Airline Passenger Satisfaction

Aadidev Sooknanan
816003022

April 16, 2021

Contents

1 Overview

1.1 Objective

The aim of this project aims to predict airline passenger satisfaction based on various factors influencing the overall airline experience.

1.2 Dataset

The data for this project was gotten from Kaggle.com and uploaded by user *TJ Klein* [2]. The parameters are a mix of categorical and numerical as follows: **Categorical**: Gender, Customer Type, Type of Travel, Class, **Numerical**: Flight Distance, Inflight Wifi, Departure Time Convenient, Ease of Online Booking, Gate Location, Food and Drink, Online Boarding, Seat Comfort, Inflight Entertainment, Onboard Service, Leg Room Service, Baggage Handling, Checkin Service, Inflight Service, Cleanliness, Departure Delay, Arrival Delay. The class to be predicted is *satisfaction*: Satisfied or Neutral/Negative. The data-set also contained some non-informative attributes such as *X* and Passenger ID, which were dropped prior to performing any analyses.

1.3 Problem Formulation

The problem will be formulated as a Generalised Linear Model, followed by sigmoid activation for the purpose of classification

2 R Implementation

2.1 Loading and Preprocessing

Firstly, the CSV file is loaded into a variable called df, and the first few rows are previewed using the head function

```
df <- read.csv("data/train.csv")
df <- df[complete.cases(df), ]
head(df)
```

```
##   X     id Gender     Customer.Type Age  Type.of.Travel    Class
## 1 0  70172   Male    Loyal Customer  13 Personal Travel Eco Plus
## 2 1   5047   Male disloyal Customer  25 Business travel Business
## 3 2 110028 Female    Loyal Customer  26 Business travel Business
## 4 3  24026 Female    Loyal Customer  25 Business travel Business
## 5 4 119299   Male    Loyal Customer  61 Business travel Business
## 6 5 111157 Female    Loyal Customer  26 Personal Travel      Eco
##   Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
## 1             460                     3                                 4
## 2             235                     3                                 2
## 3            1142                     2                                 2
## 4             562                     2                                 5
## 5             214                     3                                 3
## 6            1180                     3                                 4
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1                      3             1              5               3
## 2                      3             3              1               3
## 3                      2             2              5               5
## 4                      5             5              2               2
## 5                      3             3              4               5
## 6                      2             1              1               2
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1            5                      5                4                3
## 2            1                      1                1                5
```

```
## 3            5                     5                4               3
## 4            2                     2                2               5
## 5            5                     3                3               4
## 6            1                     1                3               4
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1                4               4                5           5
## 2                3               1                4           1
## 3                4               4                4           5
## 4                3               1                4           2
## 5                4               3                3           3
## 6                4               4                4           1
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes            satisfaction
## 1                         25                       18 neutral or dissatisfied
## 2                          1                        6 neutral or dissatisfied
## 3                          0                        0               satisfied
## 4                         11                        9 neutral or dissatisfied
## 5                          0                        0               satisfied
## 6                          0                        0 neutral or dissatisfied
```

After viewing the above, the parameter to be predicted is the level of passenger satisfaction. All discrete values (or levels) are found below

```
labels = subset(df, select = "satisfaction")
unique(labels)
```

```
##              satisfaction
## 1 neutral or dissatisfied
## 3               satisfied
```

Unnecessary columns are removed, from the above, X and id do not seem to contribute meaningful information to the dataset, and are subsequently pruned, we further explore the structure of data in order to get a better idea of datatypes and valuees

```
df <- subset(df, select = -c(X, id))
str(df)
```

```
## 'data.frame':    103594 obs. of  23 variables:
##  $ Gender                          : chr  "Male" "Male" "Female" "Female" ...
##  $ Customer.Type                   : chr  "Loyal Customer" "disloyal Customer" "Loyal Customer" "Loy
##  $ Age                             : int  13 25 26 25 61 26 47 52 41 20 ...
##  $ Type.of.Travel                  : chr  "Personal Travel" "Business travel" "Business travel" "Bus
##  $ Class                           : chr  "Eco Plus" "Business" "Business" "Business" ...
##  $ Flight.Distance                 : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
##  $ Inflight.wifi.service           : int  3 3 2 2 3 3 2 4 1 3 ...
##  $ Departure.Arrival.time.convenient: int  4 2 2 5 3 4 4 3 2 3 ...
##  $ Ease.of.Online.booking          : int  3 3 2 5 3 2 2 4 2 3 ...
##  $ Gate.location                   : int  1 3 2 5 3 1 3 4 2 4 ...
##  $ Food.and.drink                  : int  5 1 5 2 4 1 2 5 4 2 ...
##  $ Online.boarding                 : int  3 3 5 2 5 2 2 5 3 3 ...
##  $ Seat.comfort                    : int  5 1 5 2 5 1 2 5 3 3 ...
##  $ Inflight.entertainment          : int  5 1 5 2 3 1 2 5 1 2 ...
##  $ On.board.service                : int  4 1 4 2 3 3 3 5 1 2 ...
```

```
##  $ Leg.room.service                : int  3 5 3 5 4 4 3 5 2 3 ...
##  $ Baggage.handling                : int  4 3 4 3 4 4 4 5 1 4 ...
##  $ Checkin.service                 : int  4 1 4 1 3 4 3 4 4 4 ...
##  $ Inflight.service                : int  5 4 4 4 3 4 5 5 1 3 ...
##  $ Cleanliness                     : int  5 1 5 2 3 1 2 4 2 2 ...
##  $ Departure.Delay.in.Minutes      : int  25 1 0 11 0 0 9 4 0 0 ...
##  $ Arrival.Delay.in.Minutes        : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                    : chr  "neutral or dissatisfied" "neutral or dissatisfied" "satis
```

Following numeric encoding, R expects variables to be of type factor for Logistic Regression to be performed, this is done next. From the output that follows, the dataframe was successfully converted into levels with no errors (NAs introduced by coercion) thrown

```
df_enc = df

df_enc$Gender = as.numeric(factor(df_enc$Gender, levels = c("Male", "Female"), labels = c(0,
    1)))
df_enc$Customer.Type = as.numeric(factor(df_enc$Customer.Type, levels = c("Loyal Customer",
    "disloyal Customer"), labels = c(1, 0)))
df_enc$Type.of.Travel = as.numeric(factor(df_enc$Type.of.Travel, levels = c("Personal Travel",
    "Business travel"), labels = c(1, 0)))
df_enc$Class = as.numeric(factor(df_enc$Class, levels = c("Eco Plus", "Business",
    "Eco"), labels = c(0, 1, 2)))



df_enc$Age = as.numeric(df_enc$Age)
df_enc$Type.of.Travel = as.numeric(df_enc$Type.of.Travel)
df_enc$Class = as.numeric(df_enc$Class)
df_enc$Flight.Distance = as.numeric(df_enc$Flight.Distance)
df_enc$Inflight.wifi.service = as.numeric(df_enc$Inflight.wifi.service)
df_enc$Departure.Arrival.time.convenient = as.numeric(df_enc$Departure.Arrival.time.convenient)
df_enc$Ease.of.Online.booking = as.numeric(df_enc$Ease.of.Online.booking)
df_enc$Gate.location = as.numeric(df_enc$Gate.location)
df_enc$Food.and.drink = as.numeric(df_enc$Food.and.drink)
df_enc$Online.boarding = as.numeric(df_enc$Online.boarding)
df_enc$Inflight.entertainment = as.numeric(df_enc$Inflight.entertainment)
df_enc$On.board.service = as.numeric(df_enc$On.board.service)
df_enc$Leg.room.service = as.numeric(df_enc$Leg.room.service)
df_enc$Baggage.handling = as.numeric(df_enc$Baggage.handling)
df_enc$Checkin.service = as.numeric(df_enc$Checkin.service)
df_enc$Inflight.service = as.numeric(df_enc$Inflight.service)
df_enc$Cleanliness = as.numeric(df_enc$Cleanliness)
df_enc$Departure.Delay.in.Minutes = as.numeric(df_enc$Departure.Delay.in.Minutes)
df_enc$Arrival.Delay.in.Minutes = as.numeric(df_enc$Arrival.Delay.in.Minutes)

df_enc$satisfaction <- ifelse(test = df_enc$satisfaction == "satisfied", yes = 1,
    no = 0)


str(df_enc)
```

```
## 'data.frame':    103594 obs. of  23 variables:
```

```
##  $ Gender                           : num  1 1 2 2 1 2 1 2 2 1 ...
##  $ Customer.Type                    : num  1 2 1 1 1 1 1 1 1 2 ...
##  $ Age                              : num  13 25 26 25 61 26 47 52 41 20 ...
##  $ Type.of.Travel                   : num  1 2 2 2 2 1 1 2 2 2 ...
##  $ Class                            : num  1 2 2 2 2 3 3 2 2 3 ...
##  $ Flight.Distance                  : num  460 235 1142 562 214 ...
##  $ Inflight.wifi.service            : num  3 3 2 2 3 3 2 4 1 3 ...
##  $ Departure.Arrival.time.convenient: num  4 2 2 5 3 4 4 3 2 3 ...
##  $ Ease.of.Online.booking           : num  3 3 2 5 3 2 2 4 2 3 ...
##  $ Gate.location                    : num  1 3 2 5 3 1 3 4 2 4 ...
##  $ Food.and.drink                   : num  5 1 5 2 4 1 2 5 4 2 ...
##  $ Online.boarding                  : num  3 3 5 2 5 2 2 5 3 3 ...
##  $ Seat.comfort                     : int  5 1 5 2 5 1 2 5 3 3 ...
##  $ Inflight.entertainment           : num  5 1 5 2 3 1 2 5 1 2 ...
##  $ On.board.service                 : num  4 1 4 2 3 3 3 5 1 2 ...
##  $ Leg.room.service                 : num  3 5 3 5 4 4 3 5 2 3 ...
##  $ Baggage.handling                 : num  4 3 4 3 4 4 4 5 1 4 ...
##  $ Checkin.service                  : num  4 1 4 1 3 4 3 4 4 4 ...
##  $ Inflight.service                 : num  5 4 4 4 3 4 5 5 1 3 ...
##  $ Cleanliness                      : num  5 1 5 2 3 1 2 4 2 2 ...
##  $ Departure.Delay.in.Minutes       : num  25 1 0 11 0 0 9 4 0 0 ...
##  $ Arrival.Delay.in.Minutes         : num  18 6 0 9 0 0 23 0 0 0 ...
##  $ satisfaction                     : num  0 0 1 0 1 0 0 1 0 0 ...
```

```
summary(df_enc)
```

```
##      Gender      Customer.Type        Age        Type.of.Travel     Class
##  Min.   :1.000   Min.   :1.000   Min.   : 7.00   Min.   :1.00   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:27.00   1st Qu.:1.00   1st Qu.:2.000
##  Median :2.000   Median :1.000   Median :40.00   Median :2.00   Median :2.000
##  Mean   :1.508   Mean   :1.183   Mean   :39.38   Mean   :1.69   Mean   :2.378
##  3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:51.00   3rd Qu.:2.00   3rd Qu.:3.000
##  Max.   :2.000   Max.   :2.000   Max.   :85.00   Max.   :2.00   Max.   :3.000
##  Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
##  Min.   :  31    Min.   :0.00          Min.   :0.00
##  1st Qu.: 414    1st Qu.:2.00          1st Qu.:2.00
##  Median : 842    Median :3.00          Median :3.00
##  Mean   :1189    Mean   :2.73          Mean   :3.06
##  3rd Qu.:1743    3rd Qu.:4.00          3rd Qu.:4.00
##  Max.   :4983    Max.   :5.00          Max.   :5.00
##  Ease.of.Online.booking Gate.location   Food.and.drink  Online.boarding
##  Min.   :0.000          Min.   :0.000   Min.   :0.000   Min.   :0.00
##  1st Qu.:2.000          1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00
##  Median :3.000          Median :3.000   Median :3.000   Median :3.00
##  Mean   :2.757          Mean   :2.977   Mean   :3.202   Mean   :3.25
##  3rd Qu.:4.000          3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00
##  Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.00
##   Seat.comfort  Inflight.entertainment On.board.service Leg.room.service
##  Min.   :0.00   Min.   :0.000          Min.   :0.000    Min.   :0.000
##  1st Qu.:2.00   1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000
##  Median :4.00   Median :4.000          Median :4.000    Median :4.000
##  Mean   :3.44   Mean   :3.358          Mean   :3.383    Mean   :3.351
##  3rd Qu.:5.00   3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000
##  Max.   :5.00   Max.   :5.000          Max.   :5.000    Max.   :5.000
```

```
##  Baggage.handling Checkin.service Inflight.service  Cleanliness
##  Min.   :1.000    Min.   :0.000   Min.   :0.000    Min.   :0.000
##  1st Qu.:3.000    1st Qu.:3.000   1st Qu.:3.000    1st Qu.:2.000
##  Median :4.000    Median :3.000   Median :4.000    Median :3.000
##  Mean   :3.632    Mean   :3.304   Mean   :3.641    Mean   :3.286
##  3rd Qu.:5.000    3rd Qu.:4.000   3rd Qu.:5.000    3rd Qu.:4.000
##  Max.   :5.000    Max.   :5.000   Max.   :5.000    Max.   :5.000
##  Departure.Delay.in.Minutes Arrival.Delay.in.Minutes  satisfaction
##  Min.   :   0.00            Min.   :   0.00          Min.   :0.0000
##  1st Qu.:   0.00            1st Qu.:   0.00          1st Qu.:0.0000
##  Median :   0.00            Median :   0.00          Median :0.0000
##  Mean   :  14.75            Mean   :  15.18          Mean   :0.4334
##  3rd Qu.:  12.00            3rd Qu.:  13.00          3rd Qu.:1.0000
##  Max.   :1592.00            Max.   :1584.00          Max.   :1.0000
```

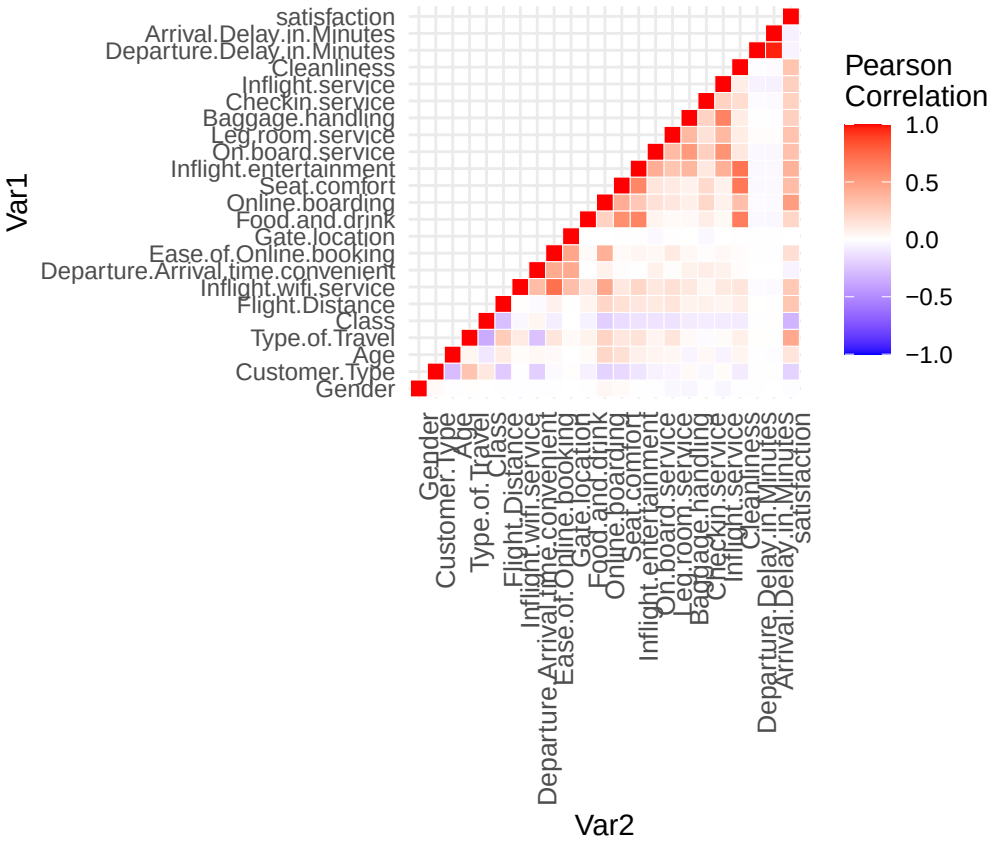
2.2 Correlation Analysis

Correlation analysis is performed to determine pairwise correlations within the dataset. Since we are concerned mainly with linear relations, the Pearson correlation coefficient is used in order to determine the extent of correlation amongst the attributes, this is visualised using a correlation heatmap below

```
library(reshape2)
library(ggplot2)

cor_matrix <- cor(df_enc)
cor_matrix[lower.tri(cor_matrix)] <- NA
cor_matrix_melted <- melt(cor_matrix, na.rm = TRUE)


ggplot(data = cor_matrix_melted, aes(Var2, Var1, fill = value)) + geom_tile(color = "white") +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,
        limit = c(-1, 1), space = "Lab", name = "Pearson\nCorrelation") + theme_minimal() +
    theme(axis.text.x = element_text(angle = 90, vjust = 1, size = 10, hjust = 1)) +
    coord_fixed()
```



2.3 Logistic GLM

Now the logit-model shall be built using the GLM syntax, with summary being called to display t-values, P-values estimated coefficients and the associated errors. Since this is essentially a binary-classification problem, the binomial distribution is used to predict the outcome.

Going through the coefficents that follow,the only factors which were not significantly correlated to satisfaction was flight distance. It seemed that all other factors exhibited some level of linear correlation with the response variable. However, a high level of significance does not alone make an attribute statistically 'interesting'. Following the levels of estimated correlation, the attributes with the highest coefficients are as follows (in descending magnitude of estimated coefficient):

2.3.1 Explanation of Coefficients

Attribute	Comments
Gender	Positively correlated, which may imply that male customers are overally more satisfied
Type of Travel	It seems that personal travel was more positively correlated with passenger satisfaction than business travel
Customer.Type	Negative correlation implies that loyal customers may be overally less satisfied with their flight
Class	Positive correlation may indicate that customers in upper classes may be less satisfied with the airline's service

Attribute	Comments
Inflight.wifi.service	the positive coefficient here may indicate that having inflight wifi may have a poistive effect on customer satisfcation
Online.boarding	The positive coefficient indicates that having the option for online boarding pre-flight may increase passenger satisfaction
Checkin.service	Similar to the above, the quality of checkin-service may positvely correlate to passenger satisfaction
Leg.room.service	the amount of leg-room available also was positively correlated to passenger satisfaction, this may translate to passengers being overally more satisfied with more legroom
Cleanliness	the level of (subjective) cleanliness was found to also positively impact a passenger's satisfaction with their flight experience
On.board.service	similar to leg-room service, the quality of service (presumably from flight-attendance during the flight) was found to positively influence passenger satisfaction
Ease.of.Online.booking	this was found to negatively correlate with passenger satisfaction
Baggage.handling	the quality of baggage handling was also found to positively impact passenger satisfaction levels
Departure.Arrival.time.convenient	strangely, the convenience of arrival time was found to negatively correlate with passenger satisfaction, this may proove an interesting area of research
Inflight.service	Finally, the quality of inflight service (media, etc) was found to positively impact a passenger's level of satisfaction on a flight

```
logmodel <- glm(satisfaction ~ ., family = binomial, data = df_enc)
summary(logmodel)
```

```
##
## Call:
## glm(formula = satisfaction ~ ., family = binomial, data = df_enc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8043  -0.5013  -0.1744   0.3919   3.9871
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.004e+01  1.108e-01 -90.538  < 2e-16 ***
## Gender                         -5.315e-02  1.936e-02  -2.745  0.00604 **
## Customer.Type                  -2.048e+00  2.947e-02 -69.500  < 2e-16 ***
## Age                            -7.601e-03  7.065e-04 -10.759  < 2e-16 ***
## Type.of.Travel                  3.041e+00  2.952e-02 103.010  < 2e-16 ***
## Class                          -1.941e-01  1.692e-02 -11.471  < 2e-16 ***
```

```
## Flight.Distance                     8.388e-05  1.069e-05   7.849 4.20e-15 ***
## Inflight.wifi.service               3.490e-01  1.135e-02  30.753  < 2e-16 ***
## Departure.Arrival.time.convenient -1.251e-01  8.163e-03 -15.329  < 2e-16 ***
## Ease.of.Online.booking            -1.291e-01  1.134e-02 -11.386  < 2e-16 ***
## Gate.location                       2.851e-02  9.120e-03   3.126  0.00177 **
## Food.and.drink                    -2.158e-02  1.064e-02  -2.028  0.04255 *
## Online.boarding                     6.418e-01  1.016e-02  63.180  < 2e-16 ***
## Seat.comfort                        8.537e-02  1.109e-02   7.701 1.35e-14 ***
## Inflight.entertainment              3.371e-02  1.417e-02   2.379  0.01737 *
## On.board.service                    3.247e-01  1.010e-02  32.144  < 2e-16 ***
## Leg.room.service                    2.594e-01  8.480e-03  30.587  < 2e-16 ***
## Baggage.handling                    1.546e-01  1.140e-02  13.562  < 2e-16 ***
## Checkin.service                     3.391e-01  8.510e-03  39.853  < 2e-16 ***
## Inflight.service                    1.438e-01  1.200e-02  11.986  < 2e-16 ***
## Cleanliness                         2.217e-01  1.208e-02  18.359  < 2e-16 ***
## Departure.Delay.in.Minutes          4.917e-03  9.802e-04   5.016 5.27e-07 ***
## Arrival.Delay.in.Minutes           -9.669e-03  9.662e-04 -10.008  < 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: 141768  on 103593  degrees of freedom
## Residual deviance:  69983  on 103571  degrees of freedom
## AIC: 70029
##
## Number of Fisher Scoring iterations: 6
```

2.3.2 Odds Ratios

```
exp(cbind(OR = coef(logmodel), confint(logmodel)))
```

```
## Waiting for profiling to be done...
```

```
##                                           OR        2.5 %       97.5 %
## (Intercept)                         4.380030e-05 3.522489e-05 5.439498e-05
## Gender                              9.482418e-01 9.129376e-01 9.849093e-01
## Customer.Type                       1.289900e-01 1.217380e-01 1.366447e-01
## Age                                 9.924276e-01 9.910537e-01 9.938021e-01
## Type.of.Travel                      2.093318e+01 1.975961e+01 2.218412e+01
## Class                               8.236043e-01 7.967737e-01 8.514060e-01
## Flight.Distance                     1.000084e+00 1.000063e+00 1.000105e+00
## Inflight.wifi.service               1.417688e+00 1.386537e+00 1.449617e+00
## Departure.Arrival.time.convenient 8.823796e-01 8.683761e-01 8.966122e-01
## Ease.of.Online.booking              8.789004e-01 8.595748e-01 8.986378e-01
## Gate.location                       1.028924e+00 1.010696e+00 1.047484e+00
## Food.and.drink                      9.786479e-01 9.584369e-01 9.992653e-01
## Online.boarding                     1.899919e+00 1.862522e+00 1.938186e+00
## Seat.comfort                        1.089124e+00 1.065715e+00 1.113046e+00
## Inflight.entertainment              1.034280e+00 1.005942e+00 1.063399e+00
## On.board.service                    1.383668e+00 1.356564e+00 1.411364e+00
## Leg.room.service                    1.296130e+00 1.274775e+00 1.317863e+00
```

```
## Baggage.handling                  1.167136e+00 1.141369e+00 1.193514e+00
## Checkin.service                   1.403727e+00 1.380535e+00 1.427363e+00
## Inflight.service                  1.154674e+00 1.127852e+00 1.182168e+00
## Cleanliness                       1.248243e+00 1.219055e+00 1.278159e+00
## Departure.Delay.in.Minutes        1.004929e+00 1.003003e+00 1.006864e+00
## Arrival.Delay.in.Minutes          9.903778e-01 9.885004e-01 9.922513e-01
```

2.3.3 Confidence Intervals

The confidence intervals for the parameters at level 0.95 are found using the confint-function as shown below

```
qnorm(1 - 0.05/2)
```

```
## [1] 1.959964
```

```
confint(logmodel, level = 0.95)
```

```
## Waiting for profiling to be done...
```

```
##                                         2.5 %        97.5 %
## (Intercept)                       -1.025376e+01 -9.8192386732
## Gender                            -9.108777e-02 -0.0152057683
## Customer.Type                     -2.105884e+00 -1.9903712089
## Age                               -8.986583e-03 -0.0062172297
## Type.of.Travel                     2.983640e+00  3.0993769312
## Class                             -2.271846e-01 -0.1608662053
## Flight.Distance                    6.293752e-05  0.0001048281
## Inflight.wifi.service              3.268094e-01  0.3712993688
## Departure.Arrival.time.convenient -1.411304e-01 -0.1091317961
## Ease.of.Online.booking            -1.513174e-01 -0.1068752139
## Gate.location                      1.063890e-02  0.0463908970
## Food.and.drink                    -4.245156e-02 -0.0007349746
## Online.boarding                    6.219316e-01  0.6617524289
## Seat.comfort                       6.364586e-02  0.1071004859
## Inflight.entertainment             5.924870e-03  0.0614700981
## On.board.service                   3.049551e-01  0.3445566583
## Leg.room.service                   2.427696e-01  0.2760117982
## Baggage.handling                   1.322284e-01  0.1769016572
## Checkin.service                    3.224714e-01  0.3558289800
## Inflight.service                   1.203145e-01  0.1673501398
## Cleanliness                        1.980762e-01  0.2454208329
## Departure.Delay.in.Minutes         2.998317e-03  0.0068407839
## Arrival.Delay.in.Minutes          -1.156619e-02 -0.0077788492
```

2.4 LRT Test

Since the LRT test approximately matches the Wald test when the sample size is relatively large, the Wald test for individual parameters is not carried out. As as aside, the main benefit of using the Wald test is not having to build a separate null model as in the LRT, hence this convenience is nullified given that LRT is carried out nonetheless.

2.4.1 LRT with Null Model

The likelihood ratio tests H_0: reduced model vs H_1: full model. Since the difference between log-likelihood statistics for two models (one of which is a special case of the other) follows an approximate χ^2 distribution, we can find the χ^2 test statistic for a full vs reduced (some parameters set to zero). The degrees-of-freedom are the number of parameters set to zero in the reduced model. The null hypotheses being tested, in essence, are that the subset of parameters set to zero are actually non-significant for the purposes of estimating the level of passenger satisfaction.

From the results of the LRT, it is shown that level of satisfaction is statistically (significantly) correlated to the attributes present in the full model, hence the null hypothesis (all attributes coefficients are zero) is rejected

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
nullmodel <- glm(formula = satisfaction ~ 1, family = binomial, data = df_enc)

lrtest(nullmodel, logmodel)
```

```
## Likelihood ratio test
##
## Model 1: satisfaction ~ 1
## Model 2: satisfaction ~ Gender + Customer.Type + Age + Type.of.Travel +
##     Class + Flight.Distance + Inflight.wifi.service + Departure.Arrival.time.convenient +
##     Ease.of.Online.booking + Gate.location + Food.and.drink +
##     Online.boarding + Seat.comfort + Inflight.entertainment +
##     On.board.service + Leg.room.service + Baggage.handling +
##     Checkin.service + Inflight.service + Cleanliness + Departure.Delay.in.Minutes +
##     Arrival.Delay.in.Minutes
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   1 -70884
## 2  23 -34992 22 71785  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.4.2 LRT Test with Reduced Model

From the following LRT using only the most significant parameters above. From the results below, we can see that the the p-value is zero with the χ^2 statistic shows that the βs for the attributes omitted contribute significantly to the fit of the model

```r
library(lmtest)

logmodel.reduced <- glm(satisfaction ~ Gender + Type.of.Travel + Customer.Type +
    Class + Inflight.wifi.service + Departure.Arrival.time.convenient + Ease.of.Online.booking +
    Online.boarding + On.board.service + Leg.room.service + Baggage.handling + Checkin.service +
    Inflight.service + Cleanliness, family = binomial, data = df_enc)
lrtest(logmodel.reduced, logmodel)
```

```
## Likelihood ratio test
##
## Model 1: satisfaction ~ Gender + Type.of.Travel + Customer.Type + Class +
##     Inflight.wifi.service + Departure.Arrival.time.convenient +
##     Ease.of.Online.booking + Online.boarding + On.board.service +
##     Leg.room.service + Baggage.handling + Checkin.service + Inflight.service +
##     Cleanliness
## Model 2: satisfaction ~ Gender + Customer.Type + Age + Type.of.Travel +
##     Class + Flight.Distance + Inflight.wifi.service + Departure.Arrival.time.convenient +
##     Ease.of.Online.booking + Gate.location + Food.and.drink +
##     Online.boarding + Seat.comfort + Inflight.entertainment +
##     On.board.service + Leg.room.service + Baggage.handling +
##     Checkin.service + Inflight.service + Cleanliness + Departure.Delay.in.Minutes +
##     Arrival.Delay.in.Minutes
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  15 -35334
## 2  23 -34992  8 685.46  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.5 Testing for Adequacy (R^2)

The standard *goodness-of-fit* statistic for OLS regression R^2 (also called the coefficient of determination). The higher this value, the better the **fit** of the model. R^2 is defined as:

$$R^2 = \frac{\text{Total Sum of Squares} - \text{Residual Sum of Squares}}{\text{Total Sum of Squares}}$$

However, R^2 is not appropriate for use with the logistic model [1], since it does not inform about the variability accounted for in the model, nor does it provide information to decide between models. Hence, a pseudo-R^2 variation is used. From the resulting value, it is seen that the model is borderline adequate

```r
ll.null <- logmodel$null.deviance/-2
ll.proposed <- logmodel$deviance/-2

r_squared <- (ll.null - ll.proposed)/ll.null

print(r_squared)
```

```
## [1] 0.5063536
```

Comparing fit to the reduced model above, the R^2 value was lower than that of the above, implying that the fit of the reduced model was worse than the original, hence the model model is used.

```
ll.null <- logmodel$null.deviance/-2
ll.proposed <- logmodel.reduced$deviance/-2

r_squared <- (ll.null - ll.proposed)/ll.null

print(r_squared)
```

```
## [1] 0.5015185
```

Saving for Future Use

```
save(logmodel, file = "model.RData")
```

2.6 Application/Evaluation

Although non-standard practice, for the purpose of prediction/application, a training tuple will be used to determine the likeliness of 'success' (passenger being satisfied)

```
model = load("model.RData")

newdata = df_enc[1, ]
predict(get(model), newdata)
```

```
##          1
## -0.9207725
```

Given the above prediction, we can re-check the original dataframe, given that the above prediction was less than zero, the logistic regression model would have correctly classified the tuple into the 'not satisfied/neutral' category

```
df_enc[1, ncol(df_enc)]
```

```
## [1] 0
```

In order to evaluate the model further, the accuracy was found to be:

$$Accuracy = \frac{53102 + 37382}{53102 + 37382 + 5595 + 7515} = \frac{90484}{103594} = 0.8734483$$

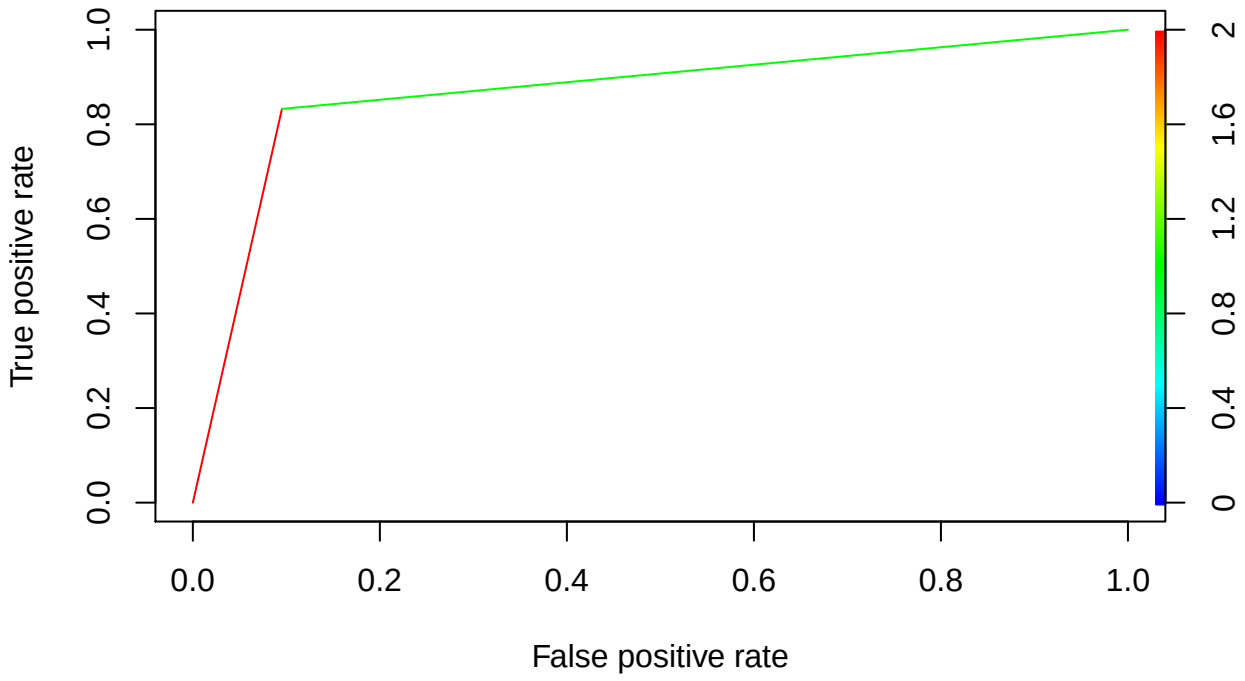
```
pred.prob = predict(logmodel, df_enc, type = "response")
pred.prob = ifelse(pred.prob > 0.5, 1, 0)
table(pred.prob, df_enc$satisfaction)
```

```
##
## pred.prob     0     1
##         0 53102  7515
##         1  5595 37382
```

2.7 ROC Curve

The ROC curve gives the ratio between True-Positive Rate and False-positive rate. The area under the curve is equivalent to the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance and is equivalent to the Wilcoxon Rank-Sum test statistic. For example, when the AUC is 0.8686484, there is roughly an 87% chance that the model will correctly discriminate between a positive and negative sample

```
library(ROCR)
pred <- prediction(pred.prob, df$satisfaction)
perf <- performance(pred, "tpr", "fpr")
plot(perf, colorize = TRUE)
```




```
perf <- performance(pred, "auc")
perf@y.values[[1]]
```

```
## [1] 0.8686484
```

3 Conclusion

It was found that reducing the original (full) model did not improve fit (according to R^2 statistic for Logistic Regression models). An interesting correlation was found that the convenience of flight departure time was found to negatively correlate with passenger satisfaction. This however, in retrospect, may be less surprising than it originally seems owing to the presence of all other predictor variables. Another unforeseen result was the relative lack of importance regarding flight distance to passenger satisfaction, in terms of both statistical significance (highest p-value) and estimated coefficient (lowest)

4 References

[1] Hilbe, J. (2017). Analysis of Model Fit. In Logistic regression models. Boca Raton: Routledge, Taylor & Francis Group
[2] https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
[3] Github for project https://github.com/aadi350/airlinepassengersatisfaction
[4] Shiny Demo App https://5b8hsq-aadidev-sooknanan.shinyapps.io/DemoPassenger/