CIND 820 Project Report

Name: Nelly Grillo

Student number: 501144764

Supervisor: Ceni Babaoglu, Ph.D.

Date: December 5th, 2022

Ryerson University

Table of Contents

Abstract

The Airline sector has played a unique role within the global transportation system providing economic growth and development in several countries around the world (Heiets & Xie, 2021).

Since the COVID-19 outbreak made its first appearance in China, it is well known that this pandemic has severely impacted in the world's airline sector during the 2020-2021 period (Heiets & Xie, 2021). Regardless worldwide society is trying to recover from this COVID-19 pandemic, it is still ongoing and posing a continuous threat to both: global economy and public health (Olaganathan, 2021). Reports show that a recovery path is expected between 2022 and 2026 in the most pessimistic scenario (Sulu et al., 2021). Therefore, to accelerate the recovery of this business sector, governments around the world are currently lifting the restrictions that were meant to contain the spread of COVID-19 (Sonntag, 2021). At the same time, private sector is managing ways to enhance and expedite the sector revival. New behavioral passenger patterns have been developed and airline companies were in the need to adjust services at a whole different level (Heiets & Xie, 2021).

One main factor to help with the airline sector recovery, is to determine the level of passenger satisfaction (Sulu et al., 2021). The degree of content shown by costumers with airline services defined the passenger satisfaction (Suhartanto, 2014 and Khudhair et al., 2021). It is important for the airline sector to provide adequate and effective strategies to meet their customer requirements (Sulu et al., 2021 and Khudhair et al., 2021) and to encourage positive post-purchase behaviours which promotes the development of customer loyalty (Razafimanjary, 2019).

Literature Review

Airline customer satisfaction and customer loyalty are two important variables that the airline service industry must consider providing adequate and effective strategies to meet their customer requirements, and to promote customer return (Maminiaina, 2019).

Airline customer satisfaction

Considering that an airline passenger is a customer in the airline industry, satisfaction can be seen as a complex level of the customer knowledge and experience when comparing the service performance with their prior expectations. (Noviantoro & Huang, 2022; Jiang & Zhang, 2016)

There are many factors that make customer satisfaction such as: product quality, price, service quality, emotional factor, and cost and convenience (Maminiaina, 2019).

In terms of marketing, the customer satisfaction highly influences the future consumer purchase behavior, profitability, and shareholder value (Suhartanto & Noor, 2012).

Customer loyalty

One important factor for the airline industry to preserve business stability is the customer loyalty.

Customer loyalty is described as customers who return to the same organization because they are very satisfied with the product or service (Maminiaina, 2019)

Airline service quality

Service quality is defined as an indispensable approach to achieve business survival and success because it can persuade customer purchase behaviour and business performance (Suhartanto & Noor, 2012).

To achieve and provide high-quality service, it is necessary for the airline industry to understand the customer needs (Jiang & Zhang, 2016). The airline customer is the one who defines the service quality by comparing their expectations and experiences (Noviantoro & Huang, 2022). When the airlines offer better quality services, it is possible to achieve customer loyalty (Jiang & Zhang, 2016).

The airline service can be seen as a group of services which can be divided on pre-flight services and in-flight services (Gao et al., 2021; Etemad-Sajadi, et al., 2016).). The pre-flight service comprises attributes such as online boarding, gate location, ease of online booking and on-board service. In addition, the in-flight service includes attributes such as seat comfort, in-flight wi-fi service and entertainment, legroom, and cleanliness.

There are several investigations on service quality that explore the problems related to the service quality in the airline industry. Some of these studies that assess the airline customer satisfaction levels uses traditional statistical testing or multiple-criteria methods (Noviantoro & Huang, 2022).

Etemad-Sajadi, et al., (2016) focus on the customer perception of the pre-flight and in-flight service quality. This study shows that there is a positive direct effect between the pre-flight service quality perceived by the airline customer and the airline customer satisfaction. Moreover,

there is a positive direct and indirect effect on the airline customer loyalty which can be improved by enhancing the pre-flight quality service.

Suhartanto and Noor (2012) aims to aims to investigates how service quality and price affect customer satisfaction in low-cost airlines and full-service airlines. By applying SERVPERF, this study concludes that the low-cost airlines use price as competitive strategy, while full-service airlines depend on quality of services. Moreover, this study highlights that both types of airlines should consider as factors: the attitude of employees when delivering the service and the price.

There are other studies that investigate the role of the customer class on the airline customer satisfaction. In this study, An and Noh (2009) concludes that there is a different perception of the service quality according to the customer seat class which can be explain in terms of a difference in income and professional status. Moreover, this study shows that the in-flight service quality impacts the customer loyalty, and the extent of the impact depends on the seat class (An & Noh, 2009). On the other hand, Jiang and Zhang (2016) suggests that, although, there is a positively association between the customer satisfaction and the customer loyalty, this is not greatly reflected among business travellers.

Using the machine learning approach, it is common to study the airline customer satisfaction by applying the sentiment analysis, which analyze tweets to detect positive or negative customer satisfaction (Khan & Urolagin, 2018).

For example, Khan and Urolagin, 2018 analyze tweets using sentiment analysis and provides a general opinion of airline customer satisfaction. Furthermore, Khan & Urolagin, 2018 use three classifiers to perform consumer loyalty prediction.

Kumar and Zymbler (2019) uses sentiment classification approaches such as SVM, ANN, and CNN. Also, this study recommends applying association analysis to provide suggestions to improve the customer satisfaction.

There are other studies that apart from applying sentiment analysis, also use more advanced methods of machine learning approach.

Tan, C. (2021) aims to classify customers satisfaction level by applying multiples machine learning models such as KNN, Logistic Regression, Decision Tree, Random Forest, and several boosting models including Gradient Boosting, Adaboost, Xgboost, LGBM and Catboost. In addition, it applies sentiment analysis on tweets and Bidirectional LSTM model to identify emotional tendency of tweets.

It concludes that the most important factors that significantly affect the airline customer satisfaction includes online boarding, in-flight wi-fi service, in-flight entertainment, and seat comfort.

Nowadays, it is usually to apply machine learning methods for customer analysis and marketing.

Machine learning and deep learning analysis allows to work with complex multi-dimensional data and enables to determine hidden correlations and important insights from the mentioned data (Park et al., 2022).

Noviantoro & Huang (2022) propose a study that uses Data mining method to examine the airline passenger satisfaction focusing on investigating the passenger most desired airline services by using feature selection. This study runs and compares classification algorithms such as decision tree, random forest, gradient boosted tree, k-NN, Naïve Bayes, rule induction,

logistic regression, neural net, deep learning, and support vector machine. This study concludes that the most critical features that the airline should improve to obtain passenger satisfaction are online boarding, inflight wi-fi service, baggage handling, and inflight entertainment.

Park et al. (2022) aim to demonstrates the prediction of customer churn risk and satisfaction by using deep learning models. Furthermore, it compares deep learning model with machine learning models concluding that, in terms of accuracy, CNN-LSTM outperformed the machine learning models.

Research from Gao et al. (2021) uses various machine learning algorithms and interpretation techniques to understand the different factors that impact the airline travel satisfaction. It compares machine learning algorithms and conventional logistic regression methods and determines that machine algorithms such as Multi-layer Perceptron, Random Forest and Support Vector Machine demonstrate superior modeling and predictive performances for modeling airline travel satisfaction. Moreover, it determines the key factors that have greater impacts on airline travel satisfaction such as type of travel, class and customer type, Wi-Fi service, online boarding, and baggage handling.

García et at. (2019) predicts airline customer satisfaction by using a combination of k-NN and ensemble regression model (BAGGING).

Hayadi et al. (2021) applies various classification algorithms such as k-NN, Logistics Regression, Gaussian Naïve, Decision Tree, Random Forest to determine airline customer satisfaction.

This study recommends optimizing the in-flight wi-fi service experience by making it more accessible or by lowering its cost.

Bellizzi et al. (2022) aims to analyse the highly educated people's satisfaction with airline services by applying the Classification and Regression Tree (CART) approach. In addition, this study uses an Importance Performance Analysis (IPA) which permits to identify the most critical factors to be considered by the airline company.

Although there are some studies that applies various machine learning algorithms to determine the level of airline customer satisfaction. This study uses a different airline customer dataset and aims to determine the most important factors that impact the airline customer satisfaction, focusing in establish if customer loyalty and customer class have a particular effect on customer satisfaction. Moreover, this study intends to apply and compare various Classification and Regression algorithms to determine which is the most accurate.

Problem Statement

The purpose of this project is to:

- Determine the most important factors that impact the airline customer satisfaction
- Determine if the customer type, loyal and first-time customer, have an effect on airline costumer satisfaction
- Determine if customer class, Business or Economy, play a key role in airline costumer satisfaction
- Use visualizations to understand the correlation between the features which impact the airline costumer satisfaction
- Predict customer satisfaction using Machine learning (ML) algorithms
- Determine which ML algorithm is the most accurate

Dataset and Data Description

The dataset used to solve the problem of this project is *Airline passenger satisfaction measurement and service quality improvement* obtained from *Kaggle*, which contains about 130,000 survey entries.

In total, the dataset has both categorical and numerical attributes and includes 24 feature columns.

The dataset consists of 4 numerical variables and 19 categorical variables including the target variable "Satisfaction", that shows the level of satisfaction of the airline passenger: satisfied, neutral or dissatisfied. Moreover, from the categorical variables, 14 variables correspond to survey entries with a score from 0 to 5 that identify the passenger's satisfaction level.

Categorical and Numerical Variables:

Variables	Type
Gender	Categorical: Nominal
Customer Type	Categorical: Nominal
Age	Numerical
Type of Travel	Categorical: Nominal
Class	Categorical: Nominal
Flight Distance	Numerical
Inflight Wi-Fi service	Categorical: Ordinal
Departure/Arrival time convenient	Categorical: Ordinal
Ease of Online booking	Categorical: Ordinal
Gate location	Categorical: Ordinal
Food and drink	Categorical: Ordinal
Online boarding	Categorical: Ordinal
Seat comfort	Categorical: Ordinal
Inflight entertainment	Categorical: Ordinal
On-board service	Categorical: Ordinal
Leg room service	Categorical: Ordinal
Baggage handling	Categorical: Ordinal
Check-in service	Categorical: Ordinal
Inflight service	Categorical: Ordinal
Cleanliness	Categorical: Ordinal
Departure Delay in Minutes	Numerical
Arrival Delay in Minutes	Numerical
Satisfaction	Categorical: Ordinal

Data Description

Gender: Gender of the passengers (Female, Male)

Customer Type: The customer type (Loyal customer, disloyal customer)

Age: The actual age of the passengers

Type of Travel: Purpose of the flight of the passengers (Personal travel, Business travel)

Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

Flight distance: The flight distance of this journey

Inflight Wi-Fi service: Satisfaction level of the inflight Wi-Fi service (0:Not applicable; 1-5)

Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient

Ease of Online booking: Satisfaction level of online booking

Gate location: Satisfaction level of Gate location

Food and drink: Satisfaction level of Food and drink

Online boarding: Satisfaction level of Online boarding

Seat comfort: Satisfaction level of Seat comfort

Inflight entertainment: Satisfaction level of Inflight entertainment

On-board service: Satisfaction level of On-board service

Leg room service: Satisfaction level of Leg room service

Baggage handling: Satisfaction level of Baggage handling

Check-in service: Satisfaction level of Check-in service

Inflight service: Satisfaction level of Inflight service

Cleanliness: Satisfaction level of Cleanliness

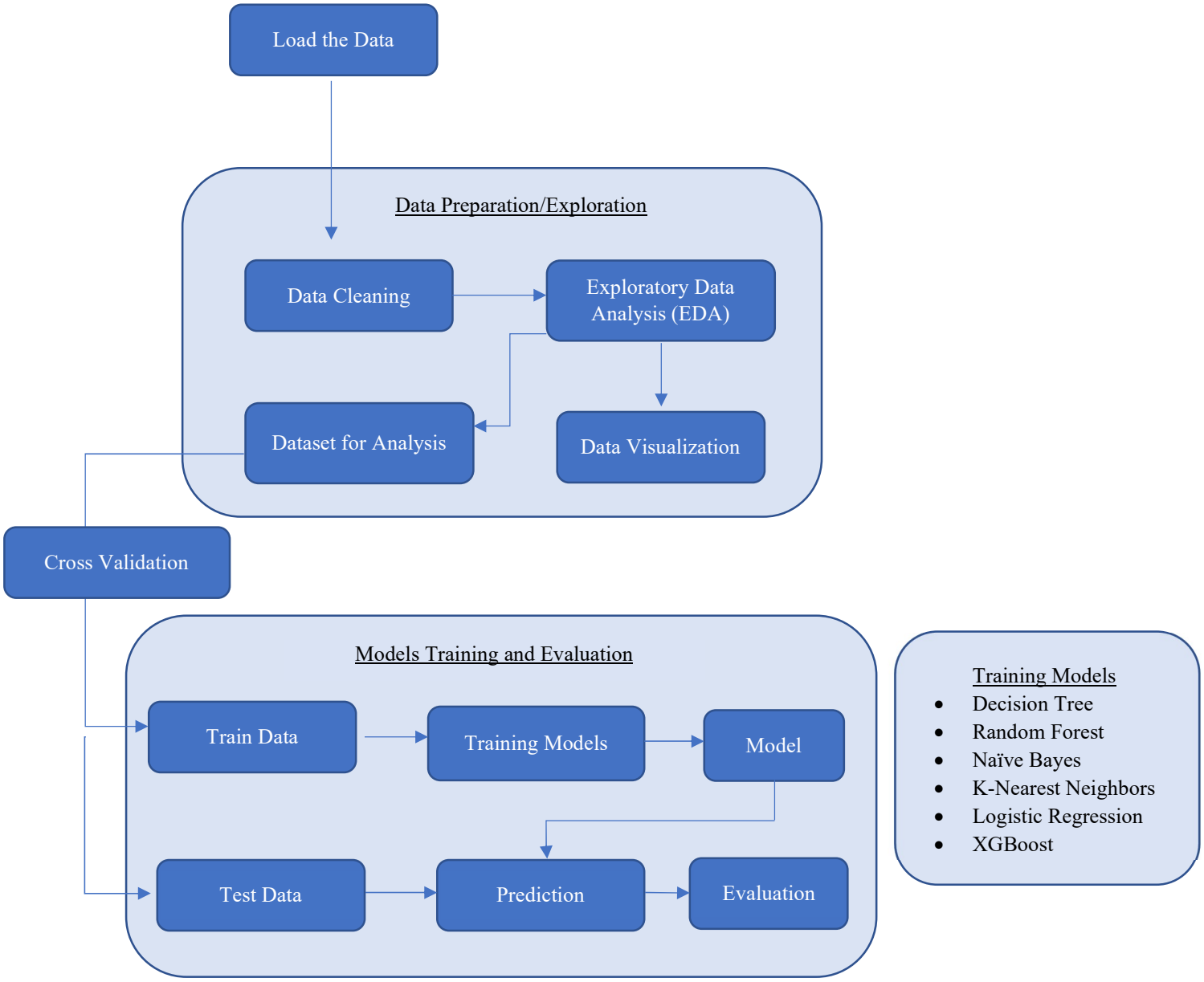
Departure Delay in Minutes: Minutes delayed when departure

Arrival Delay in Minutes: Minutes delayed when Arrival

Satisfaction: Airline Satisfaction level (Satisfied, neutral or dissatisfied) (Target variable)

Methodology

The table below shows a flow chart of the methodology for this capstone project.



Initial Analysis

This part of the project corresponds to the data preparation after importing the necessary libraries and the dataset.

The subsequent steps are completed to assign appropriate column names and dropping columns in train and test dataset:

- Elements renamed in 'Customer Type': 'disloyal customer' as 'First-time Customer'
- Elements renamed in 'Class' column: 'Eco', and 'Eco Plus' to 'Economy'
- Column 'satisfaction' corrected to 'Satisfaction'
- Column 'Leg room service' renamed
- Drop unnecessary 'Unnamed' column and sort by ascending 'id'
- Column 'Departure Delay in Minutes' type changed to float as 'Arrival Delay in Minutes'

Exploratory Data Analysis

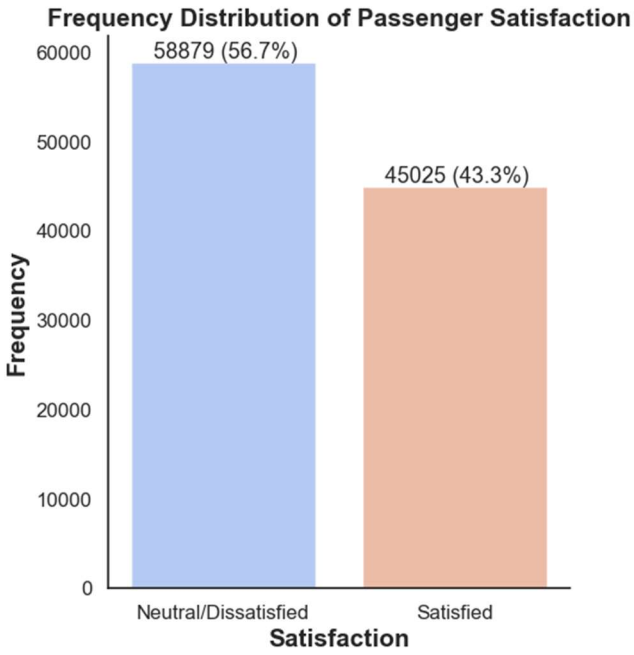
This part of the project corresponds to a Univariate, Bivariate and Multivariate Analysis.

Univariate Analysis

- **Categorical Variables**

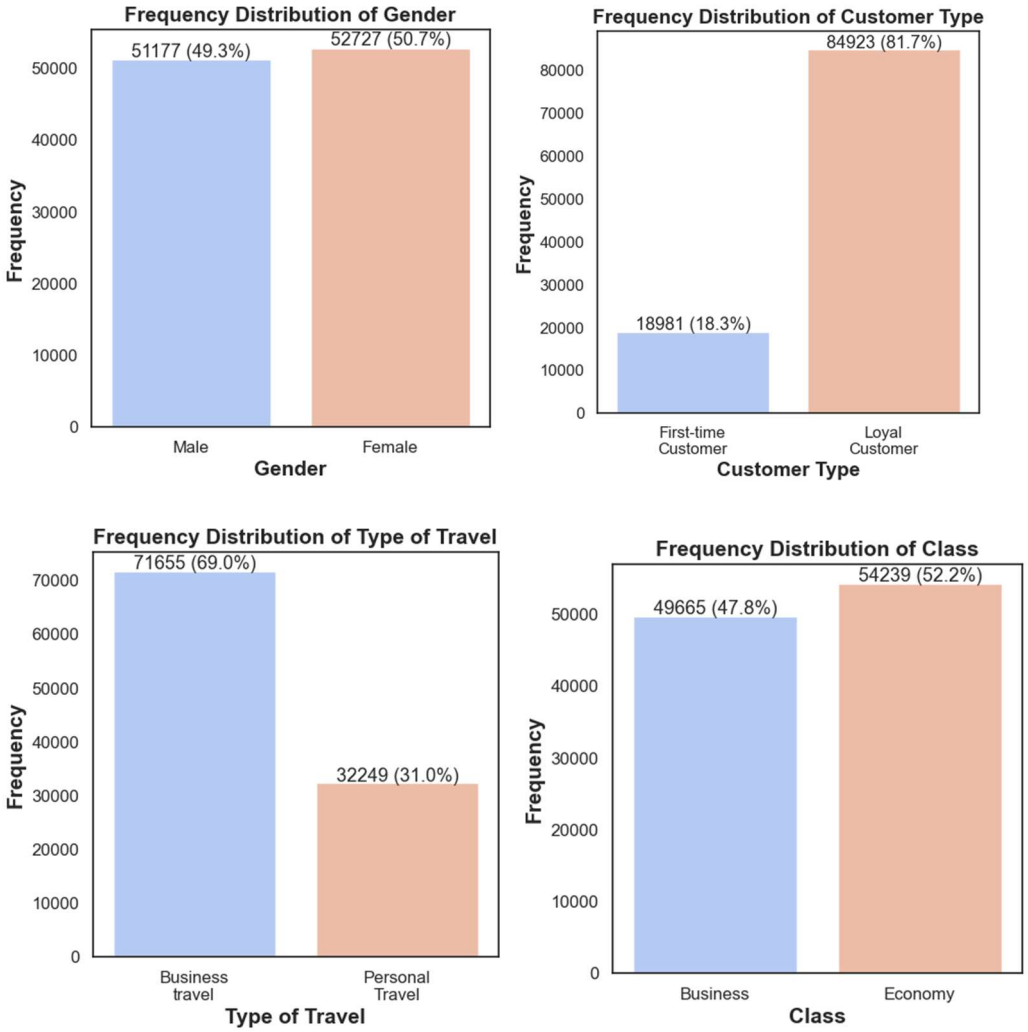
  Target variable

  Checking for Imbalance in the Target variable:



The above plot shows that the data in the target variable is relatively evenly distributed between the Neutral/Dissatisfied and the Satisfied passengers (57% and 43% respectively). It can be interpreted as a balanced data because there is no significant difference. For this reason, the dataset will not need any special treatment/resampling.

Frequency distribution of categorical variables: Gender, Customer Type, Type of Travel, and Class



- There is no significant difference in the Gender distribution: males (49%) and females (51%).
- The difference between First-time Customers and Loyal Customers is significant, with a great number of Loyal Customers (82%).
- The type of travel that is predominant is the Business customers (69%), and the airline might focus more on this group.

- This plot shows that the Economy class is the predominant after combining the Eco and Eco Plus classes, but the difference is small. Although, the Business Type of Travel was higher, there were lower customers in Business Class (48%).

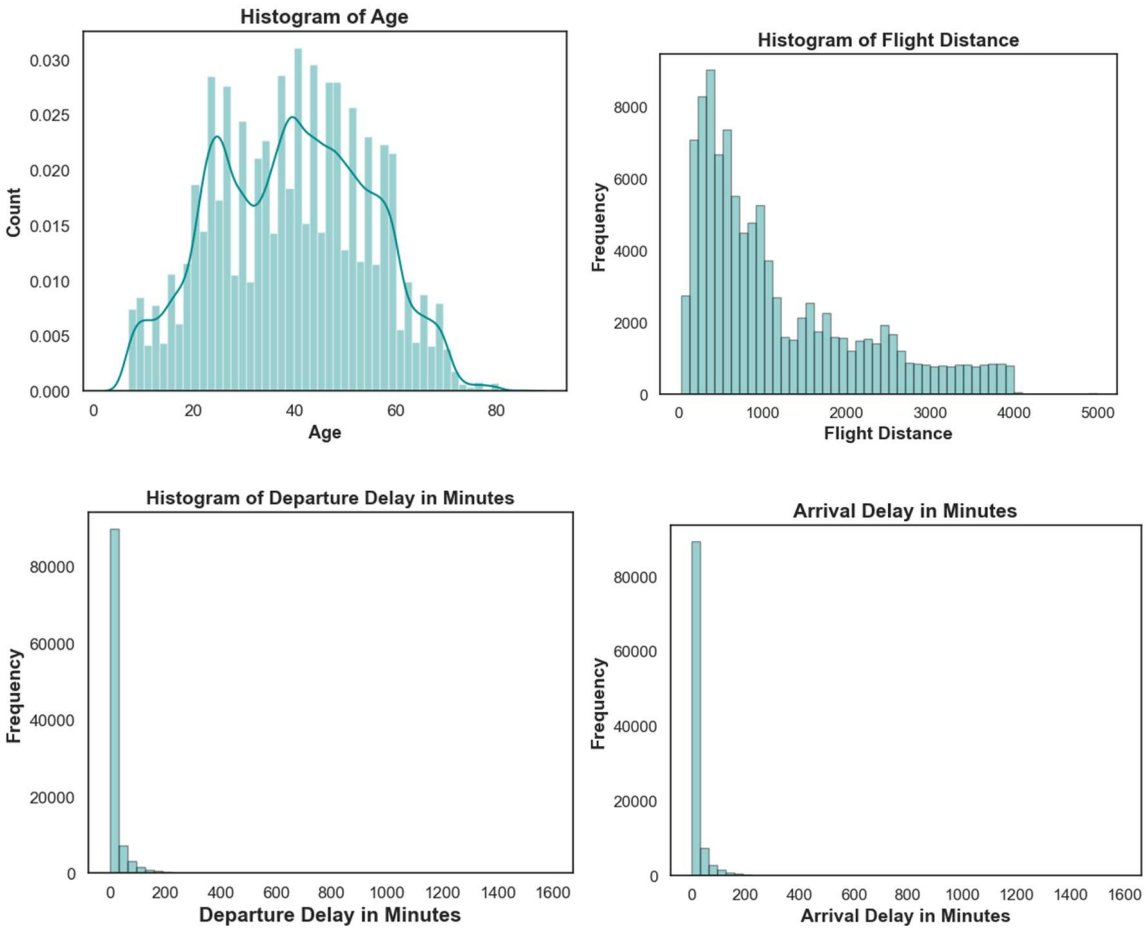
- **Numerical Variables**

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
Age	103904.0	39.379706	15.114964	7.0	27.0	40.0	51.0	85.0
Flight Distance	103904.0	1189.448375	997.147281	31.0	414.0	843.0	1743.0	4983.0
Departure Delay in Minutes	103904.0	14.815618	38.230901	0.0	0.0	0.0	12.0	1592.0
Arrival Delay in Minutes	103594.0	15.178678	38.698682	0.0	0.0	0.0	13.0	1584.0

Departure and Arrival Delay features show large values: 1592 and 1584 respectively. This will be assessed when detecting and treating Outliers in the dataset.

Histogram of Age, Flight Distance, Departure and Arrival Delay in Minutes:



- The Age distribution seems to be slightly right-skewed.
- This right-skewed histogram of Flight Distance shows that most flights possibly are domestics with a flight distance under 1000 miles.
- The histogram of Departure and Arrival Delay in minutes shows the possible presence of outliers which will be assessed in the following steps.

Data Cleaning

Missing Data - NaN Values

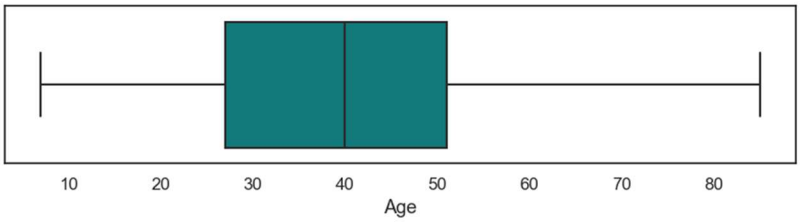
The missing values present in the train and test dataset are concentrated in the Arrival Delay in Minutes variable: there are 310 missing values in train dataset, and 83 missing values in the test dataset.

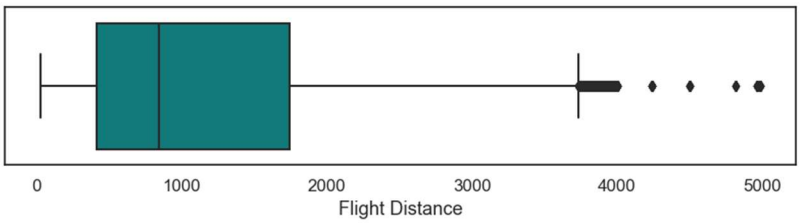
These missing values are replaced with the average values of the Arrival Delay in Minutes variable. After the imputation of the values, the shape of the train and test dataset is checked to confirm that there is no variation.

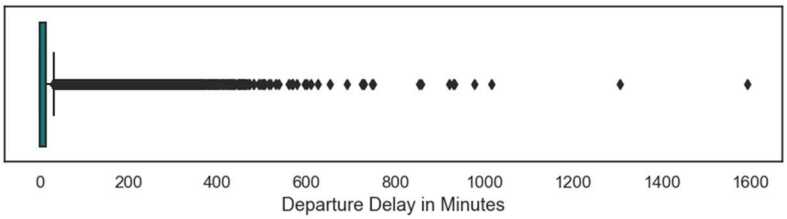
Duplicate Values

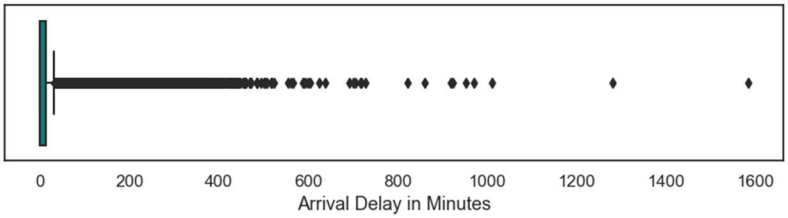
There are no duplicate values in the dataset.

Outliers

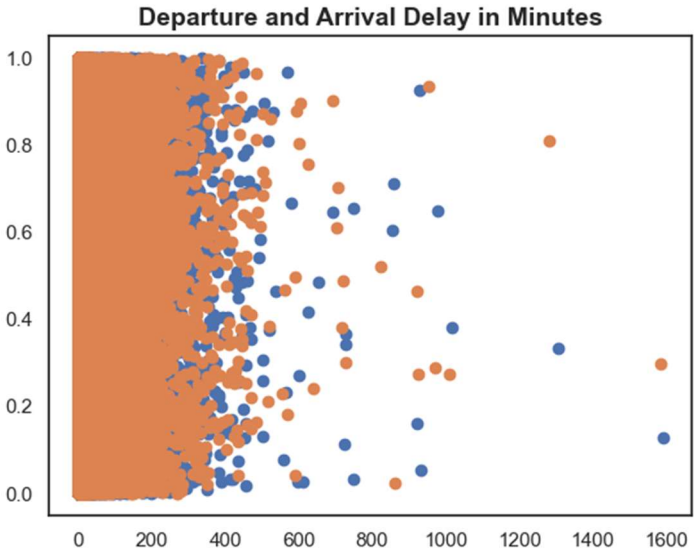








The above boxplots and the descriptive statistics table shows that there are large values for the Departure Delay in Minutes and Arrival Delay in Minutes: 1592 and 1584 respectively.



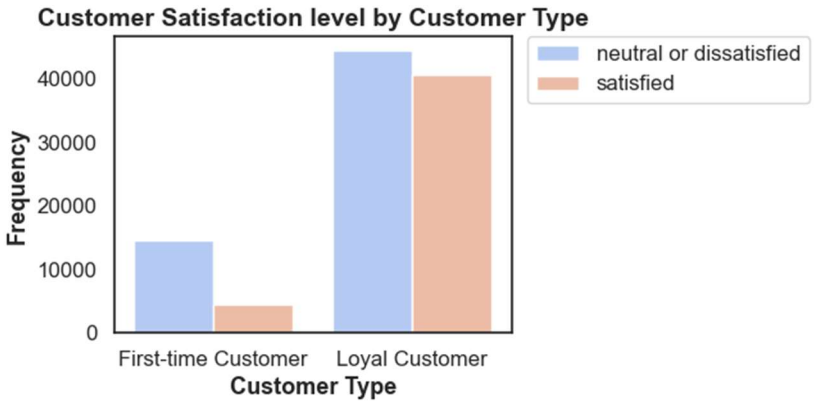
Those extreme few values are not errors, and they seem to be a natural part for the dataset. For this reason, the outliers will not be removed from the dataset.

Bivariate Analysis

- **Customer Satisfaction level in relation to the Categorical Variables**

Customer Satisfaction level by Customer type

Customer Type: Loyal and First-time Customer



Satisfaction	neutral or dissatisfied	satisfied
Customer Type		
First-time Customer	13.9%	4.3%
Loyal Customer	42.7%	39.0%

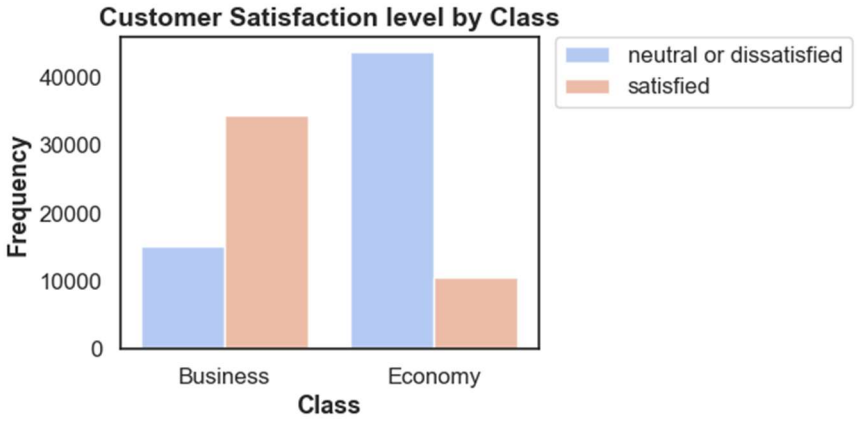
This plot shows that among the Loyal customers, there is a similar ratio between neutral/dissatisfied (43%) and satisfied (39%) airline customers.

On the other hand, for First-time customers, there is a lower ratio of satisfaction level. This could be for the higher expectation that normally first-time customers have.

Overall, Loyal customers (39%) has a higher satisfaction level than the First-time customers (4%).

Customer Satisfaction level by Class

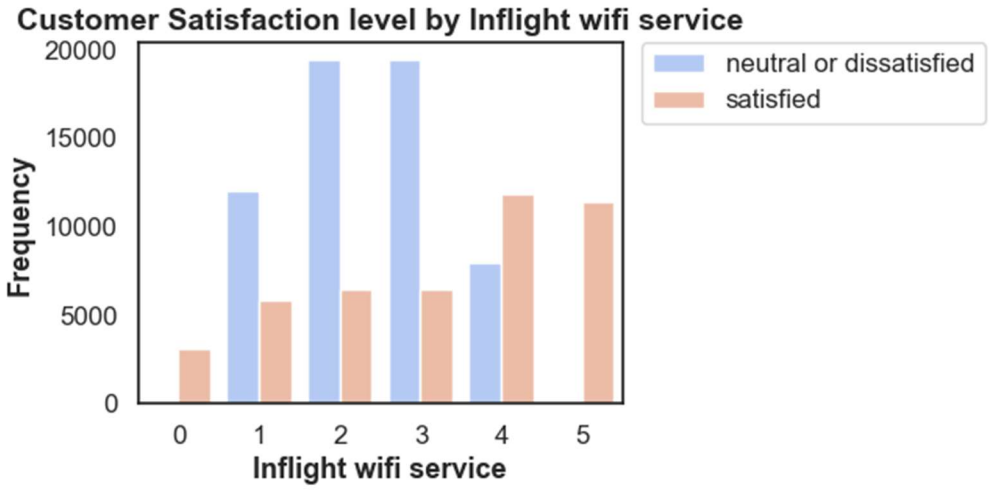
Class: Business and Economy



Satisfaction	neutral or dissatisfied	satisfied
Class		
Business	14.6%	33.2%
Economy	42.1%	10.1%

This plot shows the predominance of customers satisfied in the Business class (33%) comparing with the large number of neutral/dissatisfied customers in the Economy class (42%).

Customer Satisfaction level by Inflight Wi-Fi service



Satisfaction	neutral or dissatisfied	satisfied
Inflight wifi service		
0	0.0%	3.0%
1	11.6%	5.6%
2	18.7%	6.2%
3	18.7%	6.2%
4	7.6%	11.4%
5	0.1%	10.9%

This plot shows that there is a high level of customer dissatisfaction when the Inflight Wi-Fi service is rated as poor service (rate below 3). On the other hand, there is a high level of customer satisfaction when the service is rated good/excellent.

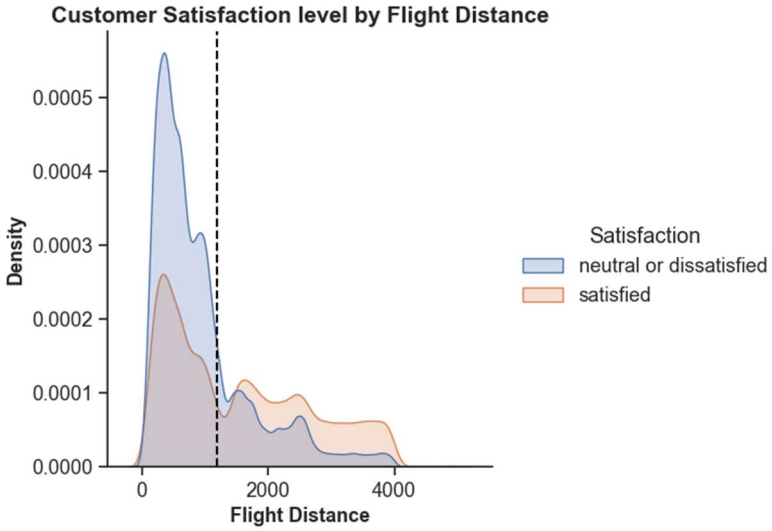
This feature could be a major factor that the airlines should consider in order to improve their inflight services.

- **Customer Satisfaction level in relation to the Numerical Variables**

Customer Satisfaction level by Age

This density plot shows that in the age range between 40 and 60, the number of satisfied customers is higher than the neutral/dissatisfied customers.

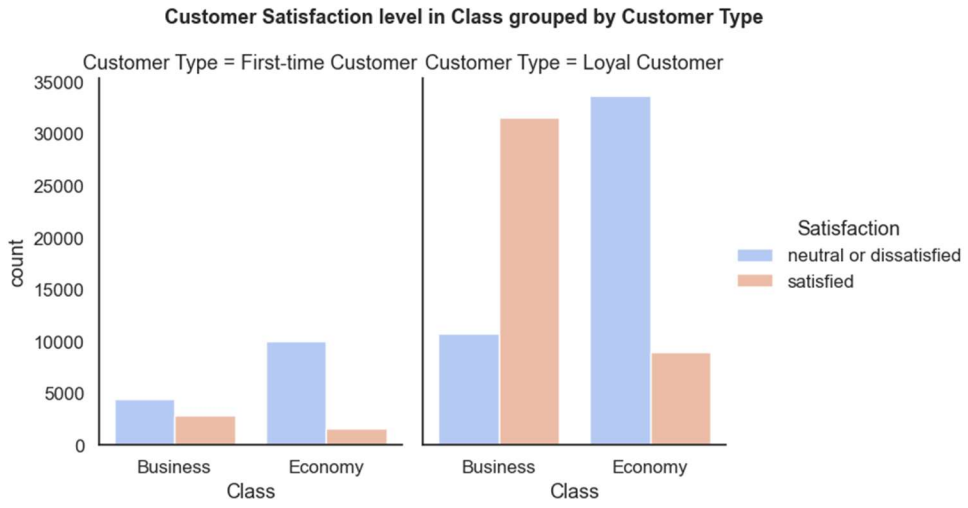
Customer Satisfaction level by Flight Distance



This plot shows that the level of customer dissatisfaction is high in customers that travelled approximately less than 1100 miles (approx. mean). On the other hand, there is a significant level of customer satisfaction in long distance flight (more than approx. 1100 miles).

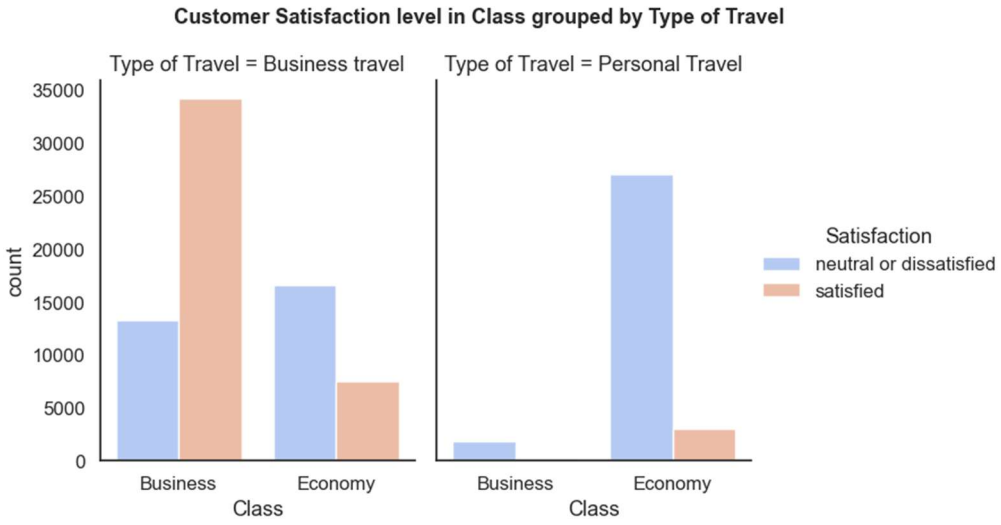
Multivariate Analysis

Customer Satisfaction level in relation to Class by Consumer Type



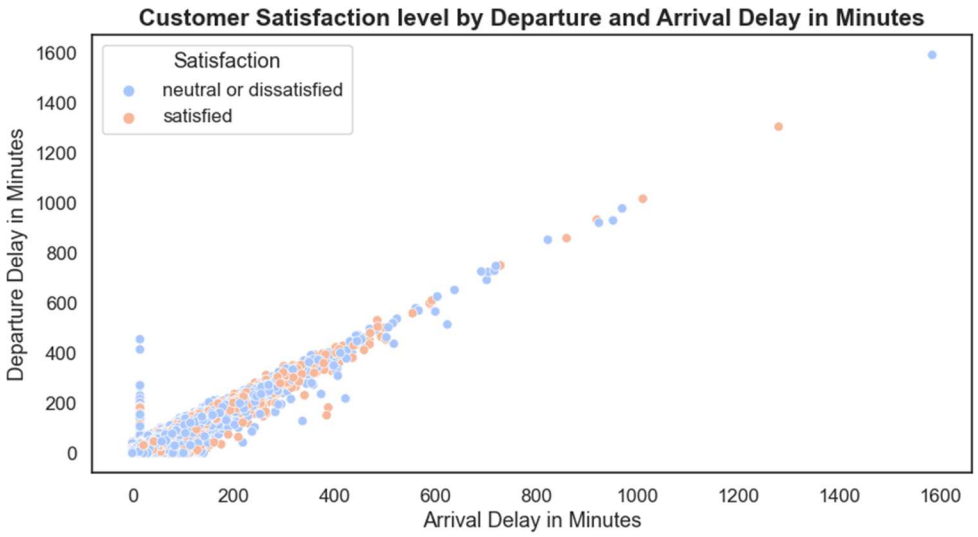
This plot confirms that Loyal Customers, who travel in Business Class, are likely more satisfied than the First-time customers who go in Business Class.

Customer Satisfaction level in relation to Class by Type of Travel



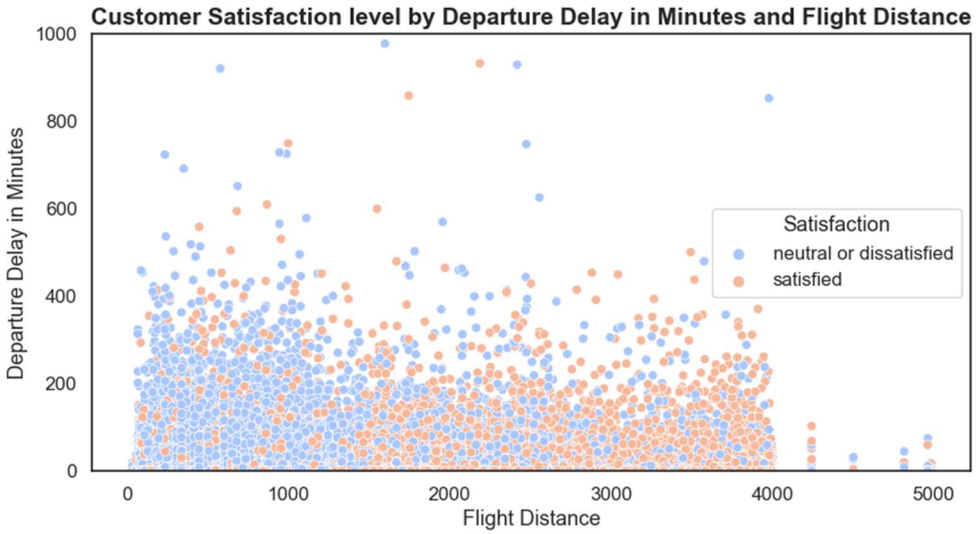
This plot confirms that customers with Business Type of Travel, and who travel in Business Class, are likely more satisfied than the customers with Personal Type of Travel who go in Business Class.

Customer Satisfaction level by Departure and Arrival Delay in Minutes



This scatterplot shows and confirms that the Departure and Arrival Delay in minutes have a linear relationship.

Customer Satisfaction level by Departure Delay in Minutes and Flight Distance



This scatterplot shows that when the flight distance is small, the airline customers are more likely to be neutral/dissatisfied with a delay in the departure.

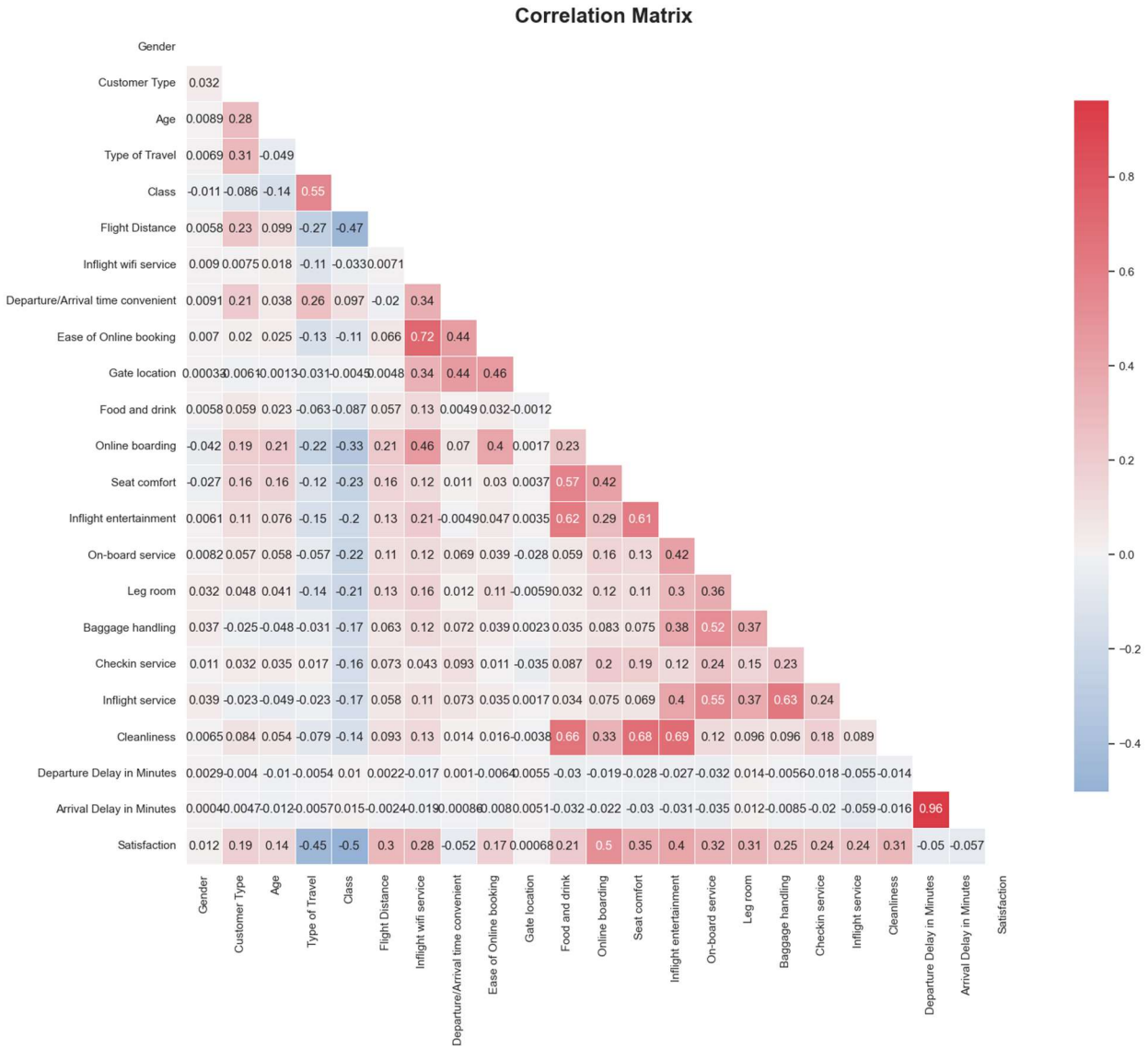
Feature Selection

Before the correlation, the categorical variables have been transformed to numeric variables using label encoding in train and test dataset.

Using Feature Selection, the model would run efficiently, and the models would have less dimensionality to reduce noise and possible overfitting.

One of the methods for feature selection is visualised in the figure below that shows the correlation matrix heatmap.

Correlation among features



This correlation matrix shows that the Satisfaction target feature has some weak correlation with all other features, having a negative correlation with the Class feature (0.5038), and A positive correlation with the Online boarding feature (0.5036). Furthermore, there is no significant correlation between the Satisfaction target with the features: Gender (0.012) and Gate location (0.00).

Moreover, it shows the following positive correlation between features:

Cleanliness and Inflight entertainment (0.69),

Cleanliness and Seat Comfort (0.68),

Ease of Online booking and Inflight Wi-Fi service (0.72),

Departure and Arrival Delay in Minutes (0.96).

Feature Importance

Class	0.503848
Online boarding	0.503557
Type of Travel	0.449000
Inflight entertainment	0.398059
Seat comfort	0.349459
On-board service	0.322383
Leg room	0.313131
Cleanliness	0.305198
Flight Distance	0.298780
Inflight wifi service	0.284245
Baggage handling	0.247749
Inflight service	0.244741
Checkin service	0.236174
Food and drink	0.209936
Customer Type	0.187638
Ease of Online booking	0.171705
Age	0.137167
Arrival Delay in Minutes	0.057497
Departure/Arrival time convenient	0.051601
Departure Delay in Minutes	0.050494
Gender	0.012211
Gate location	0.000682

The features that have relatively significant correlation with the Satisfaction target feature are: Class, Online boarding, Type of Travel, Inflight entertainment, Seat comfort, On-board service, Leg room, Cleanliness, and Flight Distance.

Feature Importance using Permutation Importance

The second method of feature selection that is applied is the feature importance using Permutation Importance. This method explains how much impact a variable has in the decision of Satisfied, or Neutral/dissatisfied.

Weight	Feature
0.1498 ± 0.0015	Inflight wifi service
0.1394 ± 0.0016	Type of Travel
0.0537 ± 0.0008	Customer Type
0.0435 ± 0.0011	Online boarding
0.0288 ± 0.0006	Checkin service
0.0266 ± 0.0002	Class
0.0244 ± 0.0002	Baggage handling
0.0232 ± 0.0007	Seat comfort
0.0198 ± 0.0004	Inflight service
0.0153 ± 0.0003	Cleanliness
0.0124 ± 0.0005	On-board service
0.0124 ± 0.0004	Leg room
0.0120 ± 0.0003	Age
0.0107 ± 0.0002	Flight Distance
0.0079 ± 0.0003	Arrival Delay in Minutes
0.0078 ± 0.0005	Inflight entertainment
0.0054 ± 0.0001	Gate location
0.0047 ± 0.0003	Ease of Online booking
0.0046 ± 0.0001	Departure Delay in Minutes
0.0040 ± 0.0002	Departure/Arrival time convenient
	… 2 more …

The most relevant features using the Permutation importance are: Inflight wifi service, Type of Travel, Customer Type, Online boarding, Check in service, Class, Baggage handling, Seat comfort, Inflight service, Cleanliness, On-board service, Leg room.

After careful consideration, the features to be dropped are Gender, Age, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Departure Delay in Minutes, and Arrival Delay in Minutes.

Modelling

The steps implemented in the Exploratory Data Analysis data-preprocessing help in the better understanding of the data and prepare them to be implemented for the prediction model in order to get high accuracy.

Before implementing the machine learning algorithms, 10-fold cross-validation strategy is applied for each classifier to avoid sample variability which can affect the model performance, and to minimize the effects of bias.

Stratified sampling of 10-fold cross-validation is used to ensure that each fold is a good representation of the whole dataset, where each fold contains the same proportion of the two classes: Satisfied, or Neutral/dissatisfied.

Machine learning algorithms

This project corresponds to a classification problem, in which a machine learning algorithm is used to train and test the dataset's attributes to predict a categorical attribute which indicates the class target: satisfaction level. The performance of the algorithm is evaluated by the accurate predicting. This project applies six supervised machine learning algorithms: Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, and XGBoost.

Logistic Regression

Logistic regression is a regression analysis technique which is a set of statistical processes used to assess the relationships between variables. This set of techniques is used to model and analyze the relationship between a dependent and the independent variables.

Logistic Regression can be used to predict a qualitative response by modelling the probability that in the database the target feature Satisfaction belongs to a particular category: neutral/dissatisfied, Satisfied (Noviantoro & Huang, 2022).

Gaussian Naïve Bayes

Naïve Bayes classifier uses statistical and probability methods that, based on previous experience, predict future opportunities. Naïve Bayes is a simple technique that generates models that assign class labels to problem instances, defined as vectors of functional meaning, in which a specific set includes class labels. The main advantage of this method is that only need a small number of training data to calculate classification parameters (Soni, 2020 and Noviantoro & Huang, 2022).

K-Nearest Neighbors

The K-Nearest Neighbors is a supervised learning algorithm that assesses new data with the nearest neighbours' previous training data sample. This simple algorithm works by extracting and classifying all available cases based on a similarity behaviour. The main advantage of this method is that the model's class decision line can be very flexible and non-linear (Soni, 2020).

Decision Tree

A Decision tree is the most common classification method because is easy to interpret for its simplicity and clarity. It is a predictive model that uses a tree structure and identifies an item’s observation (branch) to a conclusion about the target value (leaf) (Liberman, 2020).

Random Forest

The central concept of Random Forest algorithm is that to build a lot of many decision trees, it combines many trees and makes predictions with each decision tree. The process is described as: a) random points are selected from the training set, then b) decision trees are built from the data,

by choosing the number of the decision trees, then repeat steps a) and b), and the last step is a decision tree is built for all the input data (Singh, 2020; and Noviantoro & Huang, 2022).

Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems. The advantages of using Extreme Gradient Boosting include a more direct route to the minimum error, converging more quickly with fewer steps, and simplified calculations to improve speed and lower compute costs.

Performance evaluation metric

The performance evaluation metric used in this project is the prediction accuracy and F1 score which assessed each supervised machine learning algorithm's prediction performance.

Both accuracy and F1 score can be determined by the confusion matrix, shown below.

Confusion matrix.

↓Predicted\actual→	neutral or dissatisfied	satisfied
Neutral or dissatisfied	True Positive	False Positive
Satisfied	False Negative	True Negative

The average prediction accuracy level is computed based on how many data samples

of the prediction model are appropriately classified in a test dataset.

The F1 score is a balance between precision and recall. There is a perfect harmony between precision and recall when the F1 score is higher.

The accuracy and F1 score are obtained by the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

In addition to this two metrics, Precision and Recall are other scoring metrics that this project applies.

Precision is the ratio of the number of positive classes correctly predicted by the model to the total number of predicted positive classes.

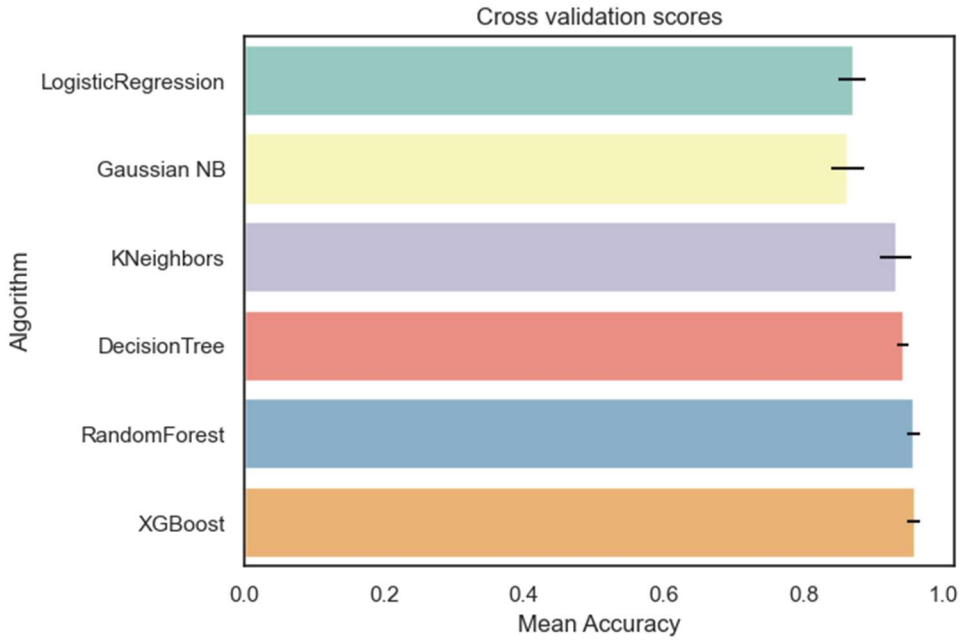
Recall is the ratio of the number of positive classes correctly predicted by the model to the total number of the actual positive classes.

Furthermore, this project also examines the receiver operating characteristic (ROC) graphs of model classifiers. The ROC curves help organising, illustrating, and evaluating the prediction models based on their performance.

The AUC (area under the ROC curve) is a numerical quantity that represents the probability of the model prediction to rank positive instances that are randomly selected higher than the negative instances (Noviantoro & Huang, 2022).

Model performance and Evaluation

Results on performance prediction after feature selection



The following table shows the performance prediction results of the supervised machine learning algorithms after feature selection.

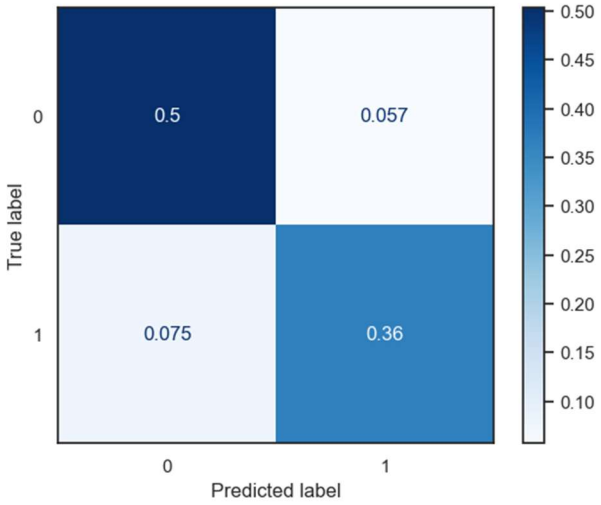
Machine Learning	Accuracy	Precision	Recall	F1 score	AUC	Time taken
Logistic Regression	86.84%	86.47%	83.02%	84.71%	0.92	0.761
Gaussian Naïve Bayes	86.19%	86.91%	80.71%	83.69%	0.92	0.063
K-Nearest Neighbor	93.07%	97.07%	86.82%	91.67%	0.98	9.361
Decision Tree	93.76%	93.79%	91.85%	92.81%	0.98	0.091
Random Forest	96.00%	96.86%	93.94%	95.37%	0.99	6.270
Extreme Gradient Boosting (XGBoost)	95.49%	95.69%	93.97%	94.83%	0.99	36.272

This project compared six machine learning algorithms, as Random Forest demonstrate the highest accuracy as 96% and highest F1 score as 95.37%, to predict airline customer satisfaction. Extreme Gradient Boosting (XGBoost) follows it with accuracy 95.49% and F1 score 94.83%.

Machine Learning algorithms:

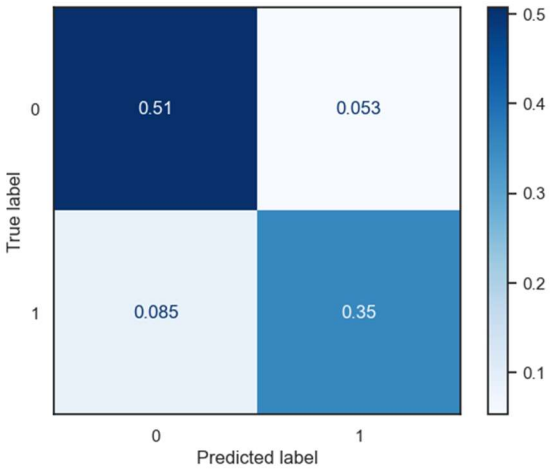
Logistic Regression

```
Accuracy = 0.8684170003079766
ROC Area under Curve = 0.864262601129349
Time taken = 0.7609250545501709
              precision    recall  f1-score   support

           0    0.87117   0.89831   0.88453     14573
           1    0.86465   0.83022   0.84708     11403

    accuracy                        0.86842     25976
   macro avg    0.86791   0.86426   0.86581     25976
weighted avg    0.86830   0.86842   0.86809     25976
```



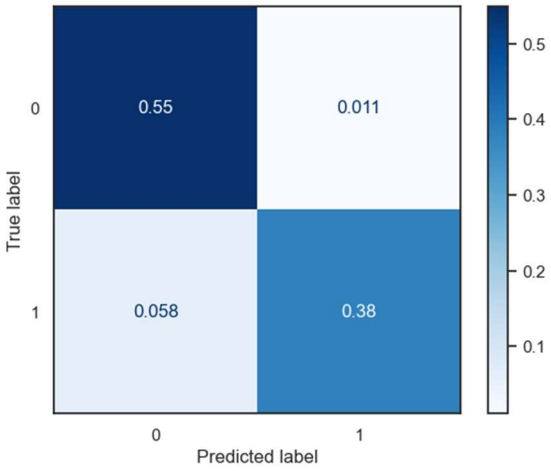
Naive Bayes

```
Accuracy = 0.8619494918386202
ROC Area under Curve = 0.8559804624880696
Time taken = 0.0630347728729248
              precision    recall  f1-score   support

           0    0.85702   0.90489   0.88031     14573
           1    0.86911   0.80707   0.83694     11403

    accuracy                        0.86195     25976
   macro avg    0.86307   0.85598   0.85862     25976
weighted avg    0.86233   0.86195   0.86127     25976
```



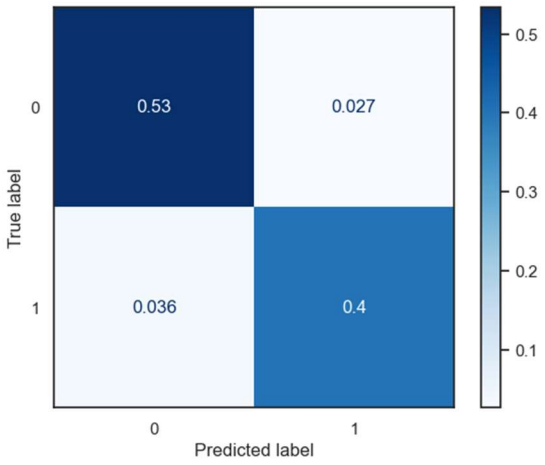
K-Nearest Neighbor

```
Accuracy = 0.9307052663997536
ROC Area under Curve = 0.9239157509939812
Time taken = 9.361132621765137
              precision    recall  f1-score   support

           0    0.90480   0.97955   0.94069     14573
           1    0.97078   0.86828   0.91667     11403

    accuracy                        0.93071     25976
   macro avg    0.93779   0.92392   0.92868     25976
weighted avg    0.93376   0.93071   0.93015     25976
```



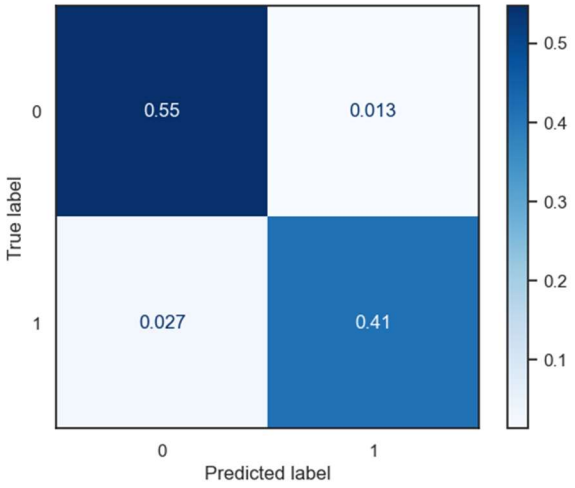
Decision Tree

```
Accuracy = 0.9375577456113335
ROC Area under Curve = 0.9354882580790782
Time taken = 0.0919337272644043
              precision    recall  f1-score   support

           0    0.93727   0.95245   0.94480     14573
           1    0.93794   0.91853   0.92813     11403

    accuracy                        0.93756     25976
   macro avg    0.93761   0.93549   0.93647     25976
weighted avg    0.93756   0.93756   0.93748     25976
```



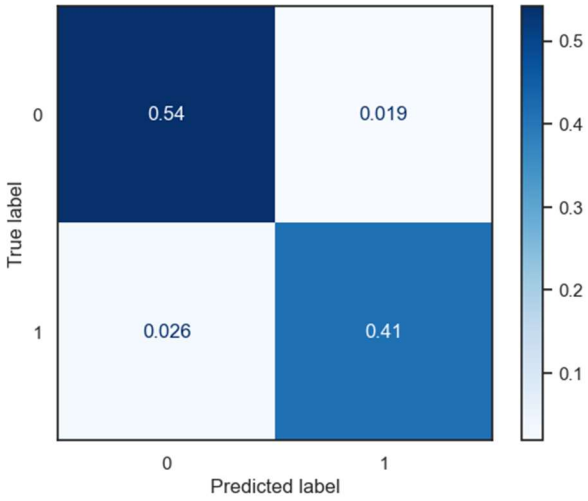
Random Forest

```
Accuracy = 0.9600400369571912
ROC Area under Curve = 0.9577953770786729
Time taken = 6.270402431488037
              precision    recall  f1-score   support

           0    0.95368   0.97619   0.96480     14573
           1    0.96862   0.93940   0.95379     11403

    accuracy                        0.96004     25976
   macro avg    0.96115   0.95780   0.95930     25976
weighted avg    0.96024   0.96004   0.95997     25976
```



Extreme Gradient Boosting (XGBoost)

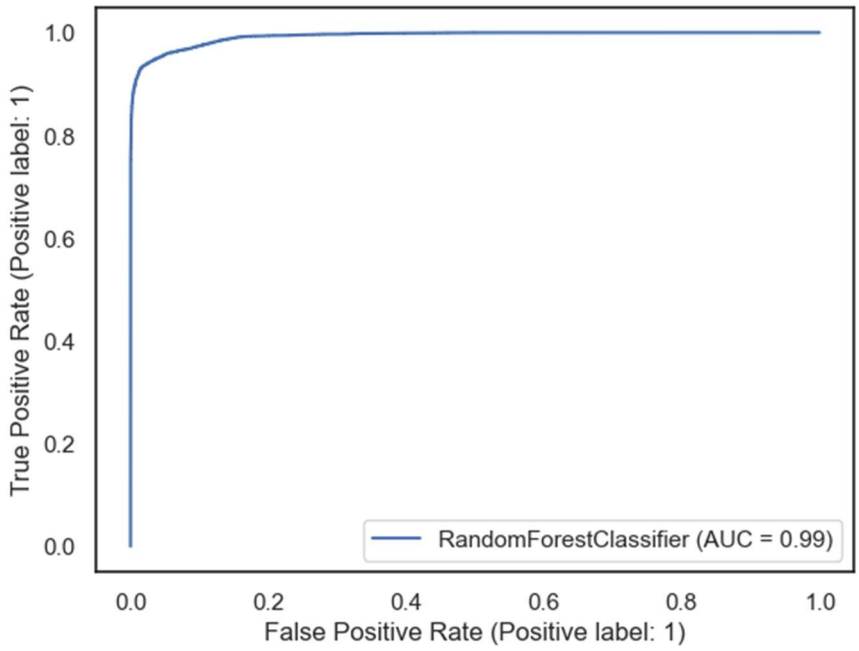
```
Accuracy = 0.9549969202340622
ROC Area under Curve = 0.9533389160916872
Time taken = 36.27225995063782
              precision    recall  f1-score   support

           0    0.95351   0.96693   0.96017     14573
           1    0.95696   0.93975   0.94828     11403

    accuracy                        0.95500     25976
   macro avg    0.95523   0.95334   0.95422     25976
weighted avg    0.95502   0.95500   0.95495     25976
```

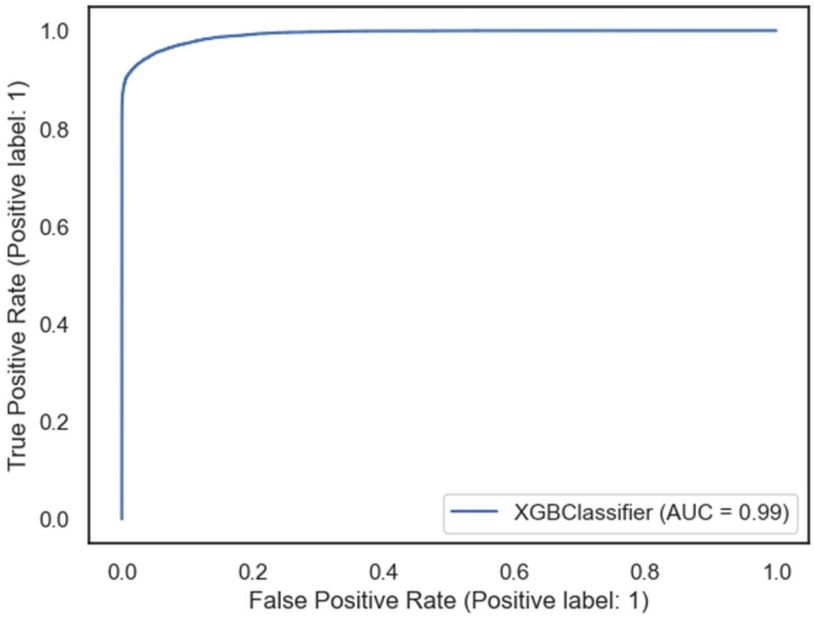


Results on ROC after feature selection

This project also evaluates the ROC graphs to ensure the visualization of performance prediction among classifiers.

The following graphs show the ROC curve result which illustrates Random Forest and Extreme Gradient Boosting (XGBoost) as the fittest classifiers with a AUC score = 0.99. Both algorithms are the fittest classifier for predicting airline customer satisfaction.




1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
XGBClassifier (AUC = 0.99)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)

Discussion

The most important services that need to be improved by the airlines to gain customer satisfaction are Inflight Wi-Fi service, Type of Travel, Customer Type, Online boarding, Check in service, Class, Baggage handling, Seat comfort, Inflight service, Cleanliness, On-board service, Leg room, Flight Distance, and Inflight entertainment.

Comparing with Nurhadi et al. (2019) study of online questionnaires research for experience using air transportation, this project findings are similar considering that customers experience dissatisfaction due to a lack of Wi-Fi service and inflight entertainment, while customers seem to experience a certain level of satisfaction using online booking.

According to Tsafarakis et al. (2018) research of multicriteria satisfaction analysis method, two important variables that can improve the airline passenger satisfaction are the inflight entertainment and inflight Wi-Fi service, which are aligned with the findings in this project.

The findings of Hayadi et al. (2021) study about the Predicting Airline passenger Satisfaction using Classification algorithms is similar with our project findings related to that Random Forest algorithm is the method which has the highest accuracy, precision, and recall. Moreover, it concludes that the Inflight Wi-Fi and the Online Booking services are critical services where the airline should concentrate on.

The results in the Gao et al. (2021) study suggest that passengers' characteristics such as type of travel, class and customer type, and service attributes such as Wi-Fi service, online boarding and baggage handling are crucial factors and have comparatively larger effects on airline travel satisfaction. These features are similar to our findings after applying feature importance using Permutation Importance.

This project findings are more related to the determination of the most important factors that the airlines providers need to consider based on the application of machine learning algorithms with the purpose of increasing the satisfaction level and the customer loyalty.

It is important that the airlines implement strategies to improve the satisfaction level by customer type, particularly of First-time customers which are less likely to have a high level of satisfaction, possibly because, as first-time customers, they have higher expectations. This will ensure a better probability for them to become returning or loyal customers.

Conclusion

The most important services that need to be improved by the airlines to gain customer satisfaction are Inflight Wi-Fi service, Type of Travel, Customer Type, Online boarding, Check in service, Class, Baggage handling, Seat comfort, Inflight service, Cleanliness, On-board service, Leg room, Flight Distance, and Inflight entertainment.

The features Customer Type (loyal and first-time customer) and customer class (Business or Economy) indeed have an impact on the airline customer satisfaction. In both cases, airlines should focus on provide a quality service to maintain loyal and business class customers, as well as airlines should implement new strategies to reach First-time customers' expectations, and to improve Economy class customers' experience.

This project assessed six machine learning algorithms on the airline customer satisfaction dataset. The different supervised machine learning algorithms are Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). The project findings show that Random Forest is the fittest machine learning algorithm and demonstrate the highest accuracy as 96% and highest F1 score as 95.37%, to predict airline customer satisfaction.

This project proposed that airlines should focus on the optimization of services such as the Inflight Wi-Fi service experience by creating tools to make it accessible or lowering the cost of accessing it for the economy class. Moreover, airlines should concentrate on the simplicity of Online Booking to make the airline experience more flexible and comfortable.

Work limitations

There are other major factors that influence the airline customer satisfaction that are not investigated in this project. The International air Transport association (IATA, 2022) list over 70 factors that affect the airline travel satisfaction.

Recommendation

The airline businesses should prioritize the most important services to meet customer satisfaction.

One of the services that should be prioritized is the online boarding because is fast and convenient by avoiding passenger queues and saving the passenger time and energy.

Other important service that should be improved is inflight wi-fi service to attract more passengers and to ensure the return of the loyal customers. The airline service should provide wi-fi service with free internet with limited access or at reasonable price. Inflight entertainment is also an important service, particularly to long distance flights.

GitHub Repository Link

https://github.com/nvgril/CIND820-Project_2022

References

1. Abhishek, M. (2020, February 27). *XGBoost an efficient implementation of gradient boosting*. Agency. https://datascience.foundation/datatalk/xgboost-an-efficient-implementation-of-gradient-boosting

2. An, M., & Noh, Y. (2009). Airline Customer Satisfaction and loyalty: Impact of in-flight service quality. *Service Business*, *3*(3), 293–307. https://doi.org/10.1007/s11628-009-0068-4

3. Bellizzi, M. G., Eboli, L., Mazzulla, G., & Postorino, M. N. (2022). Classification trees for analysing highly educated people satisfaction with airlines' services. *Transport Policy*, *116*, 199–211. https://doi.org/10.1016/j.tranpol.2021.12.008

4. Etemad-Sajadi, R., Way, S. A., & Bohrer, L. (2016). Airline passenger loyalty: the distinct effects of airline passenger perceived pre-flight and in-flight service quality. *Cornell Hospitality Quarterly*, *57*(2), 219–225. https://doi.org/10.1177/1938965516630622

5. Gao, K., Yang, Y., & Qu, X. (2021). Examining nonlinear and interaction effects of multiple determinants on airline travel satisfaction. *Transportation Research Part D: Transport and Environment*, *97*, 102957. https://doi.org/10.1016/j.trd.2021.102957

6. Garcia, V., Florencia-Juarez, R., Sanchez-Solis, J. P., Rivera-Zarate, G., & Contreras-Masse, R. (2019). Predicting airline customer satisfaction using K-NN Ensemble regression models. *Research in Computing Science*, *148*(6), 205–215. https://doi.org/10.13053/rcs-148-6-15

7. Hayadi, B., Kim, J., Hulliyah, K., & Sukmana, H. (2021). Predicting airline passenger satisfaction with classification algorithms. IJIIS: International Journal of Informatics and Information Systems, 4(1), 82–94. https://doi.org/10.47738/ijiis.v4i1.80

8. Heiets, I., Xie, Y. (2021). The impact of the COVID-19 pandemic on the Aviation Industry. *Journal of Aviation*. https://doi.org/10.30518/jav.933296

9. International Air Transport Association (IATA) (2020a) Air Passenger Market Analysis International Air Transport Association. Retrieved from https://www.iata.org/en/iata-repository/publications/economic- reports/air-passenger-monthly-analysis—june-20202/.

10. Jiang, H., & Zhang, Y. (2016). An investigation of service quality, customer satisfaction and loyalty in China's airline market. *Journal of Air Transport Management*, *57*, 80–88. https://doi.org/10.1016/j.jairtraman.2016.07.008

11. Khan, R., & Urolagin, S. (2018). Airline sentiment visualization, consumer loyalty measurement and prediction using Twitter data. *International Journal of Advanced Computer Science and Applications*, *9*(6). https://doi.org/10.14569/ijacsa.2018.090652

12. Khudhair, H.Y., Jusoh, A., Nor, K.M. and Mardani, A. (2021) Price sensitivity as a moderating factor between the effects of airline service quality and passenger satisfaction on passenger loyalty in the airline industry. *International Journal of Business Continuity and Risk Management*, Vol. 11, Nos. 2/3, pp.114–125. https://doi.org/10.1504/ijbcrm.2021.116274

13. Kohavi, R., & Edu, S. (1993). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.

14. Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0224-1

15. Liberman, N. (2020, May 21). *Decision trees and random forests*. Medium. https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

16. Maminiaina, R. (2019). A thorough literature review of customer satisfaction definition, factors affecting customer satisfaction and measuring customer satisfaction. *International Journal of Advanced Research*, 7(9), 828–843. https://doi.org/10.21474/ijar01/9733

17. Noviantoro, T., & Huang, J.-P. (2022). Investigating airline passenger satisfaction: Data Mining Method. *Research in Transportation Business & Management*, 43, 100726. https://doi.org/10.1016/j.rtbm.2021.100726

18. Nurhadi, M. I., Ratnayake, N., & Fachira, I. (2019). The practice of digitalization in improving customer experience of Indonesian commercial aviation industry. *In The 4th ICMEM 2019 and The 11th IICIES 2019*, August, 7–9. https://journal.sbm.itb.ac.id/index.php/ProceedingSBMITB/article/download/3575/1452 .

19. Olaganathan, R. (2021). Impact of covid-19 on airline industry and strategic plan for its recovery with special reference to data analytics technology. *Global Journal of Engineering and Technology Advances*, 7(1), 033–046. https://doi.org/10.30574/gjeta.2021.7.1.0050

20. Pareek, P. (2021, September 3). *Logistic regression: Essential things to know*. Medium. https://medium.datadriveninvestor.com/logistic-regression-essential-things-to-know-a4fe0bb8d10a

21. Park, S.-H., Kim, M.-Y., Kim, Y.-J., & Park, Y.-H. (2022). A deep learning approach to analyze airline customer propensities: The case of South Korea. *Applied Sciences*, *12*(4), 1916. https://doi.org/10.3390/app12041916

22. Razafimanjary M. A. (2019). A thorough literature review of customer satisfaction definition, factors affecting customer satisfaction and measuring customer satisfaction. *International Journal of Advanced Research*, 7(9), 828–843. https://doi.org/10.21474/ijar01/9733

23. Singh, J. (2020, December 26). *Random Forest: Pros and cons*. Medium. https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04

24. Soni, A. (2020, July 3). *Advantages and disadvantages of KNN*. Medium. https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336

25. Soni, A. (2020, July 4). *Pros and cons of naive bayes classifier*. Medium. https://medium.com/@anuuz.soni/pros-and-cons-of-naive-bayes-classifier-40b67249ae8

26. Sonntag, T. (2021, December 15). *How did covid-19 impact the aviation industry*. CUNY Academic Works. https://academicworks.cuny.edu/bb_etds/122/

27. Suhartanto, D. (2014, June 8). Customer satisfaction in the airline industry: *The role of service quality and Price*. Academia.edu. https://www.academia.edu/3215480/Customer_Satisfaction_in_the_Airline_Industry_The_Role_of_Service_Quality_and_Price

28. Suhartanto, D. & Noor, A.A. (2012). Customer Satisfaction in the Airline Industry: The Role of Service Quality and Price. *Asia Tourism Forum Conference*. 1–9.

29. Sulu, D., Arasli, H., Saydam, M. B. (2021). Air-travelers' perceptions of service quality during the COVID-19 pandemic: Evidence from Tripadvisor Sites. *Sustainability*, 14(1), 435. https://doi.org/10.3390/su14010435

30. Tan, C. (2021). Bidirectional LSTM model in predicting satisfaction level of passengers on airline service. *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*. https://doi.org/10.1109/icaice54393.2021.00107

31. Vishalmendekarhere. (2021, January 22). It's all about assumptions, Pros & Cons. Medium. https://medium.com/swlh/its-all-about-assumptions-pros-cons-497783cfed2d