Capstone Project-3

Airline passenger referral prediction

Ashfaque sayyed

Contents

Problem statement

Data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions. Data is scraped in Spring 2019. The main objective is to predict whether passengers will refer the airline to their friends.

Data Summary

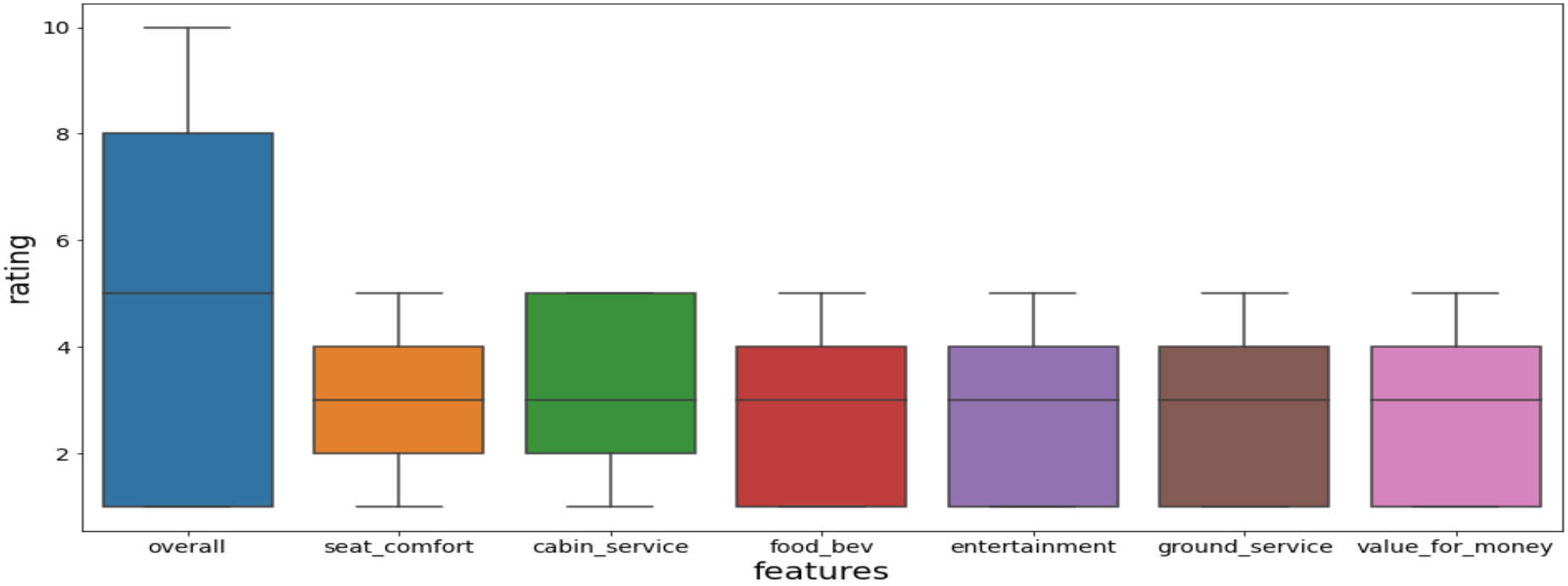
- airline: Name of the airline
- overall: Overall point is given to the trip between 1 to 10
- author: Author of the trip, reviewdate: Date of the Review customer.
- review: Review of the customers in free text format.
- aircraft: Type of the aircraft.
- Traveller type: Type of traveler (e.g. business, leisure)
- cabin: Cabin at the flight.
- date flown: Flight date
- seat comfort: Rated between 1-5.
- entertainment: Rated between 1-5
- cabin service: Rated between 1-5,foodbev: Rated between 1-5
- ground service: Rated between 1-5 value for money: Rated between 1-5.
- recommended: Binary, target variable.

Introduction

- Airline business as we know has been largely affected due to Covid-19 and most of airlines now is sitting on the verge of Bankruptcy because of this situation.
- Airline referral system generally works on customer reviews which are basically sentiment given by the customer depending upon various factors like seat comfort, their trip distance, the route they have travelled, entertainment, timing, airline frequency, ground service etc.
- These reviews are analysed and machine learning models on classification is prepared which helps airline industries to focus on factor resolution which it can actually help them in business growth better than the competitors.

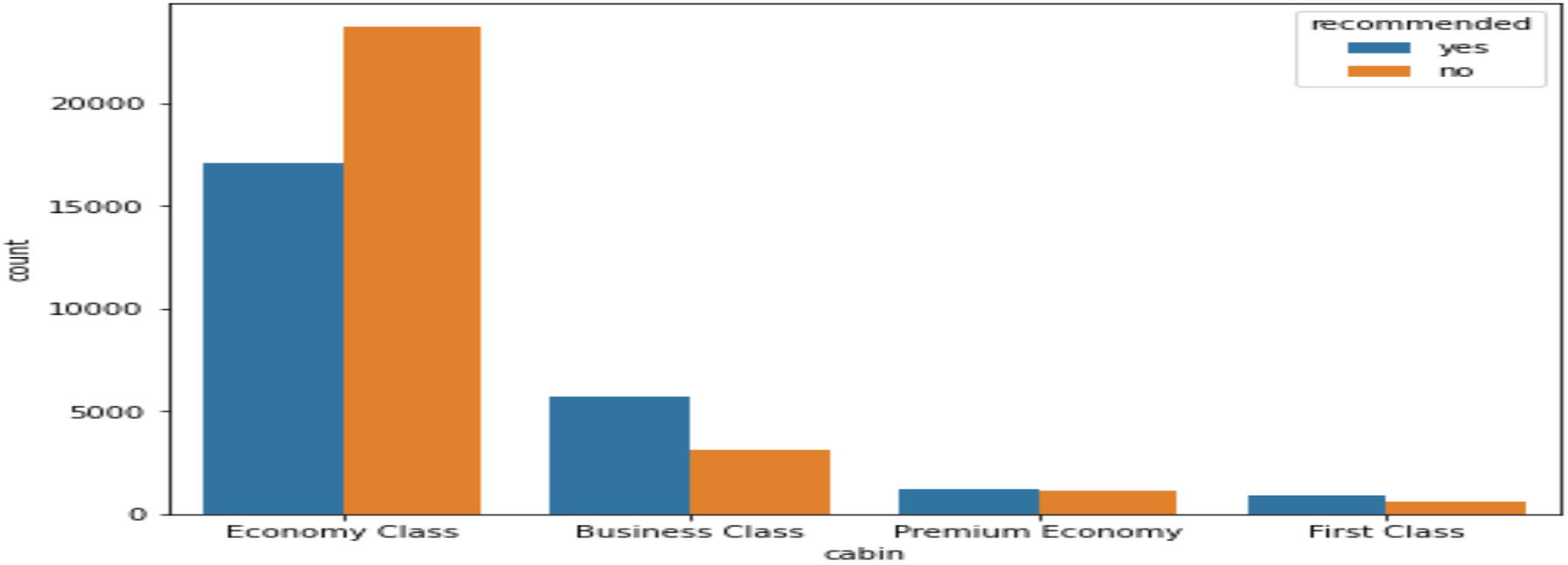
Exploratory Data Analysis

1. Outlier detection



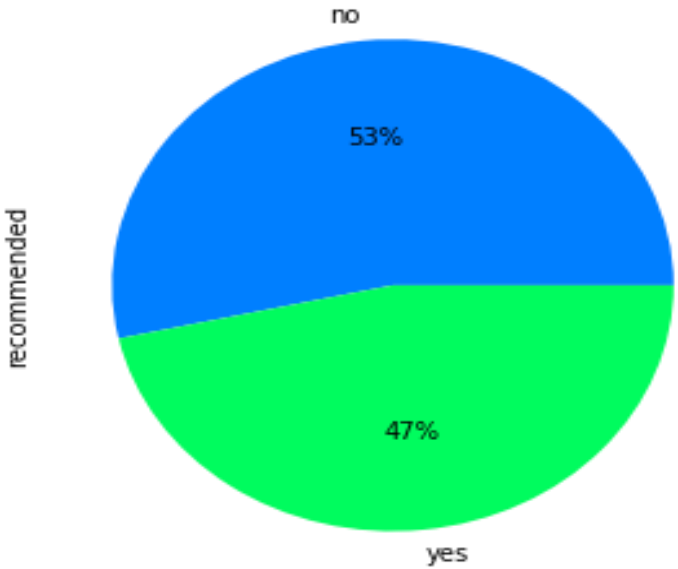
- Outliers are not present in data
- The median of 'Overall' is 5, The median of other features are approximately 3.

2. Which type of Cabin has more recommendation?



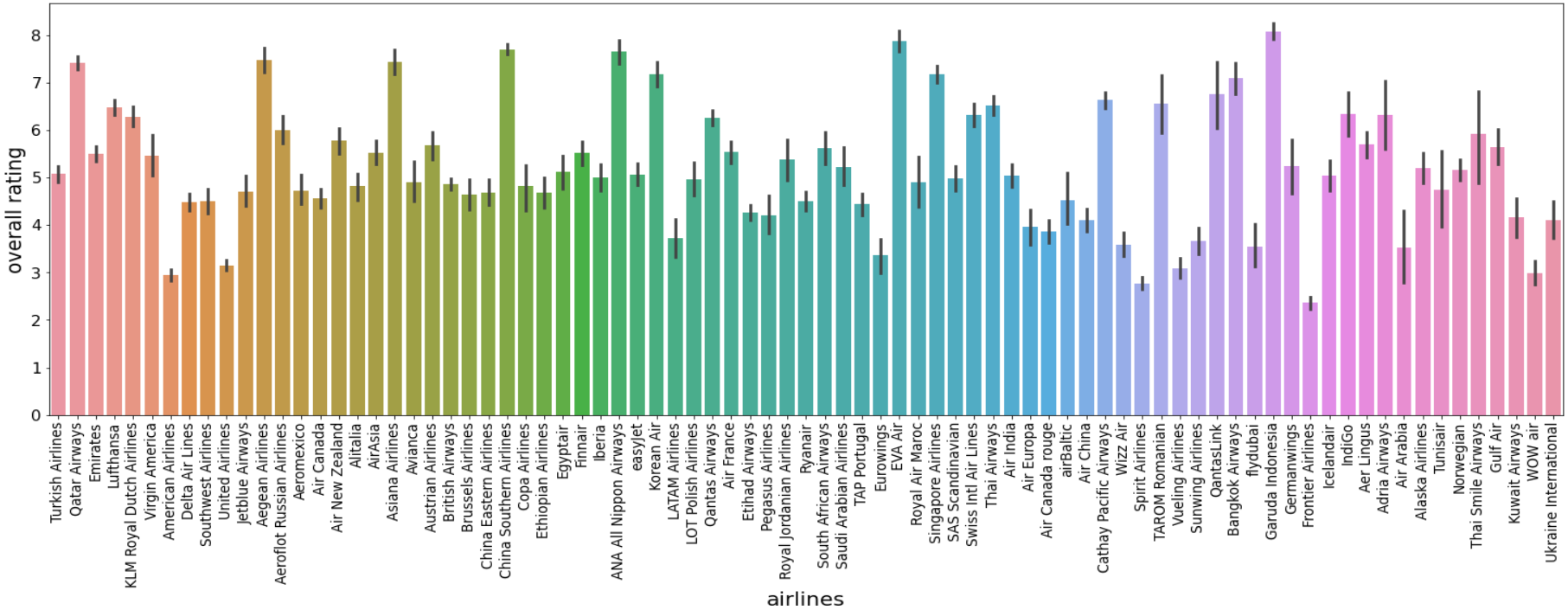
For the Economy class, the Number of 'NO' recommendations are more than 'YES' recommendations.
For business class and first class, the Number of ‘YES’ recommendations are more than ‘NO’ recommendations.
For the Premium economy of 'YES’ recommendations and 'NO' recommendations are approximately equal.

3. What is the total recommendation percentage for all airlines?



- The overall ‘’YES’ recommendation percentage for all airlines is 47% which is less than recommended 'NO' by 6%. Dependent feature have balanced data.

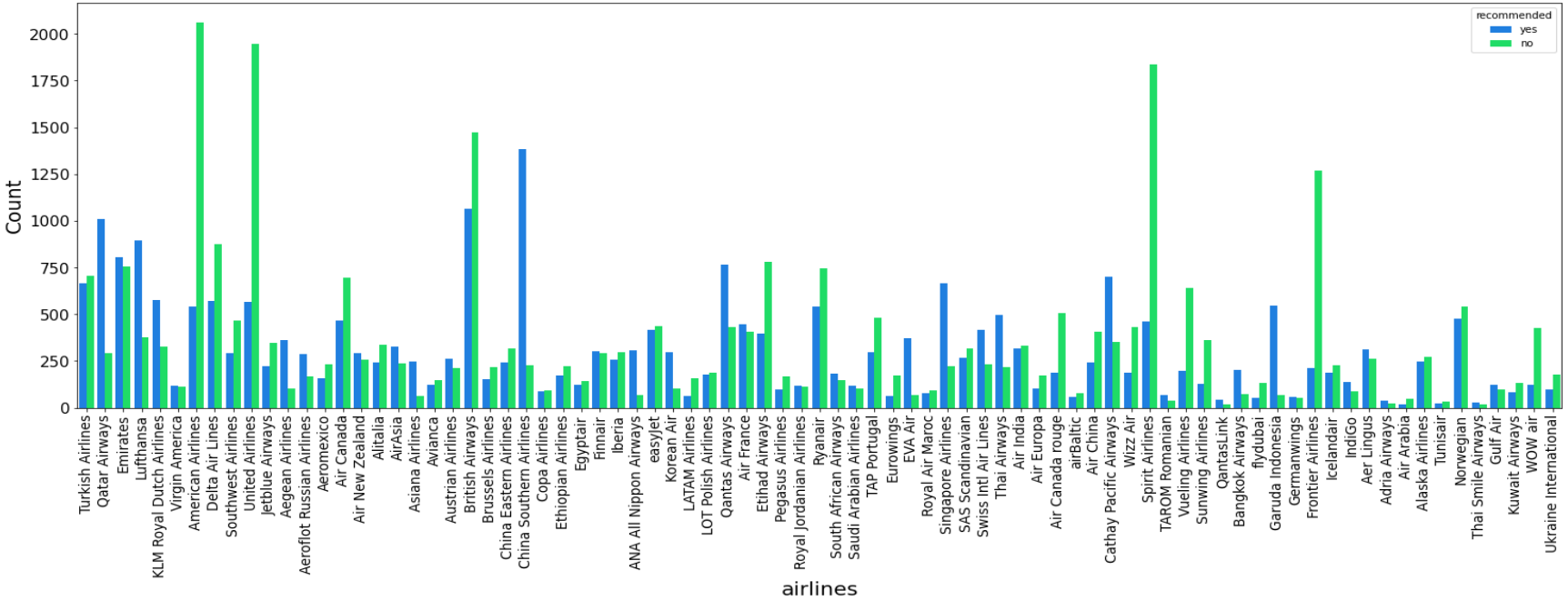
4.What is the maximum overall rating got by different airlines?



The maximum overall ratings are received by Qatar airlines, Aegean airlines,Asiana airlines, China southern Airlines etc (rating is around 7.5-8). airlines

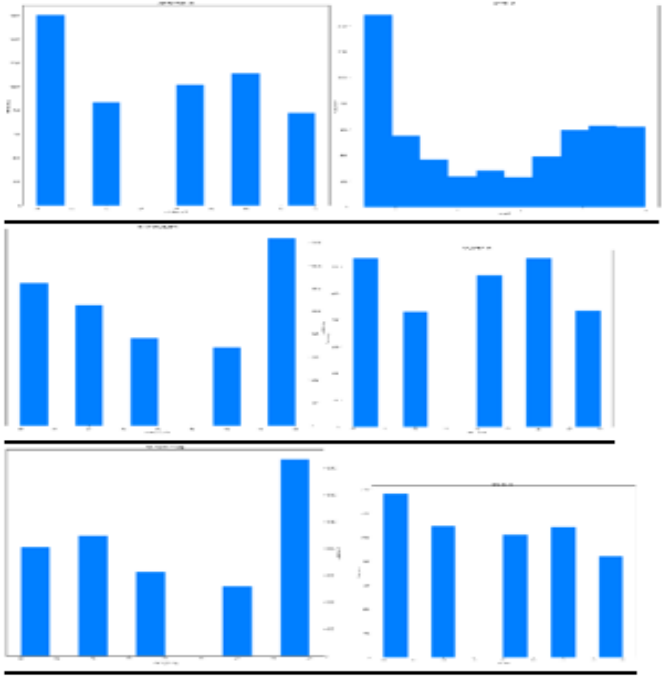
The minimum overall rating is around 2.5 received by frontier

5. Which airlines got the maximum and minmum recommendations?



- American airlines,united airlines,spirit and frontier airlines received maximum 'NO' recommendations.
- China southern airlines, Qatar and British airways received maximum 'YES' recommendations.
- Thai smile, Tunisair, Air Arabia, Adria airways received minimum 'Yes' recommendations.

Checking the frequency of values



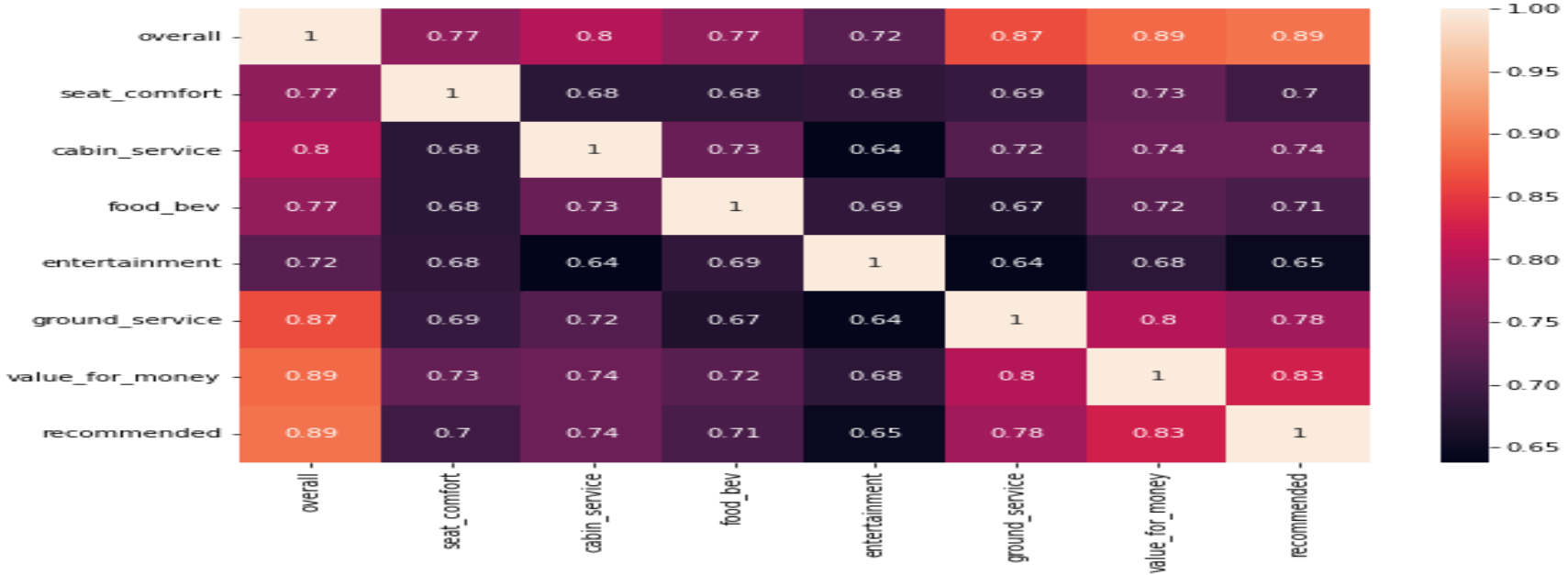
1.Cabin service got the maximum rating of 5.

2.Overall rating got by the most of the airlines are poor equal to 1.

3.Maximum customers rate food_bev as poor equal to 1.

4. Most of the customers have rated airlines as 1 indicating expensive (value for money).

6. Correlation between variables



- 'Overall','food bev','cabin_service', 'value_for_money' etc are positively correlated with recommendation
- 'Overall' is most correlatetd with recommendation.
- Entertainment has 0.45 of correlation which is less than others.
- Overall and value for money have multicollinearity.

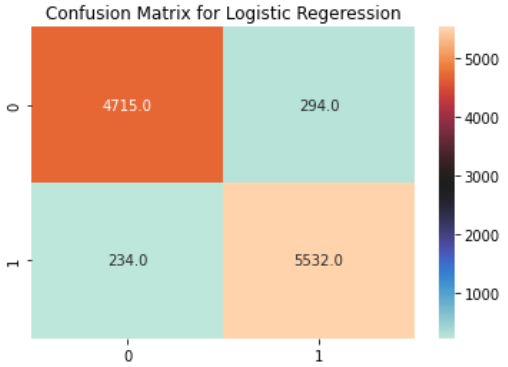
Train- test split

- The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
- In this project we have used 80% data for training purpose and 20% data for test set.
- The train-test procedure is appropriate when there is a sufficiently large dataset available.

Models

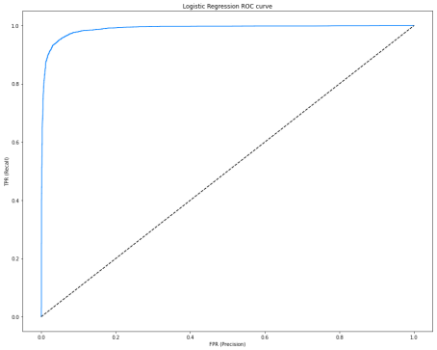
1) Logistic Regression:

Logistic regression is a classification technique that predicts the likelihood of a single-valued result (i.e. a dichotomy). A logistic regression yields a logistic curve with values only ranging from 0 to 1.



	precision	recall	f1-score	support
0	0.95	0.96	0.95	5766
1	0.95	0.94	0.95	5009
accuracy			0.95	10775
macro avg	0.95	0.95	0.95	10775
weighted avg	0.95	0.95	0.95	10775

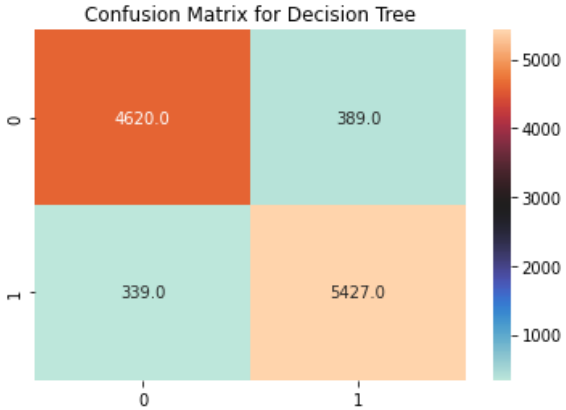
Accuracy score % of the model is 95.1%



- Accuracy score of the model is 95.1%.
- In Roc curve we can see that curve is closer to top left that means performance is better

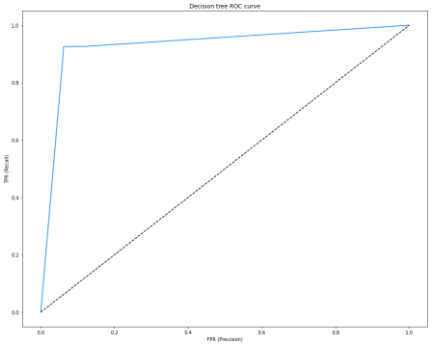
2. Decision tree

A decision tree is a supervised learning technique used to solve categorization problems.
Both categorical and continuous input and output variables are supported. The decision to make strategic splits has a significant impact on a tree's accuracy.
To decide whether to break a node into two or more sub-nodes, decision trees employ a variety of techniques.



	precision	recall	f1-score	support
0	0.93	0.94	0.94	5766
1	0.93	0.92	0.93	5009
accuracy			0.93	10775
macro avg	0.93	0.93	0.93	10775
weighted avg	0.93	0.93	0.93	10775

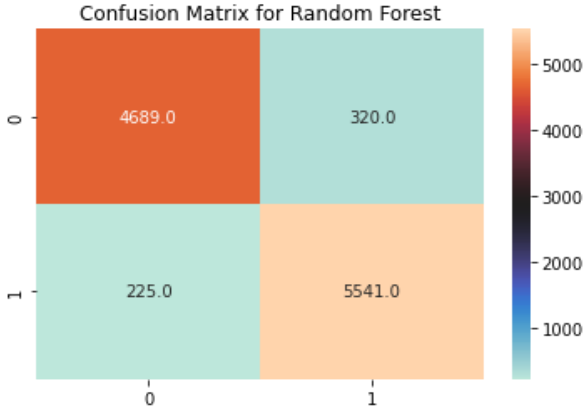
Accuracy score % of the model is 93.24%



In Roc curve we can see that curve is closer to top left that means performance is better
The accuracy score of the model is 93.24%.

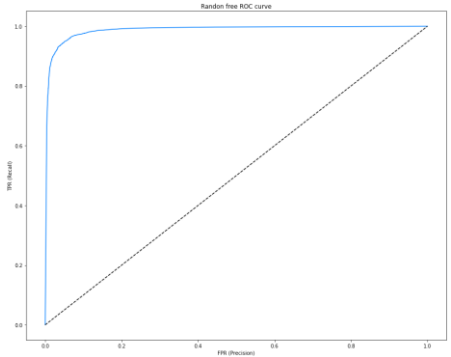
3. Random Forest

We create several trees in the Random Forest model rather than a single tree in the CART model.
From the subsets of the original dataset, we create trees. These subsets can contain a small number of columns and rows.
The classification with the highest votes is chosen by the forest.



	precision	recall	f1-score	support
0	0.94	0.96	0.95	5766
1	0.96	0.93	0.94	5009
accuracy			0.95	10775
macro avg	0.95	0.95	0.95	10775
weighted avg	0.95	0.95	0.95	10775

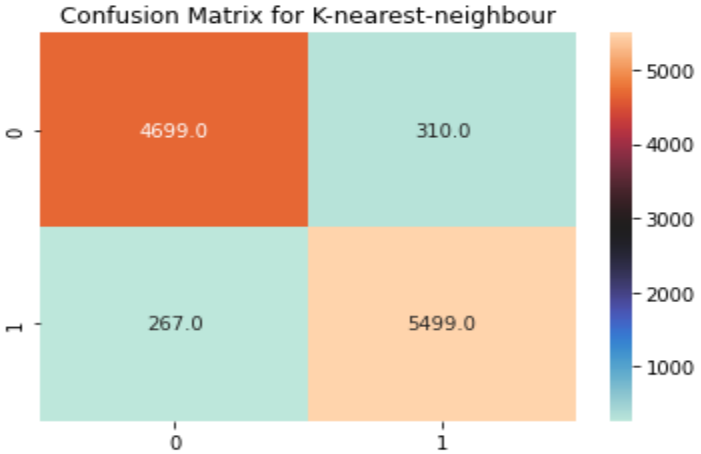
Accuracy score % of the model is 94.89%



Accuracy score for random forest model is 94.89%.
In Roc curve we can see that curve is closer to the top left that means performance is betterThe accuracy

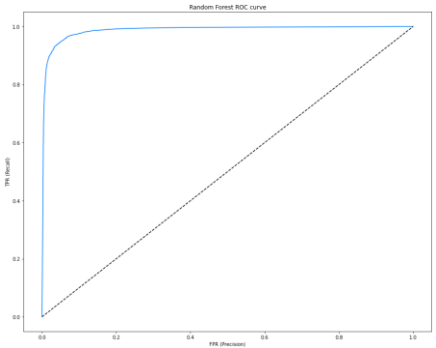
4. K-Nearest neighbour

K Nearest Neighbour is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure.
It is mostly used to classify a data point based on how its neighbours are classified.



	precision	recall	f1-score	support
0	0.95	0.95	0.95	5766
1	0.95	0.94	0.94	5009
accuracy			0.95	10775
macro avg	0.95	0.95	0.95	10775
weighted avg	0.95	0.95	0.95	10775

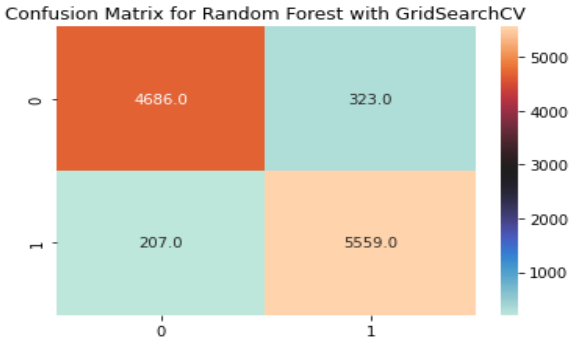
Accuracy score % of the model is 94.65%



In Roc curve we can see that curve is closer to top left that means performance is better
The accuracy score of K-Nearest model is 94.65%.

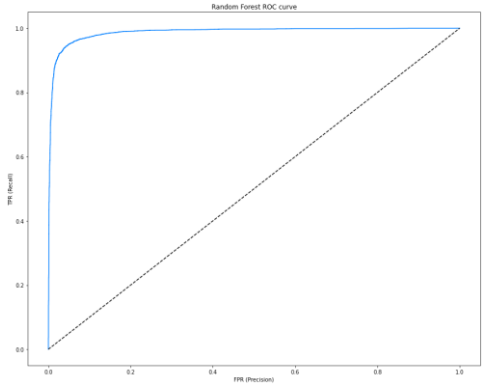
5. Random Forest with GridSearch CV

The accuracy for random forest is 94.94% and by applying GridSearchCV on we got an accuracy of 95.1%



	precision	recall	f1-score	support
0	0.95	0.96	0.95	5766
1	0.96	0.94	0.95	5009
accuracy			0.95	10775
macro avg	0.95	0.95	0.95	10775
weighted avg	0.95	0.95	0.95	10775

Accuracy score % of the model is 95.08%



In Roc curve we can see that curve is closer to top left that means performance is better
Accuracy score of the Random Forest with Gridsearch CV 95.1%.

Conclusion

1. 'Overall','food bev','cabin_service', 'value_for_money' etc are positively correlated with recommendation. these parameters should be improved to provide better service and hence it will improve recommendation chances for airlines.
2. entertainment has a 0.65 of correlation which is less than others.
3. ’Overall’ is most correlated with a recommendation.
4. Logistic regression, decision tree, random forest and Knearest neighbor gave good results in terms of accuracy. The highest accuracy obtained is 95.1% with logistic regression.
5.Random forest with Gridsearch CV also gave good accuracy approximately equal to logistic regression (95.08%).
6. American airlines, united airlines, spirit , and frontier airlines received maximum 'NO' recommendations.
7. China southern airlines, Qatar, and
British airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air arabia, Adria airways received minimum 'Yes' recommendations.
8.For the Economy class, the Number of 'NO' recommendations are more than 'YES' recommendations.
9.For business class and first class, the Number of 'YES' recommendations are more than 'NO' recommendations.
10.For Premium account number of 'YES’ recommendations and 'NO' recommendations are approximately equal.

THANK YOU