# AIRLINE PASSENGER SATISFACTION PREDICTION USING MACHINE LEARNING MODELS

*A project report submitted to ICT Academy of Kerala*

*in partial fulfilment of the requirements*

*for the certification of*

## CERTIFIED SPECIALIST

## IN

## DATA SCIENCE & ANALYTICS

submitted by

**Amrutha T G**

**Ardra Sudhakaran**

**Bhavya K**

**Jewel Anns**

**Sanjay Narayanadas**


**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**Nov 2022**

# List of Figures

# List of Tables

# List of Abbreviations

CART - Classification and Regression Tree

kNN - k Nearest Neighbours

PCA - Principal Component Analysis

SERVQUAL - Service Quality

ML - Machine Learning

# Table of Contents

# Abstract

Any industry or business is rife with rivalry. And every company that strives to stand out becomes a ‘customer favourite’. When it comes to the airline market the competition increases at a faster pace. Customer’s only point of distinction in the airline sector is the level of customer service provided to them.

Airlines always place customer satisfaction at the top of their priority list. Customers that are dissatisfied or disengaged inevitably result in fewer passengers and less money. It is critical that clients have a positive experience every time they travel.

The objective of all customer satisfaction models is to provide results that are relevant, reliable, and valid and have predictive financial capability. Customer satisfaction research is something that should be done with the greatest care. Measuring customer satisfaction must be a continuous, consistent, timely, accurate and a reliable process. This is where a new customer satisfaction approach becomes a powerful strategic business development tool for organization.

In the airline industry, customers expect ‘the best’ services including proper communications and no delays in flight departure and arrivals along with hassle free boarding-deboarding and inflight services. These factors are important for providing satisfactory customer service in the airline industry.

As part of this project, we are trying to train a model to predict airline passenger satisfaction using a dataset and we hope to test the different machine learning models to find which one suits our dataset and gives the most accurate results.

# 1. Problem Definition

## 1.1 Overview

Passenger airlines are airlines dedicated to the transport of human passengers. Passenger airlines typically operate a fleet of passenger aircraft that may be either owned outright by an airline company or leased from commercial aircraft sale and leasing companies. There is a large competition among the different airline companies. Inorder to increase revenue there must be more passengers opting for them. This can be done only through providing satisfactory customer experience throughout the journey, be it hassle free boarding to inflight services , a safe journey without delays, and even fair price availability. Maximizing profit is an essential requirement for all the airline companies. Here we try to find the possible reasons where passengers were dissatisfied and also how to improve satisfaction of passengers.

## 1.2 Problem Statement

In this project we aim to predict the satisfaction level of airline passengers based on the quality of services provided by the airline company. We also try to find out the different factors which affect the satisfaction of passengers. This is a classification problem and we aim to try the different machine learning models on classification like Logistic Regression, Decision Tree Random Forest and kNN and find the best suitable model which gives the most accurate prediction for our classification problem - predicting passenger satisfaction .

# 2. Introduction

Transportation services have become the basic needs of the community both for daily activities and travel needs. For a long-distance journey, most people prefer air transportation for the efficiency and effectiveness of time. Air transport can reach places that cannot be reached by other modes of transport such as land and sea, in addition to being able to move faster and have a straight, practically barrier-free path.Since in the late 70’s, air-passenger industry has changed considerably, the number of flights has increased, additional airports have opened, tickets are more affordable, and airline traffic continues to grow. However many argue that service offered to passengers shows little or no improvement . Regardless of the perspective taken, one fact stands clear: competition is ever increasing . Basically, a major element in this competitive battle is about the quality of services, and how to maintain passengers.

Excellent service is a profitable strategy because it results in more new customers, more business with existing customers, fewer lost customers. It also results in lower marketing costs because extra marketing money does not have to be spent convincing customers to buy despite the firm's poor service record. Passengers' expectations concerning the quality of service they receive have increased in recent years, and airlines are working very hard to meet these expectations. This means that airline management must have a good understanding of the ways in which passengers assess service quality. Satisfaction is not only considered as a customer's goal to be derived as a result of degrading services, but also as a company's goal, as a way of getting higher customer retention rates and ways of generating profits. If the service / product is provided in accordance with customer expectations, he will feel satisfied and increase the level of consumer loyalty. Conversely, if service delivery is lower

than customer expectation, service quality will be considered bad and decrease consumer loyalty.

In theory, if a customer is satisfied with the service or product provided, he will be loyal to use the product and even tell others the benefits of the product or service. Satisfied customers will continue to buy the product again. Although customer satisfaction is not the main goal, customer satisfaction is the key to the success of a company to maintain the quality of product / service and maintain the image of the company's brand so that customers will repurchase. That is why, the company must provide superior service quality to win the business competition among Airlines. However, due to the pandemic caused by COVID-19 and intense competition among airlines, many airline companies are struggling to attract new customers. Considering that airline service quality is the main factor in obtaining new and retaining existing customers, airline companies are using various approaches to improve the quality of the physical and social services.

Machine learning and big data technologies made it possible to analyze huge databases and to develop highly accurate prediction or classification models.This study will attempt to analyze service quality within the airline industry and to determine potential areas of improvement within the passenger / airline relationship using machine learning models. Machine learning techniques are used to develop a binary class classification model for Airline Satisfaction.

# 3. Literature Survey

A detailed and exhaustive review on Airline Passenger Satisfaction and Machine Learning techniques used in the predictive analysis of passenger satisfaction is done and is discussed below.

## 3.1 Airline Passenger Satisfaction

Customer satisfaction is increasingly recognized as a determinant of business performance and a strategic tool for gaining competitive advantage. High and stable customer satisfaction is considered an important determinant of an organization's long-term profitability. Research has also shown a significant moderate-to-strong association between satisfaction and a company's financial and market performance. More specifically, customer satisfaction is strongly linked to retention, revenue, earnings per share, and stock price.

For the aviation industry, some studies have used aviation services to build an index model for passenger satisfaction. Based on the combination of China's customer satisfaction index model and the actual situation of China Southern Airlines' satisfaction management, Zhang designed the China Southern Airlines customer satisfaction evaluation index and proposed nine secondary indicators with air transport characteristics: flight operation quality satisfaction degree, ticketing service satisfaction, ground service satisfaction, air service satisfaction, arrival station service satisfaction, irregular flight service satisfaction, consumption value perception, overall satisfaction, and customer loyalty.

There are also studies using flight data or text reviews to predict passenger satisfaction. Sankaranarayanan et al. used a logistic model tree (LMT) machine learning approach to predict passenger satisfaction levels based on factors such as

airport punctuality, number of flights, punctuality rankings, average delays, and queue times for inferring passenger perceptions of punctuality and delay-related event satisfaction.To achieve high levels of customer satisfaction, service providers should provide high levels of service quality, as service quality is often considered a prerequisite for customer satisfaction. Since passengers are the direct recipients of services, service quality indirectly affects enterprise development by affecting passenger satisfaction. Therefore, airlines can understand the quality of the services provided by passengers' satisfaction with each service, check the services, and then improve the service quality. For the aviation industry, it is more efficient to improve customer satisfaction by accurately understanding the main factors affecting passenger satisfaction and making improvements based on service priorities. Several studies have investigated the main factors influencing passenger satisfaction. By constructing a nested logit model of airport-airline choice in the "two-step" decision-making process of air passengers,Suzuki determined that the factors that play an important role in airline choice are ticket price, frequency of flight service provided to desired destinations and frequent flyer membership. Tsafarakis et al. proposed that the improvement of onboard entertainment, onboard Wi-Fi services can improve airline passenger satisfaction according to the multi-standard satisfaction analysis method.

In the study entitled "The impact of airline service quality on passenger satisfaction and loyalty", an exploratory analysis of UAE airports by Mohammed Arif, Aman Gupta and Aled Williams have analyzed the ways to improve the customer satisfaction with regard to the aviation industry in the country.Responses from random customers were collected to perform Chi-square test and analysed the differences between airports.The study gives a better knowledge about the public

view with regard to the innovations and ideas implemented by the government of UAE.

R.Archana and DR. M.V Subha in their study " A study on service quality and passenger satisfaction on Indian airlines" investigated the impact of the in-flight service quality passenger satisfaction and concluded that airline marketing managers have to develop various policies to provide guaranteed quality services to passengers. The study also stated that failure to provide quality services to passengers may damage the airline image and cause negative impact on passengers.

Nathalie Martel and Prianka N Seneviratne, in their research titled, ''Analysis of Factors Influencing Quality of Service in Passenger Terminal Buildings" have focused on many different factors tobe considered other than space or time when it comes to evaluating Quality of Services from the passengers' point of view. In the conclusion it is shown that 53 percent of the respondents believed that information is the most important factor. Similarly, for the waiting areas the most important factor was the availability of seats and for the processing elements it was the waiting time.HakJun Song, Wenjia Ruan and Yunmi Park [4], in their research titled, ''Effects of Service Quality, Corporate Image, and CustomerTrust on the Corporate Reputation of Airlines" have stated the cau-sal relationships among the perceived service quality, corporate image, customer trust, and corporate reputation of Asiana Airline in South Korea using SERVQUAL measures. The results showed the responsiveness and reliability of service quality significantlyaffect corporate image and customer trust, whereas tangibles,empathy, and assurance of service quality are not significant ante-cedents of corporate image and customer trust.

### 3.2 Customer Satisfaction Prediction using Machine Learning Models

Machine Learning (ML) models are increasingly being investigated as a tool to predict customer satisfaction in many industries. These approaches have the potential to provide valuable insights for formulating business strategies for improving customer satisfaction and thereby creating more profits.

There are many studies using flight data or text reviews to predict airline passenger satisfaction. V Gracia et al. uses a k-nn Ensemble Regression Model to predict airline passenger satisfaction where they focused on exploring the use of the k-nearest neighbour (k-nn) model for regression as base classifier. B. Herawan Hayadi et al. studied the competition in the aviation industry and factors influencing passenger satisfaction in detail. In their research, they used several classification models such as KNN, Logistic Regression, Gaussian NB, Decision Trees and Random Forest. The results of this study was that the Random Forest Algorithm using a threshold of 0.7 got an accuracy of 99% and was the best performing model and an important factor in getting customer satisfaction was the Inflight Wi-Fi Service.

# 4. Exploratory Data Analysis

The Dataset we have taken is from a US Airline Passenger Survey. The Data originally contains 1,29,880 survey entries, having a total of 24 columns .

There are 23 feature columns and 1 target column. Out of the 23, 14 columns are survey entries where passengers rate the flight experience on a scale of 1 to 5.

The columns included in the data are:

1. Id - numbers indexing the rows
2. Gender: male or female
3. Customer type: loyal or disloyal airline customer (changed to first time user and returning)
4. Age: the actual age of the passenger
5. Type of travel: the purpose of the passenger's flight (personal or business travel)
6. Class: business, economy, economy plus
7. Flight distance
8. Inflight wifi service: satisfaction level with Wi-Fi service on board (0: not rated; 1-5)
9. Departure/Arrival time convenient: departure/arrival time satisfaction level (0: not rated; 1-5)
10. Ease of Online booking: online booking satisfaction rate (0: not rated; 1-5)

11. Gate location: level of satisfaction with the gate location (0: not rated; 1-5)
12. Food and drink: food and drink satisfaction level (0: not rated; 1-5)
13. Online boarding: satisfaction level with online boarding (0: not rated; 1-5)
14. Seat comfort: seat satisfaction level (0: not rated; 1-5)
15. Inflight entertainment: satisfaction with inflight entertainment (0: not rated; 1-5)
16. On-board service: level of satisfaction with on-board service (0: not rated; 1-5)
17. Leg room service: level of satisfaction with leg room service (0: not rated; 1-5)
18. Baggage handling: level of satisfaction with baggage handling (0: not rated; 1-5)
19. Checkin service: level of satisfaction with checkin service (0: not rated; 1-5)
20. Inflight service: level of satisfaction with inflight service (0: not rated; 1-5)
21. Cleanliness: level of satisfaction with cleanliness (0: not rated; 1-5)
22. Departure delay in minutes
23. Arrival delay in minutes

## 4.1 Univariate Analysis

The results of the univariate analysis after plotting the bar graphs are as follows:

### 4.1.1 Target Column: Satisfaction

The target column ‘satisfaction’ contains two classes ‘satisfied’ and ‘neutral or dissatisfied’. When 56.6 % of the passengers were dissatisfied, 43.4% of passengers were satisfied. Which implies that both the classes in the target column are unbiased or the data is balanced.

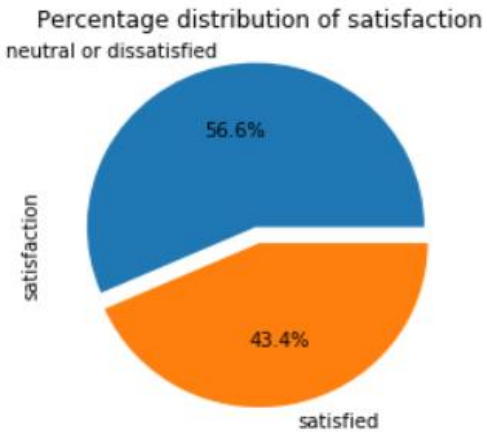


Fig 4.1 : Percentage Distribution of Target column - satisfaction

### 4.1.2 Survey Ratings

We found that the survey ratings usually range from 1 to 5. But many of the ratings included 0 which indicated that some percentage of passengers didn’t rate certain facilities. Most customers gave low ratings for Inflight wifi-service, gate location, and ease of online booking. Fig 1.2 shows the pie charts of the different survey ratings and their distribution.

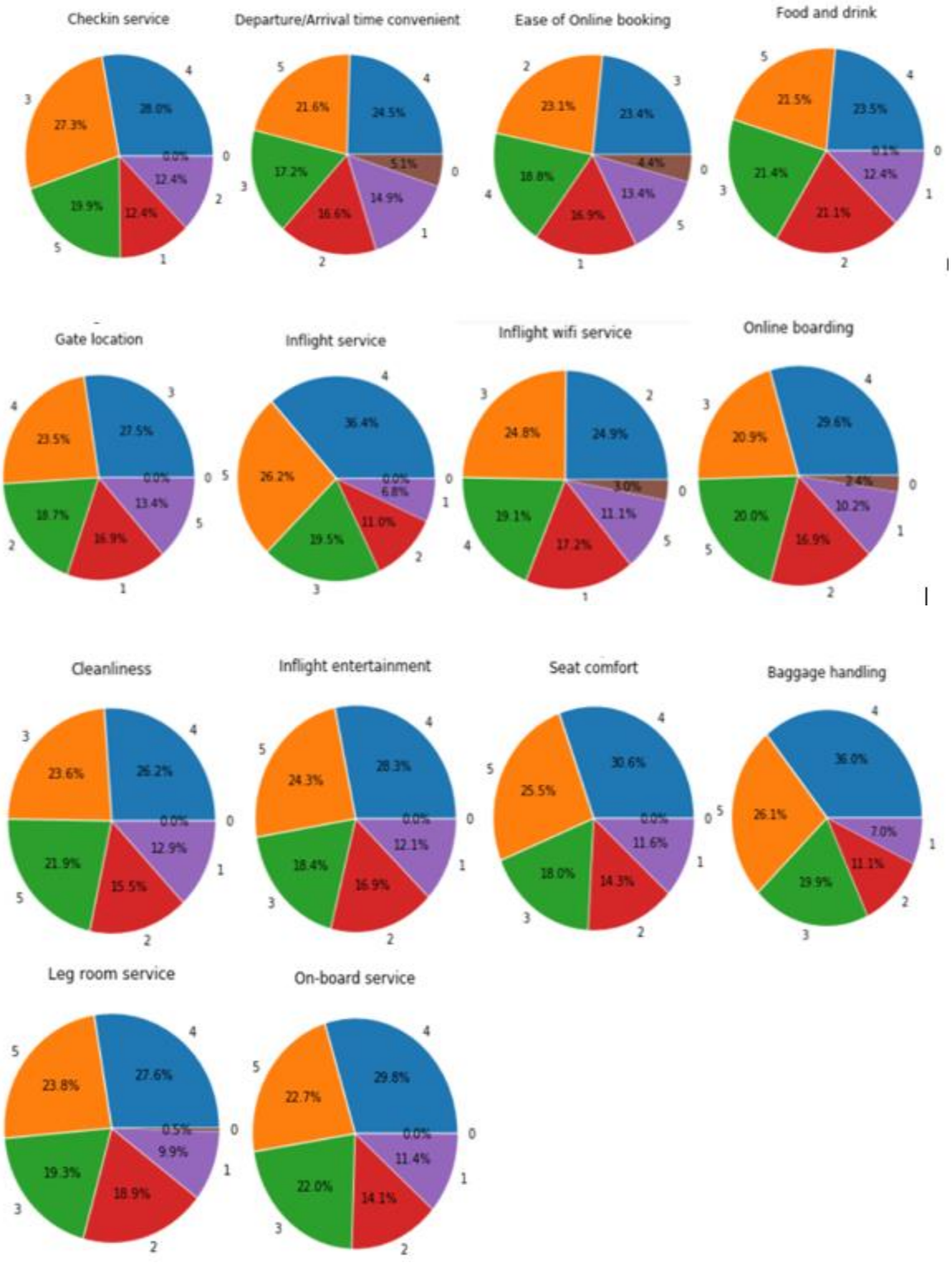


Fig 4.2 : Percentage Distribution of the different survey ratings

### 4.1.3 Other Features

By plotting and analyzing different graphs of the various features, the findings are given as below:

Gender: Data contains almost equal no. of male and female passengers.

Customer Type: Most of the passengers are returning customers(81.69%), First time customers constitute only 18.30% of the passengers.

Type of Travel: Most passengers travelled for business purposes(69%).

Class: Passengers travelling in ‘Business’ class and ‘Economy’ class are almost equal, only a small percent of passengers travelled in ‘Eco Plus’ Class.

## 4.2 Bivariate Analysis

Plots to determine the relationship between the target satisfaction and the various features and the various relationships between the features were drawn. Analyzing these graphs the conclusions are as follows:

### 4.2.1 Passenger Satisfaction vs Ratings

Bar graphs were plotted between the different survey ratings w.r.t satisfaction. We found that, Passengers who have given 5 or 4 star ratings for departure or arrival time convenience are not really satisfied with the whole service, or they were disappointed with the other services. Fig 4.3 shows some of the graphs plotted between survey rating and satisfaction.

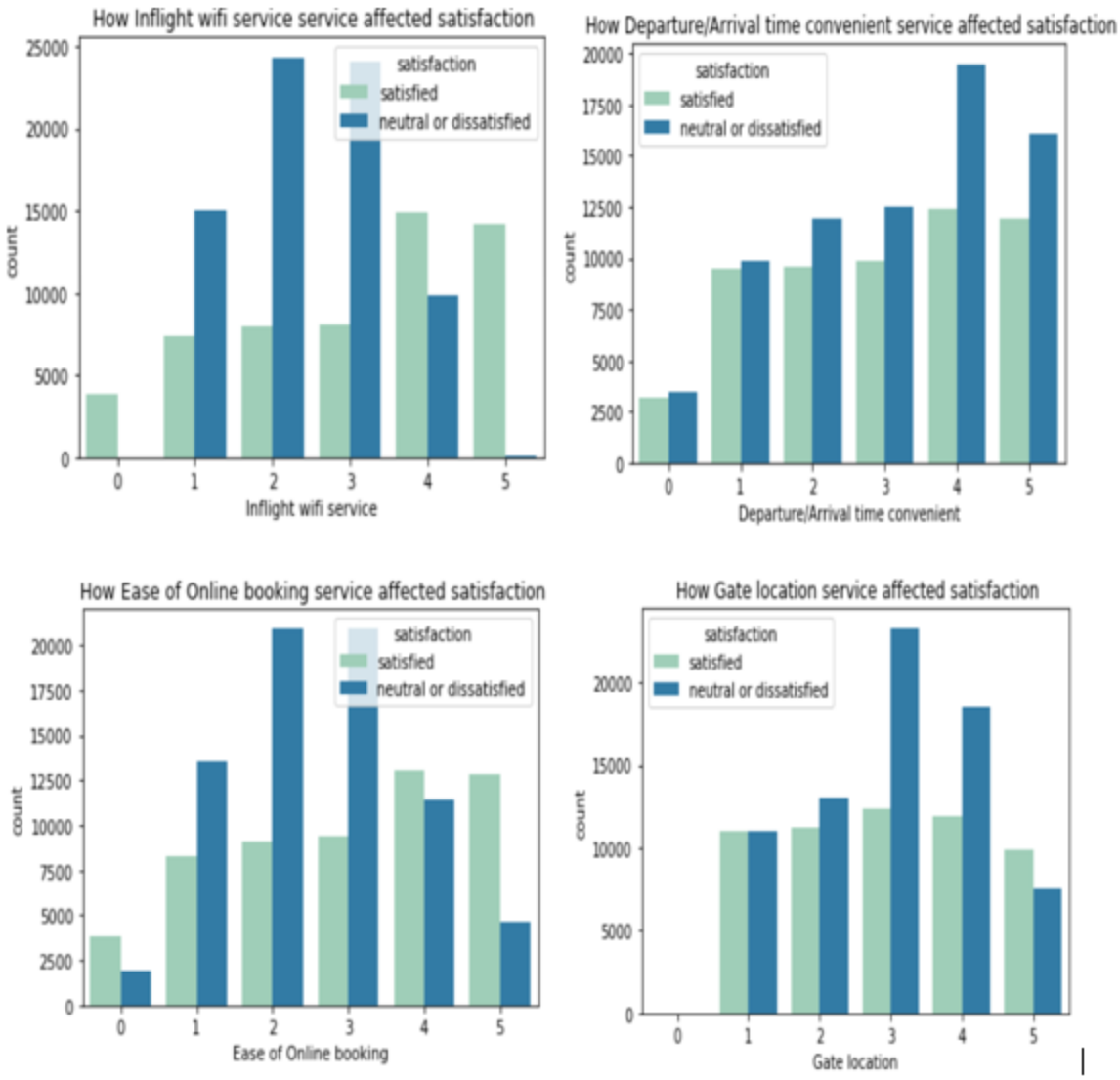


Fig 4.3 : Graph Survey Ratings w.r.t Satisfaction

### 4.2.2 Passenger Satisfaction vs Categorical Features

Bar graphs were plotted between the different categorical features such as Gender, Type of Travel, Class, Type of Customer etc., w.r.t satisfaction. We found that Satisfaction values are uniformly distributed among both the genders. Passengers who used business class were the most satisfied and in both eco and eco plus class,

more people were dissatisfied. And considering Type of Customer, both first time users and returning customers are equally satisfied or dissatisfied.

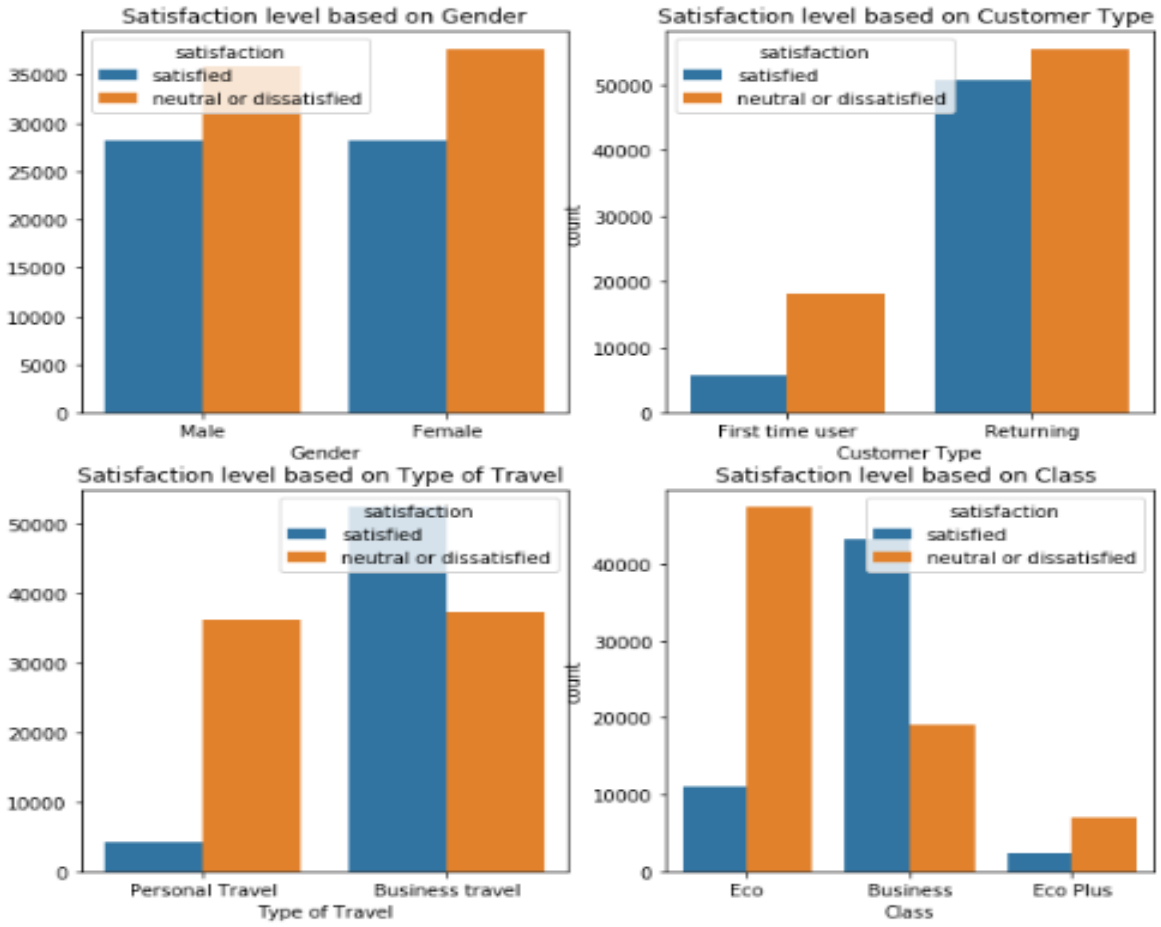


Fig 4.4 : Graph Categorical features w.r.t Satisfaction

### 4.2.3 Departure Delay in Minutes vs Arrival Delay in Minutes

There is high linear correlation between the features Departure delay in minutes and Arrival Delay in minutes. Indicating that if there is a delay in departure that time is not caught up during the flight journey and that delay is correspondingly detected in the arrival delay.


Fig 4.5 : Departure Delay vs Arrival Delay Scatter Plot

### 4.2.4 Type of Travel vs Flying Class

We found that people who travelled for Business purposes preferred Business Class over Eco and Eco plus classes and people who travelled for Personal purposes used Eco and Eco plus classes rather than Business class.

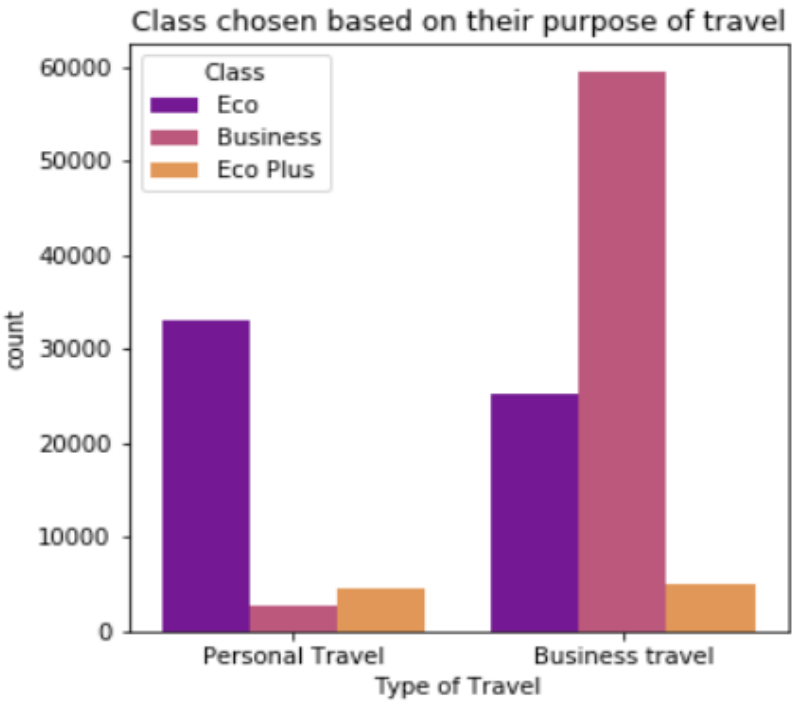


Fig 4.6 : Type of Travel w.r.t Flying Class

4.2.5 How Age and Flight distance Affected Satisfaction

To find how age and flight distance affected satisfaction we grouped the age to age classes consisting of child, youth, adult and senior and flight distance into short and long distance.

We found that people who took short distance flights were the most dissatisfied and people who took long distance flights were equally satisfied or dissatisfied. Shown in Fig 4.7

Considering age, more passengers were from the adult category (25-64) and they travelled for business purposes and age didn't have much impact on satisfaction. Shown in Fig 4.8

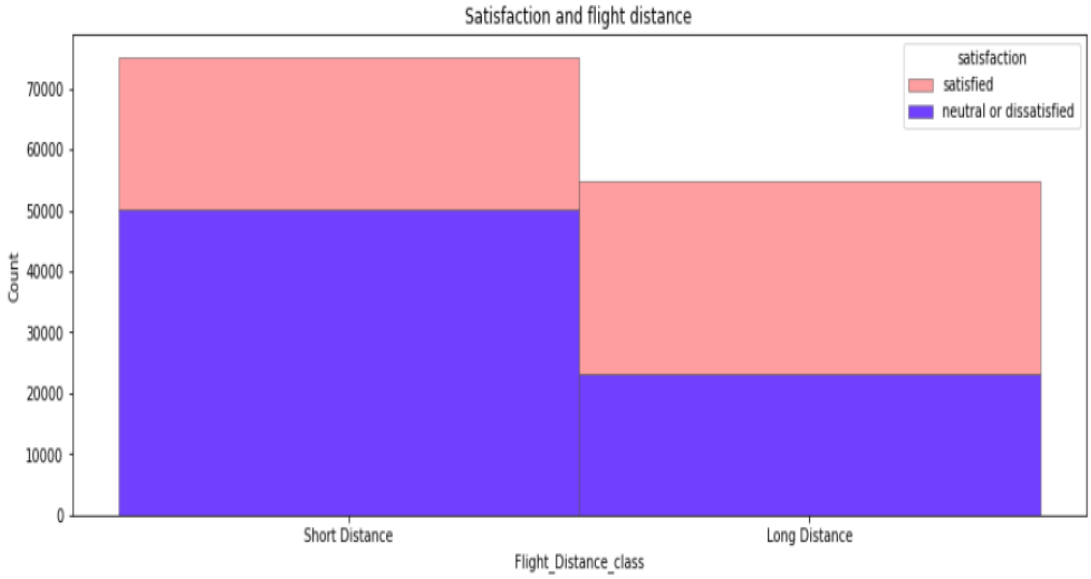


Fig 4.7 : Flight Distance vs Satisfaction

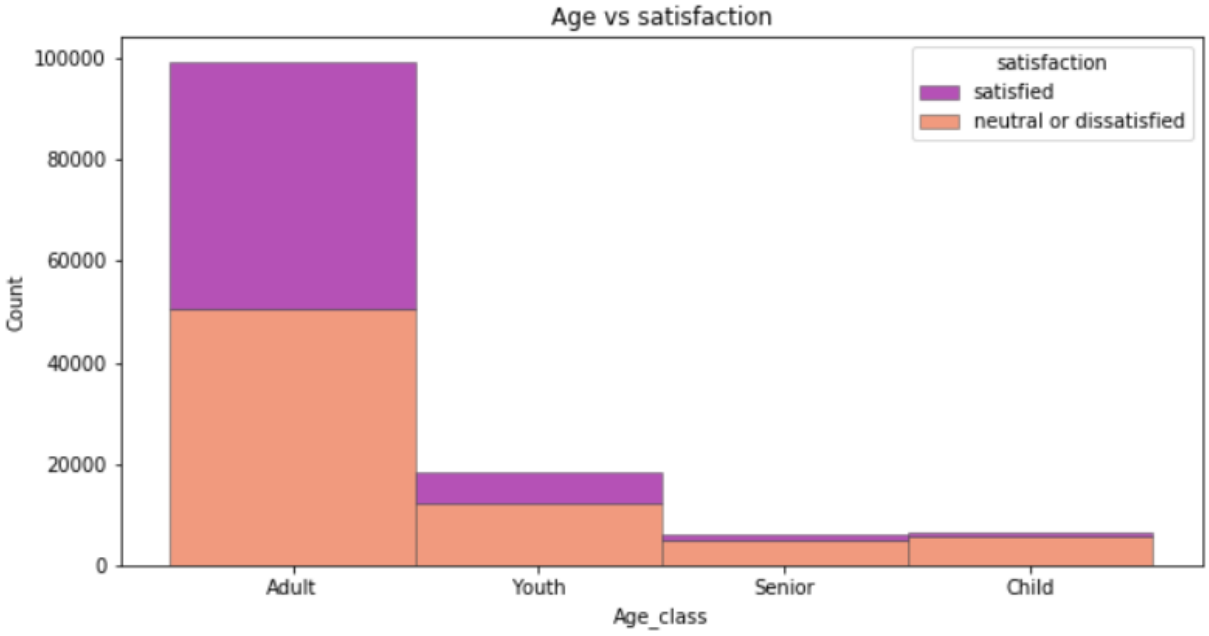


Fig 4.8 : Age groups vs Satisfaction

## 4.3 Multivariate analysis

Correlation Matrix

A Correlation matrix was plotted between the features in the dataset.

From the matrix it was found that the following Feature variables are least correlated with the target ‘satisfaction’:

1. Gender
2. Departure/Arrival Time convenient
3. Gate location
4. Departure/Arrival delay in minutes

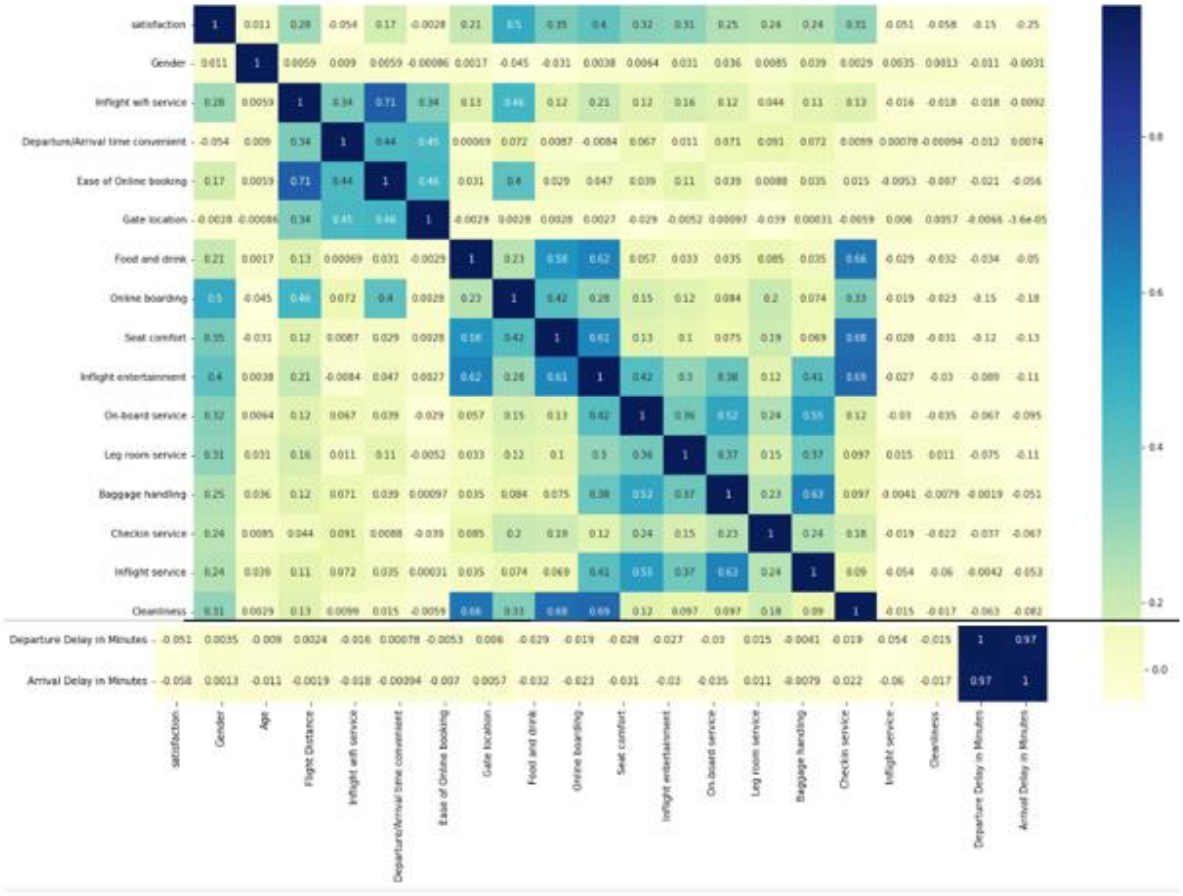


Fig 4.9 : Correlation Matrix

# 5. Preprocessing

Exploratory Data Analysis gave certain insights about all the feature columns and the target of our dataset. Next step was data preprocessing. Initially, as part of the preprocessing, the ‘id’ column was dropped from the data, since it had no meaningful relation with the target. Also, the data points ‘Eco’ and ‘Eco Plus’ in the column ‘Class’ were combined (i.e ‘Eco Plus was renamed as ‘Eco’).

## 5.1 Missing value Handling

All the columns were checked for missing/null values. It was found that the column ‘Arrival Delay in Minutes’ contained 393 missing values. To fill this column we used the values from the corresponding ‘Departure Delay in Minutes’ column since both Arrival Delay / Departure Delay in minutes have high positive linear correlation.

Out of the 14 survey results columns, all columns except Baggage Handling had ‘0’ values. These needed to be handled since they were values where the customer had not responded or rated. These values were replaced / filled with the most rated value when the customer was satisfied or dissatisfied.

## 5.2 Outlier Handling

In order to handle the outliers, at first, Boxplot of numerical columns such as 'Age', 'Flight Distance', 'Departure Delay in Minutes' and 'Arrival Delay in Minutes' are plotted. From the plots we came to an understanding that ‘Age’ doesn’t have any outliers but 'Flight Distance', 'Departure Delay in Minutes' and 'Arrival Delay in Minutes' have shown the presence of outliers.

The number of outliers detected in ‘Flight Distance’ is 2855. This data cannot be

treated as outliers and removed because they are a lot in number and moreover deletion of such columns may result in losing of some additional valid data. This is the same in the case of Departure/Arrival Delay in Minutes of the flights. The delay time is independent of the scheduled flight. Hence dropping these many columns is not possible.

## 5.3 Feature Engineering

In this step features like Flight Distance and Arrival Delay in minutes were changed. All values in Arrival Delay in Minutes were grouped into two, 'No Delay' for '0' values and 'Delayed' for values greater than '0'. Flight Distance was grouped to two 'Long Distance' and 'Short Distance'.

## 5.4 Encoding

We have seen that the dataset contains 5 columns with object data type. These columns [Satisfaction, Gender, Customer type, Type of travel & Class, Flight Distance and Arrival Delay in Minutes] were ***label encoded*** and converted to numerical ones.

After encoding a correlation matrix was plotted and we found that columns Gender, Gate location and Departure delay in minutes show least correlation, hence these columns were dropped.

| 0 | 1 |
|---|---|
| ‘neutral or dissatisfied’ | ‘satisfied’ |
| ‘Female’ | ‘Male’ |
| ‘First time user’ | ‘Returning’ |
| ‘Business Travel’ | ‘Personal Travel’ |
| ‘Business’ | ‘Eco’ |
| ‘Long Distance’ | ‘Short Distance’ |
| ‘Delayed’ | ‘No Delay’ |

Table 5.1: Values of Object Data Type columns after Label Encoding

## 5.5 Scaling

We performed scaling techniques such as standard scaling, minmax scaling, and normalization on the numerical features such as Flight distance, Age and Arrival delay in Minutes. These features were the ones where the values had a higher range than all the other feature columns and needed to be scaled down. As part of this process we tried training our data using the machine learning models such as logistic regression, kNN, random forest and decision tree.

When considering the Accuracy scores without scaling and with standard scaling, normalization, and minmax scaling we came to a conclusion that normalization, min max and standard scaling suits our data when using the different distance based algorithms and the model which produces maximum

accuracy is random forest which is not a distance based algorithm.

| | Accuracy Score |
|---|---|
| **Without Scaling** | 0.891554 |
| **MinMax Scaling** | 0.891669 |
| **Std Scaling** | 0.891592 |
| **Normalization** | 0.890668 |

Table 5.2 : Logistic Regression Model Accuracy Score Values

| | Accuracy Score |
|---|---|
| **Without Scaling** | 0.916423 |
| **MinMax Scaling** | 0.930590 |
| **Std Scaling** | 0.930551 |
| **Normalization** | 0.930012 |

Table 5.3 : kNN Model Accuracy Score Values

| | Accuracy Score |
|---|---|
| **Decision Tree** | 0.941061 |
| **Random Forest** | 0.955228 |

Table 5.4: Decision Tree & Random Forest Model Accuracy Score Values

## 5.6 Dimensionality reduction using PCA

Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset. This is accomplished by linearly transforming the data into a new coordinate system where (most of) the variation in the data can be described with fewer dimensions than the initial data. Many studies use the first two principal components in order to plot the data in two dimensions and to visually identify clusters of closely related data points. Principal component analysis has applications in many fields such as population genetics, microbiome studies, and atmospheric science.

We tried the PCA technique on our dataset after standard scaling and applied the machine learning models such logistic regression, kNN, Decision Tree and Random Forest. The results showed that the accuracy score of all the models reduced a bit. This indicated that the PCA technique was not suitable for our data set.

| | Accuracy Score |
|---|---|
| **Logistic Regression** | 0.868571 |
| **kNN** | 0.924738 |
| **Decision Tree** | 0.883546 |
| **Random Forest** | 0.926124 |

Table 5.5: Accuracy Score Values after applying PCA.

## 5.7 Cross Validation

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is

any of various similar model validation techniques for assessing how the results of a statistical analysis will generalise to an independent data set. Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of *known data* on which training is run (*training dataset*), and a dataset of *unknown data* (or *first seen* data) against which the model is tested (called the validation dataset or *testing set*). The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalise to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

We tried using stratified k-Fold Validation on our dataset. The mean accuracy score values after the cross validation is as follows:

| | Accuracy Score |
|---|---|
| **Logistic Regression** | 0.823637 |
| **kNN** | 0.837119 |
| **Decision Tree** | 0.779720 |
| **Random Forest** | 0.815599 |

Table 5.6: Cross validation scores for different models

The results of cross validation indicated that kNN, logistic regression and random forest gave the maximum accuracy. After analysing the time taken to

run and the accuracy scores, we preferred  Random Forest Classifier to build our model.

# 6. Model Building and Website Hosting

## 6.1 Machine Learning Models

A machine learning model is defined as a mathematical representation of the output of the training process. Machine learning is the study of different algorithms that can improve automatically through experience & old data and build the model. A machine learning model is similar to computer software designed to recognize patterns or behaviours based on previous experience or data. The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on new data. Based on different business goals and data sets, there are three learning models for algorithms. Each machine learning algorithm settles into one of the three models:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Supervised Learning is further divided into two categories:

1. Classification
2. Regression

Unsupervised Learning is also divided into below categories:

- Clustering

- Association Rule
- Dimensionality Reduction

### 6.1.1 Classification in Machine Learning

Classification is a supervised machine learning process that involves predicting the class of given data points. Those classes can be targets, labels or categories. For example, a spam detection machine learning algorithm would aim to classify emails as either “spam” or “not spam.”

Common classification algorithms are as follows:

**Logistic Regression**

Logistic Regression utilises the power of regression to do classification and has been doing so exceedingly well for several decades now, to remain amongst the most popular models. One of the main reasons for the model’s success is its power of explainability i.e. calling-out the contribution of individual predictors, quantitatively. Unlike regression which uses Least Squares, the model uses Maximum Likelihood to fit a sigmoid-curve on the target variable distribution. Given the model’s susceptibility to multicollinearity, applying it stepwise turns out to be a better approach in finalising the chosen predictors of the model. The algorithm is a popular choice in many natural language processing tasks e.g. toxic speech detection, topic classification, etc.

Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic

regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1*X_1 + \ldots + B_k*K_k$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1.

There are three types of logistic regression models, which are defined based on categorical response.

- Binary logistic regression: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). Some popular examples of its use include predicting if an e-mail is spam or not spam or if a tumor is malignant or not malignant. Within logistic

regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.

- Multinomial logistic regression: In this type of logistic regression model, the dependent variable has three or more possible outcomes; however, these values have no specified order. For example, movie studios want to predict what genre of film a moviegoer is likely to see to market films more effectively. A multinomial logistic regression model can help the studio to determine the strength of influence a person's age, gender, and dating status may have on the type of film that they prefer. The studio can then orient an advertising campaign of a specific movie toward a group of people likely to go see it.
- Ordinal logistic regression: This type of logistic regression model is leveraged when the response variable has three or more possible outcome, but in this case, these values do have a defined order. Examples of ordinal responses include grading scales from A to F or rating scales from 1 to 5.

**Decision Tree**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions

and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

**Strengths and Weaknesses of the Decision Tree approach**

The strengths of decision tree methods are:

- Decision trees are able to generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

The weaknesses of decision tree methods :

- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision trees are prone to errors in classification problems with many classes and a relatively small number of training examples.
- Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

**Random Forest**

A Random Forest is a reliable ensemble of multiple Decision Trees (or CARTs); though more popular for classification, than regression applications. Here, the individual trees are built via bagging (i.e. aggregation of bootstraps which are nothing but multiple train datasets created via sampling of records with replacement) and split using fewer features. The resulting diverse forest of uncorrelated trees exhibits reduced variance; therefore, is more robust towards change in data and carries its prediction accuracy to new data. However, the algorithm does not work well for datasets having a lot of outliers, something which needs addressing prior to the model building.

**K-Nearest Neighbors**

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.Similar data points are close to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.

The straight-line distance (also called the Euclidean distance) is a popular and familiar choice for calculating distance.

The kNN Algorithm:

1. Load the data.
2. Initialize K to your chosen number of neighbors
3. For each example in the data

3.1 Calculate the distance between the query example and the current example from the data.

3.2 Add the distance and the index of the example to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances

5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels

### 6.2 Model building

In this project we are dealing with Classification models in Supervised Learning.

The preprocessed data was fitted and trained using the following classification models: Logistic Regression, kNN, Random Forest and Decision Tree. Model performances were measured by finding the accuracy scores. Random Forest Classifier gave the highest accuracy score of 0.95528.

Hence Random Forest was selected to build the final model.

### 6.3 Web Application

Based on the Machine Learning predictive model, a web application was developed using Python Flask Framework. In the application, the users can enter the passenger details, flight details and the passenger experience ratings and get the prediction for passenger satisfaction.

The application is hosted using PythonAnywhere web hosting service under the url http://dsaictakprojectb1t3.pythonanywhere.com/

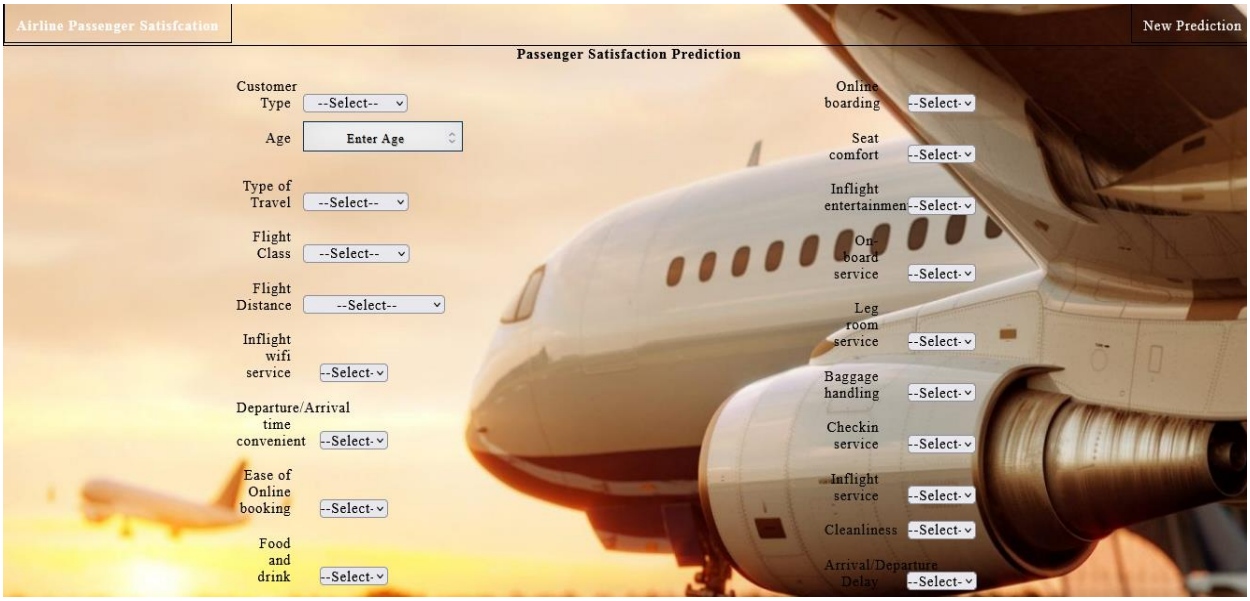


Fig 6.1 Screenshot of website hosted

# 7. Result

Random Forest model was the best considering time taken and accuracy score. The final model based on Random Forest Classifier gave the following model performances:

| | Score |
|---|---|
| **Accuracy Score** | 0.955266 |
| **Precision Score** | 0.959418 |
| **Recall Score** | 0.937059 |

Table 7.1 Scores obtained by Random Forest Model

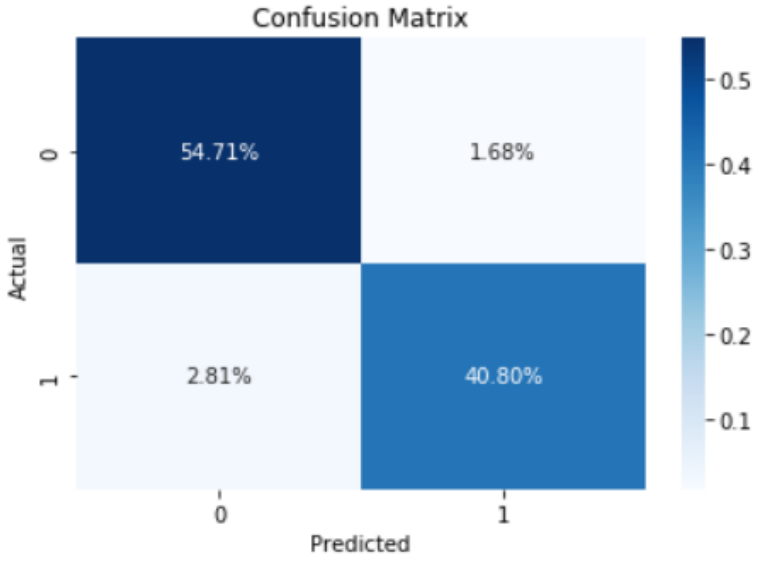


Fig. 7.1 Confusion Matrix - Random Forest Model

Feature importance from the  Random Forest was calculated and the following result was obtained:

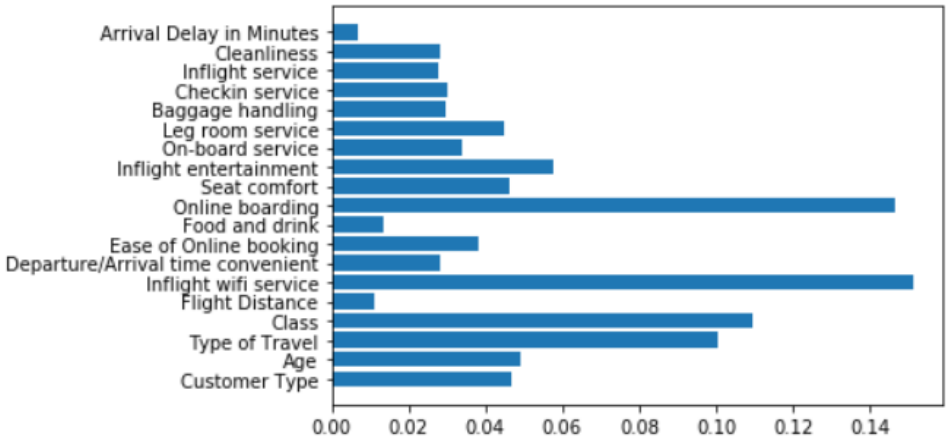


Fig. 7.2 Feature Importance - Random Forest

It was observed that features such as Online Boarding, Inflight Wifi Service, and the Flying class were important features and they influenced the target satisfaction more.

# 8. Conclusion

A classification model to predict airline passenger satisfaction was built, which would help airline companies to recognise critical bottlenecks and improve passenger satisfaction.

In-Flight Wi-Fi Service, Inflight Entertainment and Online boarding were found to influence greatly on passenger satisfaction. Travel class also had significant influence on satisfaction: Business class travellers tend to be more satisfied compared to Economy class passengers

Therefore, airline companies may focus on providing better services to economy class passengers as well, like improving leg room service and seat comfort.

# References

[1] Khodijah Hulliyah , Husni Teja Sukmana, Predicting Airline Passenger Satisfaction with Classification Algorithms, International Journal of Informatics and Information System, 2021

[2] Xuchu Jiang, Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model, ResearchGate, 2020

[3] Michel Bierlairea, Mohammed Elshafie, A systematic review of machine learning classification methodologies for modelling passenger mode choice,Journal of Choice Modelling

[4] Classification Models in Machine Learning | Classification Models (analyticsvidhya.com)

[5] Mohammed Arif, Aman Gupta, Aled Williams,Customer service in the aviation industry – An exploratory analysis of UAE airports,Journal of Air Transport Management ,Volume 32,2013,Pages 1-7,ISSN 0969-6997

[6] R. Archana, M.V. Subha, A study on service quality and passenger satisfaction on,Indian airlines, Int. J. Multidisc. Res. 2 (2) (2012)

[7] Martel,Nathalie,Seneviratne, Prianka.N, Analysis of factors influencing quality of service in passenger terminal buildings, 1990

[8] V. García , R. Florencia-Juárez , J. P. Sánchez-Solís, Predicting Airline Customer Satisfaction using k-nn Ensemble Regression Models, Research in Computing Science 148(6), 2019

[9] H. Song, W. Ruan, Y. Park, Effects of service quality, corporate image, and customer trust on the corporate reputation of airlines, Sustainability 11(12),(2019) 3302.

Data source:

- https://www.kaggle.com/datasets/johnddddd/customer-satisfaction
- https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/code?datasetId=522275