Capstone Project-3

Airline Passenger Referral Prediction

Zeeshan Ahmad

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
- travellertype: Type of traveler (e.g. business, leisure)
- cabin: Cabin at the flight
- date flown: Flight date
- seatcomfort: Rated between 1-5
- cabin service: Rated between 1-5
- foodbev: Rated between 1-5
- entertainment: Rated between 1-5
- groundservice: Rated between 1-5
- valueformoney: Rated between 1-5
- recommended: Binary, target variable.

Introduction

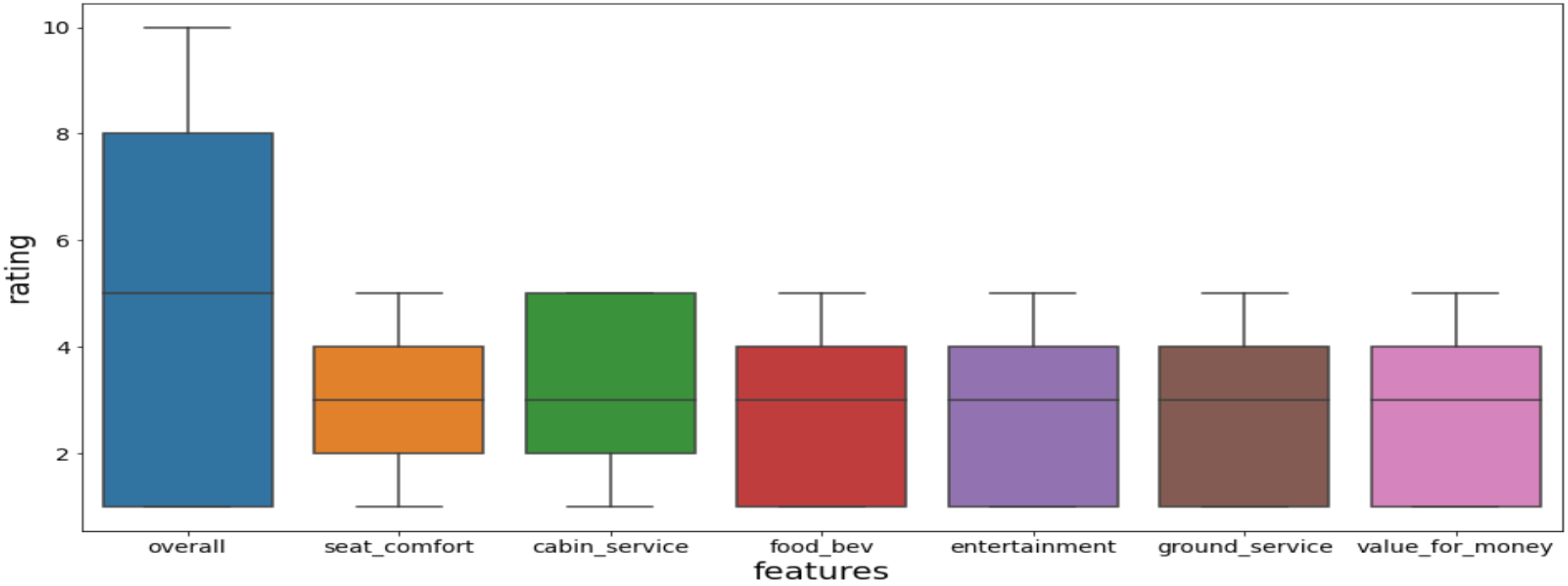
- Airline business as we know has been largely affected due to Covid-19 and most of airlines now is sitting on the verge of Bankruptcy because of this situation.

- 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.

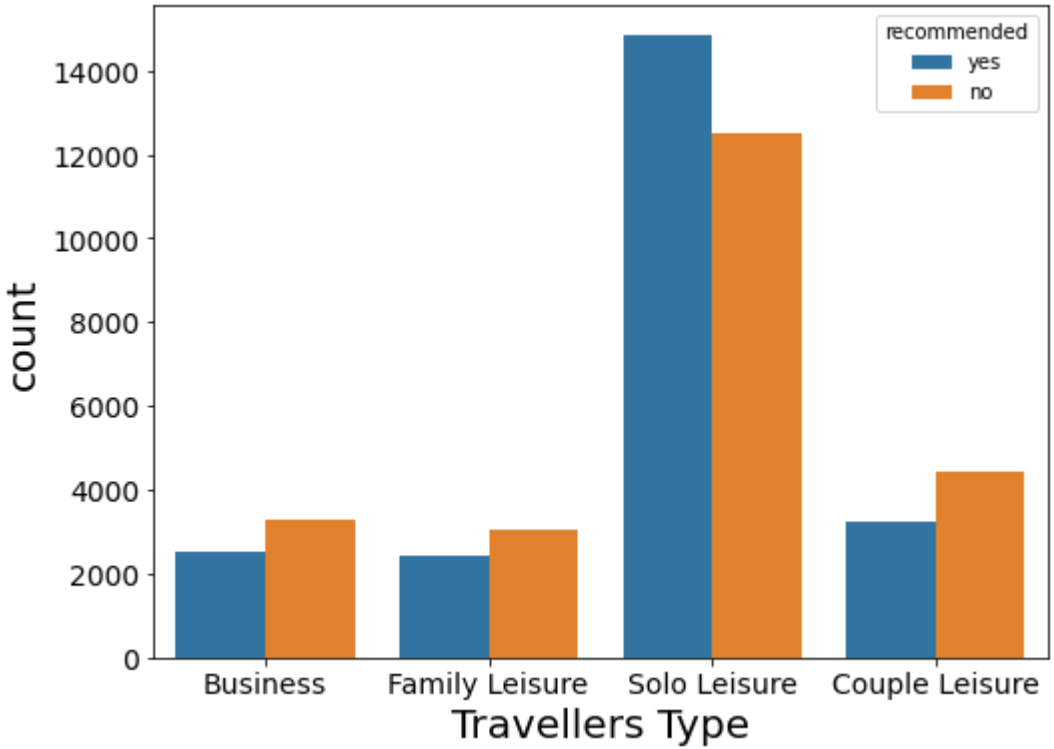
Exploratory Data Analysis

1. Outlier detection



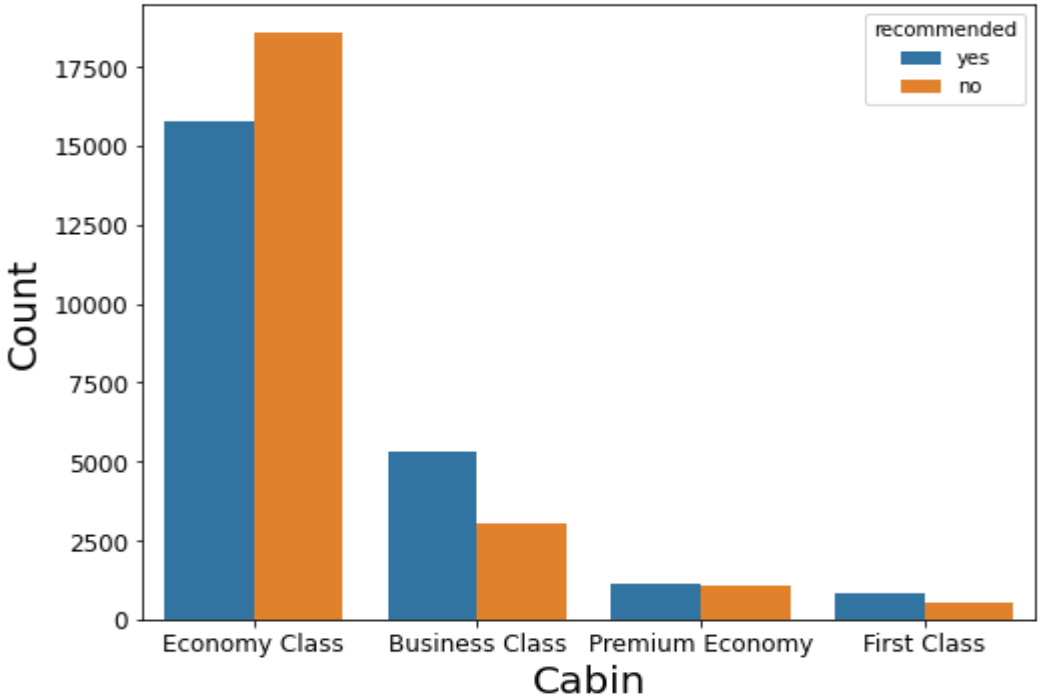
- Outliers are not present in data
- The median of 'Overall' is 5, The median of other features are approximately 3.

2. Travellers flight recommendation



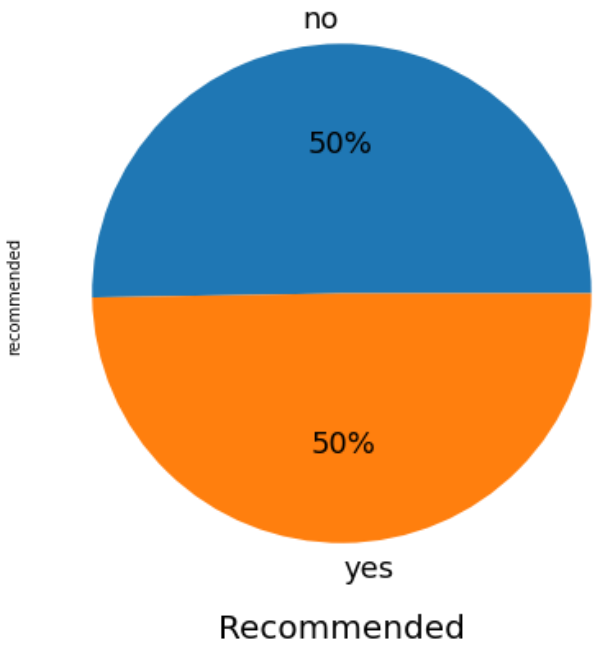
- Travellers from solo leisure are recommended yes more than others travelers.
- Travellers from business and family leisure have less recommendation
- Negative recommendation is also high from solo leisure

3. Which type of Cabin has more recommendation?



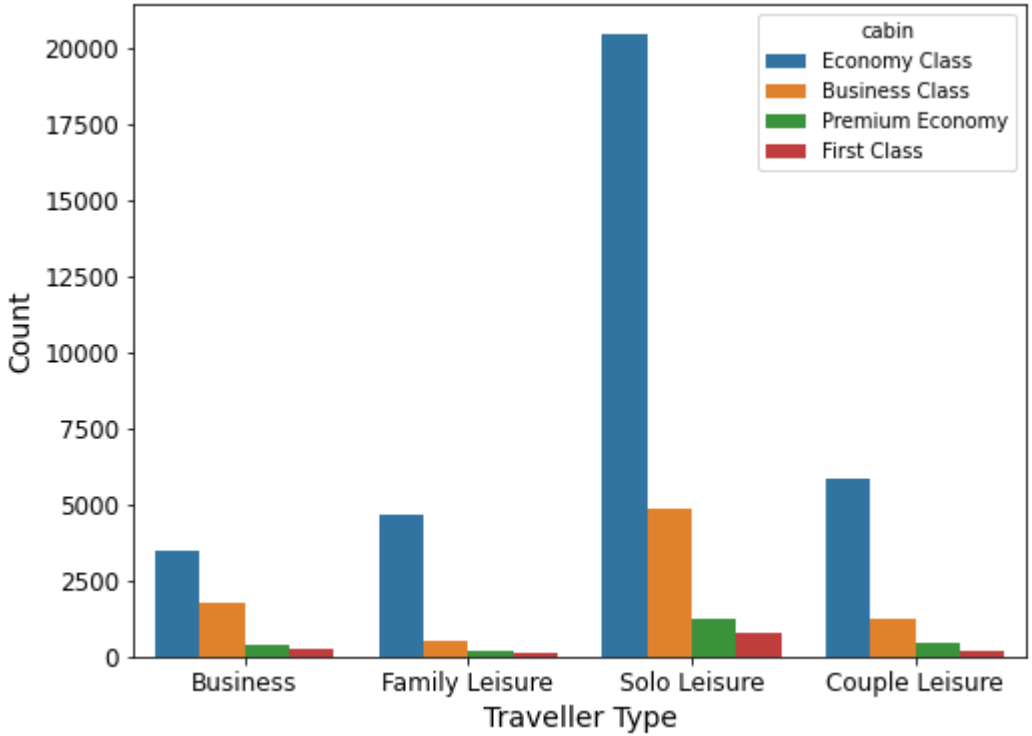
For the Economy class, the Number of 'NO' recommendations are more than 'YES' recommendations.
For business class and first class, the Number of ‘YES’ recommendations are more than ‘NO’ recommendations.
For the Premium economy of 'YES’ recommendations and 'NO' recommendations are approximately equal.

4. Total recommendation percentage for all airlines



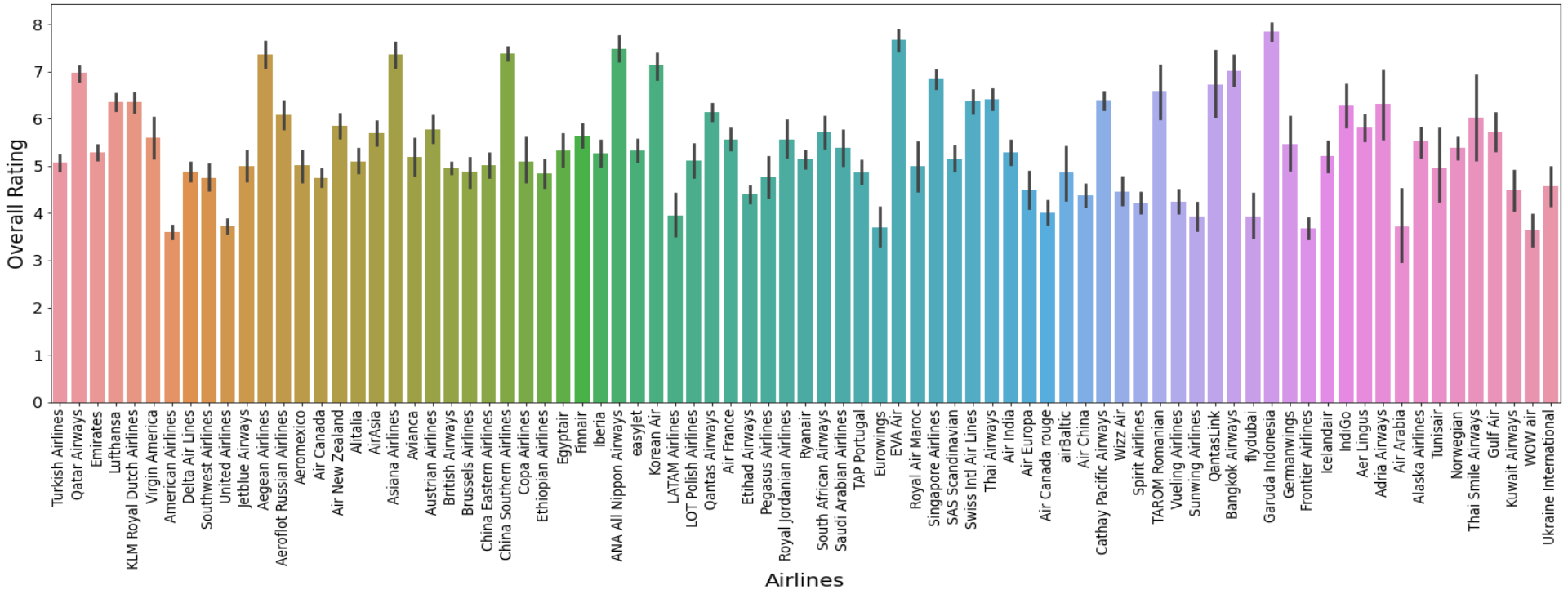
- Recommended values of yes or no are 50-50% which shows it has balanced data.

5. Travellers and their cabin



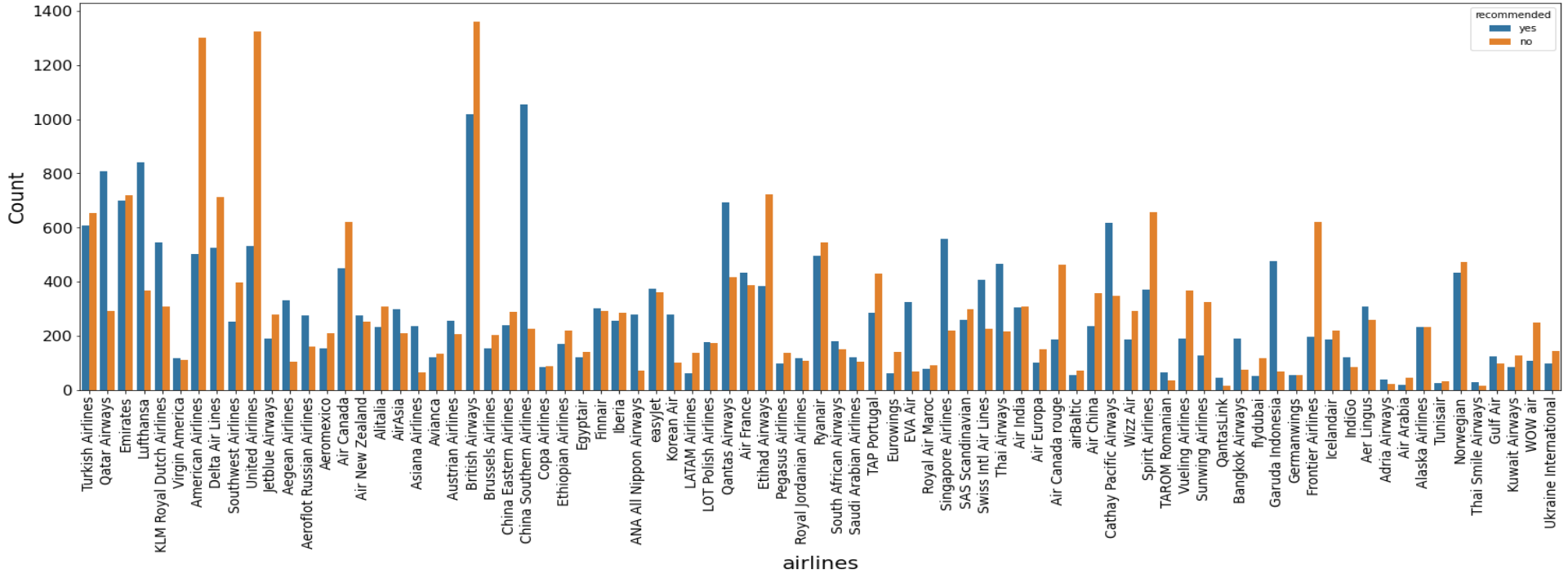
- All types of travellers mostly prefer economy class.
- Business class is less preferrable than economy class.
- First class is least preferable among all travellers.

6. Overall rating given by different customers



- The maximum overall ratings are received by Aegean airlines, Asiana airlines, China southern Airlines, ANA ALL Nippon Airways, EVA Air, Garuda Indonesia(rating is around 7.5-8).
- The minimum overall rating is received by American airlines, United airlines, Eurowings, Frontier airlines, Air Arabia and WOW air.

7.Airlines and their recommendations

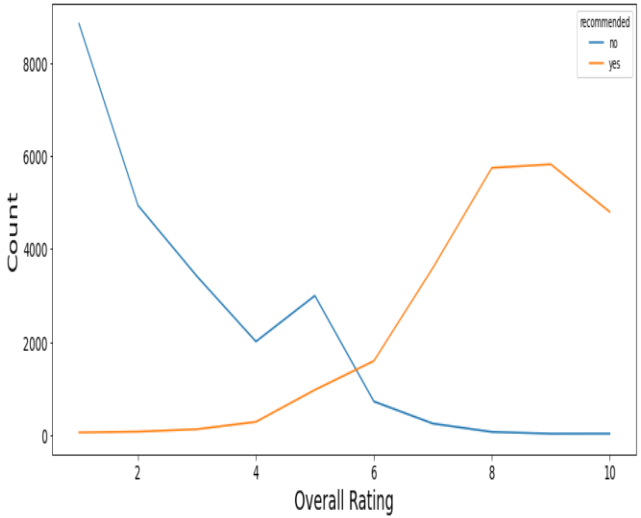


- American airlines, United airlines and British airlines received maximum 'NO' recommendations.
- China southern airlines, Lufthansa, Britsh airlines and Qatar airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air arabia, adria airways received minimum 'Yes' recommendations.

8. Variation of Recommendation with all features

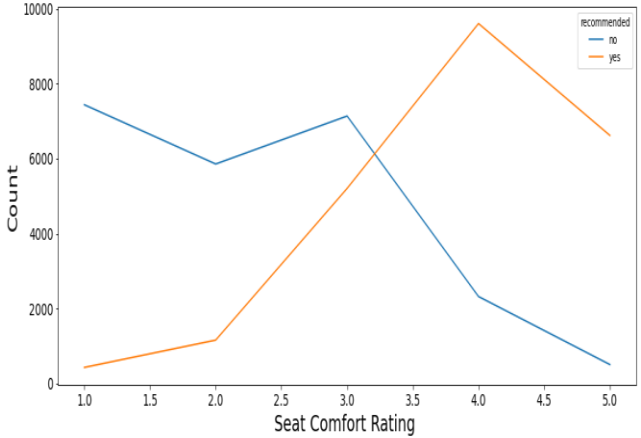
Overall Rating

- As we can see here rating less than 4 gave negative recommendation.
- Rating greater than 6 gave positive recommendation.
- We can see as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases.



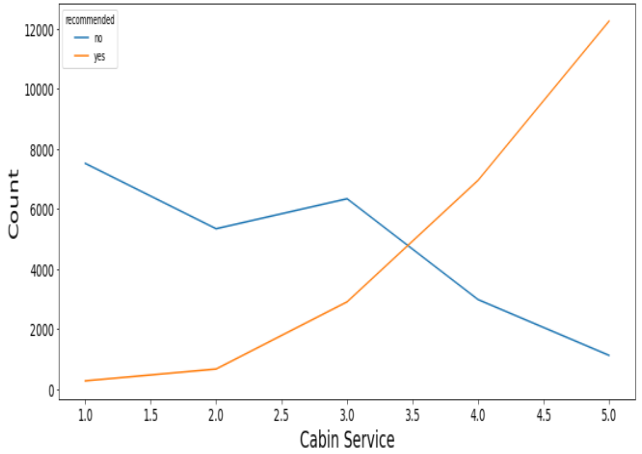
Seat Comfort

- In seat comfort we can see as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can an intersection in seat comfort rating 3.0 where we can see similar positive and negative recommendation.
- The Chances of getting positive recommendation is high when the seat comfort rating is more than 3.



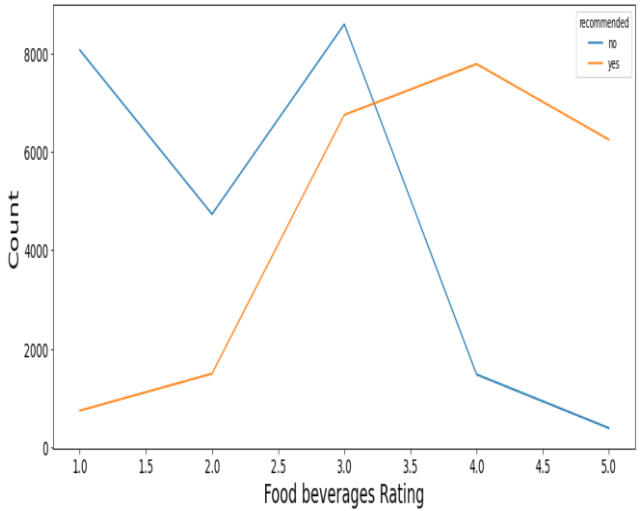
Cabin Service

- In cabin service we can see same as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can an intersection in cabin service rating 3.5 where we can see similar positive and negative recommendation.
- If the cabin service rating is greater than 3.5 then we can get 'yes' as recommendation by customer.



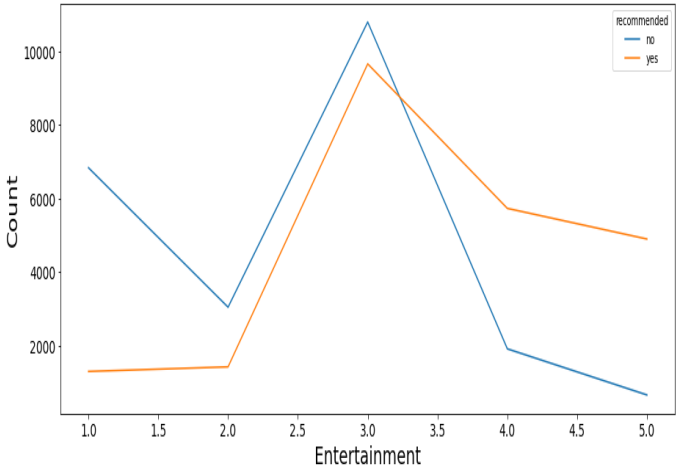
Food beverages

- If the rating is greater than 3.0 chances of getting 'yes' in recommendation is high.
- In food service we can see same as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can an intersection in food service rating close to 3.0 where we can see similar positive and negative recommendation.



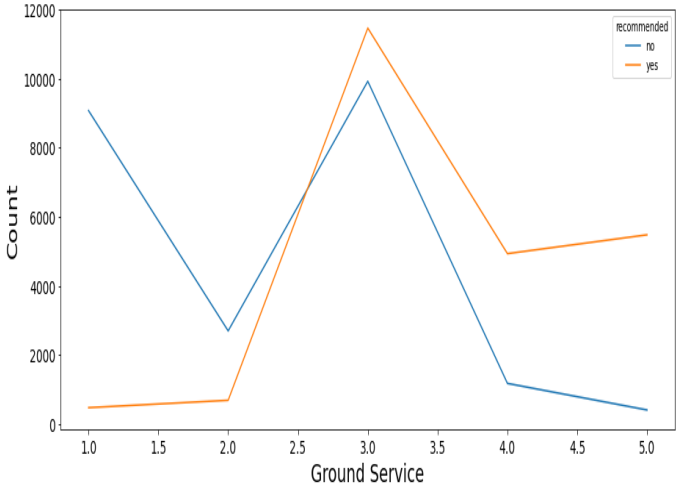
Entertainment Rating

- People giving rating 4 to 5 is less.
- Entertainment is not much affecting the recommendation
- We can see same as the positive recommendation increases with the rating and also negative recommendation on the same decreases also we can see intersection in Entertainment service rating between 3.0 and 3.5 where we can see similar positive and negative recommendation.

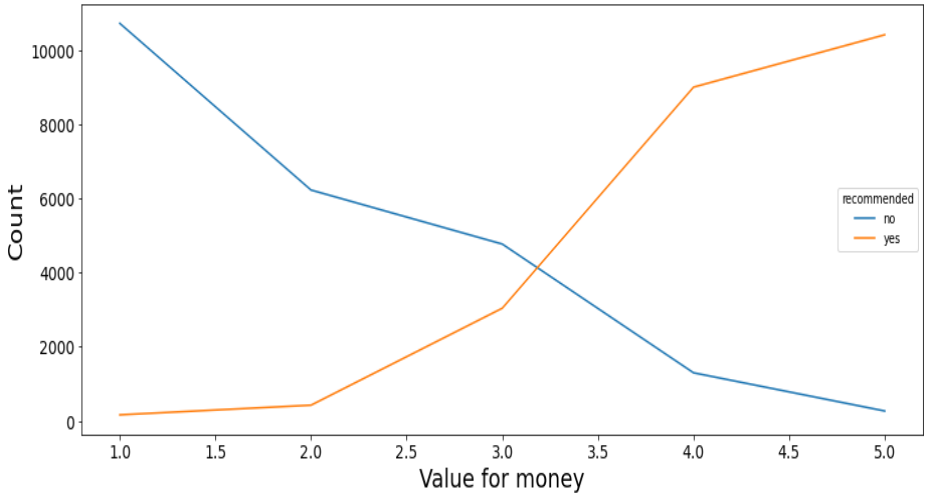


Ground Service

- 'Yes' and 'No' both recommendations are rising and dropping in same way.
- Ground Service not much affect the recommendations.

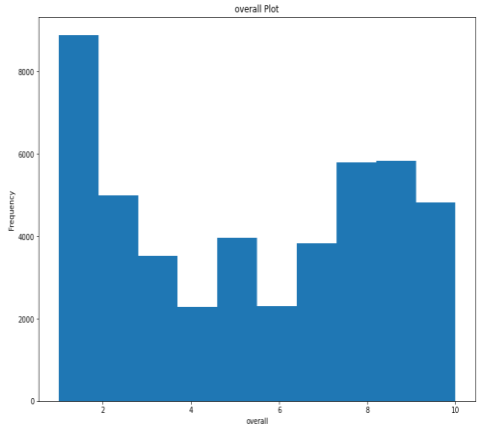


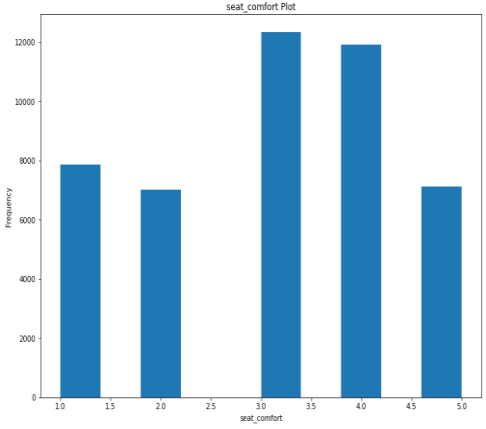
Value for money

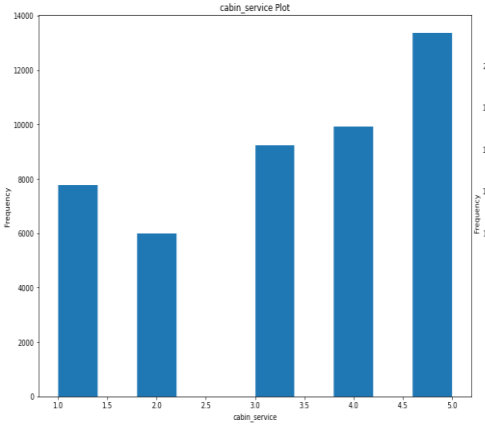


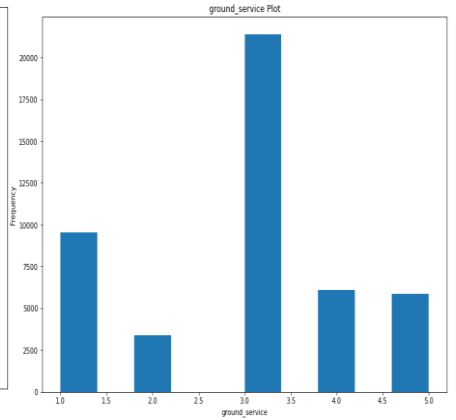
The chances of 'Yes' recommendation getting high when the value for money rating is more than 3.0.
If rating is less than 3.0 then we can get 'No' as a recommendation.

9. Frequency of feature values







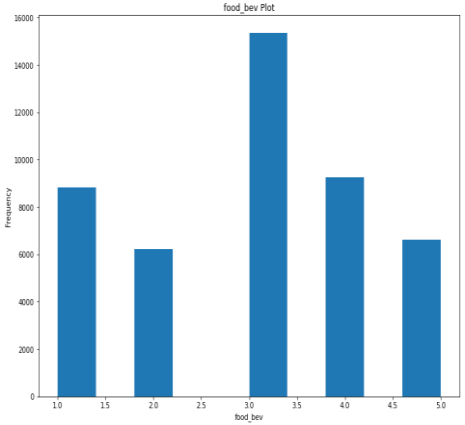


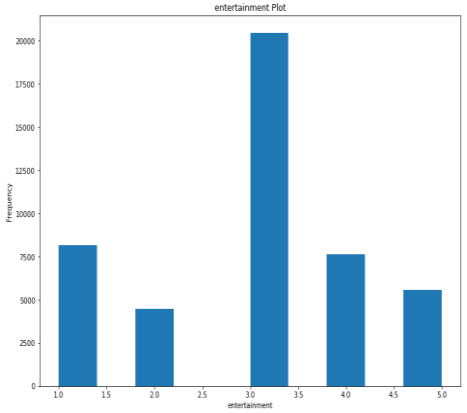
1.Cabin service got the maximum rating of 5.

2.Overall rating given to the airlines is poor equal to 1.

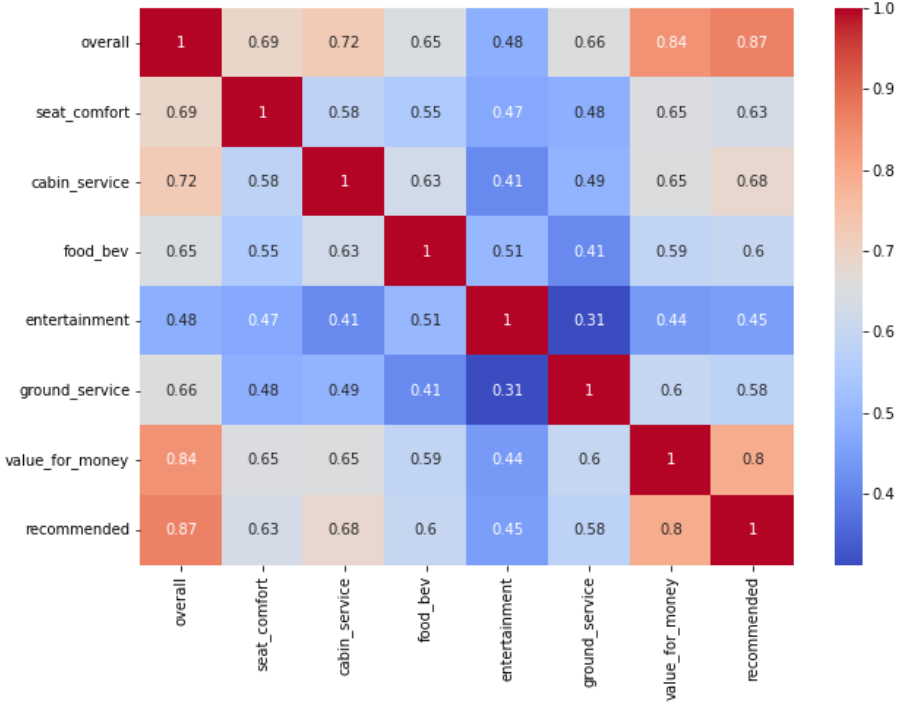
3.Maximum customers rate food_bev as poor equal to 1.

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





10. Correlation between variables

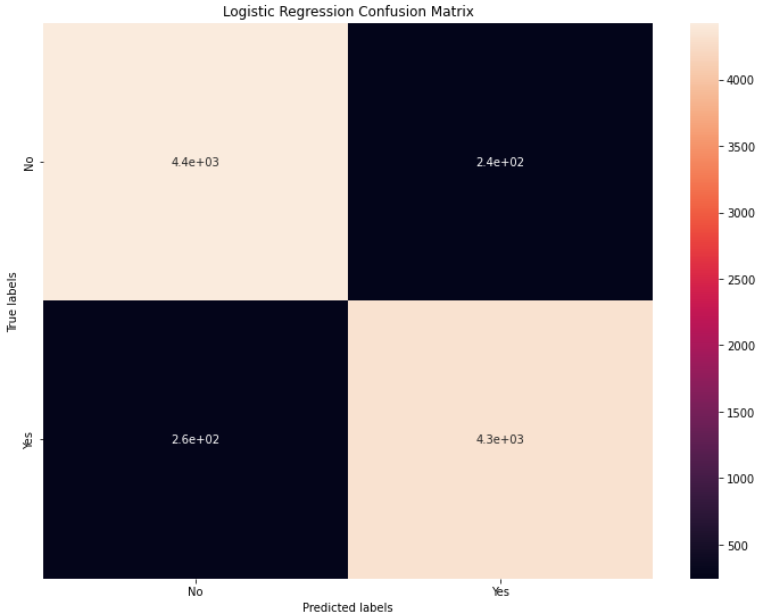


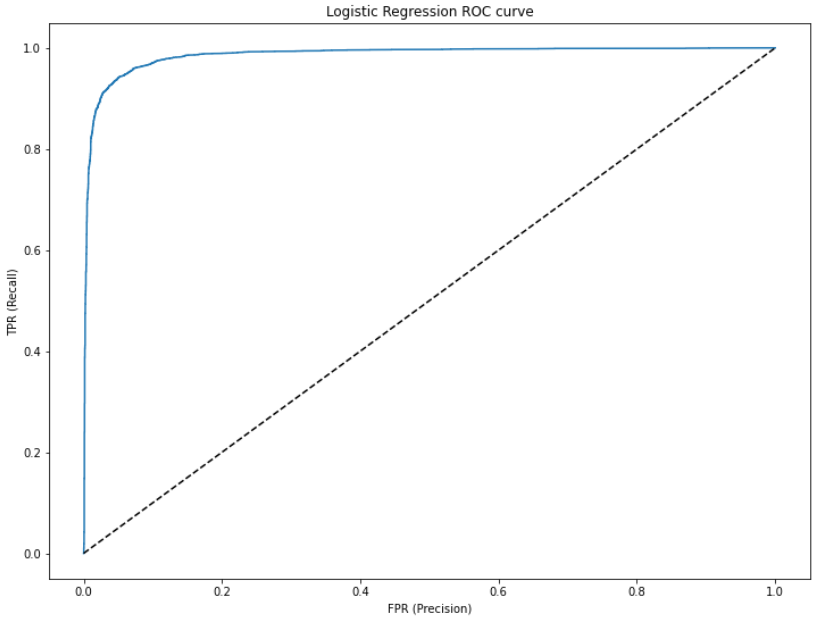
- 'Overall', 'food_bev', 'cabin_service', 'value_for_money', etc all are positively correlated with recommendation.
- 'Overall' is most correlated with recommendation.
- entertainment has 0.45 of correlation which is minimum.
- overall and value for money are multicollinear.

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.

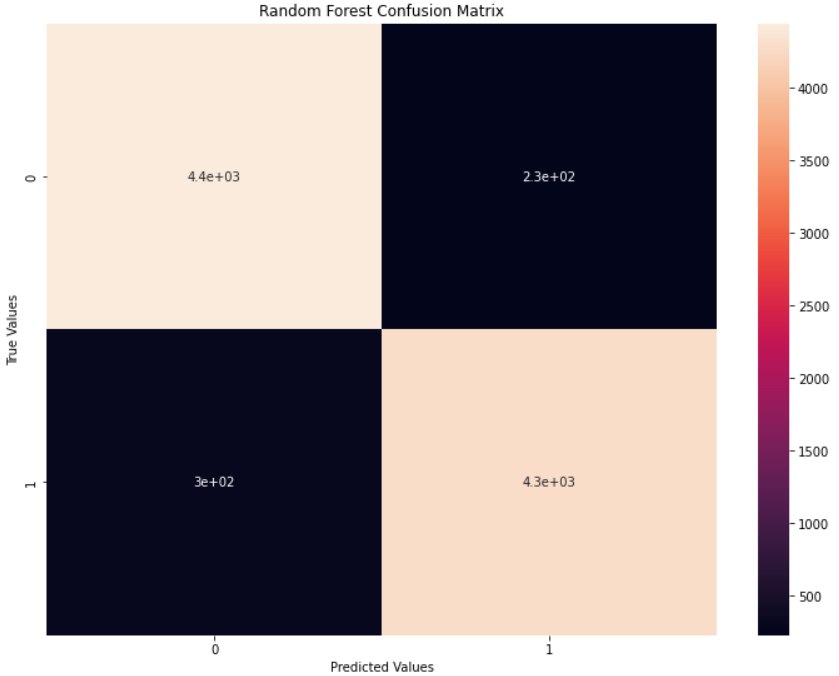


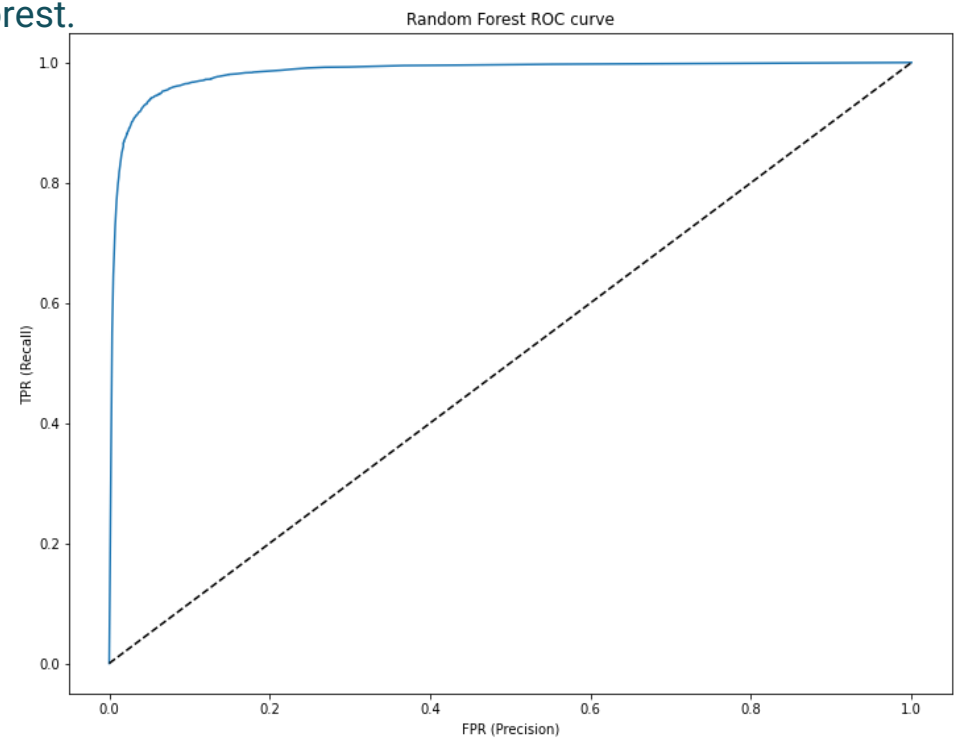


- In confusion matrix false positive and false negative is higher
- In Roc curve we can see that curve is closer to top left that means performance of the model is good.
- Accuracy score for Logistic Regression model is 94.52%.

2) 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.

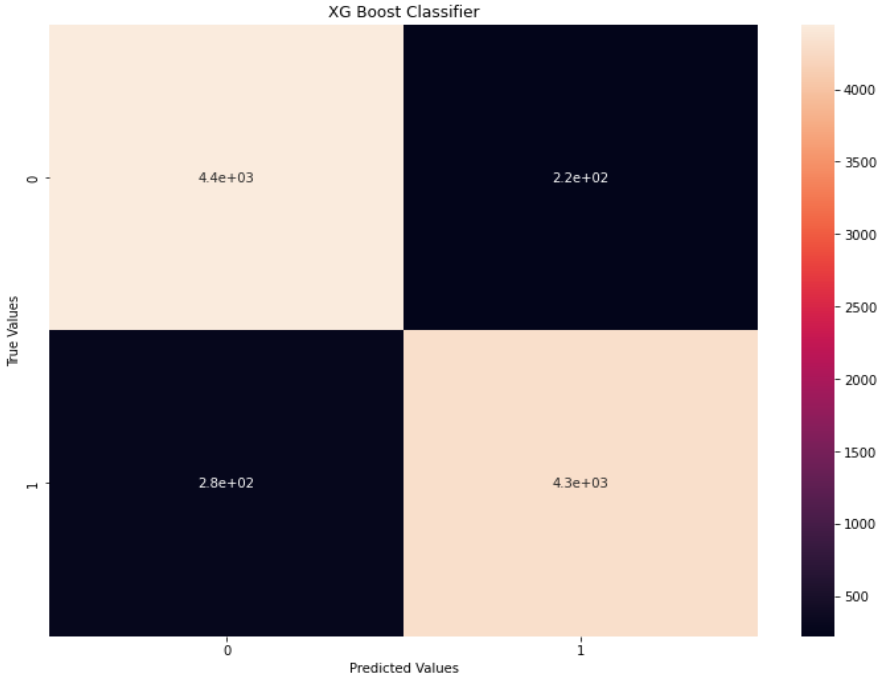


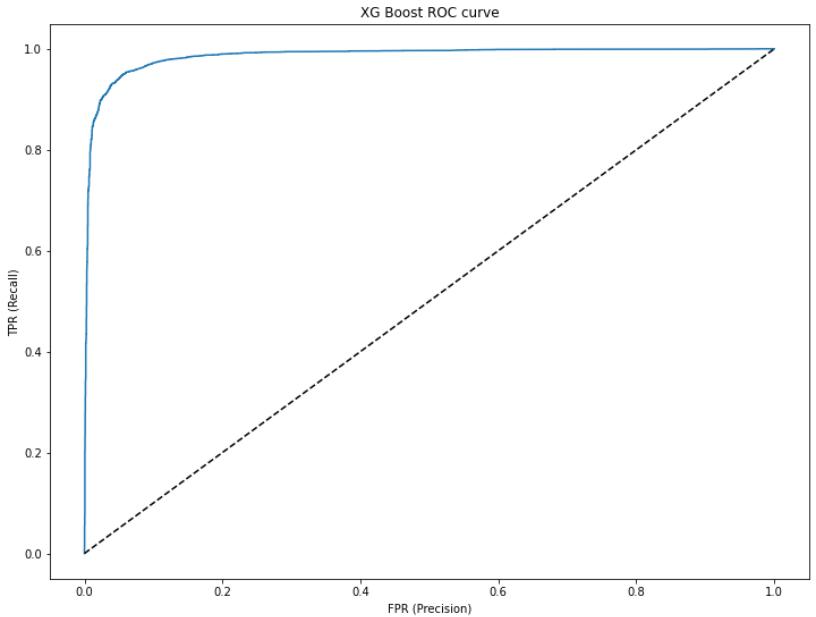


- prediction has less accuracy than logistic regression.
- Roc curve same as logistic regression
- Accuracy score for random forest model is 94.35%

3) XG Boost Classifier

It is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. XG Boost is basically designed to enhance the performance and speed of a Machine Learning model.

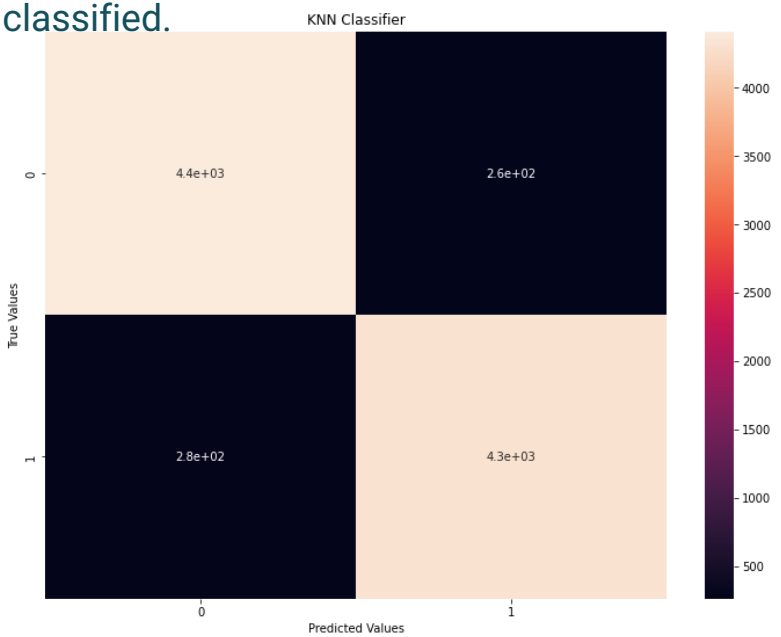


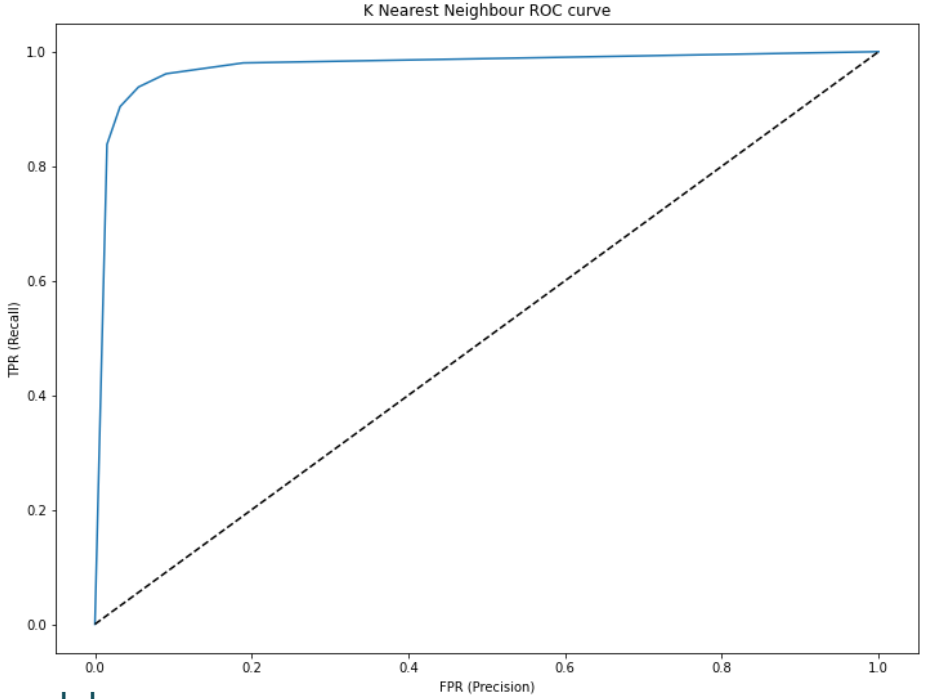


- Confusion matrix is similar to other model
- In Roc curve we can see that curve is closer to top left that means performance of model is good.
- Accuracy score for XG Boost Classifier model is 94.55%.
- We can see the small increase in accuracy.

4) K-Nearest neighbor Classifier

K Nearest Neighbor 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 neighbors are classified.

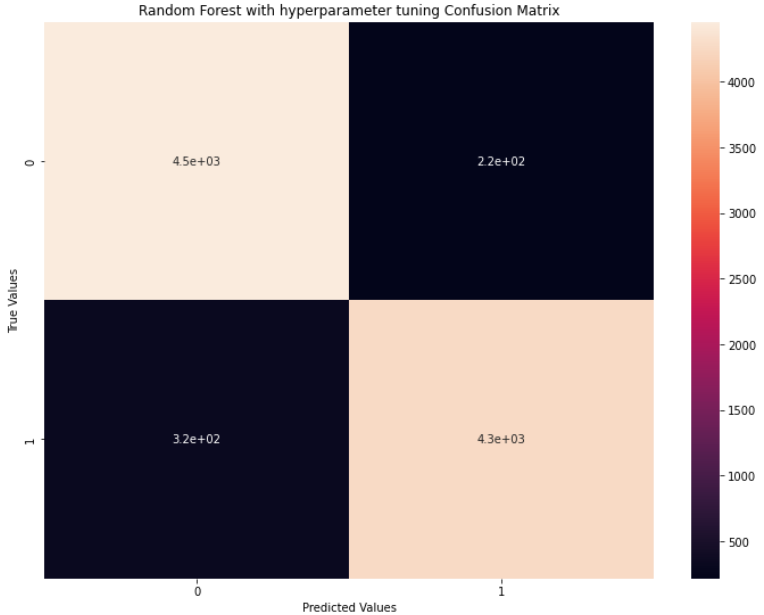


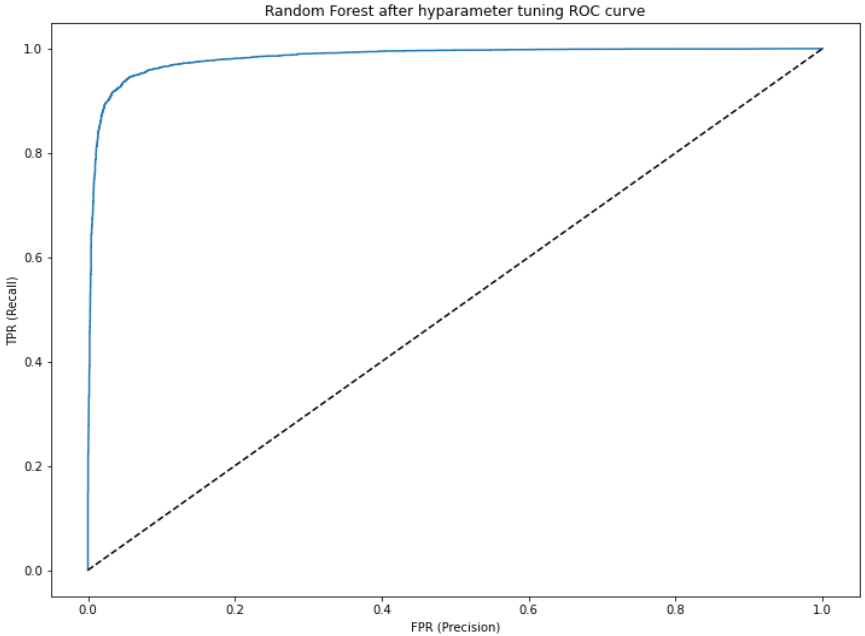


- This confusion matrix also similar to other models.
- In Roc curve we can see that small changes in the curve.
- Accuracy score for K-NN model is 94.13% which is less than other models.

Hyperparameter Tuning of Random Forest Classifier

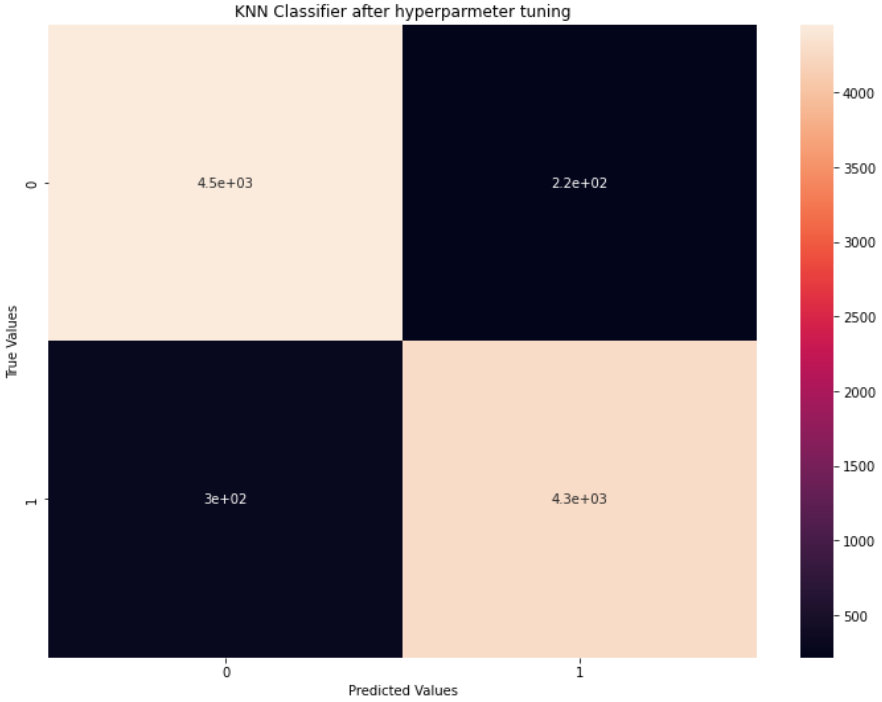
- Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.
- We used Grid Search CV for hyperparameter tuning.
- Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model's performance.

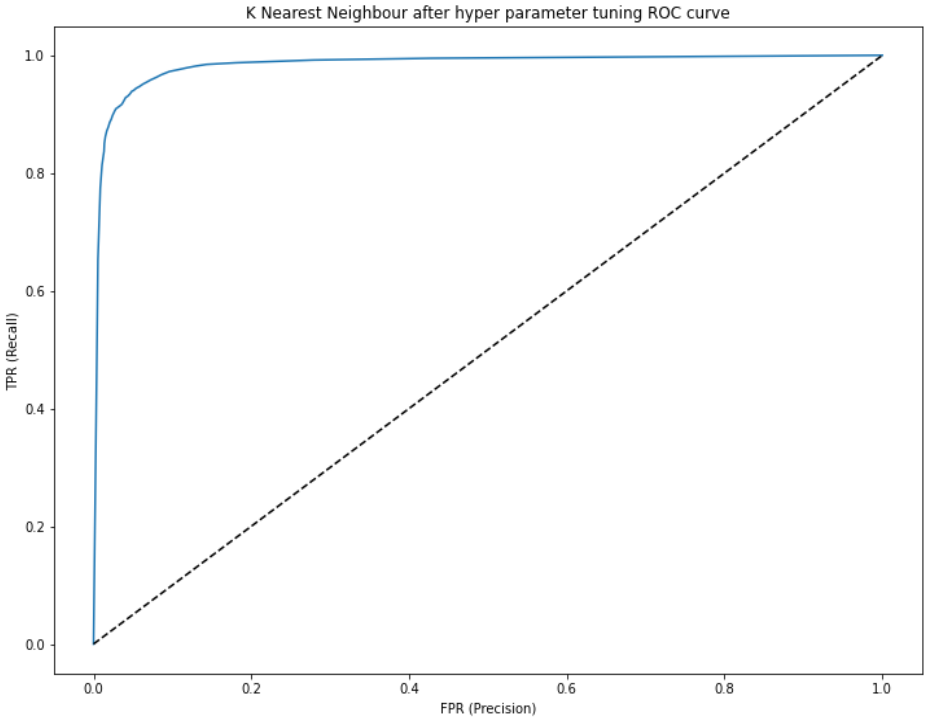




- After hyperparameter tuning values are not predicted accurately as we can see in confusion matrix.
- Accuracy is 94.25% after hyperparameter tuning of random forest

Hyperparameter Tuning of KNN Classifier





- After hyperparameter tuning of KNN accuracy not much change i.e. 94.42%

Evaluation Metrics

- Below table shows that all the model have similar accuracy but precision, f1 score and roc score high in XG boost classifier.
- Recall is high in Logistic regression.

Model	Accuracy	Recall	Precision	f1-score	roc_auc_score	Test score	Train score
Logistic Regression	0.945231	0.942639	0.946562	0.944596	0.985833	0.945231	0.947439
Random Forest Classifier	0.943502	0.935660	0.949535	0.942546	0.983916	0.945231	0.947439
Random Forest with GridSearchCV	0.942530	0.931080	0.951839	0.941345	0.983391	0.945231	0.947439
XG Boost Classifier	0.945555	0.938713	0.950740	0.944688	0.985964	0.945231	0.947439
K Nearest Neighbour Classifier	0.941342	0.938277	0.943007	0.940636	0.974454	0.945231	0.947439
K Nearest Neighbour with GridSearchCV	0.944258	0.934569	0.952011	0.943209	0.984211	0.945231	0.947439

Conclusion

- '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.
- Entertainment has 0.45 of correlation which is less than others.
- 'Overall' is most correlated with recommendation.
- Multicollinearity is present in between overall and value_for_money.
- XG boost classifier gives better accuracy.
- XG Boost, Logistic Regression gives good results in terms of accuracy. The highest accuracy obtained is 0.9455 with XG Boost Classifier.
- KNN after applying hyperparameter tuning also gave good accuracy.
- American airlines, united airlines, spirit and British airlines received maximum 'NO' recommendations.
- China southern airlines, Lufthansa and Qatar airways received maximum 'YES' recommendations. Thai smile, Tunisair, Air Arabia, Adria airways received minimum 'Yes' recommendations.
- For Economy class, Number of 'NO' recommendations are more than 'YES' recommendations.
- For business class and first class, Number of 'YES' recommendations are more than 'NO' recommendations.
- For Premium economy class number of 'YES' recommendation and 'NO' recommendations are approximately equal.

THANK YOU