Project Technical Document

INVISTICO AIRLINE CLASSIFICATION

Contents:

Classification of Invistico Airline Data

Project description:

The aim of this data science project is to analyze the customer satisfaction levels of invistico airline based on the available data. By leveraging data analysis techniques, statistical modeling and machine learning algorithms, we will extract valuable insights to determine whether the customer is satisfied or dissatisfied with the airline's service.

This project will involve data preprocessing, exploratory data analysis, feature engineering, model building, and evaluation to provide actionable recommendation for improving customer satisfaction.

Project Technical Details:

The following diagram shows the various steps that we have followed in our project.

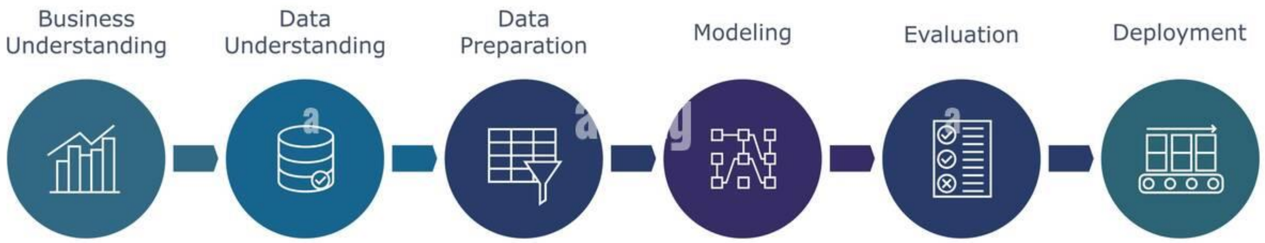
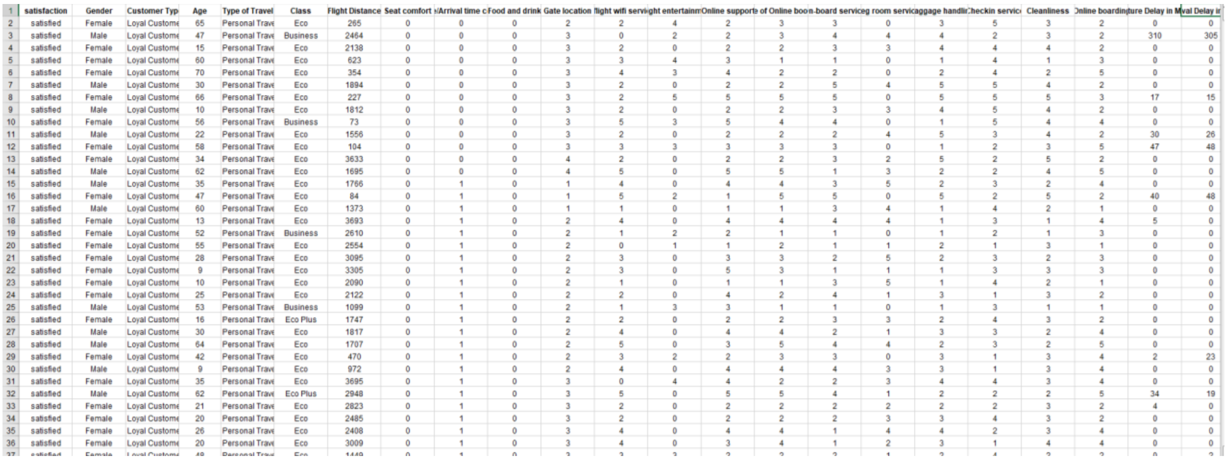


fig 1: General steps of CRISP DM process

1. Data collection:

Below are the raw data collected from Invistico airline excel data which includes various feature and observation. This dataset contain both categorical and numerical data.



1	satisfaction	Gender	Customer Typ	Age	Type of Travel	Class	Flight Distance	Seat comfort	:/Arrival time c	Food and drink	Gate location	flight wifi servi	ght entertainm	Online support	of Online boo	n-board servic	g room servic	aggage handlir	heckin servic	Cleanliness	Online boardin	ture Delay in M	val Delay ir
2	satisfied	Female	Loyal Custome	65	Personal Trave	Eco	265	0	0	0	2	2	4	2	3	3	0	3	5	3	2	0	0
3	satisfied	Male	Loyal Custome	47	Personal Trave	Business	2464	0	0	0	3	0	2	2	3	4	4	4	2	3	2	310	305
4	satisfied	Female	Loyal Custome	15	Personal Trave	Eco	2138	0	0	0	3	2	0	2	2	3	3	4	4	4	2	0	0
5	satisfied	Female	Loyal Custome	60	Personal Trave	Eco	623	0	0	0	3	3	4	3	1	1	0	1	4	1	3	0	0
6	satisfied	Female	Loyal Custome	70	Personal Trave	Eco	354	0	0	0	3	4	3	4	2	2	0	2	4	2	5	0	0
7	satisfied	Male	Loyal Custome	30	Personal Trave	Eco	1894	0	0	0	3	2	0	2	2	5	4	5	5	4	2	0	0
8	satisfied	Female	Loyal Custome	66	Personal Trave	Eco	227	0	0	0	3	2	5	5	5	5	0	5	5	5	3	17	15
9	satisfied	Male	Loyal Custome	10	Personal Trave	Eco	1812	0	0	0	3	2	0	2	2	3	3	4	5	4	2	0	0
10	satisfied	Female	Loyal Custome	56	Personal Trave	Business	73	0	0	0	3	5	3	5	4	4	0	1	5	4	4	0	0
11	satisfied	Male	Loyal Custome	22	Personal Trave	Eco	1556	0	0	0	3	2	0	2	2	2	4	5	3	4	2	30	26
12	satisfied	Female	Loyal Custome	58	Personal Trave	Eco	104	0	0	0	3	3	3	3	3	3	0	1	2	3	5	47	48
13	satisfied	Female	Loyal Custome	34	Personal Trave	Eco	3633	0	0	0	4	2	0	2	2	3	2	5	2	5	2	0	0
14	satisfied	Male	Loyal Custome	62	Personal Trave	Eco	1695	0	0	0	4	5	0	5	5	1	3	2	2	4	5	0	0
15	satisfied	Male	Loyal Custome	35	Personal Trave	Eco	1766	0	1	0	1	4	0	4	4	3	5	2	3	2	4	0	0
16	satisfied	Female	Loyal Custome	47	Personal Trave	Eco	84	0	1	0	1	5	2	1	5	5	0	5	2	5	2	40	48
17	satisfied	Male	Loyal Custome	60	Personal Trave	Eco	1373	0	1	0	1	1	0	1	1	3	4	1	4	2	1	0	0
18	satisfied	Female	Loyal Custome	13	Personal Trave	Eco	3693	0	1	0	2	4	0	4	4	4	4	1	3	1	4	5	0
19	satisfied	Female	Loyal Custome	52	Personal Trave	Business	2610	0	1	0	2	1	2	2	1	1	0	1	2	1	3	0	0
20	satisfied	Female	Loyal Custome	55	Personal Trave	Eco	2554	0	1	0	2	0	1	1	2	1	1	2	1	3	1	0	0
21	satisfied	Female	Loyal Custome	28	Personal Trave	Eco	3095	0	1	0	2	3	0	3	3	2	5	2	3	2	3	0	0
22	satisfied	Female	Loyal Custome	9	Personal Trave	Eco	3305	0	1	0	2	3	0	5	3	1	1	1	3	3	3	0	0
23	satisfied	Female	Loyal Custome	10	Personal Trave	Eco	2090	0	1	0	2	1	0	1	1	3	5	1	4	2	1	0	0
24	satisfied	Female	Loyal Custome	25	Personal Trave	Eco	2122	0	1	0	2	2	0	4	2	4	1	3	1	3	2	0	0
25	satisfied	Male	Loyal Custome	53	Personal Trave	Business	1099	0	1	0	2	1	3	3	1	1	0	1	3	1	1	0	0
26	satisfied	Female	Loyal Custome	16	Personal Trave	Eco Plus	1747	0	1	0	2	2	0	2	2	3	3	2	4	3	2	0	0
27	satisfied	Male	Loyal Custome	30	Personal Trave	Eco	1817	0	1	0	2	4	0	4	4	2	1	3	3	2	4	0	0
28	satisfied	Male	Loyal Custome	64	Personal Trave	Eco	1707	0	1	0	2	5	0	3	5	4	4	2	3	2	5	0	0
29	satisfied	Female	Loyal Custome	42	Personal Trave	Eco	470	0	1	0	2	3	2	2	3	3	0	3	1	3	4	2	23
30	satisfied	Male	Loyal Custome	9	Personal Trave	Eco	972	0	1	0	2	4	0	4	4	4	3	3	1	3	4	0	0
31	satisfied	Female	Loyal Custome	35	Personal Trave	Eco	3695	0	1	0	3	0	4	4	2	2	3	4	4	3	4	0	0
32	satisfied	Male	Loyal Custome	62	Personal Trave	Eco Plus	2948	0	1	0	3	5	0	5	5	4	1	2	2	2	5	34	19
33	satisfied	Female	Loyal Custome	21	Personal Trave	Eco	2823	0	1	0	3	2	0	2	2	2	2	2	2	3	2	4	0
34	satisfied	Female	Loyal Custome	20	Personal Trave	Eco	2485	0	1	0	3	2	0	2	2	2	3	3	4	3	2	0	0
35	satisfied	Female	Loyal Custome	26	Personal Trave	Eco	2408	0	1	0	3	4	0	4	4	1	4	4	2	3	4	0	0
36	satisfied	Female	Loyal Custome	20	Personal Trave	Eco	3009	0	1	0	3	4	0	3	4	1	2	3	1	4	4	0	0
37	satisfied	Female	Loyal Custome	48	Personal Trave	Eco	1449	0	1	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]

1. There are 24 features available in the dataset and 129880 Observations.
2. Checked the null values or missing values in the dataset. Performed data imputation, filled null values with median and mode.
3. The last column named as “@” present at the last column having no data and that was irrelevant to the dataset has been removed.
4. There are 2 features having less then 30% important importance in the whole dataset, therefore it is removed.
   Departure/Arrival time convenient, Gate location
5. Considered the main 18 features such as
   Satisfaction, Gender, Customer Type, Age, Type of Travel, Class, Flight Distance,

Departure/Arrival time convenient, Food and Drink, Inflight wifi service, Online support, Ease of Online booking, On-board service, Leg room service, Baggage handling, checkin service, cleanliness, Online onboarding, Departure delay in minutes, Arrival delay in minutes.

6. The column contains 393 null values therefore we replaced that value with the help of zero.
7. Label encoding is applied to 5 columns having string values and it is important for the model.
   Those columns were gender, customer type, age, type of travel and class.
8. From the above dataset, it is clear that our target is customer satisfaction.
9. As the data present inside the dataset which is not scalable therefore it is normalized with normalization function.
10. Outliers present in the data ,explored using standard deviation,Inter Quartile Range(IQR), Median Absolute deviation(MAD), Isolation Forest, Winsorization with 95th Percentile.
11. Separated continuous data and checked the correlation between the different features.
12. Checking the normalcy of data by using the visualization plots for all the columns.

2. Exploratory Data Analysis:

Analysis done on the basis of several charts which are shown below:

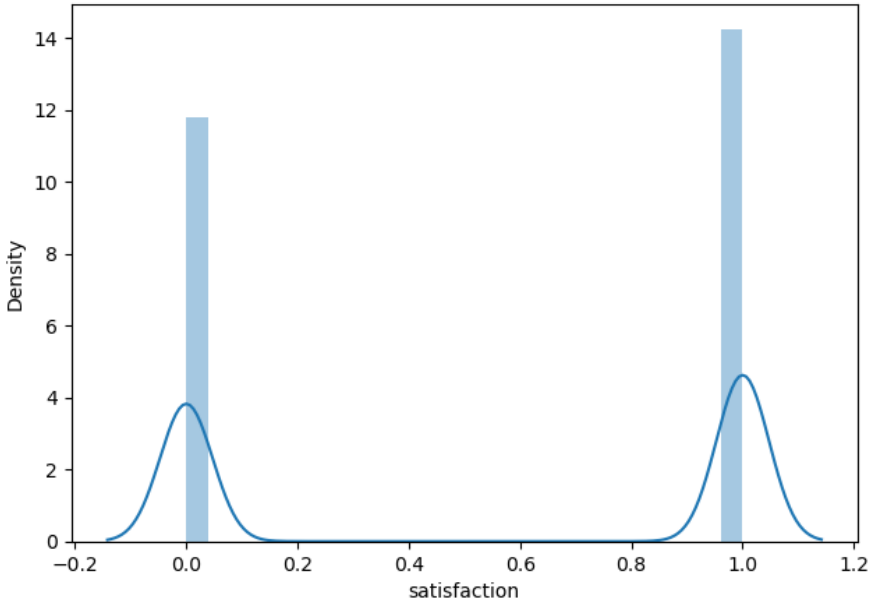
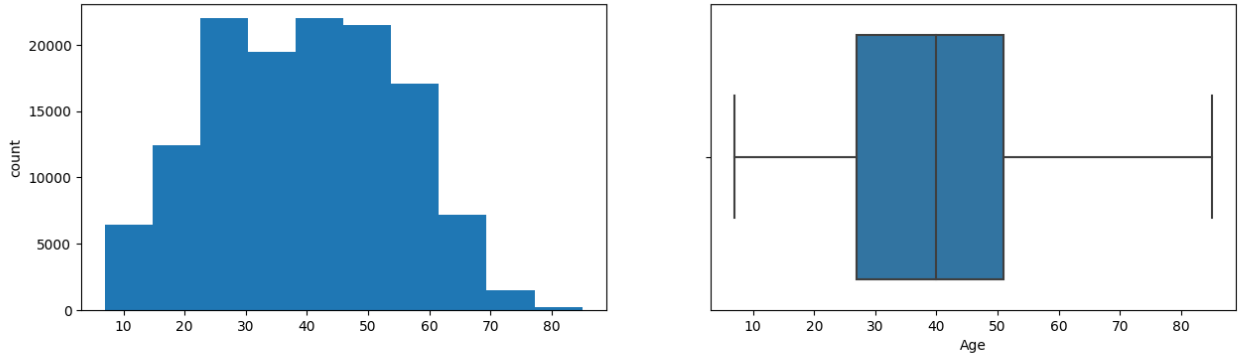


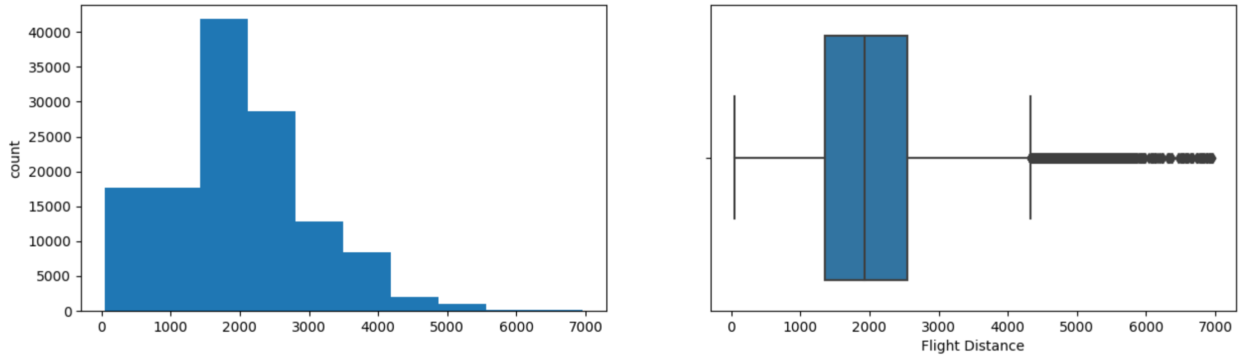
Fig.1 – customer satisfaction graph (left – dissatisfied, right- satisfied)

Checking data for each column is normally distributed.

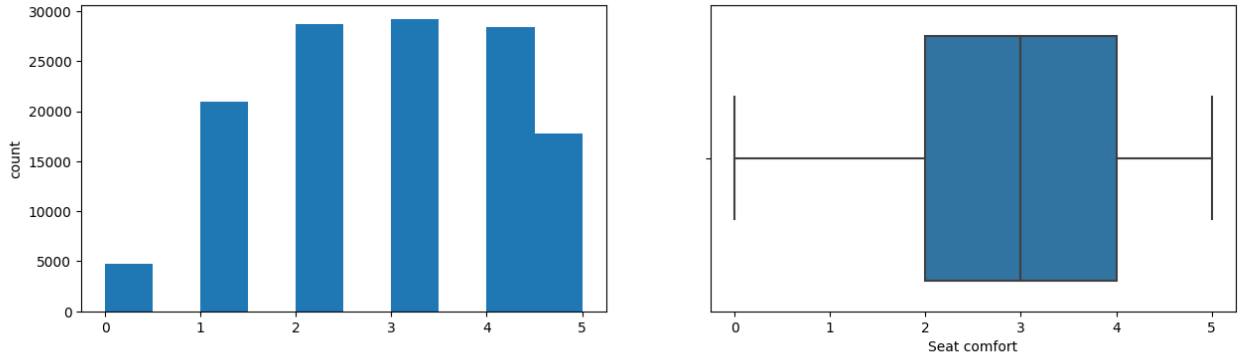
Age



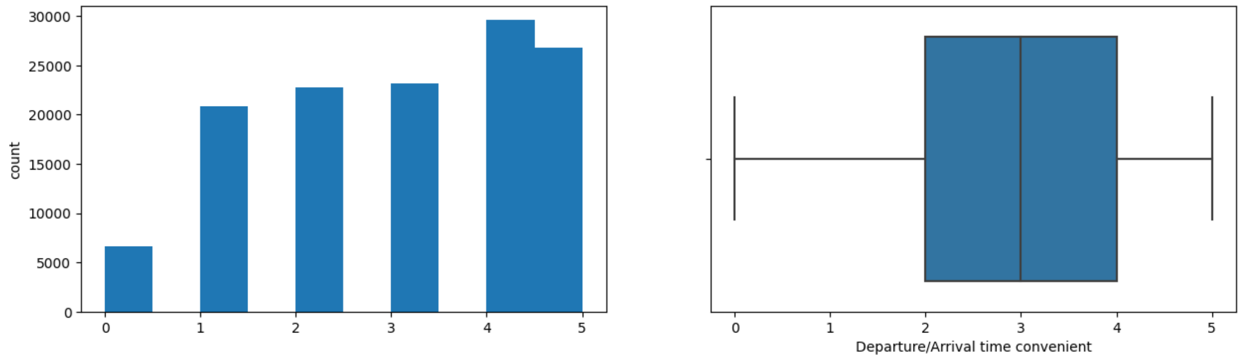
Flight Distance



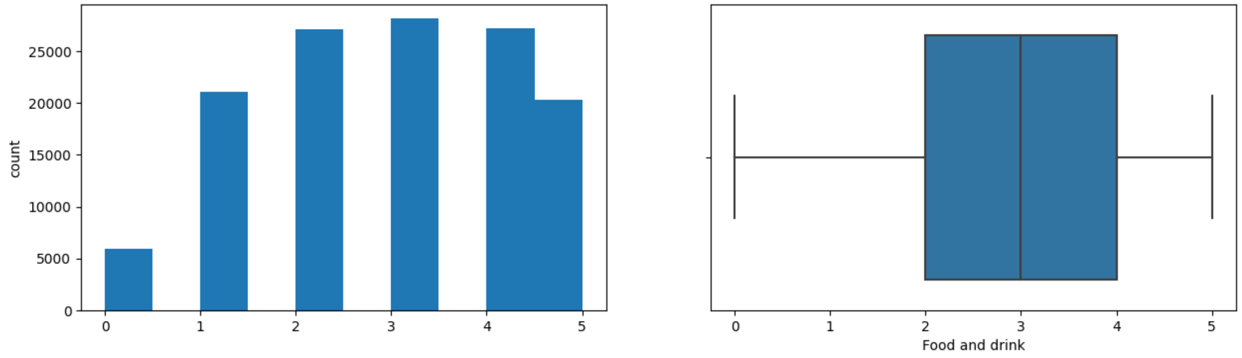
Seat comfort



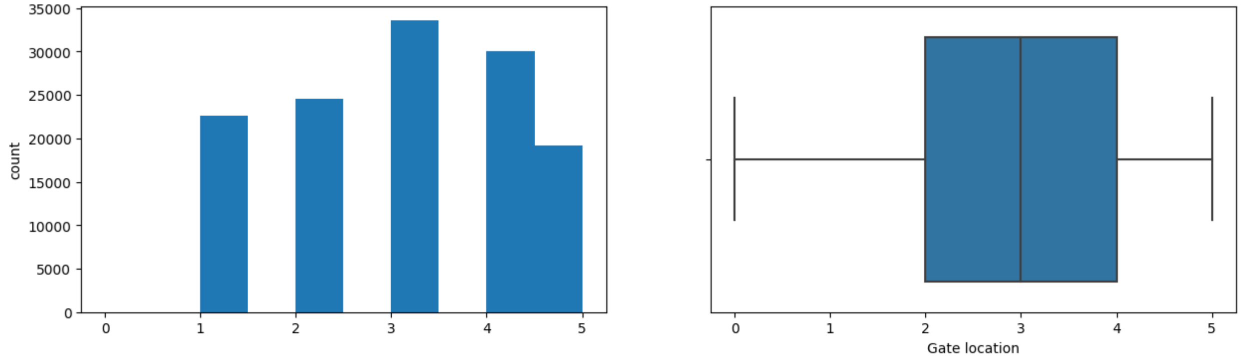
Departure/Arrival Time



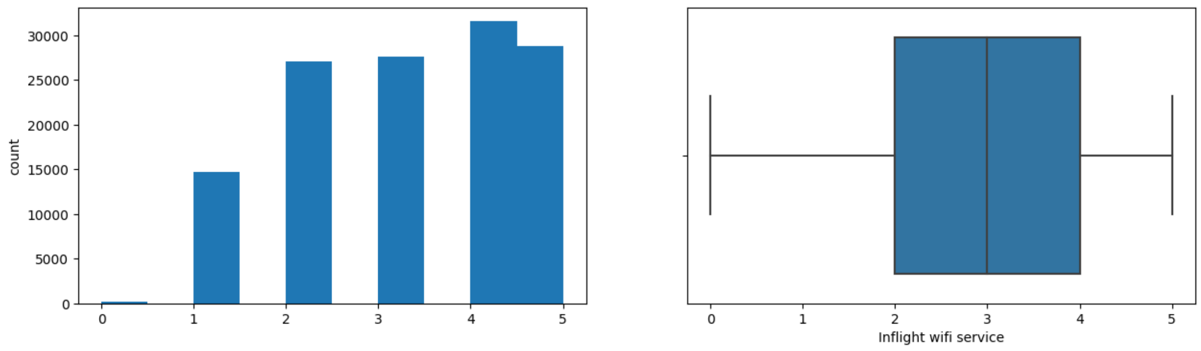
Food and Drink



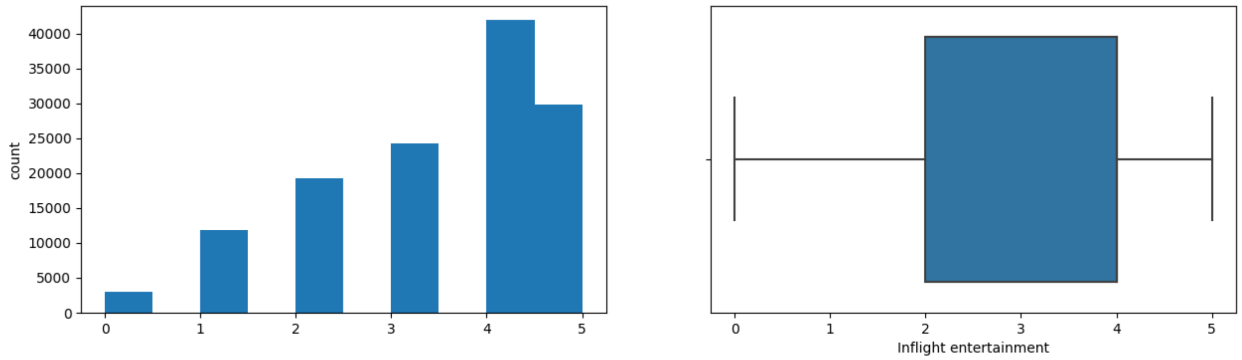
Gate location



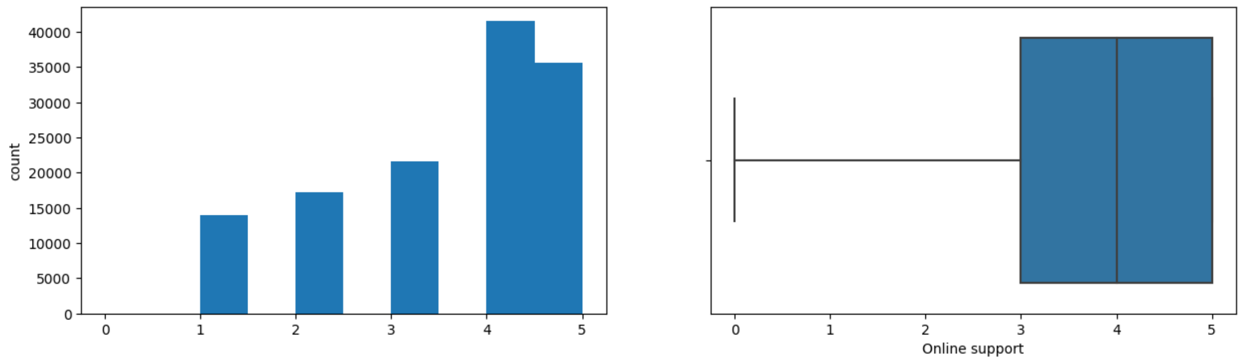
Wi-fi service



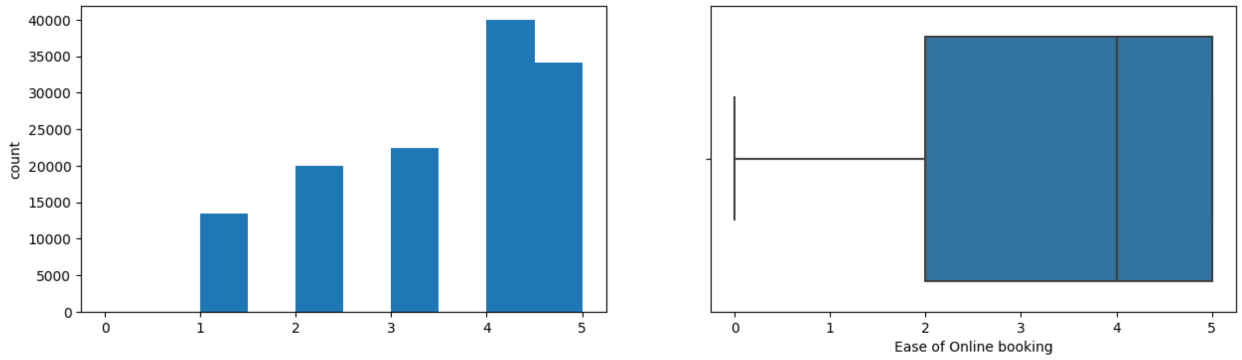
Inflight Entertainment



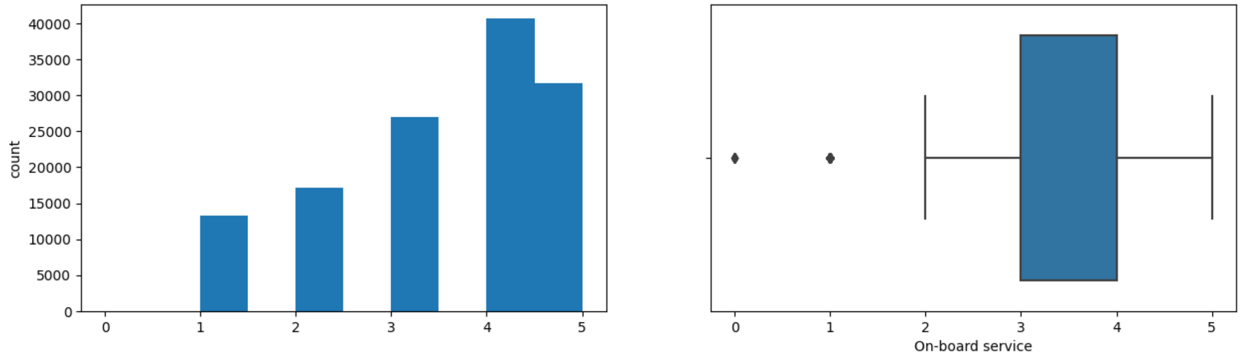
Online Support



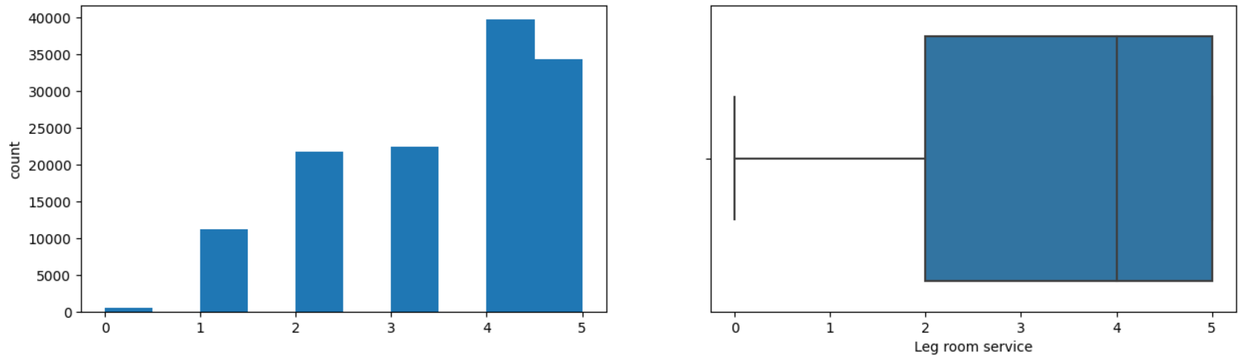
Ease of online booking



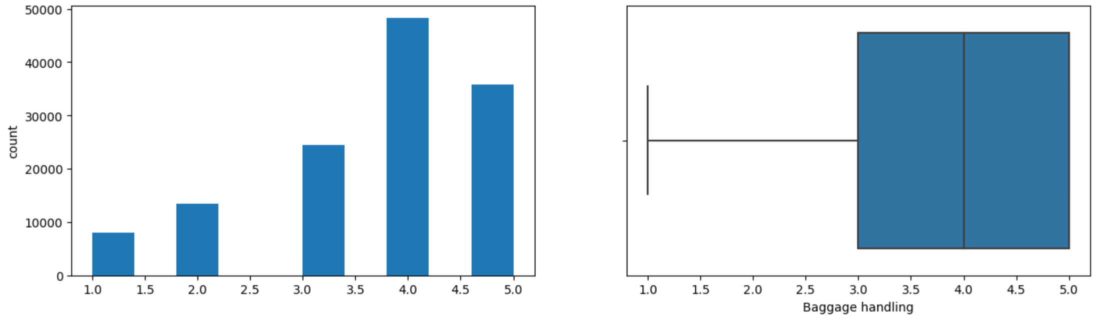
On board service



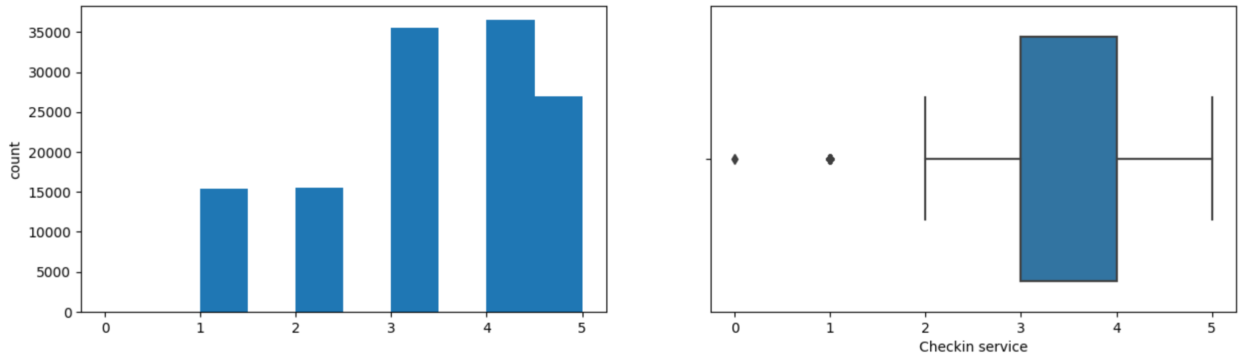
Leg room service



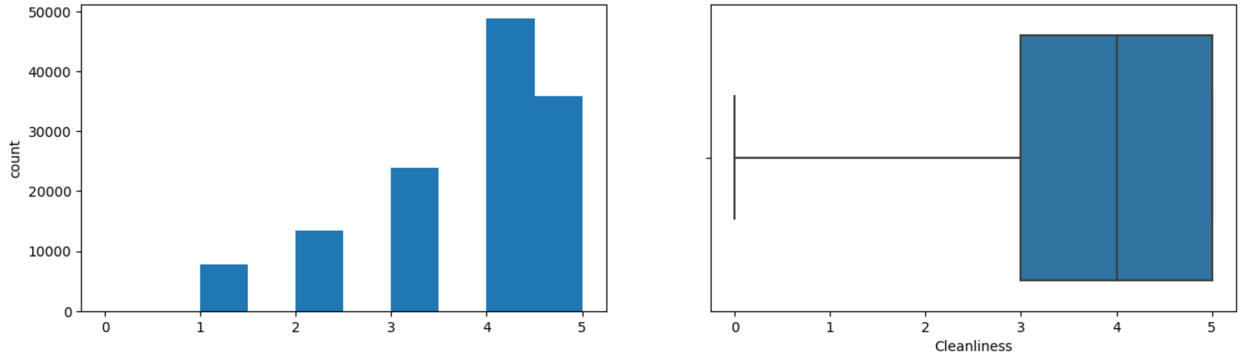
Baggage room



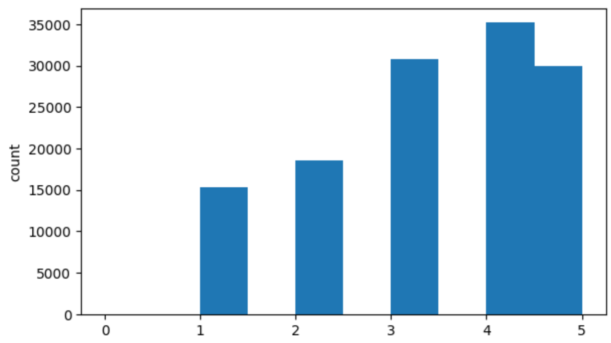
Check-in service

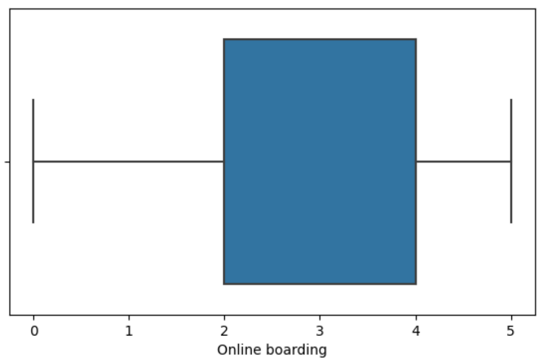


Cleanliness

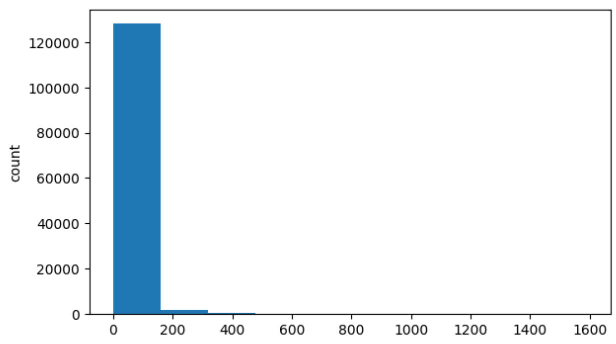


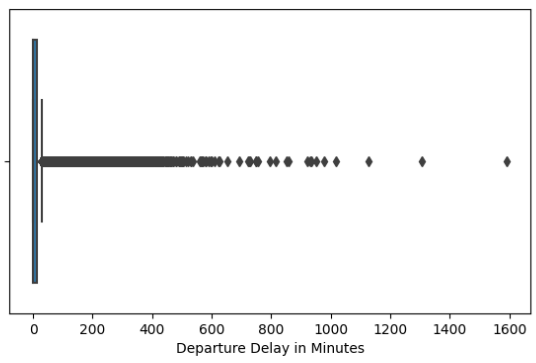
Online onboarding



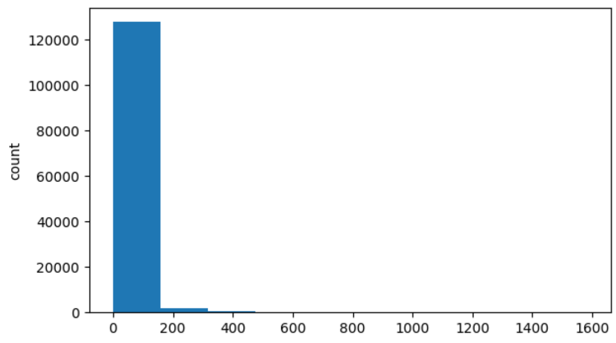


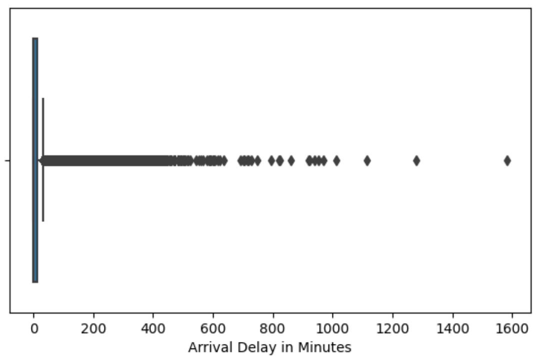
Departure Delays

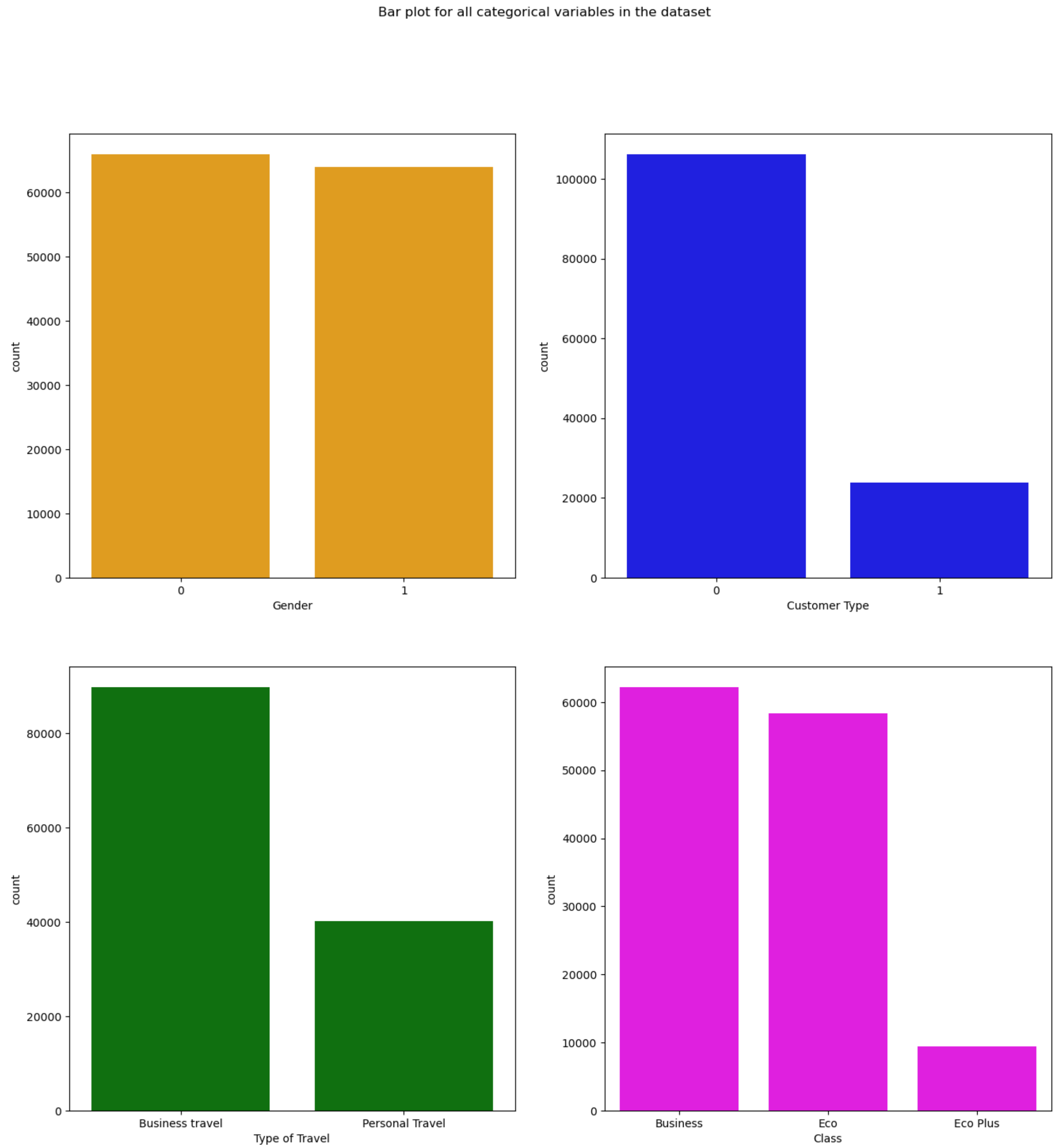




Arrival Delays





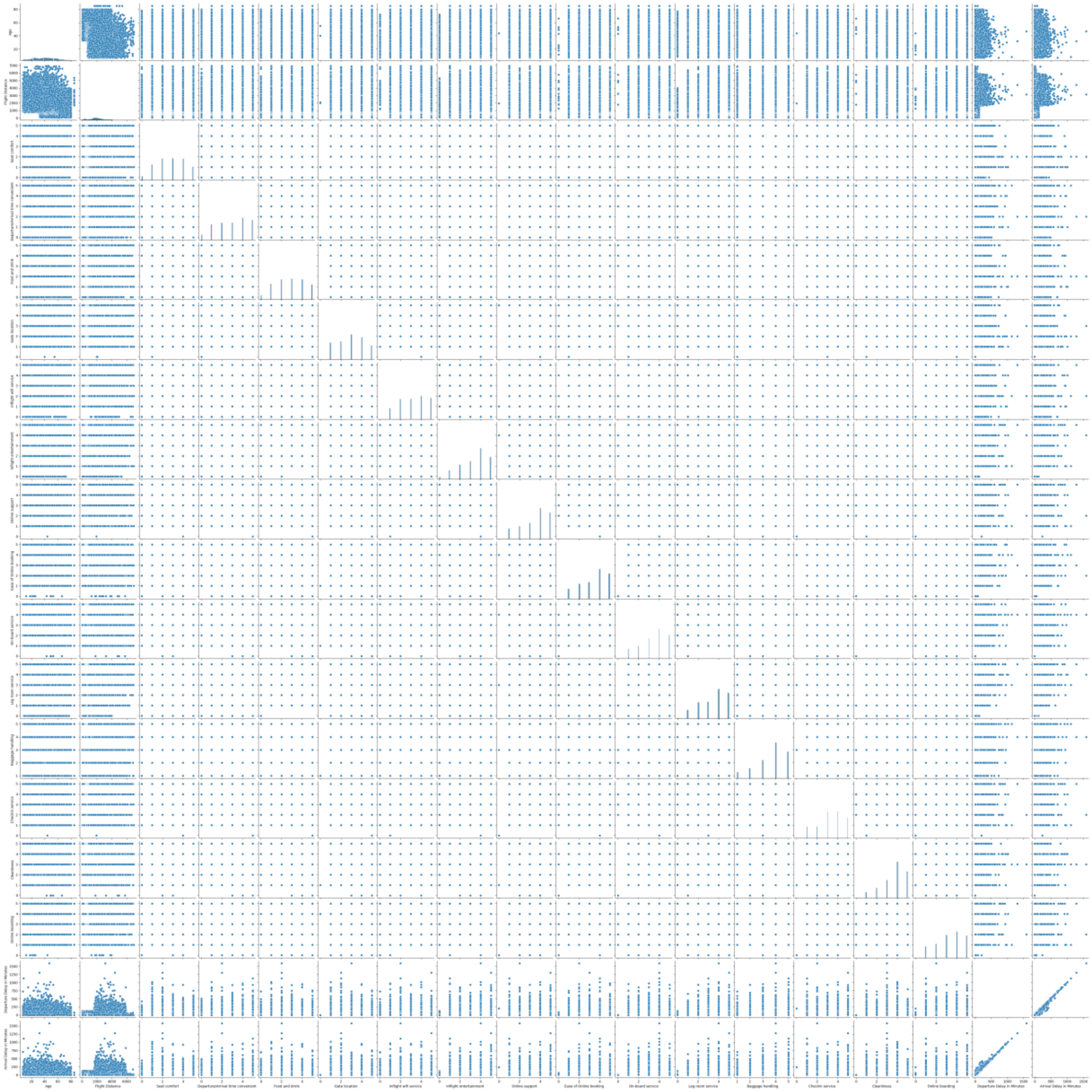


Bar plot for categorical values showing quantity

Inferences from the above charts:

- Both genders are present approx. equal.
- Loyal customers are greater than disloyal customers.
- Airlines having customer who travel for business as compared to personal travel.
- Compared classes that customers have chosen from i.e. business, economy and economy plus.

Correlation among all the columns.



3. Feature Engineering

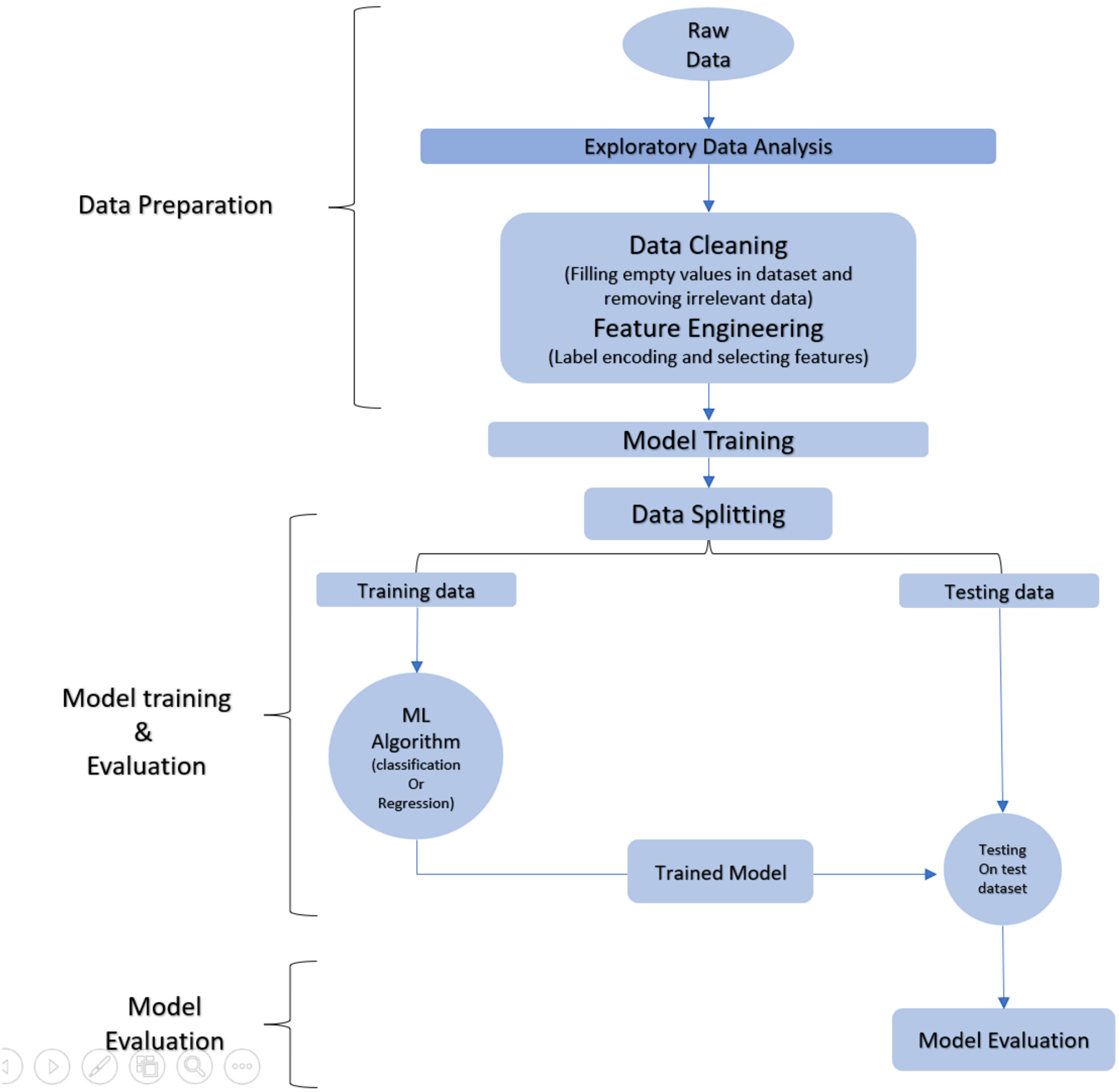
Following actions were performed in the feature engineering:

- Dropped features based on hypothesis testing (chi-square test)
- Selected features are:

  Satisfaction, Gender, Customer Type, Age, Type of Travel, Class, Flight Distance, Departure/Arrival time convenient, Food and Drink, Inflight wifi service, Online support, Ease of Online booking, On-board service, Leg room service, Baggage handling, checkin service, cleanliness, Online onboarding, Departure delay in minutes, Arrival delay in minutes.
- The 4 features have been dropped on Hypothesis Testing which are 'Departure/Arrival time convenient', 'Gate location'.

Note: Hypothesis testing is performed on survey data. features are selected on the basis of sampled data. Domain knowledge has not been taken into account.

4. Model building:


Data Preparation
Raw Data
Exploratory Data Analysis
Data Cleaning
(Filling empty values in dataset and removing irrelevant data)
Feature Engineering
(Label encoding and selecting features)
Model Training
Data Splitting
Model training & Evaluation
Training data
Testing data
ML Algorithm
(classification Or Regression)
Trained Model
Testing On test dataset
Model Evaluation
Model Evaluation

Models accuracy:

- Logistic - 76% Approx.
- Decision Tree - 54% Approx.
- KNN - 79% Approx.
- SVM - 50% Approx.
- XGB - 97% Approx.

Selected model: XGB Classifier

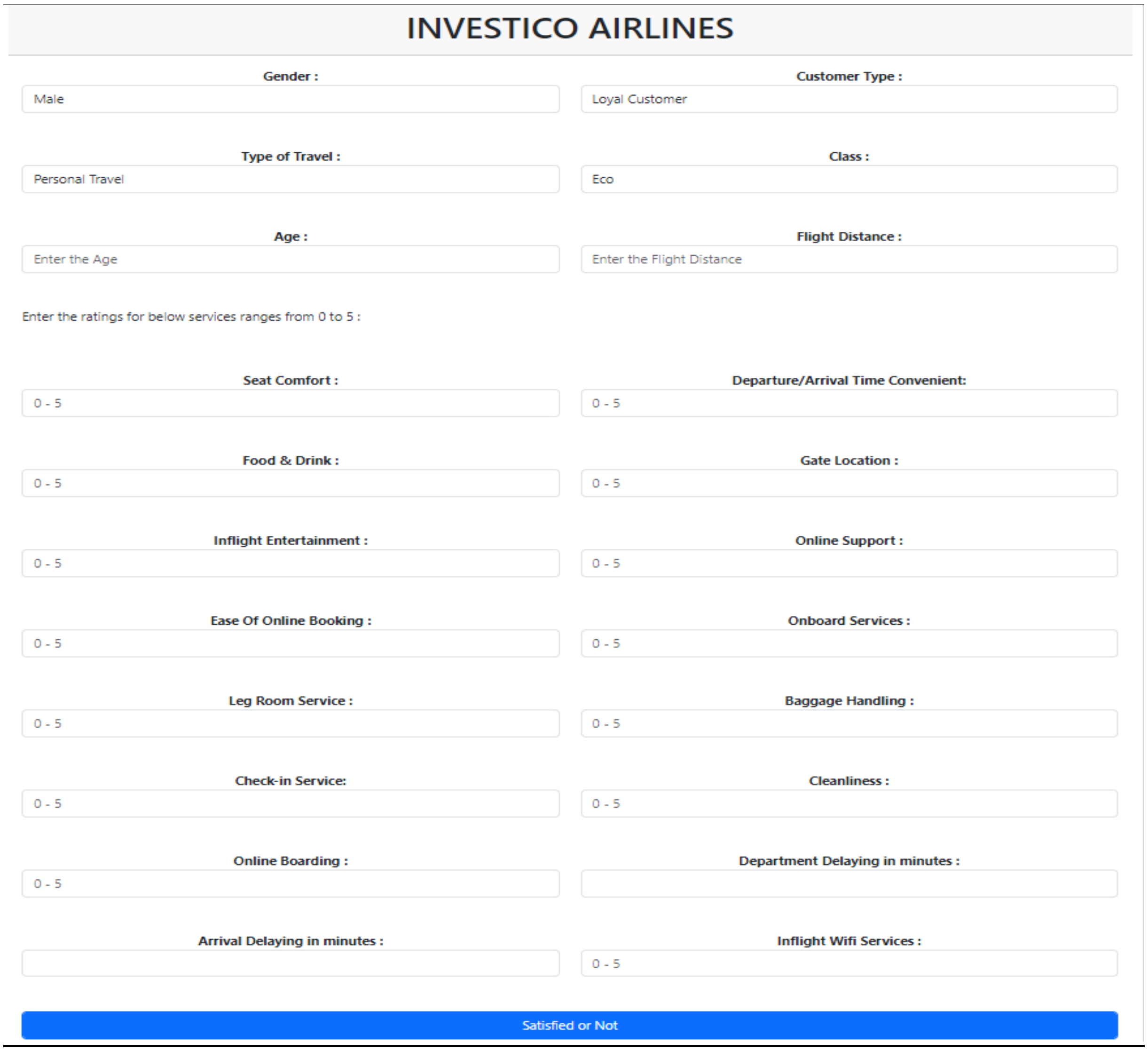
Various machine learning algorithms were explored such as Logistic, Decision Tree, KNN, SVM, Logistic Regression, Decision Tree, ensemble technique – XGB Bagging etc. but the model that gave the highest accuracy is XGB Bagging. As compared to other models, XGB Bagging giving the best model accuracy.

Main.py

```
Import pandas as pd
Import numpy as pd
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
bestfeatures=SelectKBest(score_func=chi2,k=4)

x = data1_norm.iloc[:,1:]  #independent columns
y = data1_norm[0]  #target column is species

#apply SelectKBest class to extract top 10 best features
fit=bestfeatures.fit(x,y)
dfscores=pd.DataFrame(fit.scores_)
dfcolumns=pd.DataFrame(x.columns)

#concat two dataframes for better visualization
featureScore=pd.concat([dfcolumns,dfscores],axis=1)
featureScore.columns=['Specs','Score']
featureScore
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
# importing machine learning models for prediction
import xgboost as xgb
# importing bagging module
from sklearn.ensemble import BaggingClassifier
# loading train data set in dataframe from train_data.csv file
from sklearn import preprocessing
model = BaggingClassifier(base_estimator=xgb.XGBClassifier())
model.fit(x_train,y_train)
pred_train=model.predict(x_train)
print(mean_squared_error(y_train, pred_train))
print((1-mean_squared_error(y_train, pred_train))*100)
```

Output:

Front-end



INVESTICO AIRLINES

Gender :
Male

Customer Type :
Loyal Customer

Type of Travel :
Personal Travel

Class :
Eco

Age :
Enter the Age

Flight Distance :
Enter the Flight Distance

Enter the ratings for below services ranges from 0 to 5 :

Seat Comfort :
0 - 5

Departure/Arrival Time Convenient:
0 - 5

Food & Drink :
0 - 5

Gate Location :
0 - 5

Inflight Entertainment :
0 - 5

Online Support :
0 - 5

Ease Of Online Booking :
0 - 5

Onboard Services :
0 - 5

Leg Room Service :
0 - 5

Baggage Handling :
0 - 5

Check-in Service:
0 - 5

Cleanliness :
0 - 5

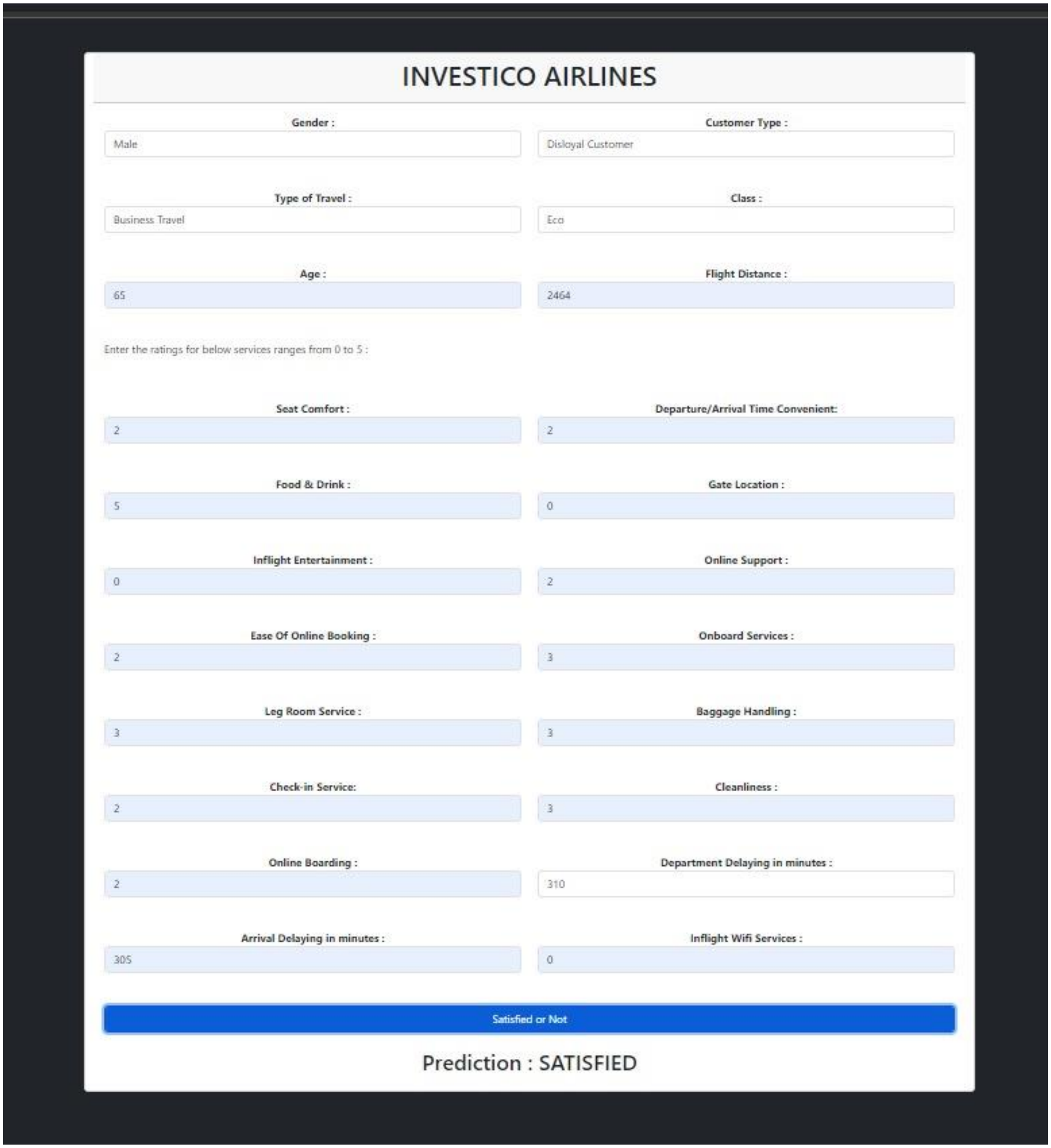
Online Boarding :
0 - 5

Department Delaying in minutes :

Arrival Delaying in minutes :

Inflight Wifi Services :
0 - 5

Result:



INVESTICO AIRLINES

Gender : Male

Customer Type : Disloyal Customer

Type of Travel : Business Travel

Class : Eco

Age : 65

Flight Distance : 2464

Enter the ratings for below services ranges from 0 to 5 :

Seat Comfort : 2

Departure/Arrival Time Convenient: 2

Food & Drink : 5

Gate Location : 0

Inflight Entertainment : 0

Online Support : 2

Ease Of Online Booking : 2

Onboard Services : 3

Leg Room Service : 3

Baggage Handling : 3

Check-in Service: 2

Cleanliness : 3

Online Boarding : 2

Department Delaying in minutes : 310

Arrival Delaying in minutes : 305

Inflight Wifi Services : 0

Prediction : SATISFIED

Invistico Airline uses the following packages and library from python:

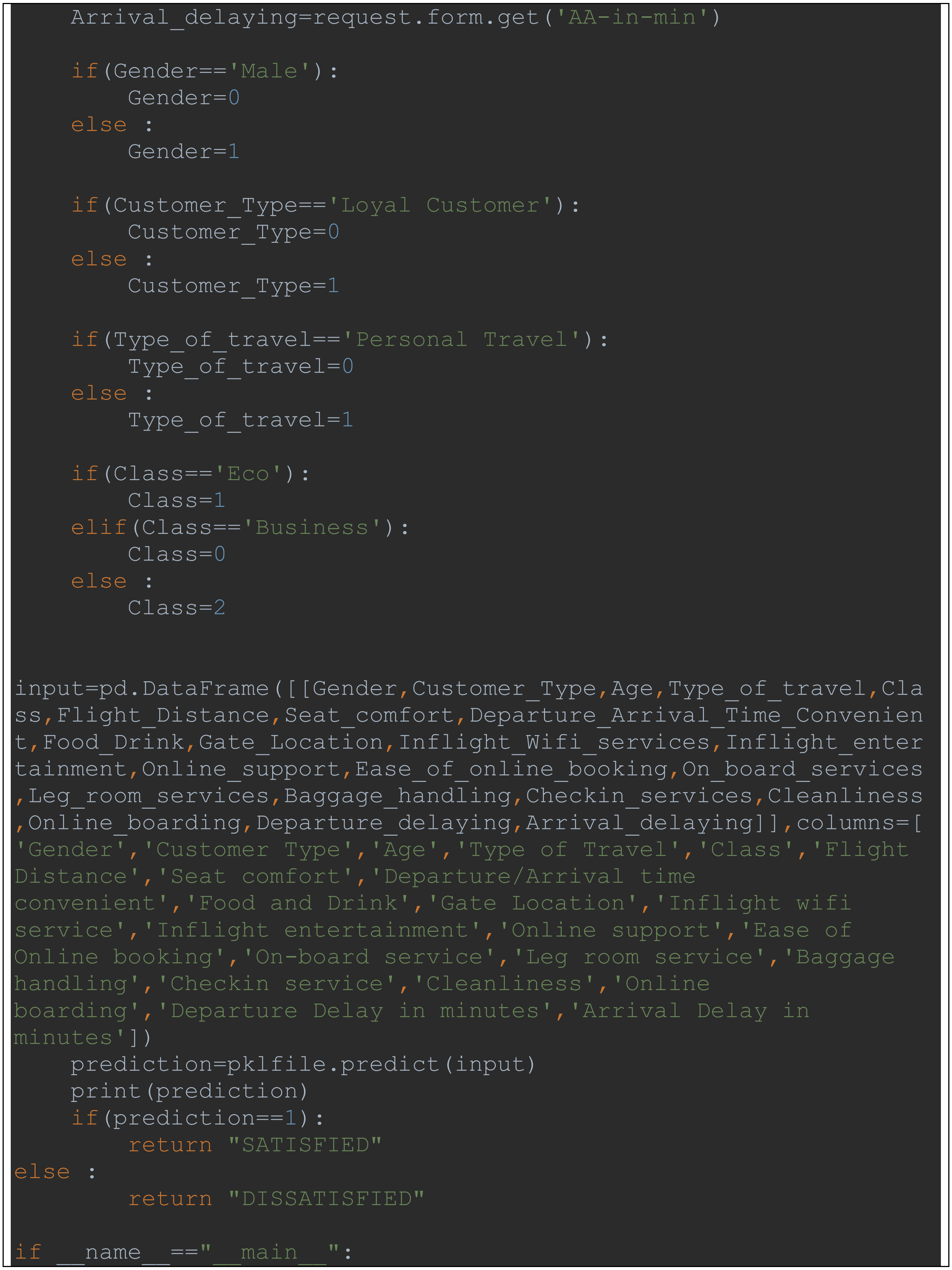
```
import pandas as pd

import numpy as np

from sklearn import preprocessing

import matplotlib.pyplot as plt

import seaborn as sns

import seaborn as sns

import matplotlib.pyplot as plt

from scipy.stats import skew

from numpy import asarray

from sklearn.preprocessing import MinMaxScaler

import warnings

from sklearn.feature_selection import SelectKBest

from sklearn.feature_selection import chi2

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error

import xgboost as xgb
```

```
from sklearn.ensemble import BaggingClassifier

from sklearn import preprocessing
```

5. Deployment using Flask:

Deployment process was done using flask technique.

```
from flask import Flask, render_template, request
import pandas as pd
import pickle

app=Flask(__name__)
data=pd.read_excel('C:\VSCODE\ML TRAINING
PROJECT\Invistico_Airline.xlsx')
pklfile=pickle.load(open('C:\VSCODE\ML TRAINING
PROJECT\mainy.pkl','rb'))
@app.route('/')

def index():
    return render_template('index.html')
# Age  Type of Travel Class  Flight Distance    Seat comfort
Departure/Arrival time convenient  Food and drink Gate location
Inflight wifi service  Inflight entertainment Online support
Ease of Online booking On-board service   Leg room service
Baggage handling   Checkin service    Cleanliness    Online
boarding    Departure Delay in Minutes Arrival Delay in Minutes

@app.route('/predict',methods=['POST'])
def predict():
    Gender=request.form.get('gender')
    Customer_Type=request.form.get('customer-type')
    Age=request.form.get('age')
    Type_of_travel=request.form.get('type-of-travel')
    Class=request.form.get('class')
    Flight_Distance=request.form.get('flight-distance')
    Seat_comfort=request.form.get('seat-comfort')
    Departure_Arrival_Time_Convenient=request.form.get('datc')
    Food_Drink=request.form.get('food-drink')
    Gate_Location=request.form.get('gate-location')
    Inflight_Wifi_services=request.form.get('Iwifi')
    Inflight_entertainment=request.form.get('inflight-ent')
    Online_support=request.form.get('online-support')
    Ease_of_online_booking=request.form.get('eoob')
    On_board_services=request.form.get('onboard-services')
    Leg_room_services=request.form.get('lrs')
    Baggage_handling=request.form.get('bh')
    Checkin_services=request.form.get('cis')
    Cleanliness=request.form.get('clean')
    Online_boarding=request.form.get('online-boarding')
    Departure_delaying=request.form.get('DD-in-min')
```



```
    Arrival_delaying=request.form.get('AA-in-min')

    if(Gender=='Male'):
        Gender=0
    else :
        Gender=1

    if(Customer_Type=='Loyal Customer'):
        Customer_Type=0
    else :
        Customer_Type=1

    if(Type_of_travel=='Personal Travel'):
        Type_of_travel=0
    else :
        Type_of_travel=1

    if(Class=='Eco'):
        Class=1
    elif(Class=='Business'):
        Class=0
    else :
        Class=2


input=pd.DataFrame([[Gender,Customer_Type,Age,Type_of_travel,Cla
ss,Flight_Distance,Seat_comfort,Departure_Arrival_Time_Convenien
t,Food_Drink,Gate_Location,Inflight_Wifi_services,Inflight_enter
tainment,Online_support,Ease_of_online_booking,On_board_services
,Leg_room_services,Baggage_handling,Checkin_services,Cleanliness
,Online_boarding,Departure_delaying,Arrival_delaying]],columns=[
'Gender','Customer Type','Age','Type of Travel','Class','Flight
Distance','Seat comfort','Departure/Arrival time
convenient','Food and Drink','Gate Location','Inflight wifi
service','Inflight entertainment','Online support','Ease of
Online booking','On-board service','Leg room service','Baggage
handling','Checkin service','Cleanliness','Online
boarding','Departure Delay in minutes','Arrival Delay in
minutes'])
    prediction=pklfile.predict(input)
    print(prediction)
    if(prediction==1):
        return "SATISFIED"
else :
        return "DISSATISFIED"

if __name__=="__main__":
    app.run(debug=True, port=5001)
```