Final Report

Batch details	APR 21
Team members	1. Akshay Joshi 2. Ashish Kumar Gupta 3. Manjunath M 4. Nitin Bhojraj Chaudhari 5. Prerana H 6. Vishak subramanyan A
Domain of Project	Airline
Proposed project title	Machine Learning based Airline Customer Satisfaction Prediction using Classification
Group Number	7
Team Leader	Prerana H
Mentor Name	Koneti Naveen Kumar Yadav

Date: 20/10/2021

prerana H

Signature of the Mentor

Signature of the Team Leader

Table of Contents

Sl. NO	Topic	Page No
1	Abstract	3
2	Problem Statement	4
3	Introduction	5
4	Overview of the Final Process	6
5	Step by Step walkthrough of the Solution	
	1. Understanding the dataset	7
	2. EDA and Preprocessing	8
6	Model Building and Training	23
7	Hyper Parameter Tuning and Feature Selection	23
8	Selecting Best Model	25
9	Business Interpretation	26
10	Limitations	26
11	Future Scope	26
12	Learnings	26
13	Final Note	27

ABSTRACT

The sheer size of the airline industry provides a reason to care about it: it affects not only millions of people directly (flyers, pilots, engineers, etcetera), but also millions more indirectly by the heft of its economic presence. Although the XYZ airline industry is strong, it must be ever-vigilant about keeping up with customer demands in order to maintain its continued growth and its continued position as industry leader across regions. Of course, success in this regard requires airlines to know what customers care about in the first place. Discovering what airline customers like and dislike about their flight experiences was the starting point for this project.

PROBLEM STATEMENT

Every business depends on customer satisfaction and the Airline industry runs on the same model. Customer satisfaction depends on various factors of the airlines such as satisfaction level in the food, drinks, online boarding, etc. Our approach to the business problem is to find out the factors which directly influence customer satisfaction. Main goal for the business is to minimize the influence/occurrences of negative factors that affect customer satisfaction.

INTRODUCTION

The machine learning field is continuously evolving. And along with evolution comes a rise in demand and importance. There is one crucial reason why data scientists need machine learning, and that is: ‘High-value predictions that can guide better decisions and smart actions in real-time without human intervention.’

Machine learning as technology helps analyze large chunks of data, easing the tasks of data scientists in an automated process and is gaining a lot of prominence and recognition. Machine learning has changed the way data extraction and interpretation works by involving automatic sets of generic methods that have replaced traditional statistical techniques.

After each trip, passengers are asked about their overall satisfaction as well as their rating of various services. Companies want to use this data to further enhance their services to maximize satisfaction. This isn’t a very straightforward task; first glances may lead to entirely inaccurate decisions as there could be hidden correlations at play. This is the task we handle in the project.

A question might arise, which aspect of flight has the most influence on customers’ overall rating? This is a classic machine learning question that is easy to ask but difficult to answer, the difficulty lying in the potentially subtle interactions among the predictor variables. To tackle this we apply various classification algorithms and deploy the model which gives the best results validated against various parameters viz. Accuracy, Precision, ROC-AUC score etc.

OVERVIEW OF THE FINAL PROCESS

The brief approach for the solution is given below

1. Solution requires model building based on the Classification model approach to predict whether the customer is satisfied with the airline services or not and to identify what actually went wrong.
2. Data cleansing and Pre-Processing are important to have a good cleaned input dataset for the model to predict the expected output. Hence the data cleansing and pre-processing steps are given in a detailed manner.
3. Visualization has been given to understand the dataset that feed into the model. This also helps to understand the structure of dataset.
4. Model Creation: Creating Models for each of the ML classification Algorithms.
5. The benchmarking of outcome has been captured. The performance of the model is tuned based on the different iterations with different parameters.
6. The business derived value based on the outcome of the model is analyzed.
7. Limitations of the model and scope of improvement has been covered.
8. The lessons learnt on each of the step of the project is noted down and summary is provided as Learnings.

STEP BY STEP WALKTHROUGH OF THE SOLUTION

1. Understanding the Dataset

The dataset has 22 features:

Gender	Details of gender
customer_type	Types of customer : Loyal or Disloyal
age	Age of passengers
type_of_travel	Business or Personal
customer_class	types of customer class
flight_distance	distance between Source and Destination
inflight_wifi_service	inflight_wifi_service ratings
departure_arrival_time_convenient	departure_arrival_time_convenient ratings
ease_of_online_booking	ease_of_online_booking ratings
gate_location	gate_location ratings
food_and_drink	food_and_drink ratings
online_boarding	online_boarding ratings
seat_comfort	seat_comfort ratings
inflight_entertainment	inflight_entertainment ratings
onboard_service	onboard_service ratings
leg_room_service	leg_room_service ratings
baggage_handling	baggage_handling ratings
checkin_service	checkin_service ratings

inflight_service	inflight_service ratings
cleanliness	cleanliness ratings
departure_delay_in_minutes	departure delay from source
arrival_delay_in_minutes	arrival delay on destination

Target	**Value count**	**Description**
neutral or dissatisfied	73452	Customer is not satisfied
satisfied	56428	Customer is satisfied

2. EDA and Preprocessing:

a. Dropped the insignificant attribute id from the dataset.

b. Check the presence of null values in the data.

c. Only arrival_delay_in_minutes has around 0.3% of null values.

d. Carried out KNN imputation of the null values for the attribute arrival_delay_in_minutes.

e. Carrying out necessary type casting for the attributes.

f. Checking for presence of skewness in the data, there is some level of skewness present in the data.

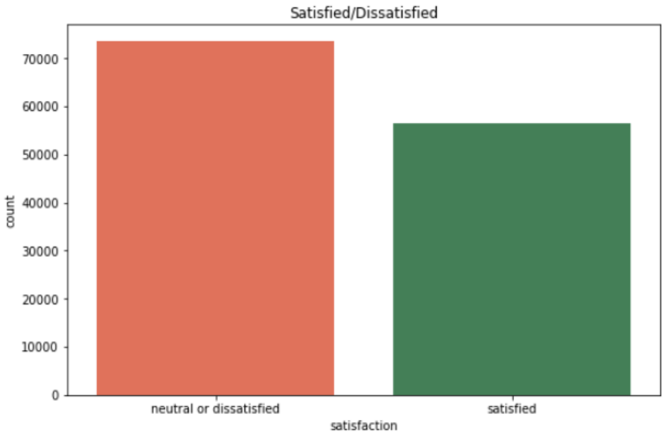
g. Separating the categorical and numerical attributes.

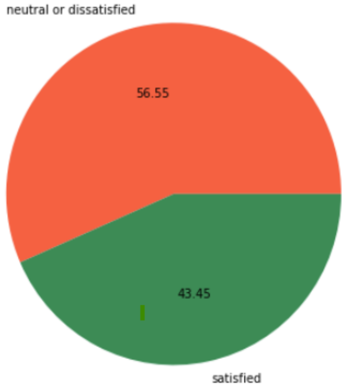
h. Applied statistical tests on the data.

i. Univariate Analysis.

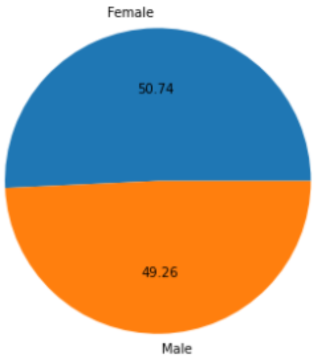
Some of the insights drawn from univariate analysis:

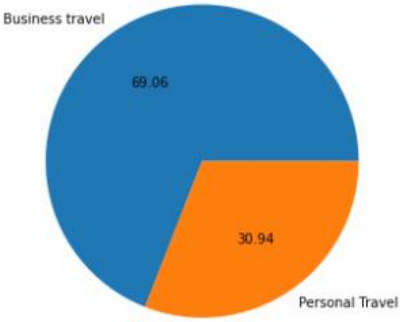
Target Feature: Satisfaction

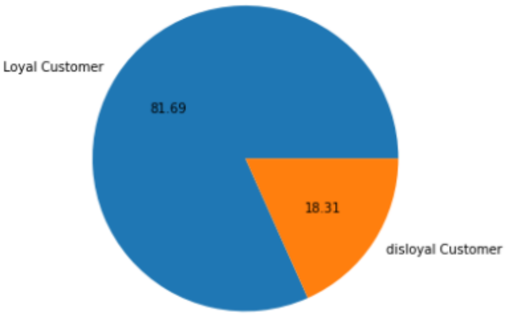


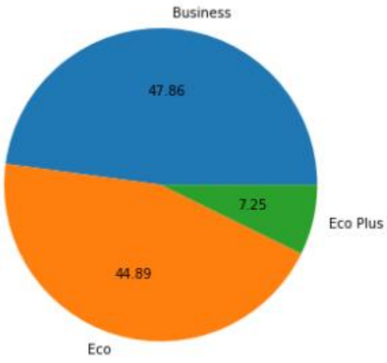


- On analyzing target feature, we can say that the dataset is balanced.
- From the dataset, we can see more passengers are dissatisfied with the current ongoing passenger facing facilities provided by the airline franchise.

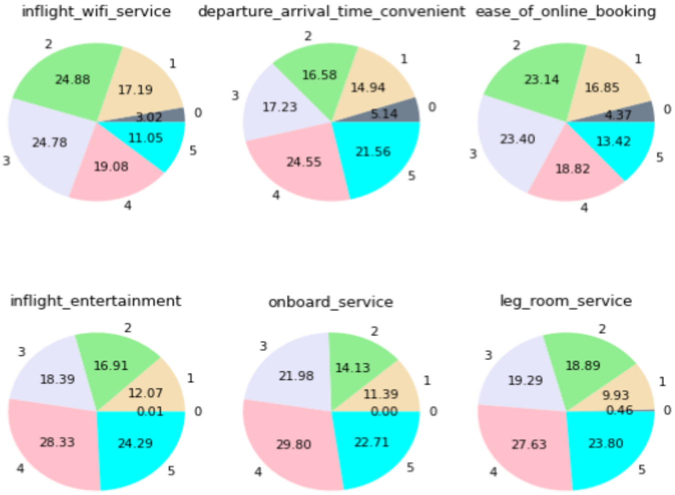


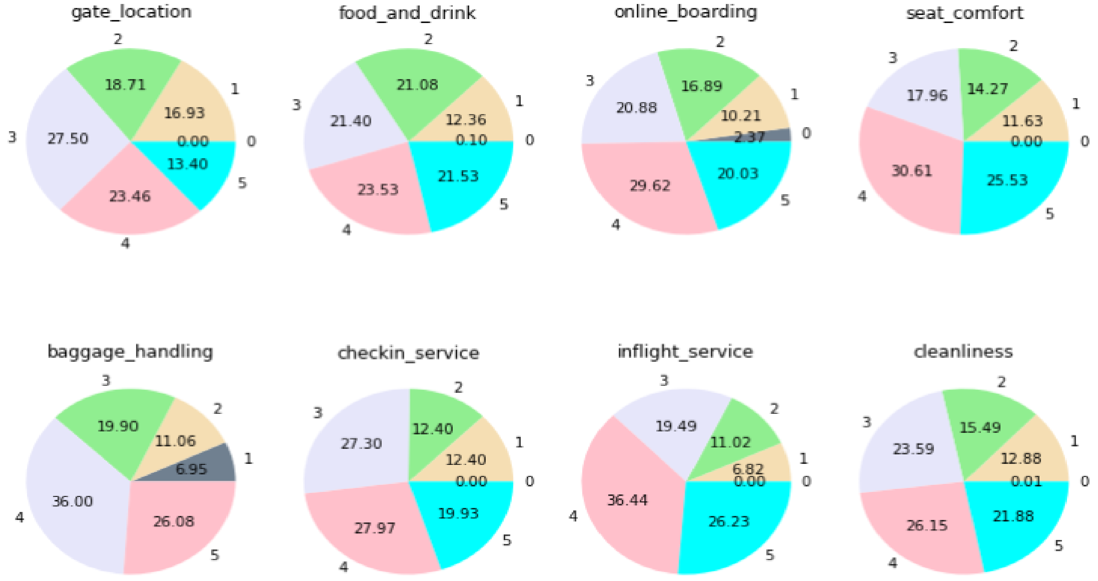






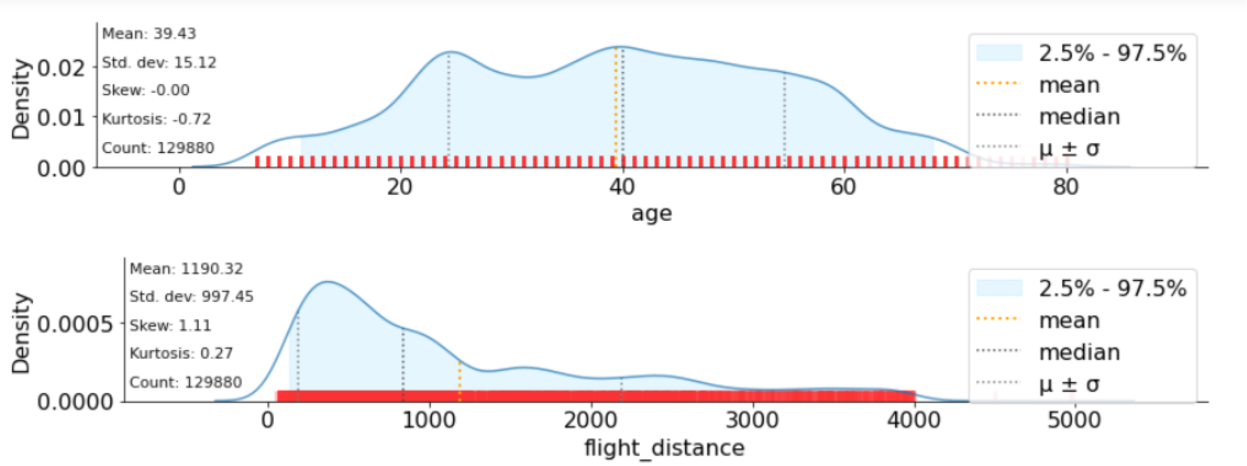
- The airline franchise is commonly used by both genders male and female. And the gender proportion for the airline passenger is also similar i.e., 50:50.
- In the past years, the airline company has more loyal customers in comparison to disloyal ones. The ratio for loyal to disloyal is approximately 81:19.
- The airline franchise has attracted more business travelers. The ratio of business travelers to personal travelers is approximately 69:31.
- The airline franchise has more customers who prefers to travel in business or economic class. There are very few people who opt for economic plus class.



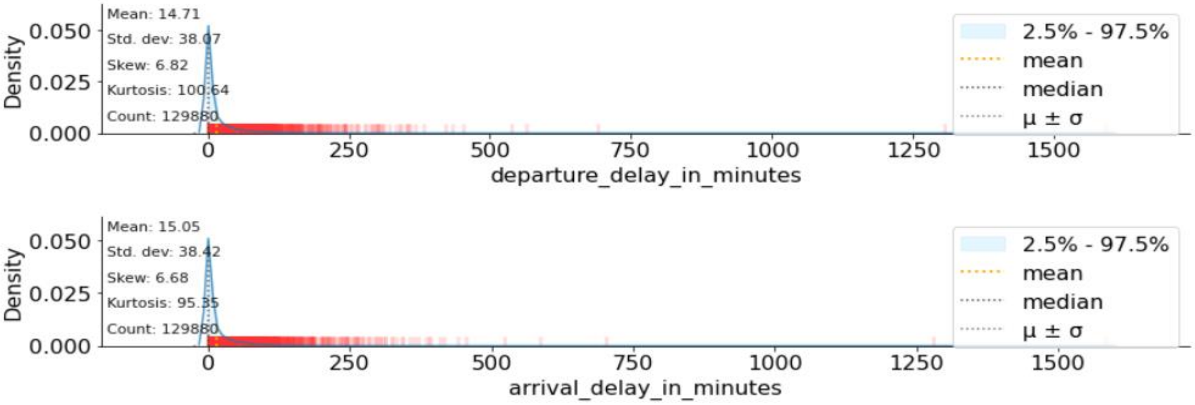


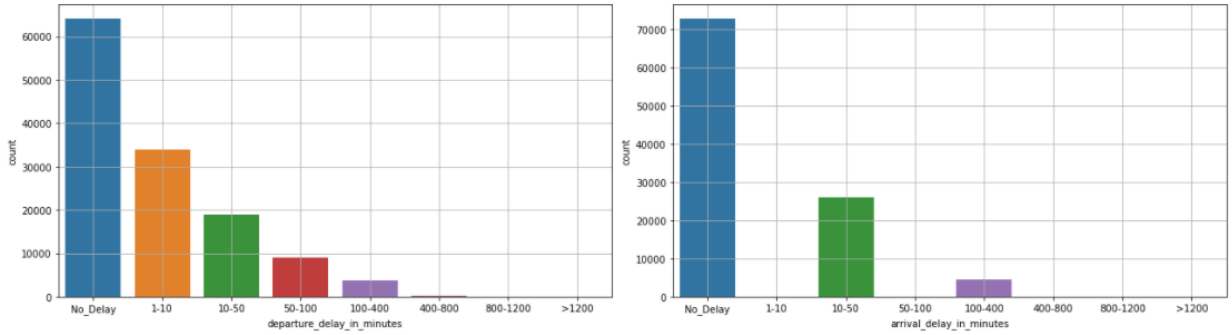
- **Inflight wifi services**: Majority of customers have rated 2 or 3 stars for Inflight wifi services, Only 11% of customers have given 5 star rating. On whole, only 30% customers are satisfied with inflight-wifi service.
- **Ease of online booking**: Majority of customers have rated 2 or 3 stars, only 32% of customers are satisfied with online booking system facility
- **departure_arrival_time_convenient**: Majority of them rated 4 and 5 implying that only 45% of customers are satisfied with airline flight timings and might have not have faced delay.
- The 55% of travelers have faced flight delay issues.
- **Gate Location**: Majority customers have given 3 or 4 as the rating, only 13% of customers has given 5 as rating.
- On whole, 36% of travelers are satisfied with the gate location.
- **Food and drink**: majority of 2, 3 and 5 rating in equal percentages, on whole, 45% of travelers satisfied with food and drink.
- **Online boarding**: Many customers have rated for 4. On whole, 49% travelers are satisfied with online boarding facility.
- **Seat comfort**: Many customers have rated for 4. On whole, 56% customers are satisfied with seat comfort.
- **Inflight Entertainment**: Many customers have rated for 4. On whole, 52 % travelers are satisfied with inflight entertainments.
- **Onboard Services**: Many customers have rated for 4. On whole, 52% travelers are satisfied with onboard service.
- **Leg Room Services**: Many customers have rated for 4.On whole, 50% travelers are satisfied with leg_room_service.
- **Baggage Handling**: Many customers have rated for 4. On whole, 62 % travelers are satisfied with baggage handling.
- **Inflight Service**: Many customers have rated for 4.On whole, 62% travelers are satisfied with inflight_service.
- **Cleanliness**: Many customers have rated for 4.On whole, 47% travelers are satisfied with cleanliness.

- **Checkin-service**: 47% travelers are satisfied with checkin-service.



- The airline franchise has more travelers in the age group 22 to 57.
- More traveler have opted the airline franchise for short distance travel (10-800 miles).

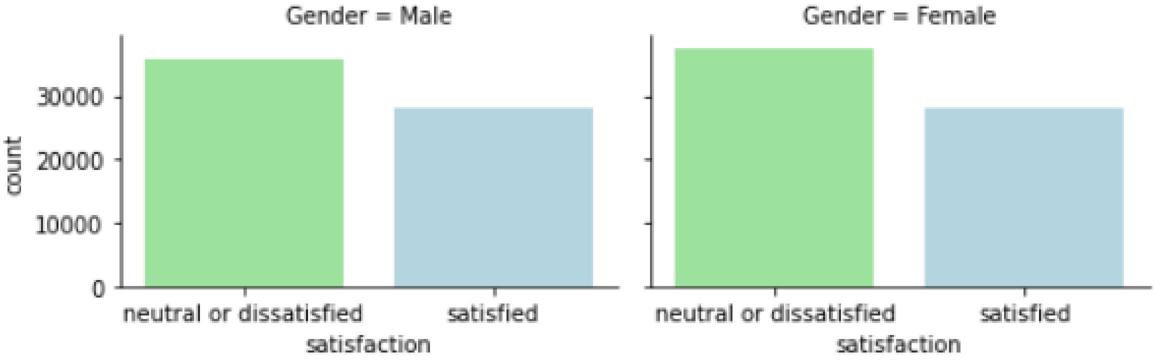




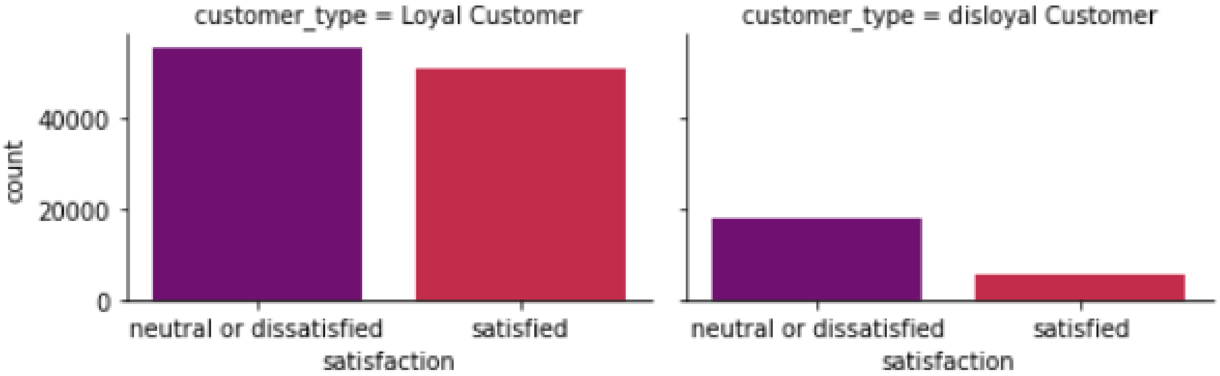
More than 50% of the travelers have not faced delay in departure and arrival time.

j. Bi-variate Analysis

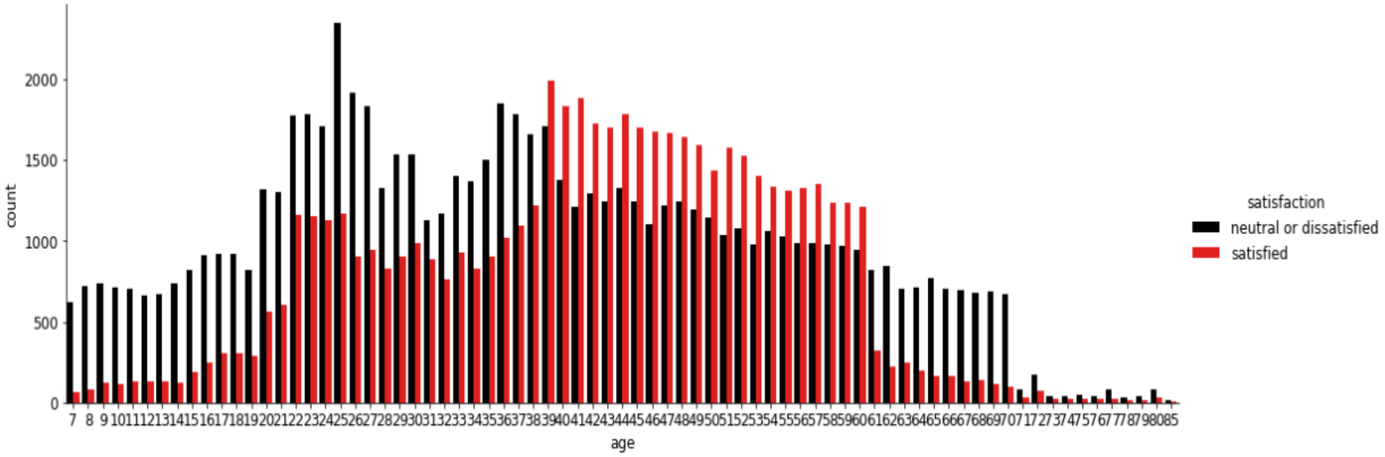
Some inferences from Bivariate Analysis are:



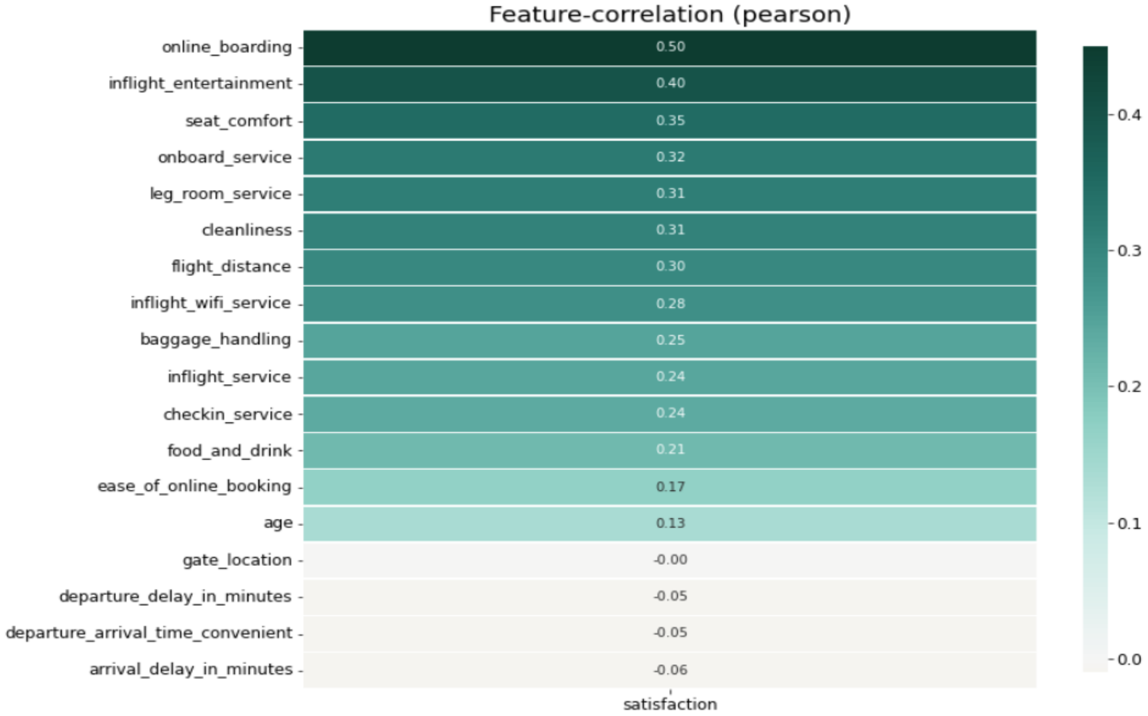
- It is observed that gender distribution of neutral/dissatisfied and satisfied customers are similar.
- For both male and female passengers, number of neutral/dissatisfied customers are on the higher when compared to number of satisfied customers.



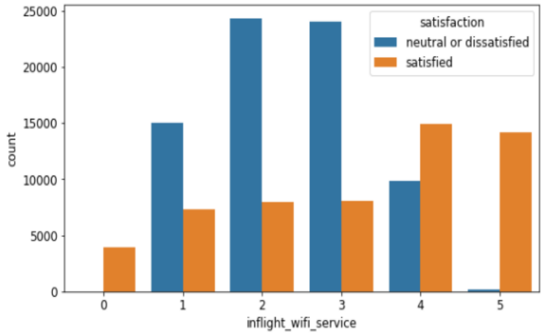
- Even among loyal passengers which are high in number, the ratio of satisfied and neutral/dissatisfied ones are almost close to 49:51.

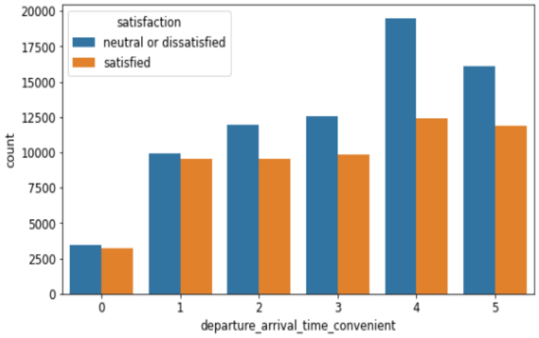


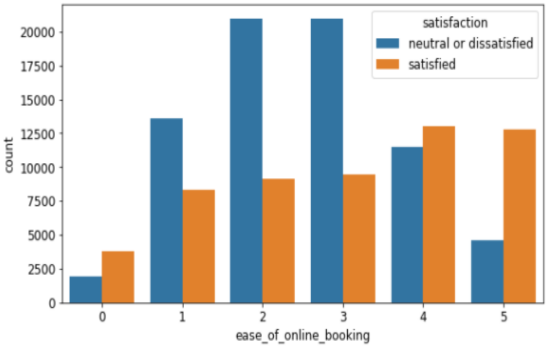
- From age 7-to-38 and from age 61-to-79, the number of neutral/dissatisfied passengers is very high compared to satisfied passengers.
- On the other hand, in age group 39-60, the number of satisfied passengers is higher compared to neutral/dissatisfied passengers

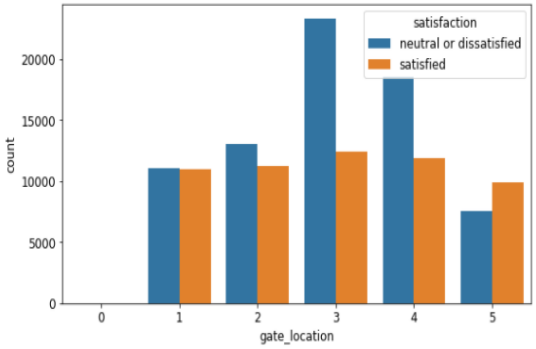


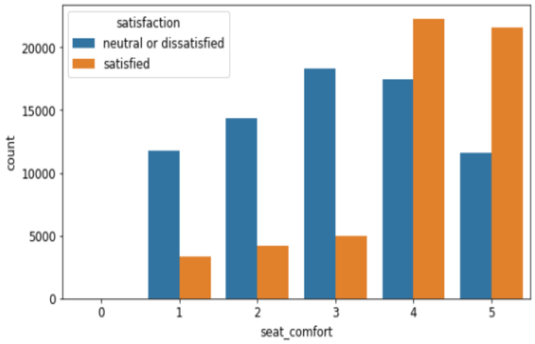
- Most of the independent rating features have moderate correlation with the target feature (satisfaction).

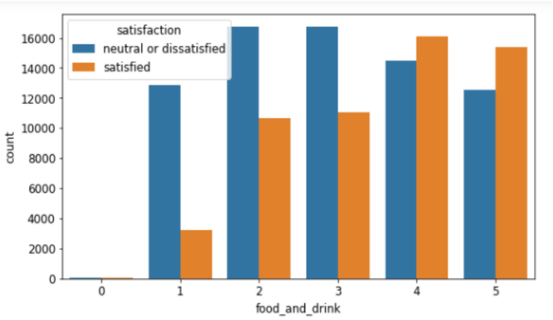


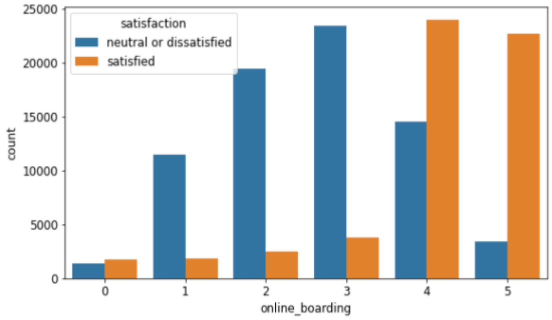


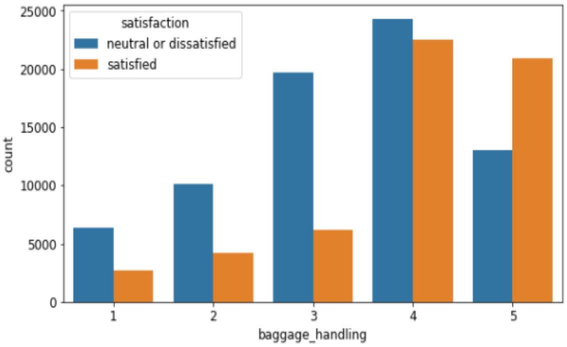


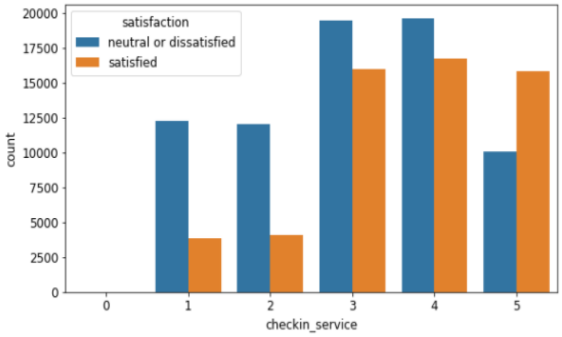


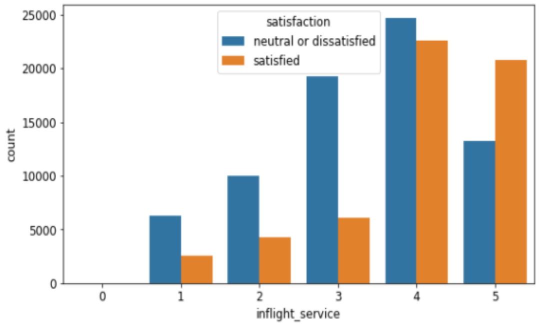


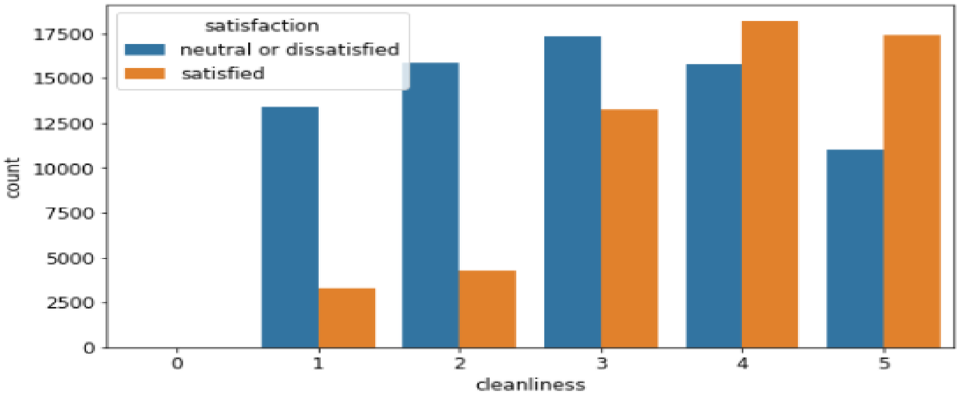


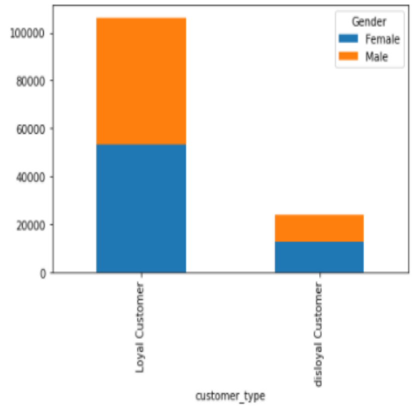


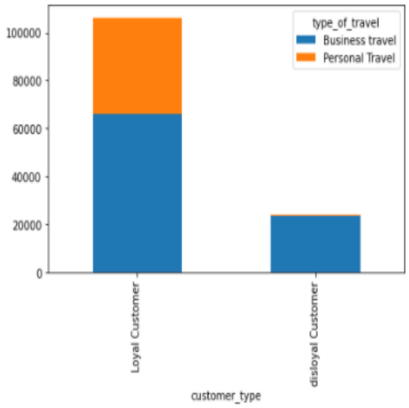


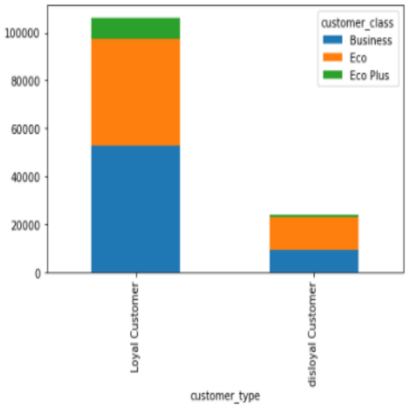

satisfaction
neutral or dissatisfied
satisfied
count
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
checkin_service

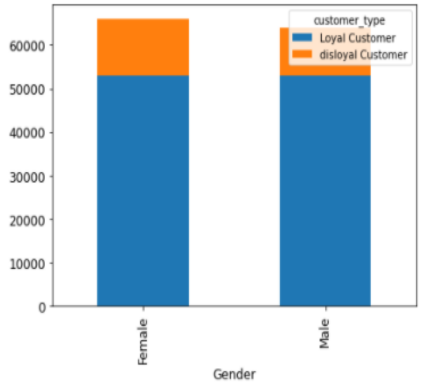

satisfaction
neutral or dissatisfied
satisfied
count
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
inflight_service

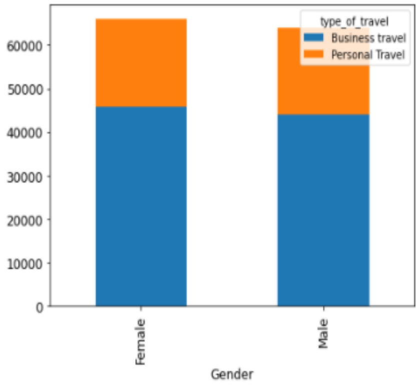

satisfaction
neutral or dissatisfied
satisfied
count
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
cleanliness

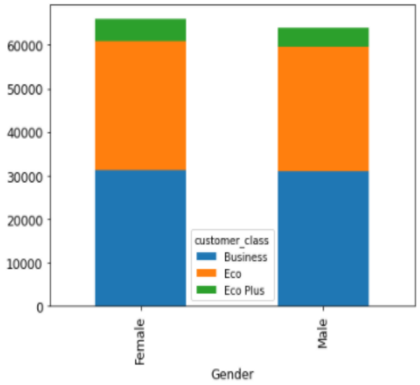

Gender
Female
Male
100000
80000
60000
40000
20000
0
Loyal Customer
disloyal Customer
customer_type

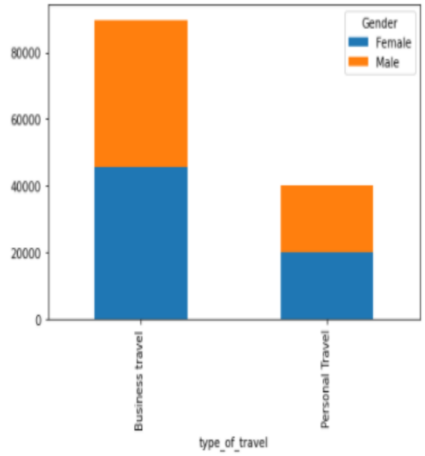

type_of_travel
Business travel
Personal Travel
100000
80000
60000
40000
20000
0
Loyal Customer
disloyal Customer
customer_type

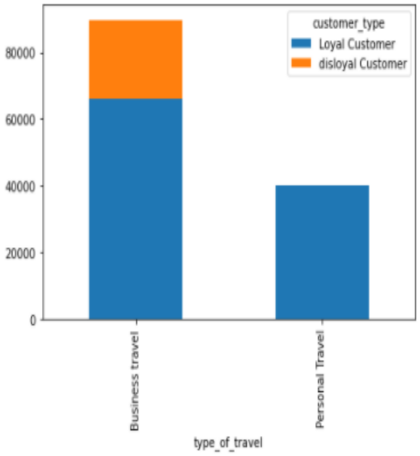

customer_class
Business
Eco
Eco Plus
100000
80000
60000
40000
20000
0
Loyal Customer
disloyal Customer
customer_type

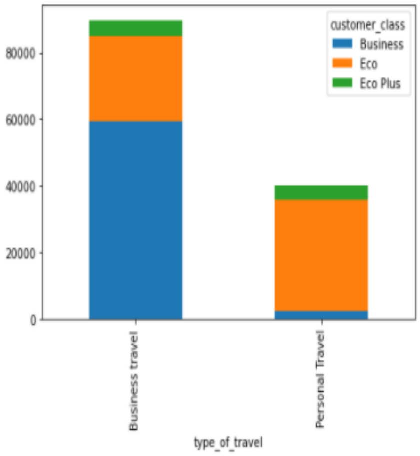

customer_type
Loyal Customer
disloyal Customer
60000
50000
40000
30000
20000
10000
0
Female
Male
Gender

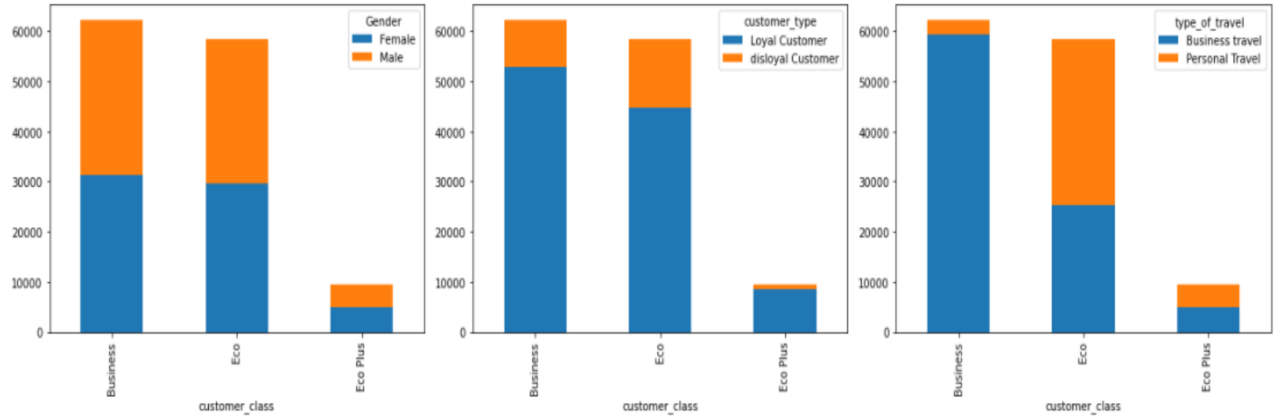

type_of_travel
Business travel
Personal Travel
60000
50000
40000
30000
20000
10000
0
Female
Male
Gender


customer_class
Business
Eco
Eco Plus
60000
50000
40000
30000
20000
10000
0
Female
Male
Gender


Gender
Female
Male
80000
60000
40000
20000
0
Business travel
Personal Travel
type_of_travel


customer_type
Loyal Customer
disloyal Customer
80000
60000
40000
20000
0
Business travel
Personal Travel
type_of_travel

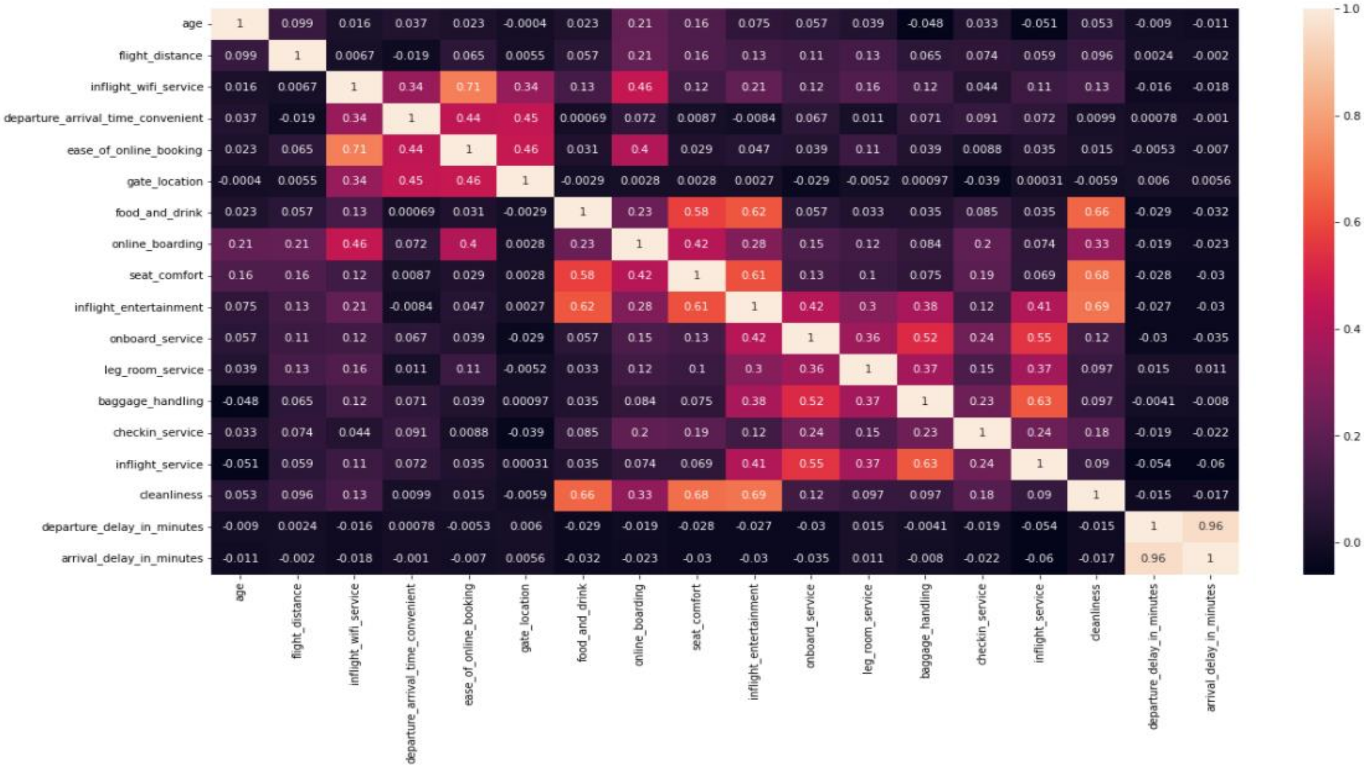

customer_class
Business
Eco
Eco Plus
80000
60000
40000
20000
0
Business travel
Personal Travel
type_of_travel



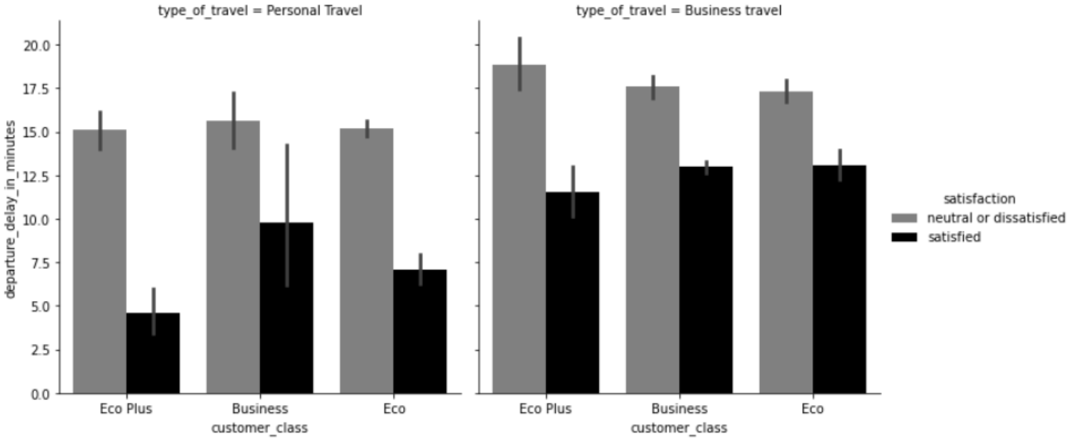
- It is observed that gender distribution in customer_types, type of travel and customer_classes are similar.
- More disloyal customers are business travelers only.
- More % of business travelers are considered as loyal customers with the airline.
- Very less % of personal travelers are considered as disloyal customer in comparison to business travelers.
- We can observe more loyal customer traveling in eco plus class.
- There is an equal proportion of Male and Female for Loyal and Disloyal Customers.

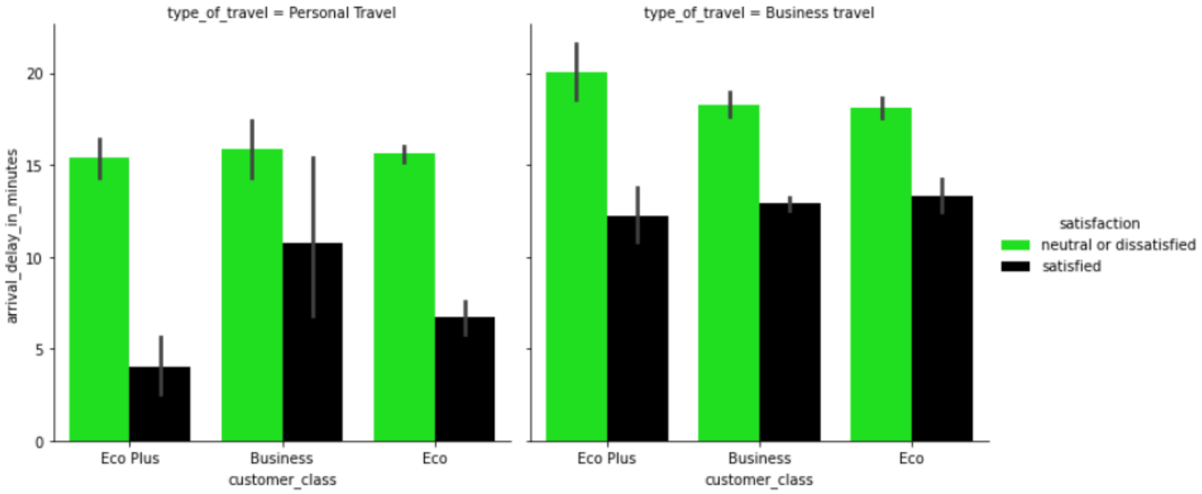
k. Multi-variate Analysis

Some inferences from Multi-variate Analysis are:

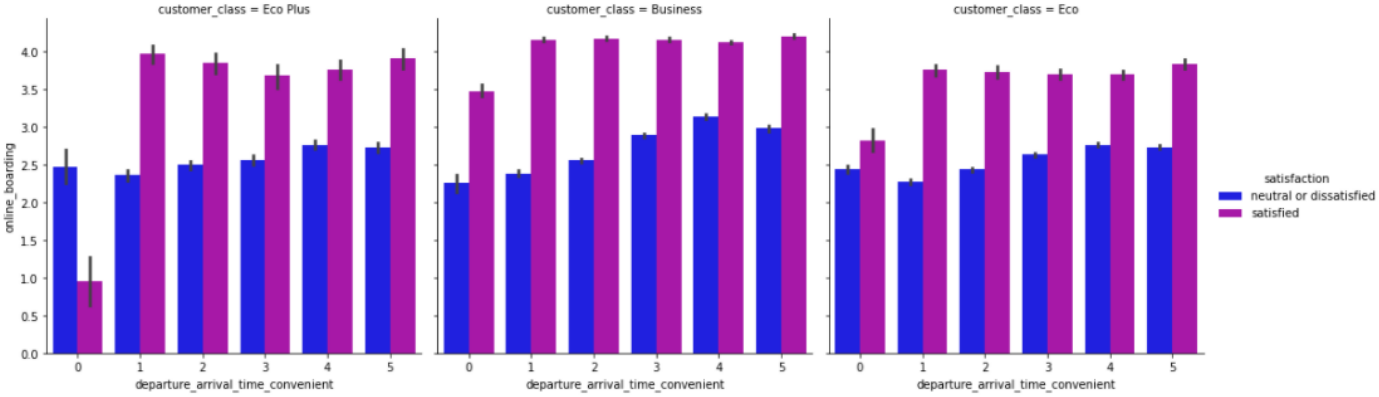


"Ease_of_Online_booking" is highly correlated with **"Inflight_wifi_service"** and **"Inflight_service"** is highly correlated with **"Baggage_handling"**. But no pair is having correlation coefficient exactly equal to 1. Therefore there is no perfect multi-collinearity. Hence we are not discarding any variable.

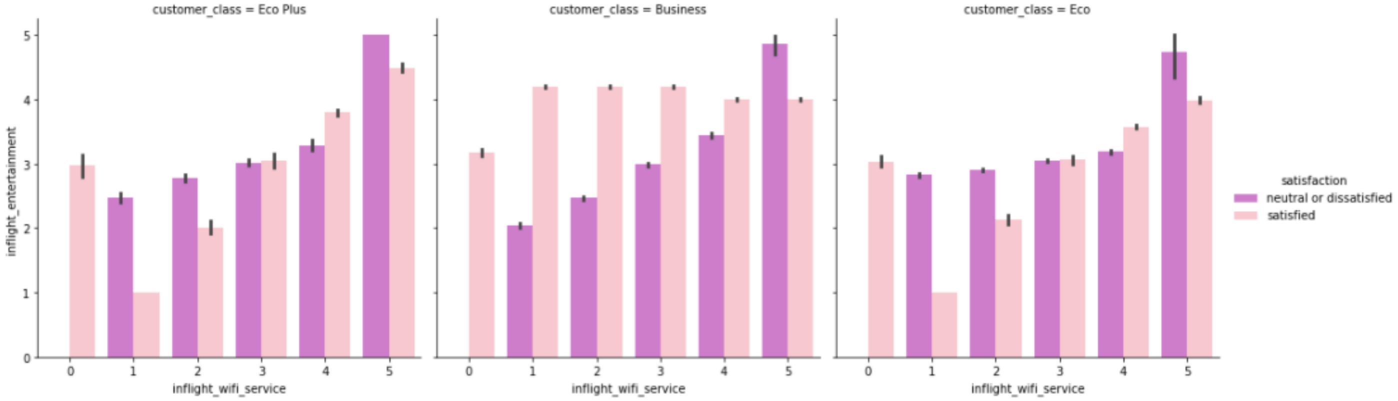




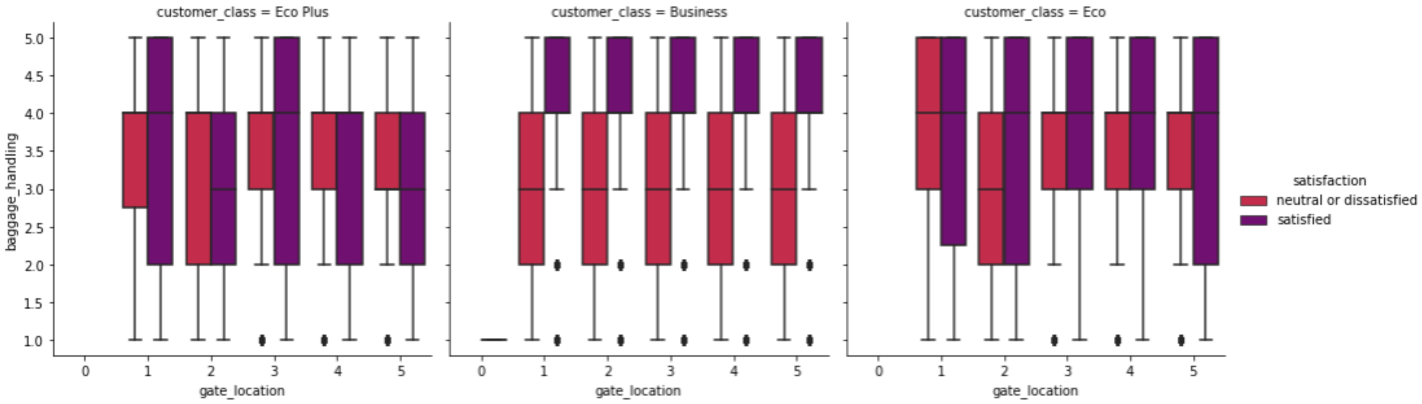
- For any customer class and for any type of travel, customers are dissatisfied when there is a delay in the departure or in arrival.



- For Eco Plus class, very inconvenient Departure/Arrival time i.e., Departure/Arrival_time_convenient = 0 has high number of neutral/dissatisfied passengers, even when online boarding is on positive side. For other combinations, the number of satisfied passengers are higher compared to number of neutral/dissatisfied passengers.

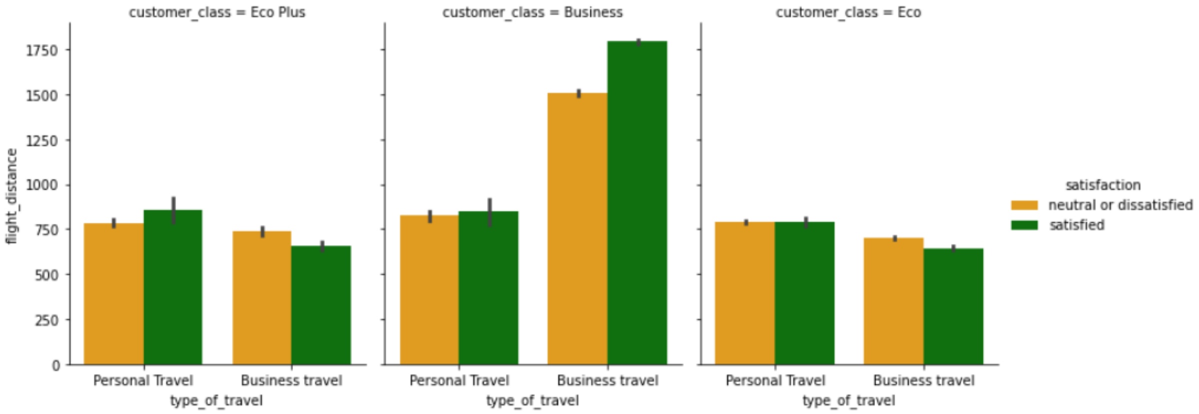


- Eco Plus passengers are mostly satisfied without inflight wi-fi service (rating 0) and with moderate level of in-flight entertainment (rating 2 - 4).
- For Business class passengers, only highest level of in-flight entertainment (rating 5) can bring satisfaction in them.
- For Eco passengers, high level of in-flight entertainment (rating 3 - 5) and very high wi-fi service availability (rating 5) can make them satisfied.



- For business class passengers, it is observed that all gate locations have higher number of neutral/dissatisfied passengers when baggage handling is not up-to the satisfactory level (rating <= 4).

- For Eco Plus class, when the gate location is 1 and for Eco class, when the gate location is 2, even when the baggages are handled in a mediocre way (rating 2 to 4), passengers remained neutral/dissatisfied.



- For business travel in business class category, the number of satisfied passengers are quite on the higher for longer flight distance.
- For other combinations, almost equal distribution of satisfied and neutral/dissatisfied passengers is present.

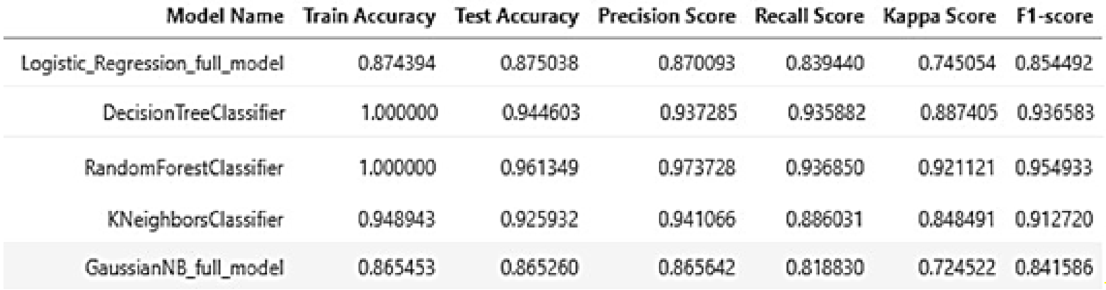
l. Data Transformation

- Applying Power Transformation which scales and transforms numerical columns.
- One-Hot Encoding of categorical columns.
- Combining one-hot encoded data and the transformed data.

MODEL BUILDING AND TRAINING

- Splitting the data to train and test dataset in the 80:20 ratio respectively.
- Fitting base model for the transformed data for various classifiers.

The values of evaluation parameters for the base models are as follows:

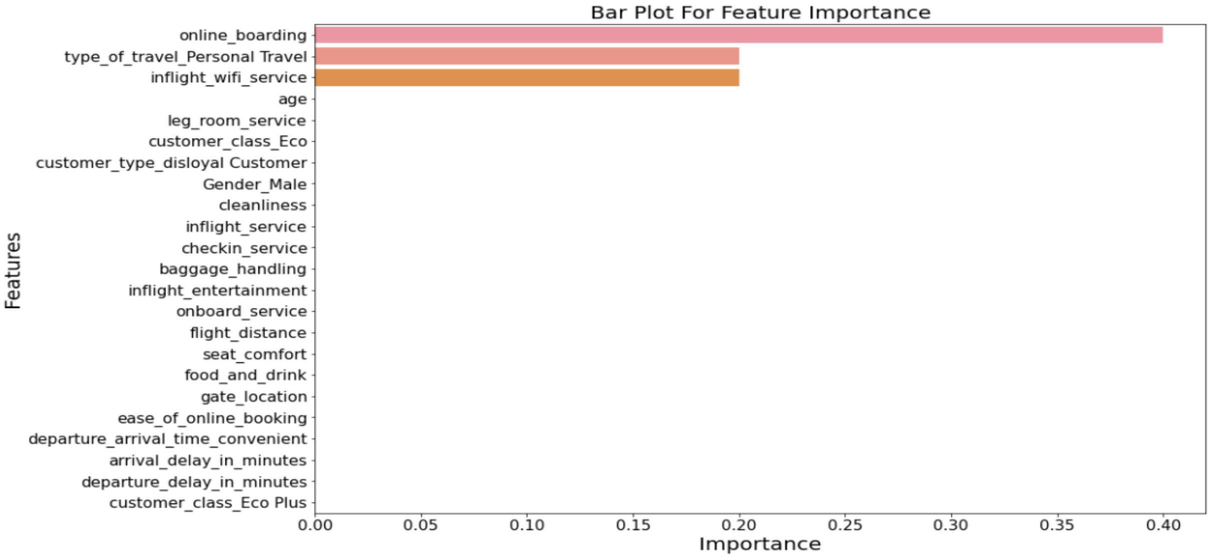


Model Name	Train Accuracy	Test Accuracy	Precision Score	Recall Score	Kappa Score	F1-score
Logistic_Regression_full_model	0.874394	0.875038	0.870093	0.839440	0.745054	0.854492
DecisionTreeClassifier	1.000000	0.944603	0.937285	0.935882	0.887405	0.936583
RandomForestClassifier	1.000000	0.961349	0.973728	0.936850	0.921121	0.954933
KNeighborsClassifier	0.948943	0.925932	0.941066	0.886031	0.848491	0.912720
GaussianNB_full_model	0.865453	0.865260	0.865642	0.818830	0.724522	0.841586

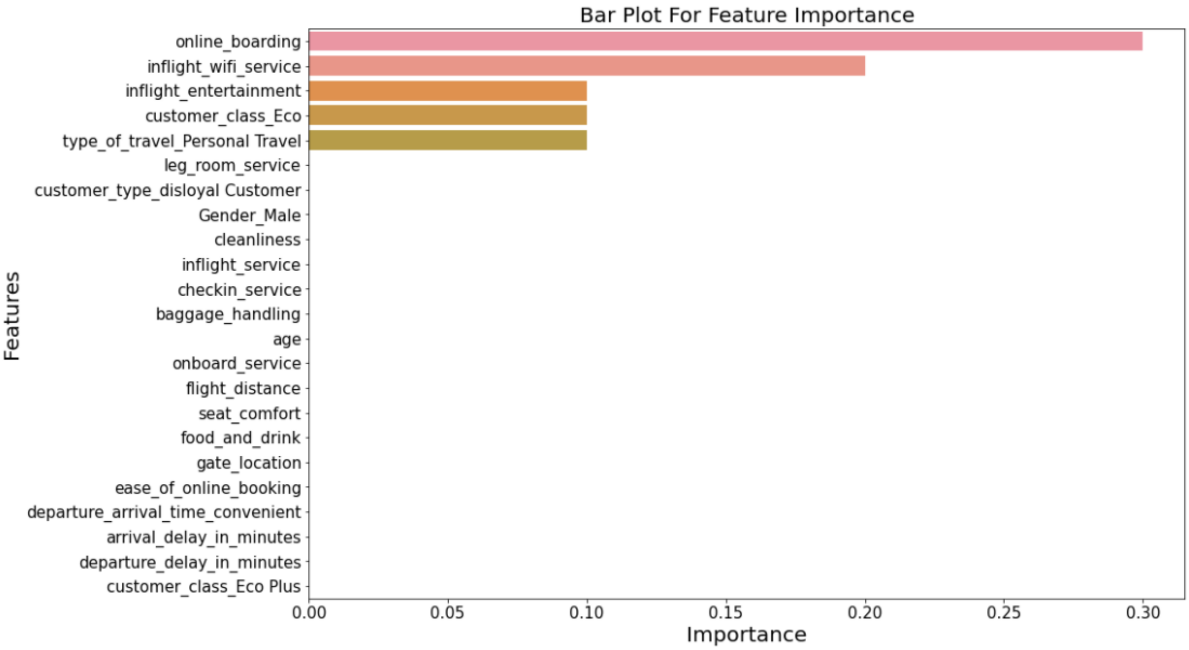
As per the values of the evaluation parameters, we can see there is presence of overfitting of the data, so we go ahead and carryout hyper parameter tuning for each classifier.

HYPER PARAMETERS TUNING AND FEATURE SELECTION

Importanat Features obtained after fitting the DecisionTree Model



Important Features obtained after fitting Tuned RandomForest Classifier



After carrying out hyper parameter tuning and feature selection as an when applicable,the values of various evaluation metrices are as follows:

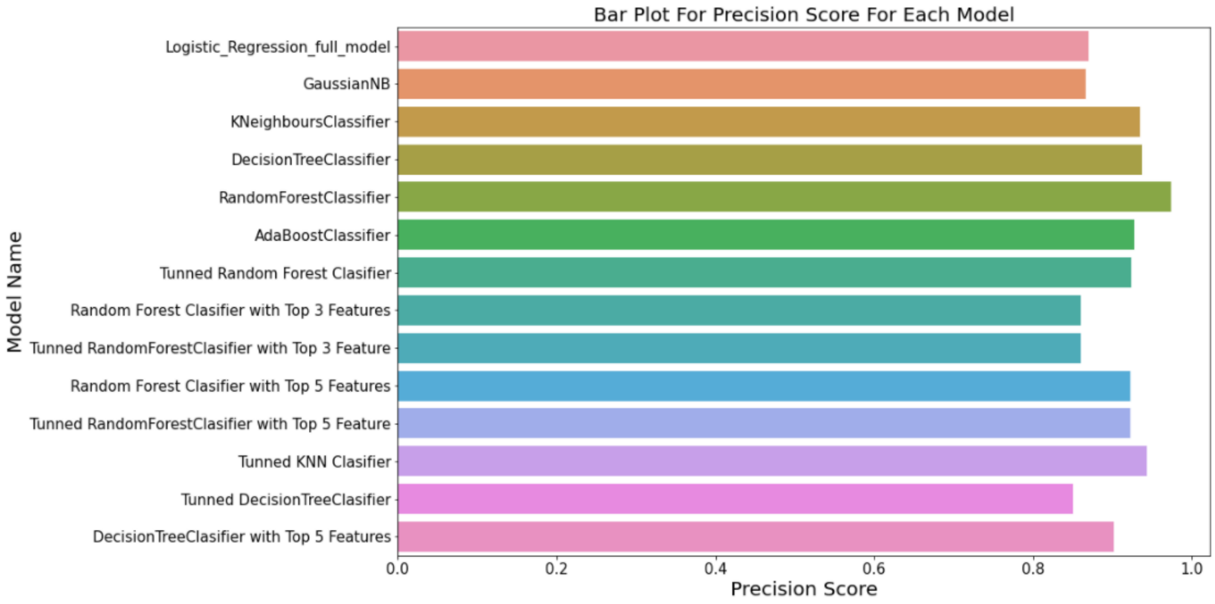
Model Name	Train Accuracy	Test Accuracy	Precision Score	Recall Score	Kappa Score	F1-score
Tunned Random Forest Clasifier	0.928857407	0.9263166	0.923546303	0.90646468	0.849953072	0.914925771
Random Forest Clasifier with Top 3 Features	0.891460249	0.890283338	0.859516009	0.89090801	0.777265503	0.874930517
Tunned RandomForestClasifier with Top 3 Feature	0.891460249	0.890283338	0.859516009	0.89090801	0.777265503	0.874930517
Random Forest Clasifier with Top 5 Features	0.925832637	0.923827122	0.922043011	0.8991897	0.844205626	0.910472973
Tunned RandomForestClasifier with Top 5 Feature	0.925832637	0.923827122	0.922043011	0.8991897	0.844205626	0.910472973
Tunned KNN Clasifier	0.943871266	0.927086541	0.942883895	0.8869121	0.850838354	0.914041935
Tunned DecisionTreeClasifier	0.875182861	0.869148445	0.850594976	0.84992073	0.734064542	0.850257721
DecisionTreeClasifier with Top 5 Features	0.918078226	0.917154296	0.902159725	0.90592231	0.831154387	0.904037101

SELECTING THE BEST MODEL

In order to select the best model for deployment, we have considered the following aspects :

- **Accuracy –** Overall performance of the model
- **Precision –** Since the problem statement is False positive sensitive
- **Bias and Variance Error –** A model is said to be the best if it has least bias and variance errors

Plot for precision scores of all the models is as follows:



The best model selected is : ***RandomForestClassifier*** ***(n_estimators=80, min_samples_split=2,min_samples_leaf=6,max_depth=6,criterion='entropy')***

BUSINESS INTERPRETATION

As shown earlier in exploratory data analysis, first-time customers have higher expectation and thus are less likely to be satisfied. However, capturing the satisfaction of first-time customer is important as this ensures a higher probability of their return to the airline for travel. Using the model, we can explore the important factors that leads to satisfaction in customers. As per the best model,RandomForestClassifier, the most important features are ***online_boarding, inflight_wifi_service, inflight_entertainment, customer_class and type_of_travel.***

LIMITATIONS

- There are many factors that affect airline customer satisfaction, but here only few of the major factors are considered.
- Efficient Hyper parameter tuning due to lack of hardware computation.
- Domain Expertise.

FUTURE SCOPE

- Applying NLP techniques on real time airline passengers' feedback data.

LEARNINGS

- Application EDA process on real time data/dataset
- Statistical Analysis
- Application of various Data Visualization tools and Techniques
- Feature Engineering
- Application of Machine Learning Algorithms
- Basics of Python Flask

FINAL NOTE

Thanks to Great Learning and team for giving us this opportunity to showcase our learnings in the form of this capstone project.

Thank You Mr. Naveen Konneti Sir for your continued guidance.

Thank You Mr. Ganesh Sir for continuous support and guidance.