

Airline Customer Satisfaction Analysis

**Submitted towards partial fulfilment of the criteria
for award of PGPDSE by Great Lakes Institute of Management**

Submitted By

Group No. 3 [Batch: 2023]

Group Members

- 1) Ardra P
- 2) Nivedhitha R
- 3) Raman R
- 4) Sneha Sanjeev Kumar
- 5) Yasho Paramesh H S

Research Supervisor

P V Subramanian

Name and signature of the team leader

nivedhitha r

Signature of the mentor






P.V. Subramanian

ACKNOWLEDGEMENT

We hereby certify that the work done by us for conceptualizing and completing this project is original and authentic.

We affirm that:

- 1. The ideas and concepts presented in this project are our own, and any external sources have been appropriately credited.*
- 2. All data, results, and findings are genuine, and any external contributions or assistance have been duly acknowledged.*
- 3. The project has not been submitted for assessment in any other context, and all aspects of the work comply with ethical standards and academic integrity.*
- 4. We take full responsibility for the content, accuracy, and authenticity of this project.*

NAME	SIGNATURE
ARDRA P	
NIVEDHITHA R	
RAMAN R	
SNEHA SANJEEV KUMAR	
YASHO PARAMESH H S	

Date: 18th January 2024

Place: Bengaluru

TABLE OF CONTENTS

Chapter NO	TOPIC	PAGE NO
1	Overview	8
2	Step-By-Step Walk through of the solution	11
3	Suggestions, Recommendations, to the Stakeholders	50
4	Limitations	52
5	Closing Reflections	53
6	Bibliography	54
7	Annexure	55
8	Data Dictionary	56

CHAPTER 1 - LIST OF FIGURES AND TABLES

1.1) List of Figures

Sl.no	Description	Page number
1	Project Flow Diagram	8
2.1	Handle missing data.	14
2.2	Barplot for Type of travel vs satisfaction level	15
2.3	Barplot for the variable classes' vs satisfaction level	15
2.4	Barplot for inflight WiFi service vs satisfaction	16
2.5	Barplot for Gender vs Satisfaction	16
2.6	Barplot for Departure/Arrival time convenient vs satisfaction	17
2.7	Barplot for Ease of Online booking vs satisfaction	17
2.8	Barplot for Gate location vs satisfaction	18
2.9	Barplot for Food and drink vs satisfaction	18
2.10	Barplot for Seat comfort vs satisfaction	19
2.11	Barplot for Inflight entertainment vs satisfaction	19
2.12	Barplot for Online boarding vs satisfaction	20
2.13	Barplot for On-board service vs satisfaction	20
2.14	Barplot for Inflight service vs satisfaction	21
2.15	Barplot for Leg room service vs satisfaction	21
2.16	Barplot for Check-in service vs satisfaction	22
2.17	Barplot for Baggage handling vs satisfaction	22
2.18	Box plot for age vs satisfaction	23
2.19	Box plot for Flight Distance vs Satisfaction	23
2.20	Box plot for departure delay in minutes vs satisfaction	24
2.21	Box plot for Arrival Delay in minutes vs Satisfaction	24
2.22	Heatmap for checking multicollinearity	25
2.23	Histogram for the variable Age	26

2.24	Histogram for the variable Flight Distance	26
2.25	Histogram for the variable Departure Delay in Minutes	27
2.26	Histogram for the variables Arrival Delay in Minutes	27
2.27	Boxplot for the variable Age	28
2.28	Boxplot for the variable Flight Distance	28
2.29	Boxplot for the variable Departure delay in minutes	29
2.30	Boxplot for the variable Arrival delay in minutes	29
2.31	Results of Chi-square tests of independence	31
2.32	Function used to calculate Skewness	32
2.33	Class distribution of target variable: “satisfaction”	33
2.34	Building base model: Step 1 to Step 3	34
2.35	Building base model: Model Summary	35
2.36	Building base model: List the significant variables at 5% level of significance	36
2.37	Building base model: Getting Odds ratio of independent features	36
2.38	Building base model: Interpretation of Odds Ratio for variables having odds ratio > 1	36
2.39	Function used to give a detailed output of the base model	37
2.40	Confusion Matrix, Accuracy of Train data	38
2.41	Confusion Matrix, Accuracy of Test data	38
2.42	Classification report of training and test data	39
2.43	Variable importance plot from XGBoost Classifier before applying SMOTE	47
2.44	Confusion matrix and ROC curve for XGBoost Classifier before applying SMOTE.	47
2.45	Variable importance plot from XGBoost Classifier after applying SMOTE	48
2.46	Confusion matrix and ROC curve for XGBoost Classifier after applying SMOTE	48
2.47	EDA done at the onset comparing most important features with target	49

2.48	Chi square test results done at the onset comparing most important features with target	50
------	---	----

1.2) List of Tables:

SL.NO	DESCRIPTION	PAGE NUMBER
1	Comparison of performance of benchmark mark model with best model	7
2.1	Results from base model	39
2.2	Model Evaluation based on default parameters	40
2.3	Model evaluation after hyperparameter tuning.	42
2.4	Performance of the best model	46
2.5	Best parameters of the best model	46
2.6	Comparison of best model with benchmark model	46

ABBREVIATIONS:

RFE: Recursive Feature Elimination

SMOTE: Synthetic Minority Oversampling Technique

XGBoost: Extreme Gradient Boosting

AdaBoost: Adaptive Boosting

EDA: Exploratory Data Analysis

SciPy: Scientific Python

NumPy: Numerical Python

Scikit: SciPy Toolkit

EXECUTIVE SUMMARY:

A customer satisfaction dataset serves as a comprehensive repository of information capturing the nuances and intricacies of customer experiences with our products and services. We obtained the data from Kaggle datasets and the data could be an anonymized real airline data not revealing the name of the airline.

This dataset is a strategic asset that goes beyond numerical ratings, delving into the qualitative aspects of customer sentiments, preferences, and interactions. The dataset has the dimension (129880 rows, 25 columns). We split the data into Training dataset and test dataset in the ratio 80% and 20%. By meticulously analysing this data, we aim to gain valuable insights into the factors influencing satisfaction, identify areas for improvement, and enhance the overall customer journey.

The target variable, Satisfaction has two levels 0 representing “Neutral or dissatisfied” and 1 representing “Satisfied” passengers. The count and percentage of these two classes are: (73452, 56.55%) and (56428, 43.45%) indicating a slight data imbalance. We have applied Synthetic Minority Over-sampling Technique (SMOTE) to treat the data imbalance.

Null values are present in the dataset. The variable, ‘Arrival Delay in Minutes’ has 310 missing values constituting 2.98% of the total observations. Multiple Imputation by Chained Equation technique was adopted to treat missing values.

We have used classification techniques to predict the outcome of the target variable which is a categorical variable. We have used Logistic Regression as our base model and the following too:

- K-Nearest Neighbour (KNN) classifier
- 2. Gaussian Naive Bayes classifier
- 3. Decision Tree classifier
- 4. Random Forest classifier
- 5. ADABOOST classifier
- 6. Gradient Boost classifier
- 7. XGBoost classifier

Since the dataset is not exactly balanced, we shall consider weighted Average-F1 score as the measures to evaluate the performance of the models and to select the best model.

Our best model is XGBoost Classifier with weighted Average F1 being 96.7% for training dataset and 96.1% for test dataset.

Comparing the best model with base model using the performance measure: **Weighted Average F1 Score**

Model	Measure for training dataset	Measure for test dataset
LOGISTIC REGRESSION	0.79	0.79
XGBOOST CLASSIFIER	0.967	0.961

Table 1 Comparison of performance of benchmark mark model with best model

Most important variables impacting the target variable, Satisfaction:

- 1) Online booking
- 2) Type of travel
- 3) Inflight WiFi service
- 4) Type of customer

CHAPTER 1 - OVERVIEW

This dataset provides comprehensive information about passengers satisfaction indicators of an airline. Customer feedback is an important part in knowing whether the passenger is satisfied or not. It encompasses various attributes related to passenger demographics, travel details and satisfaction rating. The project trains a model to take in various features or customer feedback and predict whether the customer is satisfied with the flight or not

Methodology we followed

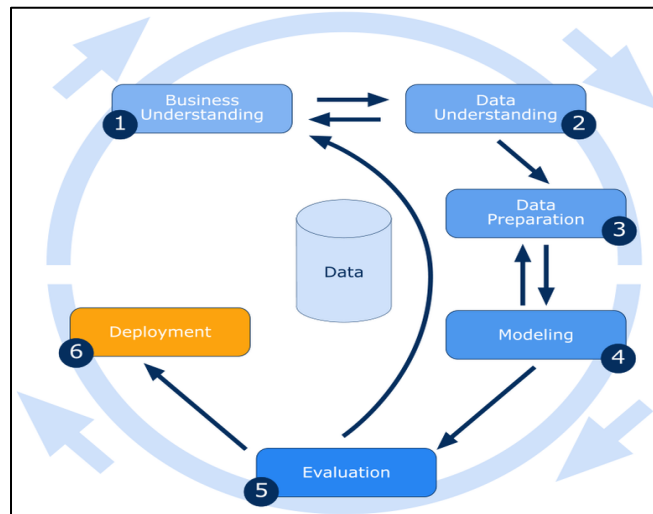


Fig 1 Project Flow Diagram

Business Understanding

Passenger satisfaction is of paramount importance in the airline business. Passenger satisfaction ensures Customer Loyalty, Positive Word of Mouth, Increased Revenue, Reduced Complaints and Disputes, Sustainable Growth etc. It is a critical component of a successful airline business. Airlines that prioritize and consistently deliver high levels of passenger satisfaction can enjoy a wide range of benefits, from increased revenue and customer loyalty to a strong brand image and competitive advantage. There is an urgent need to understand the factor leading to dissatisfaction of the passenger

Data Understanding

We shall dive deep into the minute details of the dataset. Starting with the classification of numerical and categorical columns. We shall clearly indicate the target variable.

- a) Data Description: describe the data in simple words to get a clear understanding
- b) Data Cleaning: Treat outlier, missing values, transform data etc.
- d) Data Distribution: Visualize how the data is spread

For this project, we have two datasets, namely:

- a) training dataset with the dimension, (103904 rows, 25 columns)
- b) test dataset with the dimension, (25976 rows, 25 columns)

The combined dataset has the dimension (129880 rows, 25 columns)

Training dataset and test dataset constitute 80% and 20% of the combined data.

Null Values

- Variable with missing values: **‘Arrival Delay in Minutes’**
- Count: 310
- Percentage: 2.98%
- Number of Numerical Columns: 6 ['Unnamed: 0', 'id', 'Age', 'Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes']
- Number of Categorical Columns: 19 ['Gender', 'Customer Type', 'Type of Travel', 'Class', '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', 'satisfaction']
- Number of Ordinal variables: 14 ['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']
- Our target variable is satisfaction and it has the levels with count and % as follows:
 - a) neutral or dissatisfied count: 73452 % of total: 56.55%
 - b) satisfied count: 56428 % of total: 43.45%
- Since classes in the target variables are not distributed equally, we conclude that there is some slight imbalance in the dataset (approximately, 57% and 43%). We shall consider

Recall, Precision, F1 score, Average-F1 score as the measures to evaluate the performance of the models and to select the best model.

- Using the combined dataset, we shall perform EDA to understand the data better to build useful models.
- Using the training dataset already obtained, we shall build the suitable models.
- Using both the training dataset & test dataset already obtained, we shall test the model performance using the above measures.

We shall treat the missing values using suitable techniques such as replacing the nulls with the median value.

Data Preparation

In this step we will check how the relationship between target variable and independent variables exists. We will perform Exploratory Data Analysis (EDA) to uncover the patterns related to the target variable, creating hypotheses and testing them and to discover the hidden pattern of features impacting the target variable, '**Satisfaction**'.

We noticed that 14 variables indicate the satisfaction level and we need to convert its data type from integer to categorical variable.

We will also encode 6 variables (categorical independent variables and the target variable) in order to use the machine learning algorithms.

We then drop columns that are not necessary such as 'Unnamed 0' and 'ID'. We are eliminating these columns from the dataset since it is descriptive in nature and doesn't possess predictive capabilities.

In our data set, we have observed that the target variable is not exactly equally distributed. We have (57% and 43%) as the proportion of binary classes in the target variable.

In order to balance the unbalanced dataset, we can use oversampling techniques such as SMOTE or random oversampling. We may use measures such as AUROC, Precision, Recall, F1 score and Average F1 score to measure the performance of the model and choose the best.

Modelling

First, we are going to split the given data into training and test data in the ratio 80:20 using stratified random sampling method. Using the training data, we build the model to predict the target variable, '**Satisfaction**', we shall build the suitable model and test the performance of the model using both training and test data. Before building the model, we shall check all the assumptions of the model and make sure that they are satisfied.

We will be working with various techniques such as Logistic regression, KNN, Naïve Bayes, Decision Tree, Random Forest, AdaBoost, XGBoost to build the model.

Evaluation

As our data is found to be not balanced, we shall use performance measures such as **Weighted average F1 score** as the measure of model performance. Based on the variable importance plot or odds ratio obtained from the best model we shall identify the top four important independent variables affecting the target variable '**satisfaction**'. We shall write actionable insights and recommendation and conclusion from both EDA and the best Model.

CHAPTER 2 - STEP -BY-STEP WALK THROUGH OF THE SOLUTION

2.1. Define the goal

This is a report evaluating an airline passenger satisfaction survey

2.2. Get and understand the data

2.1) Literature survey

a) **Airline customer satisfaction and loyalty**: impact of in-flight service quality This study aimed to explore the impact of in-flight service quality on airline customer satisfaction and loyalty, focusing on two passenger classes: prestige (business) and economy. Analysis of passenger data revealed distinct factors influencing satisfaction for each class. For prestige class, six key factors emerged, including alcoholic and non-alcoholic beverages, responsiveness and empathy, reliability, assurance, presentation style of food, and food quality. Economy class identified five important factors: responsiveness and empathy, food quality, alcoholic beverage, non-alcoholic beverage, and reliability. These results suggest the need for tailored in-flight service strategies based on passenger seat class, optimizing customer satisfaction and loyalty.

Ref: <https://link.springer.com/article/10.1007/s11628-009-0068-4>

b) **Service quality and customer satisfaction of a UAE-based airline**: An empirical investigation;
An investigation of Airline Service Quality, Passenger Satisfaction and Loyalty: The case of Royal Jordanian Airline: The findings suggest a significant correlation between service quality, passenger satisfaction, and subsequent loyalty. Specifically, Royal Jordanian Airline's service quality was assessed across various dimensions, and the impact on passenger satisfaction and loyalty was analyzed. The outcomes imply that improvements in specific aspects of service quality could positively influence passenger satisfaction, consequently fostering loyalty to Royal Jordanian Airline. The study underscores the importance of addressing these factors strategically to enhance the overall customer experience and build lasting relationships with passenger.

Reference: [White Rose Etheses Online - document unavailable](#)(Fahed Salim Khatib)

2.2) Source of data:

US Airline Passenger Satisfaction

<https://www.kaggle.com/datasets/johndddddd/customer-satisfaction/data>

2.3) Understanding the data:

For this project, we have two datasets, namely:

a) training dataset with the dimension, (103904 rows, 25 columns)

b) test dataset with the dimension, (25976 rows, 25 columns)

The combined dataset has the dimension (129880 rows, 25 columns)

Training dataset and test dataset constitute 80% and 20% of the combined data.

Null Values

- Variable with missing values: **‘Arrival Delay in Minutes’**
- Count: 310
- Percentage: 2.98%
- Number of Numerical Columns: 6 ['Unnamed: 0', 'id', 'Age', 'Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes']
- Number of Categorical Columns: 19 ['Gender', 'Customer Type', 'Type of Travel', 'Class', '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', 'satisfaction']
- Number of Ordinal variables: 14 ['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']
- Our target variable is satisfaction and it has the levels with count and % as follows:
 - a) neutral or dissatisfied count: 73452 % of total: 56.55%
 - b) satisfied count: 56428 % of total: 43.45%
- Since classes in the target variables are not distributed equally, we conclude that there is some slight imbalance in the dataset (approximately, 57% and 43%). We shall consider Recall, Precision, F1 score, Average-F1 score as the measures to evaluate the performance of the models and to select the best model.
- Using the combined dataset, we shall perform EDA to understand the data better to build useful models.
- Using the training dataset already obtained, we shall build the suitable models.
- Using both the training dataset & test dataset already obtained, we shall test the model performance using the above measures.

We shall treat the missing values using suitable techniques such as replacing the nulls with the median value.

2.3. Pre-process the data including cleaning, transform the data.

Pre-Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

Step 1: Remove duplicate or irrelevant observations.

There are no duplicates. We need to remove the variable, "id" which does not add any value in our analysis.

Step 2: Fix structural errors.

There are no structural errors such as dates in different formats or any word appearing with different spelling

Step 3. Transform data

We notice that the following 14 variables indicate the satisfaction level and we need to convert its data type from int to categorical variable:

1. Inflight wifi service
2. Departure/Arrival time convenient
3. Ease of Online booking
4. Gate location
5. Food and drink
6. Online boarding
7. Seat comfort
8. Inflight entertainment
9. On-board service
10. Leg room service
11. Baggage handling
12. Checkin service
13. Inflight service
14. Cleanliness

Step 4: Handle missing data.

There are missing values in the column, "Arrival Delay in Minutes" in our dataset. We need to impute them by using a suitable technique.

DSE FT B Jun23 G3 Capstone Interim report

Page 10 of 42

Imputation of missing values through Multiple Imputation by Chained Equation

- Detecting and handling missing values in the correct way is important, as they can impact the results of the analysis. It cannot be imputed with general ways of using mean, mode, or median which ignores the inherent relationship among data and also it can pollute the data.
- We observe that on a few occasions, data is missing in a dataset and is related to the other features and hence they can be predicted using other feature values. Imputing by prediction of missing values is superior to other techniques since the inherent relationship among data is not ignored.
- We are imputing missing numeric values using the IterativeImputer class in sklearn.

Ref: <https://www.numpyninja.com/post/mice-and-knn-missing-value-imputationthrough-python>
● There are no missing values after imputing and we need to use this dataset for EDA and model building.

```
lreg = LinearRegression()
imp = IterativeImputer(estimator=lreg, missing_values = np.nan, max_iter = 10, verbose = 2,\
                        imputation_order='roman', random_state = 0)
X = imp.fit_transform(data_enc)

[IterativeImputer] Completing matrix with shape (129880, 23)
[IterativeImputer] Ending imputation round 1/10, elapsed time 6.55
[IterativeImputer] Change: 503.3848173608626, scaled tolerance: 4.9830000000000005
[IterativeImputer] Ending imputation round 2/10, elapsed time 13.01
[IterativeImputer] Change: 0.0, scaled tolerance: 4.9830000000000005
[IterativeImputer] Early stopping criterion reached.

imputed_df = pd.DataFrame(X, columns = data_enc.columns)

missing_zero_values_table(imputed_df)

Your selected dataframe has 23 columns and 129880 Rows.
There are 0 columns that have missing values.
```

Zero Values	Missing Values	% of Total Values	Total Zero & Missing Values	% Total Zero & Missing Values	Data Type
-------------	----------------	-------------------	-----------------------------	-------------------------------	-----------

Fig.2.1. Handle missing data.

Step 5: Filter unwanted outliers.

We need to find numerical variables and check if there are any outliers. If an outlier is a natural part of the population you are studying, you should not remove or impute it.

1.'Age' - Age of the passenger

There are no outliers exceeding the upper threshold value or below the lower threshold value.

2.'Flight Distance' - The flight distance of this journeyThere are outliers exceeding the upper threshold value and we observe that the maximum value is 4983 miles and currently the longest flight distance is 9,585 miles.

3.'Departure Delay in Minutes' - Minutes delayed when departure There are outliers exceeding the upper threshold value and we observe that the maximum value is 1592 minutes or approximately 27 hours which is a valid value.

4.'Arrival Delay in Minutes' - Minutes delayed when Arrival. There are outliers exceeding the upper threshold value and we observe that the maximum value is 1584 minutes or approximately 26 hours which is a valid value.

So, we are not treating the data for outliers.

References:

1) <https://www.scribbr.com/frequently-asked-questions/when-to-remove-an-outlier/#:~:text=Some%20outliers%20represent%20natural%20variations,processing%20errors%2C%20or%20poor%20sampling.>

2) <https://statisticsbyjim.com/basics/remove-outliers/>

2.4 Exploratory Data Analysis

2.4.1 Relationship between variables

1) Categorical Variables

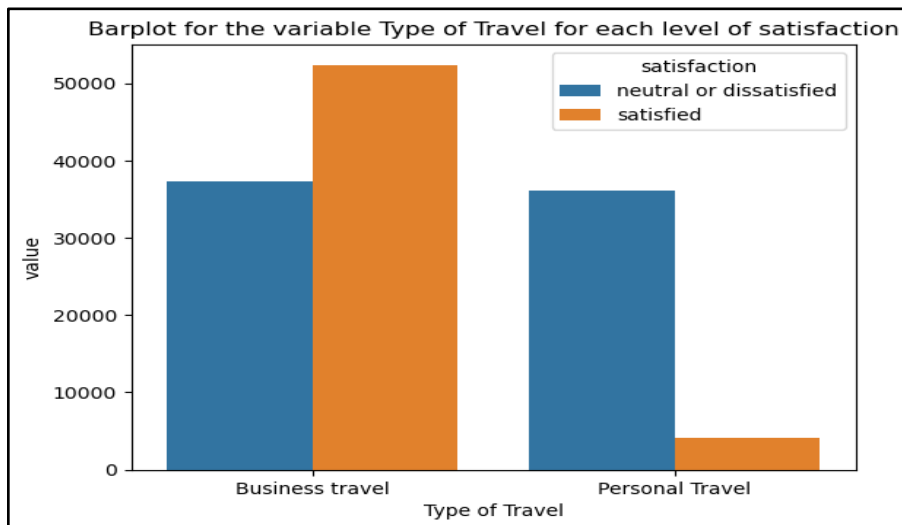


Fig.2.2. Barplot for Type of travel vs satisfaction level

Type of Travel: Business travellers contribute more to the traveling frequency and are more satisfied than personal travellers.

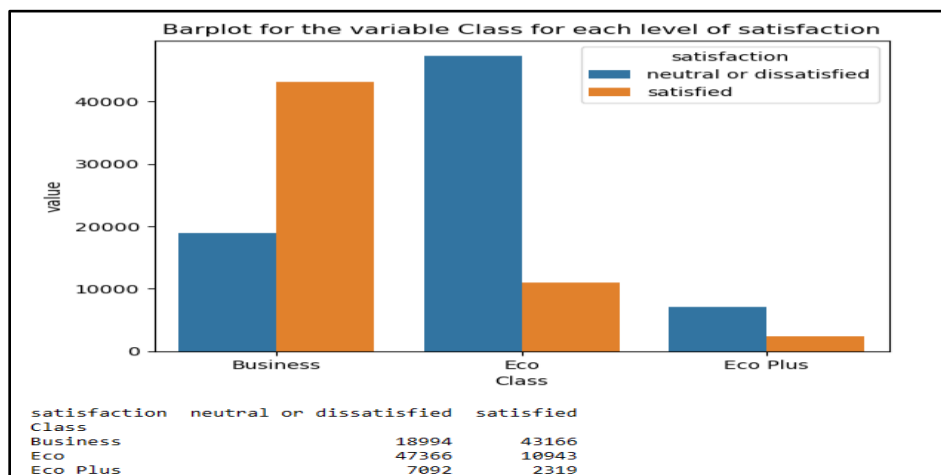


Fig.2.3. Barplot for the variable classes' vs satisfaction level

Class: Business class get the highest satisfaction, eco and eco plus have higher dissatisfaction

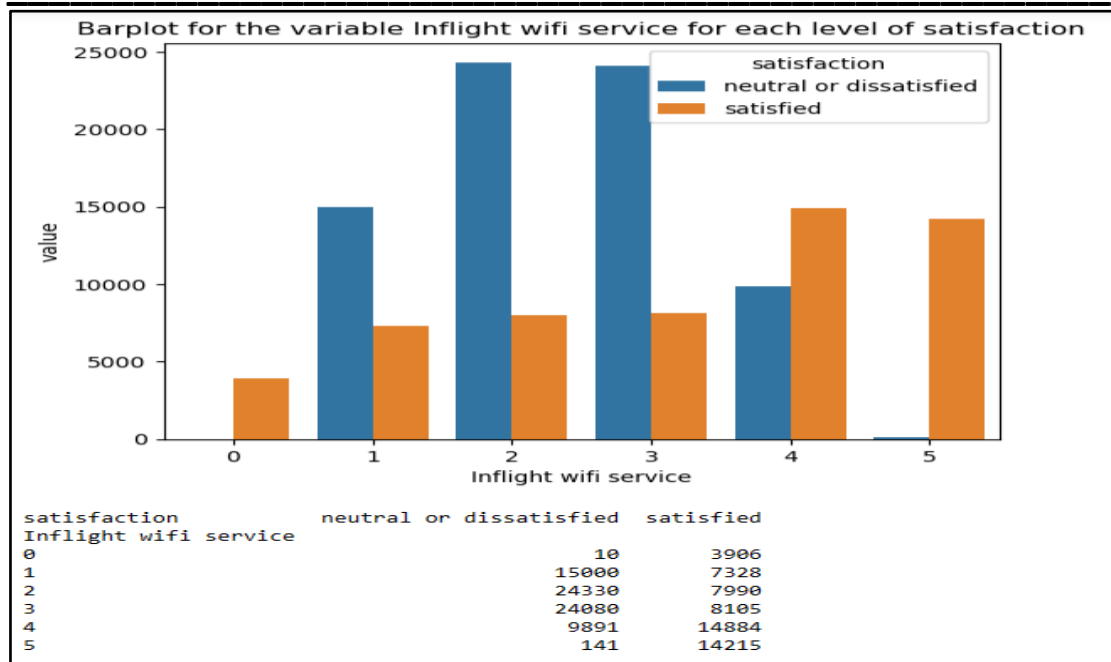


Fig 2.4. Barplot for inflight WiFi service vs satisfaction

Inflight wifi service: higher the rating of inflight wifi service higher is the people getting

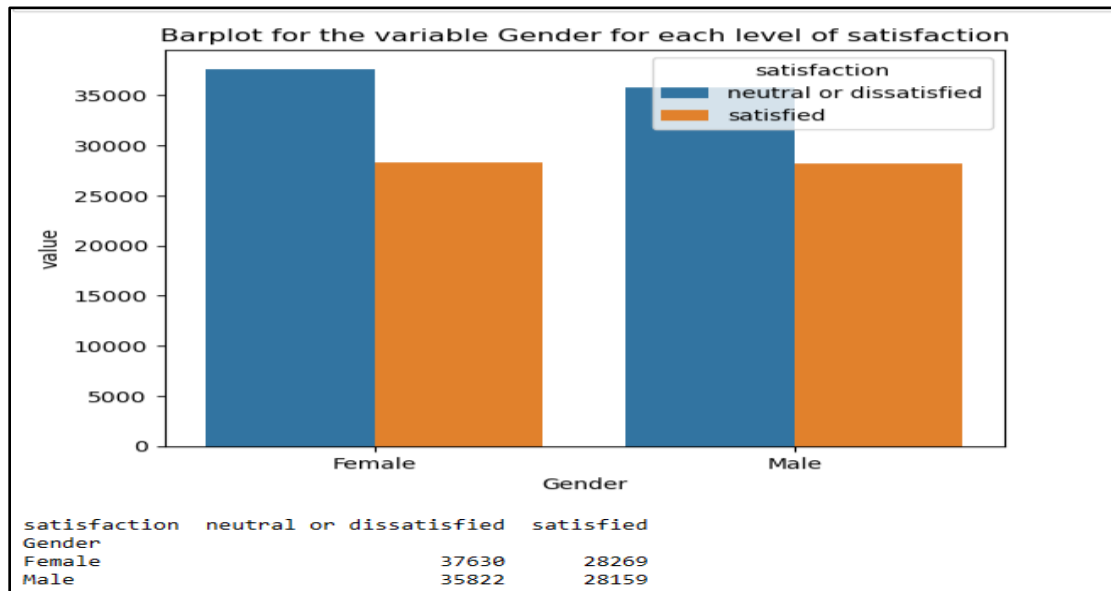


Fig 2.5 Barplot for Gender vs Satisfaction

Gender: Business class get the highest satisfaction, eco and eco plus have higher dissatisfaction

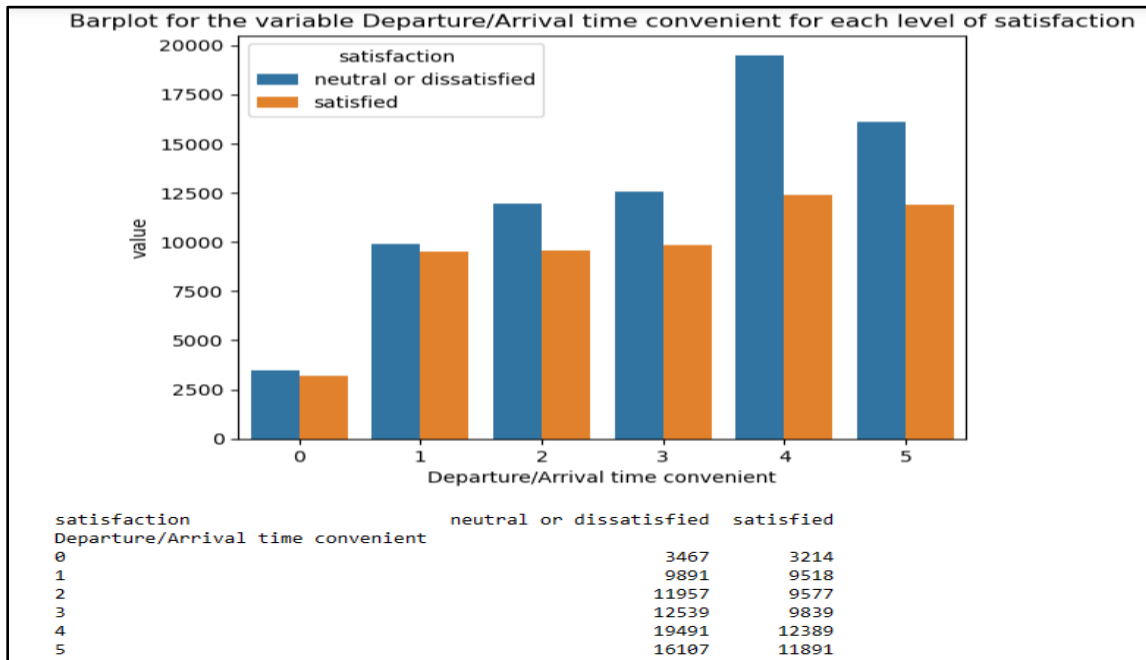


Fig 2.6. Barplot for Departure/Arrival time convenient vs satisfaction

Departure/Arrival time convenient: The customers are dissatisfied in all the levels



Fig 2.7. Barplot for Ease of Online booking vs satisfaction

Ease of Online booking: The Customers are more satisfied with lower ease of online booking

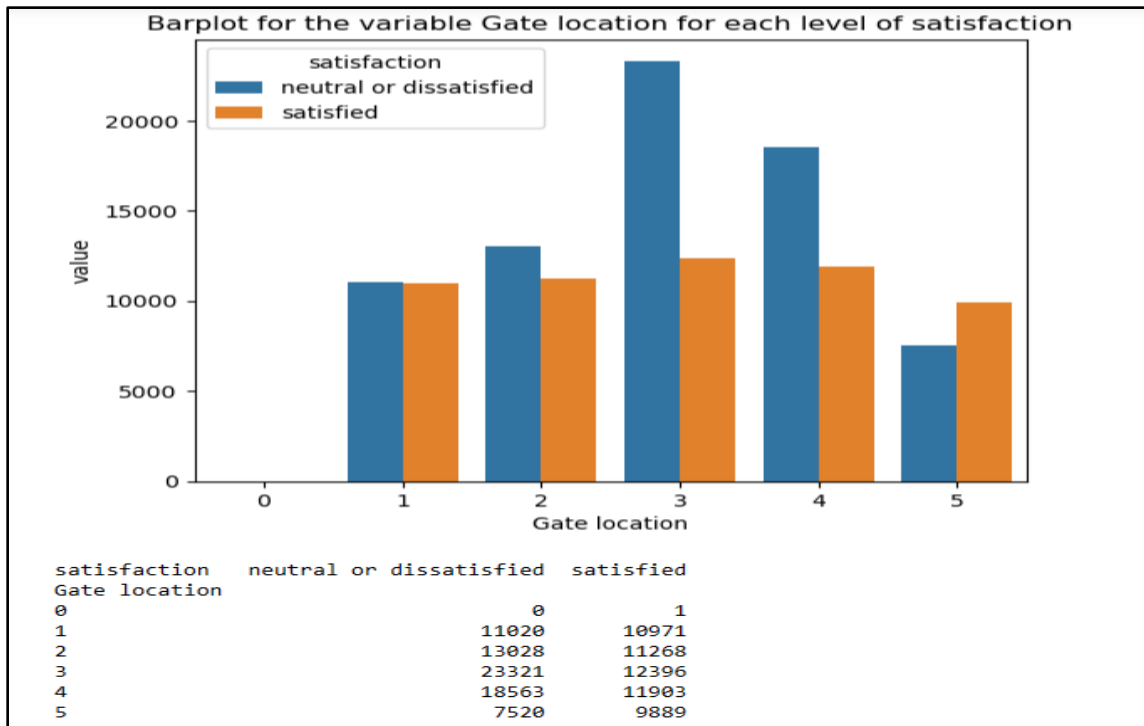


Fig 2.8 Barplot for Gate location vs satisfaction

Gate location: customers have dissatisfaction with gate location

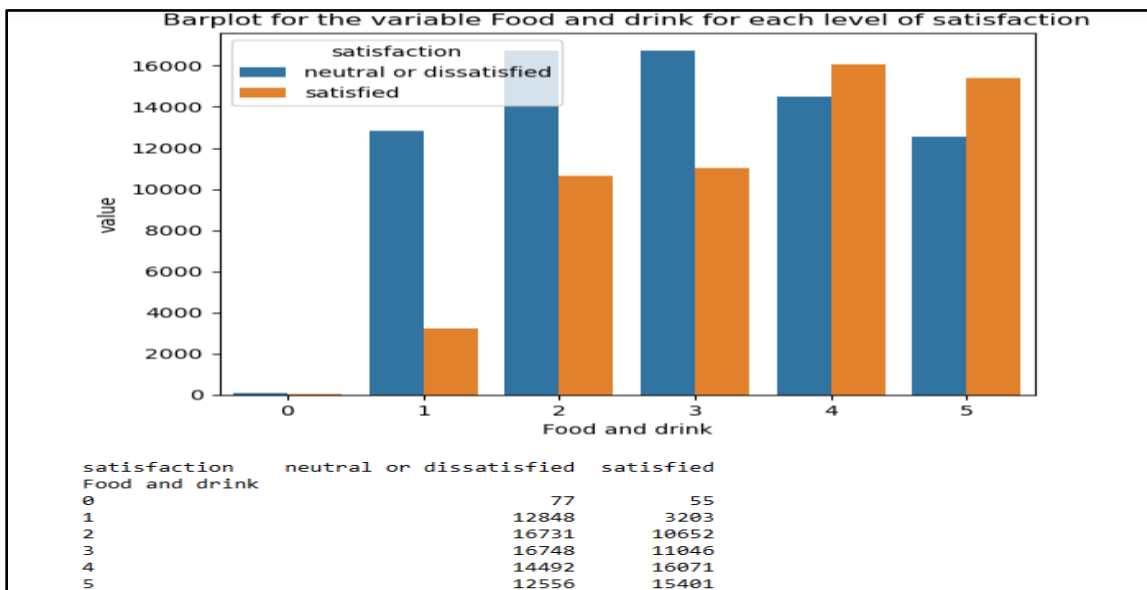


Fig 2.9 Barplot for Food and drink vs satisfaction

Food and drink: Customers are fairly more satisfied with food and drink service

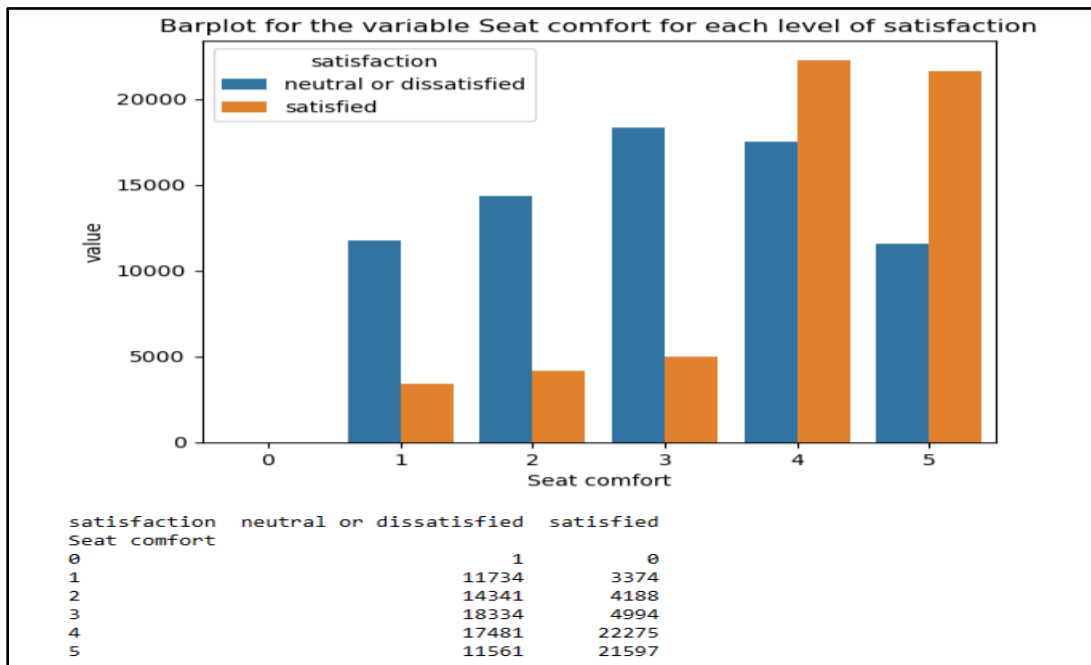


Fig 2.10 Barplot for Seat comfort vs satisfaction

Seat comfort: Customers are highly satisfied with higher levels of seat comfort

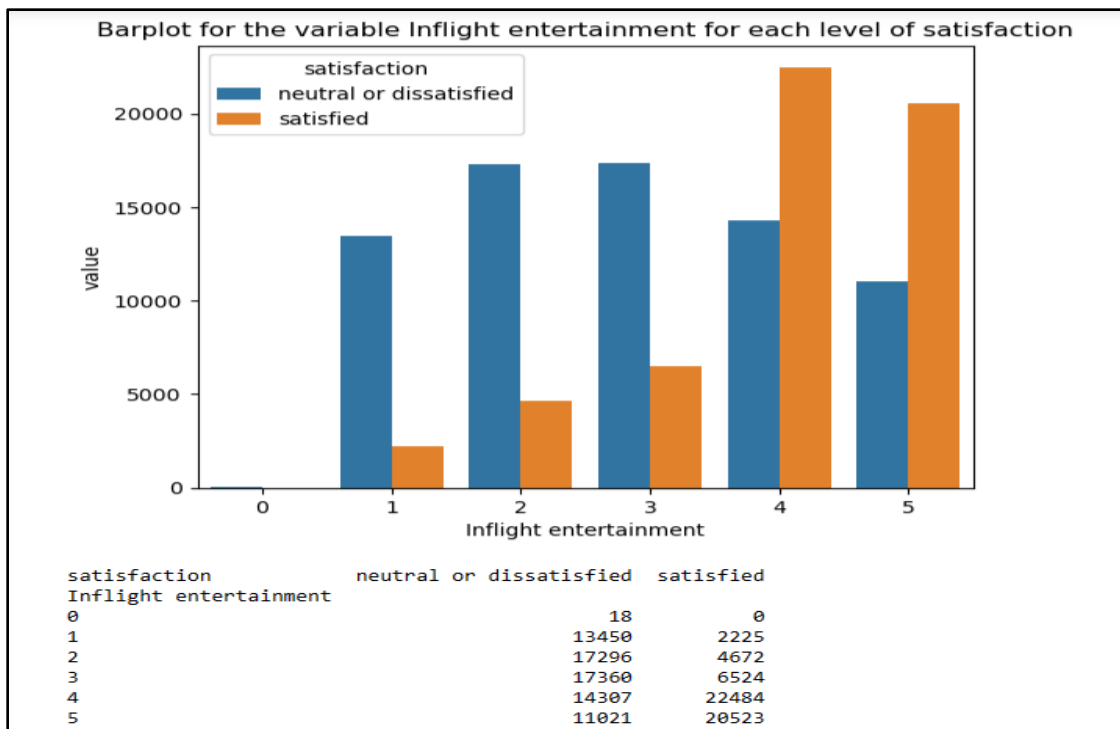


Fig 2.11 Barplot for Inflight entertainment vs satisfaction

Inflight entertainment: Customers who are satisfied like quality inflight entertainment

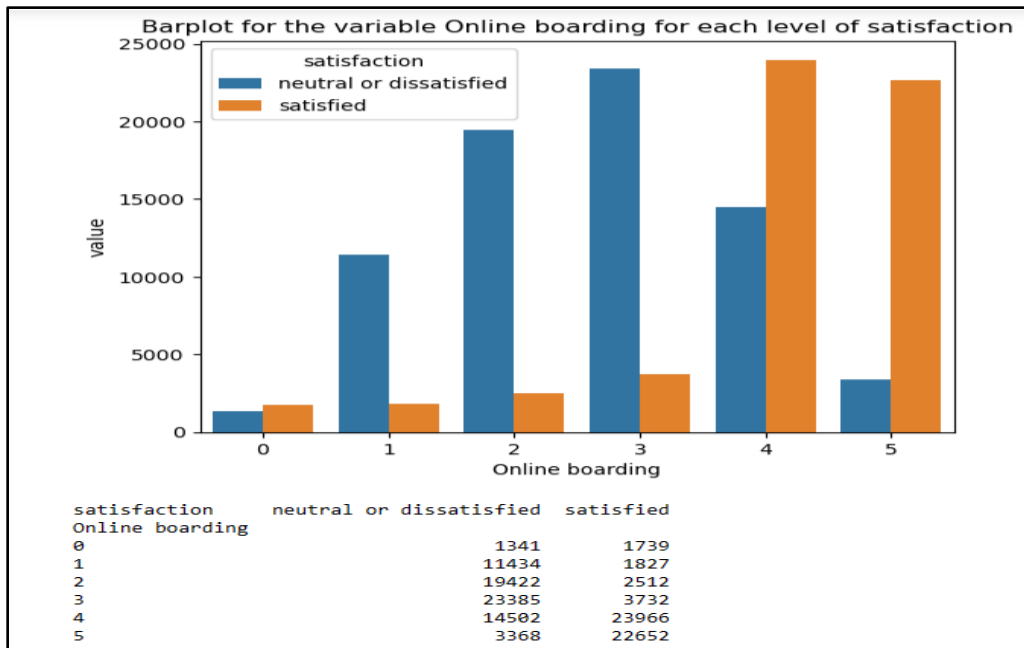


Fig 2.12 Barplot for Online boarding vs satisfaction

Online boarding: Customers are highly satisfied with online boarding.

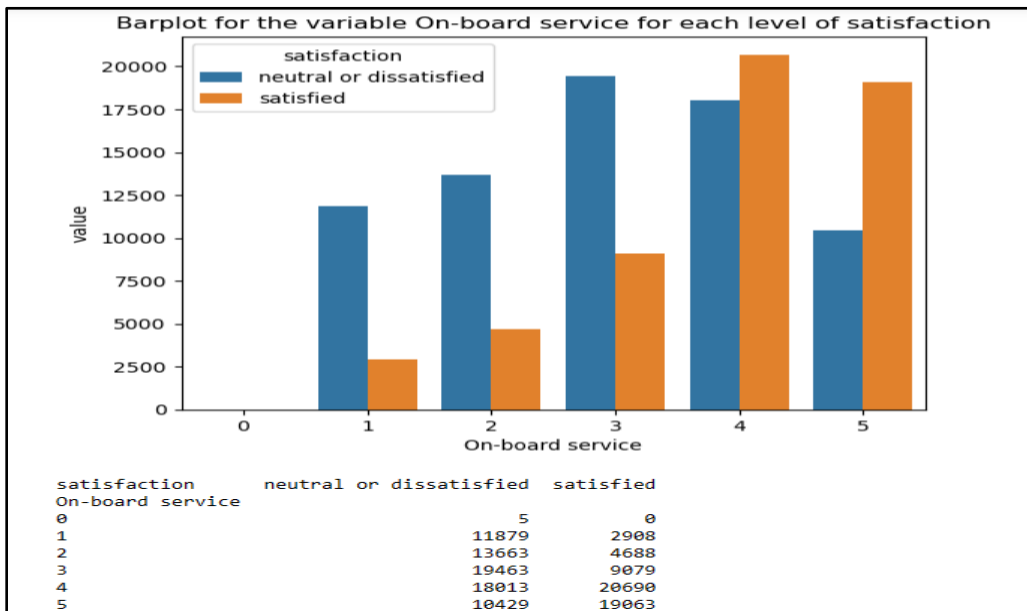


Fig.2.13 Barplot for On-board service vs satisfaction

On-board service: Customers who enjoy on board services are more likely to be satisfied

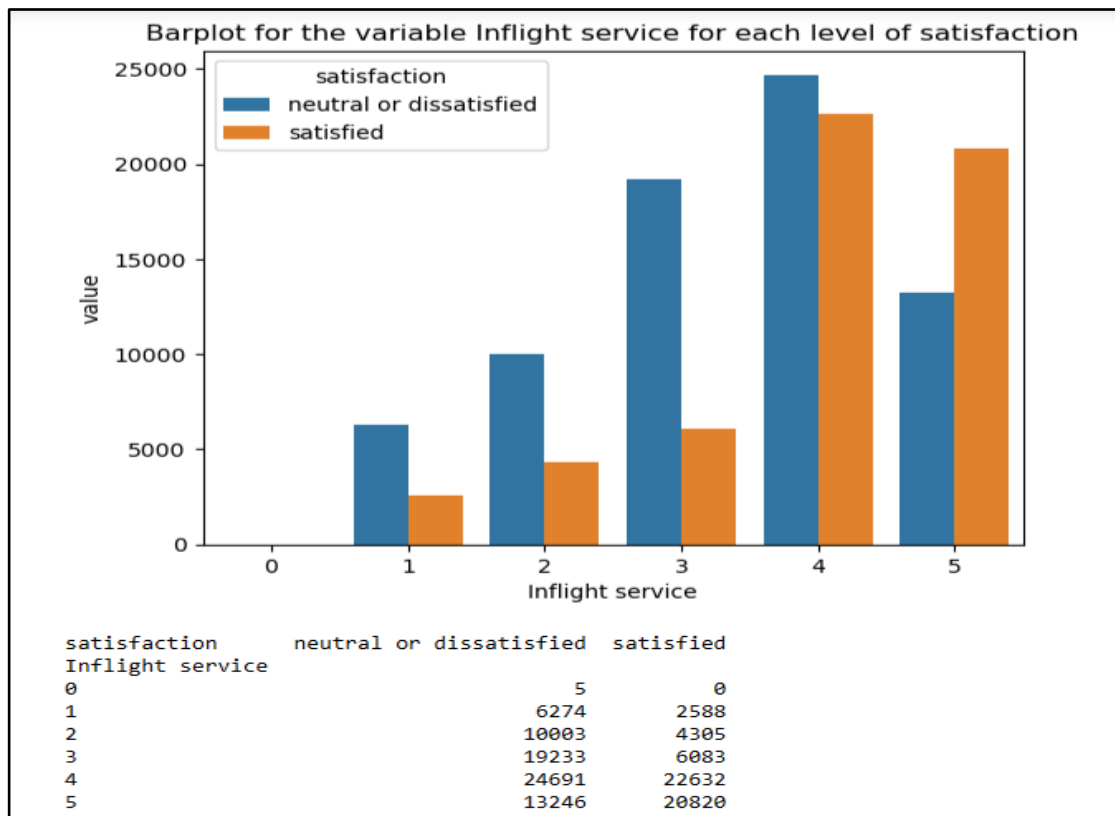


Fig 2.14 Barplot for Inflight service vs satisfaction

Inflight service: Customers are more dissatisfied with overall inflight service quality

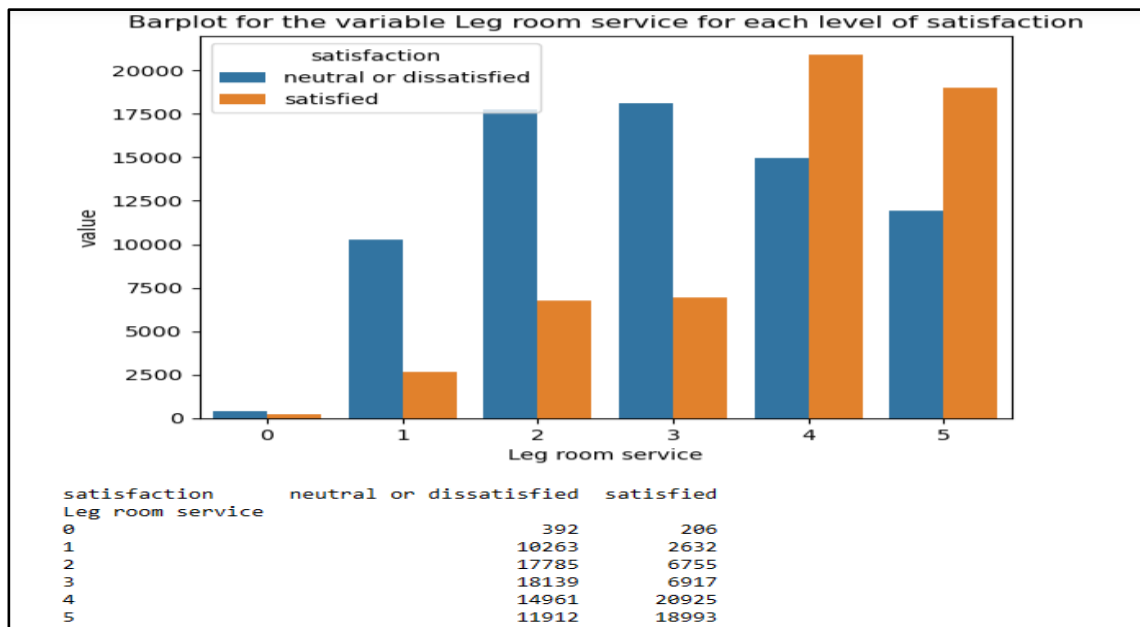


Fig 2.15 Barplot for Leg room service vs satisfaction

Leg room service: Customers are more satisfied with larger leg space

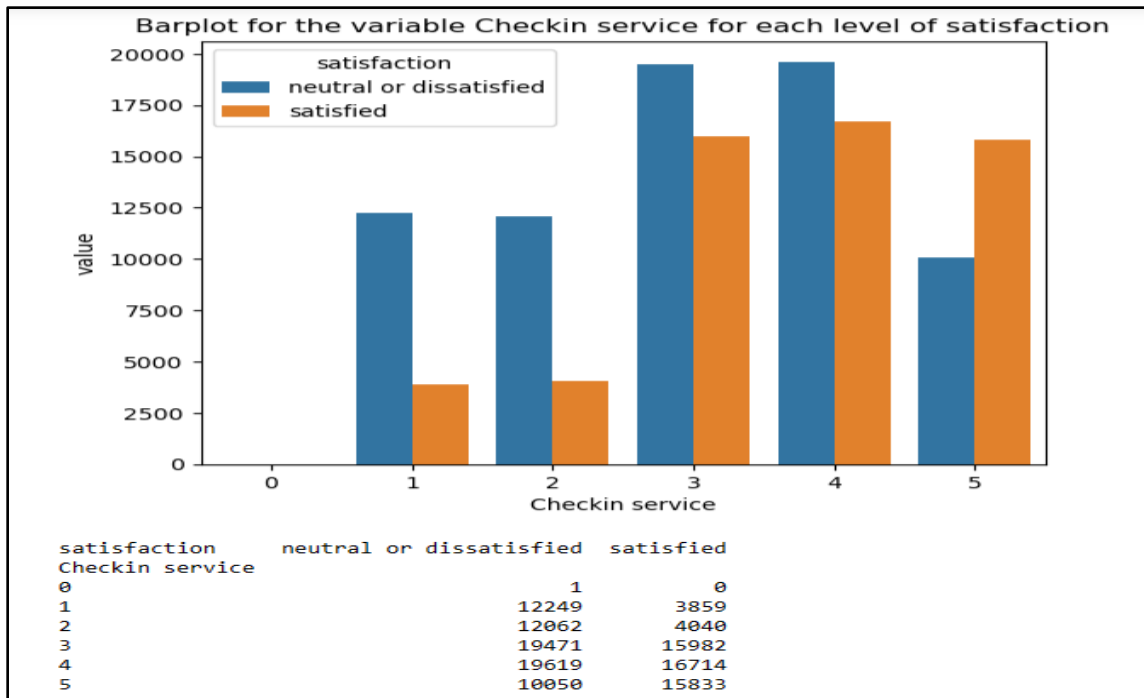


Fig 2.16 Barplot for Check-in service vs satisfaction

Check-in service: Customers are more dissatisfied with overall check-in service quality

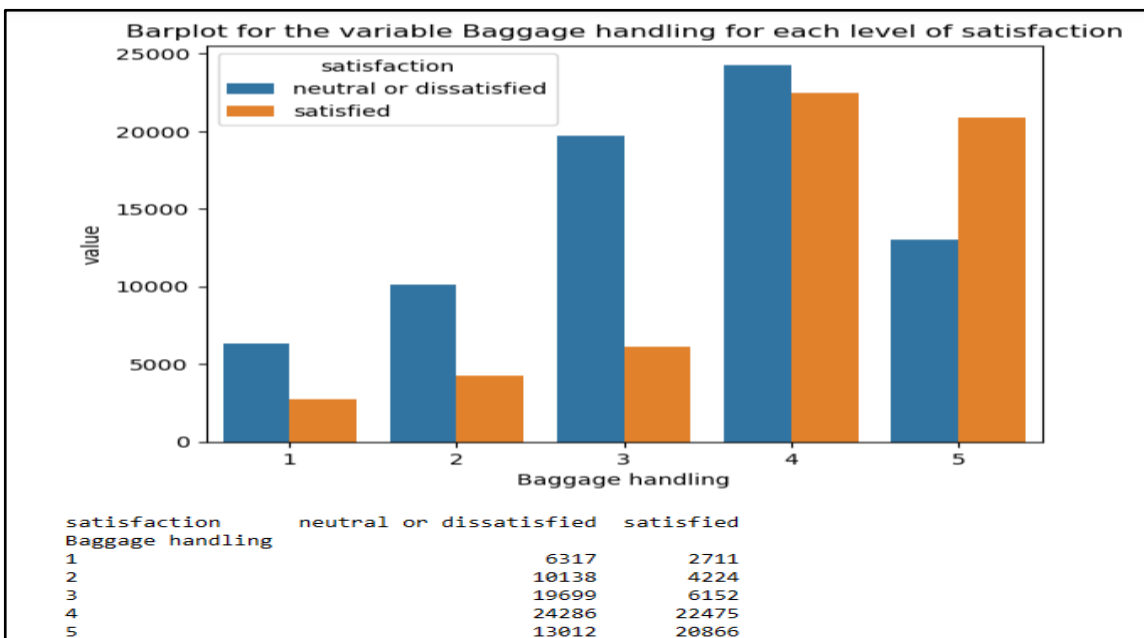


Fig 2.17 Barplot for Baggage handling vs satisfaction

Baggage handling: Customers are more satisfied with overall baggage handling quality

2) Numerical Variables:

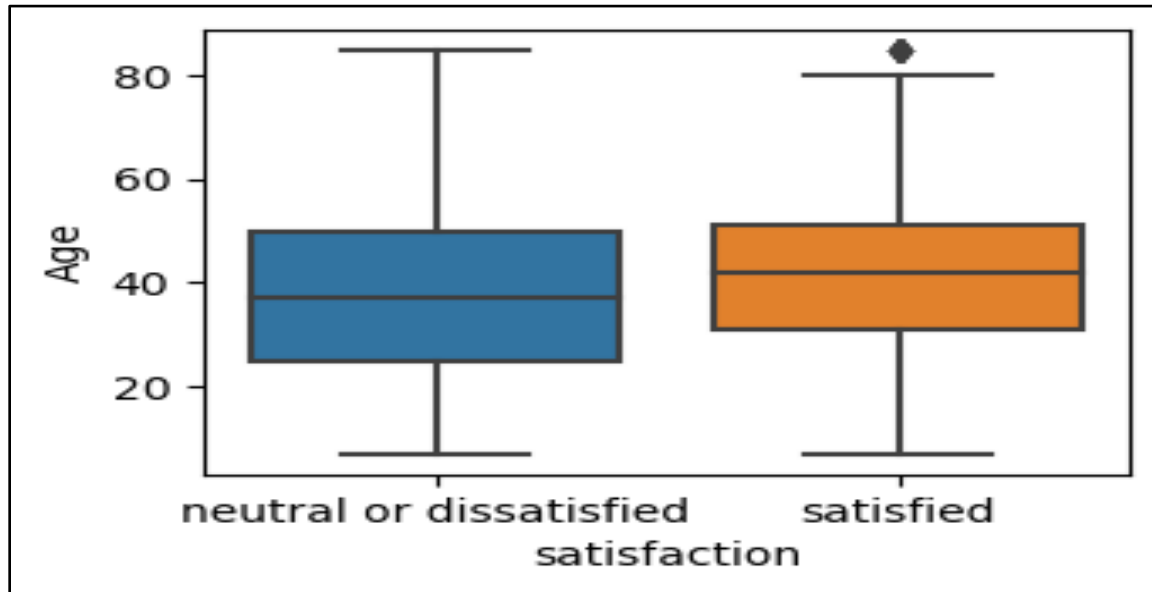


Fig 2.18 Box plot for age vs satisfaction

Age: As Age increases the median satisfaction level also increases.

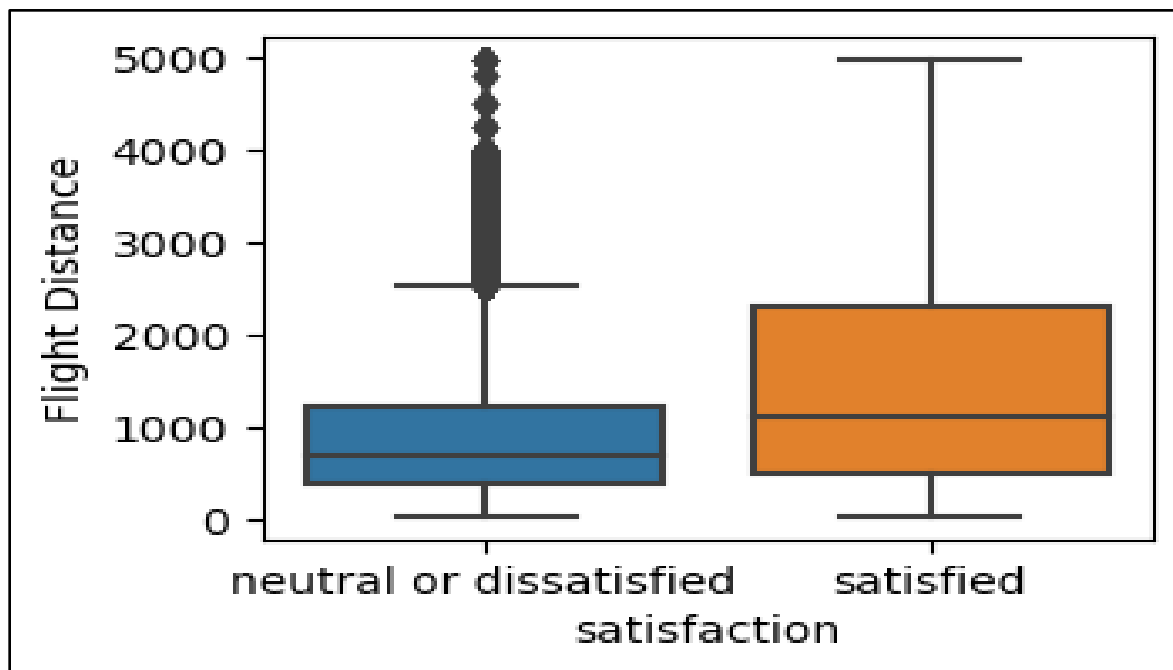


Fig 2.19 Box plot for Flight Distance vs Satisfaction

Flight Distance: As distance of travel increases customers are more satisfied

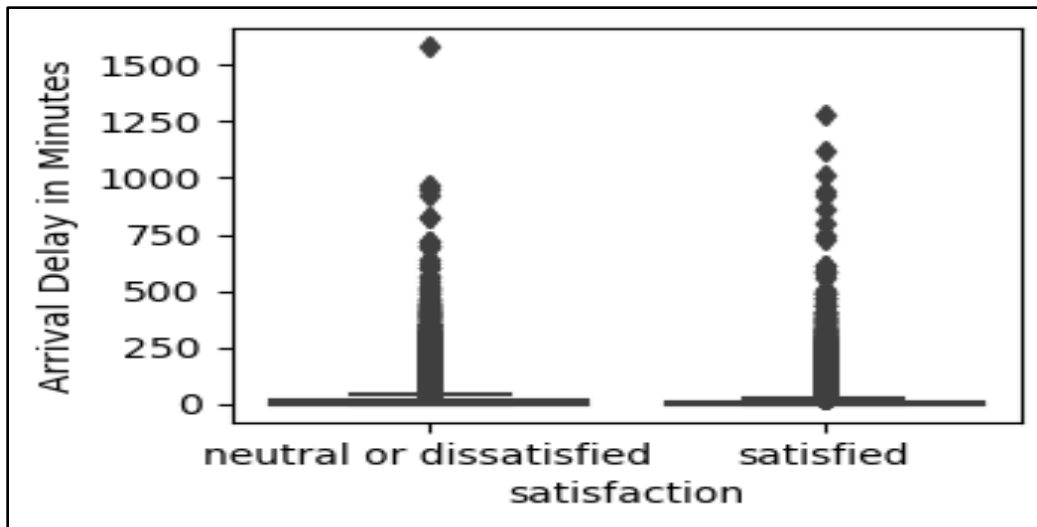


Fig 2.20 Box plot for departure delay in minutes vs satisfaction

Departure Delay in Minutes: Most of the customers are not tolerant to departure delay

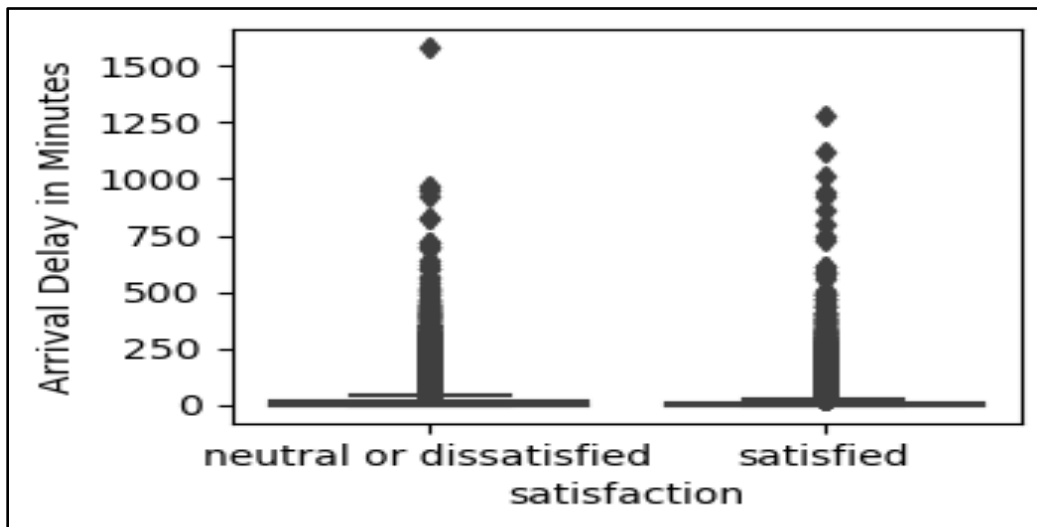


Fig 2.21 Box plot for Arrival Delay in minutes vs Satisfaction

Arrival Delay in minutes: Most of customers are not tolerant to arrival dela

Check for multicollinearity:

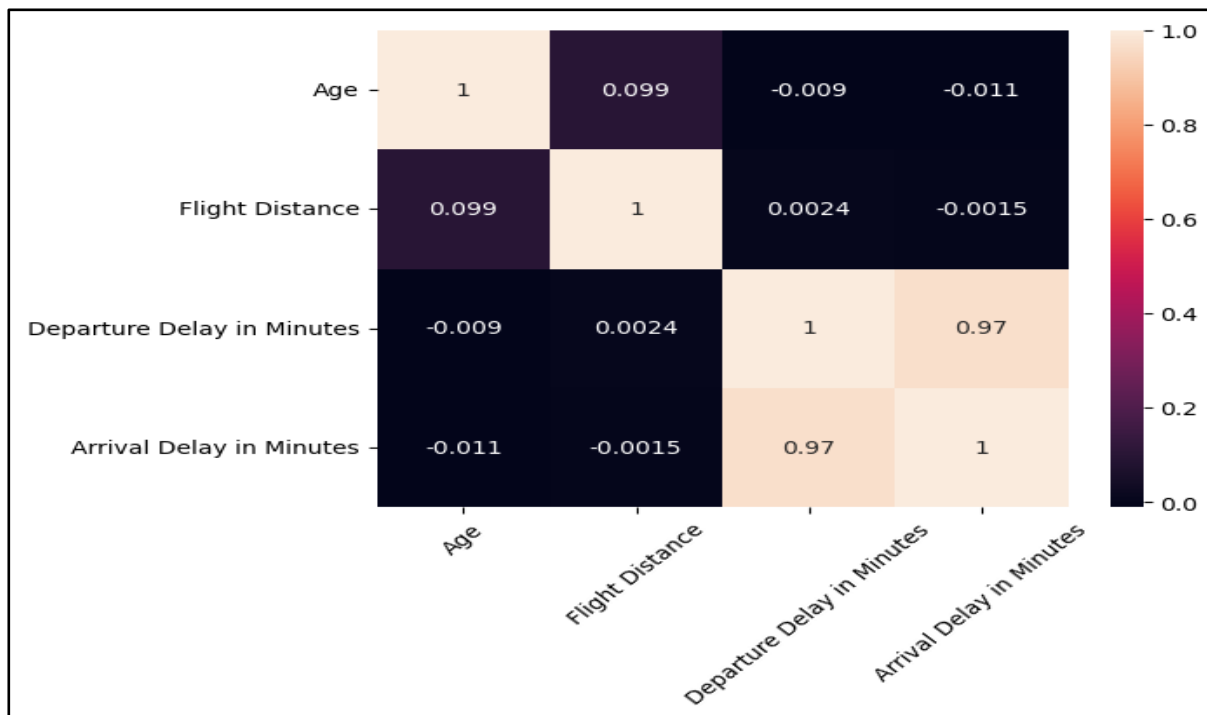


Fig 2.22 Heatmap for checking multicollinearity

Departure Delay in minutes and arrival delay in minutes have strong positive correlation (0.97).

Distribution of Variables:

We shall check the distribution of the following four numerical variables:

- 1) Age
- 2) Flight Distance
- 3) Departure Delay in Minutes
- 4) Arrival Delay in Minutes

The two well-known tests of normality, namely, the Kolmogorov–Smirnov test and the Shapiro–Wilk test are most widely used methods to test the normality of the data. The Shapiro–Wilk test is a more appropriate method for small sample sizes (<50 samples) although it can also be handled on larger sample sizes while Kolmogorov–Smirnov test is used for $n \geq 50$.

Reference:

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/#:~:text=The%20Shapiro%E2%80%933Wilk%20test%20is,taken%20from%20normal%20distributed%20population.>

The Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a normal distribution.

1) Histogram and KS test for the variable, Age

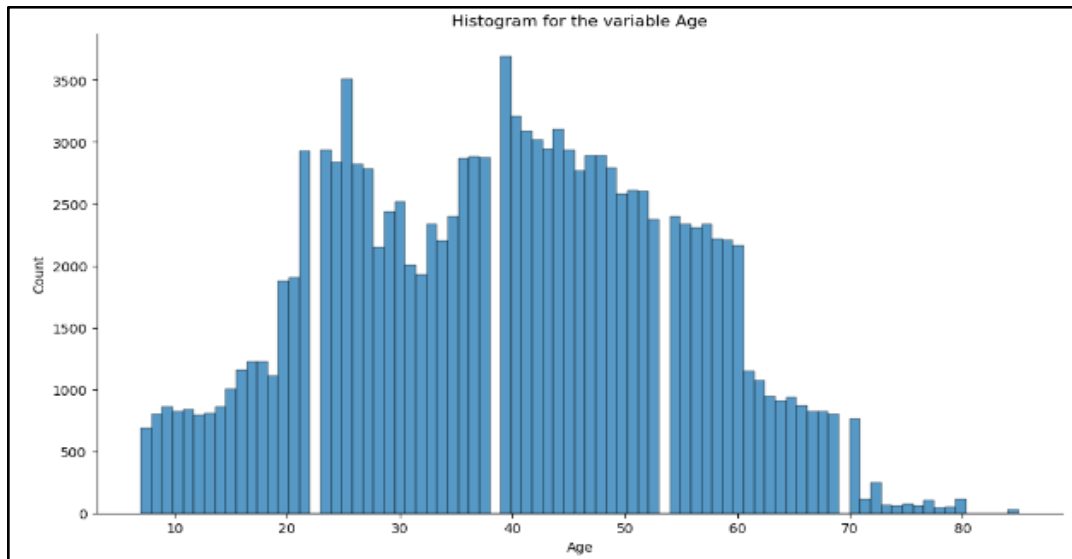


Fig 2.23 Histogram for the variable Age

Observation: From the histogram and the p-value for the KS test, we observe that the variable, Age is not normally distributed.

2) Histogram and KS test for the variable Flight Distance

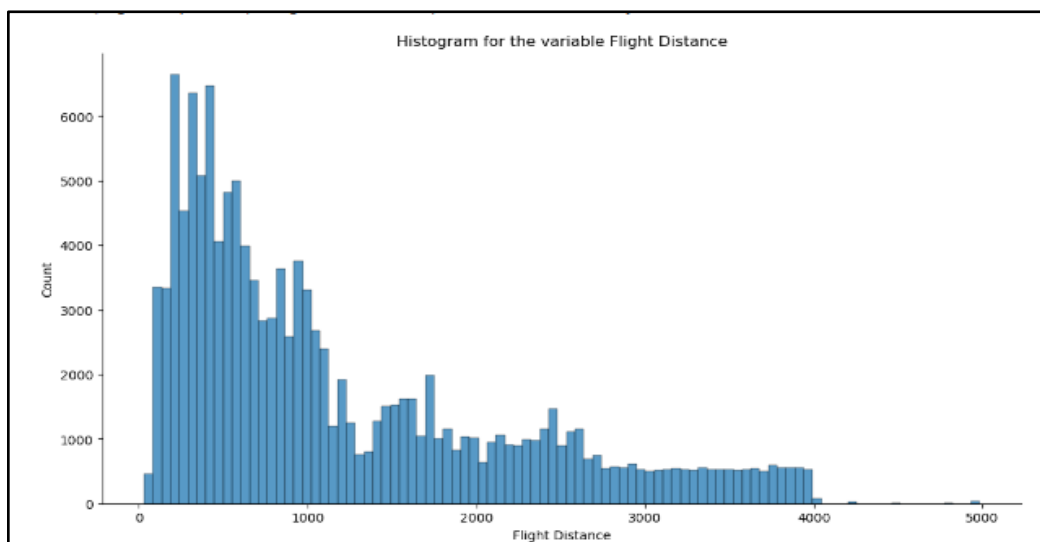


Fig 2.24 Histogram for the variable Flight Distance

Observation: From the histogram and the p-value for the KS test, we observe that the variable, Flight Distance is not normally distributed.

1) Histogram and KS test for the variable, Departure Delay in Minutes

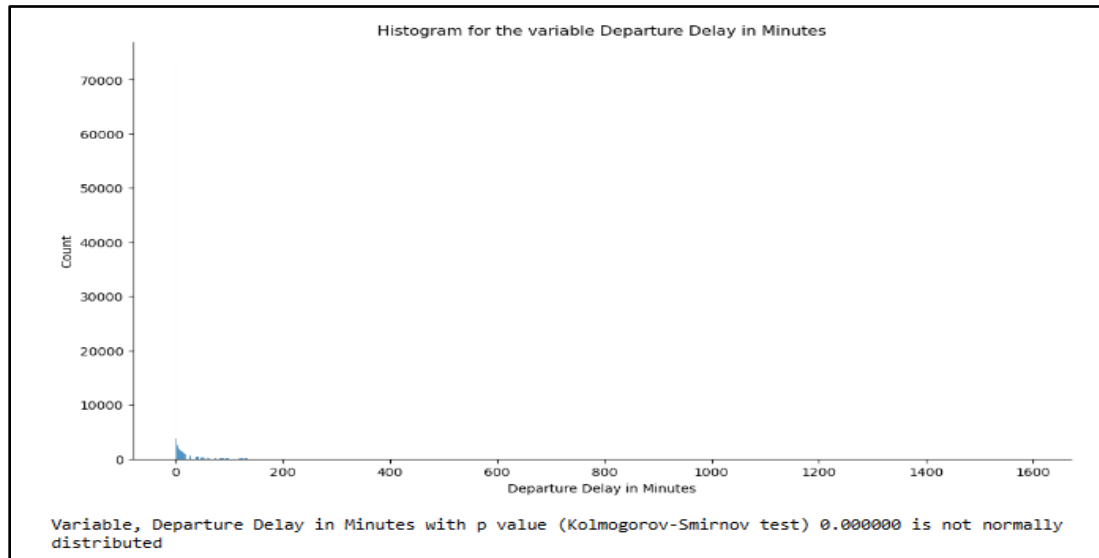


Fig 2.25 Histogram for the variable Departure Delay in Minutes

Observation: From the histogram and the p-value for the KS test, we observe that the variable, Departure Delay in Minutes is not normally distributed.

2) Histogram and KS test for the variable, Arrival Delay in Minutes

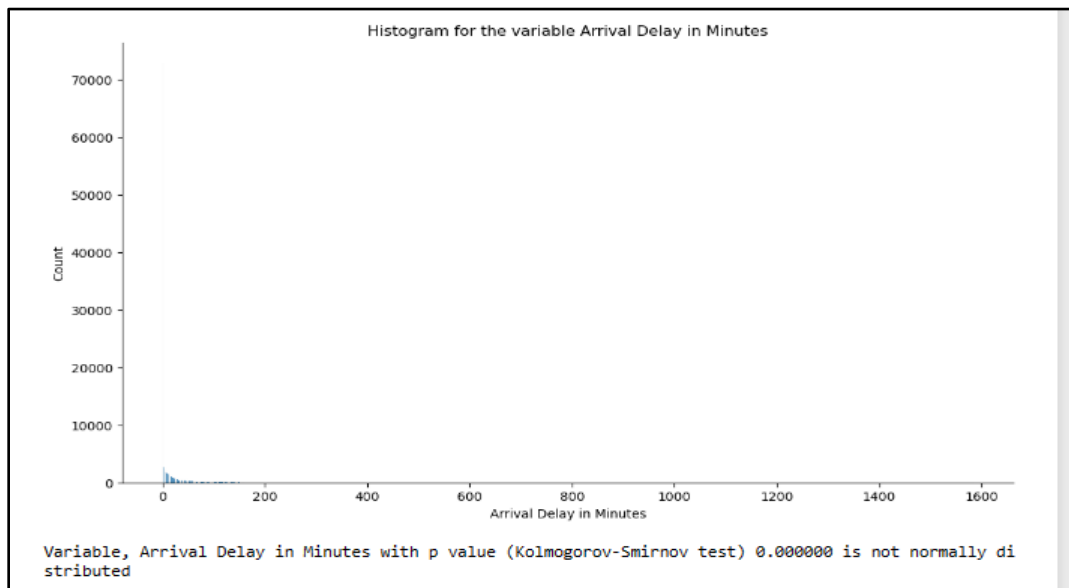


Fig 2.26 Histogram for the variables Arrival Delay in Minutes

Observation: It is not normally distributed

Presence of outliers and its treatment:

We shall draw boxplots for each of the following numerical variables to detect outliers:

- 1) Age
- 2) Flight Distance
- 3) Departure Delay in Minutes
- 4) Arrival Delay in Minutes

1) Boxplot for the variable, Age

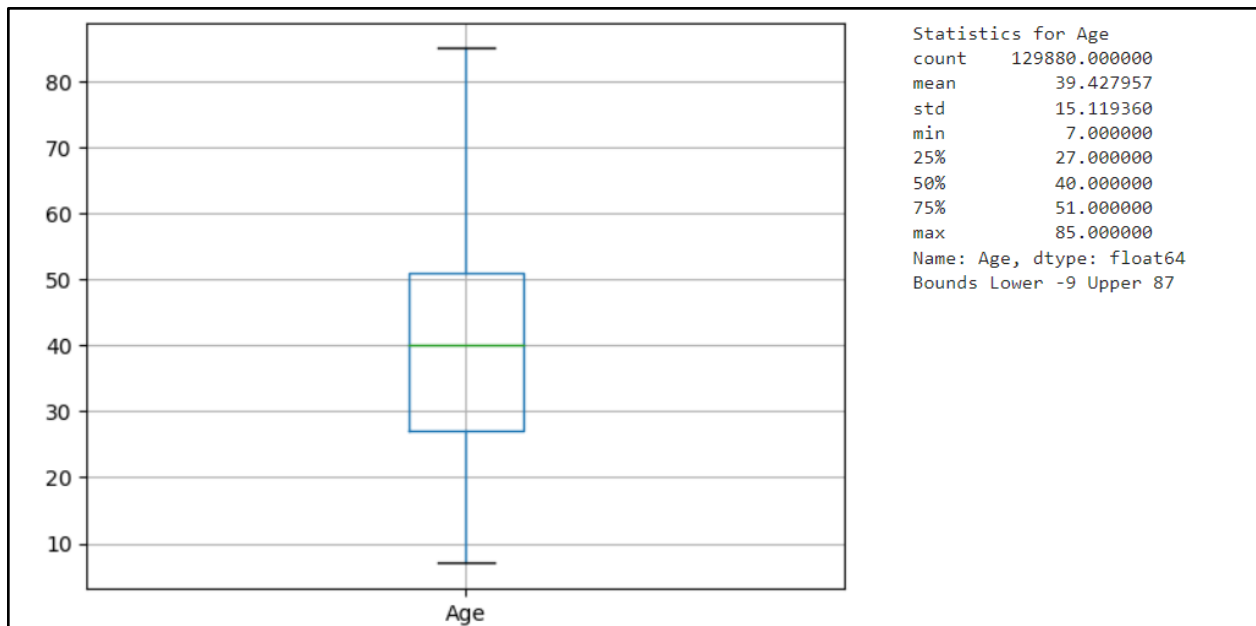


Fig 2.27 Boxplot for the variable Age

2) Boxplot for the variable, Flight Distance

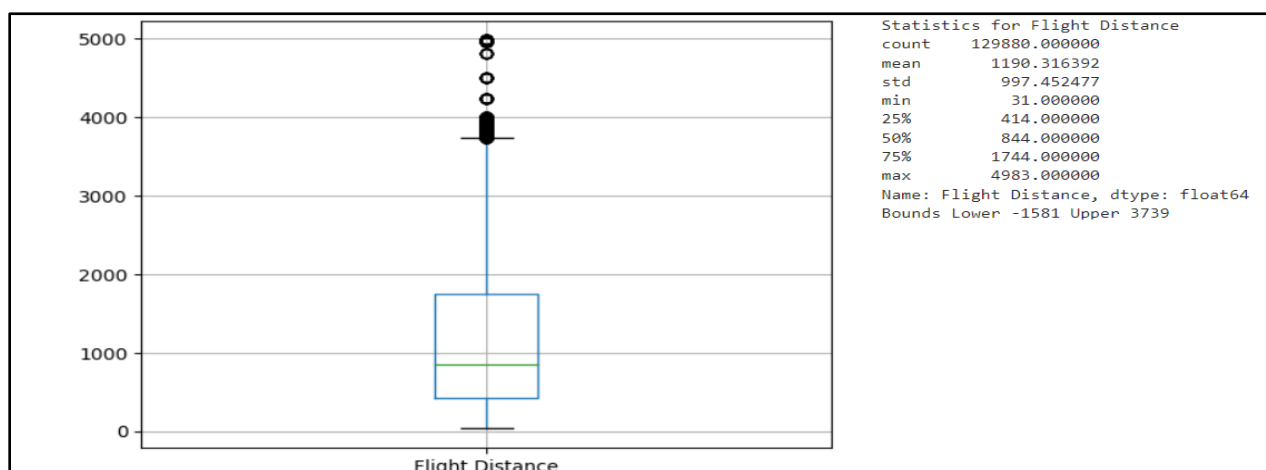


Fig 2.28 Boxplot for the variable Flight Distance

2) Boxplot for the variable, Departure delay in minutes:

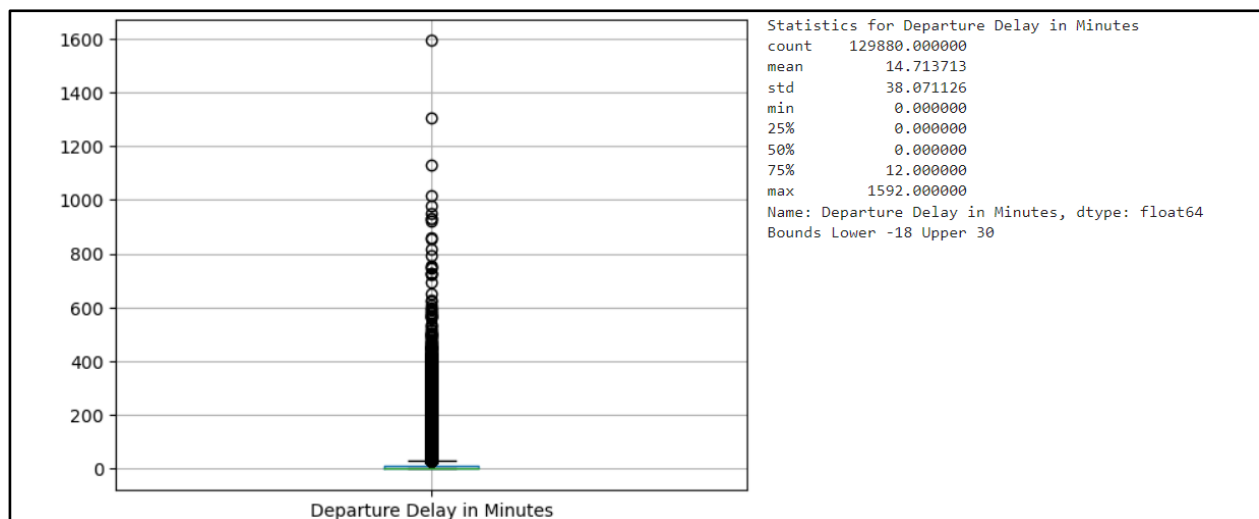


Fig 2.29 Boxplot for the variable Departure delay in minutes

3) Boxplot for the variable, Arrival delay in minutes

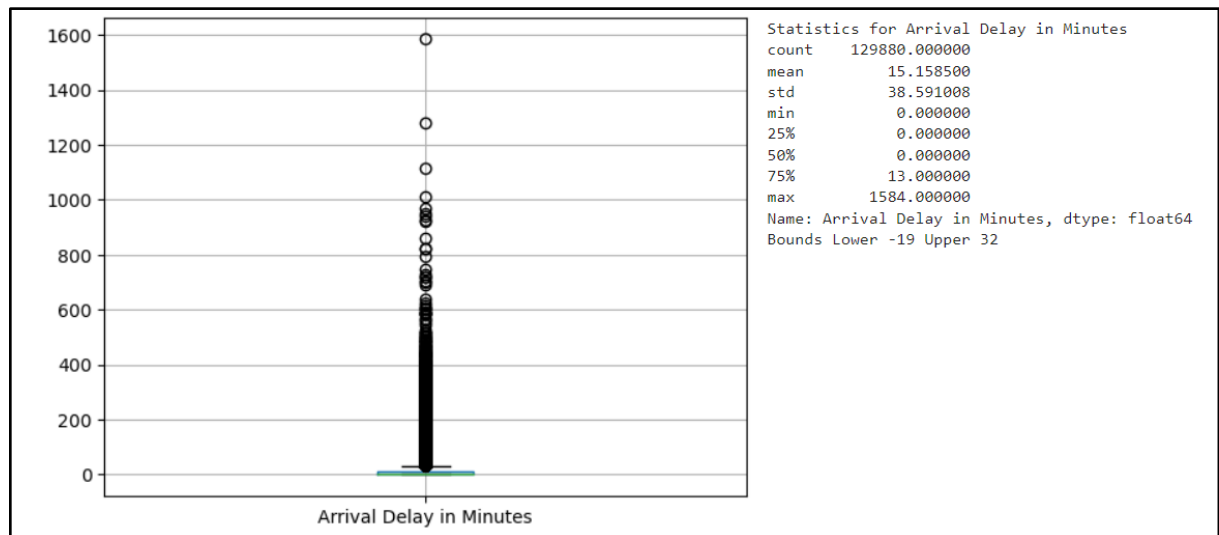


Fig 2.30 Boxplot for the variable Arrival delay in minutes

Observations:

If an outlier is a natural part of the population you are studying, you should not remove or impute it.

1) 'Age' - Age of the passenger

There are no outliers since there are no outliers exceeding the upper threshold value or below the lower threshold value.

2) **'Flight Distance' - The flight distance of this journey**

There are outliers exceeding the upper threshold value and we observe that the maximum value is 4983 miles and currently the longest flight distance is 9,585 miles.

3) **'Departure Delay in Minutes' - Minutes delayed when departure**

There are outliers exceeding the upper threshold value and we observe that the maximum value is 1592 minutes or approximately 27 hours which is a valid value.

4) **'Arrival Delay in Minutes' - Minutes delayed when Arrival.**

There are outliers exceeding the upper threshold value and we observe that the maximum value is 1584 minutes or approximately 26 hours which is a valid value.

So, we are not treating the data for outliers.

References:

1) <https://www.scribbr.com/frequently-asked-questions/when-to-remove-an-outlier/#:~:text=Some%20outliers%20represent%20natural%20variations,processing%20errors%2C%20or%20poor%20sampling.>

2) <https://statisticsbyjim.com/basics/remove-outliers/>

Statistical significance of variables:

For Categorical Variables:

We shall use the Chi-Square Test of Independence. Do we test the independence of two categorical variables?

Hypothesis:

H_0 : In the population, the two categorical variables are independent.

H_1 : In the population, the two categorical variables are dependent.

Assumptions of Chi Square Test:

- 1) Both variables are categorical.
- 2) All observations are independent.
- 3) Cells in the contingency table are mutually exclusive.
- 4) Expected value of cells should be 5 or greater in at least 80% of cells.

Assumption 1: Both variables are categorical.

We are selecting the variables which are of data type and our target variable, Satisfaction, a categorical variable, object for performing chi-square test. Hence this assumption is satisfied.

Assumption 2: All observations are independent.

It's assumed that every observation in the dataset is independent. That is, the value of one observation in the dataset does not affect the value of any other observation.

Assumption 3: Cells in the contingency table are mutually exclusive.

It's assumed that individuals can only belong to one cell in the contingency table. That is, cells in the table are mutually exclusive – an individual cannot belong to more than one cell.

Assumption 4: Expected value of cells should be 5 or greater in at least 80% of cells.

Expected value of cells should be 5 or greater in more than 95% of cells which is more than the stipulated 80% of cells.

Hence, all the assumptions of the Chi-square test are satisfied.

Now, we shall apply this test for each pair of all categorical independent variables and the dependent variable.

Results of Chi-square tests of independence			
	Column	chi2_value	P value
0	Gender	16.352081	5.259838e-05
0	Customer Type	4493.188803	0.000000e+00
0	Type of Travel	26282.520993	0.000000e+00
0	Class	32906.171859	0.000000e+00
0	Inflight wifi service	35891.433370	0.000000e+00
0	Departure/Arrival time convenient	601.462958	9.767302e-128
0	Ease of Online booking	12846.702395	0.000000e+00
0	Gate location	3069.907992	0.000000e+00
0	Food and drink	6571.202895	0.000000e+00
0	Online boarding	49531.218396	0.000000e+00
0	Seat comfort	19538.740376	0.000000e+00
0	Inflight entertainment	23071.602435	0.000000e+00
0	On-board service	14342.659690	0.000000e+00
0	Leg room service	15200.778756	0.000000e+00
0	Baggage handling	10820.213523	0.000000e+00
0	Checkin service	8143.773216	0.000000e+00
0	Inflight service	10357.930495	0.000000e+00
0	Cleanliness	12948.918125	0.000000e+00

Fig 2.31 Results of Chi-square tests of independence

Result

Since for all the 18 categorical variables, the p value of the chi-square test is less than the level of significance of 5%, we conclude that there is some relationship between each of the independent variables and the target variable.

For Numerical Variables

A deviation from a set of data's normal distribution or symmetrical bell curve is referred to as skewness. The term "skewed" refers to a curve that has been moved to the left or right. The degree to which a particular distribution deviates from the normal distribution can be expressed quantitatively as skewness.

If skewness is zero, the distribution is symmetrical. Zero skewness is shown via a normal distribution.

If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

Ref: <https://pythonguides.com/python-scipy-stats-skew/>

```
def chk_skewness(df, var):  
    from scipy import stats  
    skw = df[var].skew()  
    if skw == 0:  
        res = "Distribution is symmetrical"  
    elif (skw > 1) or (skw < -1):  
        res = "Distribution is highly skewed"  
    elif (skw > -1) or (skw < 0.5):  
        res = "Distribution is moderately skewed"  
    elif (skw > -0.5) or (skw < 0.5):  
        res = "Distribution is approximately symmetric"  
    print("Skewness for Var: %s is %f and hence %s" % (var, skw, res))  
    return
```

Fig 2.32 Function used to calculate Skewness

Skewness for 'Age' is -0.003606 and hence Distribution is moderately skewed

Skewness for 'Flight Distance' is 1.108142 and hence Distribution is highly skewed

Skewness for 'Departure Delay in Minutes' is 6.821980 and hence Distribution is highly skewed

Skewness for 'Arrival Delay in Minutes' is 6.670125 and hence Distribution is highly skewed

Result:

- 1) All the numerical variables (Age', 'Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes') are not normally distributed
- 2) From The boxplot for the variables, 'Age' and 'Flight Distance', we observe that the median for the satisfied class of the target variable, 'satisfaction' are placed to the right of the neutral or dissatisfied class of the target variable, 'satisfaction'.

Class imbalance and its treatment:

We have an imbalanced dataset (57% and 43%) but want to assign greater contribution to classes with more examples in the dataset, then the weighted average is preferred. This is because, in weighted averaging, the contribution of each class to the F1 average is weighted by its size.

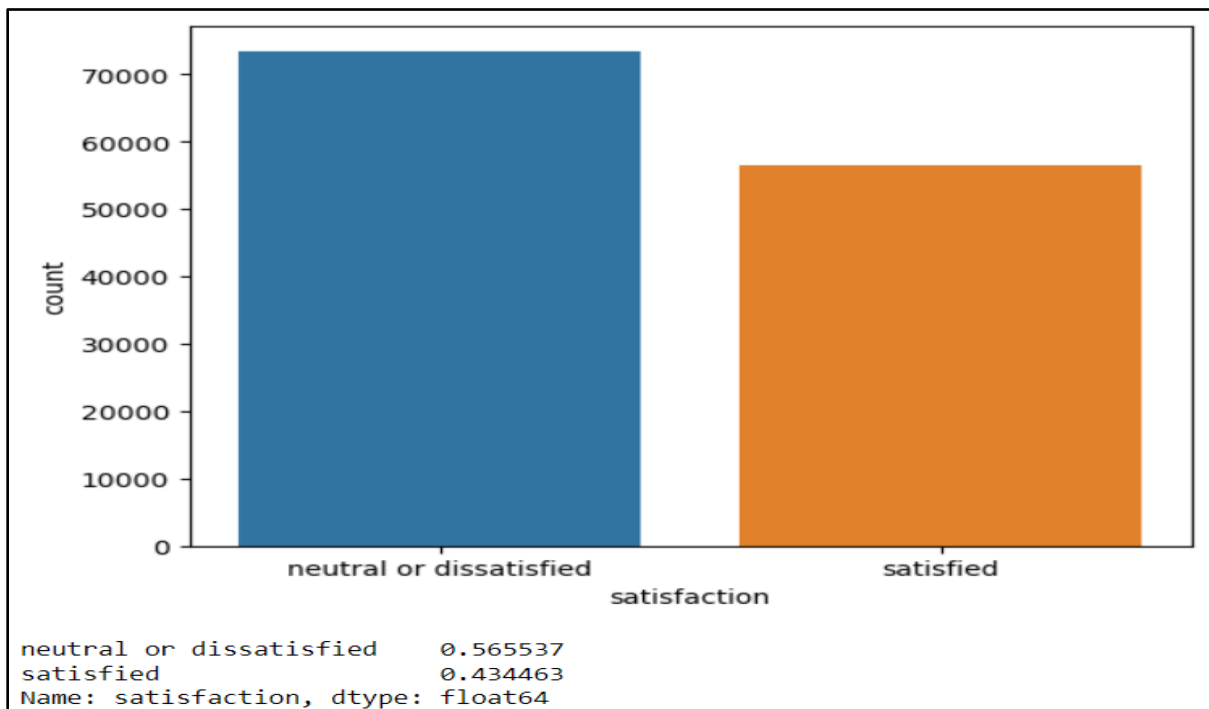


Fig 2.33 Class distribution of target variable: “satisfaction”

2.5. Build models

2.5.1) Building base model using logistic regression:

Logistic Regression is a classification algorithm that is used to predict the probability of a target variable which is categorical. Here, the target variable is a dichotomous variable that contains data coded as 1 (desired outcome like success) or 0 (Example: failure).

We have two popular options for building a logistic regression model; they are scikit-learn and StatsModels.

Good thing about statsmodels is the summary output it produces.

6.1 Building Logistic Regression from StatsModels

Steps involved:

Step 1: Define the data frame(X) with all the selected independent feature

Define the Data Frame(Y) with the target Satisfaction

Step 2: Perform a Train Test split considering X and Y with a test fraction of 0.3

and a randomstate of 12345 (for easy reference) and unpack the function to X_train,X_test,Y_train,Y_test.

Step 3: Fit a logistic regression model from StatsModels using the train data

Step 4: Print the model summary

Step 5: Report Pseudo R-square (McFadden R square), model coefficients and p-value

Step 6: List the significant variables at 5% level of significance

Step 7: Get odds ratio

Step 8: Interpret Odds Ratio for variables having odds ratio > 1

Step 9: Get the Classification report and infer from the same

```
X_      = X_noncollinear
Y_      = df['satisfaction']

X_train, X_test, Y_train, Y_test = split_train_test(X_, Y_, 0.3, 12345)

vals, counts = np.unique(Y_test, return_counts = True)
print(vals, counts)

[0. 1.] [22036 16928]

print("\nShape: Total observations %d Total features %d" %(X_train.shape[0], X_train.shape[1]))

Shape: Total observations 90916 Total features 6

vals, counts = np.unique(Y_train, return_counts = True)
print(vals, counts)

[0. 1.] [51416 39500]

logit = sm.Logit( Y_train, sm.add_constant( X_train ) )
lg     = logit.fit()

Optimization terminated successfully.
      Current function value: 0.447847
      Iterations 6
```

Fig 2.34 Building base model: Step 1 to Step 3

```
print(lg.summary())
```

Logit Regression Results						
=====						
Dep. Variable:	satisfaction	No. Observations:	90916			
Model:	Logit	Df Residuals:	90909			
Method:	MLE	Df Model:	6			
Date:	Wed, 06 Dec 2023	Pseudo R-squ.:	0.3458			
Time:	02:09:54	Log-Likelihood:	-40716.			
converged:	True	LL-Null:	-62235.			
Covariance Type:	nonrobust	LLR p-value:	0.000			
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	-0.9035	0.031	-29.408	0.000	-0.964	-0.843
Gender	0.0132	0.018	0.754	0.451	-0.021	0.048
Customer Type	-2.1139	0.023	-90.574	0.000	-2.160	-2.068
Type of Travel	-2.7643	0.026	-106.619	0.000	-2.815	-2.713
Class	-0.7478	0.016	-48.004	0.000	-0.778	-0.717
Ease of Online booking	0.1786	0.006	28.847	0.000	0.167	0.191
Baggage handling	0.6120	0.008	74.503	0.000	0.596	0.628
=====						

Fig 2.35 Building base model: Model Summary

Observation

We observe that the McFadden R square (Pseudo R square) is 34.58 % and the model fitness is good. This McFadden approach is one minus the ratio of two log likelihoods. The numerator is the log likelihood of the logit model selected and the denominator is the log likelihood if the model just had an intercept.

A goodness of fit using McFadden's pseudo r square (ρ^2) is used for fitting the overall model. McFadden suggested ρ^2 values of between 0.2 and 0.4 should be taken to represent a very good fit of the model (Louviere et al.,2000).

http://www.lifesciencesite.com/lcj/life1002/286_B01288life1002_2028_2036.pdf

List the significant variables at 5% level of significance

```
significant_vars_ = get_significant_vars( lg, sig_level = 0.05 )
significant_vars = significant_vars_.remove('const')
print(significant_vars_)
```

```
['Customer Type', 'Type of Travel', 'Class', 'Ease of Online booking', 'Baggage handling']
```

Observation

- The following 5 variables are significant at 5 % level of significance:

SNo	Significant variable	
1	Customer Type	The customer type (Loyal customer, disloyal customer)
2	Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
3	Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
4	Ease of Online booking	atisfaction level of online booking
5	Cleanliness	Satisfaction level of Cleanliness

Fig 2.36 Building base model: List the significant variables at 5% level of significance

Get Odds ratio

```
print(np.exp(lg.params))
print(type(np.exp(lg.params)))
```

```
const          0.405136
Gender          1.013308
Customer Type   0.120761
Type of Travel  0.063022
Class           0.473392
Ease of Online booking  1.195602
Baggage handling 1.844050
dtype: float64
<class 'pandas.core.series.Series'>
```

```
np.round(np.exp(lg.params),2).reindex(significant_vars).sort_values(ascending = False)
```

```
Baggage handling 1.84
Ease of Online booking 1.20
Gender           1.01
Class            0.47
const            0.41
Customer Type    0.12
Type of Travel   0.06
dtype: float64
```

Fig 2.37 Building base model: Getting Odds ratio of independent features

Odds Ratio Interpretation for variables having odd's ratio > 1

Holding other things constant:

SNo	Inference
1	For a customer, one level increase in satisfaction level of Baggage Handling, the odds of customer satisfaction increases by 84%.
2	For a customer, one level increase in satisfaction level of Ease of Online booking, the odds of customer satisfaction increases by 20%.
3	For a customer, it suggests that the odds of the outcome are higher in one gender compared to the other. the odds of customer satisfaction increases by 1%.

Fig 2.38 Building base model: Interpretation of Odds Ratio for variables having odds ratio > 1

```
"""
Function Name: logit_reg

Description: This **function** builds the logistic regression model.

Input: 1) splits for k fold
       2) random seed number
       3) Training data for predictor variables
       4) Testin  data for predictor variables
       5) Training data for target variable
       6) Testing data for target variable

Output: 1) AUROC 2) Metrics - Precision, Recall, F1
"""

def logit_reg(n_splits, random_state, X_train, X_test, Y_train, Y_test ):

    import statsmodels.api          as      sm

    from sklearn.linear_model      import LogisticRegression
    from sklearn.metrics           import classification_report
    from sklearn.metrics           import confusion_matrix
    from sklearn.model_selection   import cross_val_score

    from sklearn.model_selection   import KFold

    model = LogisticRegression(max_iter = 3000)

    model.fit(X_train, Y_train)
    predicted_train = model.predict(X_train)
    matrix         = confusion_matrix(Y_train, predicted_train)
    print("\nTraining Data")
    print(matrix)
    draw_cm(Y_train, predicted_train )

    accuracy_train = model.score(X_train, Y_train)
    print("Training Accuracy: %.3f%%" % (accuracy_train * 100.0))

    print("\nTesting Data")

    predicted_testing = model.predict(X_test)
    matrix            = confusion_matrix(Y_test, predicted_testing)
    print(matrix)
    draw_cm(Y_test, predicted_testing)

    accuracy_test    = model.score(X_test, Y_test)
    print("Test Accuracy: %.3f%%" % (accuracy_test * 100.0))

    measures_train   = classification_report(Y_train, predicted_train)
    print("\nTraining data")
    print(measures_train)

    measures_test    = classification_report(Y_test, predicted_testing)
    print("\nTesting data")
    print(measures_test)
```

Fig 2.39 Function used to give a detailed output of the base model



Fig 2.40 Confusion Matrix, Accuracy of Train data

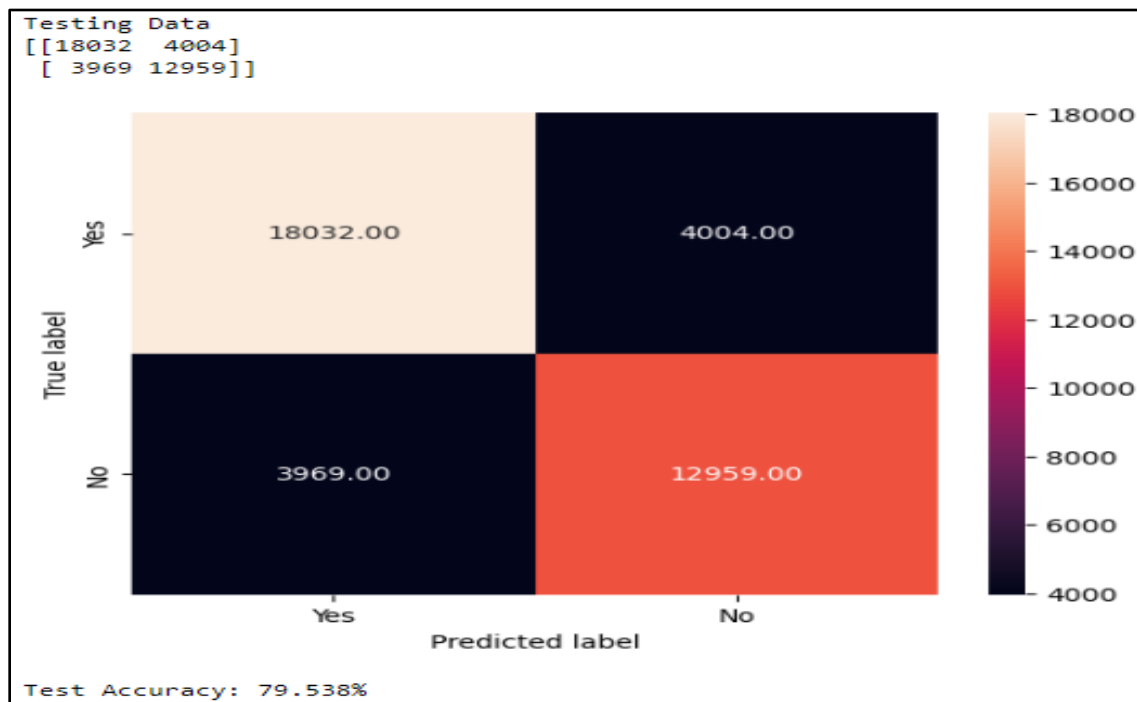


Fig 2.41 Confusion Matrix, Accuracy of Test data

Training data				
	precision	recall	f1-score	support
0.0	0.82	0.82	0.82	51416
1.0	0.76	0.76	0.76	39500
accuracy			0.79	90916
macro avg	0.79	0.79	0.79	90916
weighted avg	0.79	0.79	0.79	90916
Testing data				
	precision	recall	f1-score	support
0.0	0.82	0.82	0.82	22036
1.0	0.76	0.77	0.76	16928
accuracy			0.80	38964
macro avg	0.79	0.79	0.79	38964
weighted avg	0.80	0.80	0.80	38964

Fig 2.42 Classification report of training and test data

Inference from base model:

#	Training data -- Weighted F1 score	Test data -- Weighted F1 score
1	0.79	0.80

Table 2.1 Results from base model

The **weighted average F1 score** is a special case where we report not only the score of the positive class, but also the negative class. This is important where we have imbalanced classes. Because the simple F1 score gives a good value even if our model predicts positives all the time.

Weighted average F1 score for the training dataset is **0.79** and for the test dataset is **0.80**

2.5.2) Comparison chart of all the models built

We shall build all the models as listed before on both without and with SMOTE data and compare both the training and testing performance

2.5.2 a) Before model parameter tuning:

Performance measure: **Weighted average F1 score**

			Before applying SMOTE on training data		After applying SMOTE on training data	
Sl.No	Model name	Parameters used	Measure for training dataset	Measure for test dataset	Measure for training dataset	Measure for test dataset
1	LOGISTIC REGRESSION	'c':1.0, 'class_weight':none, 'dual' : False , 'fit_intercept':True, 'intercept_scaling' : 1, 'l1_ratio' : None, 'max_iter' : 100 , 'multi_class': 'auto' , 'n_jobs': None , 'penalty': 'l2' , 'random_state' : None , 'solver' : 'lbfgs' , 'tol' : 0.0001, 'verbose' : 0, 'warm_start': False	0.873	0.873	0.783	0.783
2	KNEIGHBOUR CLASSIFIER	algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': None, 'n_neighbors': 5, 'p': 2, 'weights': 'uniform'	0.954	0.934	0.961	0.933
3	GAUSSIAN NAÏVE BAYES	'priors' : None , 'var_smoothing' : 1e-09	0.875	0.875	0.871	0.874
4	DECISION TREE	'ccp_alpha' : 0.0 , 'class_weight' : None , 'criterion' : 'gini' , 'max_depth' : None , 'max_features' : None , 'max_leaf_nodes' : None , 'min_impurity_decrease' : 0.0 , 'min_samples_leaf' : 1 , 'min_samples_split' : 2 , 'min_weight_fraction_leaf' : 0.0 , 'random_state' : None , 'splitter' : 'best'	0.997	0.944	0.997	0.945
5	RANDOM FOREST	'bootstrap': True 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None , 'max_features' : None, 'max_leaf_nodes' : None , 'min_impurity_decrease' : 0.0 , 'min_samples_leaf': 1 , 'min_samples_split' : 2 , 'min_weight_fraction_leaf' : 0.0 , 'n_estimator' : 100 , 'n_jobs' : None ,	0.997	0.958	0.997	0.958

		'oob_score' : False , 'random_state' : None , 'verbose' : 0, 'warm_start' : False				
6	ADABOOST CLASSIFIER	'algorithm': 'SAMME.R', 'base_estimator': 'deprecated', 'estimator__ccp_alpha': 0.0, 'estimator__class_weight': None, 'estimator__criterion': 'gini', 'estimator__max_depth': None, 'estimator__max_features': None, , 'estimator__random_state': None, 'estimator__splitter': 'best', 'estimator': DecisionTreeClassifier(), 'learning_rate': 1.0, 'n_estimators': 50, 'random_state': 10	0.997	0.950	0.997	0.947
7	GRADIENT BOOST CLASSIFIER	'ccp_alpha': 0.0, 'criterion': 'friedman_mse', 'init': None, 'learning_rate': 0.1, 'loss': 'log_loss', 'max_depth': 3, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_iter_no_change': None, 'random_state': None, 'subsample': 1.0, 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': , 'warm_start': False	0.942	0.958	0.945	0.958
8	XGBOOST CLASSIFIER	'objective': 'binary:logistic', 'base_score': None, 'booster': None, 'callbacks': None, 'colsample_bylevel': None, 'colsample_bynode': None, 'colsample_bytree': None, 'device': None, 'enable_categorical': False, 'eval_metric': None, 'feature_types': None, 'gamma': None, 'grow_policy': None, 'importance_type': None, 'interaction_constraints': None, 'learning_rate': None, 'max_bin': None, 'max_cat_threshold': None, 'max_cat_to_onehot': None, 'max_delta_step': None, 'max_depth': None, 'max_leaves': None, 'min_child_weight': None, 'missing': nan, 'monotone_constraints': None, 'multi_strategy': None, 'n_estimators': None, 'n_jobs': None, 'num_parallel_tree': None, 'random_state': None, 'reg_alpha': None, 'reg_lambda': None, 'sampling_method': None, 'scale_pos_weight': None, 'subsample': None, 'tree_method': None, 'validate_parameters': None, 'verbosity': None	0.969	0.961	0.971	0.96

Table 2.2: Model Evaluation based on default parameters

2.5.2 b) After model parameter tuning

We shall tune the model with certain selected parameter with their respective ranges

Performance measure: **Weighted average F1 score**

			Before applying SMOTE on training data		After applying SMOTE on training data	
Sl.no	Model name	Parameters range and best parameters	Measure for training dataset	Measure for test dataset	Measure for training dataset	Measure for test dataset
1	LOGISTIC REGRESSION	Parameter range: 'penalty': ['l1', 'l2'], 'C': [0.001, 0.01 , 0.1, 1, 10, 100], 'solver': ['liblinear', 'saga'], 'max_iter': [100 , 200, 300, 500] Best params (BS) : 'C' : 0.01, 'max_iter': 100, 'penalty': 'l1', 'solver': 'saga' Best params (AS) : 'C' : 0.001, 'max_iter': 100, 'penalty': 'l1', 'solver': 'saga'	0.873	0.874	0.789	0.789
2	KNEIGHBOR CLASSIFIER	Parameter range: 'n_neighbors': [3, 5 , 7, 9, 11, 13], 'weights': ['uniform' , 'distance'], 'p': [1 , 2] Best params (BS) : 'n_neighbors': 7, 'p': 1, 'weights': 'uniform'	0.953	0.94	0.997	0.938

		Best params (AS) : 'n_neighbors': 5, 'p': 1, 'weights': 'distance'				
3	GAUSSIAN NAÏVE BAYES	<p>Parameter range: 'priors': [None, [0.3, 0.7], [0.5, 0.5]], 'var_smoothing': [1e-10, 1e-9, 1e-8]</p> <p>Best params (BS): 'priors': None, 'var_smoothing': 1e-10</p> <p>Best params (AS) : 'priors': None, 'var_smoothing': 1e-10</p>	0.876	0.875	0.871	0.874
4	DECISION TREE	<p>Parameter range: 'max_depth': [None, 5, 10, 15], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': [None, 'sqrt', 'log2', 0.5]</p> <p>Best params (BS) : 'max_depth': 15, 'max_features': 0.5, 'min_samples_leaf': 2, 'min_samples_split': 10</p> <p>Best params (AS): 'max_depth': 15, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 5</p>	0.957	0.947	0.966	0.952
5	RANDOM FOREST	<p>Parameter range: 'n_estimators': [50, 100, 200], 'max_depth': [None, 5, 10, 15],</p>				

		<p>'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4], 'max_features': ['auto', 'sqrt', 'log2', 0.5]</p> <p>Best params (BS) : 'max_depth': None, 'max_features': 0.5, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200</p> <p>Best params(AS): 'max_depth': None, 'max_features': 0.5, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 200</p>	0.977	0.961	0.981	0.960
6	ADABOOST CLASSIFIER	<p>Parameter range : 'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.5, 1.0], 'algorithm': ['SAMME', 'SAMME.R']</p> <p>Best params(BS) : 'algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 200</p> <p>Best params (AS) : 'algorithm': 'SAMME', 'learning_rate': 0.5, 'n_estimators': 200</p>	0.997	0.953	0.997	0.953
7	GRADIENT BOOST CLASSIFIER	<p>Parameter range: 'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 0.5], 'max_depth': [3, 4, 5], 'min_samples_split': [2, 5, 10], 'min_samples_leaf': [1, 2, 4]</p>	0.968	0.959	0.971	0.959

		<p>Best params (BS): 'learning_rate': 0.5, 'max_depth': 4, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200</p> <p>Best params (AS): 'learning_rate': 0.5, 'max_depth': 4, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 200</p>				
8	XG BOOST CLASSIFIER	<p>Parameter range: 'learning_rate': [None, 0.01, 0.1, 0.2], 'n_estimators': [None, 50, 100, 200], 'max_depth': [None, 3, 4, 5], 'subsample': [None, 0.8, 1.0], 'colsample_bytree': [None, 0.8, 1.0]</p> <p>Best params (BS): 'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': None, 'n_estimators': 200, 'subsample': 0.8</p> <p>Best params (AS): 'colsample_bytree': 0.8, 'learning_rate': 0.2, 'max_depth': None, 'n_estimators': 200, 'subsample': None</p>	0.967	0.961	0.973	0.961

Table 2.3 Model evaluation after hyperparameter tuning.

Inference:

- XGBoost Classifier has shown the best performance both on with and without applying SMOTE
- Weighted average F1 Score was found to 0.961 both on with and without applying SMOTE
- The model has a good fit

2.6) Model evaluation:

a) Performance metric: Weighted average F1 Score

		Before applying SMOTE on training data		After applying SMOTE on training data	
Sl no	Model name	Measure for training dataset	Measure for test dataset	Measure for training dataset	Measure for test dataset
1	XGBoost Classifier	0.967	0.961	0.973	0.961

Table 2.4 Performance of the best model

b) The best model is XGBoost Classifier

b) Performed well on both before and after applying SMOTE (weighted F1 score = 0.961)

d) best parameters:

Parameters tuned	Before applying SMOTE	After applying SMOTE
colsample_bytree	0.8	0.8
learning_rate	0.1	0.2
max_depth	None	None
n_estimators	200	200
Subsample	0.8	None

Table 2.5 Best parameters of the best model

e) Boosting has helped the independent features to explain the target ‘satisfaction’ very well.

f) The model has handled the binary classification very well and hence very reliable for other similar datasets.

2.7) Comparison to the bench mark model:

We shall compare the benchmark model i.e., logistic regression with the best model XGBoost Classifier.

Performance measure: **weighted average F1 Score**

Model	Measure for training dataset	Measure for test dataset
LOGISTIC REGRESSION	0.79	0.79
XGBOOST CLASSIFIER	0.967	0.961

Table 2.6 Comparison of best model with benchmark model

The model has improved **significantly** from the benchmark model.

Both the performance in training as well as performance in testing has significantly increased

This implies non parametric model(ensemble) is better suited for predictive analysis of Airline customer satisfaction dataset.

2.8) Insights from the best model and EDA

2.8.1) Visualisations.

a) Before SMOTE:

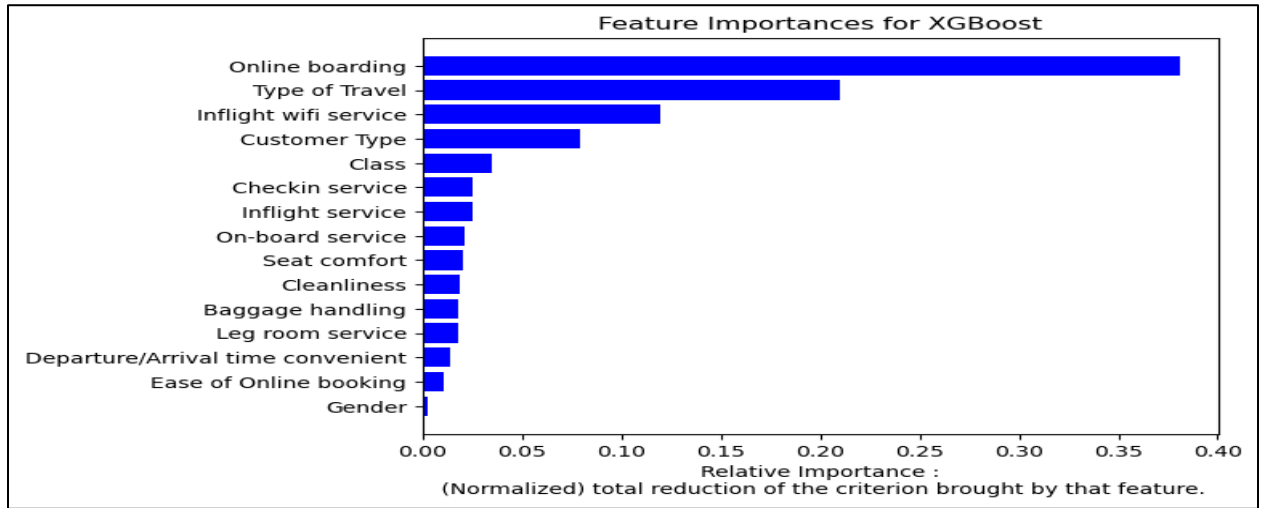


Fig 2.43 Variable importance plot from XGBoost Classifier before applying SMOTE

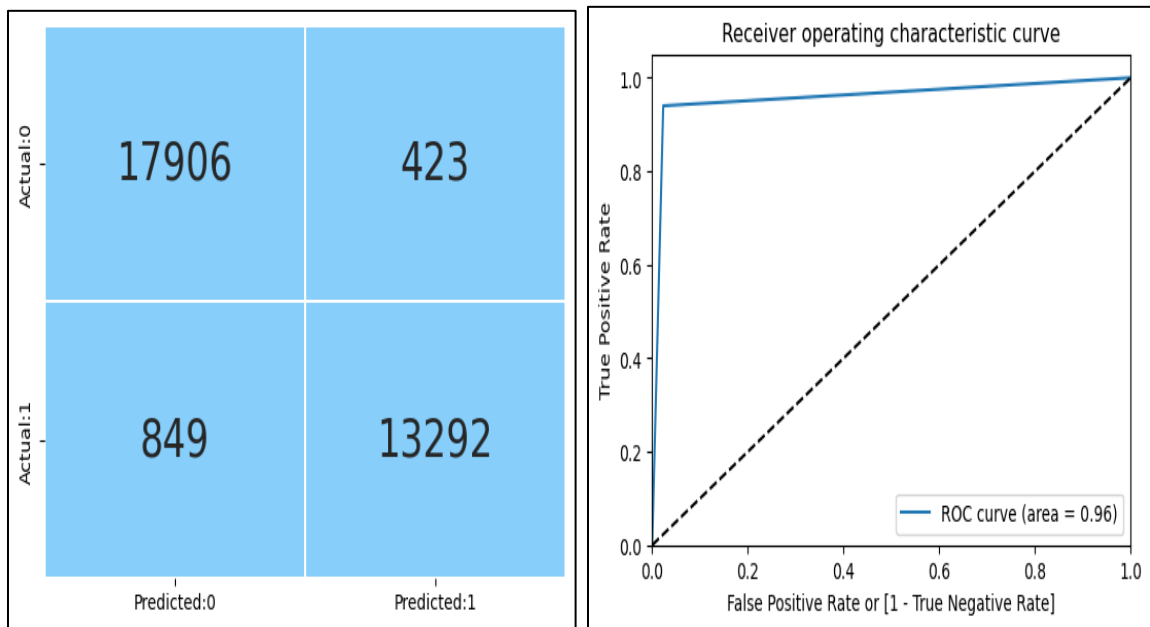


Fig 2.44 Confusion matrix and ROC curve for XGBoost Classifier before applying SMOTE.

b) After SMOTE:

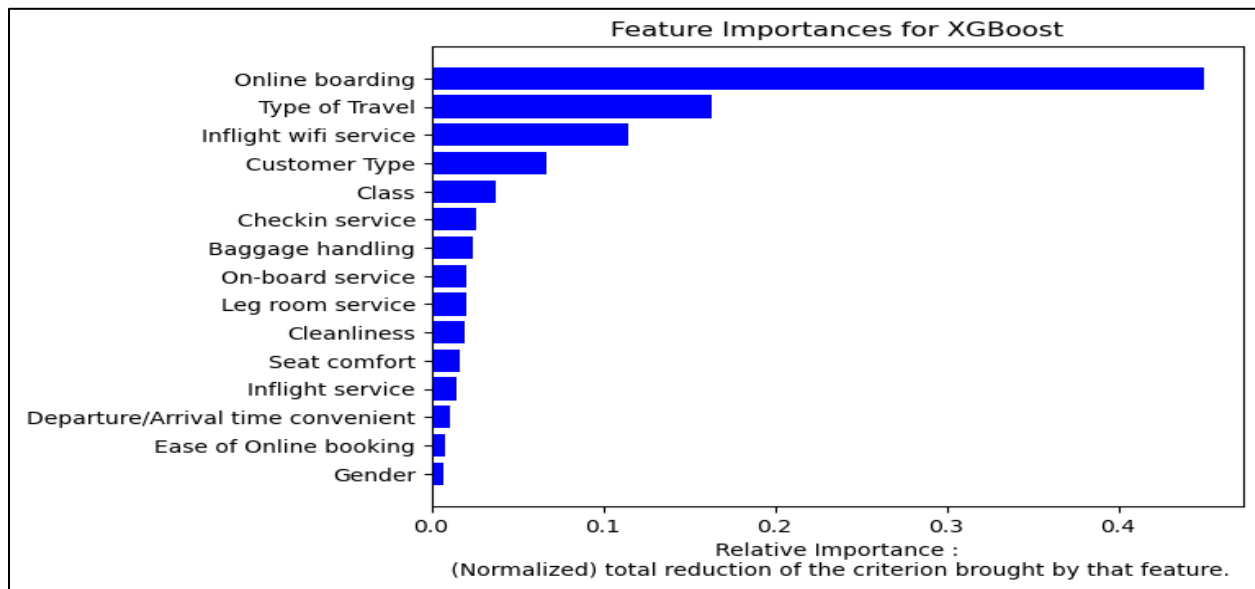


Fig 2.45 Variable importance plot from XGBoost Classifier after applying SMOTE

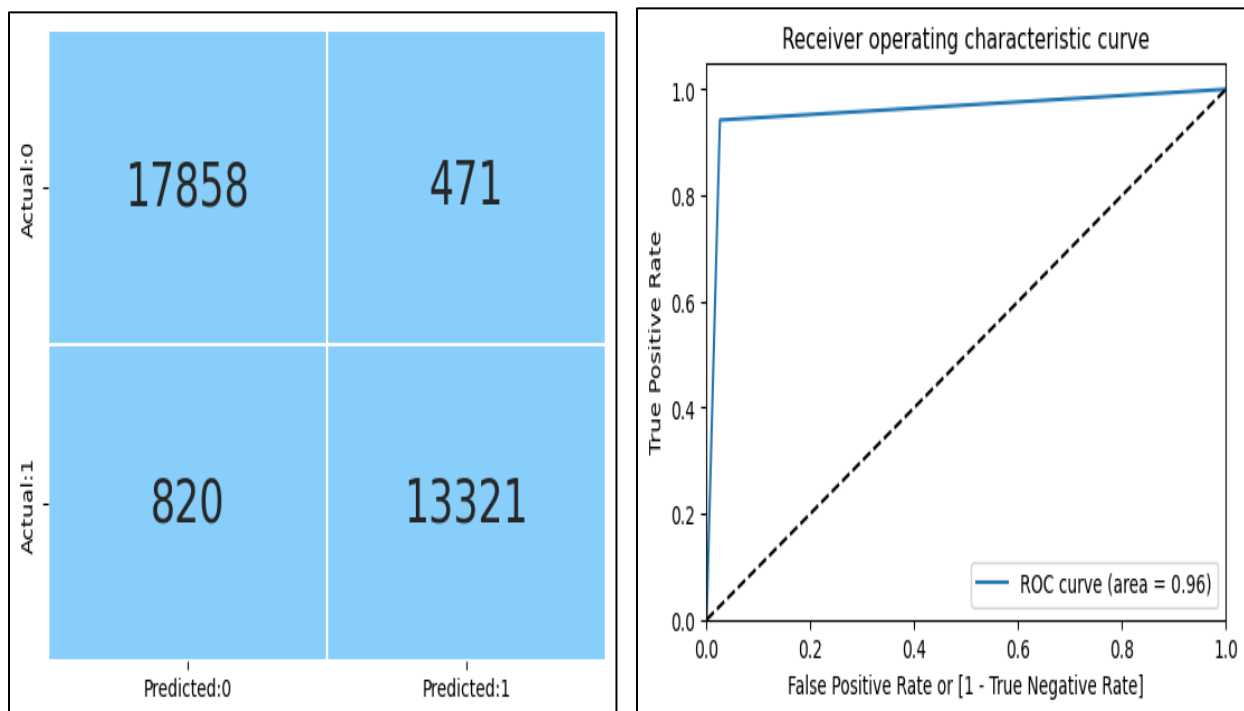


Fig 2.46 Confusion matrix and ROC curve for XGBoost Classifier after applying SMOTE

2.8.2) Interpretations:

- XGBoost model has shown the best performance both before and after applying SMOTE.
- The four most important independent features were found to be 'Online Boarding', 'type of travel', 'Inflight Wi-Fi service' and 'customer type'.
- Area under ROC is very good (0.96)
- The hyper parameter tuning though not significant has helped the model perform a little better (0.96 to 0.961)
- The model has a good fit
- Hence, we can consider that the model performs well on both with and without applying SMOTE

2.8.3)

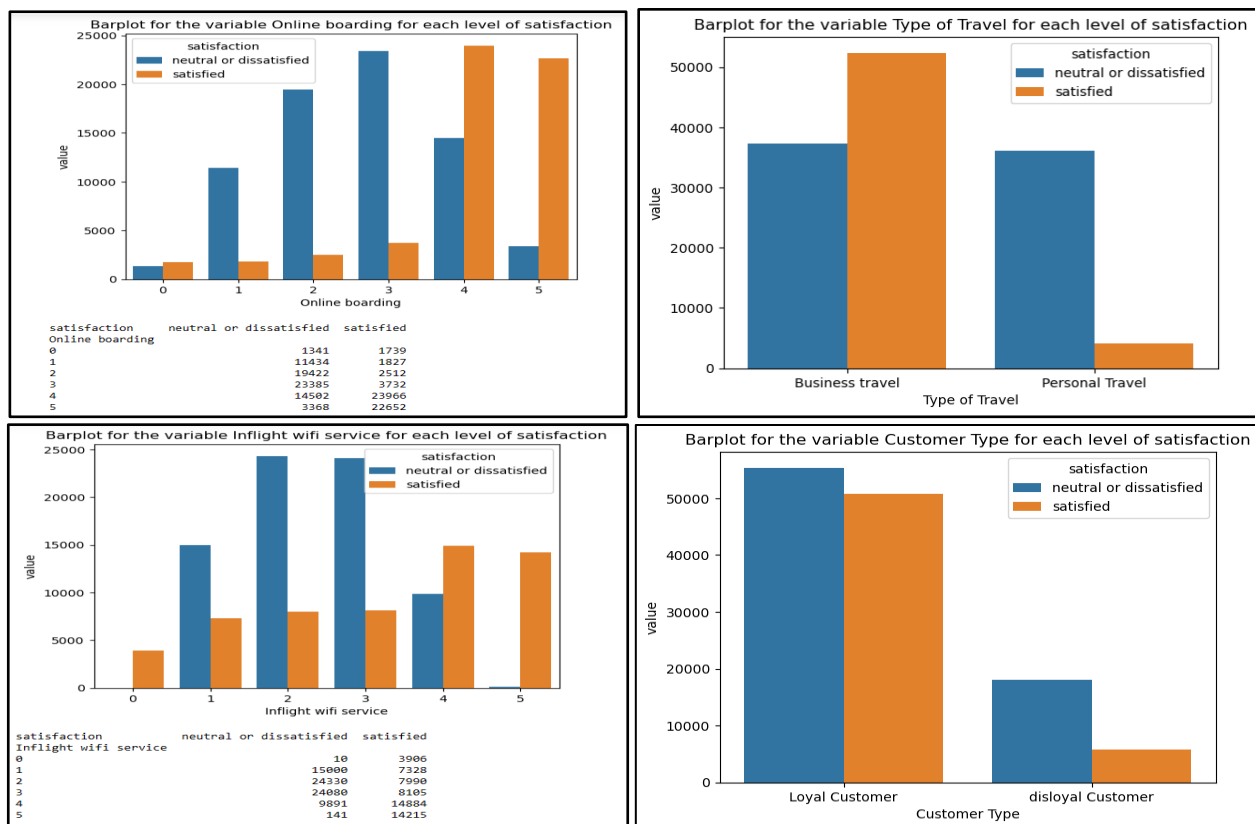


Fig 2.47 EDA done at the onset comparing most important features with target

Results of Chi-square tests of independence		
	Column	chi2_value
0	Gender	16.352081
0	Customer Type	4493.188803
0	Type of Travel	26282.520993
0	Class	32906.171859
0	Inflight wifi service	35891.433370
0	Departure/Arrival time convenient	601.462958
0	Ease of Online booking	12846.702395
0	Gate location	3069.907992
0	Food and drink	6571.202895
0	Online boarding	49531.218396
0	Seat comfort	19538.740376
0	Inflight entertainment	23071.602435
0	On-board service	14342.659690
0	Leg room service	15200.778756
0	Baggage handling	10820.213523
0	Checkin service	8143.773216
0	Inflight service	10357.930495
0	Cleanliness	12948.918125
0		

Fig 2.48 Chi square test results done at the onset comparing most important features with target

Inference: The results of feature importance from the XGBoost model is consistent with EDA and chi-square test done at the onset.

CHAPTERS 3 – Suggestions, recommendations to the stakeholder(s):

Given that online booking, type of travel, inflight WiFi service, and type of customer are identified as significant features for predicting airline passenger satisfaction, here are some suggestions for the business:

3.1. Enhance Online Booking Experience:

- Invest in improving the user interface and functionality of the online booking platform.
- Ensure a seamless and user-friendly booking process with clear navigation and minimal steps.
- Implement features such as real-time seat selection, fare transparency, and personalized recommendations to enhance the booking experience.

3.2. Tailor Services Based on Type of Travel:

- Recognize and differentiate between various types of travel, such as business and leisure.
- Customize services to meet the specific needs of each travel type. For example, business travellers might prioritize efficiency and productivity, while leisure travelers may seek more comfort and entertainment.

3.3. Optimize Inflight WiFi Service:

- Improve the reliability and speed of inflight WiFi to meet passengers' expectations.
- Consider offering complimentary or discounted WiFi for certain customer segments or travel classes.
- Promote the availability of inflight WiFi during the booking process to set expectations and attract passengers who prioritize connectivity.

3.4. Segmentation Based on Type of Customer:

- Understand the distinct preferences and expectations of different customer segments.
- Create targeted marketing campaigns and promotions based on customer types (e.g., frequent flyers, first-time travelers).
- Tailor communication strategies to address the unique needs of each segment, enhancing overall customer satisfaction.

3.5. Gather Customer Feedback:

- Implement regular surveys or feedback mechanisms to understand passenger satisfaction in more detail.
- Encourage customers to provide feedback on their experiences with online booking, inflight WiFi, and overall travel satisfaction.
- Use feedback to identify specific pain points and areas for improvement.

3.6. Continuous Monitoring and Adaptation:

- Establish a system for continuous monitoring of passenger satisfaction metrics.
- Stay updated on industry trends and technological advancements to remain competitive in providing services aligned with passenger expectations.
- Be agile in adapting strategies based on changing customer preferences and market dynamics.

3.7. Employee Training and Engagement:

- Ensure that airline staff, both online and onboard, are well-trained and customer-focused.
- Empower employees to address customer concerns promptly and effectively.

- Recognize and reward employees for exceptional service to motivate a positive customer-centric culture.

3.8. Promote Positive Experiences:

- Leverage positive customer testimonials and experiences in marketing campaigns.
- Highlight features such as excellent online booking, reliable inflight WiFi, and personalized services in promotional materials to attract new customers.

By focusing on these suggestions, the airline business can enhance customer satisfaction, build loyalty, and differentiate itself in a competitive market. Regularly assessing and adapting strategies based on customer feedback and industry trends will contribute to sustained success.

CHAPTER 4 – LIMITATIONS

The limitations of the project are as follows:

4.1 Sample Size:

the study's findings are based on a specific dataset with a limited sample size (129,880 rows). the generalization of results to a broader population may be influenced by the dataset's representativeness.

4.2 Data Quality:

The Accuracy and reliability of the study depend on the quality of the dataset. While efforts were made to clean and preprocess the data, inherent errors or biases in the original data source may impact the results

4.3 Feature selection:

The variables chosen for analysis were based on exploratory data analysis (eda) and recursive feature elimination (RFE). Alternative feature selection methods or additional variables not considered may impact the model's predictive power.

4.4 model selection:

The choice of machine learning models is subjective and may be influenced by various factors. different models may yield different results, and the selected models might not capture all intricacies of the underlying patterns.

4.5 Changing Dynamics:

External factors, such as changes in the aviation industry, economic conditions, or global events, may influence customer satisfaction dynamics. the study does not account for real-time changes beyond the dataset's timeframe.

CHAPTER 5 – CLOSING REFLECTIONS

5.1 Reflections/learnings from the project

a) Practical application of concepts:

Moving from theoretical concepts to practical implementation reinforces a deeper understanding of machine learning principles. The hands-on experience allows for a more intuitive grasp of algorithms, data preprocessing, and model evaluation.

b) Data challenges:

Dealing with real-world data presents challenges not always captured in academic exercises. Noisy, incomplete, or unstructured data requires thoughtful preprocessing and cleaning, and understanding the domain is crucial for effective data handling.

c) Model selection and tuning:

Choosing the right model and optimizing its parameters was more tough than expected. It involved experimentation, iteration, and a keen awareness of the specific problem at hand. It emphasizes the need to tailor models to the characteristics of the dataset.

d) Interpreting and communicating results:

The importance of not only building accurate models but also interpreting and communicating results became evident. Clear visualization, explanation of model predictions, and actionable insights were vital for stakeholders' understanding and decision-making.

e) Business context awareness:

Integrating Machine Learning into a real project underscored the significance of understanding the broader business context. Aligning Machine Learning goals with business objectives ensured that the work has tangible impact and value.

5.2 Improvisations:

a) Early Stakeholder Engagement:

Engaging with stakeholders early in the project could help in setting clear expectations, understanding business requirements, and refining the project scope accordingly.

b) Continuous Validation with Domain Experts:

Regular validation with domain experts can provide valuable insights and ensure that the models align with the intricacies of the industry or domain. It helps in refining assumptions and enhancing model relevance.

c) Exploration of Advanced Techniques:

Depending on the project, exploring more advanced techniques or deep learning methods might be considered for improved model performance. This includes staying updated on the latest developments in the machine learning field.

d) Ethical Considerations:

Considering ethical implications of the models, such as bias and fairness, is an integral part of responsible machine learning. This might involve implementing fairness-aware algorithms and continuous monitoring for biases.

e) Robust Documentation and Version Control:

Ensuring robust documentation and version control throughout the project facilitates collaboration, knowledge sharing, and reproducibility. It becomes crucial for the long-term maintenance and scalability of the solution.

f) User Feedback Integration:

If applicable, integrating user feedback into the model improvement process can be invaluable. Real-world users often provide unique perspectives that may not be entirely captured in the initial project phase.

CHAPTER 6 - BIBLIOGRAPHY

- 1) <https://en.wikipedia.org/wiki/SERVQUAL>
- 2) <https://link.springer.com/article/10.1007/s11628-009-0068-4> (Myungsook An)
- 3) <https://theses.whiterose.ac.uk/3647/1/489179.pdf> (Fahed Salim Khatib)
- 4) <https://www.numpyninja.com/post/mice-and-knn-missing-value-imputations-throughpython>
- 5) <https://www.scribbr.com/frequently-asked-questions/when-to-remove-anoutlier/#:~:text=Some%20outliers%20represent%20natural%20variations,processing%20errors%2C%20or%20poor%20sampling.>
- 6) <https://statisticsbyjim.com/basics/remove-outliers/>
- 7) <https://www.scribbr.com/frequently-asked-questions/when-to-remove-anoutlier/#:~:text=Some%20outliers%20represent%20natural%20variations,processing%20errors%2C%20or%20poor%20sampling.>
- 8) <https://statisticsbyjim.com/basics/remove-outliers/>
- 9) <https://pythonguides.com/python-scipy-stats-skew/>
- 10) <https://www.javatpoint.com/feature-selection-techniques-in-machine-learning>
- 11) <https://towardsdatascience.com/effective-feature-selection-recursive-featureelimination-using-r-148ff998e4f7>
- 12) <https://machinelearningmastery.com/rfe-feature-selection-in-python/>
- 13) https://www.researchgate.net/publication/215666083_Stated_choice_methods_analysi_s_and_application (Louviere et al.,2000)

CHAPTER 7 - ANNEXURE

1) Dictionary Capstone Master file:

<https://drive.google.com/drive/folders/1mw3ZPoY73ZKYVLFA1NfiNWUBX7tjNeDT?usp=sharing>

2) EDA file capstone:

https://drive.google.com/file/d/10NnGAfcCBMo18WYeohIY6wwfX0L8Vlgl/view?usp=drive_link

3) Base model building:

https://drive.google.com/file/d/10FNLS2I_5FIJ8z1Q2h5hpk8ZXwAcOm98/view?usp=drive_link

4) Final datasets notebook:

https://drive.google.com/file/d/1pttXZxxHCsapqCQLCzQYjjvvE3tZEGAJ/view?usp=drive_link

5) Without SMOTE complete file:

https://drive.google.com/file/d/1B8I0TTcx1Ovgxb1DpQ_sf4fsSao2ehc/view?usp=drive_link

6) Dictionary With SMOTE complete file:

https://drive.google.com/file/d/1hcBMbTSBAPh4LvR-C9fC7DLcDFAfFO6K/view?usp=drive_link

1. CHAPTER 8 - DATA DICTIONARY

No .	Variable Name	Types and Values	Description
1	id	Numeric	Identifier which uniquely identifies the record
2	Gender	Categorical with two levels: Female, Male	Gender of the airline passengers
3	Customer Type	Categorical with two levels: Loyal customer, disloyal customer	The customer type
4	Age	Numeric	Actual age of the passengers
5	Type of Travel	Categorical with levels (Personal Travel, Business Travel)	Purpose of the flight of the passengers
6	Class	Categorical with the levels Business, Eco, Eco Plus	Travel class in the plane of the passengers
7	Flight distance	Numeric	The flight distance of this journey
8	Inflight wifi service	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of the inflight wifi service
9	Departure/Arrival time convenient	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Departure/Arrival time convenient
10	Ease of Online booking	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the	Satisfaction level of online booking

		level of satisfaction of the airline passengers	
11	Gate location	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Gate location
12	Food and drink	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Food and drink
13	Online boarding	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of online boarding
14	Seat comfort	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Seat comfort
15	Inflight entertainment	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of inflight entertainment
16	On-board service	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of On-board service
17	Leg room service	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Leg room service

18	Baggage handling	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of baggage handling
19	Check-in service	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Check-in service
20	Inflight service	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of inflight service
21	Cleanliness	Ordinal variable with six levels Satisfaction Level:0 – Not applicable, 1 to 5 representing the level of satisfaction of the airline passengers	Satisfaction level of Cleanliness
22	Departure Delay in Minutes	Numeric	Minutes delayed when departure
23	Arrival Delay in Minutes	Numeric	Minutes delayed when Arrival
24	Satisfaction	Categorical with two levels 1. Satisfied, 2. Neutral or dissatisfied	Passenger satisfaction level with the airline – Target variable

-----**THE END**-----