Aircraft Passenger Satisfaction

Group 12
Student 1 Hanisha Kasu
Student 2 Phanindra Raja Varma Gadiraju

Northeastern University

860-987-3352 (Tel of Student 1)
857-206-1166(Tel of Student 2

kasu.h@northeastern.edu
gadiraju.p@northeastern.edu

Percentage of Effort Contributed by student 1: 50%
Percentage of Effort Contributed by student 2: 50%

Signature of Student 1: Hanisha Kasu
Signature of Student 2: Phanindra Raja Varma Gadiraju
Submission Date: 12/09/2022

Problem setting:

Airlines are one of the most important means of transportation. The task of running flights is a costly affair for the airline companies. Customers' satisfaction is the key for any company's growth. In this project, we are going to analyze the key variables responsible for customer satisfaction and predict the ways and means of increasing their sales based on customer satisfaction. The objective of this project is to predict the satisfaction of an airline customer using the given variables.

Problem Definition:

The primary objective of this project is to predict Airline Satisfaction using ML techniques and algorithms to predict the satisfaction of a customer. Accuracy of each model is compared with the baseline model and the most accurate model is selected.

The following steps are followed to carry out the prediction:

- Data Cleaning and Exploratory Data Analysis
- Data Partitioning into Train and Test Datasets
- Develop a baseline model
- Calculate results for various models to determine the most apt model.

Data source:

-Dataset

Data Description:

This data set consists of 129880 number of rows of Instances and 25 columns as Attributes.

There are 5 numerical and 20 categorical variables in this dataset. The Response Variable in this dataset is a Categorical Variable (has two categories: 1 and 0), describing customer satisfaction.

The variables which we would be using are:

Age, Gender, Delay in minutes, In-flight Service, Class, In-flight WIFI Service, Cleanliness, Seat, Comfort, Inflight Entertainment, Leg Room Service.

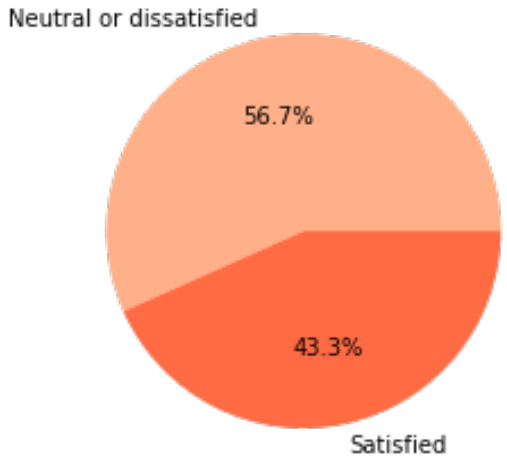
Data Cleaning:

In the Data Cleaning, the dataset has been checked for null values and duplicate values.

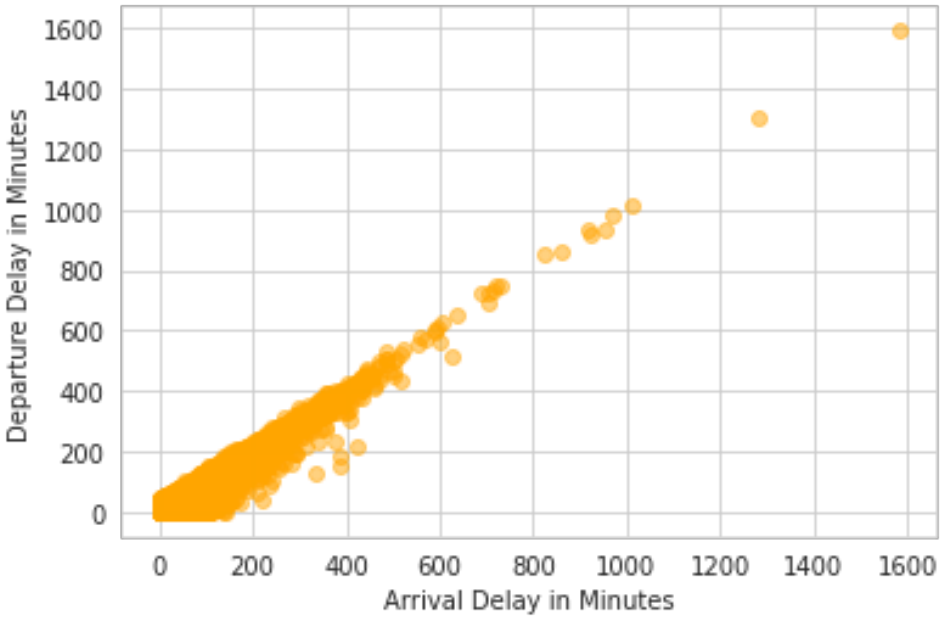
Later, duplicate records have also been removed.

Data Exploration:

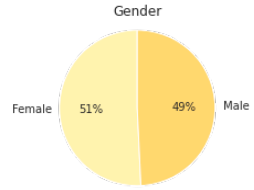
Firstly, we wanted to see the Total Satisfied Vs Dissatisfied Percentage in the Target Variable.

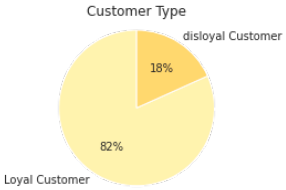


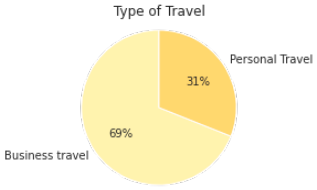
Nextly, we are trying to draw a relation between Arrival Delay in Minutes and Departure Delay in Minutes using a Scatter Plot.

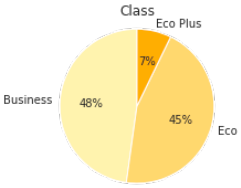


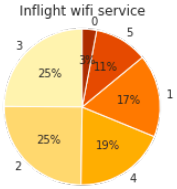
Here, we are trying to find out the percentage of each category in all the Categorical Predictors present in the dataset. Most of the deciding factors among the variables have been studied for a better understanding of the dataset based on the count of each category's occurrence.

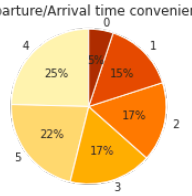


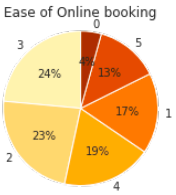


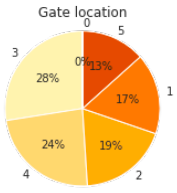


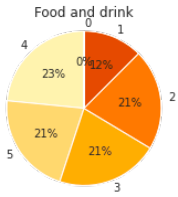


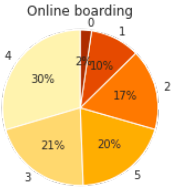


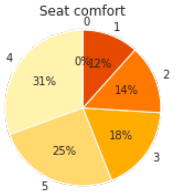


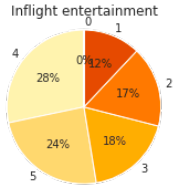


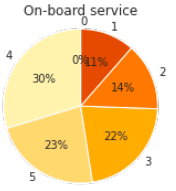


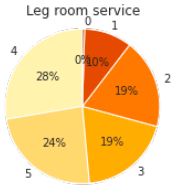


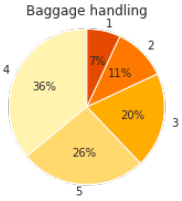


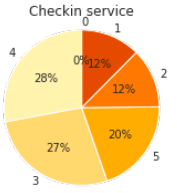


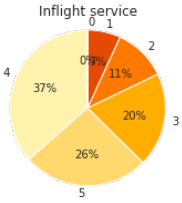


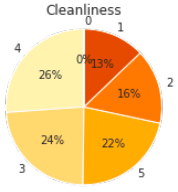




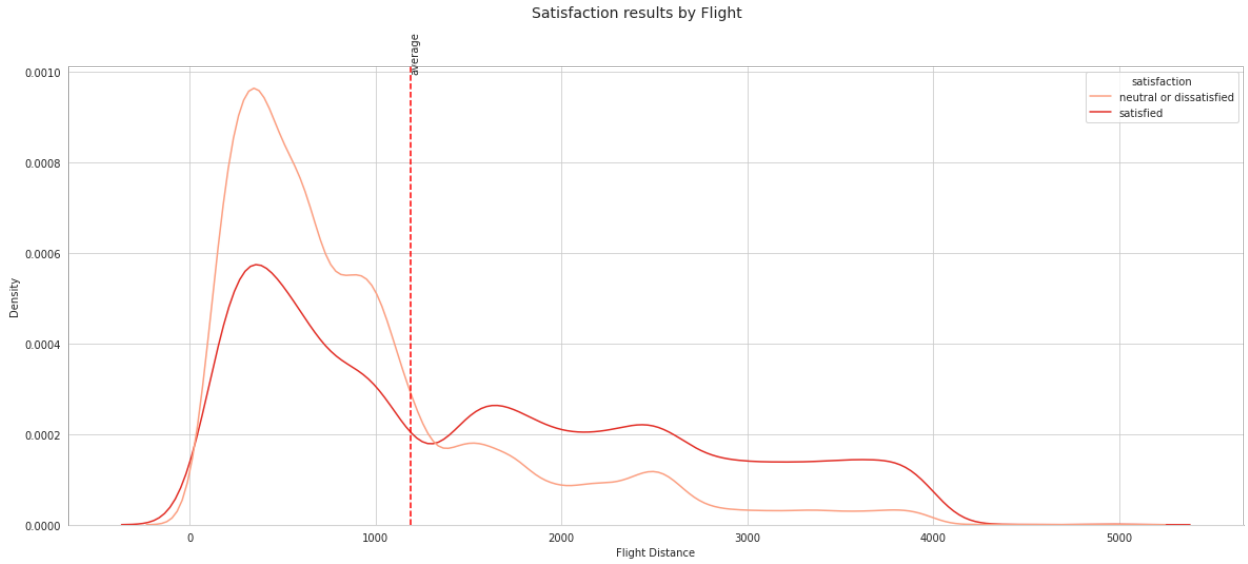






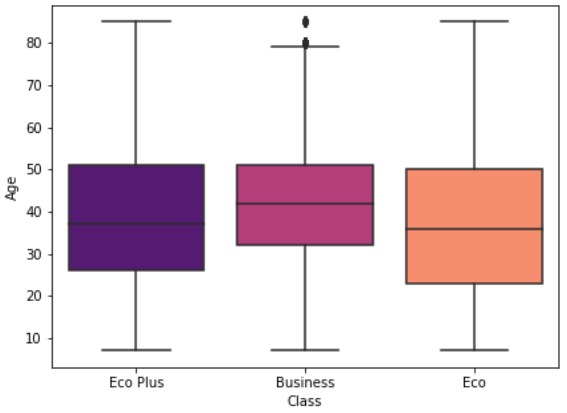


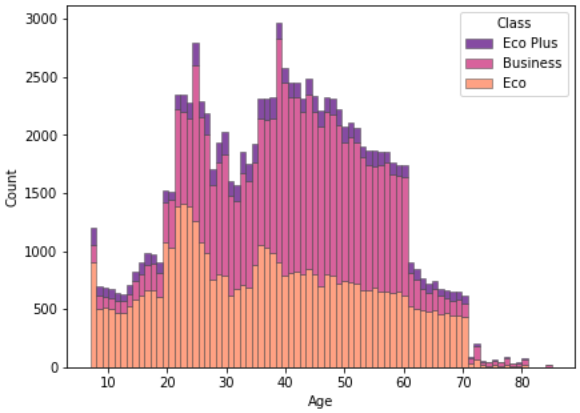
Flight Distance vs Satisfaction:

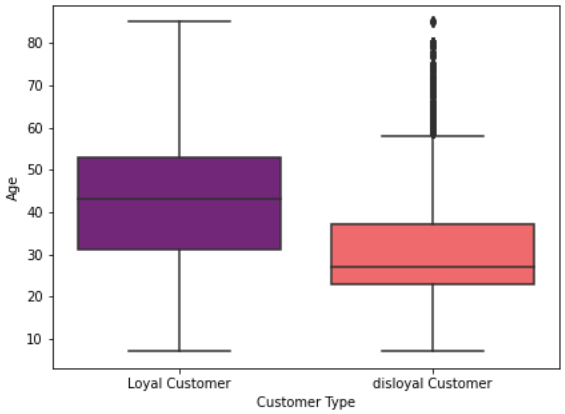


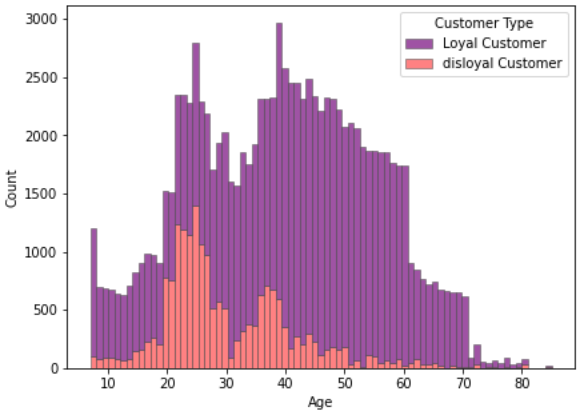
It has been observed that mostly customers are dissatisfied as the distance crosses 1000miles. So, maybe people are usually dissatisfied due to the hassle they face in long or international journeys that cover more distances.

Distribution Plots of Class vs Age of the customer









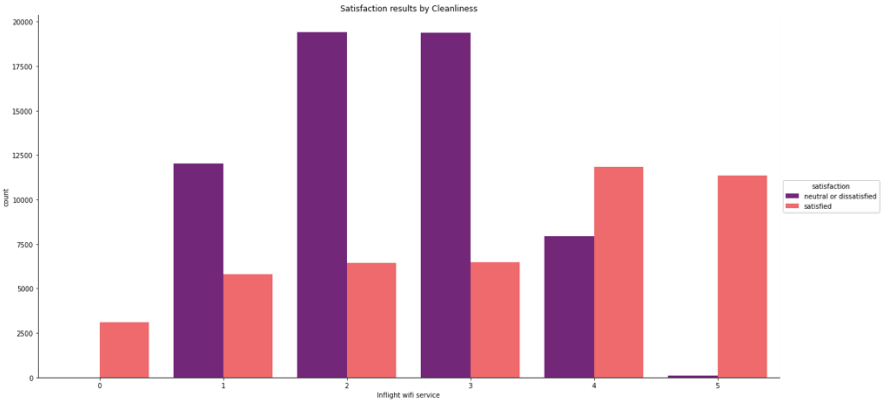
It is understood that most of the Business Class Customers belong to the age groups of early 30s and mid 50. It can be observed that their count is also phenomenally more than the other two classes.

It is observed that Loyal Customers age group is from early 30s to mid 50s. It is also observed that the count of disloyal customers is very less compared to Loyal Customers.

Flight Arrival Time Delay vs Flight Departure Time Delay.

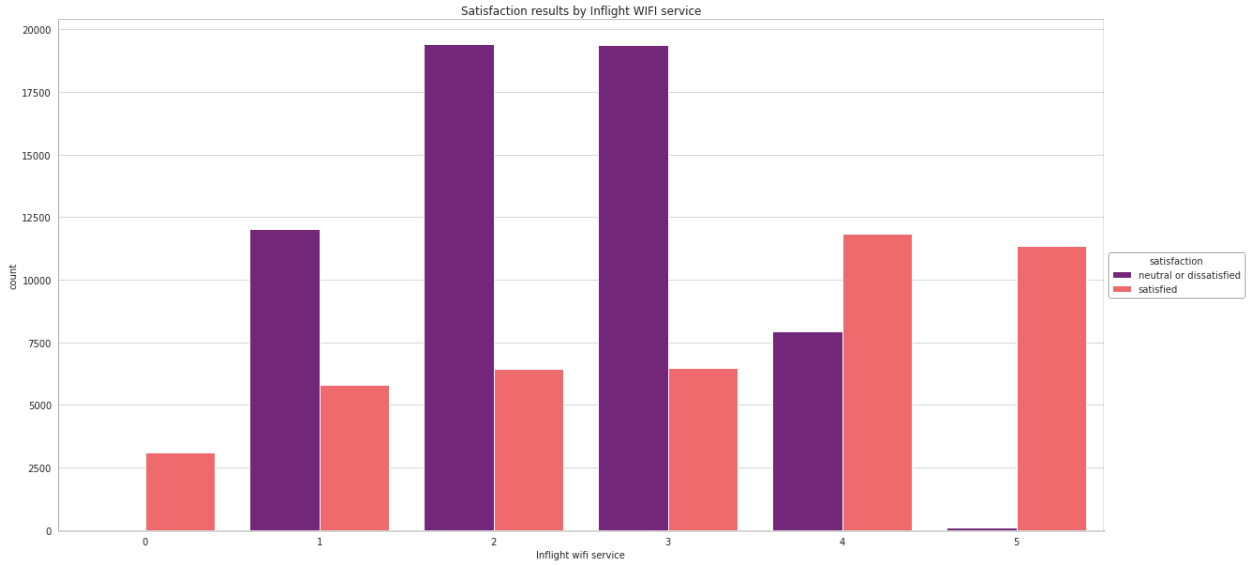
Customer Satisfaction Based on Various Categories:

Satisfaction vs Cleanliness:



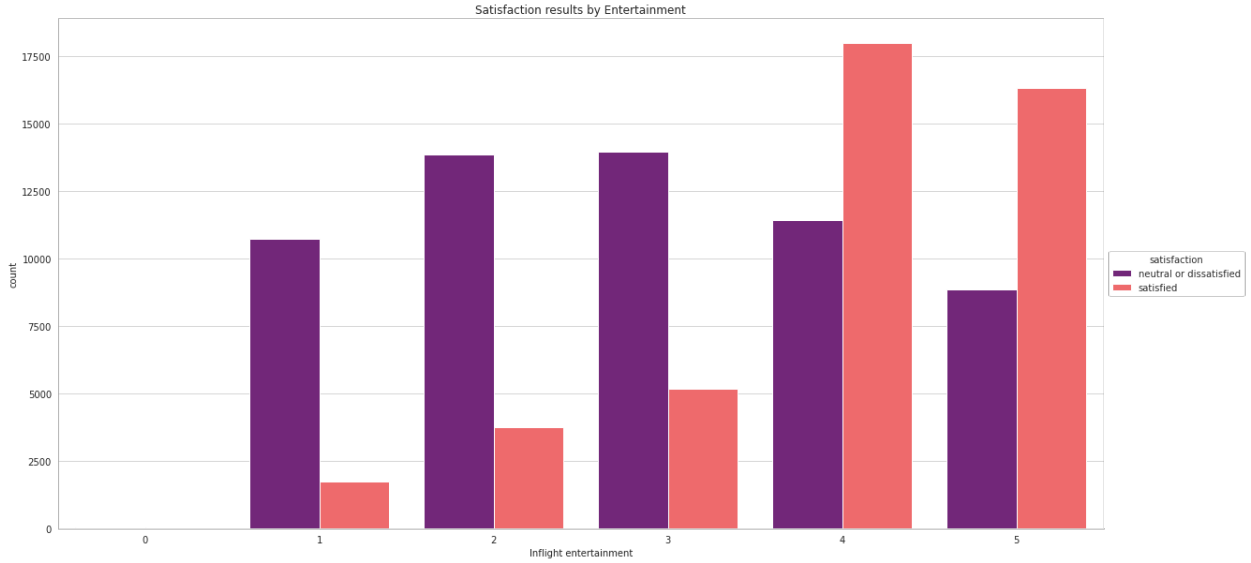
It can be observed that the satisfaction measure increases as cleanliness of the flight increases and is at peak at 4 level.

Satisfaction vs Inflight WIFI Service



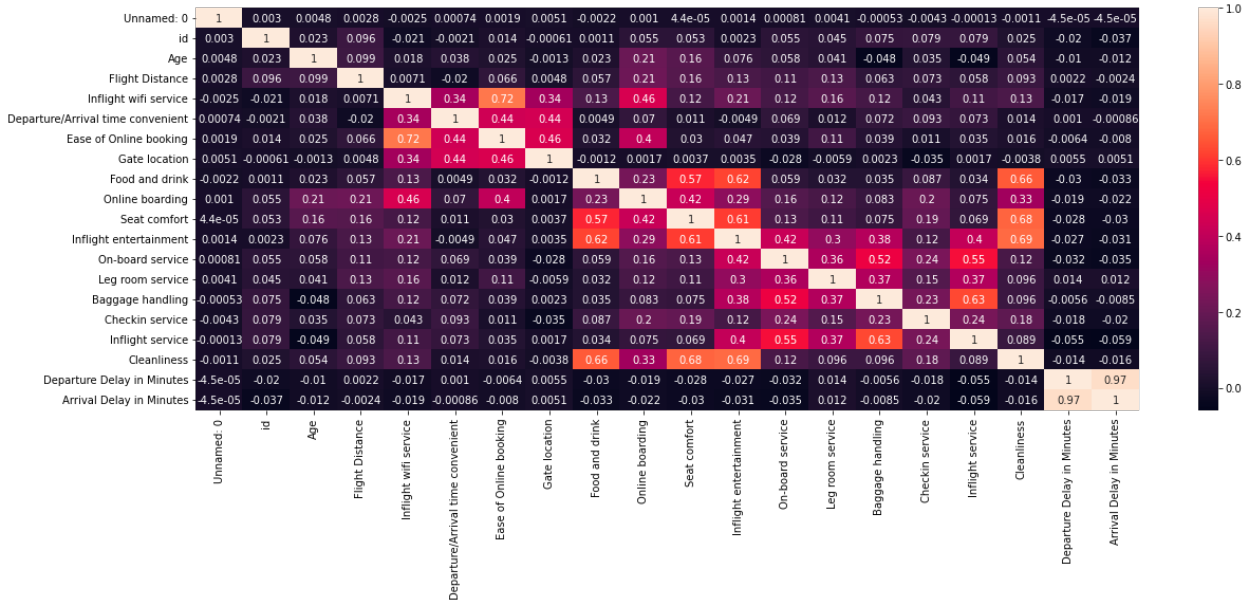
It can be observed that most of the customers are highly dissatisfied at 3 or 4 WIFI strength also.

Satisfaction based on In-flight Entertainment:



It can be observed that in-flight entertainment contributes to be one of the major factors as it follows an increasing trend.

Correlation Plot



It can be observed that most of the variables have a positive correlation. Similarly, Arrival in Delay and Departure in Delay has a positive correlation of nearly 0.97. So, maybe taking one column into consideration would make sense.

Data Mining Tasks

Data Reduction:

Using the domain knowledge, we would like to drop the first column ‘ID’ as it would be of no use for our analysis on prediction of our model.

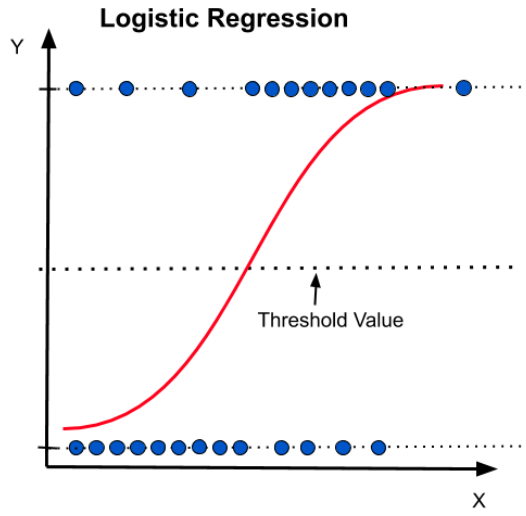
As there are no null values or missing values, there is not much use in performing Data Imputation.

Model Selection:

One of the most crucial steps in Data Mining Cycle is selecting the model which gives the highest accuracy among the trained models.

In our project, we have trained different models such as Logistic Regression, Random Forest Classifier and Naïve Bayes Classifier to classify the airline customer satisfaction.

Logistic Regression:



Logistic Regression is one of the most used supervised Machine Learning techniques.

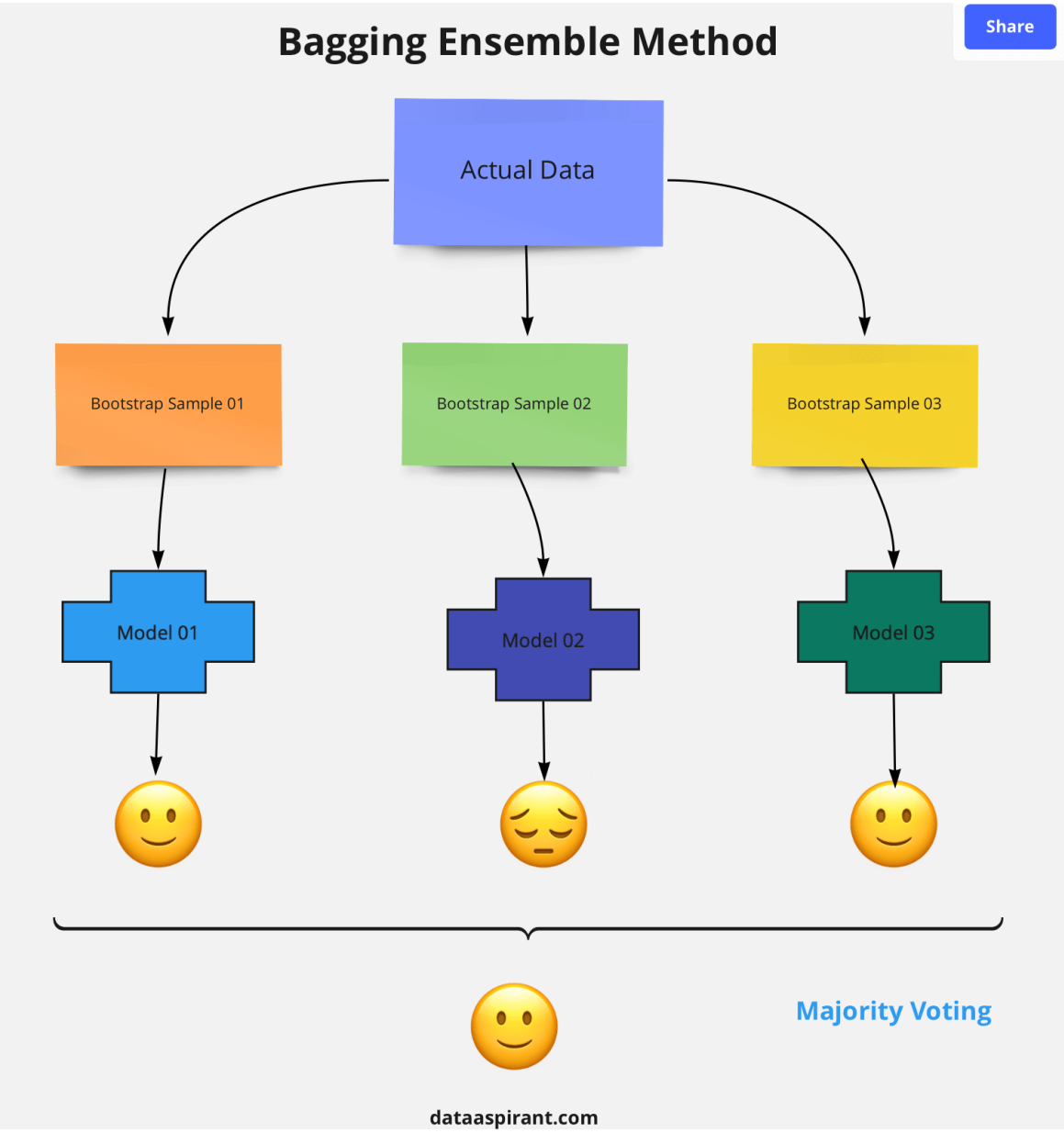
Logistic regression was performed, and the model classified the target variable as satisfied or dissatisfied.

Random Forest Classifier:

Random forest algorithm is basically a set of muliptle Decision tress working together. The original dataset is divided into multiple subsets. Each of these multiple subsets have same number of rows as the original dataset. To achieve the above-mentioned point, we sample the data with replacement which means after picking a datapoint, we are putting it back. This process of creating a series of datasets is called bootstrapping.

With the bootstrapped datasets we run the decision tress independently and consider the majority votes as depicted in the above figure.

The limitation is that this Random Forest Classifier should work better than a Random Classifier.

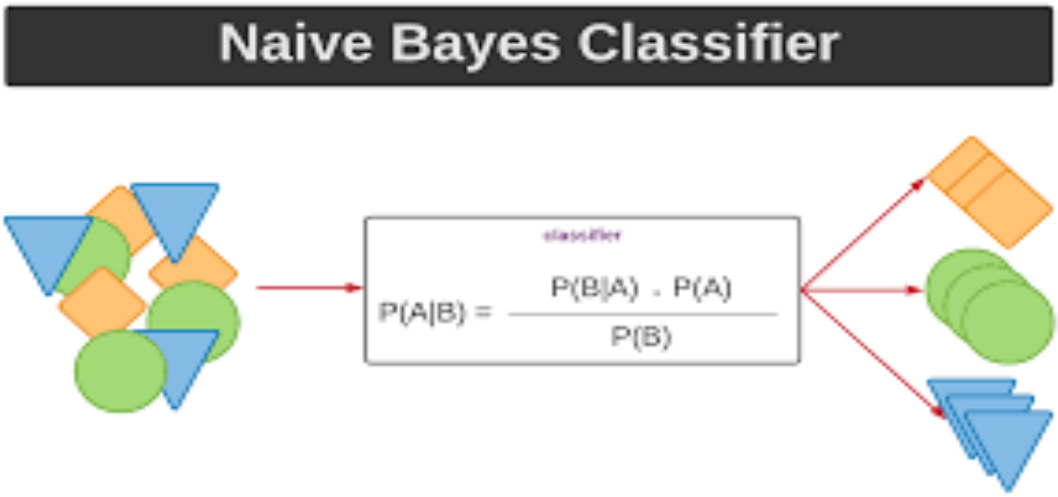


Naïve Bayes Classifier

Naïve Bayes Classifier is a powerful Machine Learning algorithm for the classification task. This classifier uses Bayes Theorem.

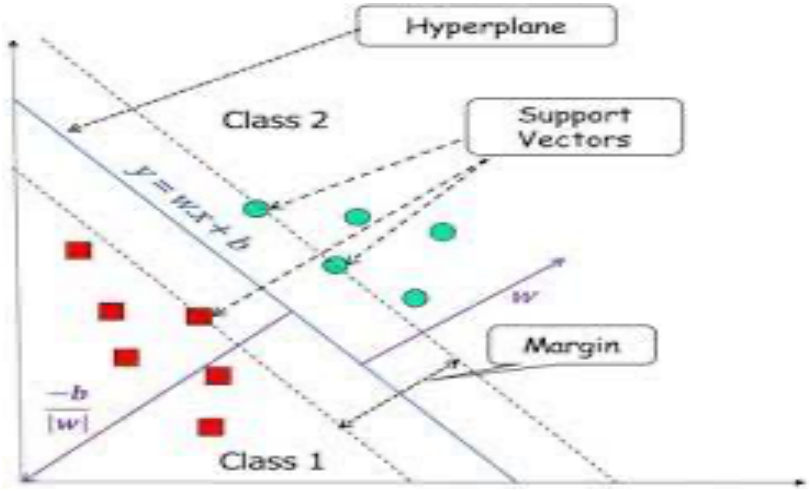
This classifier uses Bayes Theorem.

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

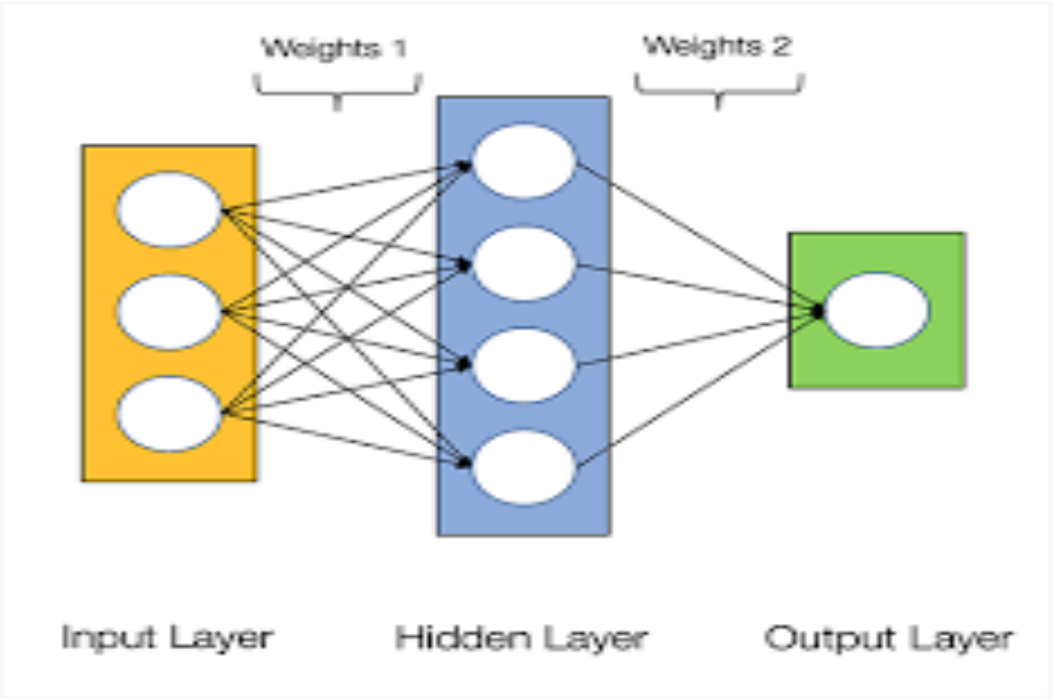


SVM:

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.



Neural Networks:



In Neural Networks, there are Input Layer, Output Layer and Hidden Layers. Accuracy of this model is determined by the number of Hidden Layers and the nodes.

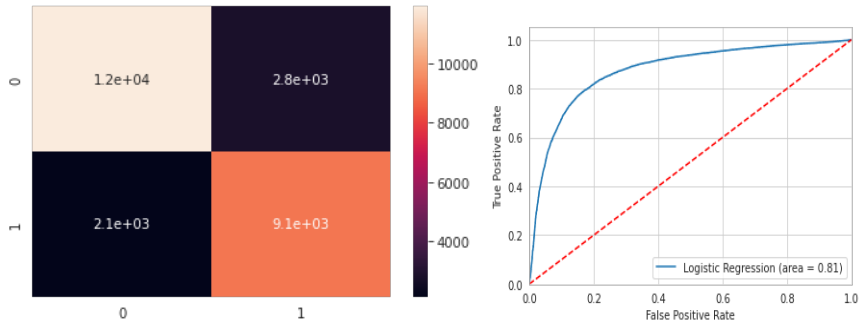
Performance Evaluation:

Logistic Regression:

An accuracy score of 0.81 was achieved while using Logistic Regression Model.

The ROC curve of the Logistic Regression model is as above.

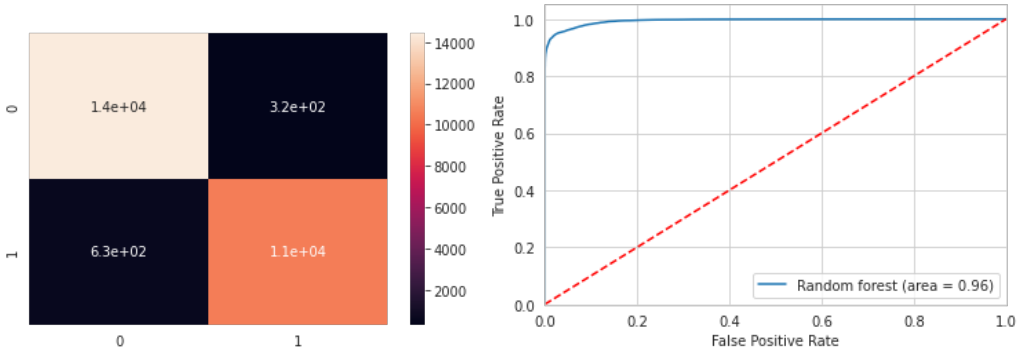
The confusion matrix is as below.



Random Forest Classifier:

We have achieved an accuracy score of 0.96 upon the implementation of Random Forest Classifier.

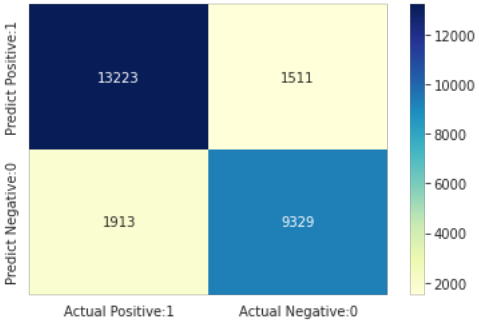
The confusion matrix is as below



Naïve Bayes Classifier:

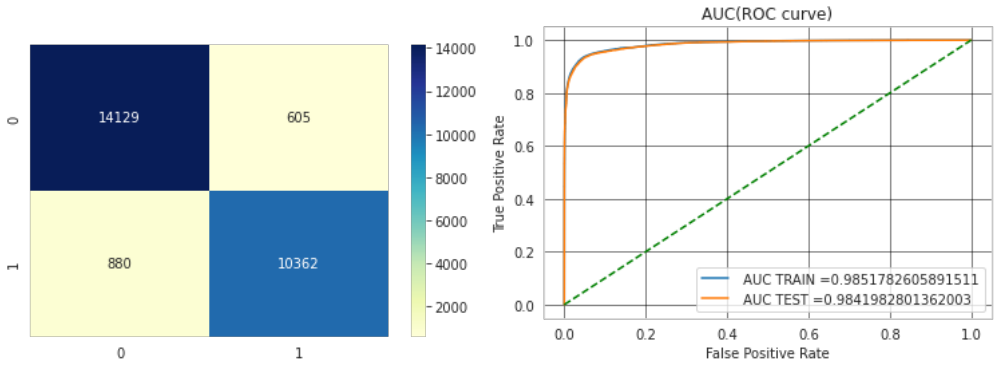
We have achieved an accuracy score of 0.86.

The confusion matrix is as below



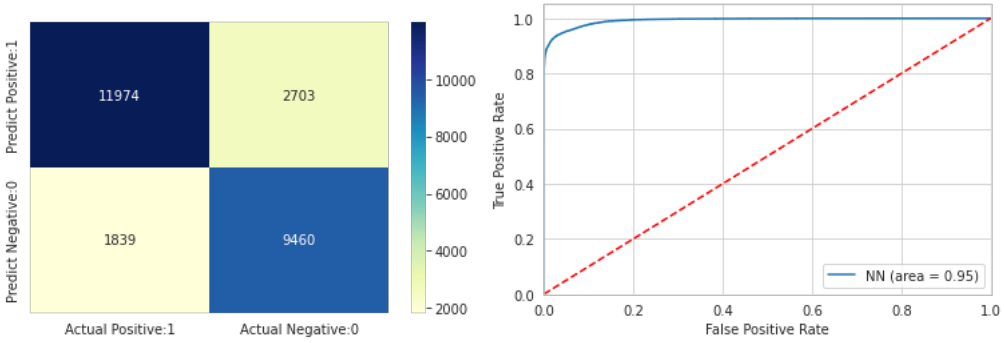
SVM:

We have achieved an accuracy score of 0.94



Neural Networks:

We have achieved an accuracy score of 0.96



Project Results:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression Class 1	0.81	0.76	0.81	0.79
Random Forest Class 1	0.96	0.97	0.94	0.96
Naïve Baye's Class 1	0.87	0.86	0.83	0.84
SVM Class 1	0.94	0.94	0.92	0.93
Neural Networks Class 1	0.96	0.96	0.94	0.95

As of now, we found that Random Forest Classifier has the highest accuracy of 0.96. We will be implementing more models and selecting the best classifier based on the accuracy score.

Impact of Project Outcomes:

Both Random Forest Classifier and Neural Networks to have an accuracy of 0.96. As, Random Forest Classifier had lesser runtime among these two, we chose Random Forest Classifier to be the most accurate for out dataset.

Our models can predict the satisfaction of a customer based on the given predictors which are very much useful to the aircraft companies.

These results can help the companies to exactly figure out which areas need improvement to increase customer satisfaction which is the key for any industry's growth.