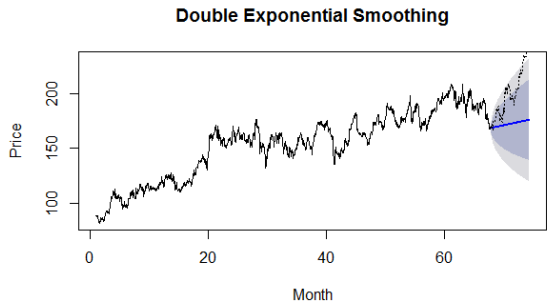
$$l_t = (1-\alpha)l_{t-1} + \alpha x_t \tag{5.2}$$

$$b_t = (1-\beta)b_{t-1} + \beta(l_t - l_{t-1}) \tag{5.3}$$

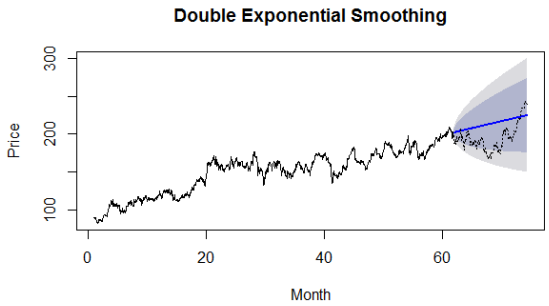
$$y_t = l_t + \mathrm{h}b_t \tag{5.4}$$

So, the value of $y_t$ at any given time t, is going to be equal to $l_t$+$b_t$. The proximity of recognizing trend in the datasets are clearly visible in the Figure 5.2, where a blue line which indicates forecasted values, clearly represents upward trend which was not resulted by Simple Exponential Smoothing. In result, the Stock data which has an upward trend, yields better results with Double Exponential Smoothing. Whereas, the second dataset of Temperature

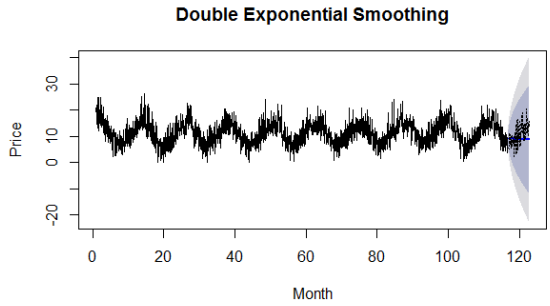
data, is a stationary dataset, with no trend and no seasonality. So, the forecasted values do not change significantly.



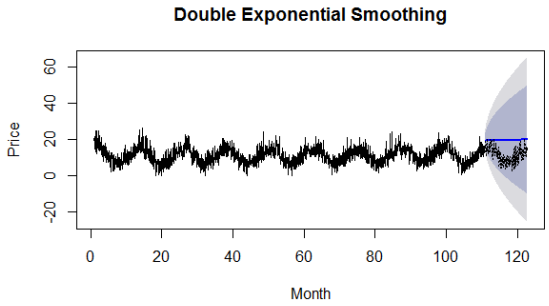

*(A) 6 months forecasting period Stock data*




*(B) 1-year forecasting period Stock data*




*(C) 6 months forecasting period Temp data*




*(D) 1-year forecasting period Temp data*

*Figure 5.2 Double Exponential Smoothing forecasting on both datasets with different time period*

The Double Exponential Smoothing gives very good results with a time series which has trend and does not have seasonality. Because $\beta$ parameter explicitly adds support for trends in the univariate time series. The method supports trends that change in different ways: either an additive or a multiplicative.

*Table 5.2 Double Exponential Smoothing forecasting performance evaluation*

| Dataset | Forecasting Period | Smoothing Parameter: $\alpha$ | Smoothing Parameter: $\beta$ | AIC | RMSE |
|---|---|---|---|---|---|
| **Stock Data** | 6 months | 0.975 | 1e-04 | 14572 | 2.323 |
| | 1 year | 0.967 | 1e-04 | 13041 | 2.254 |
| **Temperature Data** | 6 months | 0.452 | 1e-04 | 35045 | 2.610 |
| | 1 year | 0.448 | 1e-04 | 33087 | 2.622 |

Hence, from the above table 10, the value of smoothing parameters $\alpha$ and $\beta$ are clearly derived. Where value of alpha is calculated same as what was calculated by Simple Exponential Smoothing. Whereas, value of beta is calculated as 1e-04 that is equal to 0.0001. The RMSE value also remains within the same range as of Simple Exponential Smoothing. To conclude, Double Exponential Smoothing is a good approach when a time series does not have clear seasonality and has trend. But it does not capture any kind of seasonality. Hence, there is a need to use classic Exponential Smoothing method.

## 5.4 Exponential Smoothing (Holt-Winters Method)

Exponential Smoothing is a classic smoothing method that is a combination of all the above discussed smoothing methods. R studio provides function ETS () which is an autogenerated function which helps in forecasting. Exponential Smoothing is also known as Holt-Winters method. It is an extension of Double Exponential Smoothing method. With Exponential Smoothing, it is also possible to capture seasonality in a time series. This method an explicit smoothing parameter called gamma ($\gamma$) that addresses changes due to seasonality.

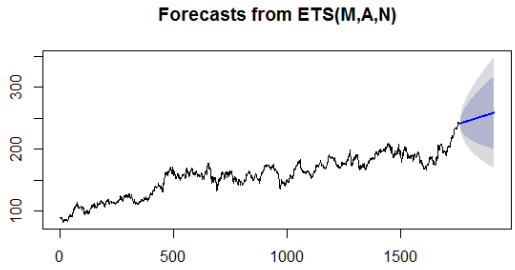
$$l_t = (1-\alpha)l_{t-1} + \alpha x_t \tag{5.5}$$

$$b_t = (1-\beta)b_{t-1} + \beta(l_t - l_{t-1}) \tag{5.6}$$

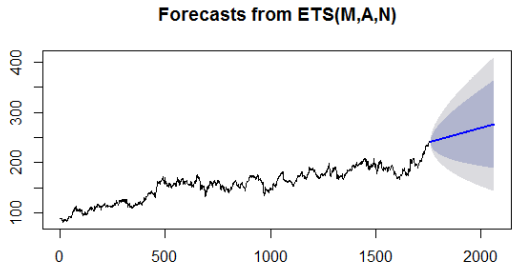
$$c_t = (1-\gamma)c_{t-L} + \gamma(x_t - l_{t-1} - b_{t-1}) \tag{5.7}$$

$$y_t = (l_t + b_t)c_t \tag{5.8}$$

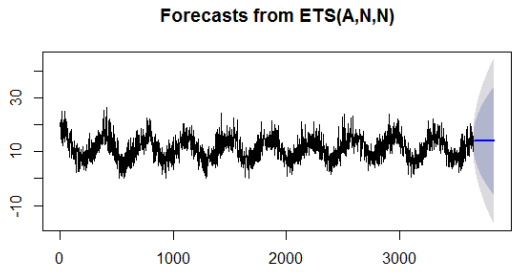
The Exponential Smoothing method contains three different components: Level ($l_t$), Trend ($b_t$) and Seasonality ($c_t$). The $c_t$ represents the seasonal factor with smoothing parameter $\gamma$ (5.7). Exponential Smoothing is mainly used for forecasting future values. While Simple and Double Exponential Smoothing methods are used for identifying trend.



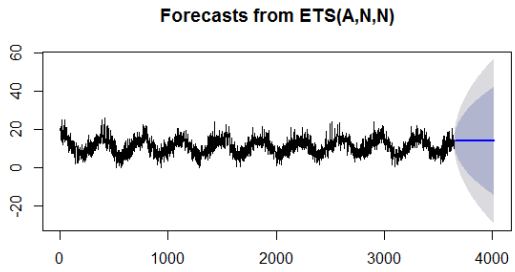

*(A) 6 months future forecasting Stock data*




*(B) 1-year future forecasting Stock data*




*(C) 6 months future forecasting Temp data*




*(D) 1-year future forecasting Temp data*

*Figure 5.3 Exponential Smoothing future forecasting on both datasets with different time period*

The results obtained after fitting Exponential Smoothing method to both datasets are seen from the Figure 5.3. As mentioned above the Exponential Smoothing is used for future forecasting for different time period of 6 months and 1-year. For both datasets, a model was fitted to forecast future values. Where the value of $\alpha$ calculated for Stock data is 0.97 and value of $\beta$ is 1e-04. And for Temp data, the value of $\alpha$ is 0.43 and $\beta$ is 1e-04. For both datasets, the model either neglected or did not calculate value of $\gamma$, because any of the datasets does not contain clear seasonality. So, in our case, the forecasted values also lie within the same range of what was calculated by Double Exponential Smoothing as no seasonality was found in the datasets.

## 5.5 Derived Conclusions after Fitting Smoothing Methods

Smoothing techniques are basically preferred for time series which do not radiate clear patterns, after applying smoothing methods, we could undoubtedly identify trend and seasonality present in the time series. So, smoothing methods would give more accurate results if it was used to smooth out the data rather than for forecast. Methods like Simple Exponential Smoothing and Double Exponential Smoothing are restricted to only selected nature of time series, whereas, classic exponential smoothing could be applied to any sort of time series.

The advantages of using smoothing methods includes; it's very simple in concept and very easy to understand. Also, it is possible to manually adjust the weights of the parameters, if the correlation between future values and observations was known. In addition to the advantages, the capability offered by smoothing methods of predicting trend in a time series is exceptional. For example, a classic exponential smoothing and double exponential smoothing methods have predicted increasing trend in the Stock dataset very well. Even, the future forecasting which has been predicted by classic exponential smoothing is also relevant to the actual trend of stock prices. Which indicates the increasing business demand in the future.

On the other hand, smoothing methods do not significantly capture the seasonality present in a time series. So, it would not efficient to fit a time series with random seasonality. So, from the results it can be said that simple and double exponential smoothing should not be used for future forecasting, but they can be applied to smooth out a time series. While classic exponential smoothing which provides flexibility to also capture seasonality can be fitted to predict future values.

# Chapter 6 Forecasting using Prophet

## 6.1 Data Modeling with Prophet

Forecasting is a typical data science task that helps in finding future demands, goal establishments and in outlier detection. However, there are some important issues associated with determining reliable and accurate forecast. To address these issues, the data science team of Facebook has introduced a forecasting procedure which is based on additive model. It allows to fit non-linear trends on yearly, weekly, and daily basis. It also incorporates adding holiday effects which might help business which generally experience distinct results during holidays such as; restaurants, amusement parks and museums. The advantage of using Prophet is that it is robust to missing data and changing shifts over the time series and provides strong results also while dealing with dataset which has outliers.

$$y_{(t)} = g(t) + s(t) + h(t) + \varepsilon_t \tag{6.1}$$

The mathematical formulation of Prophet can be demonstrated as, Eq. (6.1). Where $g(t)$ represents trend throughout the time series. $s(t)$ Describes the seasonal effects over the period. $h(t)$ Represents the holiday effects which might significantly affect the forecast. The error term $\varepsilon_t$ describes any changes and irregularities which can be considered as residuals.

Anomaly detection is also an important problem while forecasting for future. Prophet also helps in finding anomalies and experts can build the strategies on how to respond to them, once they are detected then that can be aggregated with forecasting results to accumulate changes. There are many different variants or units of forecasts that we want to make and there are associated problems with it such as; not many people are experts at forecasting, also not on demand availability of existing solutions or tools.

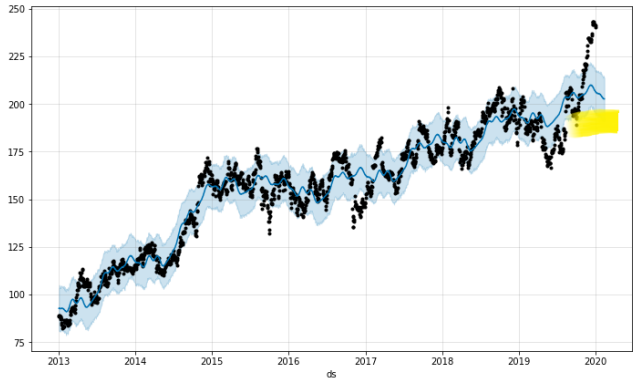
Prophet (Sean J Taylor, Benjamin Letham, 2017) [5] enables analysts with flexibility to do more forecasts. The normal modular regression model allows analysts to select the apparatuses which are appropriate to the forecasting problems and can easily make desirable modifications. There is one component which is accountable for determining and pursuing forecast accuracy and flagging forecast that should be checked manually to help build better results. Hence, critical services such as; allowing experts to adjust the model when required and when a completely model may be suitable.

Prophet requires to have a time series column names to be replaced with ‘ds’ and ‘y’. The working procedure of Prophet is ‘black-box’, which means one cannot go into details and understand how Prophet has forecasted the values.

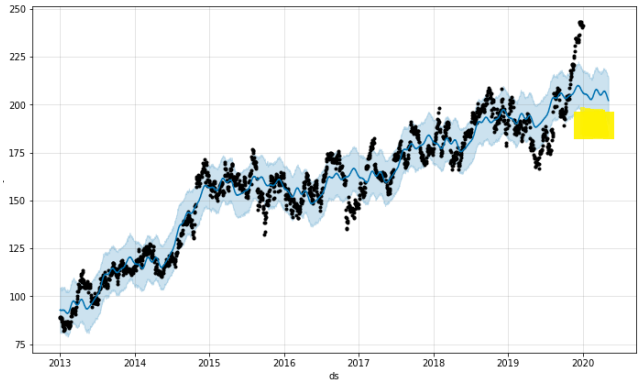
## 6.2 Fitting Prophet Model for Future Predictions

Prophet is one of the best models to forecast the future data. Because it offers a plenty of ways to enhance the performance of the model and predicts accurate future values. In this research, a Prophet model was fitted on both datasets to predict the future values in a different time

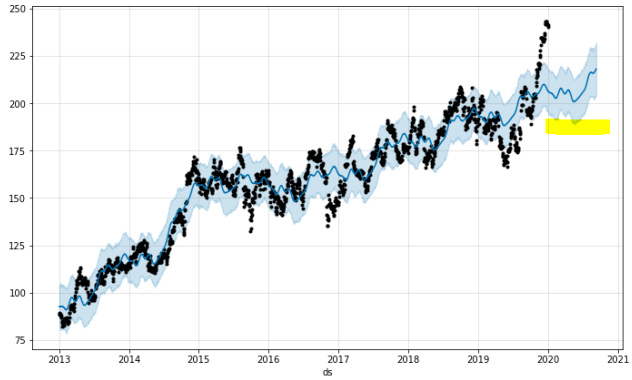
horizon. Firstly, a Prophet model was fitted on Stock data for different future forecasting time periods including: 1 month, 3 months, 6 months and 1 year in the future. Interestingly, Prophet has identified an increasing trend and forecasted future values accordingly. It can be seen clearly from the Figure 6.1 that how accurate predictions were made by Prophet. A yellow line represents the forecasted values. So, that is the strong benefit of using Prophet, its capability of capturing trend and seasonality is a way better than other models. Even for 1 year of future forecasting period, Prophet did a great job predicting an increasing trend in the Stock data.



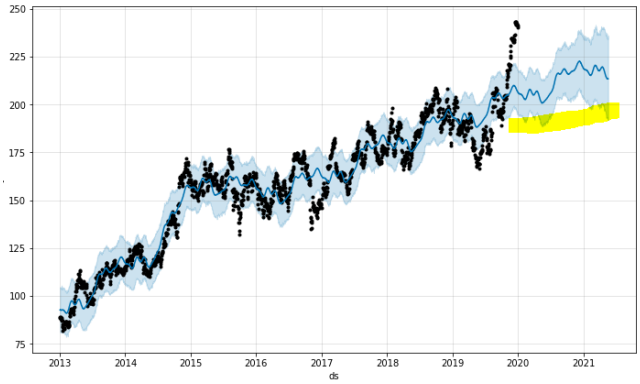

*(A) Stock prices predictions 1 month in future*




*(B) Stock prices predictions 3 months in future*




*(C) Stock prices predictions 6 months in future*

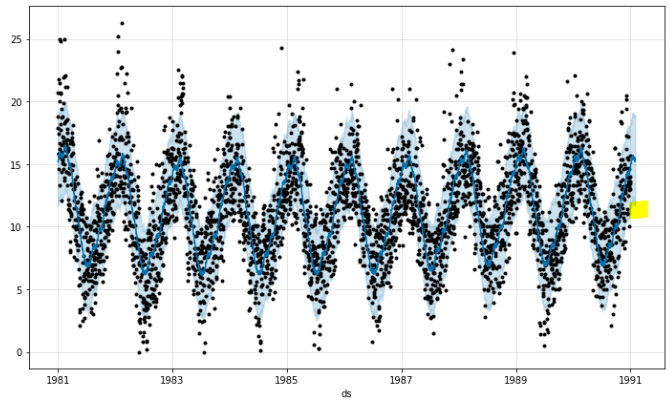



*(D) Stock prices predictions 1 year in future*

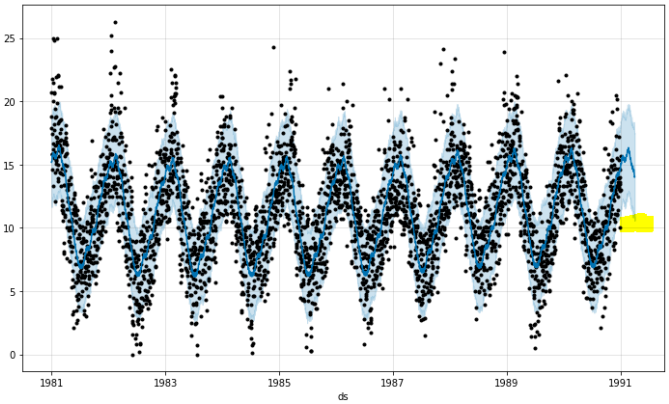
*Figure 6.1 Future Stock prices prediction using Prophet for different forecasting period*

Prophet also provides support to detect anomalies in a time series, with the help of experts, if anomalies were detected earlier, then while modeling that could be treated differently. Hence, the overall performance of Prophet could be improved. It is very difficult for other models to detect seasonality in the Stock data because it is not regular and clear. But, compare to other models Prophet did generate a forecast that is relevant to the actual values.

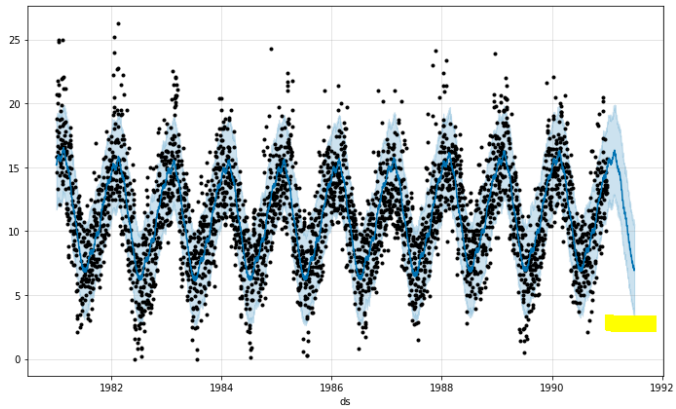
The same procedure is also applied on Temp dataset, which is a stationary dataset with no clear trend and seasonality. If Prophet did a great job with Stock data then without any doubt, it can be said that it must have performed exceptionally great with Temp data. As to capture trend and seasonality is easy with stationary time series.



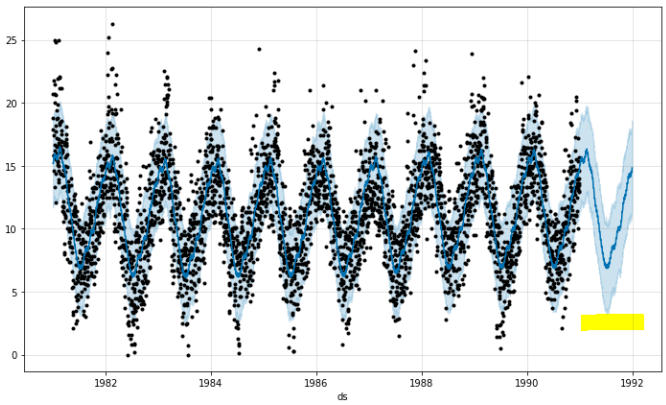

*(A) Temp. predictions 1 month in future*




*(B) Temp. predictions 3 months in future*




*(C) Temp. predictions 6 months in future*




*(D) Temp. predictions 1 year in future*

*Figure 6.2 Future Temp. predictions using Prophet for different forecasting period*

The pattern recognized by Prophet to predict the future values in the dataset is the same as what was followed throughout the time series. Figure 6.2 illustrates the forecasted values by Prophet on Temp data for different future forecasting time period. Prophet has efficiently handled the regular behaviour of the time series. As it is a temperature data, it follows yearly regular cycle, it is believed that in each year, the temperature for each individual month will remain in the same range. For Stock data, Prophet has measured increasing trend and for Temp data, Prophet has attained cyclic pattern and predicted accordingly. To summarize, Prophet is the best model for future forecasting and to determine the patterns within a time series. As we have seen from the obtained results of both datasets, that how accurately Prophet has determined patterns and predicted future values.

## 6.3 Prophet Forecasting Performance Evaluation

In order to make correct decisions, it is also important to evaluate the performance of the forecast made by Prophet. To achieve that, both datasets were divided in different train and test sizes. After that, a Prophet model was applied on train dataset and required to forecast values for the same time period as what was set to the size of the test dataset. Later, the

difference between predicted values and actual values were calculated and it is called RMSE (Root Mean Squared Error).

*Table 6.1 RMSE values according to the size of train and test split on Stock data*

| Train | Test | RMSE |
|---|---|---|
| 1057 (60%) | 705 (40%) | 14.056 |
| 1233 (70%) | 529 (30%) | 20.994 |
| 1409 (80%) | 353 (20%) | 14.731 |
| 1585 (90%) | 177 (10%) | 19.457 |

Table 6.1 illustrates the RMSE value for each train and test size split. The noticeable thing about Prophet is that the size of the dataset does not really matter when forecasting with Prophet. For example, the lowest RMSE value observed was 14.05 with 60% train and 40% test size. While for other models the lowest RMSE value was calculated while the test length was kept minimum in length (10%). For Stock data, RMSE values lie between 14 and 20, which is considered as huge difference while talking about stock market.

On the other hand, the same forecasting evaluation procedure was implemented on Temp data. Where dataset was split into train and test size according to the length of forecasting period. To measure the performance, a test dataset was kept the same in length as of the length of the forecasting period. Likewise, the lowest RMSE value is calculated for 1-year forecasting period, that again proved that Prophet does not take into consideration that what is the size of the dataset. Overall, Prophet has performed outstanding on Temp data.

*Table 6.2 RMSE values according to the size of forecasting period on Temp data*

| Forecasting Period | Train | Test | RMSE |
|---|---|---|---|
| **1 month** | 3622 | 30 | 2.589 |
| **3 months** | 3561 | 91 | 2.634 |
| **6 months** | 3470 | 182 | 2.598 |
| **1 year** | 3287 | 365 | 2.512 |

To summarize, above results obtained for both datasets clearly stated that Prophet is a very powerful tool to forecast the future prices. It identifies trend and seasonality in a time series very accurately. It also builds future predictions based on determined patterns. The disadvantage of using Prophet is that the way it builds the forecast is not available for understanding. So, this is the restrictions applied while using Prophet. For example, RMSE value obtained for Stock prices are very high, and one cannot try to implement procedures such as parameter fine tuning etc. to reduce the error rate. Thus, one can rely on the forecasted values by Prophet but do not find the procedure which sometimes can be an integral part of the forecasting process.

# Chapter 7 Discussion and Conclusions

## 7.1 Empirical Comparison of Models in Selected Applications

This research project, whose main goal was to analyze models for forecasting selected datasets in different business contexts, and to make experimental comparison between them. In order to achieve the mentioned goals, several models based on Autoregression, Moving Average, Smoothing methods and Prophet are implemented and analyzed. Chapter 4, 5 and 6 are about implementation and model fitting procedure of the models. Where, datasets are divided into different train and test sizes, in which a model is fitted on train size and performance is evaluated on test size. To evaluate the performance, RMSE (Root Mean Squared Error) and AIC (Information Criterion) metrices are used.

In order to make correct decisions, it is necessary to select an appropriate forecasting model. The future values are predicted based on several approaches; AR (Autoregression) allows to predict future values based on past observation, MA (Moving Average) gives flexibility to also add additional information from residuals which were left after applying AR model, ARIMA (Autoregressive Integrated Moving Average) model provides ability to combine AR and MA model together to have better predicted values. Whereas, forecasting with Smoothing techniques helps in assigning weight to recent observations which might have strong effects on future values. At the end, Prophet model has also fitted to predict the future values.

The datasets are chosen because of their different statistical properties; the Stock dataset is non-stationary, and values are fluctuating over time. Those fluctuations are very difficult to capture by a model. Hence, the RMSE values calculated by models on Stock data are higher than what was calculated for another dataset. The second dataset of Melbourne's daily temperature dataset is a stationary dataset, with constant fluctuations and without trend, which is very easy to model by models.

In order to analyze the performance of models, the models are used to predict the future values as well. And then the forecasted values are compared with actual values and RMSE values are derived. To forecast the future values, all models are applied on whole dataset and then they were required to predict values which were not known by them while trained. Furthermore, the comparison was also made based on how robust a model is while predicting with fluctuations and capability of capturing trend.

Firstly, for Stock data, the future values of "Close" prices are predicted for January 2020, having 21 business days. Thus, Table 7.1, with 6 columns, where column named Actual values represents actual values on that day. And other five columns are values predicted by the corresponding models. The actual values are derived from Yahoo Finance. To forecast in the future, datasets are not divided into train and test split. Instead, a model was fitted on whole dataset and asked to predict future values.

*Table 7.1 Comparison of forecasted 1-month future stock prices using all models(Prices are in US $)*

| Actual Values | AR (Autoregression) | MA (Moving Average ) | ARIMA | Exponential Smoothing | Prophet |
|---|---|---|---|---|---|
| **240.1** | 240.75 | 240.59 | 240.75 | 241.16 | 206.20 |
| **238.47** | 240.66 | 240.70 | 240.50 | 241.28 | 206.04 |
| **240.3** | 240.60 | 240.69 | 240.20 | 241.39 | 206.13 |
| **238.04** | 240.62 | 240.81 | 239.95 | 241.51 | 205.77 |
| **238.22** | 240.52 | 240.58 | 239.66 | 241.62 | 205.73 |
| **238.93** | 240.47 | 240.53 | 239.42 | 241.74 | 205.57 |
| **238.26** | 240.48 | 240.66 | 239.14 | 241.86 | 205.50 |
| **236.92** | 240.40 | 240.46 | 238.91 | 241.97 | 205.67 |
| **240** | 240.16 | 240.09 | 238.64 | 242.09 | 205.51 |
| **241.7** | 240.18 | 240.30 | 238.41 | 242.20 | 205.51 |
| **240.5** | 240.14 | 240.23 | 238.16 | 242.32 | 205.38 |
| **241.49** | 239.87 | 239.70 | 237.93 | 242.43 | 205.32 |
| **237.6** | 239.50 | 239.24 | 237.69 | 242.55 | 205.49 |
| **236.75** | 239.50 | 239.49 | 237.47 | 242.67 | 205.23 |
| **235.04** | 239.53 | 239.71 | 237.23 | 242.78 | 205.17 |
| **225.59** | 239.44 | 239.52 | 237.02 | 242.90 | 204.97 |
| **222.95** | 239.35 | 239.40 | 236.79 | 243.01 | 204.84 |
| **224.84** | 239.35 | 239.47 | 236.58 | 243.13 | 204.93 |
| **227.15** | 239.36 | 239.51 | 236.36 | 243.24 | 204.41 |
| **226.15** | 239.27 | 239.34 | 236.17 | 243.36 | 204.26 |
| **216.05** | 239.28 | 239.43 | 235.95 | 243.48 | 203.99 |

According to the procedures studied in the previous Chapters, future forecasting for one month has been generated by all the models and comparison was made against actual values. To evaluate the performance, RMSE (Root Mean Squared Error) metric has been used. According to Table 7.2, the lowest RMSE value (7.21) is calculated for ARIMA model, which is the best suited model. While the highest RMSE value was accounted for Prophet model that was around 29.9.

*Table 7.2 RMSE value for future forecasted Stock prices by all the models*

| Model | RMSE |
|---|---|
| **AR (Autoregression) AR (25)** | 8.743 |
| **MA (Moving Average)** | 8.806 |
| **ARIMA (2,1,1)** | 7.213 |
| **Smoothing** | 11.024 |
| **Prophet** | 29.981 |

In addition, it is also necessary to analyze the future predictions made by each model. From the Figure 7.1, it is illustrated that how each model has performed. Autoregression model and Moving Average model have predicted almost the same values, that too without considering the trend and seasonality. Still, the RMSE values calculated for both models are close to what

was calculated for ARIMA model. Prophet model is not derived as best forecasting model while forecasting in future for short time period. However, for long term forecasting, Prophet is one of the best forecasting models, which was discussed in Chapter 6. While Smoothing methods can also be reliable for short term forecasting.

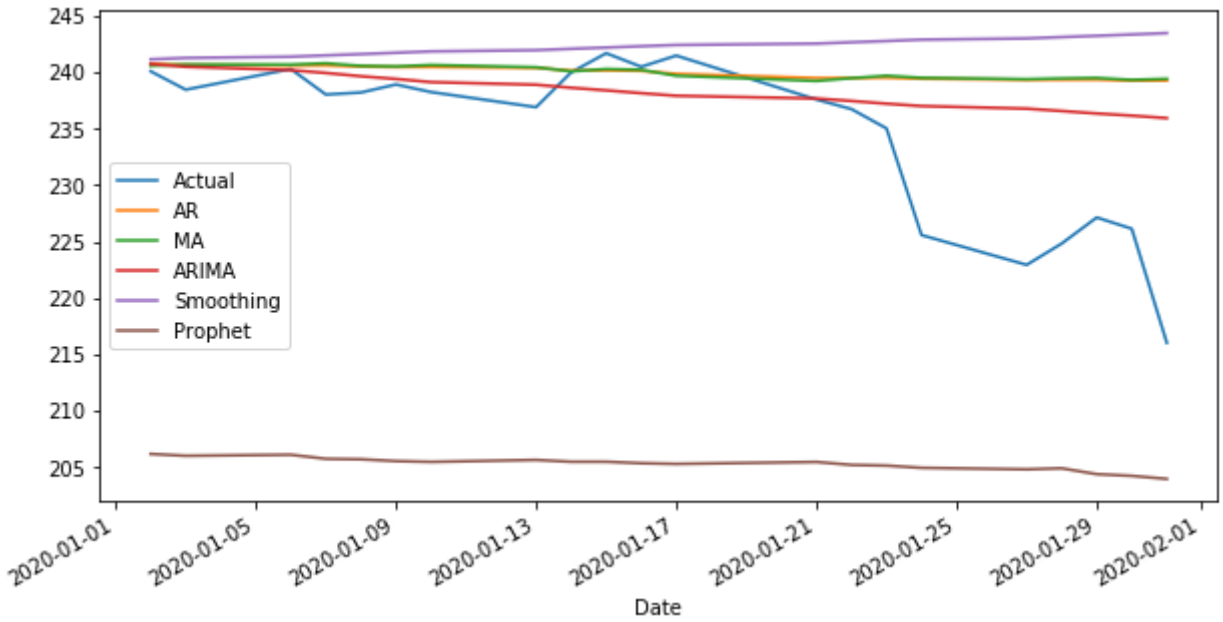



*Figure 7.1 Comparison between chosen models for future forecasting on Stock data*

Writing in accordance with the business context, Stock prices are fluctuating daily. It is difficult to determine the factors which affects the stock prices. With the help of experts with deep domain knowledge such forecasting models can be modified with exogenous factors contributing to the changes in price. Models such as ARIMA and Prophet allow to add external factors which influences the daily prices. However, it must only be performed under the guidance of experts. For example, Due to COVID-19, stock prices of Biotechnology Companies are expected to rise by 4% (Due to the increasing demand of medical appliances), whereas stock prices of Airlines Companies are expected to drop by 5% (Due to the travel restriction). One of the datasets used in this dissertation is Stock prices data of a Biotechnology Company (Amgen). The prices are expected to rise in the next upcoming month. So, observing such market scenarios help in building a better forecasting model.

On the other hand, the Temperature dataset, which is a stationary dataset, from the obtained results, there are not many contradictory or biased results discovered. From Table 7.4, by observing RMSE value calculated by all the models, it can be concluded that the RMSE value lies between 2 to 2.3. Which basically means, that stationary dataset has highest accuracy than non-stationary dataset. Even, the same behaviour was observed with each model throughout this dissertation, while Stock dataset has many biased results.

However, the graphs plotted for each model (Figure 7.2) might do not look similar to the actual values, but at the end, the future forecasted values are almost alike actual values. Therefore, we can rely on future forecasted values.

*Table 7.3 Comparison of forecasted 1-month future Temperature*

| Actual Value | AR (Autoregression) | MA (Moving Average ) | ARIMA | Exponential Smoothing | Prophet |
|---|---|---|---|---|---|
| **15.12** | 13.15 | 13.48 | 12.98 | 13.84 | 14.92 |
| **13** | 13.63 | 14.21 | 13.32 | 13.84 | 15.09 |
| **14.1** | 13.70 | 13.82 | 13.48 | 13.84 | 15.01 |
| **14.8** | 13.62 | 13.58 | 13.51 | 13.84 | 15.06 |
| **18.4** | 13.48 | 13.34 | 13.49 | 13.84 | 15.09 |
| **11.8** | 13.34 | 13.22 | 13.46 | 13.84 | 15.07 |
| **12.7** | 13.39 | 13.43 | 13.45 | 13.83 | 15.27 |
| **12.1** | 13.37 | 13.34 | 13.43 | 13.83 | 15.42 |
| **13.8** | 13.32 | 13.28 | 13.42 | 13.83 | 15.56 |
| **15.2** | 13.23 | 13.15 | 13.40 | 13.83 | 15.46 |
| **12.5** | 13.16 | 13.12 | 13.39 | 13.83 | 15.46 |
| **14.3** | 13.27 | 13.38 | 13.37 | 13.83 | 15.45 |
| **14.4** | 13.32 | 13.38 | 13.36 | 13.83 | 15.38 |
| **19.7** | 13.30 | 13.31 | 13.34 | 13.83 | 15.53 |
| **16.5** | 13.28 | 13.28 | 13.33 | 13.83 | 15.64 |
| **14.6** | 13.31 | 13.35 | 13.31 | 13.83 | 15.72 |
| **15** | 13.33 | 13.39 | 13.30 | 13.83 | 15.56 |
| **13.8** | 13.14 | 12.95 | 13.28 | 13.82 | 15.51 |
| **11.8** | 13.05 | 12.99 | 13.27 | 13.82 | 15.45 |
| **11.6** | 13.12 | 13.19 | 13.26 | 13.82 | 15.43 |
| **13.7** | 13.13 | 13.14 | 13.24 | 13.82 | 15.45 |
| **11.2** | 13.09 | 13.08 | 13.23 | 13.82 | 15.52 |
| **12.4** | 13.10 | 13.12 | 13.22 | 13.82 | 15.58 |
| **15.5** | 13.12 | 13.16 | 13.20 | 13.82 | 15.40 |
| **16.4** | 13.12 | 13.15 | 13.19 | 13.82 | 15.35 |
| **15.9** | 13.06 | 13.01 | 13.18 | 13.82 | 15.29 |
| **13.6** | 13.03 | 13.01 | 13.16 | 13.82 | 15.19 |
| **16.9** | 12.99 | 12.97 | 13.15 | 13.82 | 15.32 |
| **13.6** | 12.99 | 13.02 | 13.14 | 13.81 | 15.41 |
| **16.4** | 12.99 | 12.99 | 13.12 | 13.81 | 15.50 |
| **16.1** | 12.96 | 12.95 | 13.11 | 13.81 | 15.37 |

*Table 7.4 RMSE value for future forecasted Temperature values*

| Model | RMSE |
|---|---|
| **AR (Autoregression) AR (29)** | 2.325 |
| **MA (Moving Average)** | 2.332 |
| **ARIMA (3,0,1)** | 2.303 |
| **Smoothing** | 2.075 |
| **Prophet** | 2.216 |

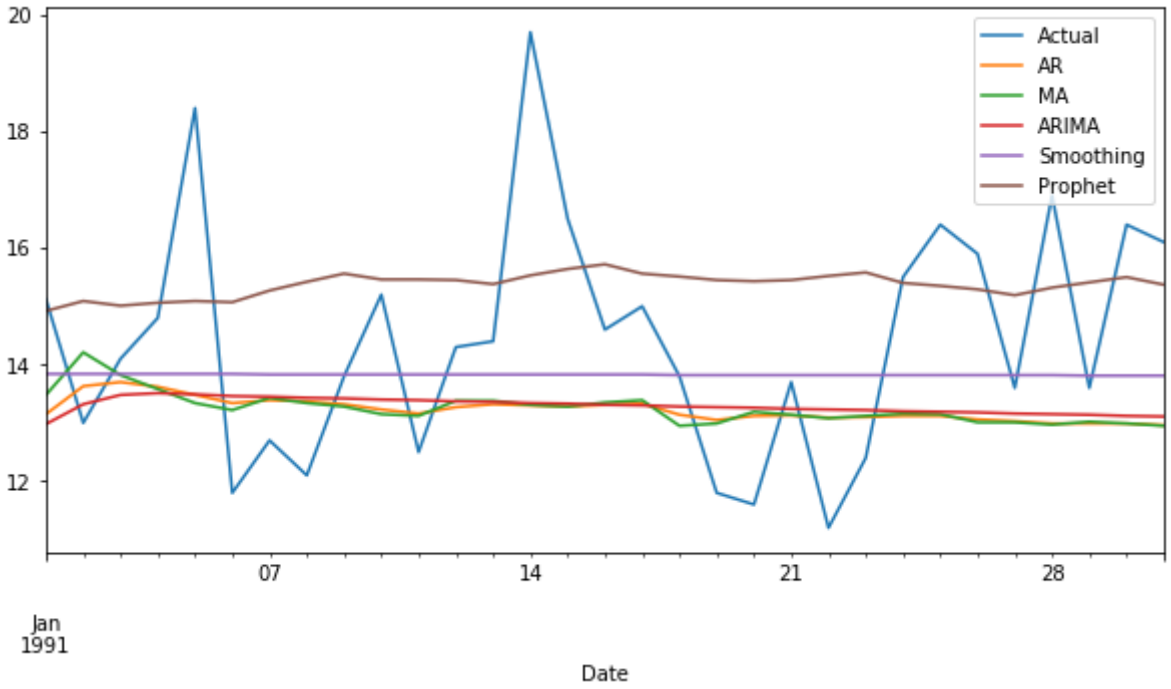



*Figure 7.2 Comparison between chosen models for future forecasting on Temp data*

*Table 7.5 Generalized Concluded analysis of all the models*

| **Model** | **AR (Autoregression)** | **MA (Moving Average)** | **ARIMA** | **Smoothing** | **Prophet** |
|---|---|---|---|---|---|
| **Stationary Time Series** | Performs good | Performs good | Performs Excellent | Performs good with Simple Expo. Smoothing | Stationarity of time series does not consider |
| **Non-stationary Time Series** | Does not perform good (Except clear trend and seasonality) | Does not perform good (Except clear trend and seasonality) | Performs good | Performs good with Expo. Smoothing | Stationarity of time series does not consider |
| **Capturing Trend** | Does not fit to the trend | Does not fit to the trend | Fits to the clear trend | Double Expo. Smoothing fits good | Performs Excellent |
| **Capturing Seasonality** | Does not fit to the seasonality | Does not fit to the trend | Fits to the regular seasonality | Only detects clear seasonality | Performs Excellent |
| **Efficiency for Short Term Forecast** | Performs good irrespective of stationarity of time series | Performs good irrespective of stationarity of time series | Performs Excellent irrespective of stationarity of time series | Performs good | Does not perform good |
| **Efficiency for Long Term Forecast** | Can not rely | Cannot rely | Performs good | Does not perform good | Performs Excellent |

As it was mentioned in the goal of this dissertation to make empirical comparison between all models, Table 7.5 describes the generalized way of performing by each model. The measurement of performance could be classified into three categories such as; performs good, performs excellent or does not perform good. These fields are derived from all the experiments performed during preparation of this dissertation and from the obtained results. Some models like Prophet and ARIMA are best choice for long term forecasting. While Autoregression and Moving Average based models are best fitted for short term forecasting. In addition, smoothing techniques can be used to smooth out the observations by assigning weight according to the influence of the observations.

To conclude, totally rely on results obtained from these statistical models are not always going to make right forecasts. But it helps in keep track of time series observations. As discussed earlier, other external variable might directly or indirectly affect the observations and changes. For example, global warming might affect the daily minimum temperature data. However, some models also give flexibility to add such exogenous variables while building forecasting model. Such models include; VAR (Vector Autoregression), SARIMAX (Seasonal Autoregressive Integrated Moving Average with exogenous). But it is necessary to perform such forecasting under the guidance of domain experts.

## REFERENCES


[1] Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting.* Hoboken, New Jersey: John Wiley & Sons.

[2] Senthamarai, K. K., Sailapathi, S. P., Mohamed, S. M., & A. P. (2010, March). Financial Stock Market Forecasting using Data Mining Techniques. *Proceedings of the International Multiconference of Engineers and Computer Scientists, I*, 4-5.

[3] Iwok, I. A., & Okoro, B. C. (2016, November). Forecasting Stocks with Multivariate Time Series Models. *International Journal of Mathematics and Statistics Invention, 4*(9), 12-13.

[4] Nau, R. (2019). *ARIMA model for time series forecasting*. Retrieved June 4, 2020, from Duke: https://people.duke.edu/~rnau/411arim3.htm

[5] Sean J Taylor, Benjamin Letham. (2017). Forecasting at Scale_Prophet. *PeerJ Preprints, 3190*(v2), 23-25.