Data Analysis and Prediction in Globe-trotting and Budget Executive Suite

Submitted By

Gondalia Ankita Dineshbhai

17MCEC06

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2019

Data Analysis and Prediction in Globe-trotting and Budget Executive Suite

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Gondalia Ankita Dineshbhai

(17MCEC06)

Guided By

Prof. Vishal Parikh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2019

Certificate

This is to certify that the major project entitled **"Data Analysis and Prediction in Globe- trotting and Budget Executive Suite"** submitted by **Gondalia Ankita Dineshbhai (17MCEC06)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering (Specialization in title case, if applicable) of Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this major project part-II, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Vishal Parikh
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. Priyanka Sharma
Professor,
Coordinator M.Tech - CSE (Specialization)
Institute of Technology,
Nirma University, Ahmedabad

Dr. Madhuri Bhavsar
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr Alka Mahajan
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Gondalia Ankita Dineshbhai**, **17MCEC06**, give undertaking that the Major Project entitled "**Data Analysis and Prediction in Globe- trotting and Budget Executive Suite**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made.It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Prof. Vishal Parikh

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Vishal Parikh**, Assistant Professor, Computer Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Madhuri Bhavsar**, Hon'ble Head of Computer Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. Alka Mahajan**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

- Gondalia Ankita Dineshbhai

17MCEC06

Abstract

Efficient usage of resources is one of the major concerns of any organization. In the recent past years, globetrotting has increased a lot and it has to be handled properly. This project mainly aims at identifying the future expenses for the organization. The past some years data is processed and prediction for the future expenses will made using this project. This data is then processed as per user requirements to retrieve relevant information from the entire available data-set by using data mining and is represented to the user in the form of indicators which are basically graphical representations of the processed information. The users can further analyze data from the indicators in order to take relevant business decisions. Thus this helps in efficient project planning.

Abbreviations

ARIMA	Auto-Regression Integration Moving Average.
ARMA	Auto-Regression Moving Average
MAPE	Mean Absolute Percentage Error
RMSE	Root Mean Square Error

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

Nowadays Globetrotting is becoming a very important for the MNCs. It can be for any reasons like Training, Meetings etc. For the same budget is allocated in the organization and according to that it is utilized by keeping in mind the same. For Expenses can vary with time and demand as a result of which we cannot predict the expenses to be done for the particular. So here by taking into consideration past years data we are going to have such thing which can predict the any expenses like flight, train, hotel, food etc. for the given date. As a result of which they can proper utilization of the budget allocated so that they do not make loss in their name.

In this an approach based on analyzing past year data and predicting the future from the same is done. In our experiment we will be using different data analysis and prediction algorithms and will compare their performances and from that they can make their appropriated decisions. Decisions are made on considering factors such as in some year say for example like during most of globetrotting happens in July month so for that quarter they can expect for expenses and act accordingly etc. So this project will be offering the most comprehensive solution to streamline business processes.

1.2 Problem Statement

The problem statement here is that we will be using past 3-4 years data for analysis. Apply some analysis algorithms and check the accuracy. By doing this we will be predicting the next years expenses so that they can improve their expenses accordingly in the required fields and departments. So initially we will be evaluating the particular characteristics of expenses for the department quarter wise and then year wise and then make decisions.

1.3 Characteristics

- It will be having processed data in graphical representation which is easy to analyze.
- You can have your filters for analyzing the data.
- It can predict future expenses of the given department for said quarter/year.
- It can also predict flight, food, accommodations etc. expenses for the specified dates in the specified places.
- Helps in keeping track of all expenses happening for the particular department in said month/quarter/year.

1.4 Purpose

The main purpose of this project is to have efficient use of the available resources in the organizations. So as to make important strategic plans appropriately and controlling the organizations over all activities. It will also help in coordinating the various operations of the different departments. From the generated reports the user can actually take proper decisions. It will help in making quarterly decisions as well as yearly also. Not only globetrotting details but also it will help in knowing accommodation details, food etc. from the previous years data. So this project will help in streamline the business needs.

Chapter 2

Literature Survey

2.1 Techniques : Time Series Analysis

A Time-Series is a period stepped informational collection in which every datum point compares to an arrangement of perceptions set aside a few minutes example. Time-series analysis accept the framework to be a black-box we simply endeavour to anticipate what is coming dependent on the past standards of conduct.

Time series data are analyzed by considering the following characteristics:

- **Trends**

  A trend exists when there is a long haul increment or diminishing in the information. It doesn't need to be straight. Once in a while we will refer to a trend, when it may go from an expanding pattern to a diminishing pattern.

- **Seasonal**

  A seasonal example exists when an arrangement is impacted via regular components (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a settled and known time period.

- **Cycle**

  A cyclic pattern exists when information display rises and falls that are not of settled period. The span of these vacillations is more often than not of somewhere around 2 years.

- **Auto-correlation**

  Here next value is dependent on the past value.

1. **Time Series Algorithms:**

   (a) **Classical Decomposition + ARMA :**

   Consists of two models : Additive Model and Multiplicative model. If the variation in the trend cycle dose not with time series additive model is preferred. However if the trend cycle varies proportional to level of time series the multiplicative model is preferred.

   Also when the multiplicative is going to be used we can transform the data series to remain stable over time period and once it is done, we can use additive model instead of using multiplicative model.

   To find the order of AR(p) and MA(q) of the stationary series obtained after the removal of the trend and seasonality component, we use ACF and PACF plots.

   Auto relationship Function/ACF. ACF is a plot of aggregate relationship between's various slack capacities. For example, stock prices at time point t is x(t). We are keen on the connection of x(t) with x(t-1) , x(t-2), etc.

In a moving normal arrangement of slack n, we won't get any relationship among's x(t) and x(t - n - 1) . Henceforth, the aggregate relationship graph cuts off at nth slack. So it ends up easy to discover the slack for a MA arrangement. For an AR arrangement this relationship will bit by bit go down with no cut off esteem.

Here is the second trap. In the event that we discover the incomplete relationship of each slack, it will cut off after the level of AR arrangement.For instance,if we have an AR(1) arrangement, in the event that we prohibit the impact of first slack (x (t-1)), our second slack (x (t-2)) is autonomous of x(t). Subsequently, the incomplete relationship work (PACF) will drop strongly after the first slack.

(b) **Auto ARIMA :**

Auto ARIMA stands for AutoRegressive Integrated Moving Average. Auto Regressive (AR) terms invokes to the slacks of the differenced arrangement, Moving Average (MA) terms invokes to the slacks of errors and I is the quantity of distinction used to make the time series stationary.

Assumption for ARIMA model :

- Data should be stationary and univariate.

Steps for following the ARIMA modelling:

- Exploratory analysis
- Fit the model
- Diagnostic measures

2. **Linear Regression:**

Linear Regression is basically said to be predictive analysis method.The general idea of regression is to look at two things: (1) does an arrangement of variable affects admirably in predicting variable? (2) Which factors specifically are huge predictors of the result variable ?

Equation for linear regression: y = b + ax

where y = estimated dependent variable , b = constant, a = regression coefficient, and x = score on the independent variable.

Chapter 3

Model Selection

- Selection of model plays a very important role in any operations. So for selecting the appropriate model there are some fundamental basics things that need to be considered.
- Hence the model selected for the prediction should be the simplest as possible. To check whether the model is simple, its complexity is also need to be considered.
- The more simple the model the less complex it is and vice-verse. Following points one can consider for checking the complexity:
  - Parameters need to for specifying the model.
  - Degree of function.
  - How depth the decision tree is
  - Size of the model used. For example, in binary encoding of model bits needed is also important thing.
- The simpler the model, more generalize it is. Also as simple as the mode less training data will be required for the model.

3.1 Over-fitting:

- Over fitting occurs with when the model explains more about errors i.e. noise etc. rather than showing the relationship among the data.
- The more complex the model it is over-fitted. As a result of which it leads to poor generalization.
- Here the model that learns completely from the given data-set will simply just memorize the data instead of learning the patterns.
- As it simply memorizes all the data, it would give almost zero error during the training phase. It focuses more on the trained data so it is not able to make generalize decisions after the training data.
- So it will give correct results for the trained data but during the testing data it wont give that good results.
- Also for the data other than trained data it will simply do random selection as during it just memorizes the data instead of learning it. This is a very good example of over-fitting.

3.2 Bias- Variance Trade-off:

- Here we will consider the same example of memorizing the data-set as discussed in over-fitting. So if we make any small changes to the data-set the model needs to be changed.
- As a result of which it becomes unstable as it is sensitive to even a single change in the training data.
- So the simple model that finds out some specific patterns from the data-set in that if we add/ remove some data points will affect very less to the pattern.
- Variance of the model is the variation in the output while considering the test data w.r.t to training data changes.

- Bias tells us the accuracy of the model for the test data-sets. Models which are not trained properly with different data-sets are very bad in prediction for testing data-set.
- If it is appropriately trained with the different data-sets then it will give quite good predictions.
- So the model which is naive will most of the time gives same outputs for the test data in spite of seeing whether the question is same or different it will simply give one same answer as it is trained for the very less or hardly just one data-set.

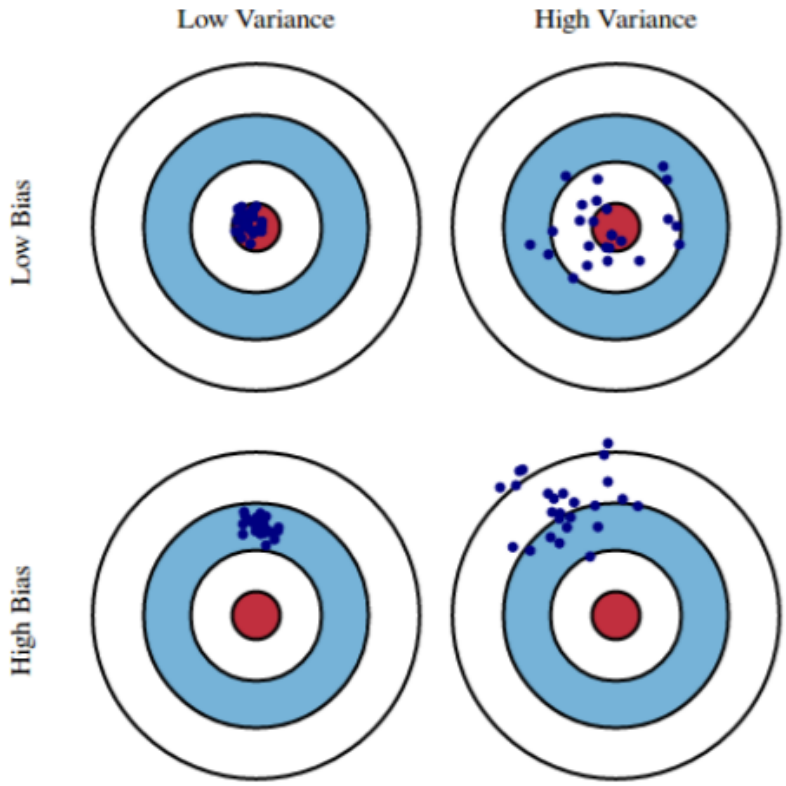

Figure 3.1: Example of Bias - Variance

- In the above figure gives the very good understanding of the Bias-Variance with shooting method. So the points that are more together (clustered together) have less variance. The ones which are at bulls eye have small variance.
- So in a brief we can say consistent shooter have small variance meaning it gives consistency where as the one who is good shooter is having small bias which means it gives the accuracy.

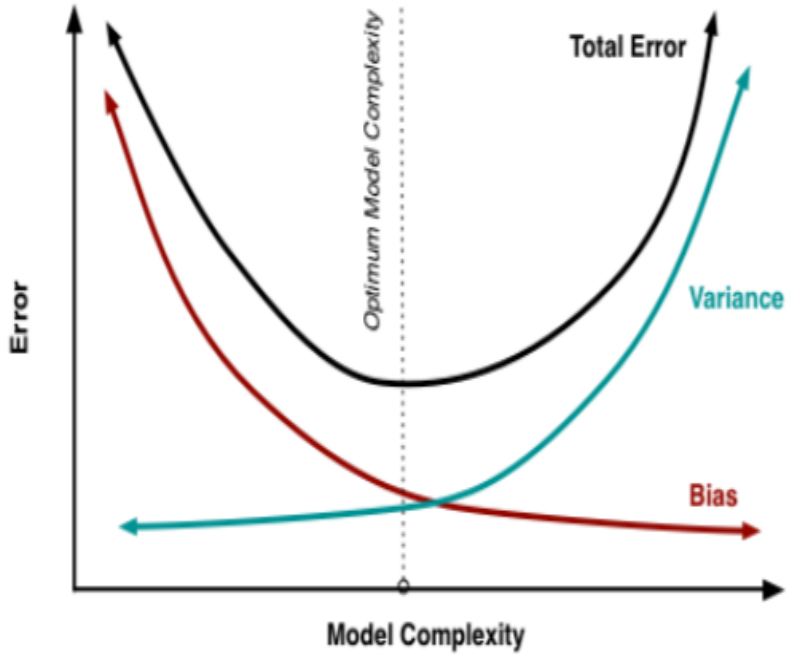

Figure 3.2: Bias - Variance Trade-off

- The figure here shows the complete trade-off between the both. The models with high variance and low bias are more complex models.
- And the ones with low variance and high bias are less complex models. Clearly from the graph one can see and tell naive models will not change after making some changes to data-sets and also its variance will remain zero with high bias.
- So the best model will be the one with the low variance and high bias as high bias will give the high accuracy whereas low variance will give the consistency to the model.

Chapter 4

Time Series Analysis

4.1 Approach:

To make efficient use of available resources and to have profit for the organization budget utilization is must. So here we are going to plan budget for the departments on the basis of quarters and years in order to have profit for the organization.

The objectives of the following analysis is:

- Find the region and department where the budget is used optimally.
- For the same will perform forecasting of the globetrotting budget for the next year.

The practical approach is divided as shown below:

Figure 4.1: Practical Approach

1. **Data Understanding and Loading:**

   Aggregation of past years data by quarter and year by department wise. Following are the major fields taken into consideration.

2. **Data Preparation and Analysis:**

   - Removal of outliers, null values, normalizing
   - Smoothing of data
   - Creation of derived variables. E.g. (budget allocated budget used) / budget allocated.

Data Analysis

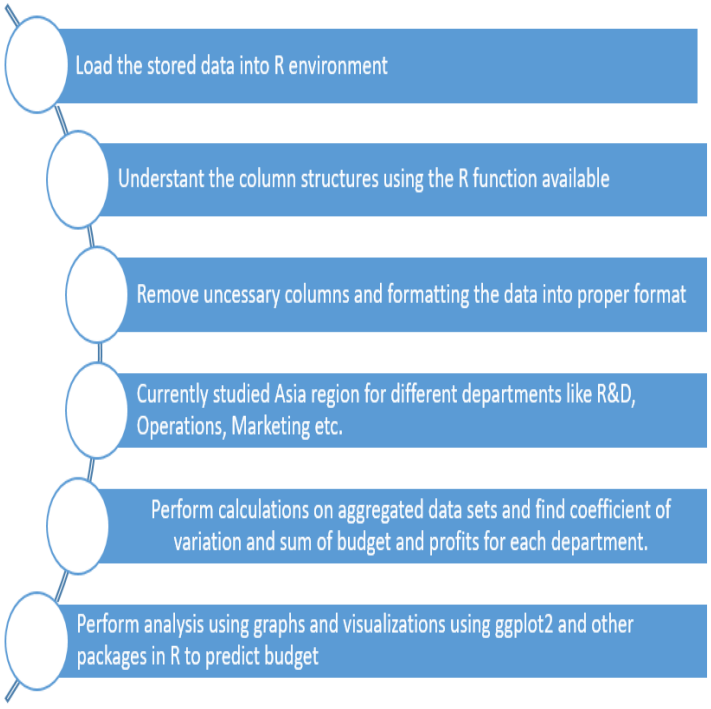

Figure 4.2: Data Analysis

3. **Exploratory Data Analysis:**

The Region attribute is a categorical variable which represents the geographical region in which Intel operates. Also, the Department attribute tells the high level departments in that region. Hence, the entire Budget is allocated is bucketed in 5*3 = 15 parts, such as APAC RD, US Operation etc.

Department
Operation
Management
R & D

Figure 4.3: Attributes for Department

Region
Africa
APAC(AsiaPacific)
US (United States)
EMEA(MiddleEast)
EU(EuropeanUnion)

Figure 4.4: Attributes for Region

4. **Time Series Modeling:**

We tried two different approach for modelling this time series data : Classical Decomposition, Auto ARIMA modelling. Below are the details for each of the approach:

- **Classical Decomposition Methods:**

  In this approach we remove the trend and seasonality from the original time series to get a stationary residual series that we model as an ARMA (p,q) process.

  To remove the trend and seasonality component we need to follow 3 steps:

  (a) Identify Additive or Multiplicative

  (b) Model the Time series data in terms of different components

  (c) Remove the Trend and Seasonality component

  (d) Finding the order of AR and MA

(a) Identify Additive or Multiplicative :

Identifying the whether the trend and seasonality can be modelled using additive or multiplicative model. For this we plot the data using ggplot2 library available in R. Below is the graph of time vs Budget Used for the APAC region and Operations Department.

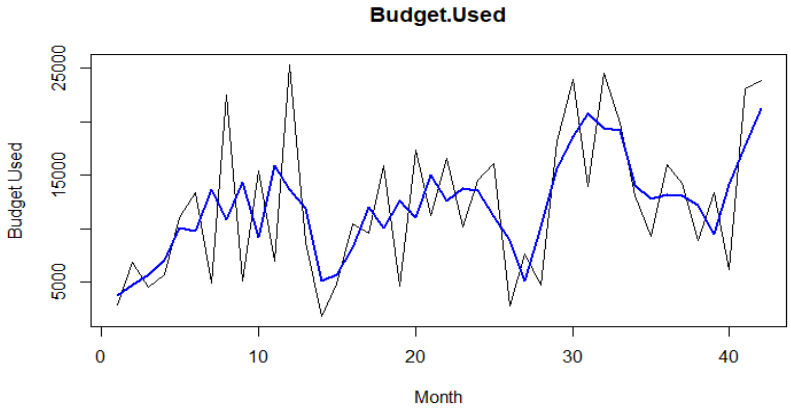

Figure 4.5: Smoothed Graph

From this we can see clearly see an increasing trend year on year in the budget used. Also we see seasonality component each year. The use of the resources decreases in the month of December each year.From the graphical representation of data we conclude that our data follows additive model for trend and seasonality.

(b)Identify Additive or Multiplicative :

Now after identifying that the series can be modelled in terms of additive model we try to separate the different components in the time series data. We tried to fit the data using linear model data.

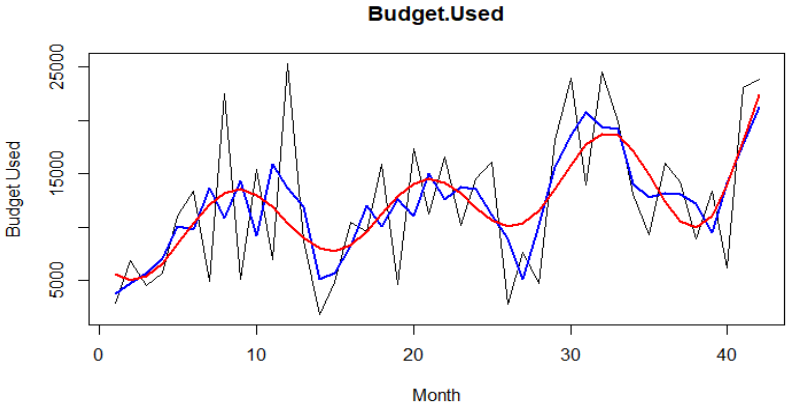

Figure 4.6: Decomposition

(c) Remove the Trend and Seasonality component:

We remove the Trend and seasonality component from the original time series for further analysis: What we get after removing is the residual series as below: resi ¡- timeser - trend

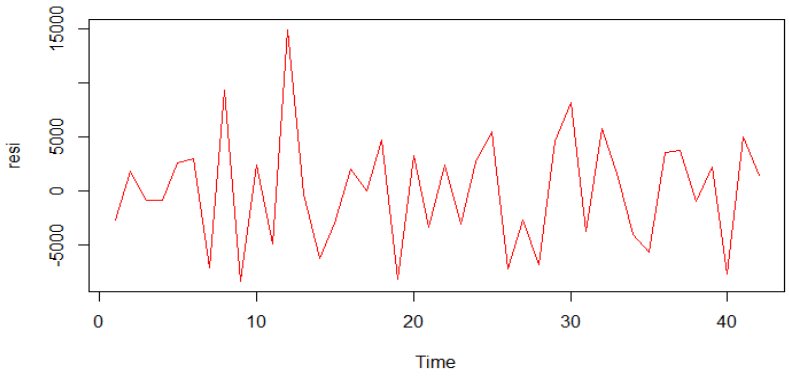

Figure 4.7: Residual Trend

Once we remove the seasonality and trend component from the original series we get a stationary series. We try to model the residual series using ARMA (Auto Regression + Moving Average). We use KPSS test to check if the residual series is not pure noise.

(d) Finding the order of AR and MA:

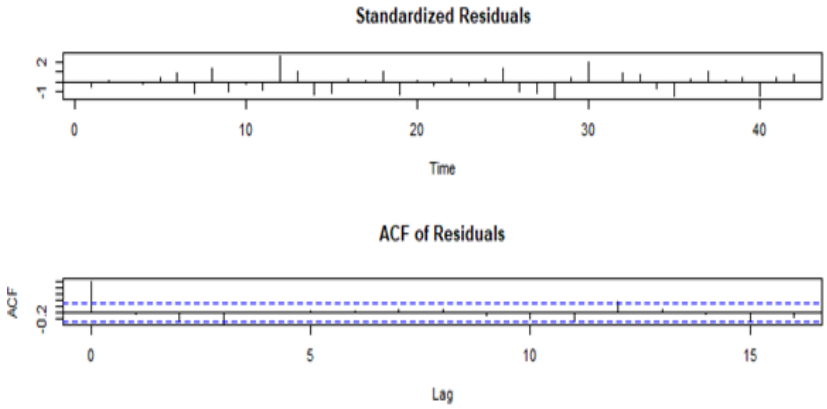

Figure 4.8: ACF Plot

As we can see in our ACF plot the order of MA is MA(0).

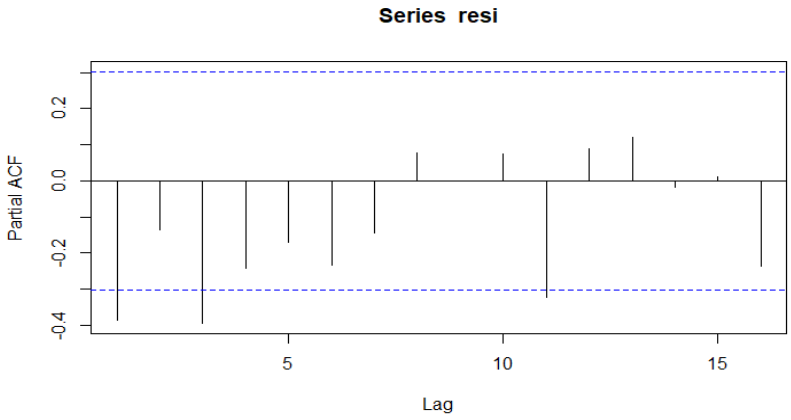

Figure 4.9: PACF Plot

In the PACF plot, it is coming here to be AR(0).

Forecasting:

We then forecast the budget which will be used for the next 6 months and compare the predicted test results with the actual available data.

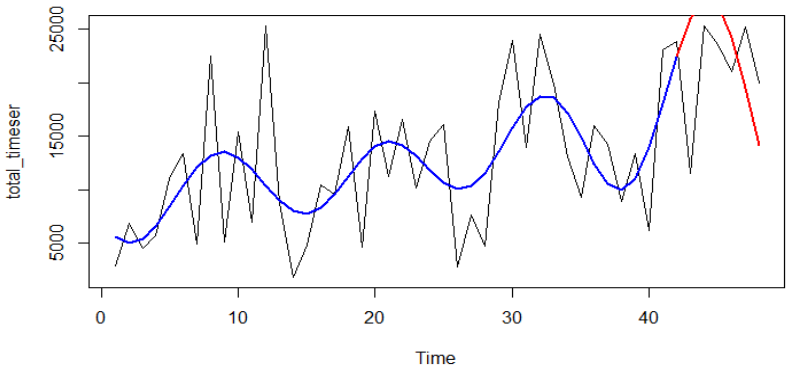

Figure 4.10: Prediction Graph

5. **Model Evaluation:**

   Evaluation of models was performed using MAPE (Mean absolute percentage error).The model with a lower MAPE value is selected out of the two models generated.

6. **Prediction:**

   Using the best model for predicting the next years budget as well as quarterly also and repeating the same for all the departments.

Chapter 5

Linear Regression

5.1 Introducion

It is basically predictive analysis method.Mostly used to study relationships among the continuous variables.One is the dependent variable and the other independent.The general idea of regression is to look at things: which factors specifically are huge predictors of the result variable? Equation for linear regression:

$$y = b + ax \tag{5.1}$$

where y = estimated dependent variable , b = constant, a = regression coefficient, and x = score on the independent variable.

Here in our case equation for the Linear Regression will be

$$BudgetAllocated = b0 + b1 * sales + b2 * GDP + b3 * year + b4 * quarter + b5 * CC \tag{5.2}$$

So simple linear regression tries to fit a straight line into given bunch of data. And using this straight line model one can predict the data for new data point available. A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

5.2 Best Fit Line

Among multiple straight lines present we need to find the best line which fits our model.One of the way for achieving this is to minimize the sum of the squared predicted errors i.e. least squared method

5.2.1 RSS, TSS and R-squared

RSS stands for Residual Sum of squares. RSS is the sum of the squared difference between the actual and the predicted values.If RSS=0 then it is the Perfect model.If units changes, its value changes since it has units of y^2.

Equation for RSS is given by:

$$e_i = \Sigma(y_i\text{-}y_{ipred})^2 \tag{5.3}$$

TSS satnds for sum of squared difference between mean actual value.

R-squared = 1 (RSS/TSS). R-square tells how better our regression model is

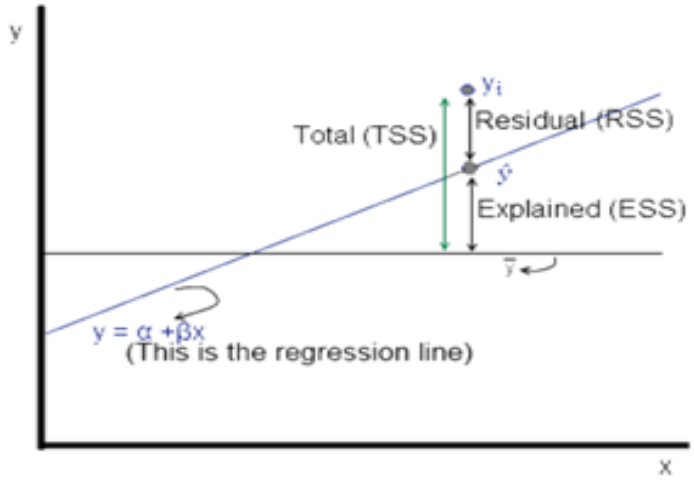

Figure 5.1: Comparison of RSS, TSS
ESS

5.3 Our Approach:

Figure 5.2: Used approach

5.4 Simple Linear Regression Model in R

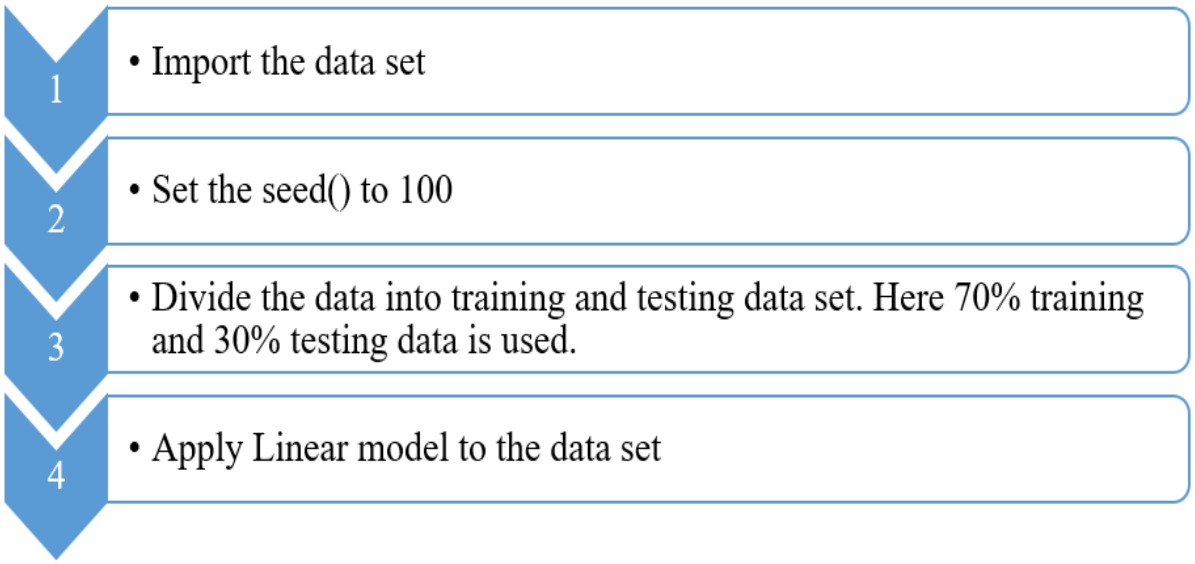

Figure 5.3: Simple Linear Regression in R

5.5 Libraries used in project

We have used following libraries to do our initial analysis and see the correlation between the different Independent variable and the dependent variable(Budget Allocated)

1. **ggplot2**

   ggplot2 is a data visualization package for the statistical programming language R.

2. **readr**

   readr is to provide a fast and friendly way to read rectangular data (like csv, tsv, and fwf). It is designed to flexibly parse many types of data found in the wild, while still cleanly failing when data unexpectedly changes.

3. **lubridate**

   Functions to work with date-times and time-spans: fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects.

4. **corrgram**

   Helps us visualize the data in correlation matrices.

5.6 Outcome of the experiment

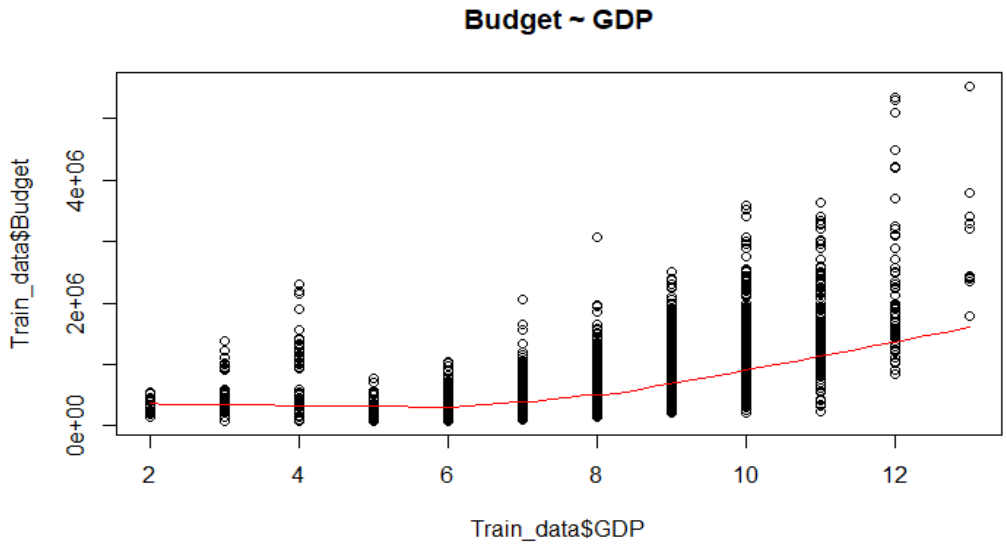

Figure 5.4: Comparison of Budget v/s GDP

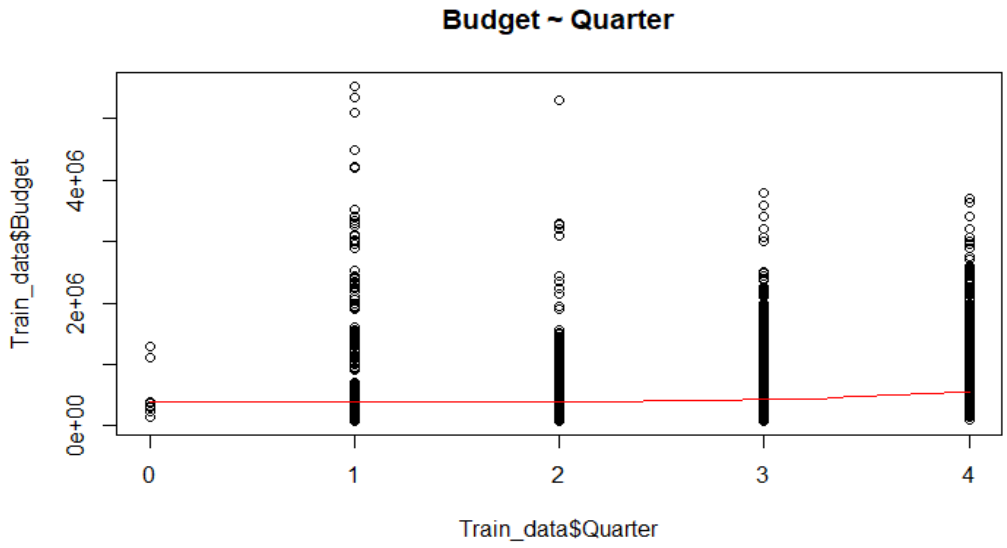

Figure 5.5: Comparison of Budget v/s Quarter

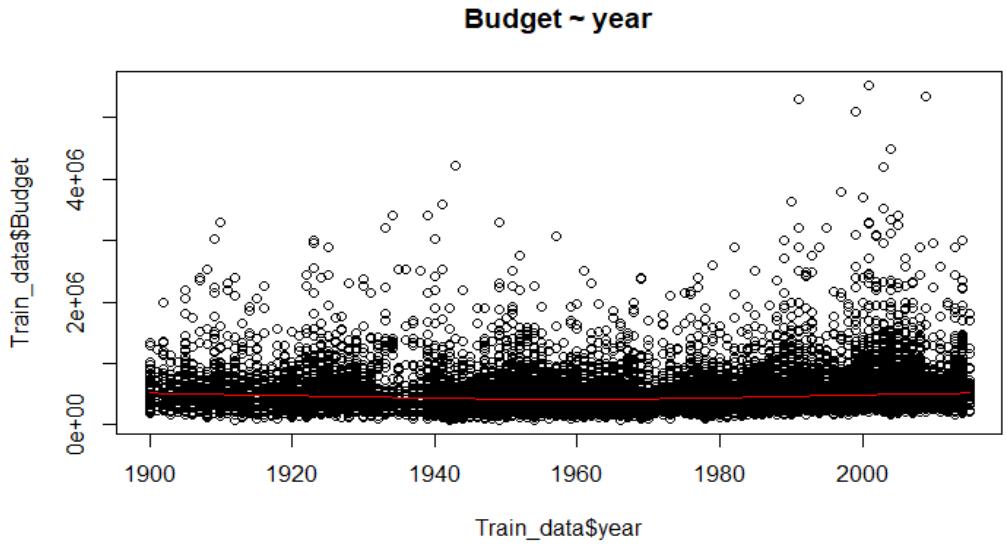

Figure 5.6: Comparison of Budget v/s Year

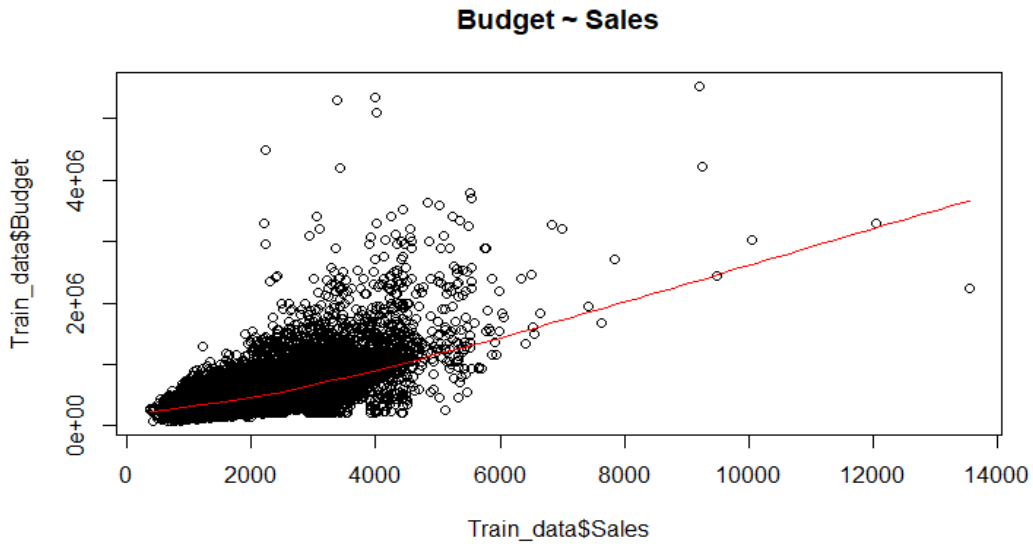

Figure 5.7: Comparison of Budget v/s Sales

Chapter 6

Smoothing and Auto - ARIMA

6.1 Smoothing

Smoothing is the process that makes curve smoother by taking out the average of noise for making trend and seasonality component apparent. In time series, it helps in better analyzing the patterns, trends etc. In short smoothing is the technique of removing the noise and making the time series curve more smoothed and clear.

There are many techniques to perform the smoothing out of them two are: (1.) Moving Average Smoothing and (2.) Exponential Smoothing.

1. **Moving Average Smoothing**

   - Here series is expected to be in the form of f(t) + e where e is the zero mean noise and f(t) is the predicted behaviour of the time series. So as the name says average so here averaging the time series values resultant series should be somewhat closer to the original f(t).
   - In Moving Average Smoothing, the values in the series are replaced by the weighted average of the values in the pre-defined window for the values.

- So from original series X_i obtained smoothed series is
- Equation of Moving Average Smoothing is given below:

$$Y_t = 1/C \sum_{i=-q}^{q} a_i * X_{t+i} \tag{6.1}$$

Where c, a_i are constants.

- Such kind operators are called as filters. Generally, filters for moving average are low pass filters as they filters out the noise having high frequency from the original time series.
- For e.g at a time instance t, having moving average window of 5 so the value will be the average of the values t-2, t-1, t, t+1, t+2.

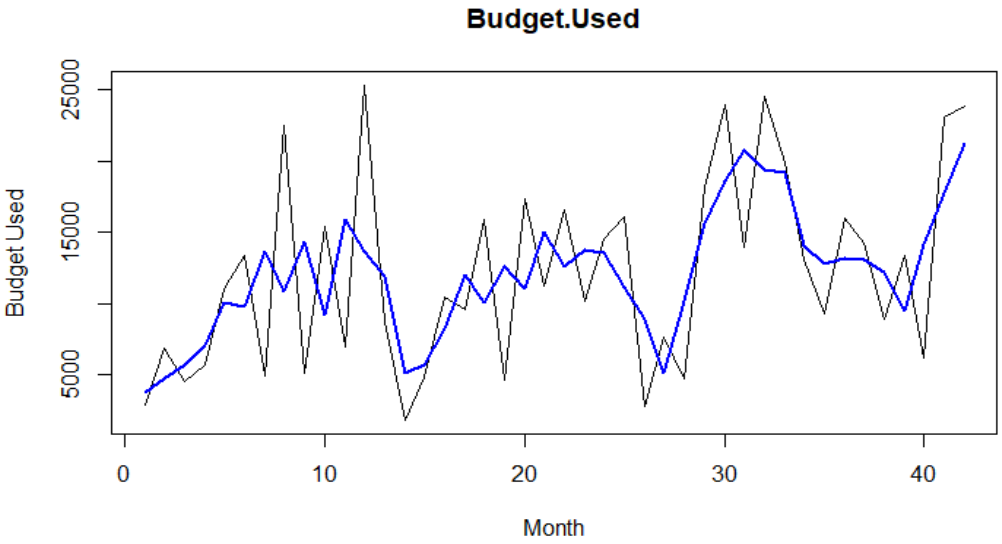

Figure 6.1: Moving Average Smoothing

- Here in the figure above black line indicates the actual time series values plotted and blue indicates smooth signal with window size = 3 using moving average smoothing.

- In our experiment, we tried with window size = 5 so the result obtained was the series losses its characteristics i.e. the curve became simpler and from that one cannot predict the appropriate data.
- Next we tried with window size = 1 over there it remained in its original form only, not much smoothing was obtained. So then we tried with window size = 3 and this met our expectations.

2. **Exponential Smoothing**

   In Exponential Smoothing, current value in the time series is replaced by the weighted sum of current value and previously smoothed value.

$$Y_t =_t +(1\text{-}\alpha)Y_t 1, Y_1 = X_1 \tag{6.2}$$

   Smaller the value of higher the level of smoothing and higher the value of (near to 1) smoothing will not occur to its best.It is also named as Holt-Winters Smoothing.It act as low pass filters.

   Used with univariate data. Whatever the forecast is obtained with help of this method is nothing but the weighted avg. of past data with weights getting decreased exponentially as the data becomes old.

   There are mainly three types of exponential smoothing present depending upon trend, seasonality, error. Simplest one has no specific structure, next one handles trend and the advanced one has support for the seasonality. They are:

   (a) **Single Exponential Smoothing:**

   - Single Exponential Smoothing is the method for the univariate data having no trend and seasonality component.
   - It contains only single parameter alpha(α) known as smoothing co-efficient or smoothing factor.

- The value of alpha is between 0 and 1. Larger the value of alpha denotes the recent past observations whereas smaller the value of alpha denotes the quite old observations are taken into consideration for making the prediction.
- The value near to 1 denotes fast learning (mainly recent values affects here) and near to 0 denotes slow learning (mostly old values affects here).

(b) **Double Exponential Smoothing:**

- It is an extension to the previous smoothing. Here one more factor other than alpha is also present i.e. beta factor. It controls the impact of the changing trend.
- The trend changes in following different way: additive trend multiplicative trend which depends on trend whether it is linear or exponential respectively.
- Additive trend is called as Holts linear trend model.
- Additive trend refers to the linear trend whereas multiplicative trend refers to exponential trend.
- It is useful in diminishing the trend with respect to the time. And this is referred to as dampen.
- Additive Dampening will dampen trend linearly where multiplicative will dampen it exponentially.
- For controlling the rate of dampening there is damping coefficient called as phi (p).

(c) **Triple Exponential Smoothing:**

- It is an extension to the single exponential smoothing. It has a support for the seasonality component of the univariate series. This method is also known as Holt-Winters Exponential Smoothing.
- Here one more new parameter is added other than alpha beta i.e. gamma. Which handles the impact for the seasonality.

- Like as trend, seasonality also comes as additive seasonality and multiplicative seasonality.
- Additive seasonality refers to linear seasonality where multiplicative refers to exponential.
- It is the most advanced among all exponential smoothing and it can be also used for developing the single and double exponential smoothing.

6.2 ARIMA

- ARIMA stands for Auto-Regressive Integrated Moving Averages.Below assumptions need to be taken before applying ARIMA model :

  1. Time series should be stationary.
  2. It should be univariate.

- Mainly used for time series forecasting.
- AR invokes to the slacks of differenced arrangements.
- MA invokes to the slacks of errors.
- I refers to the number of times differentiation performed in order to make the time series stationary.
- For checking whether the time series is stationary various tests are performed like Turning Point Test, Ljung Box Test, etc.

- ARIMA includes following steps for implementation:

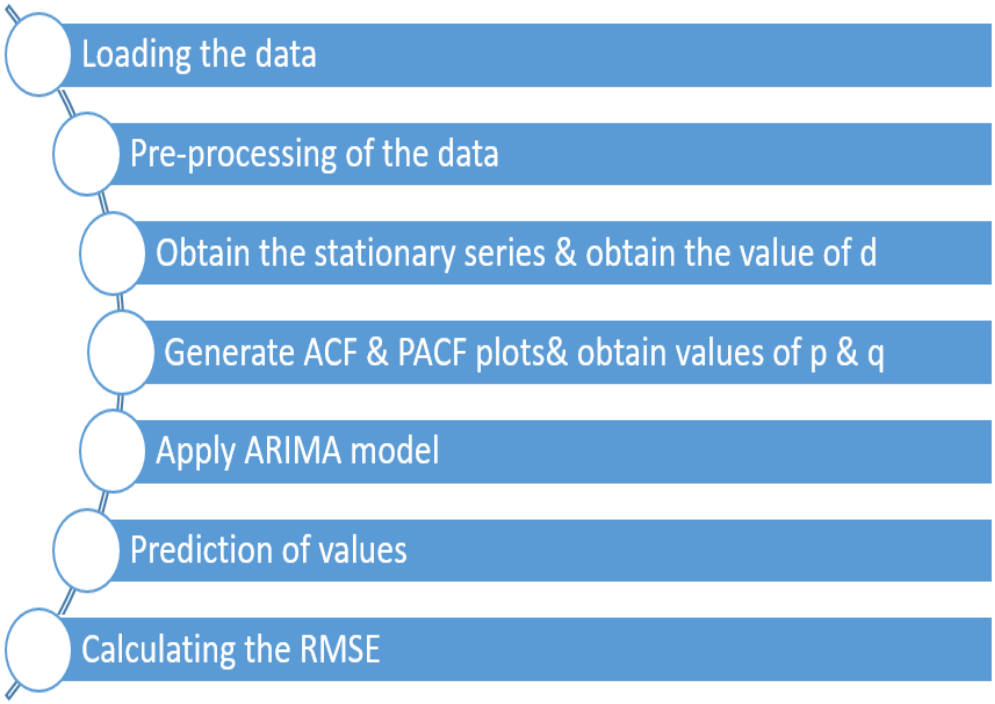

Figure 6.2: Steps for implementation of ARIMA

6.3 Auto - ARIMA

- ARIMA is one of the powerful model for the time series forecasting data. But it is very time consuming process as firstly we need to make the series stationary, then find the values of p q from the ACF PACF plots, determine the value of d etc. All these things are eliminated in Auto-ARIMA making it simpler.

- Steps included for Auto-ARIMA are:

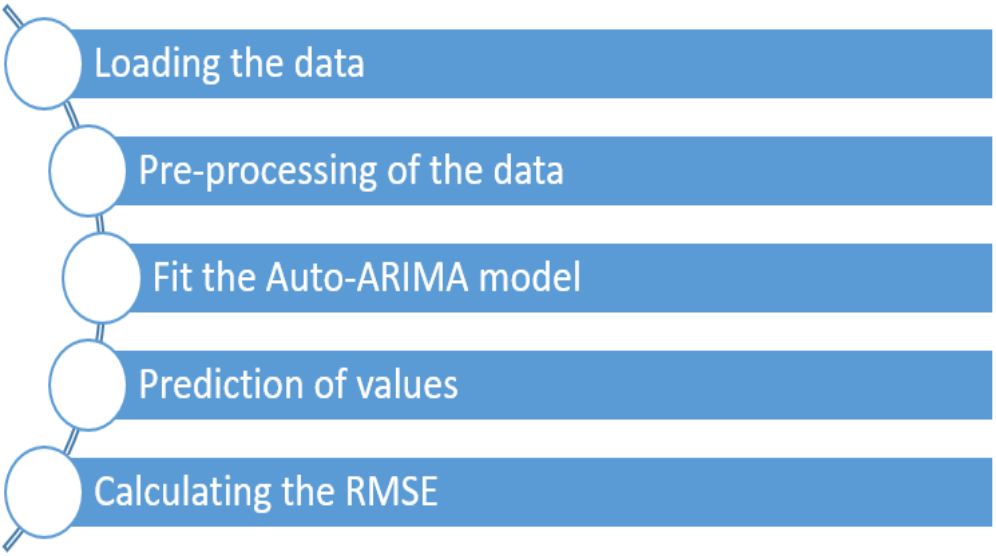

Figure 6.3: Steps for implementation of Auto - ARIMA

1. **Loading the data:**
   - Here the data in the form of .csv, .xls, etc. is loaded into the system.
2. **Pre-processing of the data:**
   - Pre-processing includes cleaning of the data. Obtained data comes from various countries having different date and time formats. So here all the date and time are converted into single format.
   - Also null values are removed from the data set. Unnecessary columns etc. are also removed. So in short here cleaning of data occurs.
   - Stationarity test is performed to check whether the series is stationary or not so that next steps can be taken accordingly.
3. **Fit the Auto-ARIMA model:**
   - After obtaining the stationary series, Auto ARIMA model is fitted to the data.

4. **Prediction of values:**
   - We predict the independent variable like sales, budget, price using the trained model.
5. **Calculating the RMSE:**
   - RMSE stands for Root Mean Square Mean.
   - It gives the model performance using the predicted value and the actual ones.

- So as we see from the steps of both the methods, steps from 3 to 6 are eliminated in Auto- ARIMA method.As a result of which much of the time is saved.

6.4 Outcome of the experiment:

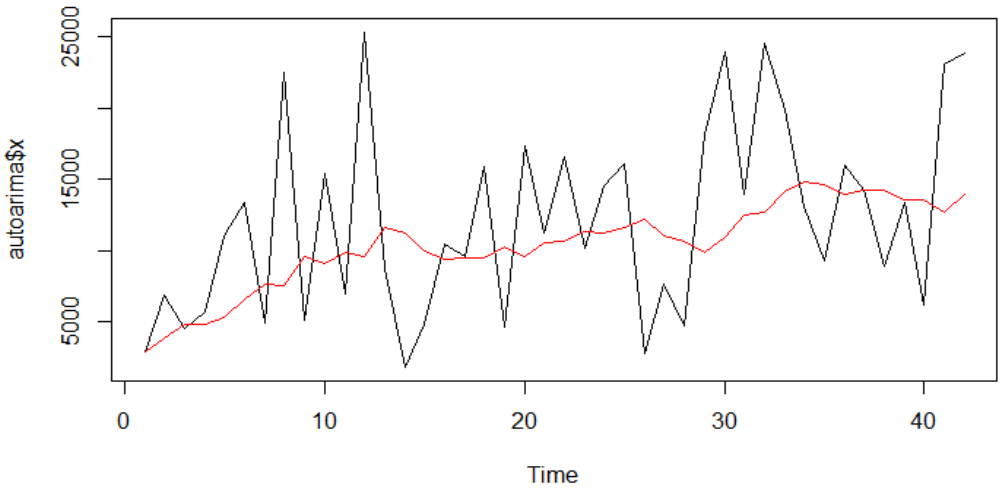

Figure 6.4: Auto - ARIMA

- In the above graph black indicates the actual values and the generated time series with that values whereas red indicates predicted values from the training data set available.

- In the above graph black indicates the actual values and the generated time series with that values whereas blue indicates predicted values from the training data set

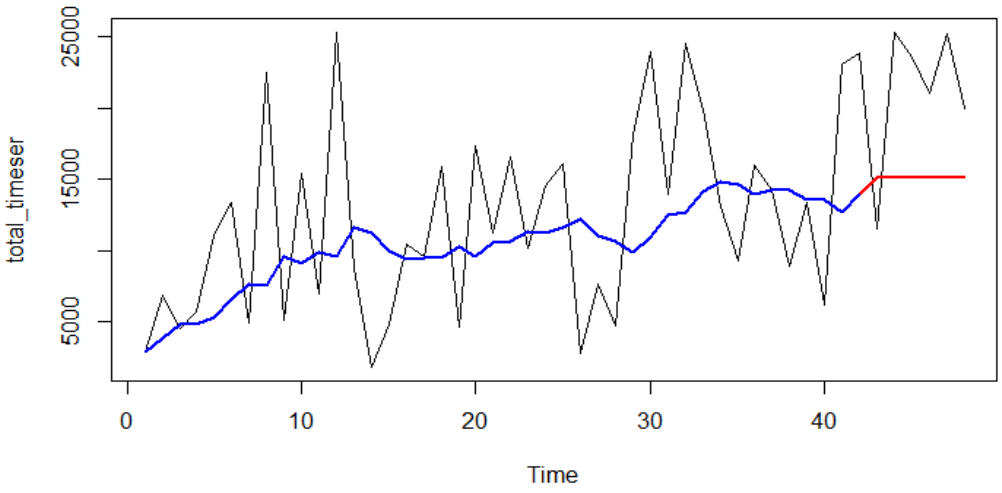

Figure 6.5: Prediction using Auto - ARIMA

available and red indicates the predicted values from the testing data set available. From the given data set, 70 % of the data was used for the training purpose whereas 30 % was used for the testing purpose.

Chapter 7

Stationarity Test

7.1 Introduction

- A series whose mean, variance, autocorrelation, etc. remains constant with time are said to be stationary series.
- Most of the forecasting methods are developed by taking into consideration that the time series is stationary then only one can apply that forecasting method to it.
- If the series is not stationary then it is first converted into the stationary series and then further models are applied according to the requirements.
- Stationary series obtained through differentiation plays an important role in fitting series into ARIMA model.
- For most of the cases in ARIMA modelling, the series is made stationary by performing the differentiation. Here in the definition of ARIMA I invokes to the number of times the series is differentiated to obtain the stationary series.
- Also stationary time series gives important statistics like mean, variance, etc. with respect to the other variables. This helps one to figure out their behaviour in future.
- For example if the series is increasing w.r.t time than the mean, variance etc. also shows increasing effect in the future.

7.2 Check stationarity of the series

To check whether the series stationary or not there are various ways. One can check by looking at the plots, from the statistics information, statistics tests etc. and predict whether the series is stationary or not.

- Using Plots:

  One can plot data of time series and check manually whether trend, seasonality etc. components are present in the plot.

- Statistics information:

  One check the statistics information of the given data on given time period, for particular months, etc.

- Statistical Test:

  There are some tests available with the help pf which can check the stationarity of the series.

7.3 Different types of the tests to check whether the series is stationary or not:

1. **Ljung-Box Test:**

   The Ljung-Box test is mainly used to check whether the autocorrelation exists in the time series or not.

$$Q = n(n+2)\sum_{j=1}^{h} \rho(j^2)/(n-j) \tag{7.1}$$

   Where n is the number of observations, (j) is the coefficient of auto-correlation. The equation returns the graph of p-values having lag j. Below is the graph generated from our data.

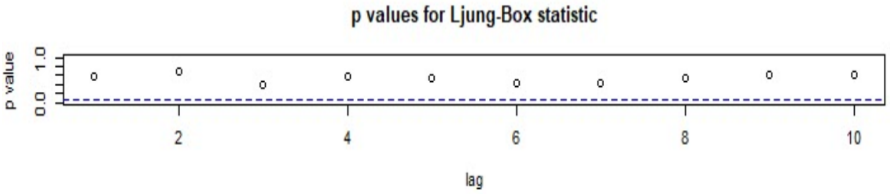

Figure 7.1: Ljung Box

2. **Turning Point Test:**

   Here in the series it is expected that the values goes up and down with given time. So a turning point says the change in directions. For example, a series having values $X_1, ..., X_n, X_t$ is said to be a Turning Point if and only if X_i ¿ X_i-1 and X_i ¿ X_i+1 or X_i ¡ X_i-1 and X_i¡ X_i+1.

   For iid series having large value of n, turning point follows the Gaussian distribution having mean and variance is given as

$$\mu_T = 2(n-2)/3, \sigma_T^2 = (16n-29)/90 \tag{7.2}$$

3. **Runs Test:**

   It is almost similar to Turning Point test as discussed above. Defined as series having either increasing values or only decreasing values. Increasing or decreasing values defines the length of the run test. Let n1 be the number of times increased value and let n2 be the number of times decreased value. So n1 + n2 = n. For the large values of n, r following Gaussian distribution having mean, variance etc. is given by

$$\mu_T = 2n_1n_2/n + 1, \sigma_r^2 = (\mu_r - 1)(\mu_r - 2)/(n-1) \tag{7.3}$$

4. **KPSS test:**

   KPSS stands for KwiatkowskiPhillipsSchmidtShin.This test helps us in figuring out whether time series is stationary w.r.t mean or linear trend component or it is non-stationary for a unit root. It is mainly based on linear regression. It consists of regression equation having 3 parameters namely deterministic trend (β_t), random walk (r_t) and a stationary error (ϵ_t).

$$X_t = r_t + \beta_t + \epsilon_t \tag{7.4}$$

   So here we are having the test for level stationarity. Here smaller the p value it indicates non-stationarity.Following is the result obtained after applying the KPSS test to the dataset

```
kpss.test (resi)

KPSS Test for Level Stationarity

Data:  resi
KPSS Level = 0.047444, Truncation lag parameter = 3, p-value =
0.1
```

Figure 7.2: Results of KPSS Tests

Chapter 8

Future Work and Conclusion

We plan to deploy the model to AWS cloud and provide a dashboard for the upper management to take decisions based on the graphs. We also plan to try out the same model for different departments and look at the accuracy.

Below table shows the comparison of accuracy's obtained from the different methods.

Method Used	Accuracy
ARIMA using Classical Decomposition	0.8754
Linear Regression	0.725
Auto - ARIMA	0.6521

Figure 8.1: Conclusion of using all methods

Bibliography

[1] Harshita Tanwar and Misha Kakkar, *Performance comparison and future estimation of time series data using predictive data mining techniques*, IEEE, 2017

[2] Li MING and Jing-xian LIU, *Prediction of the amount of vessels arriving at inland port based on time series analysis*,

[3] Selva Priya S and Lavanya Gupta , *Predicting the Future in Time Series using Auto Regressive Linear Regression Modeling*,

[4] http://www.statslab.cam.ac.uk/ rrw1/timeseries/t.pdf

[5] http://www2.hawaii.edu/ fuleky/econ427/6Timeseriesdecomposition.html

[6] https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/

[7] https://cdn.upgrad.com/UpGrad/temp/a49f0fa3-4440-430e-944b-42e14637a71f/Lecture