9/28/2020

Airline Passenger Satisfaction Prediction

Capstone Project: Final Report

Tahmid Bari

PGP IN DATA SCIENCE AND BUSINESS ANALYTICS (GREAT LEARNINIG) - THE UNIVERSITY OF TEXAS AT AUSTIN (MCCOMBS SCHOOL OF BUSINESS)

TABLE OF CONTENTS

Introduction

1.1 Defining Problem Statement

This is the dilemma of a reputed British airline carrier 'Virgin Atlantic'. They aim to determine the relative importance of each parameter with regards to their contribution to passenger satisfaction. Provided is a random sample of 90,917 individuals who travelled using their flights. The on-time performance of the flights along with the passenger's information is published. These passengers were later asked to provide their feedback on various parameters related to the flight along with their overall experience. These collected details are made available in the survey report as in (csv) labelled file 'Virgin_Atlantic_Satisfaction'.

The service sector has grown at a phenomenal rate. The last 15 years have seen a dramatic upsurge of interest in services, as academics and practitioners alike have realised the profound structural shift toward services in every advanced economy. This increasing interest in services is not surprising when one realises that services now account for over 74% of the United States' GDP and the percent of employment in the service sector has grown in every developed country in the last 25 years. Similar increases are seen in all of the industrialised countries of North America, Asia, Europe and Australia. Coinciding with the explosive growth of this economic sector, increasing emphasis has been placed on the continued development of knowledge related to service organisations, particularly the role service quality plays in creating a satisfied and loyal customer.

Customer service has become a major area of interest for both practitioners and academics. The managerial press extols the critical role of providing quality service, and academics are struggling with the problems of measuring and understanding how customers form service evaluations. The importance of service quality in any service industry cannot be disputed. Recent political, economic, and technological changes affecting the transportation industry in particular have made service quality a major concern for airlines and passengers.

There are many evolving changes in world transport. These changes have taken many forms, for example:

- The global economy: it is expected that three parts of the world; America, Asia, and Europe will dominate the commercial arena in the future. Therefore, service companies like airlines should be where the business is, and, more importantly, they should be able to offer customers what they want."
- Companies are developing into global organisations: either as multinationals or by forming strategic alliances. Airlines are not isolated from this development. The existence of a monopoly governing domestic air routes is no longer beneficial, and the rules are changing. Therefore, competition is expected to be especially harsh within the business sector where price is concerned. Airlines that survive must be able to deliver a total travel programme (i. e. offering non-stop flights, hotels, leisure programmes, financial services, etc.) through a world-wide information system. It is believed that passengers are now looking for more than cheap tickets; they ask for more comprehensive services.

Excellent service is a profitable strategy because it results in more new customers, more business with existing customers, fewer lost customers, more insulation from price competition and fewer mistakes requiring the performance of services. It also results in lower marketing costs because extra marketing money does not have to be spent convincing customers to buy despite the firm's poor service record passengers'

Passengers' expectations concerning the quality of service they receive have increased in recent years, and airlines are working very hard to meet these 2 expectations. This means that airline management must have a good understanding of the ways in which passengers service quality.

In recent years, managers have found themselves having to redefine their corporate philosophies in the face of foreign competition and rising production costs. One of the results of that soul searching was a realisation that domestic goods and services were, in many cases, shoddy by world standards. This led, in turn, to a commitment on the part of many firms to make quality their number one concern. However, one industry that has had a particularly difficult time embracing this concept is the airline industry. One reason for this is that air transport is not provided by the airline alone; it is really a joint effort involving the airlines and the government. Unfortunately, since the airline is the one entity the customer comes into direct contact with, it is, by default, blamed for the ills of the entire system despite management's best efforts and intentions. The biggest problem is that there is no consensus among the users and providers of air transport as to what quality means in the airline industry. This study examines the issue from the standpoint of the passenger, since it is considered the main element of "passenger-management-government" hence that affects the airline industry.

Efforts to understand consumer travel behaviour have become more dynamic in recent years. Currently with the increasing demand for air travel in Europe and the surrounding area, ‘Virgin Atlantic’ is confronted with the need to obtain better understanding of consumer air travel. To push for better service for air travellers appears to have some promise as a possible vehicle for improving consumer satisfaction. Thus, the airline's role in implementing this marketing goal is crucial.

Quality and satisfaction are extremely important concepts to academic researchers, particularly in-service marketing, and to practitioners as a means of creating competitive advantages. However, they have not been consistently defined and differentiated from each other in the literature. These inconsistencies result in conceptual difficulties and confusions which stunt the progress of theoretical development in the area. Customer satisfaction has long been a topic of interest in the areas of consumer behaviour, sociology and marketing, receiving a variety of interpretations and definitions from each respective discipline.

In the survey, the passengers were explicitly asked whether they were satisfied with their overall flight experience and that is captured in the data of survey report under the variable labelled ‘Satisfaction’. The objective of this exercise is to understand which parameters play an important role in swaying a passenger feedback towards the positive scale. We are expected to predict whether a passenger was satisfied or not given the rest of the details are provided. In addition, the role of service quality in passenger satisfaction will be examined to determine the linkage between these concepts.

It is important to make sure that passengers have a rich experience every time they travel. The satisfaction survey from passengers, which is a combination of categorical and continuous variable has been used for this study. This study seeks to not only explain the paramount factors which impact the passenger satisfaction in the British Airline industry but also change in those factors across different age groups.

1.2 Need of The Study/ Project

In our scenario, the airline company wants to identify a customer satisfaction level, based on his rating on various aspects of airline experience. Hence, primarily we build a model to classify the customer satisfaction level. Precisely we will classify customers' being satisfied or not and accordingly try to find out the factors related to high satisfaction.

- Intangibility: Services are often intangible; they lack precise form which makes testing them for quality in advance of sale impossible.
- Variability: Services are heterogeneous, which means that performance often varies across time, location and customer.
- Inseparability: Production and consumption of the services occur simultaneously, which makes the consumer an integral part of the process.
- Perishability: Services cannot be stored for later use, e.g. airline seats cannot be reclaimed.

The unique characteristics of service contribute to the complexities involved in assessing and managing service quality; they complicate both the consumers' assessments of service quality and the providers' ability to control it. Service quality has been increasingly identified as a major factor in differentiating service offerings and building competitive advantage. Most services involve direct contact between the customer and the service provider. This means that in addition to task proficiency; interpersonal skills like courtesy, friendliness, tolerance, and pleasantness are important dimensions of quality, particularly in high contact services where front line employees have a major influence on customer satisfaction. It has been said that for every complaint a business receives, there are twenty-six other customers who feel the same way, but do not air their feelings to the company. One satisfied customer usually tells two or three people, while the dissatisfied customer tells ten or more people. Therefore, to improve service quality, one must listen to the customer, since quality is ultimately defined by customer perceptions. Also, companies must listen to the front-line service employees in order to understand what they see as important and how they perceive the customer.

Airlines are a major transport provider. A new competitive environment has been emerged where price wars, frequent flyer programmes and other innovative marketing initiatives have become the industry norm. Therefore, airlines have been forced to introduce service development and enhancement strategies to remain competitive. The development of the consumer orientated marketing concept by the airline industry has been a response to changed environmental conditions, from a seller's market to that of a buyer's market.

Airline passengers are becoming more sophisticated about flying and therefore have higher expectations. The homogeneity of airline services forces customer service quality to emerge as a principal factor in the design of a competitive strategy.

Therefore, the benefits of offering a quality service (e. g., increase in first time customer volume, repeat business, the ability to charge higher prices that yield better profit margins and a reduction of marketing effort) are worth striving for.

The main focus of the research will be concentrated on finding answers to the following questions:

1- What are the main factors (dimensions) that can be used to measure airline service quality?

2- What is the nature of the relationship between:

- airline service quality and passenger satisfaction.
- passenger satisfaction.

3- What influence does the quality of airline services have on passenger satisfaction?

4- Do passengers rom different demographic categories:

- view airline service quality differently?
- differ in their degree of satisfaction toward a specific airline?

5. Do passengers exhibiting different psychographic and lifestyle characteristics:

- view the airline service quality differently?
- differ in their satisfaction and dissatisfaction toward a specific airline?

1.3 Understanding Business/ Social Opportunity

The purpose of this project is to review some of the consumer behaviour literature that focuses on customer satisfaction. This will help to understand the nature of satisfaction and dissatisfaction and throw some light on the differences between satisfaction and service quality.

The first section reviews the nature of satisfaction and its definition. The second section identifies some of the consumer satisfaction models, mainly focusing on reviewing the literature that covers disconfirmation models and their components. This appears to be the central model for understanding how satisfaction emerges from the purchase process. The third subsection reviews gap model and explores the difference between service quality and satisfaction.

The purpose of this project is to determine the relative importance of each parameter with regards to their contribution to passenger satisfaction. These passengers were asked to provide their feedback at the end of their flights on various parameters along with their overall experience. The two objectives of this project are:

- To understand which variables, play an important role in swaying a passenger feedback towards ‘satisfied’.
- To predict whether a passenger will be satisfied or not given the rest of the details are provided.

The major concern of this project is to investigate the passengers' perceptions of the Virgin Atlantic Airlines service quality and its influence on their satisfaction. Therefore, the main objectives of this research are:

- To identify the main attributes (dimensions) of airline service quality.
- To identify the nature of relationships between service quality, passenger satisfaction.
- To investigate the influence of both service quality and passenger satisfaction behaviour, in particular to determine whether consumers actually purchase a ticket from an airline that has the highest level of perceived service quality or from one that they are most "satisfied" with.
- To identify the influence of demographic variables on passenger' perception of service quality.
- To determine whether consumer-based variables in such categories as activities, interests, opinions and lifestyle could be used for segmenting passengers on the basis of their service quality expectations.
- To suggest effective marketing strategies that can be offered, based on the analysis of the relationship between service quality, consumer satisfaction.

The success of an airline hinges upon its knowledge of its customers and its ability to devise marketing campaigns to suit the preferences of those market segments it chooses to target. This needs a careful identification of the most important attributes (dimensions) of their services that can satisfy passenger needs and an understanding of how to provide them in the best way to achieve passenger satisfaction.

This project paper will attempt to analyse service quality within the airline industry and to determine potential areas of improvement within the passenger / airline relationship. In addition, the role of service quality in passenger satisfaction will be examined to determine the linkage between these concepts. It will also provide Virgin Atlantic airlines with a better understanding of their passengers and will aid them in developing a strategy which can best serve the passengers and the Virgin Atlantic Airlines. These are the main objectives which should be achieved through this research. The objectives are highly related and indivisible. They are all concerned with examining three phenomena: the perception of airline service quality, passenger’s satisfaction.

Data Report

2.1 Understanding How Data Was Collected in Terms of Time, Frequency and Methodology

This data set have a 90917obs. of 24 variables. I'll be cleaning the data to fill column null value at "Arrival Delay in Minutes". Probably, I can use all features for my prediction. Also, can take conclusion stating what are the features that will be used on the heat map in Exploratory Data Analysis after this part.

2.2 Visual Inspection of Data (Rows, Columns, Descriptive Details)

The data we obtained from Kaggle represents British Airline Satisfaction data consisting of 90917 observations and 24 attributes. The Satisfaction level, which is our dependent variable, is represented as a factor ("Satisfied" and "Neutral or Dissatisfied").

Variable	Variable Description	Variable Value Level
ID	General ID	Numeric numbers
Gender	Gender of the passengers	Female, Male
CustomerType	The customer type	Loyal customer, disloyal customer
Age	The actual age of the passengers	Age (in years numeric numbers)
TravelType	Purpose of the flight of the passengers	Personal Travel, Business Travel
Class	Travel class in the plane of the passengers	Business, Eco, Eco Plus
Flight_Distance	The flight distance of this journey	Numeric numbers
DepartureDelayin_Mins	Satisfaction level of Departure/Arrival time convenient	Numeric numbers
ArrivalDelayin_Mins	Satisfaction level of ArrivalDelayin_Mins	Numeric numbers
Satisfaction	Airline satisfaction level	Satisfaction, neutral or dissatisfaction
Seat_comfort	Satisfaction level of Seat comfort	Rating: 0 (least) - 5 (highest)
Departure.Arrival.time_convenient		Rating: 0 (least) - 5 (highest)
Food_drink	Satisfaction level of Food and drink	Rating: 0 (least) - 5 (highest)
Gate_location	Satisfaction level of Gate location	Rating: 0 (least) - 5 (highest)
Inflightwifi_service	Satisfaction level of the inflight wifi service	Rating: 0 (least) - 5 (highest)
Inflight_entertainment	Satisfaction level of the inflight entertainment	Rating: 0 (least) - 5 (highest)
Online_support	Satisfaction level of online support	Rating: 0 (least) - 5 (highest)
Ease_of_Onlinebooking	Satisfaction level of online booking	Rating: 0 (least) - 5 (highest)
Onboard_service	Satisfaction level of On-board service	Rating: 0 (least) - 5 (highest)
Leg_room_service	Satisfaction level of Leg room service	Rating: 0 (least) - 5 (highest)
Baggage_handling	Satisfaction level of baggage handling	Rating: 0 (least) - 5 (highest)

Checkin_service	Satisfaction level of Check-in service	Rating: 0 (least) - 5 (highest)
Cleanliness	Satisfaction level of Cleanliness	Rating: 0 (least) - 5 (highest)
Online_boarding	Satisfaction level of Online Support	Rating: 0 (least) - 5 (highest)

2.3 Understanding of Attributes (Variable Info, Renaming If Required)

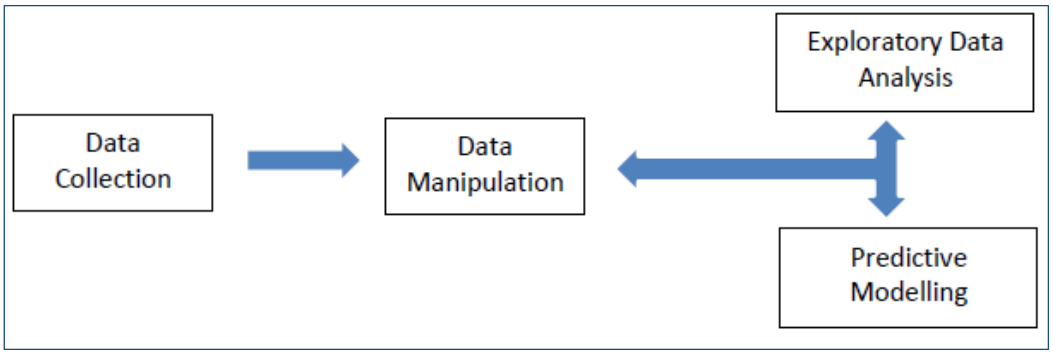


Table 1: Process flow of Exploratory data analysis

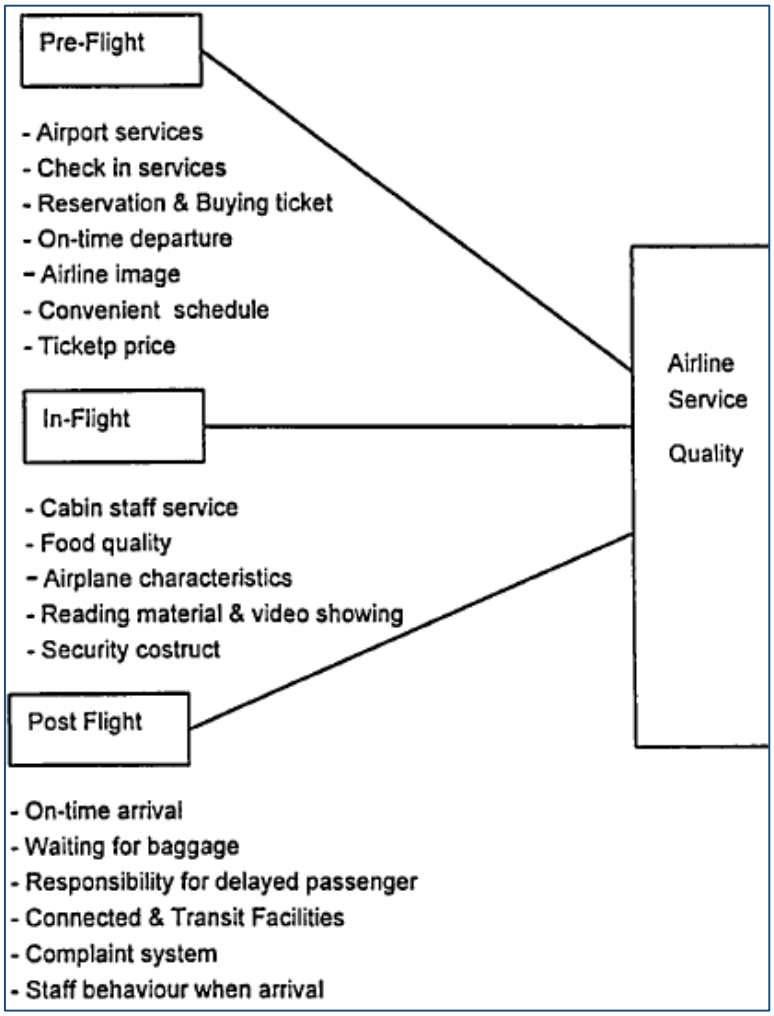


Table 2: Factors affecting airline services

We observe the dataset having 90917 observations and 24 attributes. The attribute “satisfaction” represents the satisfaction level of the customer on two different levels: “neutral or dissatisfied” and “satisfied”. Also, we drop the ID attribute as we wouldn’t require that in our analysis or model building. Hence, we have 1 dependent variable and 22 independent variables in our dataset. We check our dataset for any possible NA values, which must be dealt before we proceed with building the model.

```
> glimpse(Virgin)
Rows: 90,917
Columns: 24
$ ID                                <int> 11112, 110278, 103199, 47462, 120011, 1007...
$ Gender                            <fct> Female, Male, Female, Female, Female, Male...
$ CustomerType                      <fct> Loyal Customer, Loyal Customer, Loyal Cust...
$ Age                               <int> 65, 47, 15, 60, 70, 30, 66, 10, 56, 22, 58...
$ TypeTravel                        <fct> Personal Travel, Personal Travel, Personal...
$ Class                             <fct> Eco, Business, Eco, Eco, Eco, Eco, Eco, Ec...
$ Flight_Distance                   <int> 265, 2464, 2138, 623, 354, 1894, 227, 1812...
$ DepartureDelayin_Mins             <int> 0, 310, 0, 0, 0, 0, 17, 0, 0, 30, 47, 0, 0...
$ ArrivalDelayin_Mins               <int> 0, 305, 0, 0, 0, 0, 15, 0, 0, 26, 48, 0, 0...
$ Satisfaction                      <fct> satisfied, satisfied, satisfied, satisfied...
$ Seat_comfort                      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ Departure.Arrival.time_convenient <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...
$ Food_drink                        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ Gate_location                     <int> 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 1, ...
$ Inflightwifi_service              <int> 2, 0, 2, 3, 4, 2, 2, 2, 5, 2, 3, 2, 5, 4, ...
$ Inflight_entertainment            <int> 4, 2, 0, 4, 3, 0, 5, 0, 3, 0, 3, 0, 0, 0, ...
$ Online_support                    <int> 2, 2, 2, 3, 4, 2, 5, 2, 5, 2, 3, 2, 5, 4, ...
$ Ease_of_Onlinebooking             <int> 3, 3, 2, 1, 2, 2, 5, 2, 4, 2, 3, 2, 5, 4, ...
$ Onboard_service                   <int> 3, 4, 3, 1, 2, 5, 5, 3, 4, 2, 3, 3, 1, 3, ...
$ Leg_room_service                  <int> 0, 4, 3, 0, 0, 4, 0, 3, 0, 4, 0, 2, 3, 5, ...
$ Baggage_handling                  <int> 3, 4, 4, 1, 2, 5, 5, 4, 1, 5, 1, 5, 2, 2, ...
$ Checkin_service                   <int> 5, 2, 4, 4, 4, 5, 5, 5, 5, 3, 2, 2, 2, 3, ...
$ Cleanliness                       <int> 3, 3, 4, 1, 2, 4, 5, 4, 4, 4, 3, 5, 4, 2, ...
$ Online_boarding                   <int> 2, 2, 2, 3, 5, 2, 3, 2, 4, 2, 5, 2, 5, 4, ...
```

```
> head(Virgin)
     ID Gender   CustomerType Age      TypeTravel    Class Flight_Distance DepartureDelayin_Mins ArrivalDelayin_Mins Satisfaction
1 11112 Female Loyal Customer  65 Personal Travel      Eco            265                    0                   0    satisfied
2 110278   Male Loyal Customer  47 Personal Travel Business           2464                  310                 305    satisfied
3 103199 Female Loyal Customer  15 Personal Travel      Eco           2138                    0                   0    satisfied
4  47462 Female Loyal Customer  60 Personal Travel      Eco            623                    0                   0    satisfied
5 120011 Female Loyal Customer  70 Personal Travel      Eco            354                    0                   0    satisfied
6 100744   Male Loyal Customer  30 Personal Travel      Eco           1894                    0                   0    satisfied
  Seat_comfort Departure.Arrival.time_convenient Food_drink Gate_location Inflightwifi_service Inflight_entertainment Online_support
1            0                                 0          0             2                    2                     4              2
2            0                                 0          0             3                    0                     2              2
3            0                                 0          0             3                    2                     0              2
4            0                                 0          0             3                    3                     4              3
5            0                                 0          0             3                    4                     3              4
6            0                                 0          0             3                    2                     0              2
  Ease_of_Onlinebooking Onboard_service Leg_room_service Baggage_handling Checkin_service Cleanliness Online_boarding
1                     3              3                0                 3               5           3               2
2                     3              4                4                 4               2           3               2
3                     2              3                3                 4               4           4               2
4                     1              1                0                 1               4           1               3
5                     2              2                0                 2               4           2               5
6                     2              5                4                 5               5           4               2
> str(Virgin)
'data.frame':   90917 obs. of  24 variables:
 $ ID                               : int  11112 110278 103199 47462 120011 100744 32838 32864 53786 7243 ...
 $ Gender                           : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 2 1 2 1 2 ...
 $ CustomerType                     : Factor w/ 2 levels "disloyal Customer",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Age                              : int  65 47 15 60 70 30 66 10 56 22 ...
 $ TypeTravel                       : Factor w/ 2 levels "Business travel",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Class                            : Factor w/ 3 levels "Business","Eco",..: 2 1 2 2 2 2 2 2 1 2 ...
 $ Flight_Distance                  : int  265 2464 2138 623 354 1894 227 1812 73 1556 ...
 $ DepartureDelayin_Mins            : int  0 310 0 0 0 0 17 0 0 30 ...
 $ ArrivalDelayin_Mins              : int  0 305 0 0 0 0 15 0 0 26 ...
 $ Satisfaction                     : Factor w/ 2 levels "neutral or dissatisfied",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Seat_comfort                     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Departure.Arrival.time_convenient: int  0 0 0 0 0 0 0 0 0 0 ...
 $ Food_drink                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Gate_location                    : int  2 3 3 3 3 3 3 3 3 3 ...
 $ Inflightwifi_service             : int  2 0 2 3 4 2 2 2 5 2 ...
 $ Inflight_entertainment           : int  4 2 0 4 3 0 5 0 3 0 ...
 $ Online_support                   : int  2 2 2 3 4 2 5 2 5 2 ...
 $ Ease_of_Onlinebooking            : int  3 3 2 1 2 2 5 2 4 2 ...
 $ Onboard_service                  : int  3 4 3 1 2 5 5 3 4 2 ...
 $ Leg_room_service                 : int  0 4 3 0 0 4 0 3 0 4 ...
 $ Baggage_handling                 : int  3 4 4 1 2 5 5 4 1 5 ...
 $ Checkin_service                  : int  5 2 4 4 4 5 5 5 5 3 ...
 $ Cleanliness                      : int  3 3 4 1 2 4 5 4 4 4 ...
 $ Online_boarding                  : int  2 2 2 3 5 2 3 2 4 2 ...
```

```
> sapply(Virgin, function(x) sum(is.na(x)))
                              ID                           Gender                     CustomerType                              Age 
                               0                                0                                0                                0 
                      TypeTravel                            Class                  Flight_Distance              DepartureDelayin_Mins 
                               0                                0                                0                                0 
              ArrivalDelayin_Mins                     Satisfaction                   Seat_comfort Departure.Arrival.time_convenient 
                             280                                0                                0                                0 
                     Food_drink                    Gate_location             Inflightwifi_service           Inflight_entertainment 
                               0                                0                                0                                0 
                  Online_support             Ease_of_Onlinebooking                 Onboard_service                Leg_room_service 
                               0                                0                                0                                0 
                Baggage_handling                  Checkin_service                      Cleanliness                  Online_boarding 
                               0                                0                                0                                0 
```

We observe 280 NA values in the attribute “ArrivalDelayin_Min”. Let us analyze the dataset further to understand whether there are any NA values which are present in the data appart from this variable. To perform this, we fill the blanks with NA values. Then we take the count of the NA values of the attributes of this dataset. Following we observe there are NA’s values of attributes of original dataset and the dataset with NA values are:

CustomerType	TypeTravel	Departure.Arrival.time_convenient
9099	**9088**	**8244**
Food_drink	Onboard_service	
8181	**7179**	

By imputing the blank values of different attributes with NA’s, we observe that the attributes have some missing values which are imputed with NA’s. Hence, we can replace the NA values with mean values of the attribute or omit the missing, however before proceeding any further, let us analyze the “Arrival.Delay.in.Minutes” variable’s correlation with other variables.

Converting data into int:

```
> virgin$Onboard_service[is.na(virgin$Onboard_service)]<- Onboard_service
> virgin$Seat_comfort<-factor(virgin$Seat_comfort,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=c(0,
1,2,3,4,5),ordered = T)
> virgin$Seat_comfort <- as.integer(factor(virgin$Seat_comfort))
> virgin$Departure.Arrival.time_convenient<- factor(virgin$Departure.Arrival.time_convenient,levels=c("extremely poor","poor","need improvemen
t","acceptable","good","excellent"),labels=c(0,1,2,3,4,5),ordered = T)
> virgin$Departure.Arrival.time_convenient <- as.integer(factor(virgin$Departure.Arrival.time_convenient))
> virgin$Food_drink<-factor(virgin$Food_drink,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=c(0,1,2,
3,4,5),ordered = T)
> virgin$Food_drink <- as.integer(factor(virgin$Food_drink))
> virgin$Gate_location<-factor(virgin$Gate_location,levels=c("very inconvinient","Inconvinient","need improvement","manageable","Convinient","very
 convinient"),labels=c(0,1,2,3,4,5),ordered = T)
> virgin$Gate_location <- as.integer(factor(virgin$Gate_location))
> virgin$Inflightwifi_service<-factor(virgin$Inflightwifi_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excelle
nt"),labels=c(0,1,2,3,4,5),ordered = T)
> virgin$Inflightwifi_service <- as.integer(factor(virgin$Inflightwifi_service))
> virgin$Inflight_entertainment<-factor(virgin$Inflight_entertainment,levels=c("extremely poor","poor","need improvement","acceptable","good","exc
ellent"),labels=c(0,1,2,3,4,5),ordered = T)
> virgin$Inflight_entertainment <- as.integer(factor(virgin$Inflight_entertainment))
> virgin$Online_support<-factor(virgin$Online_support,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=
c(0,1,2,3,4,5),ordered = T)
> virgin$Online_support <- as.integer(factor(virgin$Online_support))
> virgin$Ease_of_Onlinebooking<-factor(virgin$Ease_of_Onlinebooking,levels=c("extremely poor","poor","need improvement","acceptable","good","excel
lent"),labels=c(0,1,2,3,4,5),ordered = T)
> virgin$Ease_of_Onlinebooking <- as.integer(factor(virgin$Ease_of_Onlinebooking))
> virgin$Onboard_service<-factor(virgin$Onboard_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),label
s=c(0,1,2,3,4,5),ordered = T)
> virgin$Onboard_service <- as.integer(factor(virgin$Onboard_service))
> virgin$Leg_room_service<-factor(virgin$Leg_room_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),lab
els=c(0,1,2,3,4,5),ordered = T)
> virgin$Leg_room_service <- as.integer(factor(virgin$Leg_room_service))
> virgin$Baggage_handling<-factor(virgin$Baggage_handling,levels=c("poor","need improvement","acceptable","good","excellent"),labels=c(0,1,2,3,4),
ordered = T)
> virgin$Baggage_handling <- as.integer(factor(virgin$Baggage_handling))
> virgin$Checkin_service<-factor(virgin$Checkin_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),label
s=c(0,1,2,3,4,5),ordered = T)
> virgin$Checkin_service <- as.integer(factor(virgin$Checkin_service))
> virgin$Cleanliness<-factor(virgin$Cleanliness,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=c(0,1,
2,3,4,5),ordered = T)
> virgin$Cleanliness <- as.integer(factor(virgin$Cleanliness))
> virgin$Online_boarding<-factor(virgin$Online_boarding,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),label
s=c(0,1,2,3,4,5),ordered = T)
> virgin$Online_boarding <- as.integer(factor(virgin$Online_boarding)) str(virgin)
```

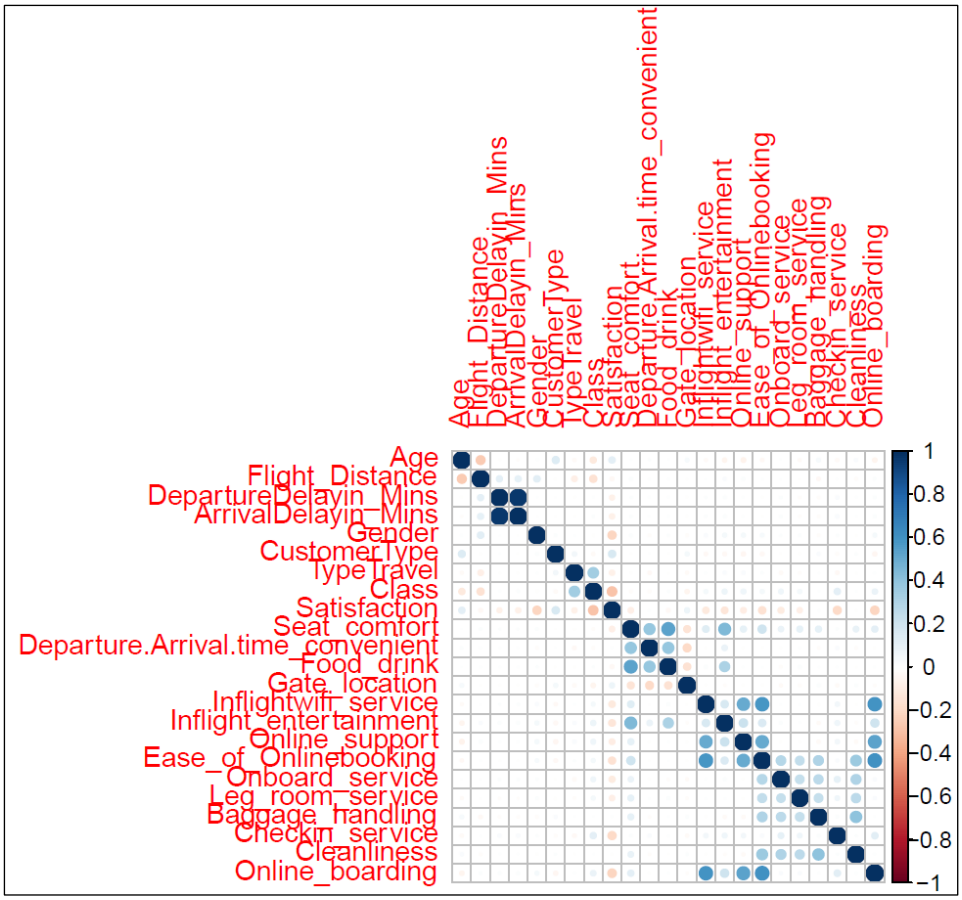


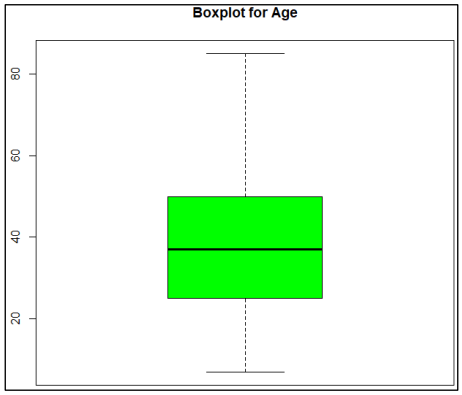
Figure 1: Plotting of the parameters of satisfaction from airline passengers

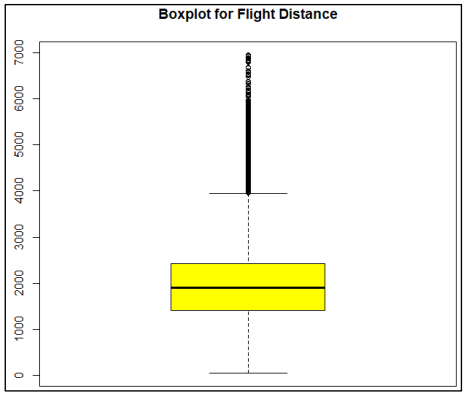
Initial Exploratory Data Analysis

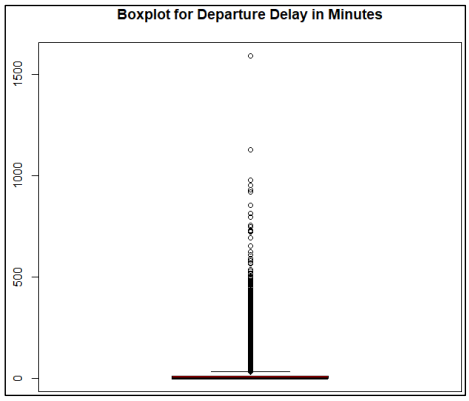
3.1 Uni-Variate Analysis (Distribution and Spread for Every Continuous Attribute, Distribution of Data in Categories for Categorical Ones)

After observing the impact of several independent variables over the “Satisfaction level”, we present below few visualizations displaying a level of satisfaction at different factor levels for the variables “Gender”, “Class”, “Seat Comfort”, “Inflight Entertainment”, “Online Support” and “Baggage Handling”.

On an inferential perspective, these variables are considered the most significant in terms of airline customer satisfaction, which is more likely the case as we observe the following visualizations.



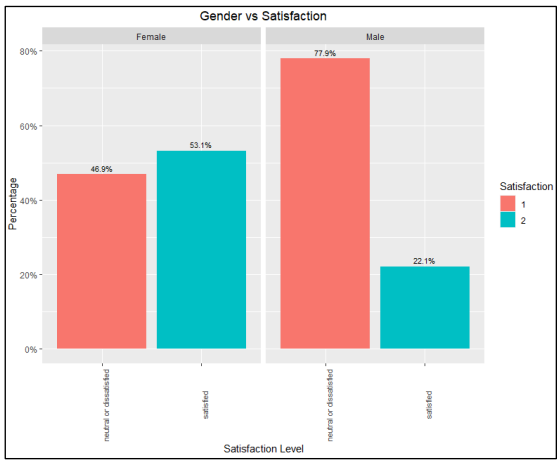




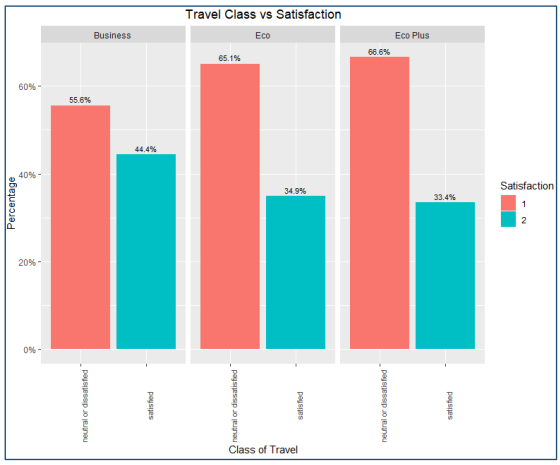
We observe that the Female customers are comparatively more satisfied than the Male customers as we can see 65% female customers are satisfied against 34% dissatisfied.

3.2 Bi-Variate Analysis (The Relationship Between Different Variables, Correlations)

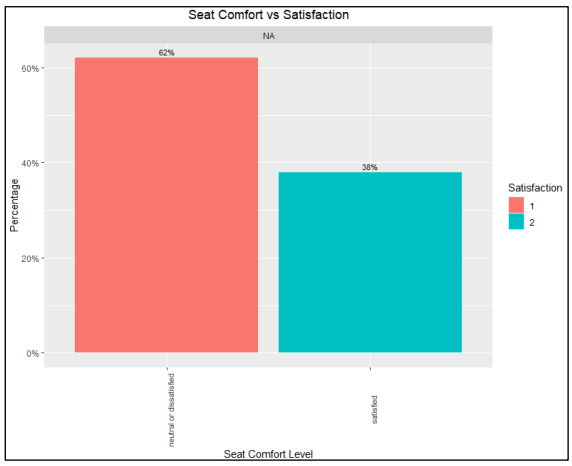
Bi-Variate Analysis (The Relationship Between Different Variables, Correlations)



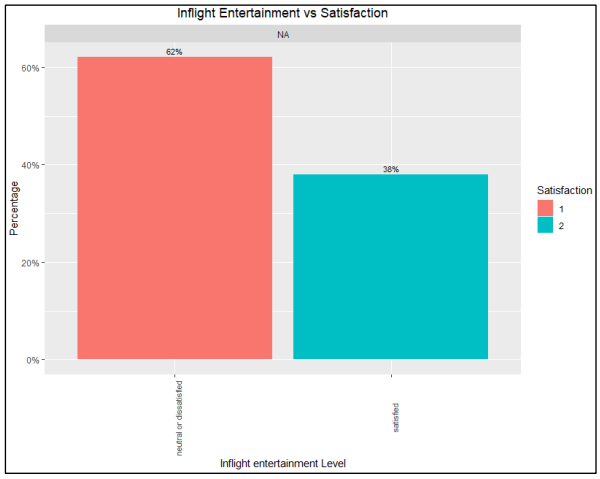
We observe that the customers traveling the Business Class are more satisfied than the customers traveling in Economy classes.



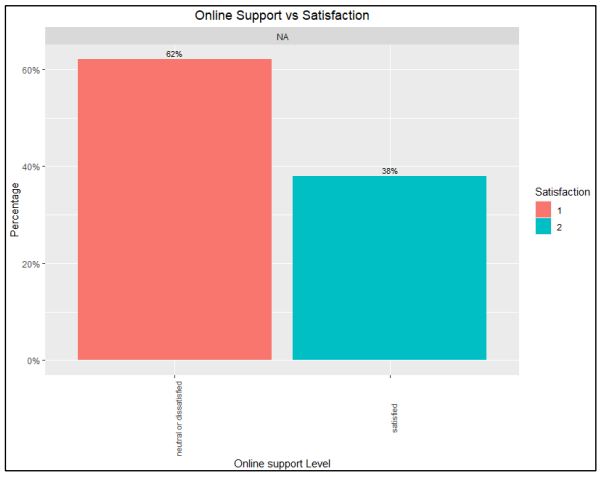
We observe that Seat Comfort is having a significant effect on the customer satisfaction level, as we see that customers rating 5 on seat comfort are 99 percent satisfied.



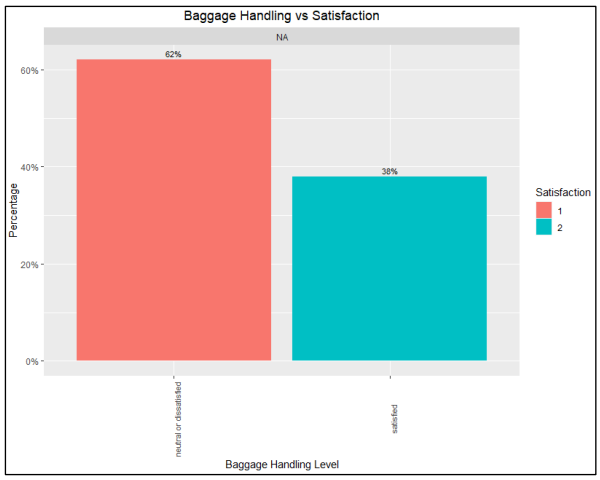
We observe that Inflight Entertainment is having a significant effect on the customer satisfaction level, as we see that customers rating 5 on seat comfort are 95 percent satisfied.



We observe that Online Support is having a significant effect on the customer satisfaction level, as we see that customers rating 0 on Online Support are 100 percent dissatisfied.



We observe that Baggage Handling is having a significant effect on the customer satisfaction Level, as we see that customers rating 5 on Baggage Handling are 73 percent satisfied.



4.1 Data Pre-Processing

This data set have a 90917obs. of 24 variables. I'll be cleaning the data to fill column null value at "Arrival Delay in Minutes". Probably, I can use all features for my prediction. Also, can take conclusion stating what are the features that will be used on the heat map in Exploratory Data Analysis after this part.

The data we obtained from Kaggle represents British Airline Satisfaction data consisting of 90917 observations and 24 attributes. The Satisfaction level, which is our dependent variable, is represented as a factor ("Satisfied" and "Neutral or Dissatisfied").

4.2 Removal of Unwanted Variables

We observe the dataset having 90917 observations and 24 attributes. The attribute "satisfaction" represents the satisfaction level of the customer on two different levels: "neutral or dissatisfied" and "satisfied". Also, we drop the ID attribute as we wouldn't require that in our analysis or model building. Hence, we have 1 dependent variable and 22 independent variables in our dataset. We check our dataset for any possible NA values, which must be dealt before we proceed with building the model.

```
> glimpse(Virgin)
Rows: 90,917
Columns: 24
$ ID                          <int> 11112, 110278, 103199, 47462, 120011, 1007...
$ Gender                      <fct> Female, Male, Female, Female, Female, Male...
$ CustomerType                <fct> Loyal Customer, Loyal Customer, Loyal Cust...
$ Age                         <int> 65, 47, 15, 60, 70, 30, 66, 10, 56, 22, 58...
$ TypeTravel                  <fct> Personal Travel, Personal Travel, Personal...
$ Class                       <fct> Eco, Business, Eco, Eco, Eco, Eco, Eco, Ec...
$ Flight_Distance             <int> 265, 2464, 2138, 623, 354, 1894, 227, 1812...
$ DepartureDelayin_Mins       <int> 0, 310, 0, 0, 0, 0, 17, 0, 0, 30, 47, 0, 0...
$ ArrivalDelayin_Mins         <int> 0, 305, 0, 0, 0, 0, 15, 0, 0, 26, 48, 0, 0...
$ Satisfaction                <fct> satisfied, satisfied, satisfied, satisfied...
$ Seat_comfort                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ Departure.Arrival.time_convenient <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...
$ Food_drink                  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ Gate_location               <int> 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 1, ...
$ Inflightwifi_service        <int> 2, 0, 2, 3, 4, 2, 2, 2, 5, 2, 3, 2, 5, 4, ...
$ Inflight_entertainment      <int> 4, 2, 0, 4, 3, 0, 5, 0, 3, 0, 3, 0, 0, 0, ...
$ Online_support              <int> 2, 2, 2, 3, 4, 2, 5, 2, 5, 2, 3, 2, 5, 4, ...
$ Ease_of_Onlinebooking       <int> 3, 3, 2, 1, 2, 2, 5, 2, 4, 2, 3, 2, 5, 4, ...
$ Onboard_service             <int> 3, 4, 3, 1, 2, 5, 5, 3, 4, 2, 3, 3, 1, 3, ...
$ Leg_room_service            <int> 0, 4, 3, 0, 0, 4, 0, 3, 0, 4, 0, 2, 3, 5, ...
$ Baggage_handling            <int> 3, 4, 4, 1, 2, 5, 5, 4, 1, 5, 1, 5, 2, 2, ...
$ Checkin_service             <int> 5, 2, 4, 4, 4, 5, 5, 5, 5, 3, 2, 2, 2, 3, ...
$ Cleanliness                 <int> 3, 3, 4, 1, 2, 4, 5, 4, 4, 4, 3, 5, 4, 2, ...
$ Online_boarding             <int> 2, 2, 2, 3, 5, 2, 3, 2, 4, 2, 5, 2, 5, 4, ...
```

```
> sapply(Virgin, function(x) sum(is.na(x)))
                          ID                       Gender                 CustomerType                          Age
                           0                            0                            0                            0
                  TypeTravel                        Class              Flight_Distance        DepartureDelayin_Mins
                           0                            0                            0                            0
        ArrivalDelayin_Mins                 Satisfaction                Seat_comfort Departure.Arrival.time_convenient
                         280                            0                            0                            0
                  Food_drink                Gate_location         Inflightwifi_service       Inflight_entertainment
                           0                            0                            0                            0
              Online_support         Ease_of_Onlinebooking              Onboard_service             Leg_room_service
                           0                            0                            0                            0
            Baggage_handling              Checkin_service                  Cleanliness              Online_boarding
                           0                            0                            0                            0
```

Data	
Virgin	90917 obs. of 24 variables
Virgin.mis	90917 obs. of 24 variables

4.3 Missing Value Treatment

```
# Generate 10% missing values at Random
Virgin.mis <- prodNA(Virgin, noNA = 0.1)

# Check missing values introduced in the data
summary(Virgin.mis)

# I've removed categorical variable.
# Let's here focus on continuous values.
# To treat categorical variable, simply encode the levels and follow the procedure below.

# Remove categorical variables
Virgin.mis <- subset(Virgin.mis, select = -c(Species))
summary(Virgin.mis)

# mice package has a function known as md.pattern().
# It returns a tabular form of missing value present in each variable in a data set.

md.pattern(Virgin.mis)
mice_plot <- aggr(Virgin.mis, col=c('navyblue','yellow'),
                  numbers=TRUE, sortVars=TRUE,
                  labels=names(iris.mis), cex.axis=.7,
                  gap=3, ylab=c("Missing data","Pattern"))
```

```
> # Generate 10% missing values at Random
> Virgin.mis <- prodNA(Virgin, noNA = 0.1)
> # Check missing values introduced in the data
> summary(Virgin.mis)
       ID              Gender                CustomerType         Age                TypeTravel          Class
 Min.   :     1   Female:41935   disloyal Customer:21191   Min.   : 7.00   Business travel:45794   Business:27253
 1st Qu.: 32584   Male  :39981   Loyal Customer   :60581   1st Qu.:25.00   Personal Travel:36046   Eco     :47613
 Median : 64957   NA's  : 9001   NA's             : 9145   Median :37.00   NA's            : 9077   Eco Plus: 6997
 Mean   : 64934                                           Mean   :37.59                           NA's    : 9054
 3rd Qu.: 97527                                           3rd Qu.:50.00
 Max.   :129880                                           Max.   :85.00
 NA's   :9134                                             NA's   :8991
 Flight_Distance DepartureDelayin_Mins ArrivalDelayin_Mins                  Satisfaction    Seat_comfort
 Min.   :  50    Min.   :   0.00       Min.   :   0.00     neutral or dissatisfied:50798   Min.   :0.000
 1st Qu.:1408    1st Qu.:   0.00       1st Qu.:   0.00     satisfied              :30953   1st Qu.:2.000
 Median :1901    Median :   0.00       Median :   0.00     NA's                   : 9166   Median :3.000
 Mean   :1938    Mean   :  15.36       Mean   :  15.72                                     Mean   :2.649
 3rd Qu.:2427    3rd Qu.:  13.00       3rd Qu.:  14.00                                     3rd Qu.:4.000
 Max.   :6951    Max.   :1592.00       Max.   :1584.00                                     Max.   :5.000
 NA's   :8911    NA's   :9235          NA's   :9207                                        NA's   :9132
 Departure.Arrival.time_convenient   Food_drink    Gate_location   Inflightwifi_service Inflight_entertainment
 Min.   :0.00                      Min.   :0.000   Min.   :0.000   Min.   :0.00         Min.   :0.000
 1st Qu.:2.00                      1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00         1st Qu.:2.000
 Median :3.00                      Median :3.000   Median :3.000   Median :3.00         Median :3.000
 Mean   :2.99                      Mean   :2.772   Mean   :2.986   Mean   :3.06         Mean   :3.042
 3rd Qu.:4.00                      3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00         3rd Qu.:4.000
 Max.   :5.00                      Max.   :5.000   Max.   :5.000   Max.   :5.00         Max.   :5.000
 NA's   :9069                      NA's   :9219    NA's   :9172    NA's   :9257         NA's   :9072
 Online_support  Ease_of_Onlinebooking Onboard_service Leg_room_service Baggage_handling Checkin_service  Cleanliness
 Min.   :0.000   Min.   :0.000         Min.   :0.000   Min.   :0.000    Min.   :1.000    Min.   :0.000   Min.   :0.000
 1st Qu.:2.000   1st Qu.:2.000         1st Qu.:2.000   1st Qu.:2.000    1st Qu.:3.000    1st Qu.:2.000   1st Qu.:3.000
 Median :3.000   Median :3.000         Median :3.000   Median :3.000    Median :4.000    Median :3.000   Median :4.000
 Mean   :3.224   Mean   :3.062         Mean   :3.192   Mean   :3.224    Mean   :3.486    Mean   :3.226   Mean   :3.501
 3rd Qu.:4.000   3rd Qu.:4.000         3rd Qu.:4.000   3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000         Max.   :5.000   Max.   :5.000    Max.   :5.000    Max.   :5.000   Max.   :5.000
 NA's   :8875    NA's   :9047          NA's   :9162    NA's   :9093     NA's   :9131     NA's   :9128    NA's   :9169
 Online_boarding
 Min.   :0.0
 1st Qu.:2.0
 Median :3.0
 Mean   :3.1
 3rd Qu.:4.0
 Max.   :5.0
 NA's   :9011
```

```
 [ reached getOption("max.print") -- omitted 16944 rows ]
> md.pattern(Virgin.mis)
     Online_support Flight_Distance  Age Gender Online_boarding Ease_of_Onlinebooking Class Departure.Arrival.time_convenient
7309              1               1    1      1               1                     1     1                                  1
859               1               1    1      1               1                     1     1                                  1
846               1               1    1      1               1                     1     1                                  1
82                1               1    1      1               1                     1     1                                  1
783               1               1    1      1               1                     1     1                                  1
96                1               1    1      1               1                     1     1                                  1
98                1               1    1      1               1                     1     1                                  1
5                 1               1    1      1               1                     1     1                                  1
821               1               1    1      1               1                     1     1                                  1
86                1               1    1      1               1                     1     1                                  1
74                1               1    1      1               1                     1     1                                  1
7                 1               1    1      1               1                     1     1                                  1
101               1               1    1      1               1                     1     1                                  1
6                 1               1    1      1               1                     1     1                                  1
9                 1               1    1      1               1                     1     1                                  1
812               1               1    1      1               1                     1     1                                  1
86                1               1    1      1               1                     1     1                                  1
109               1               1    1      1               1                     1     1                                  1
9                 1               1    1      1               1                     1     1                                  1
96                1               1    1      1               1                     1     1                                  1
11                1               1    1      1               1                     1     1                                  1
3                 1               1    1      1               1                     1     1                                  1
1                 1               1    1      1               1                     1     1                                  1
100               1               1    1      1               1                     1     1                                  1
22                1               1    1      1               1                     1     1                                  1
7                 1               1    1      1               1                     1     1                                  1
6                 1               1    1      1               1                     1     1                                  1
1                 1               1    1      1               1                     1     1                                  1
818               1               1    1      1               1                     1     1                                  1
90                1               1    1      1               1                     1     1                                  1
104               1               1    1      1               1                     1     1                                  1
12                1               1    1      1               1                     1     1                                  1
85                1               1    1      1               1                     1     1                                  1
15                1               1    1      1               1                     1     1                                  1
17                1               1    1      1               1                     1     1                                  1
3                 1               1    1      1               1                     1     1                                  1
90                1               1    1      1               1                     1     1                                  1
8                 1               1    1      1               1                     1     1                                  1
11                1               1    1      1               1                     1     1                                  1
1                 1               1    1      1               1                     1     1                                  1
```

We observe 280 NA values in the attribute “ArrivalDelayin_Min”. Let us analyse the dataset further to understand whether there are any NA values which are present in the data apart from this variable. To perform this, we fill the blanks with NA values. Then we take the count of the NA values of the attributes of this dataset. Following we observe there are NA’s values of attributes of original dataset and the dataset with NA values are:

CustomerType	TypeTravel	Departure.Arrival.time_convenient
9099	**9088**	**8244**
Food_drink	Onboard_service	
8181	**7179**	

By imputing the blank values of different attributes with NA’s, we observe that the attributes have some missing values which are imputed with NA’s. Hence, we can replace the NA values with mean values of the attribute or omit the missing, however before proceeding any further, let us analyze the “Arrival.Delay.in.Minutes” variable’s correlation with other variables.

```
 [ reached getOption("max.print") -- omitted 16944 rows ]
> imputed_Data <- mice(virgin.mis, m=5, maxit = 50, method = 'pmm', seed = 500)

 iter imp variable
  1   1  ID  Gender  CustomerType  Age  TypeTravel  Class  Flight_Distance  DepartureDelayin_Mins  ArrivalDelayin_Mins  Satisfacti
on  Seat_comfort  Departure.Arrival.time_convenient  Food_drink  Gate_location  Inflightwifi_service  Inflight_entertainment  Onli
ne_support  Ease_of_Onlinebooking  Onboard_service  Leg_room_service  Baggage_handling  Checkin_service  Cleanliness  Online_board
ing
  1   2  ID  Gender  CustomerType  Age  TypeTravel  Class  Flight_Distance  DepartureDelayin_Mins  ArrivalDelayin_Mins  Satisfacti
on  Seat_comfort  Departure.Arrival.time_convenient  Food_drink  Gate_location  Inflightwifi_service  Inflight_entertainment  Onli
ne_support  Ease_of_Onlinebooking  Onboard_service  Leg_room_service  Baggage_handling  Checkin_service  Cleanliness  Online_board
ing
  1   3  ID  Gender  CustomerType  Age  TypeTravel  Class  Flight_Distance  DepartureDelayin_Mins  ArrivalDelayin_Mins  Satisfacti
on  Seat_comfort  Departure.Arrival.time_convenient  Food_drink  Gate_location  Inflightwifi_service  Inflight_entertainment  Onli
ne_support  Ease_of_Onlinebooking  Onboard_service  Leg_room_service  Baggage_handling  Checkin_service  Cleanliness  Online_board
ing
  1   4  ID  Gender  CustomerType  Age  TypeTravel  Class  Flight_Distance  DepartureDelayin_Mins  ArrivalDelayin_Mins  Satisfacti
on  Seat_comfort  Departure.Arrival.time_convenient  Food_drink  Gate_location  Inflightwifi_service  Inflight_entertainment  Onli
ne_support  Ease_of_Onlinebooking  Onboard_service  Leg_room_service  Baggage_handling  Checkin_service  Cleanliness  Online_board
ing
  1   5  ID  Gender
```

4.4 Variable Transformation

```
> Virgin$Onboard_service[is.na(Virgin$Onboard_service)]<- Onboard_service
> Virgin$Seat_comfort<-factor(Virgin$Seat_comfort,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=c(0,
1,2,3,4,5),ordered = T)
> Virgin$Seat_comfort <- as.integer(factor(Virgin$Seat_comfort))
> Virgin$Departure.Arrival.time_convenient<- factor(Virgin$Departure.Arrival.time_convenient,levels=c("extremely poor","poor","need improvemen
t","acceptable","good","excellent"),labels=c(0,1,2,3,4,5),ordered = T)
> Virgin$Departure.Arrival.time_convenient <- as.integer(factor(Virgin$Departure.Arrival.time_convenient))
> Virgin$Food_drink<-factor(Virgin$Food_drink,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=c(0,1,2,
3,4,5),ordered = T)
> Virgin$Food_drink <- as.integer(factor(Virgin$Food_drink))
> Virgin$Gate_location<-factor(Virgin$Gate_location,levels=c("very inconvinient","Inconvinient","need improvement","manageable","Convinient","very
 convinient"),labels=c(0,1,2,3,4,5),ordered = T)
> Virgin$Gate_location <- as.integer(factor(Virgin$Gate_location))
> Virgin$Inflightwifi_service<-factor(Virgin$Inflightwifi_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excelle
nt"),labels=c(0,1,2,3,4,5),ordered = T)
> Virgin$Inflightwifi_service <- as.integer(factor(Virgin$Inflightwifi_service))
> Virgin$Inflight_entertainment<-factor(Virgin$Inflight_entertainment,levels=c("extremely poor","poor","need improvement","acceptable","good","exc
ellent"),labels=c(0,1,2,3,4,5),ordered = T)
> Virgin$Inflight_entertainment <- as.integer(factor(Virgin$Inflight_entertainment))
> Virgin$Online_support<-factor(Virgin$Online_support,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=
c(0,1,2,3,4,5),ordered = T)
> Virgin$Online_support <- as.integer(factor(Virgin$Online_support))
> Virgin$Ease_of_Onlinebooking<-factor(Virgin$Ease_of_Onlinebooking,levels=c("extremely poor","poor","need improvement","acceptable","good","excel
lent"),labels=c(0,1,2,3,4,5),ordered = T)
> Virgin$Ease_of_Onlinebooking <- as.integer(factor(Virgin$Ease_of_Onlinebooking))
> Virgin$Onboard_service<-factor(Virgin$Onboard_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),label
s=c(0,1,2,3,4,5),ordered = T)
> Virgin$Onboard_service <- as.integer(factor(Virgin$Onboard_service))
> Virgin$Leg_room_service<-factor(Virgin$Leg_room_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),lab
els=c(0,1,2,3,4,5),ordered = T)
> Virgin$Leg_room_service <- as.integer(factor(Virgin$Leg_room_service))
> Virgin$Baggage_handling<-factor(Virgin$Baggage_handling,levels=c("poor","need improvement","acceptable","good","excellent"),labels=c(0,1,2,3,4),
ordered = T)
> Virgin$Baggage_handling <- as.integer(factor(Virgin$Baggage_handling))
> Virgin$Checkin_service<-factor(Virgin$Checkin_service,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),label
s=c(0,1,2,3,4,5),ordered = T)
> Virgin$Checkin_service <- as.integer(factor(Virgin$Checkin_service))
> Virgin$Cleanliness<-factor(Virgin$Cleanliness,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),labels=c(0,1,
2,3,4,5),ordered = T)
> Virgin$Cleanliness <- as.integer(factor(Virgin$Cleanliness))
> Virgin$Online_boarding<-factor(Virgin$Online_boarding,levels=c("extremely poor","poor","need improvement","acceptable","good","excellent"),label
s=c(0,1,2,3,4,5),ordered = T)
> Virgin$Online_boarding <- as.integer(factor(Virgin$Online_boarding)) str(Virgin)
```

5 Exploratory Data Analysis

5.1 Relationship Among Variables

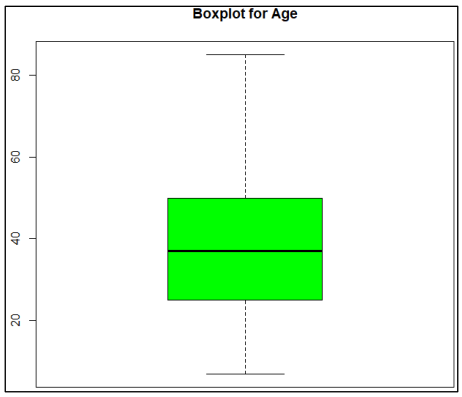
After observing the impact of several independent variables over the “Satisfaction level”, we present below few visualizations displaying a level of satisfaction at different factor levels for the variables “Gender”, “Class”, “Seat Comfort”, “Inflight Entertainment”, “Online Support” and “Baggage Handling”.

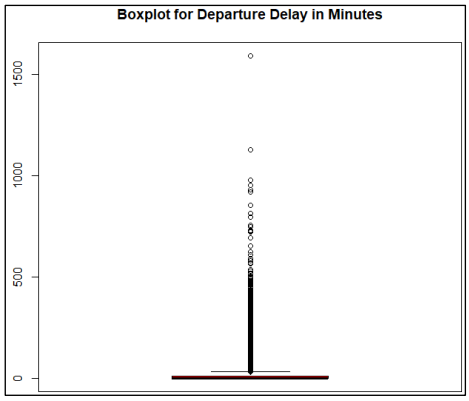
Uni-Variate Analysis – these techniques will be used if there is a single measurement of each of "n" sample objects, or if there are several measurements on each of the "n" observations, but each variable is to be analysed in isolation. Examples of the techniques that can be used with the univariate data are: the central tendency measures (mean, median, mode), measures of dispersion (standard deviation, relative and absolute frequencies) and the single sample t-test.

Bi-Variate Analysis – these techniques allow the researcher to examine the interaction between variables taken two at a time. Among the available bivariate techniques are: linear correlation coefficient, the rank correlation coefficient, the Man Whitney utest, the Kolmogorov-Smirnov test and the chi-square test of association.

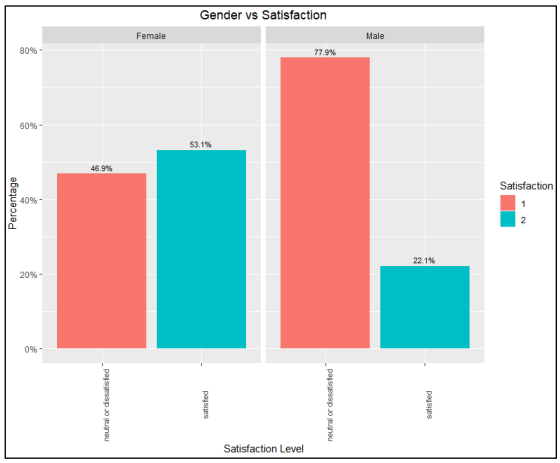
Multi-Variate Analysis – these techniques are concerned with the investigation of interactions among a set of variables. These techniques are distinguished from the univariate and bivariate methods by their focus on more than two variables at a time. The multivariate techniques can be classified as either dependent or independent. The dependent methods require that one or more variables are specified as being predicted by a set of independent variables, while the independent methods require that no variables selected as being a dependent variable. The dependence methods include analysis of variance (ANOVA), analysis of variance and covariance (ANCOVA), multiple regression and discriminant analysis (DA). The independence methods include factor analysis, cluster analysis, latent structure analysis and non-metric multidimensional scaling.

On an inferential perspective, these variables are considered the most significant in terms of airline customer satisfaction, which is more likely the case as we observe the following visualizations.

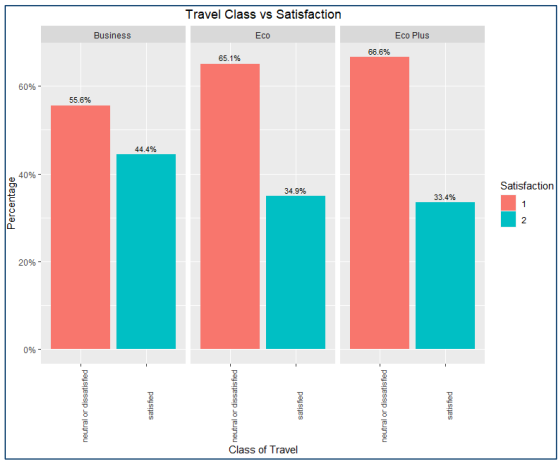




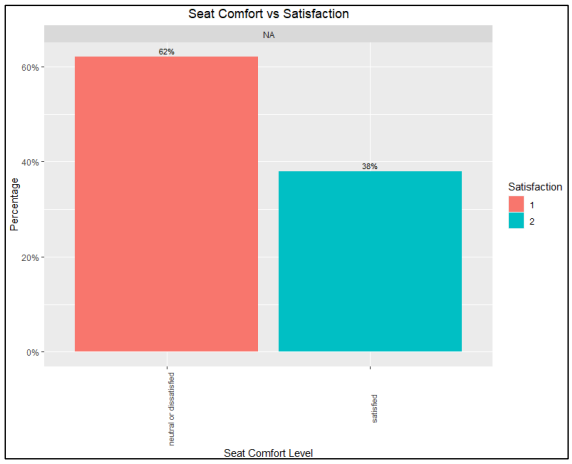
We observe that the Female customers are comparatively more satisfied than the Male customers as we can see 65% female customers are satisfied against 34% dissatisfied.



We observe that the customers traveling the Business Class are more satisfied than the customers traveling in Economy classes.



We observe that Seat Comfort is having a significant effect on the customer satisfaction level, as we see that customers rating 5 on seat comfort are 99 percent satisfied.



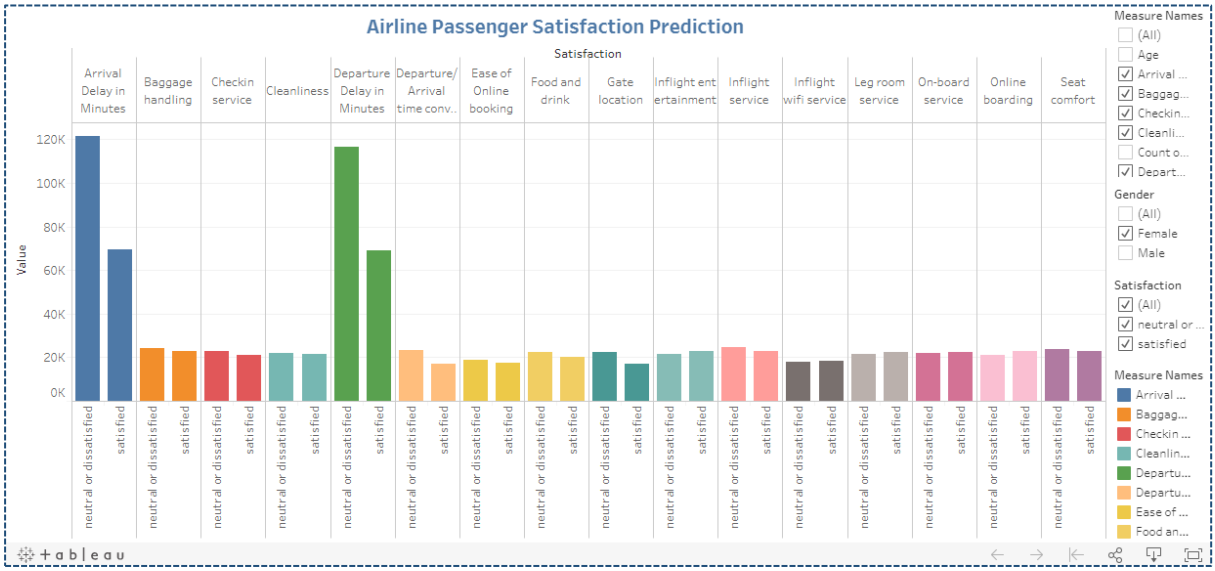
Important Variables

Inflight Entertainment was most important variable in predicting the customer satisfaction index. Inflight Wi-Fi service, Seat comfort, Ease of online booking, Leg room were the other important variables in the prediction.

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_REP_Inflight_entertainment	Imputed: Replacement: Inflight entertainment	1	1.0000	1.0000	1.0000
Class	Class	1	0.5206	0.5250	1.0085
IMP_REP_Inflight_wifi_service	Imputed: Replacement: Inflight wifi service	1	0.4219	0.4109	0.9739
IMP_REP_Seat_comfort	Imputed: Replacement: Seat comfort	1	0.3580	0.3561	0.9946
IMP_REP_Ease_of_Online_booking	Imputed: Replacement: Ease of Online booking	2	0.3333	0.3345	1.0036
IMP_REP_Leg_room_service	Imputed: Replacement: Leg room service	1	0.2320	0.2343	1.0102
IMP_REP_Online_boarding	Imputed: Replacement: Online boarding	1	0.2099	0.2019	0.9621
IMP_REP_Cleanliness	Imputed: Replacement: Cleanliness	1	0.1781	0.1736	0.9749
Type_of_Travel	Type of Travel	1	0.1772	0.1705	0.9623

5.2 Insightful Visualizations

We have seen the highest number of dissatisfactions by passengers for the 'Arrival Delay' and 'Departure Delay' – in those 2 criteria. Number of passengers are more dissatisfied than satisfied in these 2 parameters. Rest of the parameters, we can say that passengers are almost equally satisfied and dissatisfied with the services.



Source:

https://public.tableau.com/profile/tahmid.bari#!/vizhome/AirlinePassengerCustomerSatisfaction_16015564531270/Cmrsatisf

6 Analytical Approach

We have used the 'Mice' package to impute the missing data. PMM – 'Predictive Mean Matching'. This function is used internally only but might help others to implement an efficient way of doing predictive mean matching on top of any prediction based missing value imputation. It works as follows: For each predicted value of a vector x-test, the closest k predicted values of another vector x-train are identified by k-nearest neighbour. Then, one of those neighbours is randomly picked and its corresponding observed value in y-train is returned.

6.1 Multicollinearity Analysis

- Multicollinearity – It's a state of very high intercorrelations or among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.
- There are certain reasons why multicollinearity occurs:
    - It is caused by an inaccurate use of dummy variables.
    - It is caused by the inclusion of a variable which is computed from other variables in the data set.
    - It can also result from the repetition of the same kind of variable.
    - Generally, occurs when the variables are highly correlated to each other.

- Multicollinearity can result in several problems, such as:
  - The partial regression coefficient due to multicollinearity may not be estimated precisely. The standard errors are likely to be high.
  - Multicollinearity results in a change in the signs as well as in the magnitudes of the partial regression coefficients from one sample to another sample.
  - Multicollinearity makes it tedious to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.

In the presence of high multicollinearity, the confidence intervals of the coefficients tend to become very wide and the statistics tend to be very small. It becomes difficult to reject the null hypothesis of any study when multicollinearity is present in the data under study.

6.2 PCA/ Factor Analysis

- check for correlation
  - pairs(cst.d)
  - pairs.panels(cst.d)
- run bartlett test
  - cortest.bartlett(cor(cst.d), nrow(cst.d))
  - basicially this is a chi sq. test
  - results are siginificant proceed with pca
- run pca using prcomp
  - cst.prcomp <- prcomp(cst.d, scale. = T)
  - summary(cst.prcomp)
  - print(cst.prcomp)
  - biplot(cst.prcomp)
  - plot(cst.prcomp)
  - pairs.panels(cst.prcomp$x)
- scree plot
  - screeplot(cst.prcomp, type = "l")
- eigen value
  - ev <- eigen (cor(cst.d))
  - ev
- percentage of variance explained
  - part.pca <- (ev$values/sum(ev$values)) * 100
  - plot(part.pca, type = "b")
- pca using princomp
  - cst.comp2 <- princomp(cst.d, score = T, cor = T)
  - summary(cst.comp2)
  - biplot(cst.comp2)
- pca using principal
  - cst.comp3 <- principal (cst.d, nfactor =4, rotate = "none")
  - print(cst.comp3)
  - cst.data1 <- principal (cst.d, nfactor =4, rotate = "varimax")
  - cst.data1
  - fa.diagram(cst.data1)

6.3 Multiple Linear Regression

- Scores for all the rows:
    - head(fal$scores)
    - regdata <- cbind (cst1[12], fal$scores)
- Labeling the data
    - names(regdata) <- c("Satisfaction", "Prod")
    - head(regdata)
- Splitting the data 70:30
    - set.seed(100)
    - indices= sample(1:nrow(regdata), 0.7*nrow(regdata))
    - train=regdata[indices,]
    - test = regdata[-indices,]
- Regression Model using train data
    - model1 = lm(Satisfaction~., train)
    - summary(model1)
- Regression model without post_purchase:
    - model2 <- lm(Satisfaction ~ Purchase+ Marketing+ Prod, data= train)
    - summary(model2)
- The factors Purchase, Marketing, Prod are highly significant and Post_purchase is not significant in the model.

7. Modelling Process

For the purposes of this section of the project, a number of data models will be built namely, a decision tree model, a logistic regression, a random forest classifier and a gradient boosting classifier in order to try and predict the satisfaction levels of the customers who flew Airlines based on the independent variables of our dataset. The pre-processing step often is crucial for obtaining a good fit of the model and better predictive ability.

7.1 Clean and Transform Dataset

Based on the training data above, there are several things we need to do in order to prepare the data for use in a model. There are several categorical variables that need to be encoded, including our target variable 'Satisfaction'. There are also a couple of columns that are unnecessary, such as 'id'. We can drop these.

The general purpose of factor analysis is to summaries the information contained within a large number of variables, into a smaller number of factors.

Factor analysis refers to a diverse set of techniques used to discern the underlying dimensions in phenomena. The mathematical aim of factor analysis is to determine linear combinations of variables that best describe the interrelationships between them. This study purpose is to discover the basic structure of a given domain and to add substantive interpretation to the underlying dimensions. Factor analysis accomplishes this by combining the original variables to create new, more abstract variables called factors. Therefore, the goal of factor analysis is parsimony: to reduce a large number of variables to as few a dimension or constructs as possible. In this research, factor analysis will be used to identify those factors that may measure specific attributes or dimensions of service quality and to identify the factors (dimensions) of passenger satisfaction.

Many previous researchers had considered a rule of thumb approach that takes factor loadings greater than ±0.30 as a significant loading. The loadings ± 0.40 are considered more important, and if the loadings are ± 0.50 or greater, they are considered very significant. The significance of loadings varies according to the number of variables under investigation, and should be adjusted downwards for larger samples (Smith 1995). In this research, a rule of thumb approach was adopted; whereby, a factor loading of ± 0.30 was considered to be acceptable in all the solutions presented in this study (chapter six) for both identifying quality factors. However, when attempting to give labels or simple meaningful interpretations for some factors that include many different items attention was often focused on variables with loadings greater than ± 0.50 i. e. the label was chosen to represent mainly those items with higher loadings.

Regression Analysis – is used to estimate a linear relationship between a dependent variable and one or more independent variables. This technique, contrary to correlation analysis, assumes a causal relationship between variables. It is assumed that the dependent (criterion) variable is predictively linked to the independent (predictor) variable. Moreover, regression analysis attempts to predict the values of a continuous, interval scaled dependent variable from the specific values of the independent variable. Multiple regression analysis is an extension of bivariate regression analysis which allows for the simultaneous investigation of the effect of two or more independent variables on a single interval- scaled dependent variable. Thus, a continuous, interval-scaled dependent variable is required in multiple regression, as it is bivariate regression. Interval scaling is also a

requirement for the independent variables. Stepwise regression is a multiple procedure for obtaining coefficients for a multiple regression equation that includes all the variables we want to introduce into the model. What the system does at each step is to add to the equation the variable that is most significantly and strongly related to the dependent variable as possible. Then it looks at the remaining set of variables, whatever variables are left out of the model, and from that pool it selects the variable that is most strongly correlated to the dependent variable.

The objective of this particular statistical procedure is to identify those variables which best predict the variance in airline service quality, and how behavioral and attitudinal dimensions of satisfaction explain the variance in passenger satisfaction toward a specific airline.

The dataset is firstly normalized and once normalization complete; we took a look at a correlation heatmap to see which features correlate well with customer satisfaction.

Best features - Online Booking, Class, and Type of Travel

Worst features - Gate location, Gender, and Departure/Arrival Time Convenient

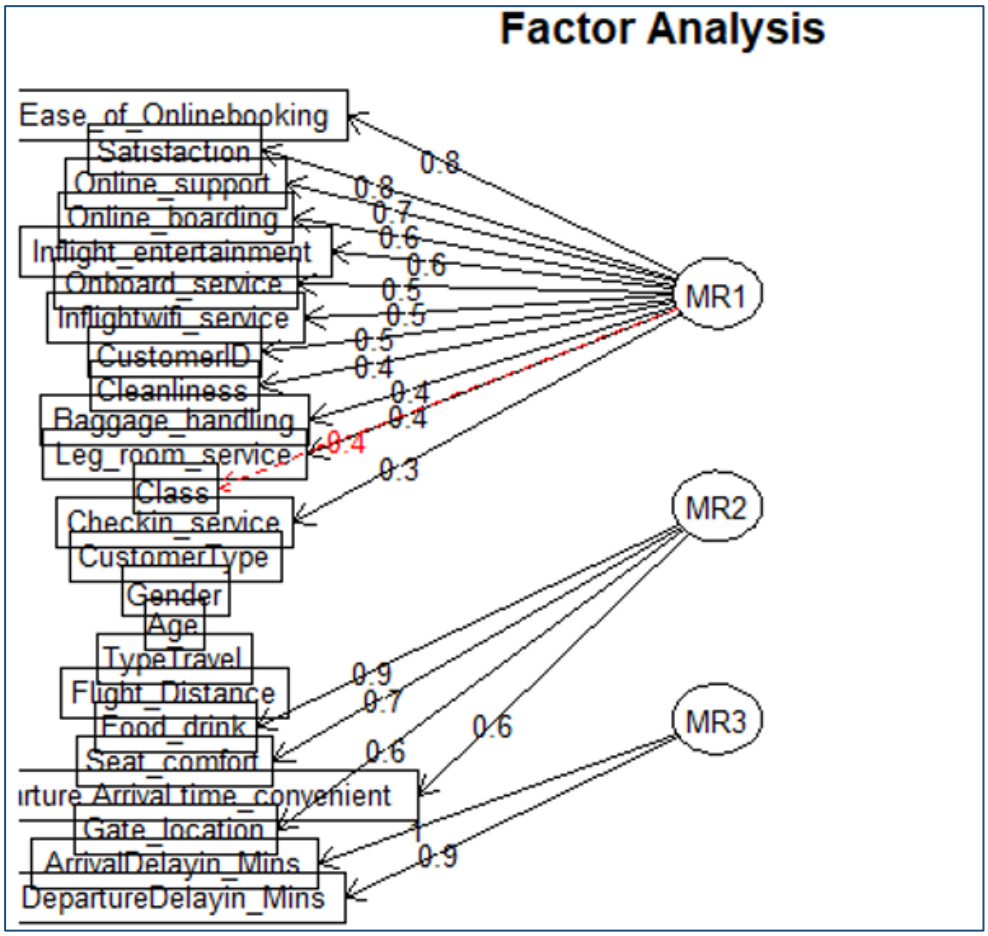
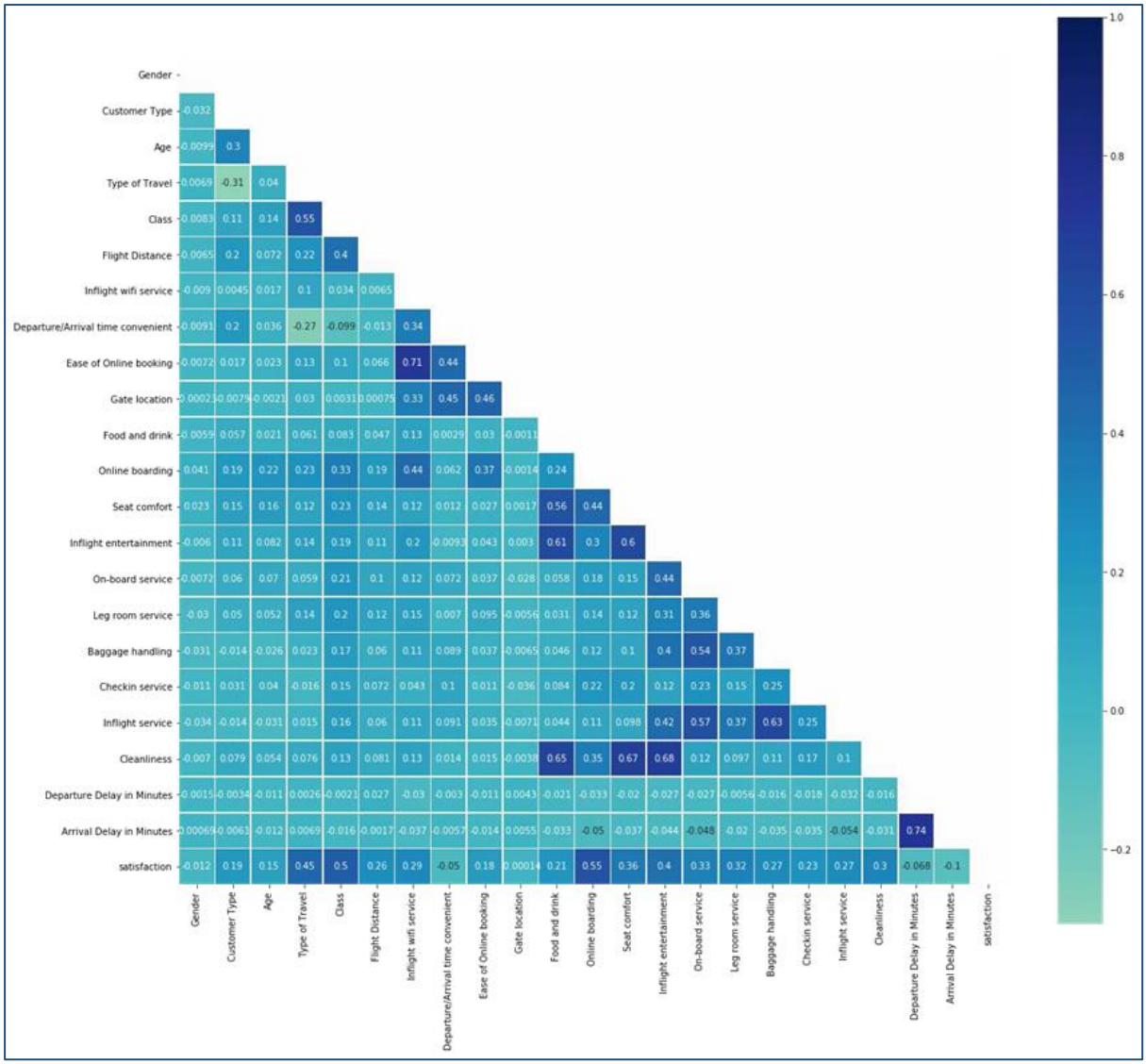


Figure 2: Factor analysis of passenger satisfaction



The dataset was then partitioned into a 80-20 split for training and testing. The model has been built using the training data and the same has been evaluated using the validation data.

7.2 Model Evaluation

Now to the models! We will be trying out a few different models to see which one is the best choice for our problem. I have created a small function below which will train, predict, and evaluate all of our models. We will be evaluating performance of our models with the ROC_AUC metric. This metric is good for classification of a dataset which a relatively balance dataset in terms of our target. We will also be looking at the confusion matrix for our model to best understand how our model is mischaracterizing predictions (Are we seeing majority false positives? etc.)

7.3 Logistic Regression Model

To build a classification model to predict satisfaction, we start with the go-to classification model: Logistic Regression model. Here, we utilize the Logistic regression classification algorithm from the GLM package in R to predict “satisfaction_v2” from the set of independent variables available. Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both.

```
Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -8.438e+00  9.592e-02 -87.972  < 2e-16 ***
Gender                                -9.585e-01  2.352e-02 -40.747  < 2e-16 ***
CustomerType                           2.013e+00  3.571e-02  56.357  < 2e-16 ***
Age                                   -8.049e-03  8.164e-04  -9.859  < 2e-16 ***
TypeTravel                             8.374e-01  3.327e-02  25.172  < 2e-16 ***
Class                                  3.512e-01  1.505e-02  23.330  < 2e-16 ***
Flight_Distance                       -1.105e-04  1.229e-05  -8.989  < 2e-16 ***
DepartureDelayin_Mins                 -4.838e-03  3.122e-04 -15.496  < 2e-16 ***
Seat_comfort                           2.892e-01  1.319e-02  21.922  < 2e-16 ***
Departure.Arrival.time_convenient     -1.883e-01  9.703e-03 -19.403  < 2e-16 ***
Food_drink                            -2.206e-01  1.347e-02 -16.379  < 2e-16 ***
Gate_location                          1.117e-01  1.096e-02  10.191  < 2e-16 ***
Inflightwifi_service                  -5.693e-02  1.263e-02  -4.506 6.60e-06 ***
Inflight_entertainment                 7.039e-01  1.189e-02  59.188  < 2e-16 ***
Online_support                         7.761e-02  1.291e-02   6.014 1.81e-09 ***
Ease_of_Onlinebooking                  2.274e-01  1.665e-02  13.657  < 2e-16 ***
Onboard_service                        3.103e-01  1.178e-02  26.333  < 2e-16 ***
Leg_room_service                       2.233e-01  1.001e-02  22.306  < 2e-16 ***
Baggage_handling                       1.100e-01  1.334e-02   8.243  < 2e-16 ***
Checkin_service                        3.036e-01  9.895e-03  30.682  < 2e-16 ***
Cleanliness                            6.530e-02  1.387e-02   4.710 2.48e-06 ***
Online_boarding                        1.597e-01  1.421e-02  11.244  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 87655  on 63641  degrees of freedom
Residual deviance: 48979  on 63620  degrees of freedom
AIC: 49023
```

Outcome:

From the above summary statistics of the model we built, we observe the coefficient weights and the importance of the variable from p-value. We observe that all the independent variables are significantly important in predicting the “satisfaction”.

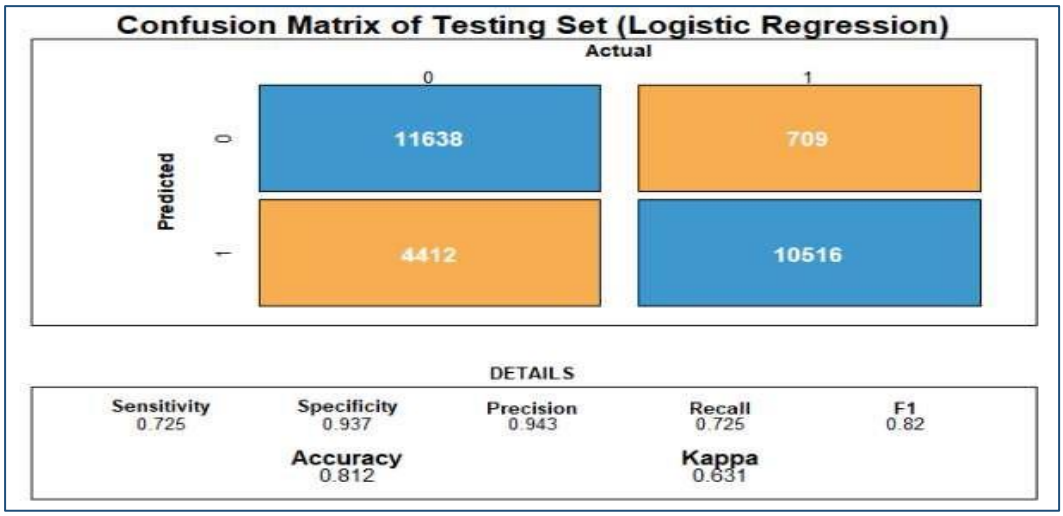
We shall predict the test dataset using the model built. Note that we obtained the predictions in terms of probabilities. Hence, by considering a cutoff of 0.5 (practically used default value), we classify the predicted “satisfaction” into “Yes” or “No”. Let us build the confusion matrix, using the predicted and observed values to assess the performance of the model. As we can see, summary() returns the estimate, standard errors, z-score, and p-values on each of the coefficients. Looks like all the coefficients are significant here. It also gives us the null deviance (the deviance just for the mean) and the residual deviance (the deviance for the model with all the predictors). There's a significant difference between the 87655, along with 48979 degrees of freedom.

```
Confusion Matrix and Statistics
Reference
Prediction neutral or dissatisfied satisfied
neutral or dissatisfied 10020 2267
satisfied 2326 12661
Accuracy : 0.8316
95% CI: (0.8271, 0.836)
No Information Rate: 0.5473
P-Value [Acc > NIR]: <2e-16
Kappa: 0.66
Mcnemar's Test P-Value : 0.3921
Sensitivity: 0.8481
Specificity: 0.8116
Pos Pred Value: 0.8448
Neg Pred Value: 0.8155
Prevalence: 0.5473
Detection Rate: 0.4642
Detection Prevalence: 0.5495
Balanced Accuracy: 0.8299
'Positive' Class: satisfied
```

From the above results, we can observe that:

- The overall accuracy of 83.6% and Kappa value of 66%. That means, our model performed better than a random prediction of the dependent variable.
- Sensitivity value of 84 percent, indicating that our model has correctly identified 84% of the satisfied customers correctly.
- Specificity value of 81 percent, indicating that our model has correctly identified 81 percent of the "neutral or dissatisfied" customers correctly.

Model Performance



Logistic Regression provided an accuracy of 81.2% with low F1 and Sensitivity value. Although the logistic regression performed well on the minority class, it fared badly to determine the majority of the predicted class. Out of 14928 records, 4412 records have been predicted incorrectly.

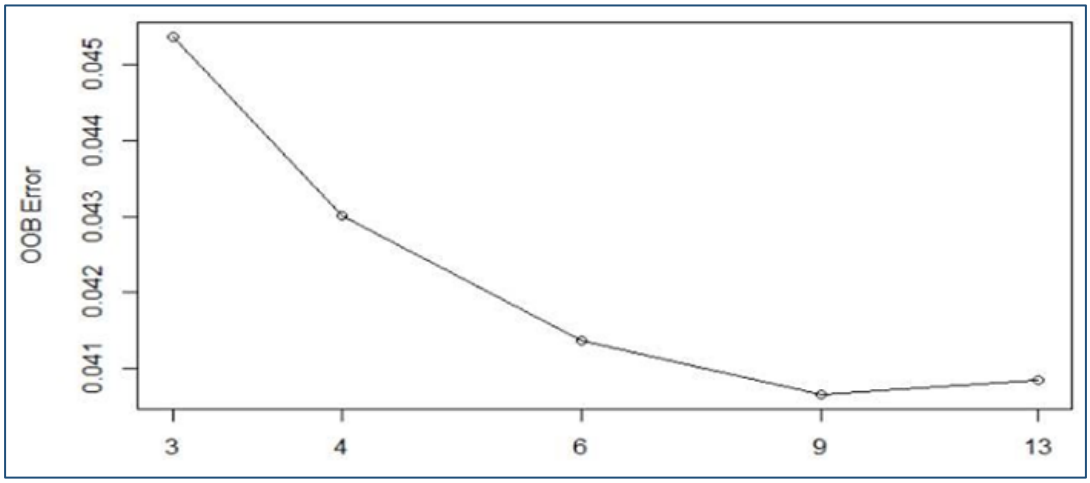
Variables of Importance for Logistic Regression

From the above figure, we can see that Inflight_entertainment, CustomerType, Gender, Checkin_service, Onboard_service, TypeTravel, Class, Leg_room_service, Seat_comfort, DepartureDelayin_Mins, Ease_of_OnlineBooking and Online_boarding are the most important factors driving the Satisfaction of the customers.

7.4 Random Forest

The random forest algorithm works by aggregating the predictions made by multiple decision trees of varying depth. Every observation is fed into every decision tree. The most common outcome for each observation is used as the final output. A new observation is fed into all the trees and taking a majority vote for each classification model. We use a statistical resampling technique called bootstrapping to fetch random samples to build each decision tree. We use Random Forest to increase the accuracy of our predictions.

The optimal number of the columns that needs to be applied to make the trees as much as different from one another is shown below.

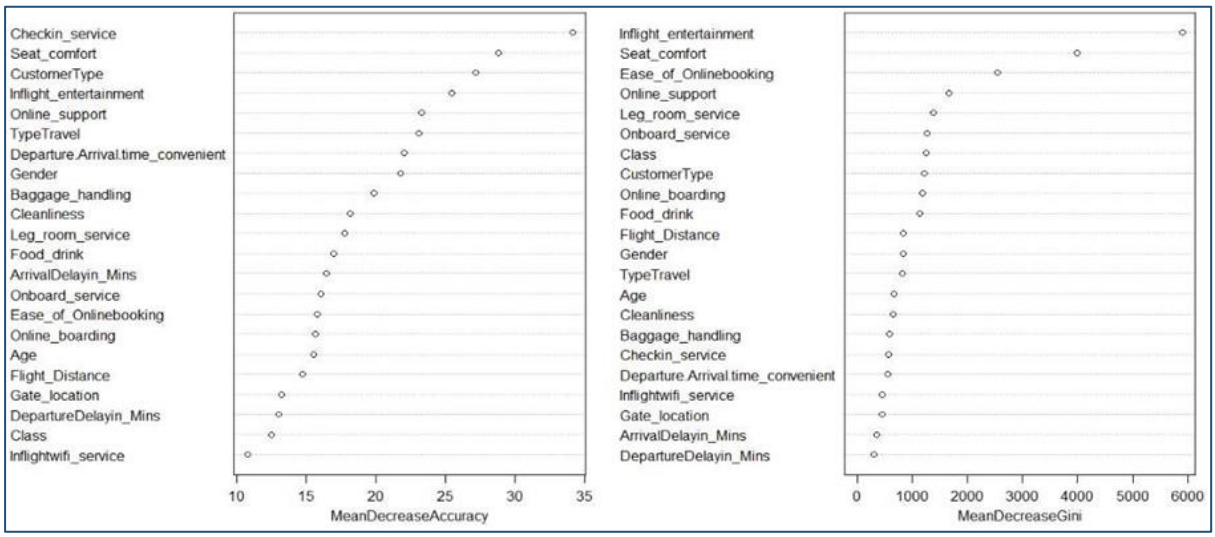


The above figure shows that at mtry = 9, the accuracy predicted is maximum at 98.2%.

Using the above mtry to tune the Random Forest to build 500 tree initially. After we plot the OOB accuracy with 501 trees, we observe that it reaches maximum at ~50 trees. So, we further tune the tree to 51 trees(to choose odd number).

```
Confusion Matrix and Statistics
Reference
Prediction neutral or dissatisfied satisfied
neutral or dissatisfied 10024 2322
satisfied 2243 12685
Accuracy: 0.982
95% CI: (0.8281, 0.837)
No Information Rate: 0.5502
P-Value [Acc > NIR]: <2e-16
Kappa: 0.662
Mcnemar's Test P-Value: 0.2483
Sensitivity: 0.8453
Specificity: 0.8172
Pos Pred Value: 0.8497
Neg Pred Value: 0.8119
Prevalence: 0.5502
Detection Rate: 0.4651
Detection Prevalence: 0.5473
Balanced Accuracy: 0.8312
'Positive' Class: satisfied
```

Variables of Importance for Random Forest



The above plot shows that if we remove the variables Checkin_service, Seat_comfort, Customer_Type, Inflight_entertainment, Online_support, Departure.Arrival.time_convenient, Gender and Leg_room_service the accuracy of the model will reduce drastically.

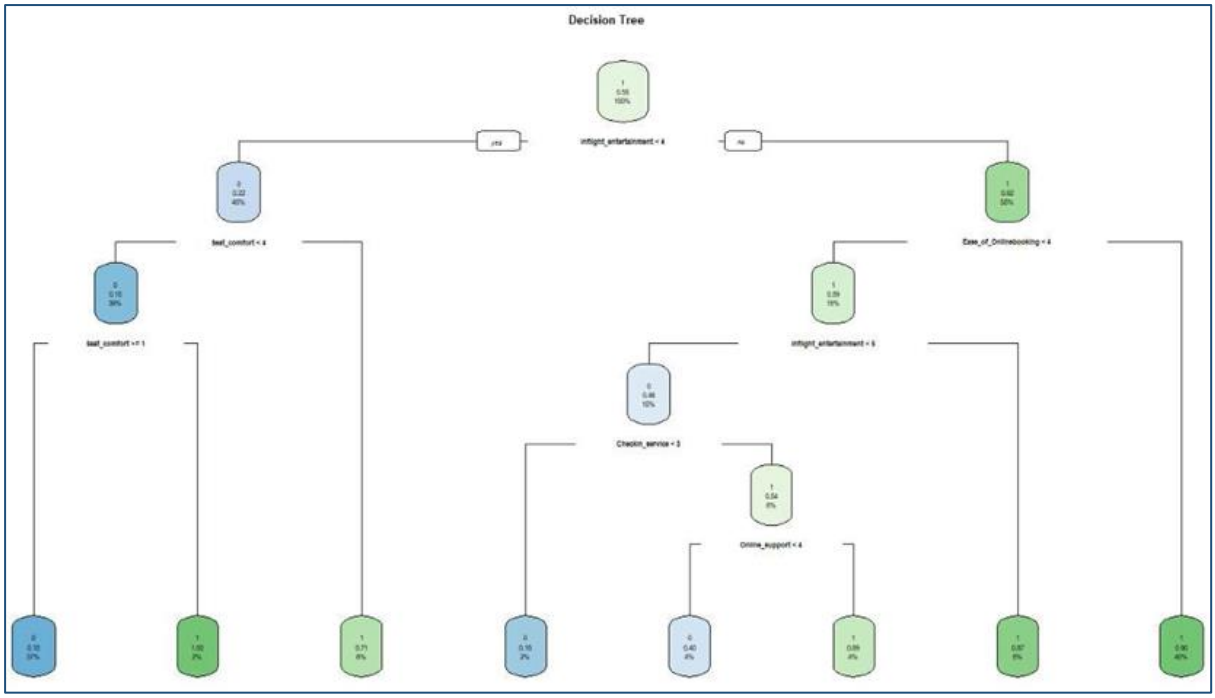
The MeanDecreaseGini denotes that if we remove the top variable, the purity of the dataset will be reduced.

7.5 Decision Tree

We use a decision tree algorithm, CART (Classification and Regression Tree) for classifying the dependent variable based on the set of independent variables. This algorithm repeatedly partitions the data into multiple-subsets so that the leaf nodes are pure or homogeneous. Here we ensure that our categorical variables are in factor form so that the CART algorithm uses classification tree rather than regression tree. Following we visualize the model we built. Decision Trees are based on separating records into subgroups by creating splits on predictors. These splits create logical rules that are transparent and easily understandable, for example, “IF Age < 55 AND Education > 12 THEN class = 1.”

The resulting subgroups should be more homogeneous in terms of the outcome variable, thereby creating useful prediction or classification rules. We discuss the two key ideas underlying trees: recursive partitioning (for constructing the tree) and pruning (for cutting the tree back).

To plot the Classification Tree, we first split the dataset into Training Set and Testing Set and validate if the data is balanced after splitting. We see that the data is evenly split according to the target variable:



Observation:

We started with a 55% and 45% of customers who are overall satisfied and unsatisfied with Virgin airlines services. The Gini has helped in getting a better percentage of people who are likely to be satisfied with Virgin airlines services. This number stands at 90.2%. Also, most importantly, the people likely to be satisfied with Virgin airlines' services are those who rate Ease of Online booking with a rating greater than 4 to mean more than acceptable.

IF (Inflight_entertainment >= 4) AND (Ease_of_OnlineBooking >= 4) THEN "Satisfied" IF (Inflight_entertainment < 4) AND (Seat_comfort >= 1) THEN "Satisfied"
IF (Inflight_entertainment >= 4) AND (Ease_of_OnlineBooking < 4) AND (Inflight_entertainment = 5) THEN "Satisfied"

IF (Inflight_entertainment >= 4) AND (Ease_of_OnlineBooking < 4) AND (Inflight_entertainment = 5) AND (Checkin_service >= 3) AND (Online_support >= 4) THEN "Satisfied"

The algorithm built a classification tree, as we observe that the first or the more valuable variable we have "Inflight entertainment" being rated 1-3 we classify them to be "neutral or dissatisfied" class on left node, which is further classified based on the rating given on "seat comfort". If seat comfort is rated 1-3, we finally decided them to be "neutral or dissatisfied", else classified them to be "satisfied". Similarly, the tree progresses finally into 7 leaf nodes. Note that the factor levels are normalized (Normalized Value (Rating Value): 0(0), 0.2(1), 0.4(2), 0.6(3), 0.8(4), 1(5)).

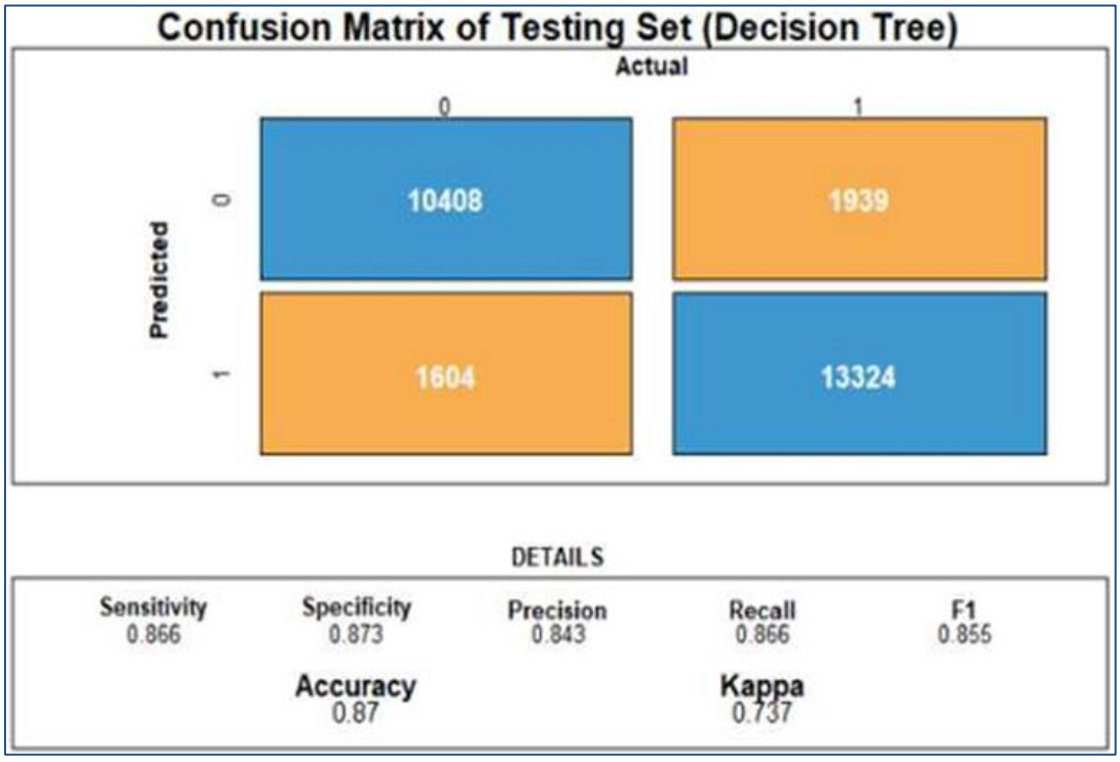
We use this model to predict the test data set and observe the confusion matrix to understand the model performance:

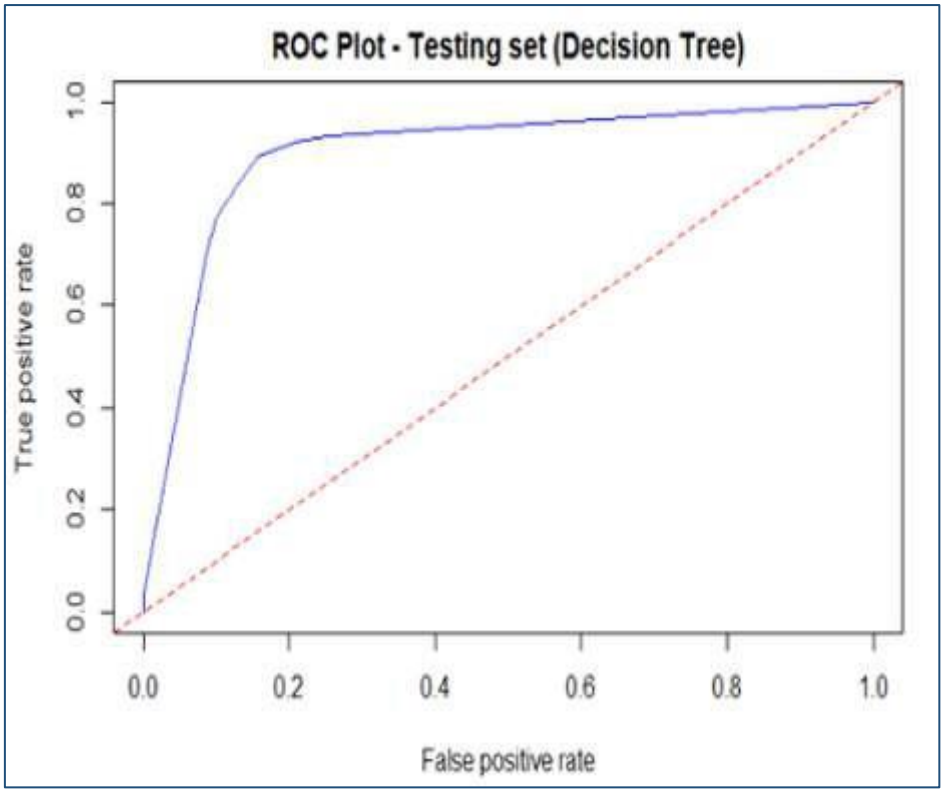
```
Confusion Matrix and Statistics
Reference
Prediction neutral or dissatisfied satisfied
neutral or dissatisfied 10638 1997
satisfied 1708 12931
Accuracy: 0.8642
95% CI: (0.86, 0.8682)
No Information Rate : 0.5473
P-Value [Acc > NIR] : < 2.2e-16
Kappa: 0.7264
Mcnemar's Test P-Value: 2.229e-06
Sensitivity: 0.8662
Specificity: 0.8617
Pos Pred Value: 0.8833
Neg Pred Value: 0.8419
Prevalence: 0.5473
Detection Rate: 0.4741
Detection Prevalence: 0.5367
Balanced Accuracy: 0.8639
'Positive' Class: satisfied
```

From the above results, we can observe that:

- The overall accuracy of 86% and Kappa value of 72%. That means, our model performed better than a random prediction of the dependent variable.
- Sensitivity value of 86 percent, indicating that our model has correctly identified 86% of the satisfied customers correctly.
- Specificity value of 86 percent, indicating that our model has correctly identified 86 percent of the "neutral or dissatisfied" customers correctly.

Model Performance



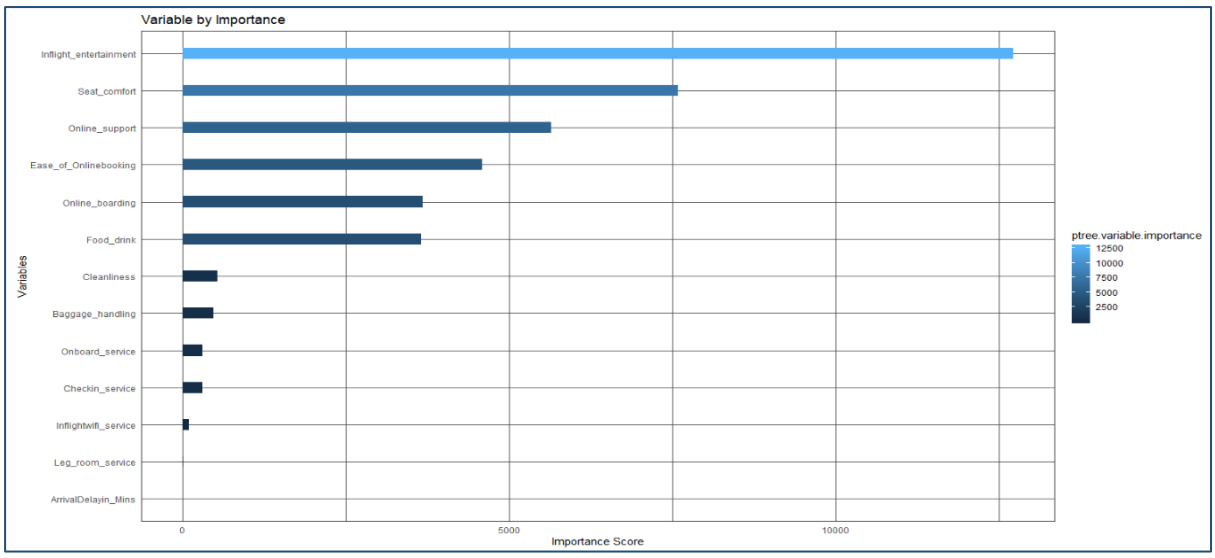


Confusion matrix on the CART Model on the Training and Test data shows that the model performed equally well in both the Training and Test Set, with an Accuracy of 87% in the Test Set. The Sensitivity and Specificity are also very close to each other. Out of 14928 records for the predicted class in the Test set, only 1604 records have been predicted incorrectly. However, the F1 and Precision values are pretty low compared to Accuracy.

A Receiver Operating Characteristic (ROC) curve plots the false alarm rate against the hit rate for a probabilistic forecast for a range of thresholds. The true positive rate (Sensitivity) is plotted to the false positive rate (1-Specificity) for different cut-off points of a parameter. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups.

Variables of Importance for Decision Tree

The overall importance of a feature in a decision tree can be computed in the following way: Go through all the splits for which the feature was used and measure how much it has reduced the variance or Gini index compared to the parent node. The sum of all importance is scaled to 100. This means that each importance can be interpreted as share of the overall model importance.



From the above figure, we can see that Inflight_entertainment, Seat_comfort, Ease_of_OnlineBooking, Online_support and Online_boarding are the most important factors driving the Satisfaction of the customers.

7.6 KNN Model

For K-NN, the critical part is to choose the k predictor points. As we have 63642 rows in total, sqrt (63642) approx. 205 should be the optimal 'k'. However, we try to loop through k = 3:19 and try to figure out the optimal 'k' in this scenario and determine the accuracy and precision on the test data for the obtained classifier. This initial model is built considering all the independent variables.

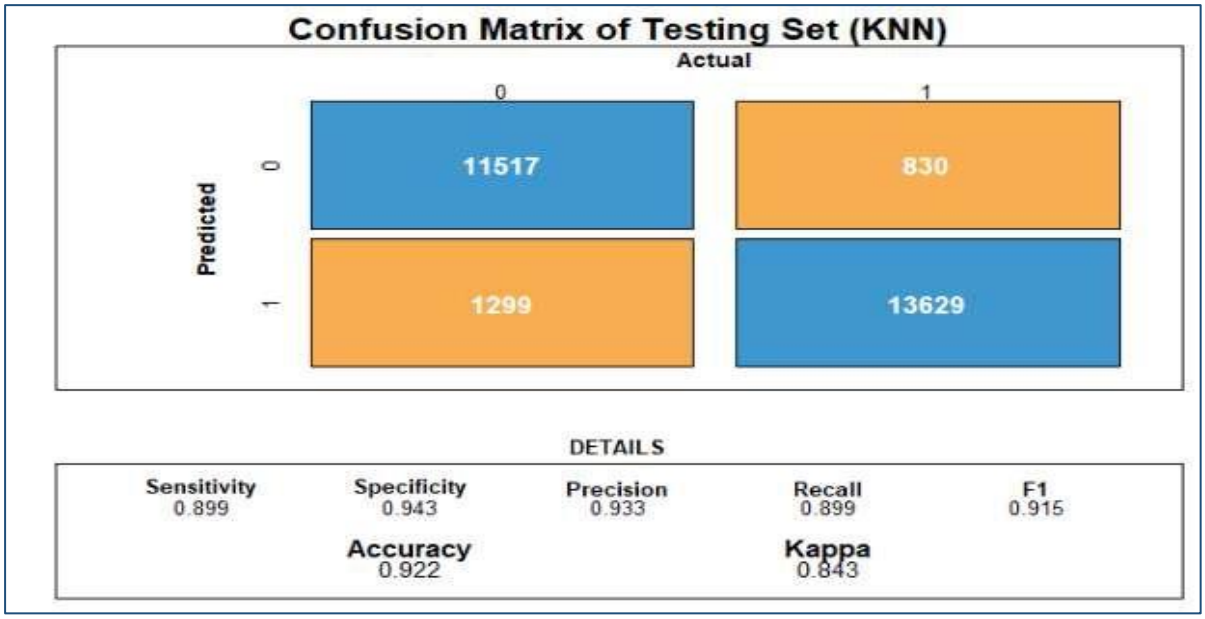
```
[1] "The value of K is :" "3"
knn_fit
0 1
0 11459 888
1 1271 13657
[1] "Accuracy is :" "0.920843263061412"
0 1
"Precision is :" "0.928079695472584" "0.914857984994641"
[1] "The value of K is :" "5"
knn_fit
0 1
0 11501 846
1 1279 13649
[1] "Accuracy is :" "0.922089825847846"
0 1
"Precision is :" "0.93148133149753" "0.914322079314041"
[1] "The value of K is :" "7"
knn_fit
0 1
0 11517 830
1 1299 13629
[1] "Accuracy is :" "0.921943171402383"

0 1
"Precision is :" "0.932777192840366" "0.91298231511254"
[1] "The value of K is :" "9"
```

```
knn_fit
0 1
0 11530 817
1 1324 13604
[1] "Accuracy is :" "0.921503208065995"
0 1
"Precision is :" "0.933830080181421" "0.911307609860665"
[1] "The value of K is :" "11"
knn_fit
0 1
0 11536 811
1 1354 13574
[1] "Accuracy is :" "0.920623281393217"
0 1
"Precision is :" "0.934316028184984" "0.909297963558414"
[1] "The value of K is :" "13"
knn_fit
0 1
0 11548 799
1 1372 13556
[1] "Accuracy is :" "0.920403299725023"
0 1
"Precision is :" "0.935287924192111" "0.908092175777063"
[1] "The value of K is :" "15"
knn_fit
0 1
0 11536 811
1 1378 13550
[1] "Accuracy is :" "0.91974335472044"
0 1
"Precision is :" "0.934316028184984" "0.907690246516613"
[1] "The value of K is :" "17"
knn_fit
0 1
0 11530 817
1 1384 13544
[1] "Accuracy is :" "0.919303391384051"
0 1
"Precision is :" "0.933830080181421" "0.907288317256163"
[1] "The value of K is :" "19"
knn_fit
0 1
0 11536 811
1 1396 13532
[1] "Accuracy is :" "0.919083409715857"
0 1
"Precision is :" "0.934316028184984" "0.906484458735263"
```

We see that k=5, gives us the best result on the test set with high accuracy and precision of the predictor class.

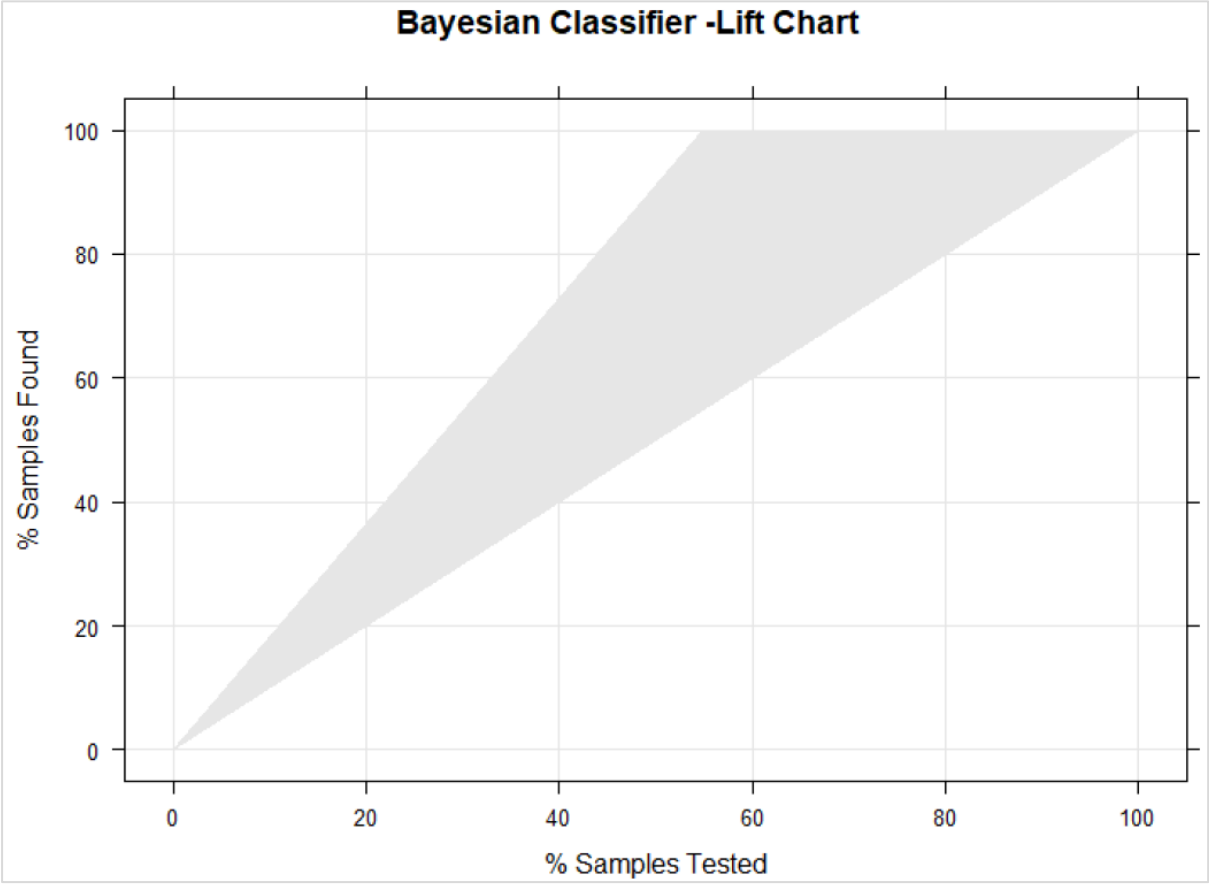
Model Performance



Outcomes:

Out of 14928 records in the Test set, only 1299 records have been predicted incorrectly, providing a high Accuracy, F1 and Precision values. Overall, K-NN performed really well and provided much better performance than CART and Logistic Regression.

7.7 NAÏVE BAYES – Classification Algorithm



```
Confusion Matrix and Statistics
pc 1 0
1 13071 1949
0 1792 10329
Accuracy: 0.8622
95% CI: (0.858, 0.8662)
No Information Rate: 0.5476
P-Value [Acc > NIR]: < 2e-16
Kappa: 0.7215
Mcnemar's Test P-Value: 0.01076
Sensitivity: 0.8794
Specificity: 0.8413
Pos Pred Value: 0.8702
Neg Pred Value: 0.8522
Prevalence: 0.5476
Detection Rate: 0.4816
Detection Prevalence: 0.5534
Balanced Accuracy: 0.8603
'Positive' Class: 1
```

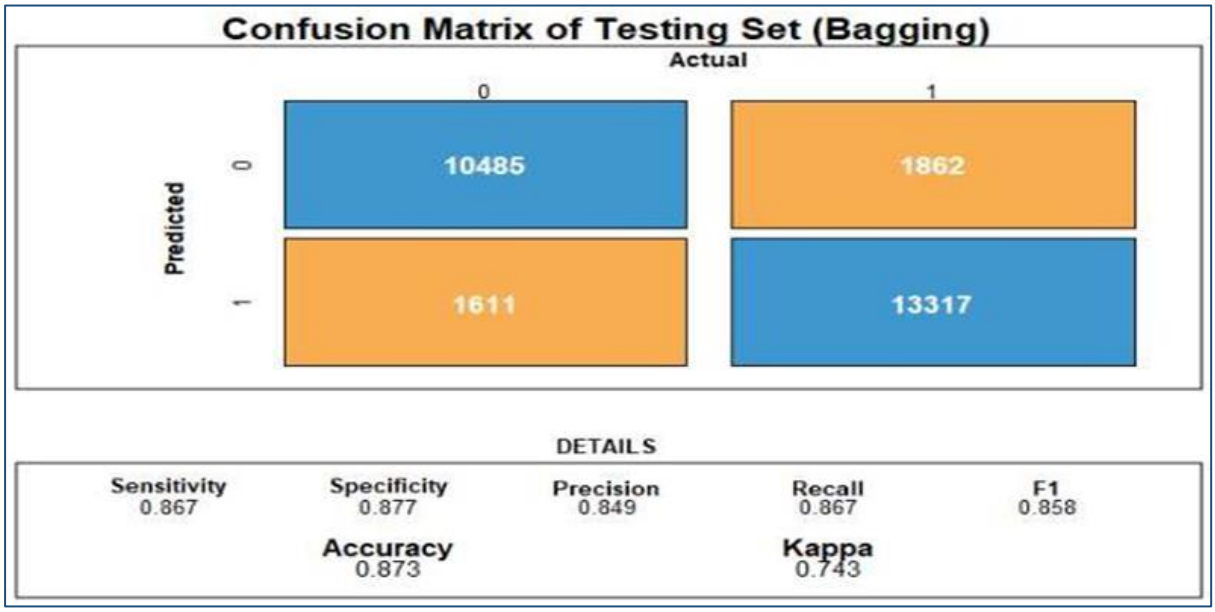
From the above results, we can observe that:

- The overall accuracy of 86% and Kappa value of 72%. That means, our model performed better than a random prediction of the dependent variable.
- Sensitivity value of 87 percent, indicating that our model has correctly identified 87% of the satisfied customers correctly which is higher than all the models.
- Specificity value of 84 percent, indicating that our model has correctly identified 84 percent of the "neutral or dissatisfied" customers correctly which is lower than CART model.
- This model has performed good in predicting the satisfied customers but failed to perform that well when comes to neutral or dissatisfied customers.

7.8 Bagging

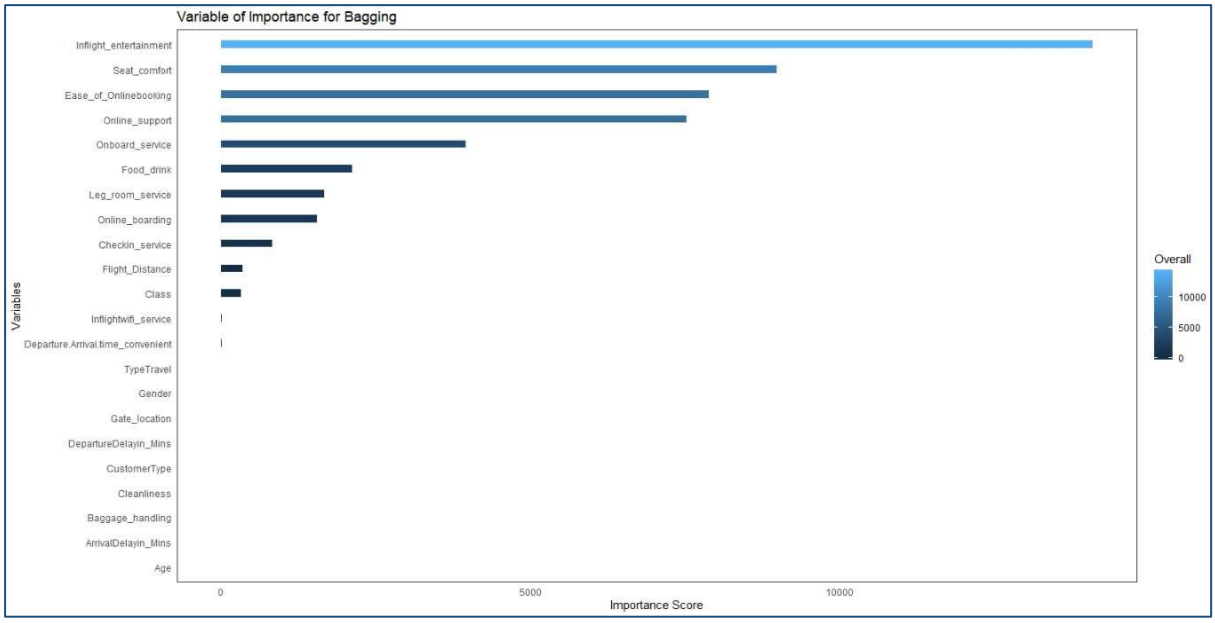
Bagging (aka Bootstrap Aggregating): is a way to decrease the variance of prediction by genera ting additional data for training from the original dataset using combinations with repetitions to produce multisets of same cardinality as the original data.

Model Performance



We see that we have slightly improved on the Accuracy over the CART Algorithm. We also see that prediction of customers being unsatisfied increased over K-NN, however prediction of "Satisfaction" decreased. The model can predict 13317 records correctly for the prediction class in the test set. Lower values of F1 and Precision makes this model unstable.

Variables of Importance for Bagging



From the above figure, we can see that Inflight_entertainment, Seat_comfort, Ease_of_OnlineBooking, Online_support, Onboard_service, Food_drink, Leg_room_service, Online_boarding and Checkin_service are the most important factors driving the Satisfaction of the customers.

7.9 Boosting

The aim of Boosting is to train the weak learners sequentially.

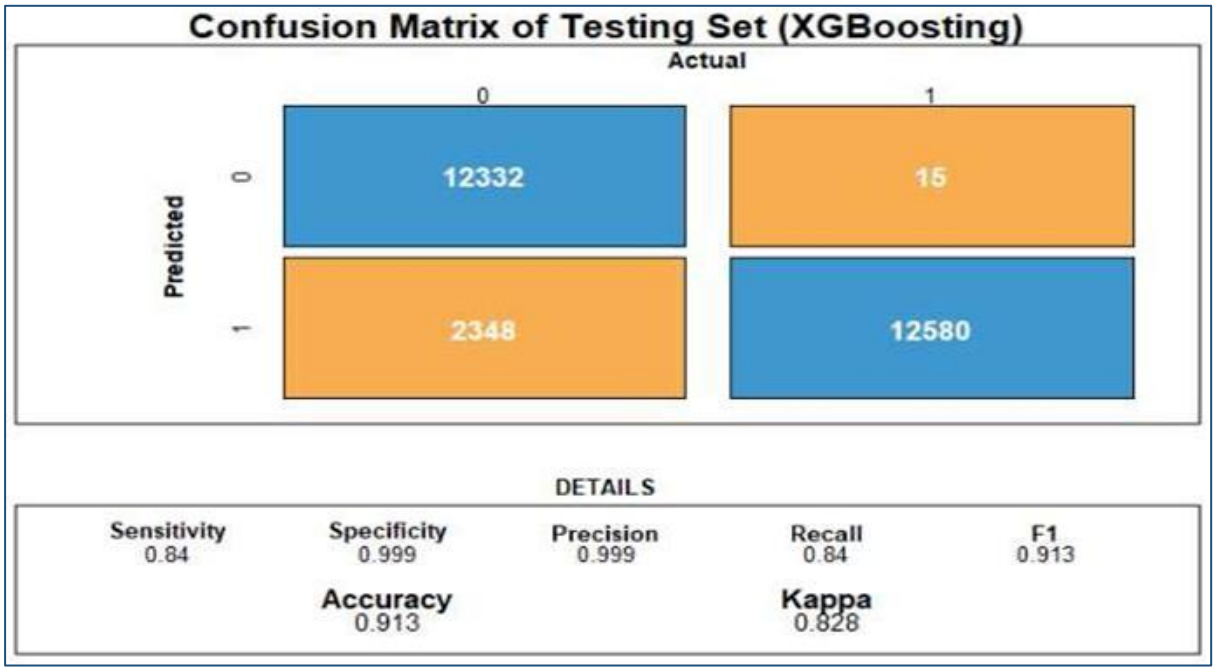
Applying **XGBoost** (Extreme Gradient Boosting) Algorithm: Since this method has higher performance over Gradient Boosting, we are choosing this algorithm.

To find the optimal parameters let us run a code in a loop to figure out the values at 90% confidence:

```
## Findng the best fit parameters
tp_xgb<-vector()
lr <- c(0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 1)
md<-c(1,3,5,7,9,15,17,19)
nr<-c(2, 50, 100)
for (i in lr) {
for (j in md) {
for (k in nr) {
xgb.fit <- xgboost(
data = XGBoost_features_train,
label = XGBoost_label_train,
eta =i,
max_depth = j,
nrounds = k,
nfold = 10,
objective = "binary:logistic", # for regression models
```

```
verbose = 1, # silent,
early_stopping_rounds = 10 # stop if no improvement for 10 consecutive trees
)
testdata_XGBoost$xgb.pred.class <- predict(xgb.fit, XGBoost_features_test)
tp_xgb<-cbind(tp_xgb,sum(testdata_XGBoost$Satisfaction==1 & testdata_XGBoost$x
gb.pred.class>=0.90))
#if your class=1 and our prediction=0.5, we are going to display it with the n
ext line compare the same algorithm for different values
}
}}
```

Model Performance



This is an improvement from the previous logistic regression model and KNN with an accuracy of 91.3% where 2348 out of 14928 predictions were incorrect for 'Satisfied' customers even at 90% confidence which was not the case for the modified logistic regression, K-NN or CART model also. However, the prediction for the majority class is low compared to other models. The Accuracy is high as the determination of unsatisfied customers were very high.

Variables of Importance for XGBoost

Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
Inflight_entertainment	0.385132353	0.11307644	0.05126119
Seat_comfort	0.222696118	0.20493127	0.14646054
Ease_of_Onlinebooking	0.056470341	0.06740250	0.05777055
CustomerType	0.044134846	0.06173930	0.04556550
TypeTravel	0.043242034	0.05324744	0.08299430
Class	0.027376297	0.02797142	0.04068348
Gender	0.025903945	0.02708683	0.02685110
Departure.Arrival.time_convenient	0.024799798	0.03599658	0.04312449
Online_support	0.021437847	0.04104650	0.03580146
Checkin_service	0.020026436	0.04069456	0.03010578
Gate_location	0.019254664	0.01898804	0.03742880
Onboard_service	0.017275510	0.04172357	0.02766477
Leg_room_service	0.016966397	0.02065531	0.03336046
Cleanliness	0.015958848	0.04099385	0.04231082
Online_boarding	0.012654087	0.02911587	0.02522376
Flight_Distance	0.010555022	0.02548481	0.05126119
Baggage_handling	0.008551961	0.03173575	0.03336046
Age	0.008345169	0.04362060	0.07241660
Food_drink	0.008343372	0.02252887	0.03580146
ArrivalDelayin_Mins	0.006856850	0.03121815	0.02766477
Inflightwifi_service	0.002773947	0.01001901	0.02766477
DepartureDelayin_Mins	0.001244157	0.01072334	0.02522376

8. Model Results Comparison

From all the model outputs that we see above, Random Forest outperformed all other models. It provided an Accuracy of 95% on the test set and an Accuracy of 98% on the training set. The KS value is very high at 90% with a ROC value of 99%

	Logistic Regression	*Decision Tree*	*Random Forest*	*KNN*	*NAÏVE BAYES*	*Bagging*	*XGBoosting*
Accuracy	*0.937*	*0.870*	*0.982*	*0.922*	*0.862*	*0.873*	*0.913*
Kappa	*0.631*	*0.737*	*0.962*	*0.843*	*0.721*	*0.743*	*0.828*
Sensitivity	*0.725*	*0.866*	*0.945*	*0.899*	*0.879*	*0.867*	*0.840*
Specificity	*0.937*	*0.873*	*0.979*	*0.943*	*0.841*	*0.877*	*0.999*
Precision	*0.943*	*0.843*	*0.949*	*0.933*	*0.870*	*0.849*	*0.999*

Table: Comparison of the model performance of the models we built on a common metrics

9. Interpretation from the Best Model

Reasons for choosing Random Forest as the model:

- It can handle binary features, categorical features, and numerical features.
- There is very little pre-processing that needs to be done.
- The data does not need to be rescaled or transformed.
- Random forest is great with high dimensional data since we are working with subsets of data which is the case in our dataset.
- Random forest handles outliers by essentially binning them. It is also indifferent to non-linear features.

10. Conclusions – Business Insights

Since the problem statement is to determine the “Satisfaction” of the customers flying by Virgin Airlines and to find out the drivers of satisfaction, we can clearly see that the most important factors driving the satisfaction are:

- ✓ Checkin_service
- ✓ Seat_comfort
- ✓ Customer_Type
- ✓ Inflight_entertainment
- ✓ Online_support
- ✓ Departure.Arrival.time_convenient
- ✓ Gender
- ✓ Leg_room_service
- ✓ Ease_of_OnlineBooking
- ✓ Class
- ✓ Food_drink

If Virgin Airlines can concentrate and improve on these parameters, the Satisfaction can go up by a huge extent. However, a better way of understanding the customer satisfaction would be to get a written feedback from customers and run text analysis on them. A lot of interesting aspects can come out which the airlines might not have thought about. Also, the like scale can often be misleading as customers would not concentrate when the rating is from 0-5 (with 5 being highest) and when 1 is highest.

Virgin Airlines should aim at offering a complex travel concept, sensitive to individual needs at a relatively high price. The customer is buying a concept and will himself be measuring the relational quality between the different services. It would be meaningless to divide the core, supporting and facilitating services into small parts and measure the "quality" of the individual pieces. For example, it does not help much that the "checkin" and "reservation" services were perfect and the dinner was delicious if a delay caused the customer a two or three-hours wait at the airport. The quality is not measured vertically, but horizontally for the entire service package. The important thing is not the average quality from a check list, but the total quality of the service package; and that can only be known by the customer in his own subjective way.

Service quality is a matter of controlling details in the service delivery. Quality development means improving all the parts of service chain and seeing the whole. Therefore, the following points are very important for quality development:

- Highlighting the personal quality of airline staff.
- Being sensitive to signals of dissatisfaction by "reading" customers and thus discovering quality defects before the customer complaints and making it easier for the customer to complain.
- Looking after the customer and correcting faults which have arisen.
- Providing generous compensation
- Providing clear, rapid and "truthful" information.

11. Recommendations

In order to be successful, the Virgin Atlantic Airlines will have to achieve high scores on the following issues:

- ✓ Building on the results of the psychographic segmentation the following recommendations can be made:
  - The "Punctual Passengers", the airline should concentrate on the importance of convenient flight schedules and supporting aspects. Therefore, Virgin Atlantic Airlines should focus on promotional strategy, that shows the appropriateness of its flights schedule and its ability to reach many destinations all over the world.
- ✓ Improve the quality of services provided in first and business classes. It was found that the quality of services in business class is not perceived to be of a of high standard.
- ✓ Examine those areas of weak evaluations of the quality of services as perceived by different categories of passengers. This will help the airline to solve the problems and to provide certain kinds of services that were not covered before. Moreover, it will affect passenger's satisfaction and their willingness to fly with the airline in the future. Achieve a high degree of sensitivity to individual passengers' needs. To succeed in achieving that, the airline should take into consideration the following principles:
  - The passenger is buying a travel concept according to individual needs and utility and not a number of services that are measurable. This is individual service matched with mass services.
  - Advanced information systems are imperative in order to offer a complex travel programme and allow for the possibility of sensitivity to individual needs and utility. The historical information pertaining to the travel habits of customer is of crucial analytical value for evaluating both service quality and the profit of each service in the system, as well as for offering new service opportunities.
- ✓ In-flight services need to include the following:

  (i) Cabin-staff services: This factor includes services related to the contact personnel during the flight. It consists of the special attention and personal touch given to passengers by personnel. This factor is also linked to variables that describe the knowledge and skills required to achieve effective service delivery, and to the smart appearance of employees. As seen, this factor contains activities representing direct interaction between passengers and airline employees.
  The nature of such interaction has always been recognised as an important determinant of satisfaction with a service.

  (ii) Tangibles: this factor is associated with menu selection, food quality and quantity, and includes also plane characteristics. Many researchers have pointed out that flight attendants and meals are quite visible to passengers and would affect overall customer-perceived service quality. In addition, plane characteristics such as plane size had been recognised as an important quality of service variable. The existence of this factor is also supported by Jones and Cooke (1981) who found that there exists evidence that air travellers indeed discriminate among types of aircraft according to a number of flight-specific attributes.

(iii) Communications: This factor includes variables such as assistance in case of delays, clear announcements, clear signs at airport, and interesting in-flight entertainments. The "Communications" factor is considered important by the passengers who contributed to this study. A high percentage of them (25.2 %) perceived a problem with the quality of audio-visual materials (one of the communications' variables), and the majority of complaints concerned with the quality of sound and music in the plane.

- ✓ Achieve a high degree of customer satisfaction with the total service package, including the comparative measurements of the definitive concepts of service from month to month, creating the capability for quick trouble-shooting (i. e. eliminating the undesirable elements of the service package).
- ✓ Airline administrators should focus their effort on specific areas of quality that had greater influence in explaining the consumer's intent to behave and their satisfaction. If only limited resources are available to implement service quality improvements, ensuring that the promised service is performed accurately, dependably, and with great care of individual's needs, will offer the best return in passenger satisfaction and passenger intentions for repeat business (satisfaction) in the future.
- ✓ Service quality is a matter of controlling details in the service delivery. Quality development means improving all the parts of service chain and seeing the whole. Therefore, the following points are very important for quality development:
    - Highlighting the personal quality of airline staff.
    - Being sensitive to signals of dissatisfaction by "reading" customers and thus discovering quality defects before the customer complaints and making it easier for the customer to complain.
    - Looking after the customer and correcting faults which have arisen.
    - Providing generous compensation
    - Providing clear, rapid and "truthful" information.
- ✓ This factor includes five items measuring cleanliness of airport facilities, efficient check-in procedures, quick baggage handling, efficient security procedures and safety perceptions during the flight. The presence of the last item with the other four airport services can be explained by the fact that assuring safety during the flight depends heavily on efficient security procedures and good maintenance programmes which actually were conducted on the ground (at the airport). This factor was found to explain most of the variance (30%) and contains many important services that were identified in the previous literature. For example, baggage handling and check-in procedures are an important aspect of the ground services; they affect punctuality in departure and, therefore, affect passenger satisfaction as will be seen later.
- ✓ It is important here to emphasise that better co-ordinated activities between the airline and the companies providing ground transport is required and the offering or service system should be so designed that the different customer needs are satisfied. The core service is often described in terms of the flight, but a number of support services are required to meet the needs of the customer in connection with the flight. Thus, transport to and from the airport is an essential part of the total service package and is often experienced as an integrated part of the core service.

- ✓ Finally, according to the theoretical and practical contributions, the airline can benefit by using the methodology involved in producing such significant results. For example, the airline can use the instrument in several ways:
  - The airline could do a market study of a route that is losing revenue. By applying the current survey assessment, they can identify what customers on this route need from an airline and how they evaluate the service provided by an airline. This will allow the airline to identify where problems occurred.
  - The survey assessment (questionnaire) can be used as an in-flight survey instrument. The airline can use the information gathered to continuously assess passengers' needs. The information gathered can be statistically compared on a monthly or quarterly basis. Furthermore, the instrument can be modified so that an airline can examine a particular dimension of customer service.
  - The airline can use the questionnaire during problematic flight operations, such as delays relating to weather or mechanical reasons. Passengers tend to be honest in their opinions when they feel that they have been inconvenienced. Therefore, airlines can use the survey in several ways.
    - First, they can give it to passengers so that they can express their frustration. When passengers are allowed to express their opinions and they know that the information will be sent directly to the upper management, they are more likely to remain rational about the situation.
    - Second, the airlines can use the information gathered to assess how passengers view their performance during a crisis situation. Also, the airline can use the information gathered to see how the airline actually performed and compare this with a performance criterion of how the airline would like to perform during a crisis situation.