Airline Customer Satisfaction

What variables affects the customer satisfaction levels?

GROUP 8

KYUNGJIN BAIK, SHRUTANVI DATAR, TRANG NGUYEN, VRINDA SEHGAL

Table of Contents

Introduction

Airline companies operate commercial air transportation services for passengers and cargo, both domestically and internationally. Airline companies are a vital part of the global transportation industry and play a significant role in the economic growth of countries by connecting people and businesses across different regions of the world. However, the airline industry is facing many challenges in recent years, like rising fuel costs, increasing competition, customer satisfaction etc.

This data set includes data for two low-cost airlines called Primera Air and WOW Air. Both these airlines are out of operation and filed for bankruptcy in 2018 and 2019 respectively. The problems faced by them were similar like high fuel costs, a highly competitive market, and the impossibility of tapping customers' loyalty. Below is more information about the two carriers.

Primera Air was a low-cost airline based in Denmark and Latvia. The airline was founded in 2003 however, in October 2018, Primera Air filed for bankruptcy.

Wow Air was a low-cost Icelandic airline that operated from 2012 to 2019. Wow Air was known for its cheap fares. Wow Air also ceased operations in March 2019 .

Goal

The goal of this project is to help the airlines in improving customer satisfaction and their overall performance.

Data Description

This dataset includes information about two airlines and the variables containing consumer data and how these variables reflect the customer's relationship with the airlines and its impact on customer satisfaction. There are a total of 22 variables with 10600 rows of customer reviews about the airline that they flew in. Among these flyers, 5430 customers i.e., over 51% are overall satisfied with the airline service while the rest are not satisfied with the service provided by the airline and want improvement with the overall experience.

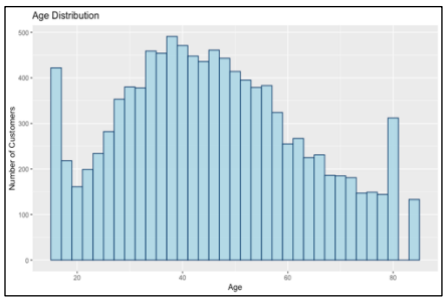
Below is the table representing all the variables from the data and their description.

Variable	Description
Customer ID	Unique ID for each customer
Satisfied (string)	A binary measure indicating if a customer is satisfied or not satisfied
Airline Status (string)	Customer airline status (Blue,Silver, Gold,Platinum)
Age (numeric)	Customers age
Gender Identity (String)	Customer's gender identity
Type of travel	Business, Mileage, or Personal
Shopping amount at airport	$ amount spent by customer while shopping
Eating and drinking amounts at airport	$ amount spent by customer eating and drinking
Class	Travel class (Business, Eco, Eco plus)
Flight date	Date of flight
Day of flight	Day of month for travel
Month of flight	Month
No of flights	Number of previous flights by customer
Airline name	Name of Airline
Flight time	Number of minutes from origin to destination
Flight canceled	Canceled or not
Arrival delay	Number of minutes flight is delayed for arrival
Departure delay	Number of minutes flight is delayed for departure
Origin State	Flight origin state
Destination State	Flight destination state
Scheduled Departure Hour	The hour flight is scheduled for departure (24 hour military time)
Flight Distance	Flight distance in miles

Exploratory Data Analysis

Graphical representation of the Data

EDA 1: Age Distribution of the flyers



From the above graph, we can interpret the age of the flyers. The most frequent flyers are in their mid-thirties (35-45 years old), followed by ones of 45 to 60 years old. It is interesting to know that the flyers who are 15-17 years old also fly a lot, could be the kids who are accompanied by their parents (mid-thirties). The flyers in the age group of 75-85 years are not much into flying, but the count of 80-year-old flyers is higher.

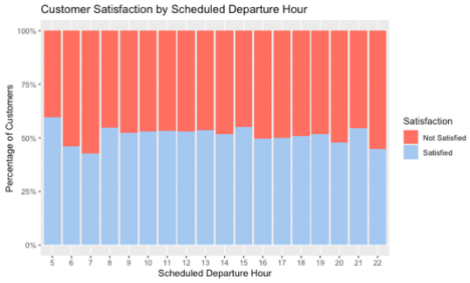
EDA 2: Customer Satisfaction based on the different classes.

The battery graph represents the class that a customer is flying in and their level of satisfaction.
The highest satisfaction percentage can be seen in the Business class. Among the people flying through Business Class, around 60% are satisfied with the airline. While that of Economy class is about 55% and the least number of satisfied flyers are from Economy Plus Class.
The possible explanation for the dissatisfaction among the Economy Plus class is they are not getting value for the extra money spent for this class.

EDA 3: Count of Customer satisfaction based on their ages

The graph represents the age of the flyers and their satisfaction. The most satisfied customers are between the ages of 25-55. The elder flyers (60+) tend to be least satisfied with the airlines.

EDA 4: Customer satisfaction levels by departure hour



The graph indicates the satisfaction level of the flyers in comparison to the scheduled departure time. Mostly the flyers who are traveling in the early morning flights (6-7 am) are not satisfied, followed by the ones having late night flights (8- 10 pm). Most of the flyers are satisfied through other times of the data.

EDA 5: Satisfaction levels of customers by the type of travel

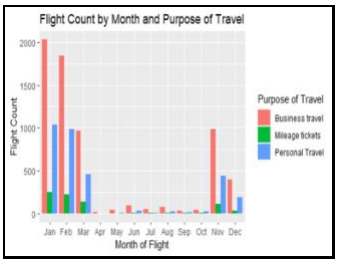


Fig 1

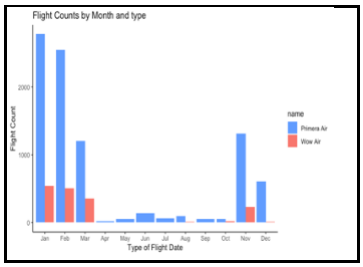


Fig 2

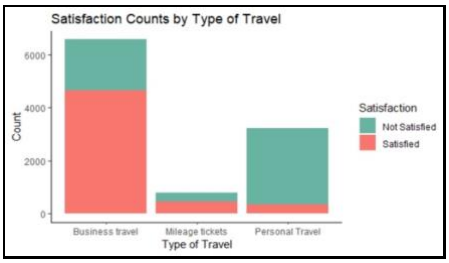


Fig 3

Fig4

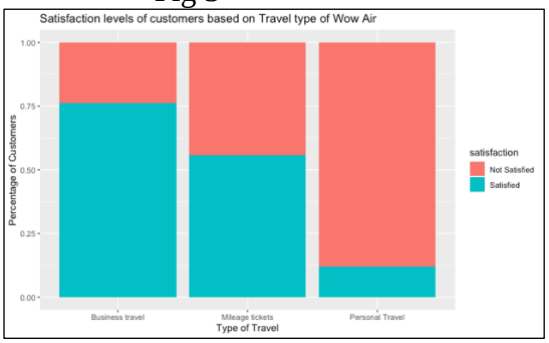


Fig5

The graphs above represent the relationship between the type of travel through Primera and Wow airlines. Fig2 represents the count of flyers by both the airlines. Fig 1 provides information about the reason for travel month on month. To analyze whether a particular airline was having an impact on the satisfaction level, we decided to plot separate graphs for both the airlines and Fig 4 and Fig5 represents the overall satisfaction rate of the flyers traveling for different purposes by Primera and Wow airlines respectively.
From Fig3 we can interpret that the most unsatisfied flyers are the ones flying for personal reasons followed by business travelers and lastly the mileage ticket owners (Fig3).
Maximum traveling takes place in the 1st quarter of the year (Jan-Mar) and not during the holidays (Nov-Dec) like we expected. We can see that there is hardly any traveling between the months of April to October. We can see some spike during the Holiday season however it is less in comparison to what we expected.

Data Understanding and Business Problem

The main objective of this project is to help the airlines to improve the customer satisfaction and overall business performance.
From the given data and through its analysis, we understood that most of the people prefer to travel by Primera Air over Wow Air. During EDA it appeared that there were relationships among satisfaction and some variables, such as females, teenagers, personal travel etc. We further used the regression model to identify which all variables influenced customer satisfaction. From that, we made the recommendations for the purpose of improving customer satisfaction and the overall business performance.

Technique of Data Modelling

Logistic regression

To find which variables have any impact on customer satisfaction, we decided to use Logistic Regression as our dependent variable is a binary variable. The regression had the following features:

- The dependent variable is satisfied and is denoted as 1 for satisfied and 0 for not satisfied.
- The predictor variables include all or some columns among airline_status, age, gender, type_of_travel, shopping_amount_at_airport, eating_and_drinking_amounts_at_airport, class, day_of_flight_date, month_of_flight_date, no_of_flights, airline_name, scheduled_departure_hour, flight_distance, flight_cancelled. We drop the columns origin_state, destination_state, and flight_date since the distance is highly correlated with origin state and destination state combined.

The data had some missing values in columns flight_time, arrival_delay, and departure_delay (below 4% of the total records), which mostly corresponded to canceled flights. This led us to check whether cancellation was a reason for dissatisfaction. The result turned out that it had no significant relationship, as shown in Graph 5(Appendix 2). It turned out that the customers were still satisfied if flights were canceled. Our focus was on the actual flight data to make suggestions to improve customer satisfaction. This approach helped to deal with most of the missing values as well.

Regarding the 25 missing values left of flight_time and arrival_delay, they are in the same records, corresponding to 0.24% of the total records of *df*, so we dropped these rows.

In the next step we checked the correlation between arrival_delay and departure_delay, flight_time and flight_distance. We saw that both correlation coefficients are about 0.96 and p_value are less than 0.05. These results are shown in Table 1 and Table 2. So, they are strongly correlated, thus we kept one of each variable pair to avoid multicollinearity.

After preprocessing data, we split data into training and test data and trained the model for training data using the **Forward Stepwise** Regression, beginning with the null model and then picking up the most important variable, which has the lowest AIC one by one. As a result, 12 variables were added to the model in the sequence of type_of_travel, airline_status, arrival_delay, gender, no_of_flights, scheduled_departure_hour, airline_name, class, flight_time, shopping_amount_at_airport, age, and flight_distance. However, between flight_time and flight_distance, flight_time had more impact (see Table 3).

We reran the model after removing statistically insignificant variables and checked which model explained the dependent variable better using ANOVA. Table 4 did not display a significant chi-square value with the p_value of 0.2908. It means that the latter fitted as well as the previous.

Final model:

satisfied ~ type_of_travel + airline_status + arrival_delay + gender + no_of_flights + scheduled_departure_hour + airline_name + class + flight_time + shopping_amount_at_airport + age.
The three terms of scheduled_departure_hour and all terms of other variables had p_value less than 0.05 (see Table 5), so all variables are statistically significant.

In reference to the regression coefficients from the model, we see that the odds of being satisfied increase with airline_status, genderMale, airline_nameWow Air, and shopping_amount_at_airport while decreasing with type_of_travel, arrival_delay, no_of_flights, scheduled_departure_hour, class, flight_time, and age.
We can more easily interpret the effects of variables from exponentiated coefficients (see Table 6).
For example:
The odds of being satisfied are 416% higher for a customer with silver airline status compared to one with blue status.
A minute increase in arrival_delay leads to 0.64% decrease in the odds of being satisfied.
Now we tested for an overall effect of categorical variables, including type_of_travel, airline_status, class, and scheduled_departure_hour using the Wald test. The chi-squared test statistic of the four variables above is associated with a p-value less than 5% indicating that the overall effects of these variables are statistically significant (shown in Table 8).
Lastly, we employed this model to calculate the probabilities of satisfaction on test data and labeled them as satisfied (the value of 1) using the threshold of 0.5.

Model Assessment

We evaluated the model using different metrics, in which.

- **The accuracy of the model reached 76.86%.**
- **AUC (the area under the curve of ROC) equaled 0.8264428.**
- Confusion matrix (see detail in Table 9)
- **Specificity equaled 0.6571.**

We can see Logistic Regression is a reasonable approach to help answer the question of which and how variables affect customer satisfaction. The measurements above also indicate that our model can make predictions with fairly high accuracy and offer companies a tool to recognize the potentially dissatisfied customers. As a result, we can use these results to form the basis for diverse strategies for each specific segment.

Conclusions and Recommendations

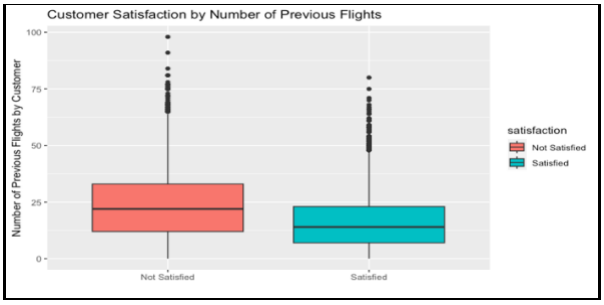
1. As we saw in the EDA, the economy plus flyers is the most unsatisfied out of all the classes. So, to make the economy plus flyers feel more satisfied, the airline could offer a better choice of foods and entertainment in comparison to the regular economy class. Also, providing free WIFI facilities to the business and economy plus class can be a game changer. (ref. EDA 2).
2. To increase the overall satisfaction for all the classes, the company should offer on time arrival and departure of flight.
   Secondly, offering distinctive feature differentiation between the three classes will provide the value of money to all customers. (ref. EDA 2).
3. To increase the overall performance, the company should offer differential pricing to the various age groups, leading to more sales and satisfaction (price reduction for infants and senior citizens). (ref. EDA 3).
4. To improve the satisfaction level for the older customers, we can provide them with better facilities, like free wheelchair assistance. Early check-in and boarding for these customers. Different and faster queues for check-ins and lounge facilities. (ref. EDA 3).
5. Along with older customers, younger customers can be provided with better amenities like access to the lounge with a separate kids' section available. Also, we can provide them with priority boarding and checking in like for older customers. This will improve their overall experience flying with the airline and thus, improve the satisfaction level. (ref. EDA 3).
6. Since there are more female flyers than male, we can increase their satisfaction level by improving the experience for this group which can be done by added benefits like providing free hygiene kits when necessary. (ref. Graph 2, Table 5).
7. The airlines should roll out more loyalty programs based on classes (Blue, Platinum, Gold, and silver). All classes had a positive relation to the regressor(satisfaction) and thus will help in the overall satisfaction.

References

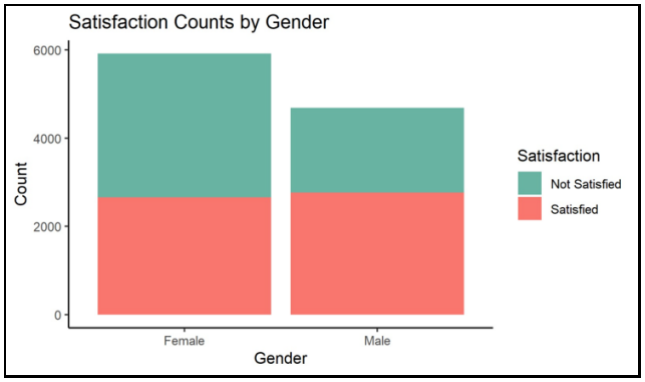
- Sujata Ramnarayan (2020). R for Fundamental Data Analysis in Market Research
- Joseph F. Hair, Dana E. Harrison, Haya Ajjan. Essentials of Marketing Analytics(Edition 1). McGraw Hill LLC.

Appendix A: Graphs

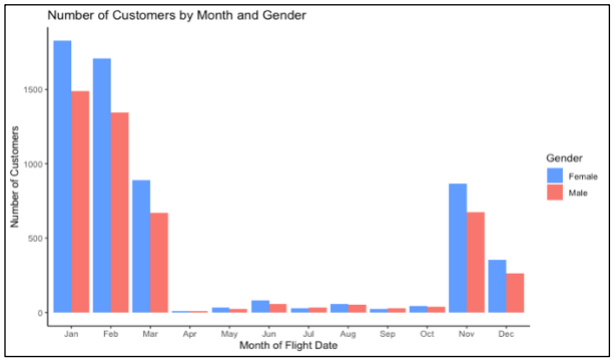
Graph 1: Customer Satisfaction by Number of Previous Flights



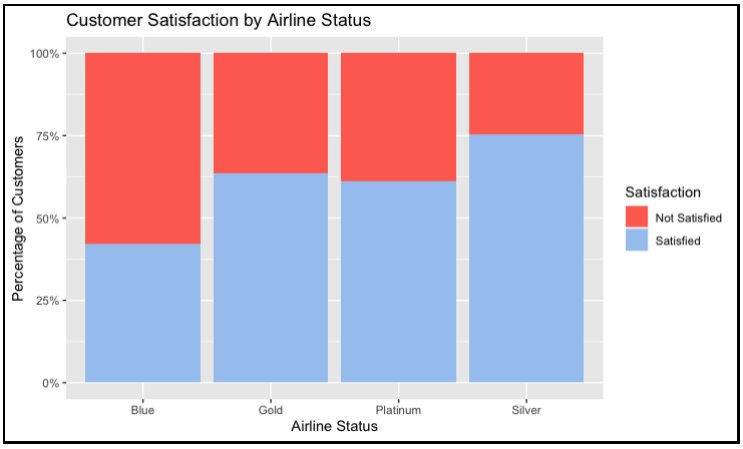
Graph 2: Customer Satisfaction by Gender



Graph 3: Flight count based on Month and Gender



Graph 4: Customer Satisfaction by customer status with the Airlines



Graph 5: Customer satisfaction based on the flight status

Appendix B: Model results

Table 1: Pearson's correlations between arrival_delay and departure_delay

```
        Pearson's product-moment correlation

data:  df$arrival_delay and df$departure_delay
t = 363.49, df = 10180, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9621513 0.9649314
sample estimates:
      cor
0.9635674
```

Table 2: Pearson's correlations between flight_time and flight_distance

```
        Pearson's product-moment correlation

data:  df$flight_time and df$flight_distance
t = 343.15, df = 10180, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9578137 0.9609061
sample estimates:
      cor
0.9593887
```

Table 3: Forward Stepwise Regression

```
Step:  AIC=7814.37
satisfied ~ type_of_travel + airline_status + arrival_delay +
    gender + no_of_flights + scheduled_departure_hour + airline_name +
    class + flight_time + shopping_amount_at_airport + age +
    flight_distance

                                          Df Deviance    AIC
<none>                                         7748.4 7814.4
+ eating_and_drinking_amounts_at_airport  1   7748.3 7816.3
+ departure_delay                         1   7748.4 7816.4
+ month_of_flight_date                   11   7731.4 7819.4
+ day_of_flight_date                     30   7717.4 7843.4
```

Table 4: Analysis of Deviance Table

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	8166	7751.1			
2	8122	7702.4	44	48.662	0.2908

Table 5: Model result

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4976  -0.5397   0.3515   0.8223   3.4299

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    1.5012646  0.3388306   4.431 9.39e-06 ***
type_of_travelMileage tickets -0.5315528  0.0944386  -5.629 1.82e-08 ***
type_of_travelPersonal Travel -3.0118393  0.0824666 -36.522  < 2e-16 ***
airline_statusGold             0.8753955  0.1018152   8.598  < 2e-16 ***
airline_statusPlatinum         0.4340337  0.1565099   2.773 0.005551 **
airline_statusSilver           1.6417397  0.0818402  20.060  < 2e-16 ***
arrival_delay                 -0.0063848  0.0007180  -8.892  < 2e-16 ***
genderMale                     0.5300727  0.0582061   9.107  < 2e-16 ***
no_of_flights                 -0.0156959  0.0021394  -7.337 2.19e-13 ***
scheduled_departure_hour6     -0.6587366  0.3227763  -2.041 0.041266 *
scheduled_departure_hour7     -0.6739976  0.3258016  -2.069 0.038571 *
scheduled_departure_hour8     -0.1049308  0.3264570  -0.321 0.747890
scheduled_departure_hour9     -0.1026458  0.3289744  -0.312 0.755027
scheduled_departure_hour10    -0.1527041  0.3270915  -0.467 0.640604
scheduled_departure_hour11    -0.1305005  0.3276403  -0.398 0.690406
scheduled_departure_hour12    -0.2553524  0.3230461  -0.790 0.429264
scheduled_departure_hour13    -0.0432117  0.3290067  -0.131 0.895506
scheduled_departure_hour14    -0.2325217  0.3238415  -0.718 0.472751
scheduled_departure_hour15    -0.0206641  0.3290867  -0.063 0.949932
scheduled_departure_hour16    -0.3241828  0.3239011  -1.001 0.316890
scheduled_departure_hour17    -0.1785646  0.3223170  -0.554 0.579577
scheduled_departure_hour18    -0.1758175  0.3296388  -0.533 0.593781
scheduled_departure_hour19    -0.0922298  0.3315398  -0.278 0.780870
scheduled_departure_hour20    -0.4051304  0.3392735  -1.194 0.232434
scheduled_departure_hour21    -0.1856255  0.3471440  -0.535 0.592842
scheduled_departure_hour22    -0.9246409  0.5378076  -1.719 0.085564 .
airline_nameWow Air            0.3042076  0.0805704   3.776 0.000160 ***
classEco                      -0.3041816  0.0995585  -3.055 0.002248 **
classEco Plus                 -0.4732445  0.1294630  -3.655 0.000257 ***
flight_time                   -0.0025169  0.0008054  -3.125 0.001777 **
shopping_amount_at_airport     0.0013053  0.0005397   2.419 0.015576 *
age                           -0.0045217  0.0018790  -2.406 0.016110 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 6: Exponentiating coefficients

(Intercept)	type_of_travelMileage tickets	type_of_travelPersonal Travel
4.4873602	0.5876917	0.0492011
airline_statusGold	airline_statusPlatinum	airline_statusSilver
2.3998243	1.5434709	5.1641460
arrival_delay	genderMale	no_of_flights
0.9936356	1.6990559	0.9844266
scheduled_departure_hour6	scheduled_departure_hour7	scheduled_departure_hour8
0.5175047	0.5096671	0.9003868
scheduled_departure_hour9	scheduled_departure_hour10	scheduled_departure_hour11
0.9024466	0.8583836	0.8776561
scheduled_departure_hour12	scheduled_departure_hour13	scheduled_departure_hour14
0.7746435	0.9577086	0.7925326
scheduled_departure_hour15	scheduled_departure_hour16	scheduled_departure_hour17
0.9795480	0.7231180	0.8364700
scheduled_departure_hour18	scheduled_departure_hour19	scheduled_departure_hour20
0.8387710	0.9118956	0.6668898
scheduled_departure_hour21	scheduled_departure_hour22	airline_nameWow Air
0.8305846	0.3966738	1.3555504
classEco	classEco Plus	flight_time
0.7377269	0.6229777	0.9974863
shopping_amount_at_airport	age	
1.0013061	0.9954886	

Table 7: Coefficients and confidence intervals for the coefficient estimates

	coef	2.5 %	97.5 %
(Intercept)	4.4873602	2.30981817	8.7177434
type_of_travelMileage tickets	0.5876917	0.48838608	0.7071895
type_of_travelPersonal Travel	0.0492011	0.04185806	0.0578323
airline_statusGold	2.3998243	1.96568622	2.9298453
airline_statusPlatinum	1.5434709	1.13573501	2.0975865
airline_statusSilver	5.1641460	4.39881821	6.0626291
arrival_delay	0.9936356	0.99223828	0.9950349
genderMale	1.6990559	1.51587196	1.9043764
no_of_flights	0.9844266	0.98030751	0.9885630
scheduled_departure_hour6	0.5175047	0.27489487	0.9742311
scheduled_departure_hour7	0.5096671	0.26913103	0.9651824
scheduled_departure_hour8	0.9003868	0.47484127	1.7072998
scheduled_departure_hour9	0.9024466	0.47358507	1.7196695
scheduled_departure_hour10	0.8583836	0.45212717	1.6296797
scheduled_departure_hour11	0.8776561	0.46178136	1.6680625
scheduled_departure_hour12	0.7746435	0.41126761	1.4590805
scheduled_departure_hour13	0.9577086	0.50255362	1.8250904
scheduled_departure_hour14	0.7925326	0.42010975	1.4951042
scheduled_departure_hour15	0.9795480	0.51393307	1.8670023
scheduled_departure_hour16	0.7231180	0.38326937	1.3643138
scheduled_departure_hour17	0.8364700	0.44472714	1.5732840
scheduled_departure_hour18	0.8387710	0.43959668	1.6004143
scheduled_departure_hour19	0.9118956	0.47614354	1.7464347
scheduled_departure_hour20	0.6668898	0.34297618	1.2967141
scheduled_departure_hour21	0.8305846	0.42062431	1.6401114
scheduled_departure_hour22	0.3966738	0.13824545	1.1381938
airline_nameWow Air	1.3555504	1.15753479	1.5874398
classEco	0.7377269	0.60694773	0.8966851
classEco Plus	0.6229777	0.48336301	0.8029188
flight_time	0.9974863	0.99591298	0.9990621
shopping_amount_at_airport	1.0013061	1.00024759	1.0023658
age	0.9954886	0.99182912	0.9991615

Table 8: Wald test for the overall effects of each variable

```
> # Test for an overall effect of type_of_travel
> wald.test(b = coef(reduced_model), Sigma = vcov(reduced_model), Terms = 2:3)
Wald test:
----------

Chi-squared test:
X2 = 1334.3, df = 2, P(> X2) = 0.0
> # Test for an overall effect of airline_status
> wald.test(b = coef(reduced_model), Sigma = vcov(reduced_model), Terms = 4:6)
Wald test:
----------

Chi-squared test:
X2 = 433.8, df = 3, P(> X2) = 0.0
> # Test for an overall effect of scheduled_departure_hour
> wald.test(b = coef(reduced_model), Sigma = vcov(reduced_model), Terms = 10:26)
Wald test:
----------

Chi-squared test:
X2 = 51.2, df = 17, P(> X2) = 2.7e-05
> # Test for an overall effect of class
> wald.test(b = coef(reduced_model), Sigma = vcov(reduced_model), Terms = 28:29)
Wald test:
----------

Chi-squared test:
X2 = 13.9, df = 2, P(> X2) = 0.00097
```

Table 9: Confusion matrix and statistics

```
          Reference
Prediction   0   1
         0 621 135
         1 324 904

               Accuracy : 0.7686
                 95% CI : (0.7495, 0.787)
    No Information Rate : 0.5237
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.532

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8701
            Specificity : 0.6571
         Pos Pred Value : 0.7362
         Neg Pred Value : 0.8214
             Prevalence : 0.5237
         Detection Rate : 0.4556
   Detection Prevalence : 0.6190
      Balanced Accuracy : 0.7636

       'Positive' Class : 1
```