A Major Project-II Report

On

TIME SERIES ANALYSIS AND PREDICTION

Submitted in Partial fulfilment of the Requirement for the Degree of

Master of Technology

in

Software Engineering

Submitted By

Ashish Kumar

2K17/SWE/06

Under the Guidance of

Mr Sanjay Kumar

Assistant Professor
(COE Department)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahabad Daulatpur, Main Bawana Road, Delhi-110042

June 2019

DECLARATION

I at this moment declare that the Major Project-II work entitled **"**TIME SERIES ANALYSIS AND PREDICTION**"** which is being submitted to Delhi Technological University, in partial fulfilment of requirements for the award of the degree of Master of Technology (Software Engineering) is a bona fide report of Major Project-II carried out by me. I have not submitted the matter embodied in this dissertation for the award of any other degree or diploma.

Place: Delhi   Name: Ashish Kumar

Date: Roll No: 2K17/SWE/06

CERTIFICATE

This is to certify that **Ashish Kumar** (2K17/SWE/06) have completed the major II project titled **“**TIME SERIES ANALYSIS AND PREDICTION**”** under my supervision in partial fulfilment of the master of technology degree in software engineering at Delhi Technological University.

PLACE: DELHI

DATE:

SUPERVISOR

Mr Sanjay Kumar

Assistant Professor

Delhi Technological University

Bawana Road, Delhi -110042

ACKNOWLEDGEMENT

First of all, I would like to express my deep sense of respect and gratitude to my project supervisor **Mr Sanjay Kumar** for providing the opportunity of carrying out this project and being the guiding force behind this work. I am deeply indebted to him for the support, advice and encouragement he provided without which the project could not have been a success.

Secondly, I am grateful to **Dr Rajni Jindal,** HOD, Computer Science & Engineering Department, DTU for her immense support. I would also like to acknowledge the Delhi Technological University library and staff for providing the right academic resources and environment for this work to be carried out.

Last but not least I would like to express sincere gratitude to my parents and friends for continually encouraging me during the completion of work.

Ashish Kumar

Roll No – 2K17/SWE/06

M. Tech (Software Engineering)

Delhi Technological University

ABSTRACT

Nowadays, in the world of digitalisation that promotes ease to do any task as well as in more inferior time. Also, by utilising the latest techniques, there is a provision to execute things accurately in the present by taking responsibility for the future. Various fields-maintained records, these records required to analysed timely, such time-series data set now analysing using various Machine learning models. When we have a time-series data set, the various models can be used for forecasting. Now, the challenge is to choose the most appropriate when we have non-constant and inconsistent data sets.

In this project, different kinds of time series prediction and forecasting models have applied on the same sets of data in order to identify the best. Also, before applying any model, test data-set for the time series components such as trend and seasonality. This project is done in several stages and also, removing the inconsistency present in the acquired data sets before applying.

The three data sets selected are agriculture value-added percentage of GDP is data set 1, India's monthly inflation rate is dataset 2 and Revenue generation of a restaurant is data set 3. The selected data sets are essential and secure relevance place in this analysis model.

For this purpose, chosen most generalised time series models Autoregressive integrated moving average (ARIMA), Exponential Smoothing Models, Holt Linear Trend model, Seasonal Decomposition, Cross-validation moreover, and Neural Network model and examine the best suitable model when the data set contains trend, or seasonality or both.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ARIMA	Autoregressive integrated moving average
NN	Neural network
SVN	Support vector machine
GDP	Gross domestic product
MAsE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
RMSE	Root mean square error

y_i	Absolute value
$\hat{y}_i$	Forecast of y_i
ACF	Autocorrelation Function
PACF	Partial autocorrelation function

CHAPTER 1

INTRODUCTION

When looking at Time appears as a quantitative measure for life that is infinite and irredeemable. Time gives a chance to analyse things in terms of past, present and future. Time is simple when analysing things in terms of past that already known and recorded, present that is currently recording. However, in the case of the future, there is no way to determine the exact thing. However, in many cases, there is a chance to determine the future by the prediction that has based on the collective information of the prior records. It has planned, scheduled or arranged when something should happen or have done. The project has concentrated on the periodical data that have collected on recorded time.

1.1 Background

The forecasting methods classified into major two parts are a qualitative model and a quantitative model. The quantitative technique for forecasting utilises numerical realities and verifiable information to anticipate up and coming occasions. The two principle kinds of quantitative forecasting utilised by business analysts are the informative technique that endeavours to relate at least two factors and the time series strategy that utilizations past patterns to make conjectures — the qualitative prediction model based on either linear or non-linear time series [9]. The linear time series models are simple model, exponential smoothing, ARIMA, or Seasonal decomposition whereas, Neural network (NN), Support vector machine (SVN) and clustering are the non-linear time series analysis.

The improvement of systems in non-linear time series analysis has risen from its time series framework and created in the course of the most recent couple of decades into a scope of procedures which intend to fill a hole in the capacity to show and figure specific kinds of informational collections such chaotic determinate frameworks [1].

1.2 Motivation

For time series analysis prediction or forecasting presents several types of models can be linear or non-linear, but the requirement to choose the best model that analysis data set more precisely and accurately. From existing work, it has realised that their present different nature of data sets, that are showing different properties. Some data sets are changes with time; some remain stationary; some grow upward with time, and some are functional. However, there is not always the most suitable design that gives the most dependable prediction result. So, the requirement to identify the model best suits any particular data set.

1.3 Problem statement

We have numerous sorts of data sets present that requires time series analysis for future prediction. For this purpose, collected three data sets are agriculture value-added percentage of GDP is data set 1, India's monthly inflation rate is dataset 2 and Revenue generation of a restaurant is data set 3. These three data sets are distinct from among and may show different properties. There present many kinds of model that has applied to time series data for analysis, but at that time, there must be required to know that the which model is best suitable, for these purposes choose some linear or some non-linear time series model. The linear models are Autoregressive integrated moving average (ARIMA), Exponential Smoothing Models, Holt Linear Trend model, Seasonal Decomposition, Cross-validation moreover, the linear model Neural Network model.

1.4 Research and objective

Aim of this project is to identify among the several possible approaches in time series analysis and prediction, which suits best for the given data set. For this purpose, selected three data sets are agriculture value-added percentage of GDP is data set 1, India's monthly inflation rate is dataset 2 and Revenue generation of a restaurant is data set 3. There present some common or distinct properties in data set such as residuals, stationary, Trend, De-trending, seasonality.

Now, the first requirement to identify that the data set has contained which property. That may contain trend, or seasonality, or both.

Secondary, apply all possible models on the data sets and identify which results are more accurate.

Nowadays, in the world of digitalisation that promotes ease to do any task as well as in more inferior time. Also, by utilising the latest techniques, there is a provision to execute things accurately in the present by taking responsibility for the future. Various fields-maintained records, these records required to analysed timely, such time-series data set now analysing using various Machine learning models and Artificial model. When we have time-series data set the various models used for forecasting. Now, the challenge is to choose the most appropriate. Also, data sets are inconsistent.

In this project, the different time series prediction and forecasting models have applied on the same sets of Data set in order to identify the best. Also, before applying any model, test data-set for the time series components such as trend and seasonality. This project is done in several stages and removing the inconsistency present in the acquired data sets.

The three data sets selected are agriculture value-added percentage of GDP is data set 1, India's monthly inflation rate is dataset 2 and Revenue generation of a restaurant is data set 3. the collected data sets are essential and secure relevance place in the analysis model. Finally, from the experiments and results identify that among the several approaches which are best when the Data set showing trend property, seasonality, or both. The project consists of many stages such as Data Acquisition, Data Pre-position, Trend removal/ seasonality removal/ Non-stationarities removal, Filtering / smoothing, Forecasting.

CHAPTER 2

LITERATURE SURVEY

2.1 Time series

Time-series is the observation of the data that depend on values that fluctuate with time. The stock prices, weather forecasting, communication, sales, loans are some examples of time series.

The time series used for prediction model where only one variable is present that is time, the time is taken should have equal interval if one data has presented in months then next should also present in months, it is a discrete format. This represented as a line graph where the x-axis is the time interval, and the y-axis is the magnitude of the data.

2.2 characteristics of time series

The first defining characteristic of a time series is a list of observations where the ordering matters. Ordering is significant because there is a dependency on time, and changing the order could change the meaning of the data. To accurately forecast future values, we will need the measurement of data to be taken across sequential and equal intervals and with each time unit having at most one data point. Once the data have collected, there present two objectives. First, identify patterns represented by the sequence of observations and, secondly forecasting or predicting future values of time series. There presents another aspect that needed to consider before even starting the time series forecasting process. The first requirement to define what necessitated from the forecast, why wanted the forecast and what the effects of it. There have several levels to do this. First, needed to define the time frame to be used for the forecast, this could be yearly, monthly, weekly, daily or hourly anything could be possible. Before that, it is needed to determine which type of data using.

2.3 Measuring Parameters of time series: -

There are five types of measuring parameters for time series given and explained below:

2.3.1 Mean Absolute Error (MAsE): -

It measures the forecast error compared to the error of a negative forecast. Just always picking the last value observed. Lower the MASE value better the model is.
If x represents MASE value then, 0 < x < 1
X =1 means naïve forecast i.e. always picking the last observed value.
X=0.5 means the model has double the prediction accuracy as a naïve last value approach.
X>1 means model needs much improvement.

2.3.2 Mean Absolute Percentage Error (MAPE): -

Mean absolute percentage error is very similar to MASE, but it measures the difference of forecast errors and divides it by actual observation value.

$$\text{MAPE} = \sum_{i=1}^{n} |P_i| = \frac{\sum_{i=1}^{n} \left|\frac{100e_i}{y_i}\right|}{n} \qquad (2.1)$$

It does not allow for zero values. It puts much more weight on extreme values and positive errors. Independently It is scaled as we can use it to compare a model on a different dataset.

2.3.3 Root mean square error (RMSE): -

In order to prevent positive and negative error rates cancelling each other out, we use the root mean squared error. There we would square all error rates before adding them up so that we can finally get the root of it.

$$\text{RMSE} = \frac{\sqrt{\sum_{i=1}^{n} E_i^2}}{n} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \qquad (2.2)$$

Where,

y_i = absolute value

$\hat{y}_i$ = forecast of y_i

2.3.4 Mean absolute error: -

In mean absolute error, we took summation of absolute difference of actual value and predicted the value and take the average of it. The mathematical formula for MAE is,

$$\text{MAE} = \frac{\sum_{i=1}^{n} |e_i|}{n} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n} \qquad (2.3)$$

2.3.5 Akaike information criterion (AIC): -

It is a widespread measure in forecasting statistics modelling and machine learning. It is great to compare the complexity of different models. It penalises more complex models lower the AIC score better the model.

2.4 Components of time series

2.4.1 Residuals: - (X_i – fitted value)

Residuals seem to be little relevance. After all, they are called residuals, which is like a by-product but to make no mistake, they are important in many aspects. Most importantly, they can tell a lot about our modelling quality whenever we do time series analysis. We want all the patterns in the model; the only randomness should stay in the residuals. So basically, residuals are the container of randomness anything happening by chance or stuff. We cannot explain in mathematical terms stays in the residuals. That means in an ideal world; our model has a means of zero and constant variance. Of course, we want our residuals to be uncorrelated because in this case there would still be information left in the residuals.

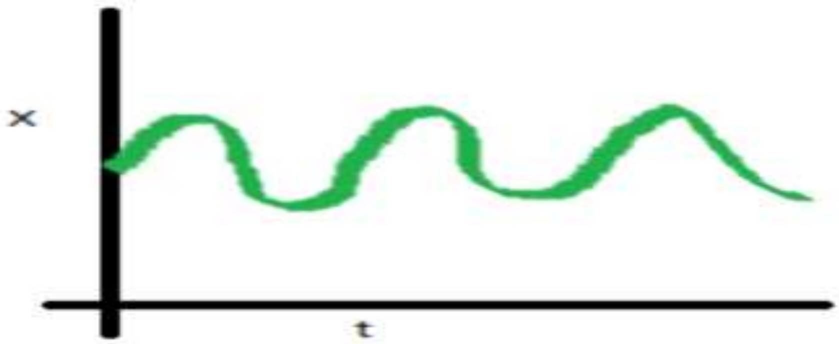
Moreover, ideally, we want them to be of a normal distribution which is not always possible fixing a non-zero mean is relatively simple. We add or subtract the non-zero mean, and have done expecting correlations is possible. They are modelling tools such as differencing, ensuring normal distribution and hence, the constant variance is tricker and sometimes impossible. We could try transformations like logarithms to work on this issue.

2.4.2 **Stationarity: -**

Stationarity is a key statistic when it comes to time series analysis and working with time data. The question behind it is that the data has the same statistical properties throughout the time series. The statistic properties, in this case, contain variance, mean, or autocorrelation. So, most analytical procedures in time series required us to have stationary data. However, the good thing is that if the data lacks that stationarity there are transformations that can be used to make the data stationary or we can use a procedure called differencing to change our data. Differencing adjusts our data according to the period that differs, for example, variance or means so that the statistics are the same throughout the dataset. After our differencing procedure, especially in the ARIMA model, it is used extensively.

Stationarity is the key statistics when it comes to time series analysis and working with time data. Time series data must be stationary without this property; we cannot apply any time series concept and models. For the series to be stationary it must follow three conditions: -

- Mean should be constant according to time
- Variance should be equal at a different time interval
- Co-variance should be equal



Stationary series

Figure 1 Graph of Stationary series

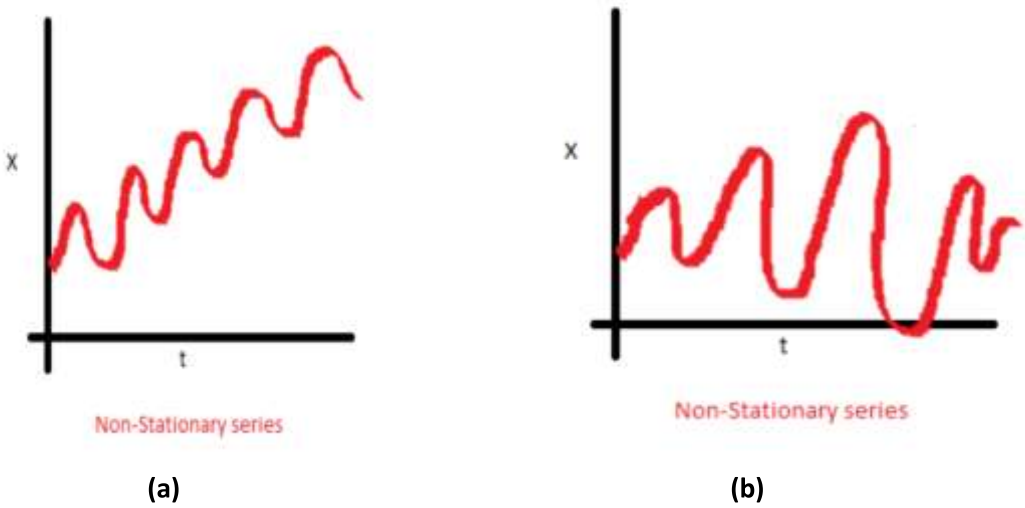


Figure 2 Graph of Non-stationary series (a) and (b)

In the first figure, it is seen that mean value increases with time, which results in an increasing trend in the dataset and makes the data non-stationary for stationary it should be constant throughout the time. In the second figure we certainly not see any fixed trend in it but the variance is the function of time which make it non-stationary time series

Figure 3 Graph of non-stationary series

In the third figure, indeed there is no constant mean no variance any trend but if we see the spread become closer concerning time as the time increases the spreads seems to be closer this show that covariance is present concerning time in the dataset that’s why it becomes non-stationary time series.

2.4.3 **Trend: -**

The trend is the pattern that noticed over a long period, which is comparatively increasing or decreasing at a time. The trend increased with the increase in time is called uptrend and the trend that decreased with the decrease in time is called downtrend, and if cannot determine whether increasing or decreasing called a stationary trend. The trend is something that appears for some time then disappears with time like a wear-out stage of hardware and repeating.

2.4.4 **De-trending: -**

A lot of time series has a trend in it, and if we check that sort of data out, we will find that the mean will change as a result of the trend and predictions tend to be underestimated as a result of this. We can easily de-trend our dataset and see if we get stationarity. If we put the trend component out of dataset, which gives a trend stationarity. If it is not enough, we can use differencing which gives difference stationarity. We can also use the unit-root test that can be used to see if we have trend stationarity or difference stationarity.

2.4.5 **Autocorrelation: -**

It is a statistical term which describes the correlation or the lack of such in a time series data set. It is a key statistic when it comes to time-series data because as soon as we work with that sort of data, there is always the question whether previous observations influenced the record one, so basically, we are talking about correlation on a time scale. Basic idea about autocorrelation is that in the ideal situation as soon as the lags increase correlation value decreases. This implies that values of a long-time back data have no more impact on the values of the current period. If it impacts, the influence will be very less. So, the most recent lags have a stronger influence on the current period time series value. The older lags could not impact the current period values.

2.5 Methods to get autocorrelation calculated

2.5.1 **ACF ( ) :- Autocorrelation Function**

It shows the autocorrelation between different time lags in a time series. If we need a measure for autocorrelation, this one is probably the best and gives a perfect idea even between those time-lapse.

2.5.2 **PACF ( ) :-Partial autocorrelation function**

In this function, the correlation coefficient adjusted for all shorter lags in a time series. ACF and PACF are used to determine whether our data is stationary or not. If it is stationary, we will develop our model if it is not, we will try to convert it into stationary data.

2.6 Time series models

2.6.1 **ARIMA Model: -** Autoregressive integrated moving average

This a system based on autoregression which allows us to model nearly any univariate time series. ARIMA is a standard modelling system for time series. An analyst with its money knows about this tool and how to use it. This is the most widely used system today. It is an extremely versatile system that can capture nearly all univariate time series data set. ARIMA is available for seasonal time series and multivariate time series as well. ARIMA model is also known as the Box-Jenkins model because it was published and described by the team of Jauch box and Goolam Jankin's around 50 years ago. Generally, an ARIMA model is all about three parameters p, d, q it means model consist of 3 parts. Autoregressive part represented by p, integration part represented by d, and Moving Average part q.

ARIMA models require stationary time series, or in other words, the model will make it stationary for us if it is not and only then can the other two parameters 'p' and 'q' be identified. Therefore, if we start with a non-stationary time series, it will be differenced until we obtain a stationary time series. The order of differences or how often we differenced a data set has described in the middle parameter 'D' of the model

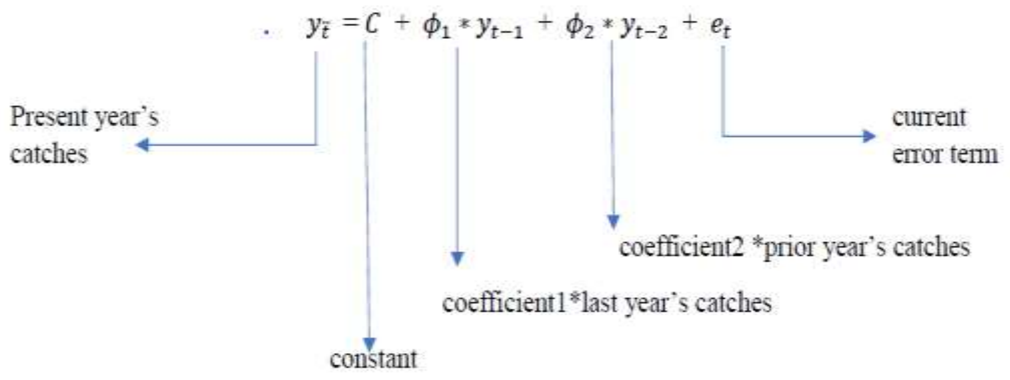


Figure 4: Mathematical equation of ARIMA model

Identify ARIMA model parameters: -

It is understandably not easy to figure out the ARIMA parameters right away when we are new to this. Some rules can help us in identifying a good model; we will divide this rule into three categories.

- parameter 'd'
- parameters 'p' & 'q'
- seasonal parameters 'P', 'D', 'Q'

2.6.1.1 Parameter 'd': -

If we see the significant lag in PACF plots out to even relatively high numbers let us say leg 8 or higher, we should add a differencing step in this case. It is simply impractical to have an ARIMA model with that high often order correlation parameters. In general, adding a differences step decreases the parameter number a lot and simplifies the model. On the other hand, if lag number one on the PACF plot is non-significant, then this is a good sign that we are on the right track and no more differencing is required. The same also applies if the auto-correlation in the PACF plot appears pattern-less and somewhat random if there is no differencing step in the model the data is expected to be stationary.

If we have one differencing step, which is d=1, then the original data has a constant trend, and if there are two differencing steps, then the trend in the data set varies throughout the timeline. Whereas if the d=0 then these will likely to be a constant in the model.

Order of d tells about the data: -

- d=0: Stationary data (constant)
- d=1: Trending data
- d=2: Varying trend (no-constant)

2.6.1.2 Parameter ‘p’ and ‘q’: -

As we know that PACF plot is the one responsible for the ‘p’ the auto-correlation parameter. If the PACF plot shows a significant positive auto-correlation at lag 1 or there is sharp cut-off between significant lags and not significant lag, then it is time to add at least one ‘p’ order. Similarly, we would look at the ACF plot, which is responsible for the moving average parameter ‘Q’. If the ACF plot is significant and negative with lag 1 or there is a sharp cut-off between significant and non-significant lags, then we would add at least one ‘Q’ order for the moving average part, but we are also aware that MA and AR affect each other. That means never add ‘p’ and ‘q’ orders at the same time to the parameters always test them individually; it is even possible that AR and MA terms cancel each other. Therefore, always vary with models that feature multiple AR and MA terms at the same time.

Further there is a problem when the summation of the coefficient is close to 1. This is true for either the summation of the MA coefficient or the summation of the AR coefficient. If we encounter such a situation, always add a differencing step and then determine a new AR or MA order.

Seasonal Parameters (P, D, Q): -

As we can imagine the more parameters, there are the more complicated the whole thing gets, so this is something we should only model once we gained some non-seasonal ARIMA experience. In general, it is recommended to get ACF and PACF plots for at least three full seasonal cycles plus some extra buffer. That means if we have monthly data and there is some seasonality yet this monthly level, we should take quint three years times twelve months plus a buffer. That means 40 lags should give us a reasonable picture now the fundamental rule with these seasonal models is if we see a healthy seasonal pattern which stays constant over several cycles, we need to perform seasonal differencing. That means the practice we put the parameter ‘d’ to one. In general, it is best to never use more than one differencing step for ‘D’. If we identify positive auto correlation at a given seasonal lag. That means this lag is significant in all seasonal cycles. Then we should add an AR order. If the auto-correlation is negative at a given seasonal lag, then we should add an MA order, so this is Q.

The situation with the positive seasonal auto-correlation and an added AR or P order is likely to occur if and only if we have a data set with a non-constant seasonal effect and no seasonal differencing. These are some general rules which we can use to identify the best models of course in the real world we would compare and test different models with information criteria and 0residual diagnostics.

2.6.2 **Exponential Smoothing Models: -**

Exponential smoothing forecasts use weighted averages of past observations, giving more weight to the most recent observation with weights gradually getting smaller as the observation gets older. ‘E’, ‘T’, and ‘S’ terms represent how error, trend and seasonality are applied in the smoothing method calculation. Selecting whether a time series exhibits additive or multiplicative behaviour in its terms relies on the analyst. Ability to see trend, seasonality, and error patterns while visualising the data does not seem very scientific. An effective way to determine how to apply a model.

The additive method is useful when the trend and seasonal variation are relatively constant over time

$$\text{Data}(y_t) = \text{Seasonality} + \text{Trend} + \text{Residuals} \qquad \text{(Additive models)}$$

Whereas,

The multiplicative method has used when the trend and seasonal variation increases or decreases in magnitude over time.

$$\text{Data}(y_t) = \text{Seasonality} * \text{Trend} * \text{Residuals} \qquad \text{(Multiplicative models)}$$

With the help of the diagram, we can see that the additive trend show linear behaviour. While multiplicative exhibits exponential behaviour.

2.6.3 **Neural Network model: -**

Neural networks are not single defined types of the algorithm like, for example, logistic regression or exponential smoothing. Neural networks do come in many different flavours, and there are multiple algorithms used in this powerful toolset for us. We are interested in the time series analysis side of neural networks. There is just a feel of these algorithms that do work for

this time series data set. The most important one of these algorithms being the auto regression-based neural network the most basic idea of any neural network is to simplify the input variable or variables over a single layer, or over several layers until we have left with one output value in a very primitive model. We have several input values, and they are combined to produce the output value such a model only has an input layer and an output layer, which is the result. So, this is pretty much like linear regression. Now usually a neural network also has a hidden layer which is an intermediary layer between the input and the output layers. Such an intermediary layer considered an artificial layer since we, as the analyst, compute this layer. So, the data contained in that hidden layer has produced with the help of the neural network. This is in contrast to the input layer, which consists of the original dataset; mostly, this neural network structure is called a multi-layer feed-forward network, which is a non-linear type of model.
This structure happens to be very suitable to process auto regression models for a time series the left value has used as the input; hence, the name NNAR (Neural Network Auto Regression model). Like the ARIMA model, there is a specific notation to these models, which is essential to keep in mind. NNAR (p, k) means that we have a neural network auto regression model where 'p' lag value used as input and 'k' nodes present in the hidden layer.

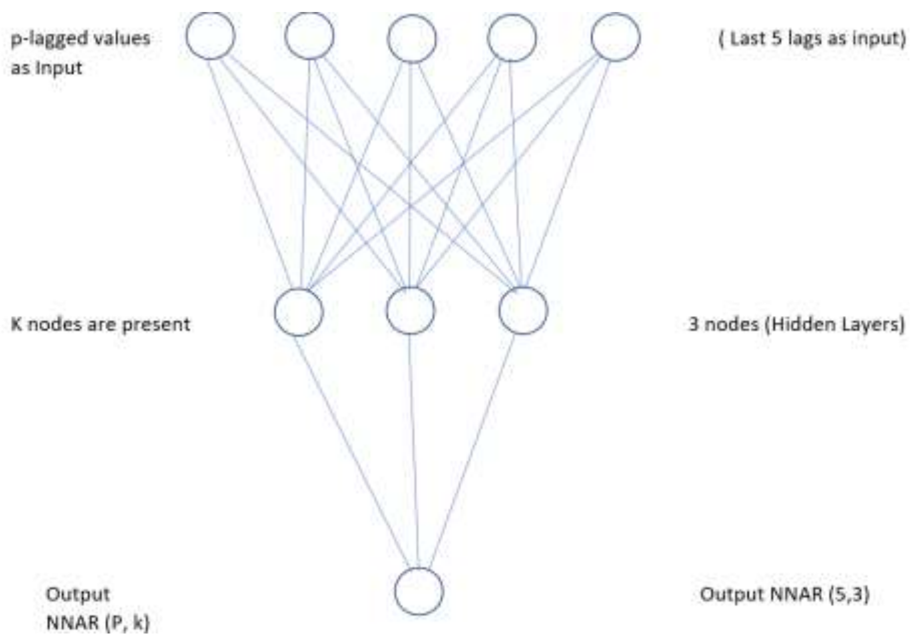


Figure 5: Neural Network Auto Regression model NNAR (p,k)

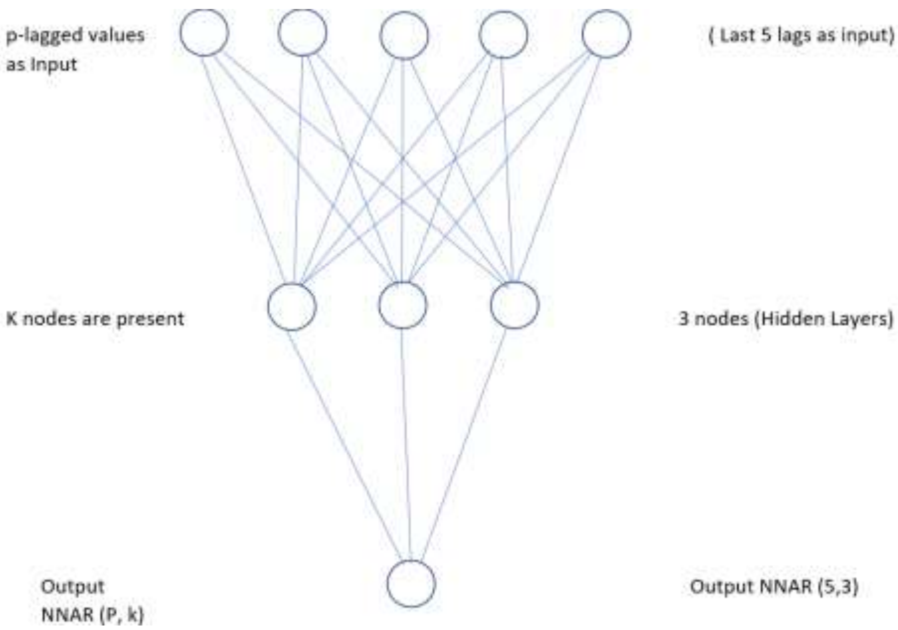


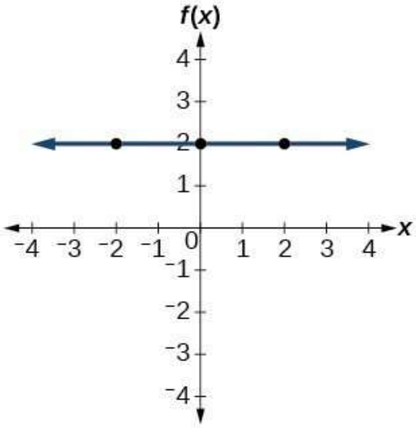
Figure 6: Neural Network Auto Regression model NNAR(p,P,k)

2.7 When Time series analysis is not applicable

There are some types of data that can not be applied on time series analysis are given below:

2.7.1 **Values are constant**

If the results are always remaining constant the there is no requirement of analysis the going to be same every time.



x	f(x)
−2	2
0	2
2	2

Figure 7: graph of constant time function

2.7.2 Values in the form of functions

Here the results can easily appear by using sin or cos functions. Here is no requirement of time series process, easily get results by putting values in functions**.**

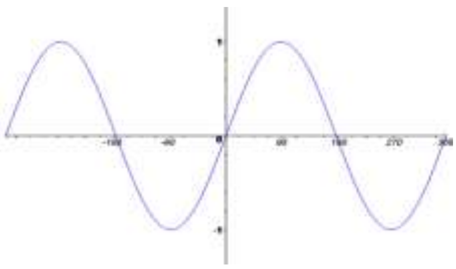


Figure 8: Sin function graph

CHAPTER 3

RELATED WORK

The forecasting of building energy utilisation is not another subject and has broadly examined lately [4-6]. The utilisation of forecasting as far as encouraging compelling building energy the executives has very much talked about [7,8], for example, to help energy utilisation target arrangement for building office the executive's office so the limit of building energy utilisation with various span could be better comprehended. Another fundamental purpose of building energy utilisation forecasting lies in the commitment to the information parameters and calendar data of some simulation tools.

Deb, C et al.[3] give a review on energy consumption forecast. There the various existing work has done on identifying the energy consumption on the data set. This paper shows an extensive audit of nine of the most famous machine learning systems. These nine models are Fuzzy time series, Grey prediction model, Autoregressive Integrated Moving Average (ARIMA), Support Vector Machines (SVM), Moving average and exponential smoothing (MA & ES), Nearest Neighbor prediction method (kNN), Hybrid models, Artificial neural network (ANN), Case-Based Reasoning (CBR). The extent of this survey paper, be that as it may, is confined over some examination of time series prediction and forecasting methods for the consumption of building energy which structure a necessary some portion of the building streamlining and control process.

Machine Learning strategies and Information based procedures have simultaneously utilised for sentiment analysis as a centre segment. Neural networks, including a myriad of deep learning variants like convolutional neural networks (CNN) [10], restricted Boltzmann machines (RBM) [11], long short-term memory (LSTM) networks [12], are experimented with prediction algorithms. Sometimes these models are also applied together with classic time series models, for example, autoregressive integrated moving average (ARIMA) [13,14].

In the paper [15], the issue of assault forecast is an intriguing examination issue that has moved toward numerous times by various specialists. Although numerous arrangements have proposed, there is still no definite answer on the most effective method to viably and decisively anticipate digital assaults. Assault expectation is not yet utilised by and by and some of the time seen as somewhat deceptive [19], yet it is as yet an open and an underlying, attractive research issue [16], [17], [18].

In paper [20], Foreseeing urban traffic is characterised in past surveys as an extremely unpredictable undertaking. Sign, collaborations with close connections, wellsprings of traffic, and an increasingly perplexing birthplace goal the connection has made a few analysts express that traffic stream in urban arterials - even single area traffic-can't be Anticipated as precisely as in expressways [21]. Once more, reproduction models are the most appropriate to address this issue, parametrising going before information sources; in any case, a few scientists have focused on this issue by joining spatial data to time arrangement [21], neural systems [22-24] or other non-parametric methodologies [25], [26]. Blood vessel traffic forecast has developed in intrigue, yet it establishes a slight segment of takes a shot at traffic estimating [27] and typifies alongside system expectation one of the primary difficulties of determining.

CHAPTER 4

PROPOSED WORK

In this section there is a detailed discussion about the data sets that have taken for the analysis purpose. However, the models that have used for the prediction modelling. This process has required to be done in various levels that is described in section 4.1.

In section 4.2, 4.3, and 4.4 given the three 1. data sets agriculture value-added % of GDP, 2. India's monthly inflation rate and 3. Revenue generation of a restaurant.

4.1 Proposed work stages

In this section given the levels in which data is analysed, this consists of Data Acquisition, Data Pre-position, Trend removal/ seasonality removal/ Non-stationarities removal, Filtering / smoothing, forecasting levels, this level is explained details and the flow chart is shown in figure 9.

Data Acquisition is the process of gathering data and information from an informatic source.in this step, we collect data or gather data set for further analysis. This is the prior stage where the data is gathering from a source, in this stage the three data sets collected are: -

1. data sets agriculture value-added % of GDP,
2. India's monthly inflation rate and
3. Revenue generation of a restaurant.

Data pre-processing and cleaning is an immensely talented yet overlooked part of the analysis. It often takes the most significant chunks off of our work just by looking at the data we can already exclude some modelling techniques just because the nature of the data does not allow it. Moreover, once we cleaned the data, we should always get a chart or line chart to be precise. This represents the presence of trend, seasonality or another interesting pattern. With this information, we can narrow down the possible types of model for this data. After that, data should also be clean and ready for a detailed analysis. In the next step, we will decide which model or models we are using from the ones that were deemed possible in the previous step.

This required research but based on the experience, we can right way find which model will be a useful fit. Based on the project, it is required to provide several forecast models and to calculate some common denominator quite often the average, but of course, only one model may be asked for the best model. Once we have identified the best model, we can start to fit and to run them. This is the stage where we generate the forecasts as well as the relevant graphs to be used for.

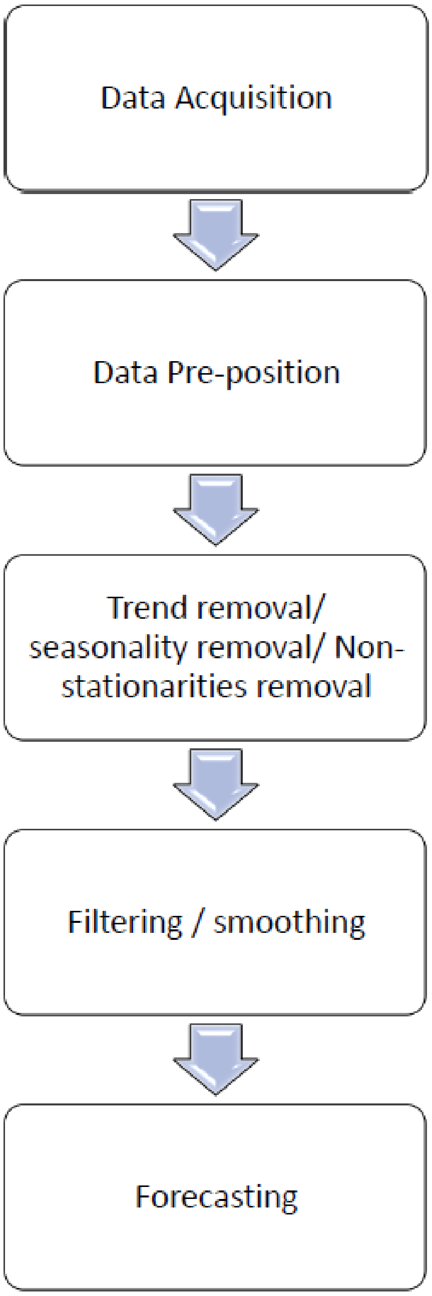


Figure 9: Steps of time series analysis and prediction

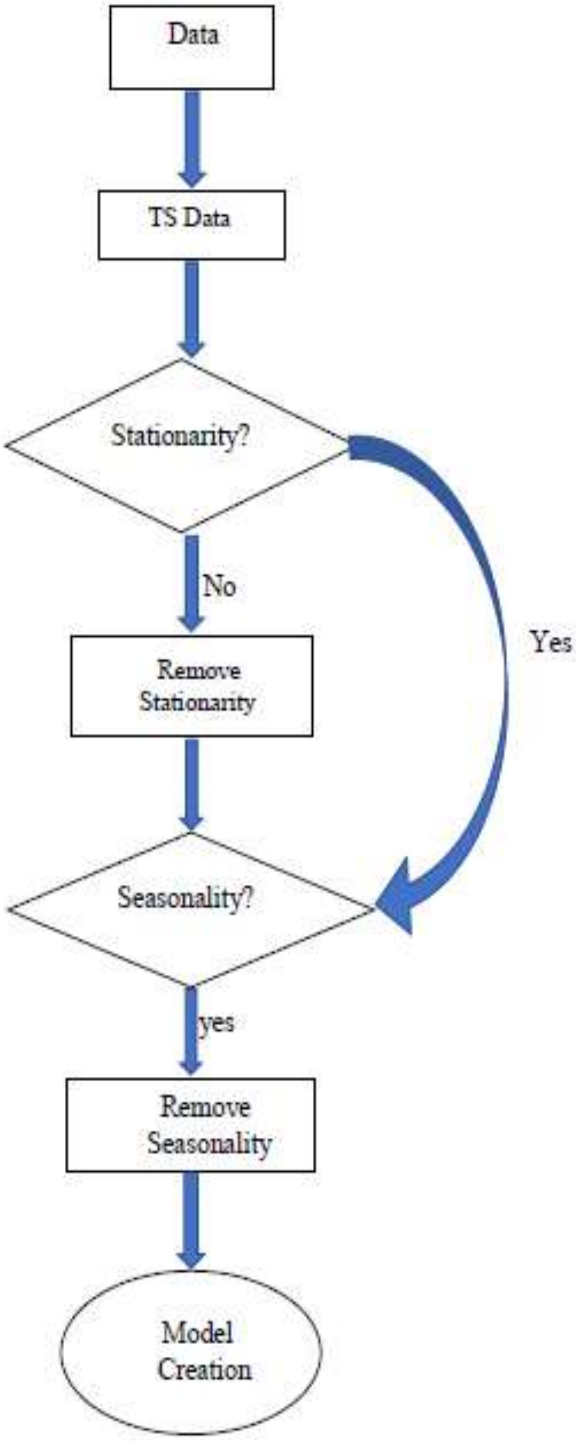
4.2 Data flow diagram

Every data set required to be pre-processing before applying any model. Also, there is requirement of identifying whether stationary and seasonality is present in data set or not. The DFD explains the flow of data processing before applying any model, in the first step the for any type of data it has prerequired to convert into Time series data given in figure 10.

```
Time Series:
Start = 1961
End = 2011
Frequency = 1
 [1] 41.77398 39.88881 41.07580 42.95914 40.91358
 [6] 41.81414 44.52621 43.52206 43.28513 41.95456
[11] 40.28153 40.27730 43.30959 40.30985 37.61800
[16] 35.75332 37.09044 35.46785 33.62536 35.38943
[21] 34.06881 32.87935 33.54486 32.21044 30.89263
[26] 29.73841 29.18245 30.19922 28.97086 29.02334
[31] 29.38953 28.73566 28.67698 28.27200 26.25641
[36] 27.12743 25.88702 25.79099 24.64728 23.12310
[41] 23.00178 20.74587 20.77125 19.02853 18.81054
[46] 18.28822 18.25627 17.78433 17.71924 17.74228
[51] 17.21640
```

Figure 10: Sample of the time series dataset

Now, time series data is check for stationary if stationary present then remove before moving to next step and if not present then move direction for checking seasonality. Only after this pre-processing stage apply intended model for prediction.



Flow chart

Figure 11: Data flow diagram of Time series analysis

4.3 DATA SET-1(agriculture value-added % of GDP)

India the land of farming, if we look after the history and geographical condition of India found that our motherland is one of the fertile countries in the world. India's geographical condition is so perfect that it is very suitable for agriculture and makes the soil fertile. It is vital to increase the agricultural area and production at the same rate as the population is increasing to keep balancing our economy. If our cultivation area and production have not increased as per the population rate, then it became very challenging to fulfil the need for food to our compatriots. As we all know that India's GDP has completely based on agriculture sector so, it may be possible that if we are not increasing our cultivation land and farming capability, our GDP disintegrated and we all going to face a serious problem. So, it has significant to analyse the contribution of agriculture in our GDP growth.

According to the world, bank Farming relates to ISIC divisions 1-5 and incorporates ranger service, chasing, and angling, just as the development of yields and animals' generation. Worth included is the net yield of an area in the wake of including all yields and subtracting middle information sources.

It has determined without making findings for the devaluation of created resources or exhaustion and corruption of natural assets. The starting point of significant worth includes controlled by the International Standard Industrial Classification (ISIC), modification.

This dataset contains the percentage of agriculture value-added in the gross domestic product of a country. In this dataset, the values of the percentage of agriculture in the GDP of India from 1961 to 2011 years of 50 long years. Using this dataset, we can analyse the decrement or increment of agriculture production and contribution of it in the GDP of our country.

The Data set I converted into time series data set represented in figure 12. The starting time of data set is 1961 and the end time is 2011, and distributed with the frequency of 1.

Moreover, the time series data set is diagrammatically represented in figure 13. using line graph. The X axis represents the Time from 1961 to 2011, and Y axis percentage of GDP that is agriculture value added.

```
Time Series:
Start = 1961
End = 2011
Frequency = 1
 [1] 41.77398 39.88881 41.07580 42.95914 40.91358
 [6] 41.81414 44.52621 43.52206 43.28513 41.95456
[11] 40.28153 40.27730 43.30959 40.30985 37.61800
[16] 35.75332 37.09044 35.46785 33.62536 35.38943
[21] 34.06881 32.87935 33.54486 32.21044 30.89263
[26] 29.73841 29.18245 30.19922 28.97086 29.02334
[31] 29.38953 28.73566 28.67698 28.27200 26.25641
[36] 27.12743 25.88702 25.79099 24.64728 23.12310
[41] 23.00178 20.74587 20.77125 19.02853 18.81054
[46] 18.28822 18.25627 17.78433 17.71924 17.74228
[51] 17.21640
```

Figure 12: Time series of data set I.

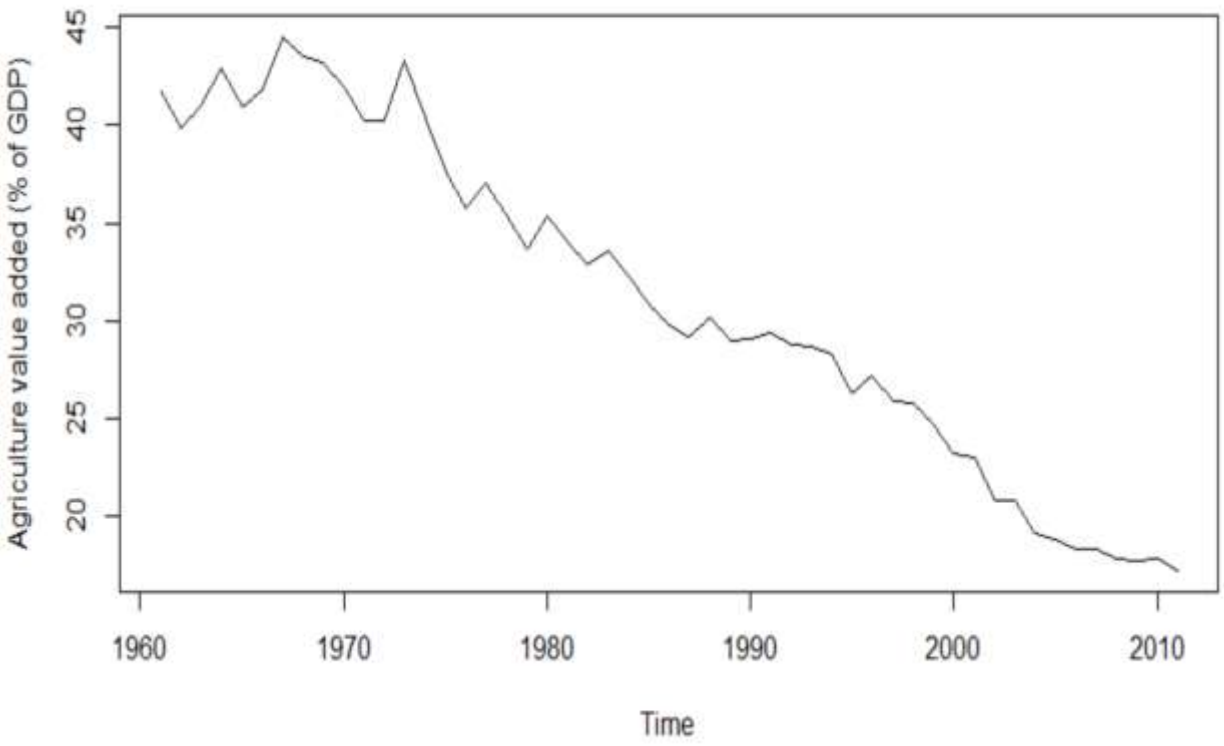


Figure 13: Graph plot of data set-1

4.3.1 Holt Linear Trend model:

Function for this model called Holt, which is part of the forecast package that goes in this package. The first argument is the time series with a data set, and that uses the argument (h) to specify the forecast length.

The simple problem with this specific data set as the trend is decreasing, and it becomes impossible to go beyond the zero per cent mark. The participation of agriculture in the country's GDP can be zero per cent, but it cannot be below zero per cent, i.e. negative. Therefore, the participation rate likely comes to holt long before the zero per cent mark. That means the curve has likely to flatten out in the range of 5 to 10 per cent participation.

If someone encounters a similar situation in proposed work as a statistical programmer keeps in mind that the corresponding expert always determines these sorts of threshold or parameters.

In this requires the input of the macroeconomist or a social scientist who perform to service to get an exact idea about the relevant threshold. As an analyst, we need to take care of these problems, and we can tackle these problems by using holt function. However, needed to use the damped argument, a damped holt linear trend model. Also, it assumes that a trend cannot be constant forever at some point growth needs to come to an end the curve needs to flatten out.

However, easily adjust the halt model with the damping parameters. It is possible to calculate a parameter value for us, or we can set it manually with the phi argument. By definition, when we use a damped parameter, the slope of the trend cannot be constant.

Smoothing parameters of the Holt linear trend model: -

Parameters	
Alpha (α)	Smoothing parameters for the level
Beta (β)	Smoothing parameters for the trend
Phi ( )	Damping parameters (0<phi<1)

Table 1: Smoothing parameters of Holt linear trend model

```
Holt's method

Call:
 holt(y = india, h = 15)

  Smoothing parameters:
    alpha = 0.7065
    beta  = 1e-04

  Initial states:
    l = 42.2863
    b = -0.4904

  sigma:  1.31

     AIC     AICc      BIC
233.9019 235.2352 243.5610

Error measures:
                       ME     RMSE       MAE        MPE
Training set -0.003281516 1.257598 0.9412211 -0.1894533
                 MAPE      MASE       ACF1
Training set 2.902127 0.8369204 0.04891844
```

```
Forecasts:
     Point Forecast      Lo 80    Hi 80     Lo 95
2012      16.680320 15.001461 18.35918 14.112727
2013      16.189930 14.134272 18.24559 13.046072
2014      15.699540 13.326084 18.07300 12.069652
2015      15.209150 12.555614 17.86269 11.150916
2016      14.718761 11.811938 17.62558 10.273159
2017      14.228371 11.088566 17.36818  9.426453
2018      13.737981 10.381266 17.09470  8.604329
2019      13.247591  9.687101 16.80808  7.802292
2020      12.757201  9.003930 16.51047  7.017069
2021      12.266811  8.330138 16.20348  6.246189
2022      11.776422  7.664470 15.88837  5.487734
2023      11.286032  7.005927 15.56614  4.740176
2024      10.795642  6.353700 15.23758  4.002278
2025      10.305252  5.707122 14.90338  3.273019
2026       9.814862  5.065636 14.56409  2.551548
```

Figure 14: Analysis of Holt Linear Trend model

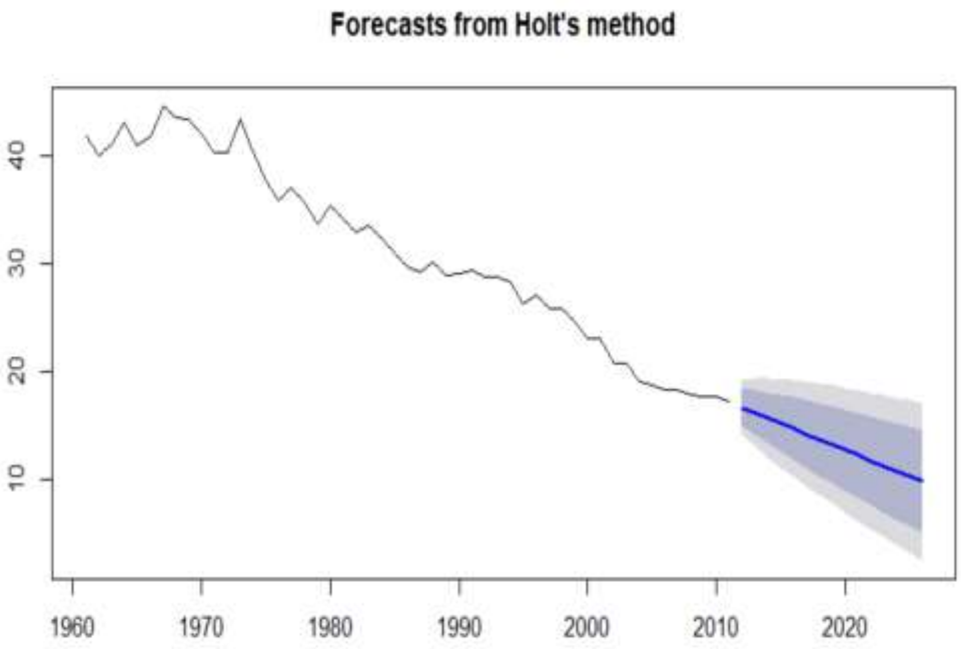


Figure 15: Forecast graph of Holt's method

ARIMA model: -

It is prevalent and well implemented in R. the basic idea of the ARIMA model is as follows; there are three parameters. Just picture them as three radio buttons as we wish in order to capture the structure of a dataset optimally.

4.3.1.1 **Auto-regressive (AR):** Seasonality or trend has captured via this first button

4.3.1.2 **Integration (I):** It differences the whole dataset, means from now on the data set consist of the differences between the observation instead of the original values.

4.3.1.3 **Moving Average (MA):** if there has some randomness or movement along with a constant. Mean, and this is likely covered with this button altogether.

In this way, we can eliminate a massive portion of the chaos in the dataset. It is a simplification process that helps to capture the patterns further. ARIMA models are very flexible, and thus very general random walks exponential Smoothing or Autoregressive models can all be explained with that system of ARIMA. For example, if the model only has an autoregressive component a simple AR model. Denoted as AR (1) or ARIMA (1,0,0). If it is moving average only, denoted as MA (1) or ARIMA (0,0,1).

```
Series: india
ARIMA(0,1,1) with drift

Coefficients:
          ma1    drift
      -0.2788  -0.4877
s.e.   0.1473   0.1305

sigma^2 estimated as 1.679:  log likelihood=-82.92
AIC=171.84   AICc=172.36   BIC=177.58

Training set error measures:
                       ME    RMSE       MAE        MPE
Training set -0.003358659 1.25704 0.9419477 -0.1898422
                 MAPE      MASE       ACF1
Training set 2.90193 0.8375665 0.03671365
```

Figure 16: Measured parameters of Moving Average MA(1)

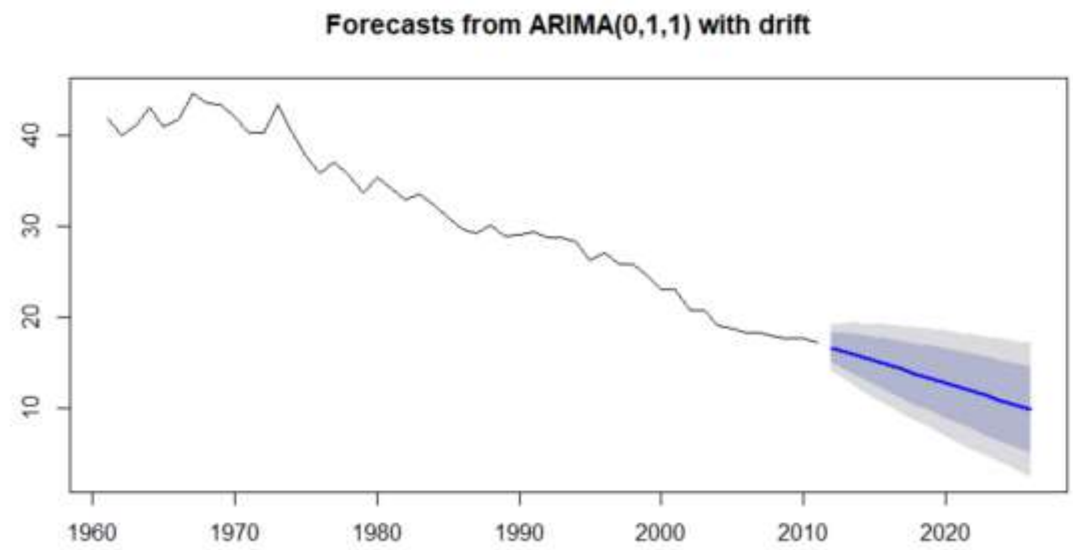


Figure 17: Forecast from ARIMA(0,1,1)

4.4 DATA SET-2(India's monthly inflation rate)

This dataset has totally based on seasonal time series. Besides trend seasonality is something we have regularly encounter in our analytics work. There are significant differences between a dataset with and without seasonality primarily the time series needs of frequency in order to work with these seasonal methods. So, keep the frequency argument in the time series function in mind whenever we use this one. The path to seasonal methods is at least theoretically open to us. As we learnt in the initial data import, this data set deals with Indian monthly inflation rates. It has given 50 years of inflation data from 1968 to 2018 has 12 months of frequency in time series. The data set has no trend, but there are some seasonal patterns. Therefore, we use models like Seasonal Decomposition, Exponential Smoothing and Seasonal ARIMA. Moreover, finally, we will compare all these models with others to find the best suitable model for this type of data set.

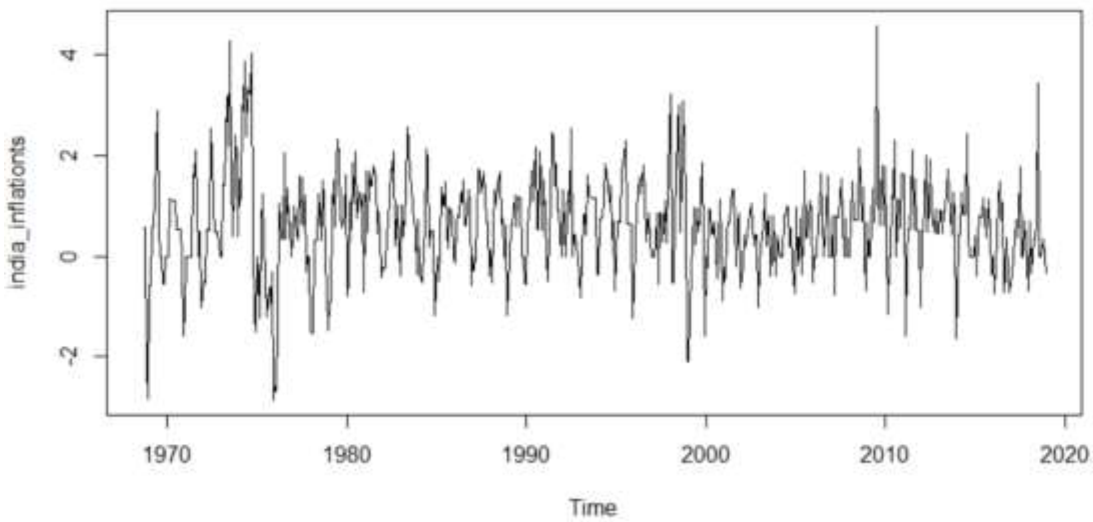


Figure 18: Graph plot of Data set-2

4.4.1 Seasonal Decomposition:

The single components of the time series look like this is called seasonal decomposition. Which is quite old, and the method in its purest form is 100 years old with seasonal decomposition. Basically, in this, we divide data into trend, seasonality and remainder, which should be random. We either can choose a method which adds these three components up, or we can opt for a multiplicative model. Which multiplies that in general if just seasonal cycles. It is best to opt for an additive decomposition. The strength of the method, however, is its ease of using its simplicity. It is something we like to get a quick idea about data. It provides a starting point for further data exploration.

Now do keep in mind that the seasonal decomposition has several drawbacks which put this method into disadvantage towards a tool like ARIMA or exponential smoothing. For example, the first few observations result in an NA. since the model uses moving averages, which requires some initial data. This method is also slow to catch fast rises in the data set and on top of that model assumes that the seasonal component stays constant. This is an assumption that gets problematic when we have a very long series with a higher chance of seasonal changes. The primary function of this method has decomposed, which is even part of our base.

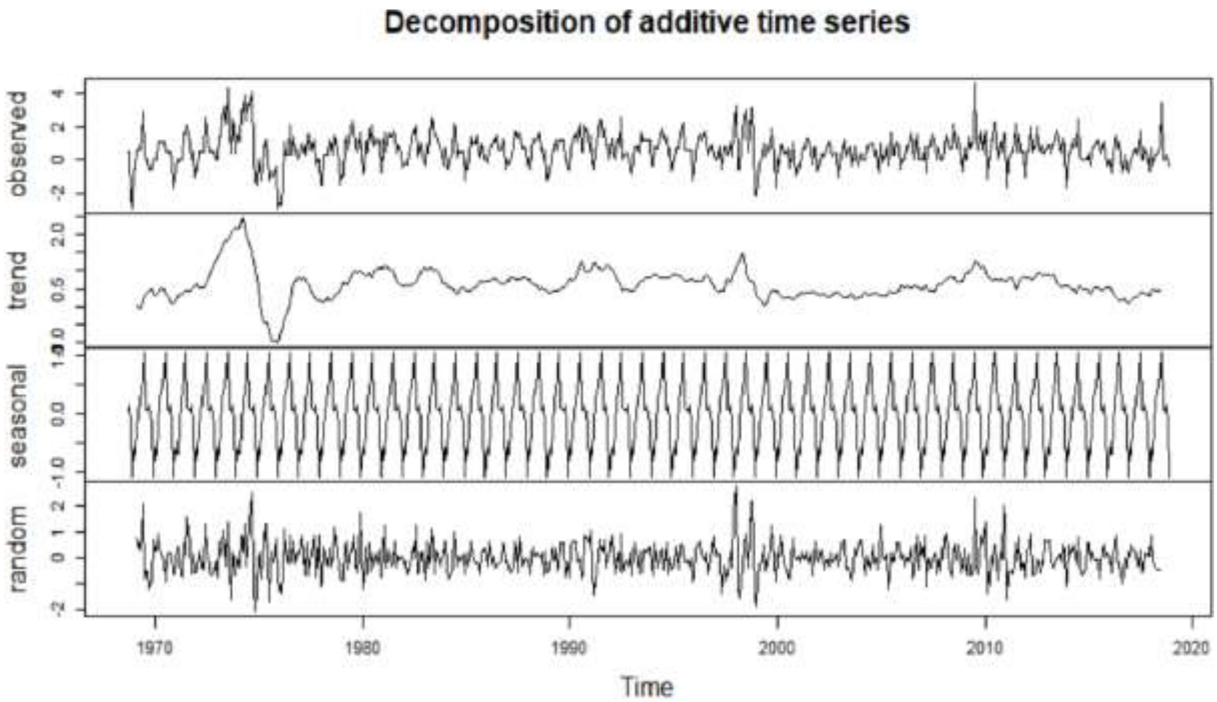


Figure 19: Decomposition of Additive time series

4.4.2 Seasonal ARIMA:

ARIMA models can not only be used on the non-seasonal dataset as well, but the whole thing gets significantly more complicated. We find out the parameters step by step using ACF and PACF plots and differencing the dataset if required. This process is theoretically the same with the seasonal dataset, but as we might imagine, the process complexity is a lot higher, but luckily for us, there is also the automated method with the auto. Arima function. The main difference between a non- seasonal and seasonal ARIMA model is the effect that we get another set of parameters. They do the same as the first set but just for the seasonal part, given in figure 20.

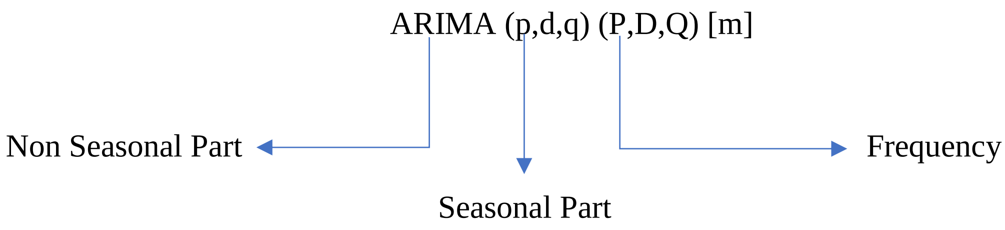


Figure 20: Parameters of seasonal ARIMA model

Autocorrelation, differencing and moving average part of function stay the same. However, now we need to multiply the seasonal and non-seasonal part with each other in order to compute the results.

So mostly each non- seasonal component gets a big uppercase letter of the seasonal part, and these letters multiplied if they are present in the model.

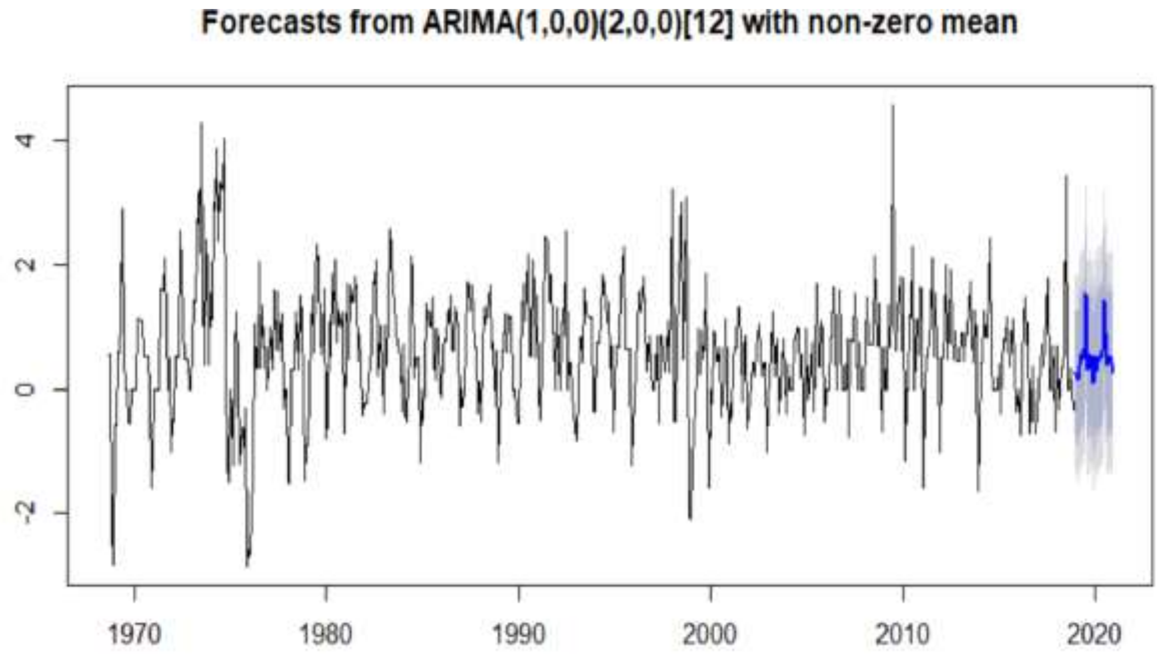


Figure 21: Forecasts from ARIMA (1,0,0) (2,0,0) [12] with non-zero mean

4.4.3 Exponential Smoothing with ETS:

If we want to model this dataset with the exponential smoothing system, there are two main ways in how to do this the classic exponential smoothing method used for a seasonal dataset is called Holt's winter seasonal method. The standard alternative to this method would be the ETS method with the ETC function no matter which of these two systems we go for the underlying principle is always the same. As usually exponential smoothing the three parameters error, trend, and seasonality have a corresponding smoothing coefficient.

The model will be fully auto-generated by default. That is what stated with the argument model. All three components of the function are put to set. Z means auto-generated if it has set to 'A' this means the model will be additive for that component. M means multiplicative, and N means non-existent. So, it is A, M, N and Z that we have available here in the model part. So, if we have a model with "MZM" given in figure 22.

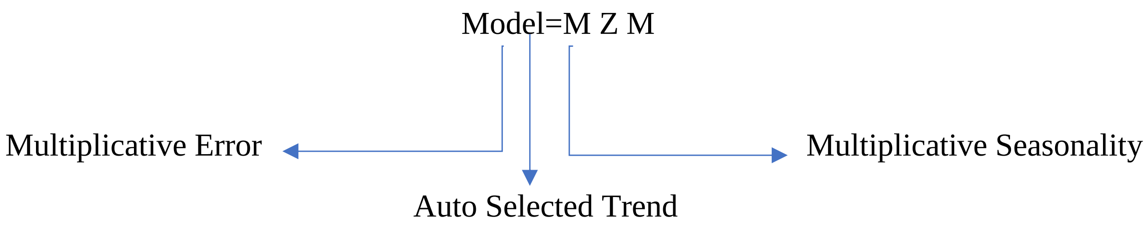


Figure 22: Exponential smoothing parameters

So, we can mix fixed settings with the auto setting set by the way this argument plays together with the additive only argument which we can use to only get additive components in our model settings, and it also is related to allow to multiplicative dot trend which is used to restrict the trend component to non-multiplicative somewhat. This model here which we are going to use allows a damping effect. This model is the same as in a damped hold trend model. This model is used to slightly decrease the trend over time the corresponding coefficient argument for this is phi ().

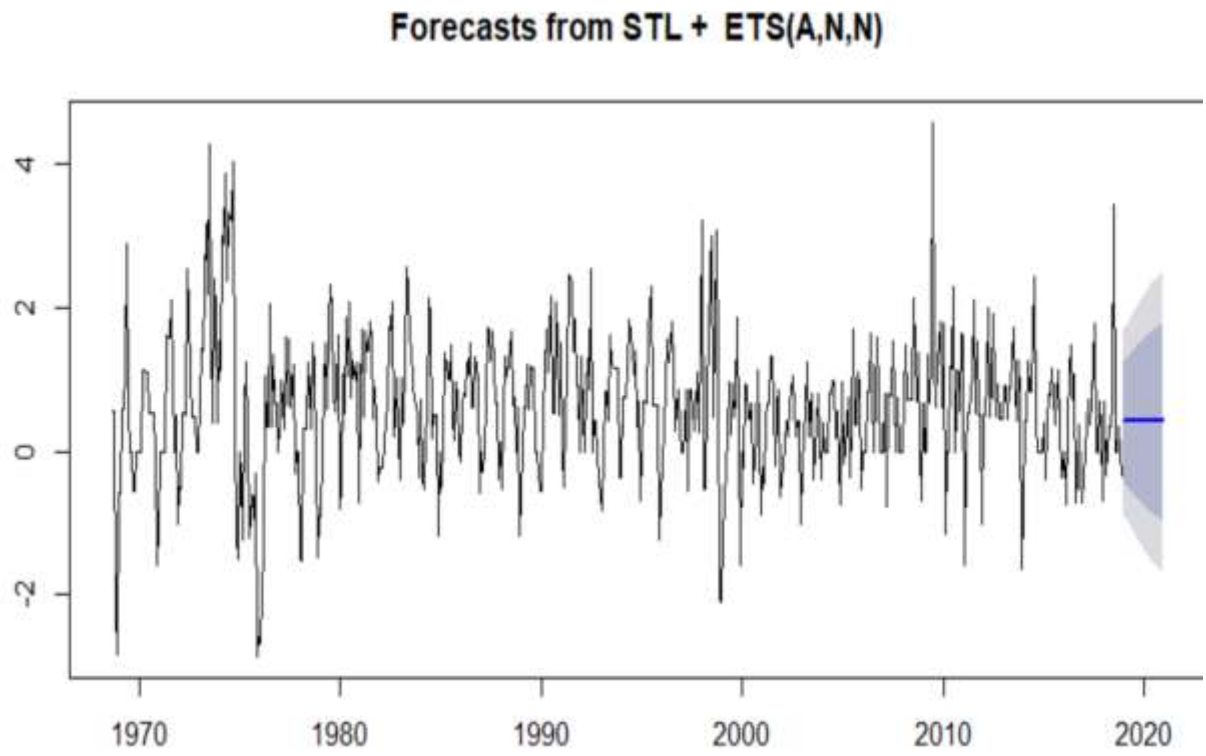


Figure 23: Forecast graph of STL +ETS(A,N,N)

4.4.4 Cross-validation: -

We have several models calculated it is, of course, paramount to know which of the models is the best one. It is not a good idea to use info criteria to compare different systems of models. ARIMA and Exponential Smoothing cannot be compared just with information criteria. The underlying statistics used to calculate the information criteria are different for both model systems. Therefore, we have to concentrate on prediction or forecast accuracy Mean Square Error (MSE). From the forecast, residuals have proven to be a measure of choice to compare ARIMA & Exponential methods. In the forecast package, there are several functions available that allow us to compute relevant statistics. The general idea of cross-validation is to produce an error rate for each time point in the series. The error rate is essentially a measure for the discrepancy between the calculated value for the task set and the actual value. The time series splits into training and test sets. However, the test set is only one observation, and the training set is all the data before this test set time point. Data after the test set observation has not used for that specific error point. That means error rates have computed one after another and along the timeline is called a rolling forecast origin.

```
ARIMA(2,0,2)(1,0,1)[12] with non-zero mean : Inf
ARIMA(0,0,0)            with non-zero mean : 1677.581
ARIMA(1,0,0)(1,0,0)[12] with non-zero mean : 1448.639
ARIMA(0,0,1)(0,0,1)[12] with non-zero mean : 1484.027
ARIMA(0,0,0)            with zero mean     : 1876.555
ARIMA(1,0,0)            with non-zero mean : 1503.254
ARIMA(1,0,0)(2,0,0)[12] with non-zero mean : 1424.017
ARIMA(1,0,0)(2,0,1)[12] with non-zero mean : Inf
ARIMA(1,0,0)(1,0,1)[12] with non-zero mean : Inf
ARIMA(0,0,0)(2,0,0)[12] with non-zero mean : 1564.776
ARIMA(2,0,0)(2,0,0)[12] with non-zero mean : Inf
ARIMA(1,0,1)(2,0,0)[12] with non-zero mean : 1425.611
ARIMA(0,0,1)(2,0,0)[12] with non-zero mean : 1444.371
ARIMA(2,0,1)(2,0,0)[12] with non-zero mean : 1424.483
ARIMA(1,0,0)(2,0,0)[12] with zero mean     : 1445.078

Best model: ARIMA(1,0,0)(2,0,0)[12] with non-zero mean
```

Figure 24: Time series of cross-validation

4.5 DATE SET -3 (Revenue generation of a restaurant)

The dataset we are using required several cleaning steps. We have raw data there are so many missing data probably there are some outliers, and on top of that, there are quotations which are totally out of place and prevent the proper usage of the dataset. Commonly, we will encounter problematic dataset like this. Necessary to clean the dataset. The dataset shows the revenue of the restaurant in us dollars. The dataset ranges from January 1997 to December 2016. So, we have 20 years of data. Dataset has many problems even at first glance we can see that the format is not usable in the current form there are many quotations around the revenue numbers. These should not be there. So, the person providing this dataset did a terrible job in formatting the data. In fact, in our consulting work, we get data presented in shapes that we could not imagine. If we scroll through data, we can see that there is some text in it and some values make no sense at all. Usually, revenue is around 10 to 40000 us dollars per month for the restaurants. However, the data shows a month with over 3 million revenue only 3 dollars per month. So, these are clear outliers which needed to address in the dataset cleaning, and there are missing values as well.

1997	20442	20052	19414	18176	15561	17246
1998	23944	21866	20923	19960	17726	16123
1999	21205	20427	20468	18063	16330	16392
2000	324	22947	19306	17567	15644	16240
2001	21852	20978	20385	18141	16359	16057
2002	NA	NA	NA	18787	17336	16636
2003	23477	23400	821	18221	16572	18821
2004	NA	23625	21673	20204	19373	19378
2005	25446	24875	23260	21606	19290	19907
2006	25933	25186	25110	21642	19645	19435
2007	27050	27652	26085	24043	21412	22724
2008	27842	27006	25577	23116	24367	25590
2009	30899	31474	28837	25597	24228	23616
2010	32718	32567	30874	27278	24373	23177
2011	31637	29600	29259	26501	23642	24734
2012	31477	31405	29027	26228	24739	3334333
2013	30624	31047	30076	27232	3	24669
2014	32201	32439	30640	28841	25591	26752
2015	32601	32945	31860	29742	28788	28598
2016	33916	34366	33185	29267	29369	28964

```
> summary(revenue_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      3   18980   23218   36912   26816 3334333       4
```

Figure 25: Time series of dataset-2

4.5.1 Cleaning and pre-processing:

It is imperative to clean the data. Many times, cleaning of data or pre-processing takes the most time by far. The dataset needs to convert into a proper numeric time-series format and the missing data as well as outliers should be replaced with plausible values. Interesting functions for these tasks are separate from TTR in order to extract the numbers and disk clean from the forecast, as a generally easy to use a cleaning too. Data have converted from its messy character column to a proper time series.

With the dataset like this, we have to think individually about each function we use on the dataset with that we mean it is perfectly fine to use general modelling techniques like ARIMA or ETS to see if there are general patterns in the data set. However, since this is irregularly spaced data, there is a lower chance that any seasonal pattern can be detected. For that, there have to pre-process the information with the goal that we can regularise the information. It needs to have equal intervals in order to provide a frequency for seasonality.

```
> summary(revenue_ts)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14209   19280   23267   23282   26658   34366
```

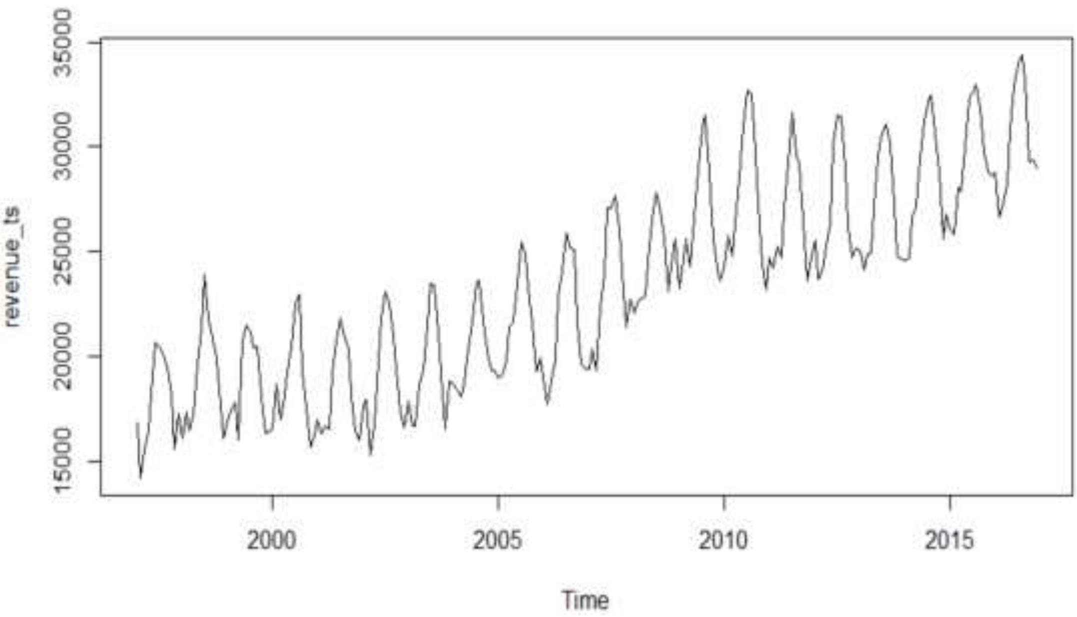


Figure 26: Graph plot of time series dataset-2

CHAPTER 5

EXPERIMENTS AND RESULTS

5.1 Dataset-1

The first dataset is trending model data. It is of India's agriculture value-added in GDP percentage growth, in which we compared the exponential smoothing models with ARIMA models and found that ARIMA models show the better result. In this, we avoid damping parameter option (ϕ) to avoid an unrealistic linear trajectory the general idea of this first dataset is to how to handle a data that contain trend. The dataset contains trend in isolated instance means there is no seasonality present in it.

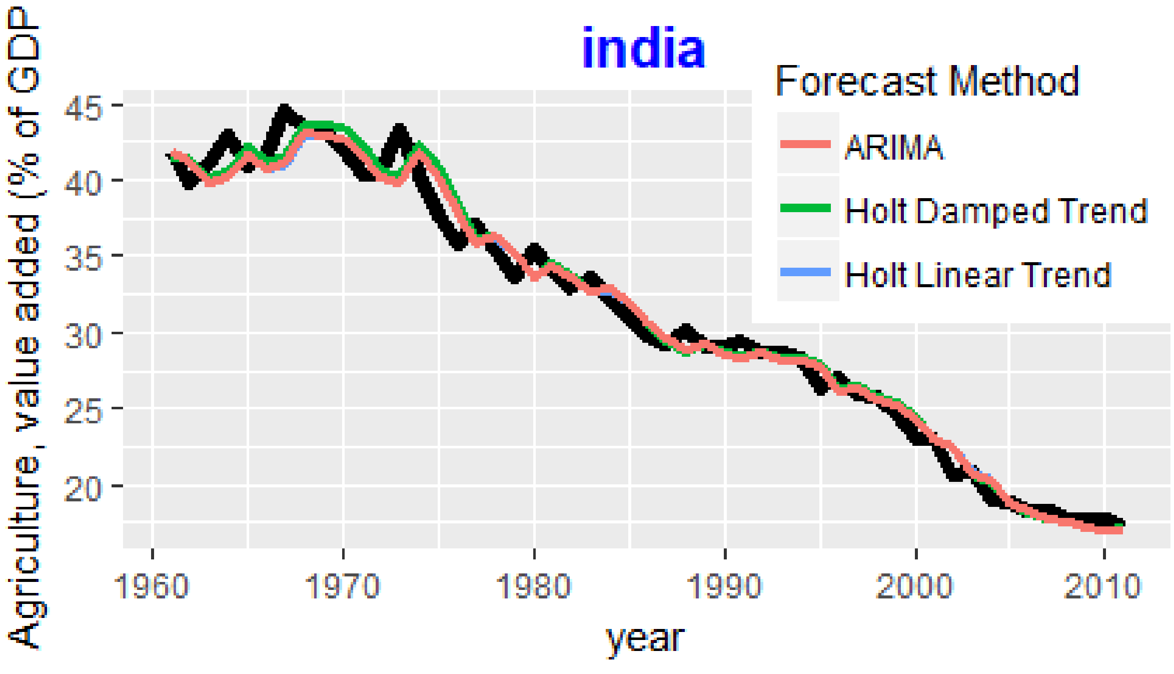


Figure 27: Comparison graph plot of dataset-1

5.2 Dataset-2

Second Dataset contain a seasonality without trend. It was India's monthly inflation rate data here we use exponential smoothing and ARIMA models with different parameters and functions and then use time series cross-validation. We try to find out which model forecast accurate results and we found that ARIMA model perform best. Even in seasonal dataset the output of seasonal ARIMA looks different to a standard one, we get a second set of parameters for this seasonal part of the model.

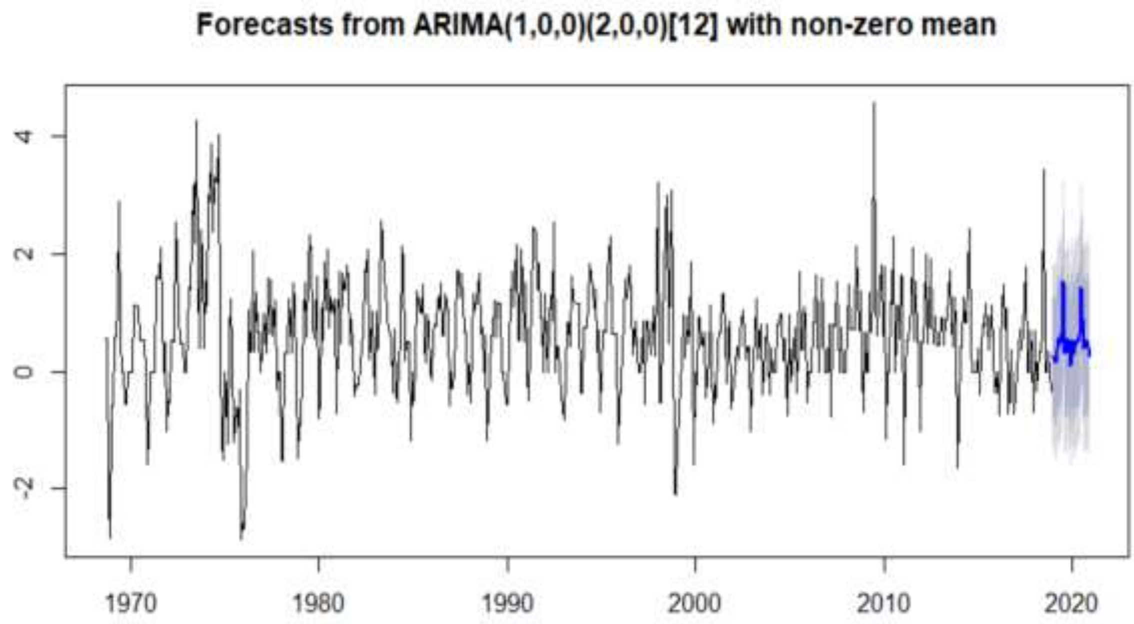


Figure 28: Forecast result graph plot of dataset-2

5.3 Dataset-3

The third dataset is slightly different from the other two dataset. The third dataset is of revenue generation of restaurant annually. It is more towards a pre-processing one as a lot of data is missing and data need to be normalised, it has no seasonality and trend. On this dataset we apply some time series models and found that the neural network model shows the best result.

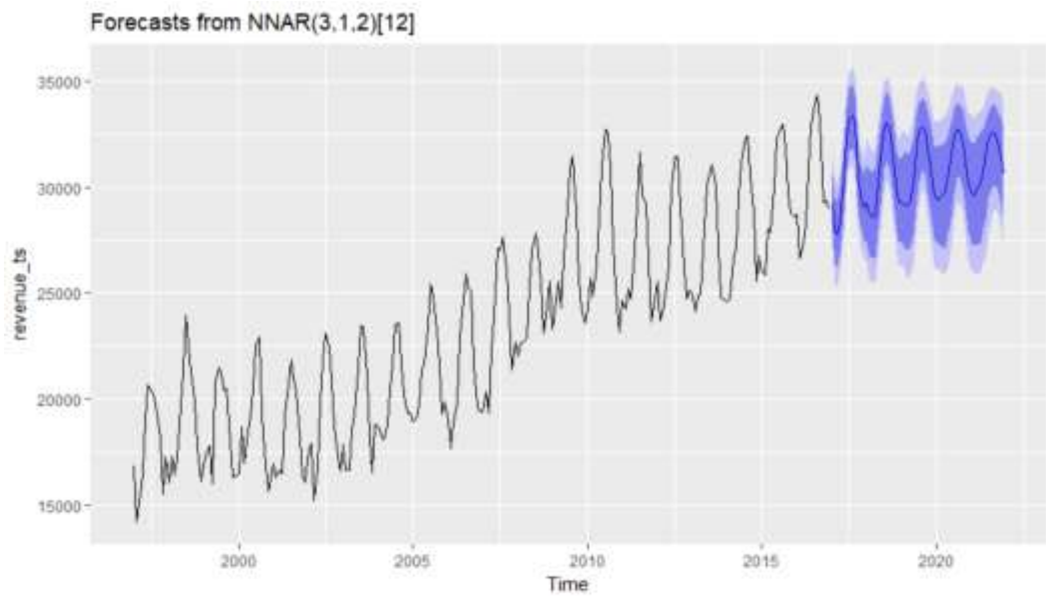


Figure 29: Forecast result of graph plot dataset-3

From analysis results of the three data set, we conclude the following: -

1. In data set-1, the trend is present, but seasonality is not present and shown in table 2, and ARIMA is the best model suitable for this purpose.

2. In data set-2, seasonality is present, but the trend is not present and shown in table 2, and seasonal ARIMA is the best model suitable for this purpose.

3. In data set-3, both trend and seasonality are present and shown in table 2, and the Neural Network model is the best model suitable for this purpose.

Datasets	Trend	Seasonality
Dataset-1 (India's agriculture growth in GDP %)	√	X
Dataset-2 (India's monthly inflation rate)	X	√
Dataset-3 (revenue generation of restaurant)	X	X

Table 2: Table of data set contained trend or seasonality.

Trend	Seasonality	Model
√	X	ARIMA
X	√	Seasonal ARIMA
X	X	Neural Network Model

Table 3. Table for different models and properties showing.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Time series analysis is significant in light of the fact that to every single association to succeed then it needs to comprehend itself in regards to execution, accomplishments, typically, and so on. Every single association ought to guarantee utilisation of this strategy when fundamental of estimating on the off chance that it is to succeed. An active business is the one which is in the situation to conjecture its prospects. For this situation, it will be in a situation to dodge any conceivable misfortune that may happen. Time arrangement is finished by guaranteeing that legitimate arranging and setting up better approaches.

During determining, time arrangement information is generally significant. Numerous associations depend on information delivered by bookkeepers to estimate the prospects of the firm. Organisation cherishes this arrangement of anticipating since they can be in the situation to cure any future activity that can be negative to organisation tasks. Along these lines, time arrangement information is significant in anticipating of something that continues changing over the extensive stretch at model costs of stock, marketing projections, and benefit. Continually, anything that is seen now and then successively is called time series. The significant point of estimating time arrangement information is to have the option to decide and obtain some much-needed education or comprehension on how and for to what extent the perception will keep on future. There are various techniques for ascertaining time arrangement; they incorporate.

REFERENCES

1. Camilleri, M. (2004, September). Forecasting using non-linear techniques in time series analysis: An overview of techniques and main issues. In Computer Science Annual Research workshop (pp. 19-28).
2. Rani, S., & Sikka, G. (2012). Recent techniques of clustering of time series data: a survey. International Journal of Computer Applications, 52(15).
3. Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74, 902-924.
4. H. Zhao, F. Magoulès, A review on the prediction of building energy consumption, Renew. Sustain. Energy Rev. 16 (2012) 3586–3592, http://dx. doi.org/10.1016/j.rser.2012.02.049.
5. S. Ahmad, M.Y. Hassan, M.P. Abdullah, H. Rahman, F. Hussin, H. Abdullah, et al., A review on applications of ANN and SVM for building electrical energy consumption forecasting, Renew. Sustain. Energy Rev. 33 (2014) 102–109,
6. A. Foucquier, S. Robert, F. Suard, L. Stéphan, A. Jay, State of the art in building modelling and energy performances prediction: a review, Renew. Sustain. Energy Rev. 23 (2013) 272–288.
7. C. Deb, L.S. Eang, J. Yang, M. Santamouris, Forecasting diurnal cooling energy load for institutional buildings using artificial neural networks, Energy Build. 121 (2016) 284–297.
8. C. Deb, F. Zhang, J. Yang, S.E. Lee, K.W. Shah, A review on time series forecasting techniques for building energy, Consumption (2017),
9. Barão, S. M. M. (2008). Linear and Non-linear time series analysis: forecasting financial markets (Doctoral dissertation).
10. Ding, X.: Research on methodology of market trends prediction based on social media. Ph.D. thesis, Harbin Institute of Technology (2016)
11. Yoshihara, A., Seki, K., Uehara, K.: Leveraging temporal properties of news events for stock market prediction. Artificial Intelligence Research 5(1), 103–110 (2016)
12. . Li, L., Qin, B., Ren, W., Liu, T.: Truth discovery with memory network. CoRR abs/1611.01868 (2016)

13. Liu, C., Hoi, S.C.H., Zhao, P., Sun, J.: Online arima algorithms for time series prediction. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
14. Zhang, G.P.: Time series forecasting using a hybrid arima and neural network model. Neurocomputing 50, 159–175 (2003)
15. Husák, M., Komárková, J., Bou-Harb, E., & Čeleda, P. (2018). Survey of attack projection, prediction, and forecasting in cyber security. IEEE Communications Surveys & Tutorials, 21(1), 640-660.
16. A. Kott, Towards Fundamental Science of Cyber Security. New York, NY: Springer New York, 2014, pp. 1–13.
17. S. J. Yang, H. Du, J. Holsopple, and M. Sudit, Attack Projection. Cham: Springer International Publishing, 2014, pp. 239–261.
18. E. Vasilomanolakis, S. Karuppayah, M. Muhlh ¨ auser, and M. Fischer, ¨ "Taxonomy and survey of collaborative intrusion detection," ACM Comput. Surv., vol. 47, no. 4, pp. 55:1–55:33, May 2015.
19. A. Chuvakin, "Sad hilarity of predictive analytics in security?" http://blogs.gartner.com/anton-chuvakin/2016/03/31/sad-hilarity-ofpredictive-analytics-in-security/, March 2016, Published: 2016-03-31, Accessed 2017-07-15.
20. Lana, I., Del Ser, J., Velez, M., & Vlahogianni, E. I. (2018). Road traffic forecasting: recent advances and new challenges. IEEE Intelligent Transportation Systems Magazine, 10(2), 93-109.
21. A. Stathopoulos and M. G. Karlaftis, "A multivariate state space approach for urban traffic flow modeling and prediction," Transport. Res. Part C Emerging Technol., vol. 11, no. 2, pp. 121–135, 2003.
22. H. Yin, S. Wong, J. Xu, and C. Wong, "Urban traffic flow prediction using a fuzzy-neural approach," Transport. Res. Part C Emerging Technol., vol. 10, no. 2, pp. 85–98, 2002.
23. M. Annunziato and S. Pizzuti, "A smart-adaptive-system based on evolutionary computation and neural networks for the on-line short-term urban traffic prediction," in Proc. European Symp. Intelligent Technologies Hybrid Systems and Their Implementation on Smart Adaptive Systems, Aachen, 2004.
24. H. Dia, "An object-oriented neural network approach to short-term traffic forecasting," Eur. J. Oper. Res., vol. 131, no. 2, pp. 253–261, 2001.

25. L. Dimitriou, T. Tsekeris, and A. Stathopoulos, “Adaptive hybrid fuzzy rule-based system approach for modeling and predicting urban traffic flow,” Transport. Res. Part C Emerging Technol., vol. 16, no. 5, pp. 554–573, 2008.
26. W. C. Hong, Y. Dong, F. Zheng, and C. Y. Lai, “Forecasting urban traffic flow by SVR with continuous ACO,” Appl. Math. Model., vol. 35, no. 3, pp. 1282–1291, 2011.
27.] L. Moreira-Matias, J. Mendes-Moreira, J. de Sousa, and J. Gama, “Improving mass transit operations by using AVL-based systems: A survey,” IEEE Trans. Intell. Transport. Syst., vol. 16, no. 4, pp. 1636– 1653, 2015