GROUP 3

Capstone Project – Final Report

Airline Passenger Satisfaction Prediction

Mentor: Srikar Muppidi

- **Akshaya ES**
- **Nikita Ann George**
- **Pala Sheshu**
- **Ujjwal Kumar Sharma**
- **Venkat Paritala**
- **Viresh Raj Sah**

Contents

PROBLEM STATEMENT 3
PROJECT OUTCOME 3
BUSINESS IMPACT 3
DATA SET AND DOMAIN 4
DATA DESCRIPTION: 5
PROJECT METHODOLOGY 9
PRE-PROCESSING DATA ANALYSIS 12
EXPLORATORY DATA ANALYSIS & BUSINESS INSIGHTS 15
Univariate Analysis 15
Bivariate Analysis 18
Inference From Statistical Tests: 24
MODEL BUILDING: 25
Understanding The Metrics: 25
BUSINESS IMPACT 26
Preparing the data for model building: 26
Encoding 26
Train-Test Split 27
BASE MODEL 28
Logistic regression: 28
Logistic regression after transformation: 29
Decision Tree Algorithm: 30
Tuned Decision Tree: 32
Random Forest Model: 33
Tuned Random Forest Model: 34
Boosting: 35
Model Comparison: 36
INFERENCES AND RECOMMENDATIONS 37
CONCLUSION 37
REFERENCES 38

PROBLEM STATEMENT

Predicting customer satisfaction based on their holistic experience while traveling by air.

PROJECT OUTCOME

For this project, we intend to explore survey results provided by passengers (anonymized), and understand what causes them to be satisfied or dissatisfied. Ultimately, we hope our findings enable us to make recommendations on how to address the factors that cause dissatisfaction.

BUSINESS IMPACT

Airline businesses around the world are decimated by Covid-19 as most international air travel has been grounded. In fact, some airlines such as Thai Airways have already filed for bankruptcy. Nonetheless, once the storm is over, demand for air travel is expected to surge as people rush back for overseas holidays. What can airlines prepare to give themselves a competitive edge when the crowd finally arrives?

Customer satisfaction is always top of mind for airlines. Unhappy or disengaged customers naturally mean fewer passengers and less revenue. It's important that customers have an excellent experience every time they travel. On-time flights, good in-flight entertainment, more (and better) snacks, and more legroom might be the obvious contributors to a good experience and more loyalty.

While we might hear about those aspects the most, the customer experience is not about just the flight itself. It's everything from purchasing the ticket on the company's website or mobile app to checking bags in at the airport or via a mobile app to waiting in the terminal. This mindset has been, and continues to be, adapted to the post-security, onboard, and post-flight experience. So how can we determine which of these factors contribute to the satisfaction of the customer?

To answer this, we intend on building a classification problem to predict the customer satisfaction

DATA SET AND DOMAIN

- A dataset is a collection of data, and it can be structured or unstructured.
- A structured data is represented in a tabular format, where every column of the table representsa particular variable, and each row corresponds to a given record of the dataset in question.
- Unsupervised data is not represented in a tabular form, data that we fetch from Facebook, Twitter, Netflix, etc. with the help of recommendation systems are all our unsupervised data.
- This dataset contains information about travelers of a certain airline, information about them in terms of their relationship with the airline (loyal/disloyal customer), details about their trip (distance, class, etc.) and takes their ratings on various aspects of their experience in their trip.
- With this information we also have information about if these travelers would rate their trip as 'Satisfactory' or 'neutral or unsatisfactory'. This dataset provides an interesting insight into customer satisfaction and aspects which affect it in terms of the aviation industry, which is notoriously competitive and heavily relies on customer satisfaction.

DATA DESCRIPTION:

The data set consists of 129880 observations and 25 features before the cleaning and contains information regarding the customers as well as their experience in the flight based off multiple factors such as their reviews for the food/drinks,seat comfort,inflight entertainment etc. The description for each variable along with the datatype as given in the dataset is as follows:

VARIABLE	DATATYPE	DESCRIPTION
Unnamed:0	numeric	Index value
1. Id	numeric	Unique identifier
2. Gender	object	Gender of the passengers (Female, Male)
3. Customer Type	object	The customer type (Loyal customer, disloyal customer)
4. Age	numerical	The actual age of the passengers
5. Type of Travel	object	Purpose of the flight of the passengers (Personal Travel, Business Travel)
6. Class	object	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
7. Flight distance	numeric	The flight distance of this journey
8. Inflight wifi service	numeric	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)

9.Departure/Arrival time convenient	numeric	Satisfaction level of Departure/Arrival time convenient
10. Ease of Online booking	numeric	Satisfaction level of online booking
11. Gate location	numeric	Satisfaction level of Gate location
12. Food and drink	numeric	Satisfaction level of Food and drink
14. Online boarding	numeric	Satisfaction level of online boarding
15. Seat comfort	numeric	Satisfaction level of Seat comfort
16. Inflight entertainment	numeric	Satisfaction level of inflight entertainment
17. On-board service	numeric	Satisfaction level of On-board service
18. Leg room service	numeric	Satisfaction level of Leg room service
19. Baggage handling	numeric	Satisfaction level of baggage handling
20. Check-in service	numeric	Satisfaction level of Check-in service

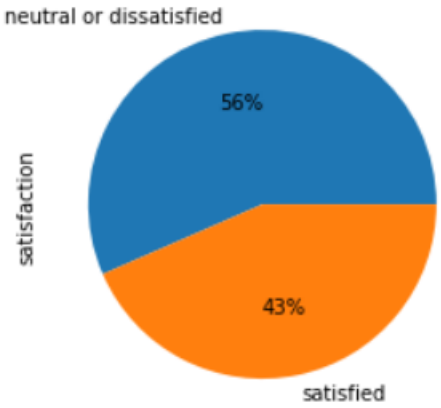
21. Inflight service	numeric	Satisfaction level of inflight service
22. Cleanliness	numeric	Satisfaction level of Cleanliness
23. Departure Delay in Minutes	numeric	Minutes delayed when departure
24. Arrival Delay in Minutes	numeric	Minutes delayed when Arrival
25. Satisfaction	object	Airline satisfaction level(Satisfaction, neutral, or dissatisfaction)

Data Types-independent variables:

- The satisfaction levels across columns like 'Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking','Gate location', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment, 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness' work on a rating system and hence can be classified as categorical data.
- The variables: Gender, Type of Travel, Customer Type and Class are also part of categorical data
- The variables: Age, Distance, Departure Delay in Minutes and Arrival Delay in Minutes are numerical

Target Variable: Satisfaction-Categorical

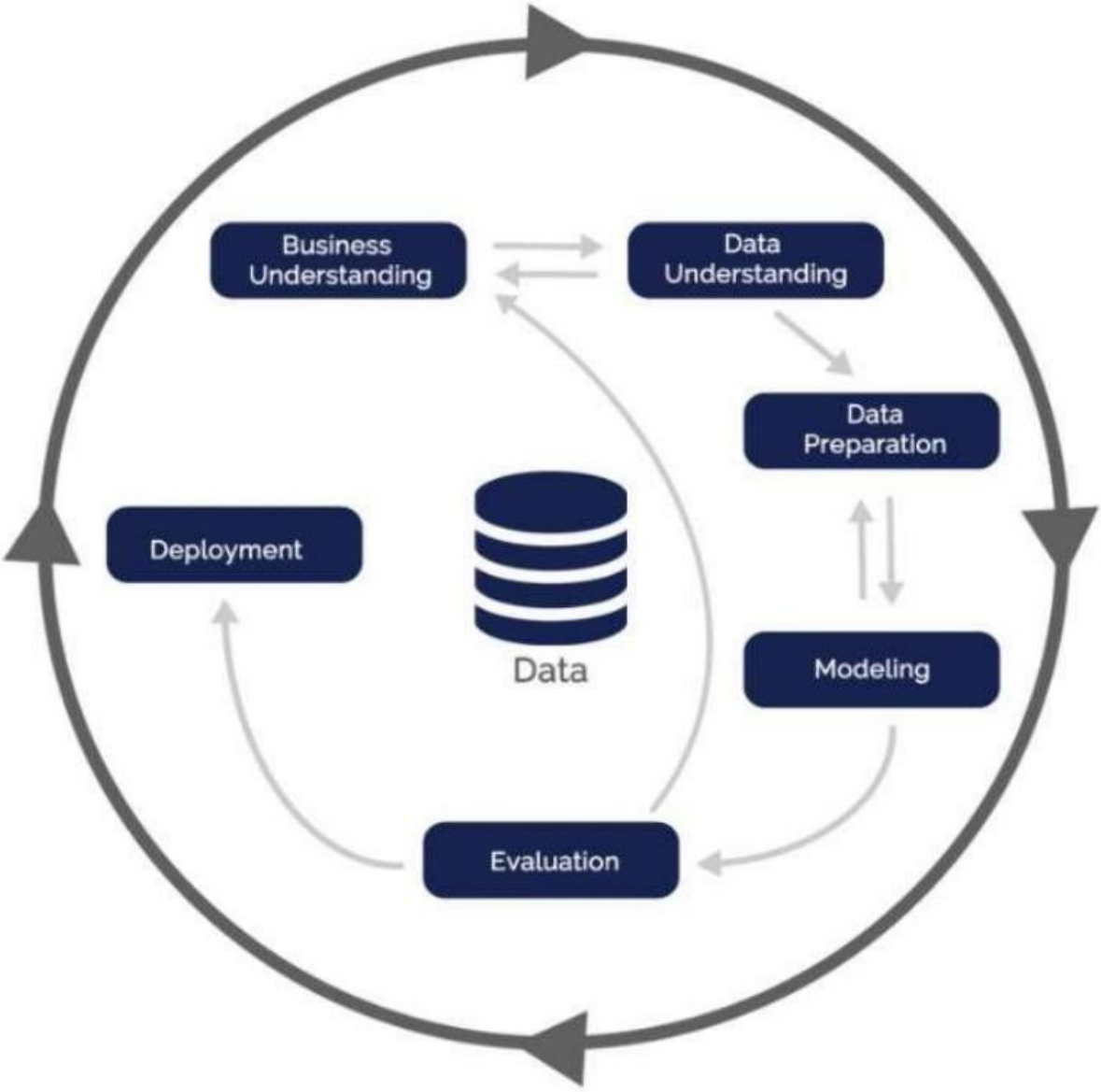
There is a slight imbalance in the distribution of classes in the Target variable.



Despite the imbalance the data still captures enough data for both satisfied and unsatisfied customers for us to build a model upon. Hence we can proceed with this distribution of classes in the target variable.

PROJECT METHODOLOGY

CRISP-DM which stands for Cross Industry Standard Process for Data Mining is a methodology created to help shape data mining projects. It describes the different phases/tasks involved in the project and provides an overview of data mining life cycle.



1. Business Understanding - It focuses on determining the business requirements/objectives and understanding what outcome to achieve. Also, determine the business units being affected. Convert this business problem into a data mining problem and carve out an initial plan.

- Determine the business objectives: Understand what is needed to be accomplished for the customer.
- Assess the situation: Determine resource availability, and project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.
- Determine data mining goals: Convert a business problem to a data mining problem and recognize the data mining problem type such as classification, regression or clustering, etc.
- Produce a project plan: Devise a step-to-step plan for executing the project.

2. Data understanding - This phase starts with collecting the data and then examining the data for its surface properties like data format, number of records, etc. The next step is to better understand the data by understanding each attribute and performing basic statistics on them. Understand the relationship between different attributes. Determine the quality of data by checking the missing values, outliers, duplicates, etc.

- Collect initial data: Acquire the data and load it into the analysis tool to be used.
- Describe data: Examine the data and document its surface properties like data format, number of records, or field identities. Understand the meaning of each attribute and attribute value in business terms. For each attribute, compute basic statistics so as to get a higher-level understanding.
- Explore data: Find insights from the data. Query it, visualize it, and identify relationships among the data.
- Verify data quality: Identify special values, missing attributes, and null data. Determine how to clean/dirty is the data.

3. Data preparation - This stage, which is often referred to as data wrangling,

has the objective to develop the final data set for EDA and modeling. Covers all activities to construct the final dataset from the initial raw data. Some of the tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

- Select data: Determine which attributes/features will be used and document reasons forinclusion/exclusion.
- Clean data: Correct, impute and remove the improper data.
- Extract data: Derive new attributes from the existing ones
- Integrate data: Create features by combining data from multiple sources.
- Format data: Re-format data as necessary. For example, convert string values to numericvalues so as to perform mathematical operations.

4. Modeling - In this stage, we build and assess different models built using various techniques from the training dataset.

- Select modeling technique: Determine the algorithms to be used to model the data based onthe business requirement.
- Generate test design: In order to build and test the model, we need to divide the dataset intotraining and testing data sets. In this step, we divide the data into train and test data set.
- Build model: Based on the modeling technique selected, build the model on the input dataset.
- Assess model: Compare the results of different models based on the confusion matrix. Theoutcome of this step frequently leads to model tuning iterations until the best model is found.

5. Evaluation - Evaluate the models and review the steps executed to construct the model to be certain it properly achieves the business objectives.

PRE-PROCESSING DATA ANALYSIS

Data Preparation:

Data preprocessing is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training Machine Learning models.

Data Preparation is the process of collecting, cleaning, and consolidating data into one file or data table,primarily for use in the analysis.

Acquire the dataset:

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Missing/Null Values:

Impute or drop features with missing values based on the percentage of missing values and relevance formodel building.

```
df.isnull().sum()
```

```
Unnamed: 0                             0
id                                     0
Gender                                 0
Customer Type                          0
Age                                    0
Type of Travel                         0
Class                                  0
Flight Distance                        0
Inflight wifi service                  0
Departure/Arrival time convenient      0
Ease of Online booking                 0
Gate location                          0
Food and drink                         0
Online boarding                        0
Seat comfort                           0
Inflight entertainment                 0
On-board service                       0
Leg room service                       0
Baggage handling                       0
Checkin service                        0
Inflight service                       0
Cleanliness                            0
Departure Delay in Minutes             0
Arrival Delay in Minutes             393
satisfaction                           0
dtype: int64
```

In the dataset we have Null values in one column("Arrival Delay in Minutes"):We calculated the percentage of null values present in the variable to understand their impact on the data

```
In [55]: #Checking percentage of null values
         df["Arrival Delay in Minutes"].isnull().sum()/len(df)

Out[55]: 0.003025870033877425
```

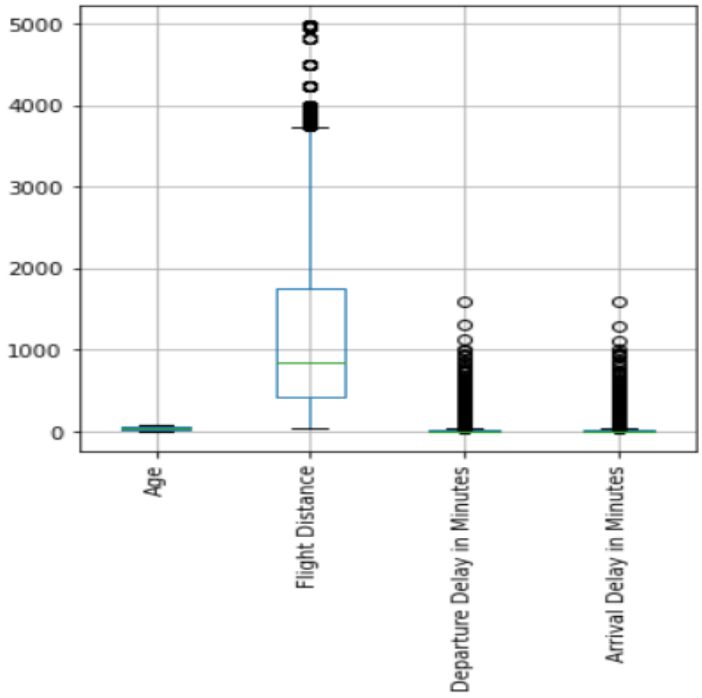
As the percentage of null values was found to be negligible, we dropped the observations containing these null values.
There are no null values in the data now

```
df.isnull().sum()/len(df)
```

```
Unnamed: 0                          0.0
id                                  0.0
Gender                              0.0
Customer Type                       0.0
Age                                 0.0
Type of Travel                      0.0
Class                               0.0
Flight Distance                     0.0
Inflight wifi service               0.0
Departure/Arrival time convenient   0.0
Ease of Online booking              0.0
Gate location                       0.0
Food and drink                      0.0
Online boarding                     0.0
Seat comfort                        0.0
Inflight entertainment              0.0
On-board service                    0.0
Leg room service                    0.0
Baggage handling                    0.0
Checkin service                     0.0
Inflight service                    0.0
Cleanliness                         0.0
Departure Delay in Minutes          0.0
Arrival Delay in Minutes            0.0
satisfaction                        0.0
dtype: float64
```

OUTLIERS

```
: df_num.boxplot()
  plt.xticks(rotation=90)
  plt.show()
```



There are outliers present in the Flight Distance, Departure Delay in Minutes and Arrival Delay in Minutes column. As these variables play a role in customer satisfaction (more the delay,we can assume less satisfied the customer), we are not excluding the outliers . As we proceed with model building we can transform the extreme outliers. But for the base model we choose to leave the outliers as it is

INSIGNIFICANT COLUMNS

The below are the redundant features that are dropped from the dataset.

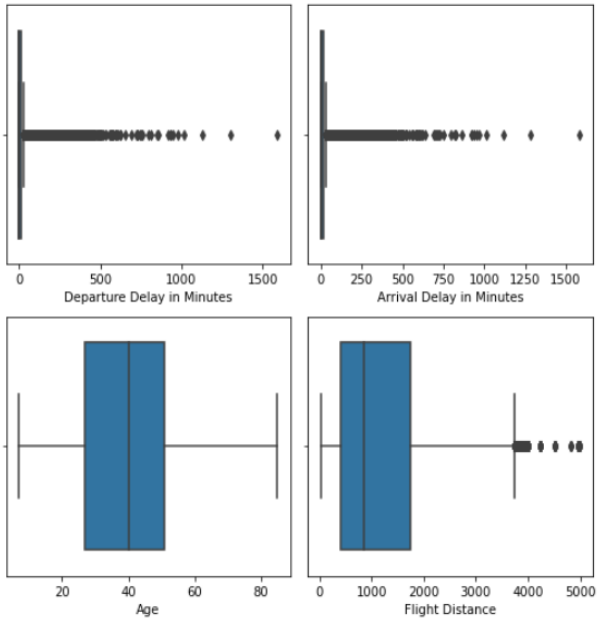
- Unnamed: 0- Index value
- Id- Unique identifiers of each row, has no impact on the model

EXPLORATORY DATA ANALYSIS & BUSINESS INSIGHTS

Univariate Analysis

1. Overview of the Numerical variables

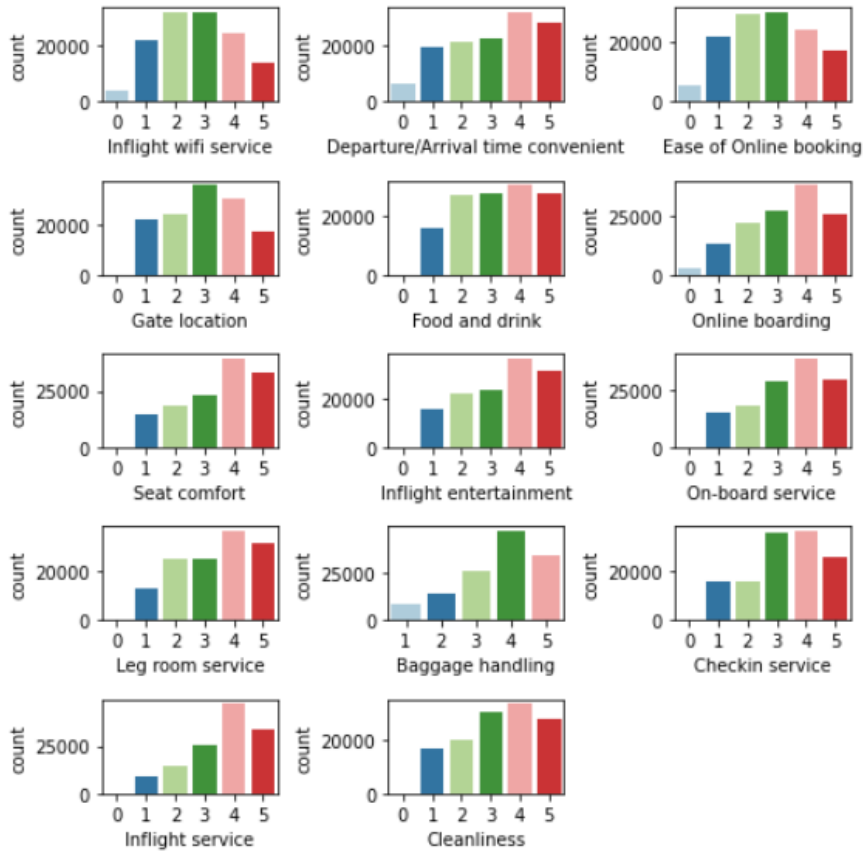
```
l=df[['Departure Delay in Minutes', 'Arrival Delay in Minutes','Age', 'Flight Distance']]
plt.rcParams["figure.figsize"]=[7,7]
for (i,j) in zip(l,range(1,len(l)+1)):
    plt.subplot(2,2,j)
    sns.boxplot(df[i])
plt.tight_layout()
plt.show()
```



- ***Heavy presence of outliers in variables: 'Departure Delay in Minutes', 'Arrival Delay in Minutes'***
- ***The average age of passengers was found to be ~40***
- ***The average distance was found to be 1190 miles***

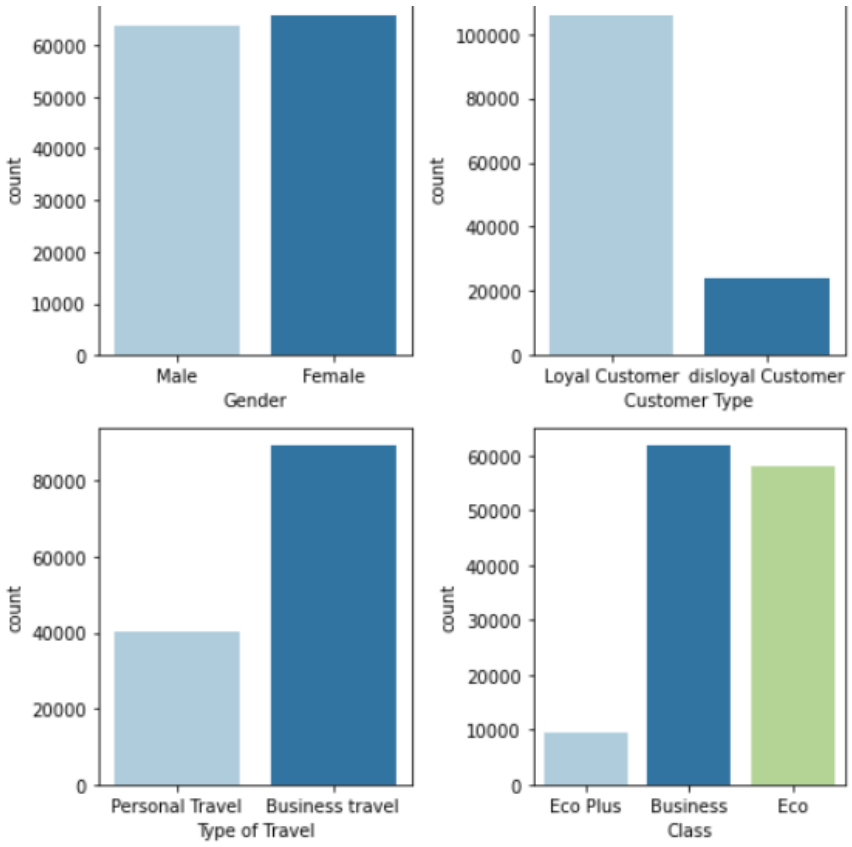
2. Overview of Rating columns

```
l=df[[ 'Inflight wifi service',
       'Departure/Arrival time convenient', 'Ease of Online booking',
       'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort',
       'Inflight entertainment', 'On-board service', 'Leg room service',
       'Baggage handling', 'Checkin service', 'Inflight service',
       'Cleanliness']]
plt.rcParams["figure.figsize"]=[7,7]
for (i,j) in zip(l,range(1,len(l)+1)):
    plt.subplot(5,3,j)
    sns.countplot(df[i],palette="Paired")
plt.tight_layout()
plt.show()
```



- ***A large population of passengers gave an average rating of 4 out of 5 for various factors like ease of online booking, on-board service, checkin service etc.***
- ***The factor with the largest amount of dissatisfaction (rating 1) was in-flight service***
- ***There were some factors that were voted zero ,which we can assume to be for the cases where the customer refused to provide a rating***

3. Overview of Categorical columns



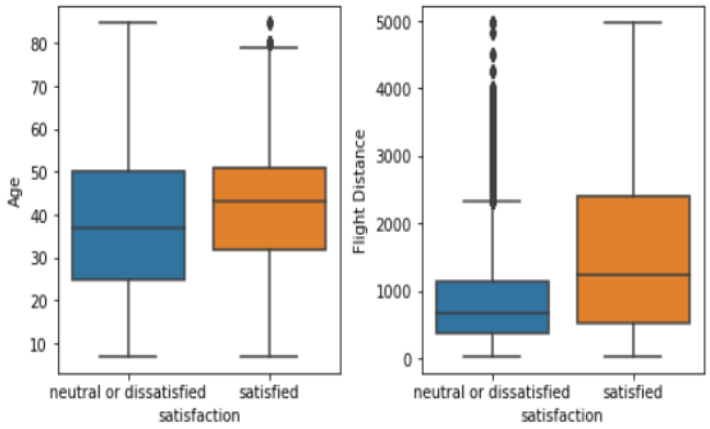
- ***Majority of passengers chose Business class.***
- ***The data contains information predominantly of loyal customers***
- ***Both genders were more or less equally captured for this project***

Bivariate Analysis

Relationship with the target variable:

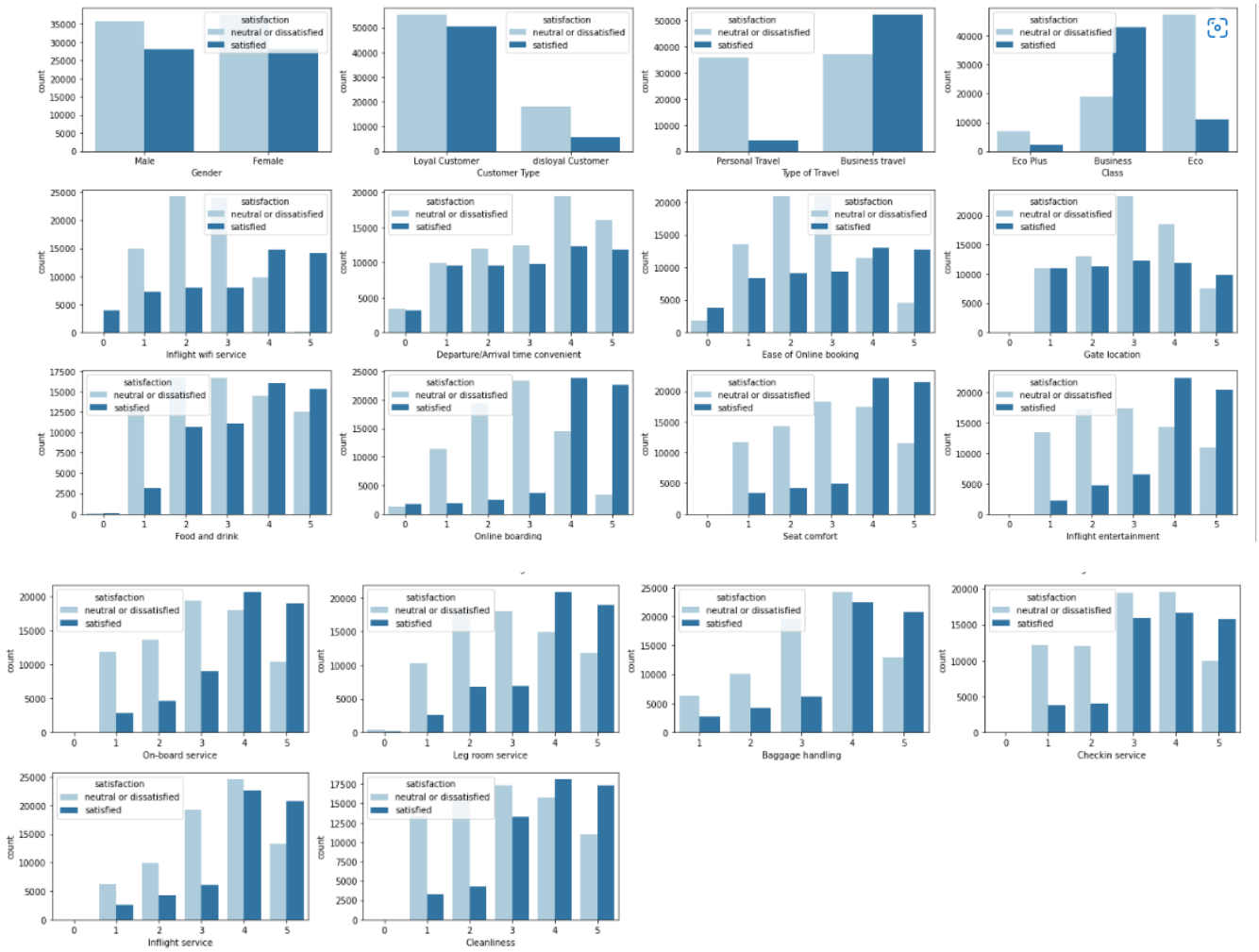
1. Numerical independent variables vs target variable

```
l=df[['Age', 'Flight Distance']]
plt.rcParams["figure.figsize"]=[7,7]
for (i,j) in zip(l,range(1,len(l)+1)):
    plt.subplot(2,2,j)
    sns.boxplot(df["satisfaction"],df[i])
plt.tight_layout()
plt.show()
```



- ***The average age for satisfied and dissatisfied were found to be 37 and 41 respectively***
- ***The average distance travelled by satisfied customers was higher than that of neutral/dissatisfied***

2. Categorical independent variables vs target variable



- ***Loyal customers were found to be satisfied in comparison with disloyal ones***
- ***Business travelers' are highly satisfied***
- ***Moreover, passengers in business class get a high satisfied count, while passengers in the eco and eco plus class were found to be more dissatisfied***
- ***The Main Causes of dissatisfaction are: In-Flight Wifi, Online Booking,and Online boarding***

- ***Less But Observable Dissatisfaction is caused by Leg-Room Services and Cleanliness***
- ***Some Dissatisfaction is also caused by Gate Location, Food, and Drinks, Seat Comfort, Inflight Entertainment, On-Board Services***
- ***Online Boarding Shows generally satisfactory ratings but is one of the main causes of dissatisfaction among the dissatisfied customers***

Statistical Tests:

So far we have observed that the satisfaction of the customer has a significant relationship with other variables such as the type of travel, their rating of factors like inflight wifi/seat comfort/onboard service, etc. We have made these observations by studying the graphs above.

We can further validate these observations using statistical tests like Chi-squared contingency(Test of independence) and Kruskal Wallis test.

Chi-Square test of independence

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables.

The hypothesis to test the independence of attributes

H0**:** The attributes are independent**.**
H1**:** The attributes are dependent

1. Gender vs Satisfaction:

To understand the impact of the variable “gender” on the satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the customer’s gender on their satisfaction

H0:The variables Gender and satisfaction are independent

H1:The variables Gender and satisfaction are dependent

```
from scipy .stats import chi2
from scipy.stats import chi2_contingency
from scipy import stats

x=pd.crosstab(df['Gender'],df['satisfaction'])
#For α = 0.05
# performing the test of independence
test_stat, p1, dof, expected_value = chi2_contingency(observed = obs_val, correction = False)
print("p-value:", p1)
```

```
p-value: 5.135589964552752e-05
```

2. Customer Type vs Satisfaction :

To understand the impact of the variable “customer type” on the satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the customer’s loyalty on their satisfaction

H0:The variables Customer Type and satisfaction are independent

H1:The variables Customer Type and satisfaction are dependent

```
x=pd.crosstab(df['Customer Type'],df['satisfaction'])

obs_val=x.values
test_stat, p1, dof, expected_value = chi2_contingency(observed = obs_val, correction = False)
print("p-value:", p1)
```

```
p-value: 0.0
```

3. Type of Travel vs Satisfaction

To understand the impact of the variable “Type of Travel” on the satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the customer’s reason for travel on their satisfaction

H0:The variables Type of Travel and satisfaction are independent

H1:The variables Type of Travel and satisfaction are dependent

```
]: x=pd.crosstab(df['Type of Travel'],df['satisfaction'])

   obs_val=x.values
   test_stat, p1, dof, expected_value = chi2_contingency(observed = obs_val, correction = False)
   print("p-value:", p1)
```

```
p-value: 0.0
```

3. Class vs Satisfaction:

To understand the impact of the variable “Class” on the satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the Class the customer is using (Business/Eco/Ecoplus) on their satisfaction

H0:The variables Class and satisfaction are independent

H1:The variables Class and satisfaction are dependent

```
: x=pd.crosstab(df['Class'],df['satisfaction'])

  obs_val=x.values
  test_stat, p1, dof, expected_value = chi2_contingency(observed = obs_val, correction = False)
  print("p-value:", p1)
```

```
p-value: 0.0
```

5. Rating columns vs Satisfaction:

To understand the impact of the rating variables (example: seat comfort, inflight service, food, and drink quality)on the satisfaction we conducted a test of independence. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the way these factors were rated by the customer on their satisfaction.

```
for i in df_cat.columns:
    x=pd.crosstab(df[i],df['satisfaction'])
    obs_val=x.values
    test_stat, p, dof, expected_value = chi2_contingency(observed = obs_val, correction = False)
    print(i,':',p)
```

```
Gender : 3.521830031647996e-05
Customer Type : 0.0
Type of Travel : 0.0
Class : 0.0
Inflight wifi service : 0.0
Departure/Arrival time convenient : 2.473802119547831e-127
Ease of Online booking : 0.0
Gate location : 0.0
Food and drink : 0.0
Online boarding : 0.0
Seat comfort : 0.0
Inflight entertainment : 0.0
On-board service : 0.0
Leg room service : 0.0
Baggage handling : 0.0
Checkin service : 0.0
Inflight service : 0.0
Cleanliness : 0.0
```

Kruskal Wallis test

It is used to check the equality of population medians for more than two independent samples. It is a non-parametric test that works similar to ANOVA but doesn't require the tests of normality and equal variances to be satisfied.

The null and alternative hypothesis is given as:

Ho: The median of all treatments is the same.

H1: At least one treatment has a different median.

To understand the impact of the numerical variables (example: age, flight distance, departure delay in minutes, arrival delay in minutes)on the satisfaction we conducted a Kruskal Wallis test. If the result of the test indicates dependency (reject the null hypothesis) we can state that there is some impact of the way these factors were rated by the customer on their satisfaction.

	p_value
Age	0.000000e+00
Flight Distance	0.000000e+00
Departure Delay in Minutes	3.580163e-137
Arrival Delay in Minutes	7.087659e-287

Inference From Statistical Tests:

Gender and satisfaction

For the variable Gender the p_value < α , so we reject the null hypothesis. This means that the Gender and the target variable satisfied are dependent. Gender plays a role in customer satisfaction.

Customer type and satisfaction

For the variable Customer Type the p_value < α, so we reject the null hypothesis. The variable customer type also makes a difference in customer satisfaction.

Type of Travel and satisfaction

For the variable Type of travel the p_value < α, so we reject the null hypothesis. Business travelers travel more frequently than people who travel for personal affairs. So they are more familiar with the flight experiences than the other group. This can create an effect on customer satisfaction.

Class and satisfaction.

For the variable Class the p_value < α, so we reject the null hypothesis. The variable satisfaction is dependent on the variable class.

Rating variables and satisfaction.

For all the variables the p_value < α, so we reject the null hypothesis. The variable satisfaction is dependent on the rating variables.

Numerical variables and satisfaction.

For all the numerical variables the p_value < α, so we reject the null hypothesis. The variable satisfaction is dependent on the numerical variables.

MODEL BUILDING:

Understanding The Metrics:

PERFORMANCE METRICS:

- **Confusion Matrix:**

  It is the performance measure for the classification problem. It is a table used to compare predicted and actual values of the target variable.

- **ROC:**

  ROC curve is the plot of TPR against the FPR values obtained at all possible threshold values.

PERFORMANCE EVALUATION METRICS:

- **Accuracy:**

  Accuracy is the fraction of predictions that our model got correct. Higher the accuracy of the model better is the model. The accuracy of the base model was found to be 0.83

- **Precision:**

  Precision is the proportion of positive cases that were correctly predicted.

- **Recall:**

  A recall is the proportion of actual positive cases that were correctly predicted.

- **F1 score:**

  F1score is the harmonic mean of precision and recall values for a classification model.

BUSINESS IMPACT

- The main motive is to improve the performance of the model. As per the business scenario, the model is predicting the customers are satisfied but in reality, they aren't. In this scenario, we won't be able to establish satisfaction of the customer and this might result in us losing a potentially loyal customer
- In the second scenario the model has predicted customer is not satisfied but in reality, they are satisfied.
- It is important to minimize the two types of false predictions. Being unable to classify a satisfied customer accurately may lead to a wrong interpretation of how the business is being carried out (Unable to identify our strengths).On the other hand, falsely classifying a dissatisfied customer as satisfied have a huge impact on the business as we will never be able to identify areas of improvement. This can ultimately result in the loss of customers.
- Our goal is ultimately to reduce both types of False predictions. Hence we will be looking at the F1 score and Accuracy as our metrics for model consideration.
- We will also proceed with feature selection to obtain the features that contribute most to the prediction of satisfaction. Reducing the features will thereby reduce the complexity, making the model more efficient and making it more practical for real-time applications.

Preparing the data for model building:

Encoding:

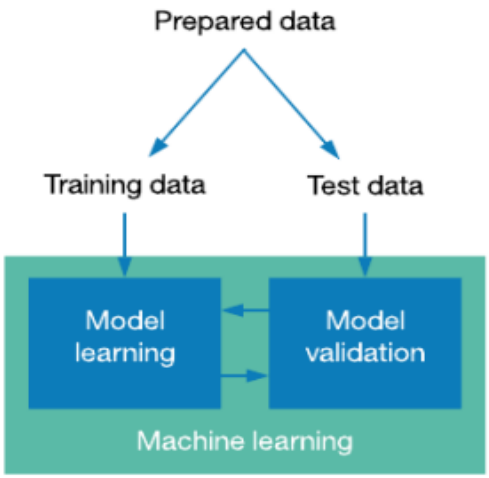
```
from sklearn.preprocessing import LabelEncoder
LE=LabelEncoder()
df["satisfaction"]=LE.fit_transform(df["satisfaction"])
```

```
df_cat=df[["Gender",'Customer Type','Type of Travel', 'Class']]
df=pd.get_dummies(df,columns=df_cat.columns,drop_first=True)
df.head(2)
```

On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction	Gender_Male	Customer Type_disloyal Customer	Type of Travel_Personal Travel	Class_Eco	Class_Eco Plus
5	5	5	2	5	5	50	44.0	1	0	0	0	1	0
4	4	4	3	4	5	0	0.0	1	0	0	0	0	0

Train-Test Split

After encoding the categorical features we split the data into train data and test data. The model uses train data to learn and test data to evaluate/validate the trained model. (Note: We also perform K-Fold cross-validation to overcome any bias that may have been created due to the split)



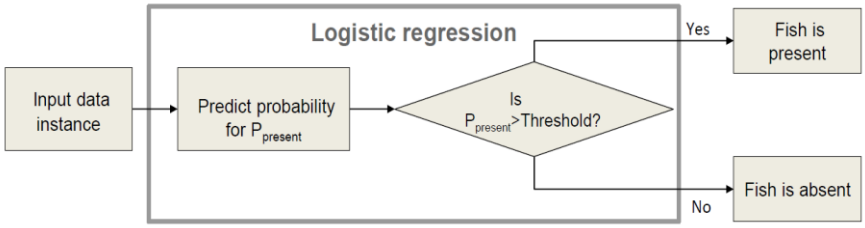
Performing a 70-30 split:

```
X=df.drop("satisfaction",axis=1)
Y=df["satisfaction"]
from sklearn.model_selection import train_test_split
xtrain,xtest,ytrain,ytest=train_test_split(X,Y,train_size=0.7,random_state=100)
```

BASE MODEL

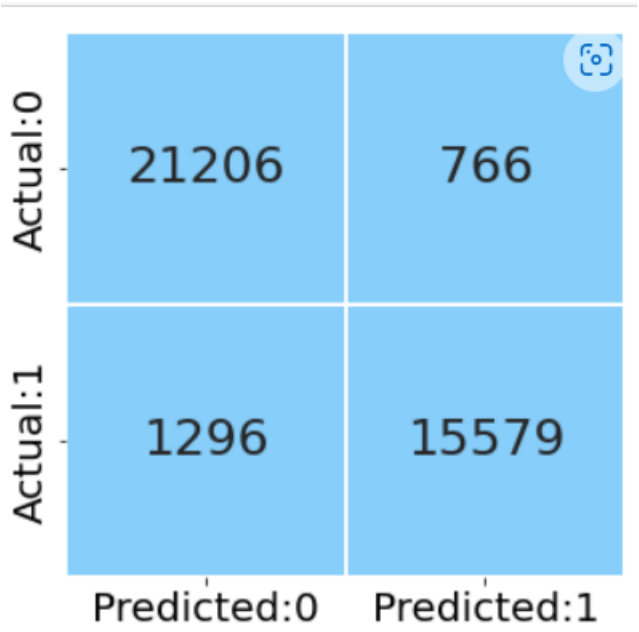
Logistic regression:

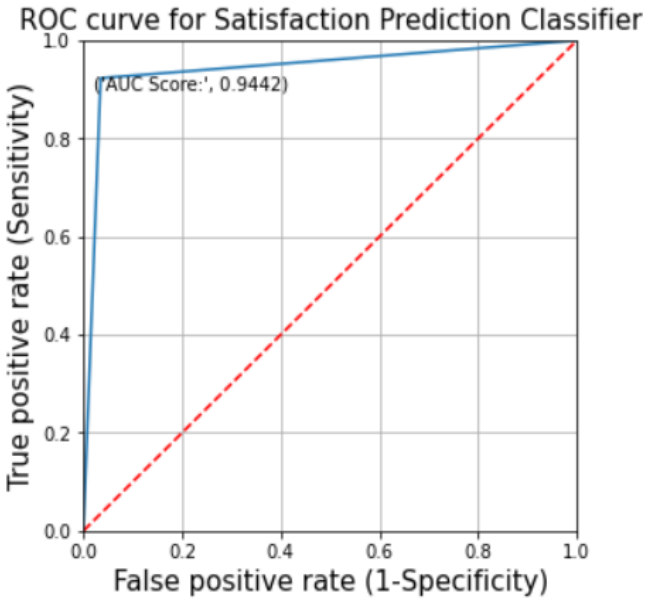
- Logistic Regression is a binary classification algorithm. It predicts the probability of occurrence of a label class.
- Consider that logistic regression is used to identify whether the product falls under the advantage category or not.




```
from sklearn.linear_model import LogisticRegression
LR=LogisticRegression()
model1=LR.fit(xtrain,ytrain)
ypred=model1.predict(xtest)
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

```
              precision    recall  f1-score   support

           0       0.85      0.81      0.83     21972
           1       0.76      0.82      0.79     16875

    accuracy                           0.81     38847
   macro avg       0.81      0.81      0.81     38847
weighted avg       0.81      0.81      0.81     38847
```



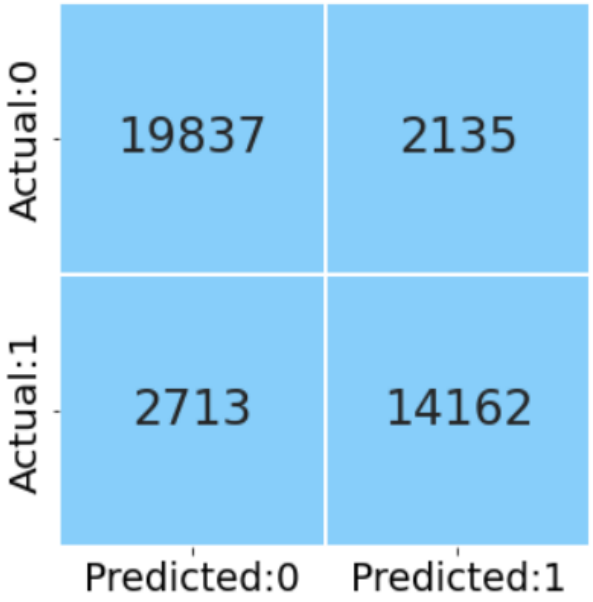


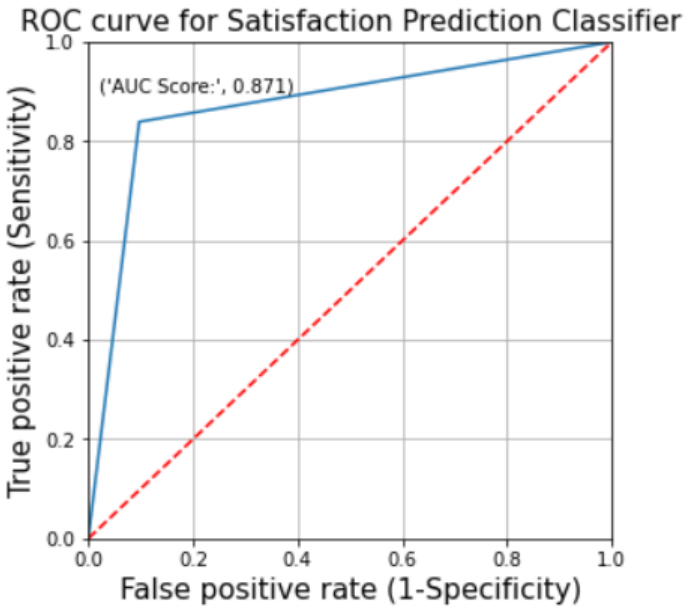
Logistic regression after transformation:

Transforming the numerical variables to reduce skewness and impact of an outlier.

```
LR=LogisticRegression()
model2=LR.fit(xtrain,ytrain)
ypred=model2.predict(xtest)
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

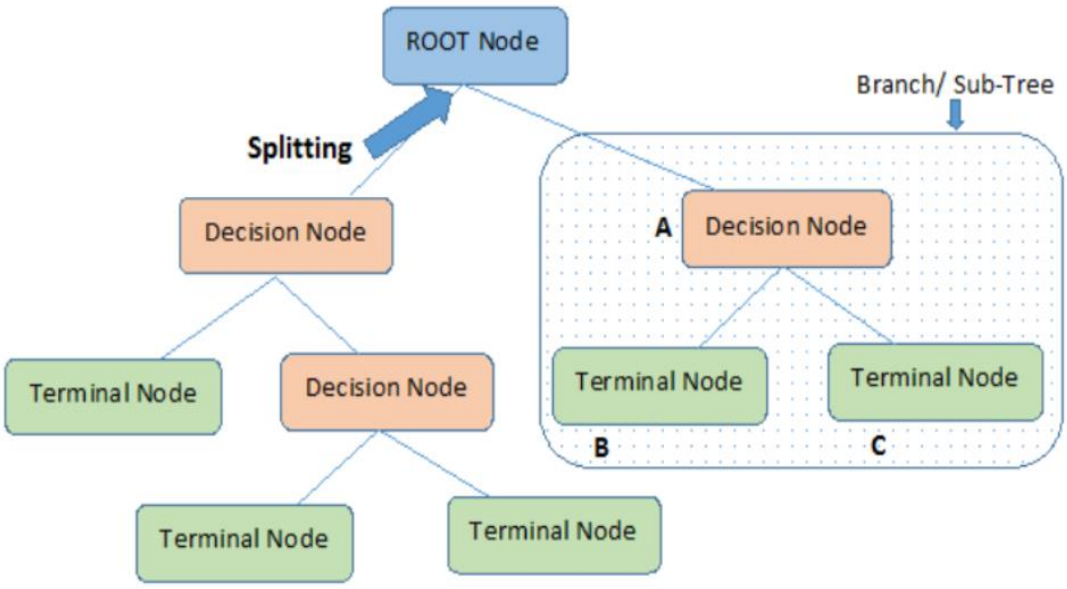
```
              precision    recall  f1-score   support

           0       0.88      0.90      0.89     21972
           1       0.87      0.84      0.85     16875

    accuracy                           0.88     38847
   macro avg       0.87      0.87      0.87     38847
weighted avg       0.88      0.88      0.87     38847
```





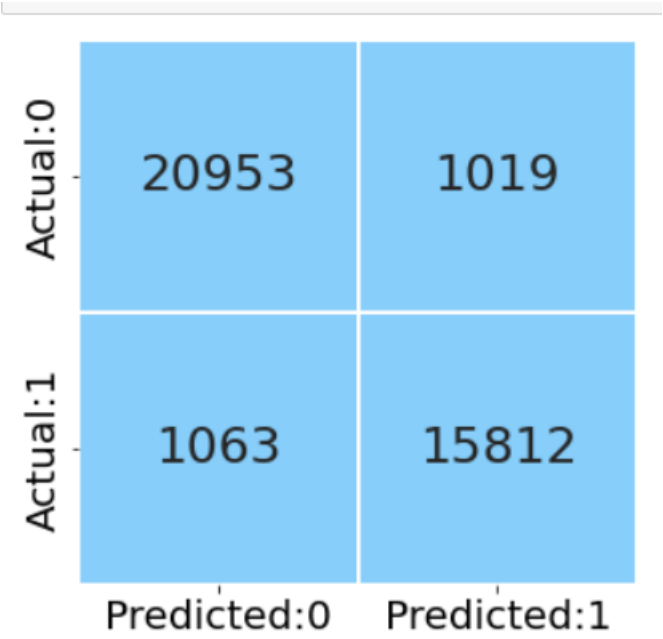
Decision Tree Algorithm:

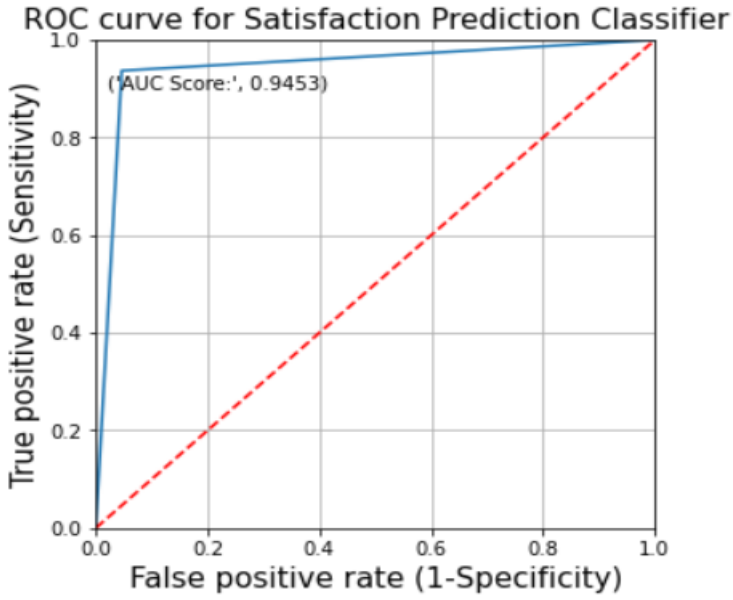
- Decision trees can be used for classification as well as regression problems.
- The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits.
- It starts with a root node and ends with a decision made by leaves.



```
dt=DecisionTreeClassifier(criterion="entropy",random_state = 10)
model3=dt.fit(xtrain,ytrain)
ypred=model3.predict(xtest)
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

```
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     21972
           1       0.94      0.94      0.94     16875

    accuracy                           0.95     38847
   macro avg       0.95      0.95      0.95     38847
weighted avg       0.95      0.95      0.95     38847
```



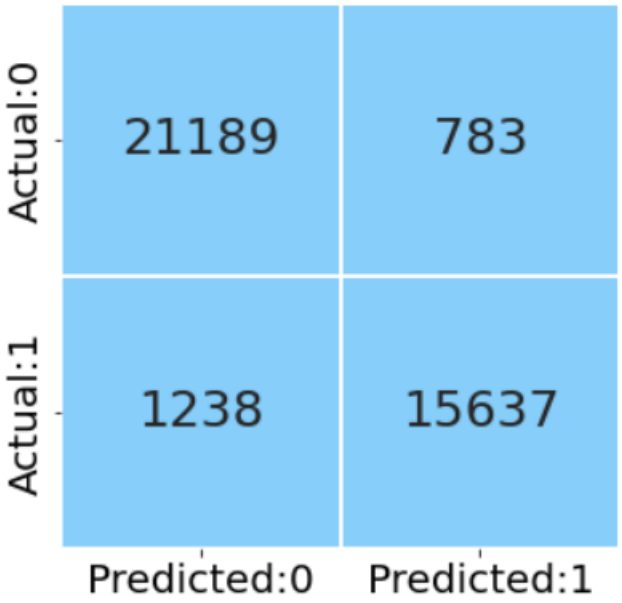


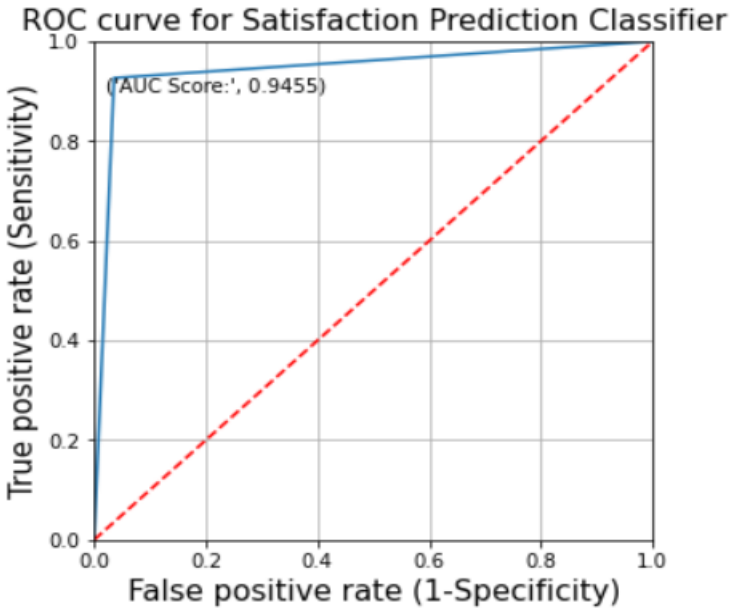
Tuned Decision Tree:

We used GridSearchCV to obtain best hyper parameters. We then built our model with these hyperparameters: criterion="entropy", max_depth=11,min_samples_split=5.

```
model4=dt.fit(xtrain,ytrain)
ypred=model4.predict(xtest)
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

	precision	recall	f1-score	support
0	0.94	0.96	0.95	21972
1	0.95	0.93	0.94	16875
accuracy			0.95	38847
macro avg	0.95	0.95	0.95	38847
weighted avg	0.95	0.95	0.95	38847



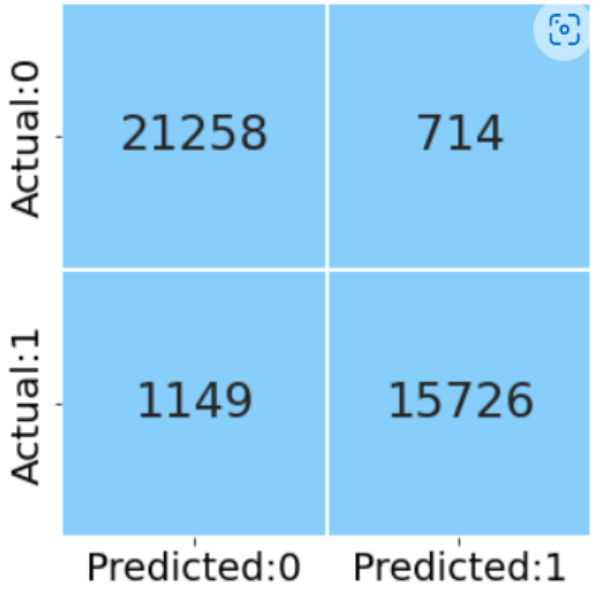


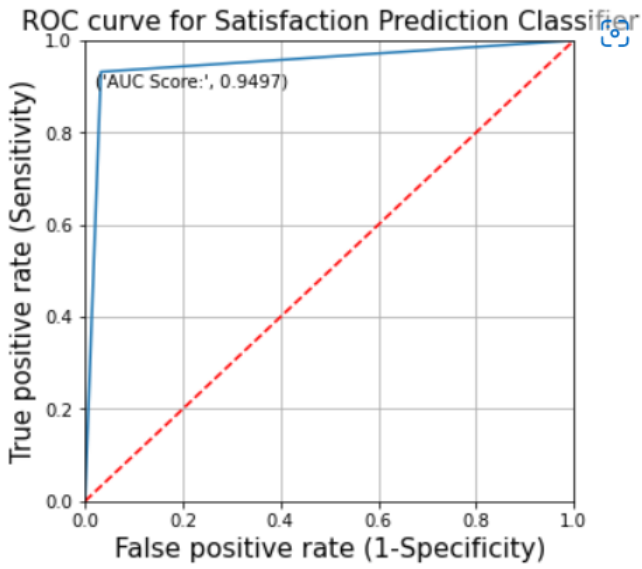
Random Forest Model:

- Random Forest consists of several independent decision trees that operate as an ensemble.
- It is an ensemble learning algorithm based on bagging
- For the base random forest model we chose n_estimators as '5', for ease of computation

```
model5=rf.fit(xtrain,ytrain)
ypred=model5.predict(xtest)
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

	precision	recall	f1-score	support
0	0.95	0.97	0.96	21972
1	0.96	0.93	0.94	16875
accuracy			0.95	38847
macro avg	0.95	0.95	0.95	38847
weighted avg	0.95	0.95	0.95	38847



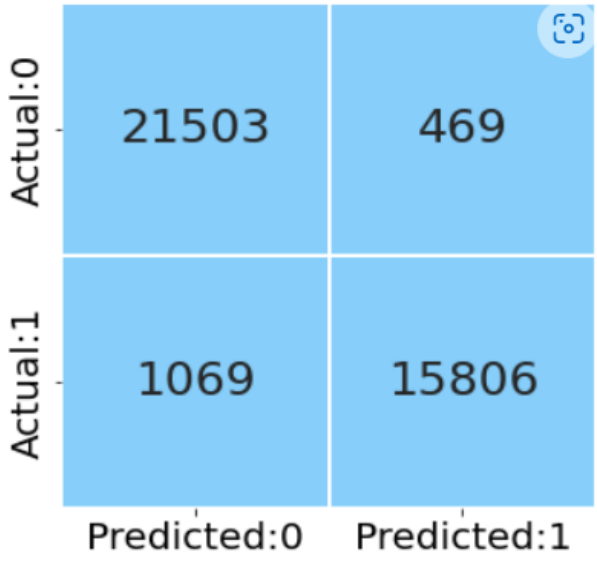


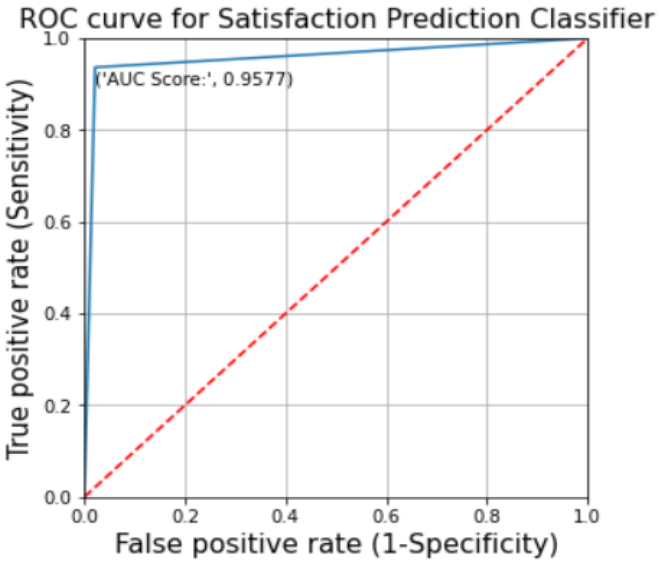
Tuned Random Forest Model:

- We used GridSearchCV to obtain best hyperparameters.We then built our model with these hyperparameters: criterion="entropy", max_depth=20, min_samples_split=25, random_state = 10, n_estimators=100

```
model6=rf.fit(xtrain,ytrain)
ypred=model6.predict(xtest)
from sklearn.metrics import classification_report
print(classification_report(ytest,ypred))
```

```
              precision    recall  f1-score   support

           0       0.95      0.98      0.97     21972
           1       0.97      0.94      0.95     16875

    accuracy                           0.96     38847
   macro avg       0.96      0.96      0.96     38847
weighted avg       0.96      0.96      0.96     38847
```



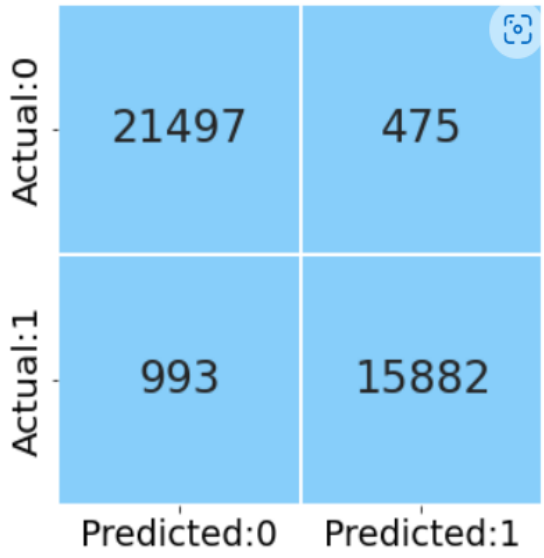


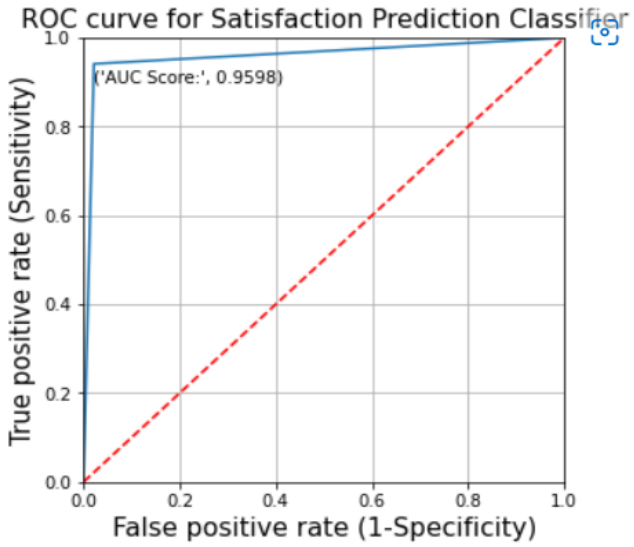
Boosting:

We used XGBoost as our boosting technique.

- In this algorithm, decision trees are created in sequential form.
- Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results.
- The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree.
- These individual classifiers/predictors then ensemble to give a strong and more precise model

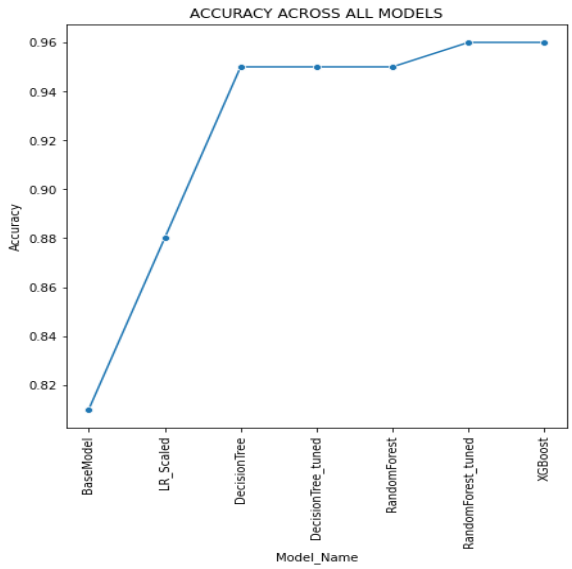
	precision	recall	f1-score	support
0	0.96	0.98	0.97	21972
1	0.97	0.94	0.96	16875
accuracy			0.96	38847
macro avg	0.96	0.96	0.96	38847
weighted avg	0.96	0.96	0.96	38847

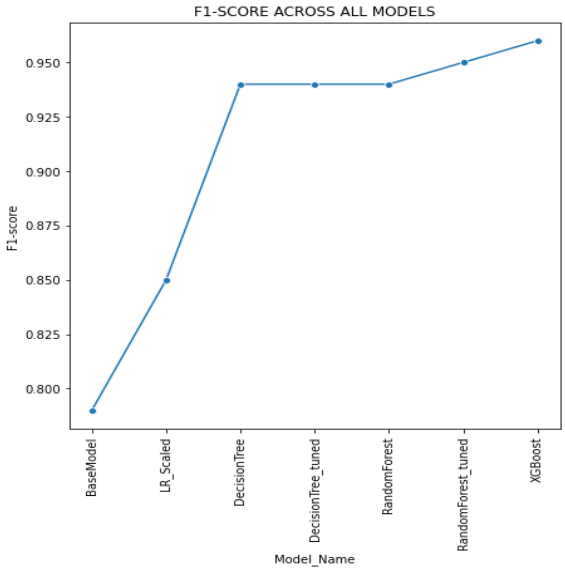




Model Comparison:

	Base_Model	LR_scaled	DecisionTree	DecisionTreeTuned	RandomForest	RandomForestTuned	XGBoost
Model_Name	BaseModel	LR_Scaled	DecisionTree	DecisionTree_tuned	RandomForest	RandomForest_tuned	XGBoost
Accuracy	0.81	0.88	0.95	0.95	0.95	0.96	0.96
Kfold_accuracy	0.81	0.81	0.95	0.95	0.95	0.96	0.96
F1-score	0.79	0.85	0.94	0.94	0.94	0.95	0.96





- It was found that both the Random Forest Tuned model and the XGBoost model had an accuracy of 96%
- The F1-score was found to be highest for the XGBoost model-96%.
- Overall we see a remarkable improvement from the base model which had an accuracy of 81% and an F1-Score of 79%.
- We ultimately see the most improvement overall with the XGBoost Model.
- The confusion matrix as well as other metrics such as precision and recall have seen an improvement using this model indicating that it is effective in reducing the false predictions.

INFERENCES AND RECOMMENDATIONS

We observed the following features based have the most influence on customer satisfaction based on the final model: Online Boarding, Class, Inflight Wi-Fi service, Inflight Entertainment, and Cleanliness. After going for a deep dive into how these features were influencing the customer satisfaction we have come up with the following recommendations:

- It was observed that the customers who were ultimately unsatisfied with their overall experience also rated the Online boarding Process poorly. We might improve upon Customer satisfaction by directing our efforts on improving this process.
- Large proportion of people who traveled in the Economy class were found to be dissatisfied in comparison with the people who traveled in business class. A further deep dive on their ratings for different features showed that they were largely unsatisfied with the inflight wi-fi service, food, and drinks, seat comfort, cleanliness, etc. By making changes to these aspects we might help improve the experience of those traveling by economy class
- We also see that a large number of customers who were unsatisfied overall gave a poor rating to the Wi-fi service. Changing the internet service provider to a more efficient version might improve customer satisfaction
- Similarly that a large number of customers who were unsatisfied overall gave a poor rating to Inflight Entertainment. By looking into different resources like tie-ups with different OTT platforms that provide better content for example we may be able to enhance customer experience.
- Cleanliness also had a huge impact on customer satisfaction. By focusing more on this aspect by increasing the frequency of cleaning before and after the boarding process we may be able to improve customer experience.

By focusing on these aspects for starters we may be able to ultimately provide a more satisfying experience for the customers which will ultimately help in increasing the loyal customer base of the airline.

CONCLUSION

From our business perspective, it is crucial to minimize the two types of false predictions. Being unable to classify a satisfied customer accurately may lead to a wrong interpretation of how the

business is being carried out (Unable to identify our strengths).On the other hand, falsely classifying a dissatisfied customer as satisfied has a huge impact on the business as we will never be able to identify areas of improvement. This can ultimately result in the loss of customers. We, therefore, looked at Accuracy and F1-Score as our target metrics.

After running different models to improve upon these metrics we found that by using the boosting technique with XGBoost we were able to create the most efficient model with accuracy and an F1 score of 96%. This was an improvement on our base model which had an accuracy and F1 score of 81% and 79% respectively.

We Further performed cross-validation on all the models to see if there was an agreement on the overall accuracy across different datasets(avoiding the bias that may occur due to train test split) and we see that our final model is able to replicate the results across different datasets, making it a reliable model for satisfaction prediction.

REFERENCES

1. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
2. https://towardsdatascience.com/predicting-satisfaction-of-airline-passengers-with-classification-76f1516e1d16
3. https://www.researchgate.net/publication/350552031_Predicting_Airline_Passenger_Satisfaction_with_Classification_Algorithms
4. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
5. https://www.knowledgeisle.com/wp-content/uploads/2019/12/2-Aur%C3%A9lien-G%C3%A9ron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-O%E2%80%99Reilly-Media-2019.pdf