Capstone Project

Airline Passenger Referral Prediction

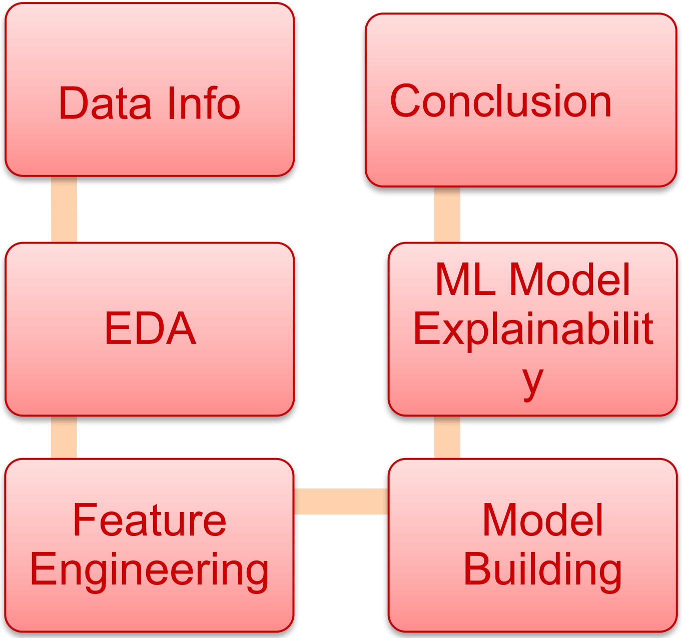
By
Saugata Deb
Shreyash Movale
Ankit Patil
Naga Sai Kiran

Objective

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

Methodology

The process from getting the data to drawing the conclusion is as follows:



Data Insights…

- The data set has 17 variables, in which 'recommended' is a Dependent variable and the rest are independent variables.
- The size of the data is (131895,17) i.e., we have 131895 rows with 17 columns
- There are lots of null values and duplicates in the data set so we must have to clean the data first.
- Data Set is a mixture of categorical and numerical data so we have to arrange and encode the data before feeding it to the ML model.

```
df.info()
```

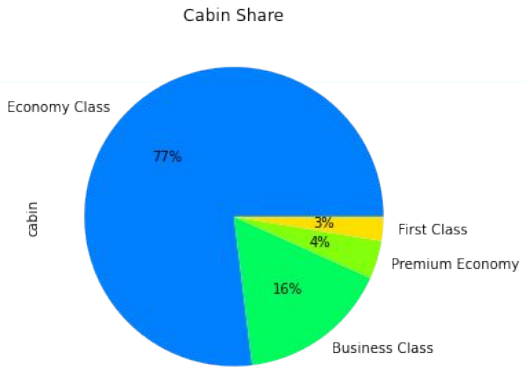
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 131895 entries, 0 to 131894
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   airline          65947 non-null  object 
 1   overall          64017 non-null  float64
 2   author           65947 non-null  object 
 3   review_date      65947 non-null  object 
 4   customer_review  65947 non-null  object 
 5   aircraft         19718 non-null  object 
 6   traveller_type   39755 non-null  object 
 7   cabin            63303 non-null  object 
 8   route            39726 non-null  object 
 9   date_flown       39633 non-null  object 
 10  seat_comfort     60681 non-null  float64
 11  cabin_service    60715 non-null  float64
 12  food_bev         52608 non-null  float64
 13  entertainment    44193 non-null  float64
 14  ground_service   39358 non-null  float64
 15  value_for_money  63975 non-null  float64
 16  recommended      64440 non-null  object 
dtypes: float64(7), object(10)
```

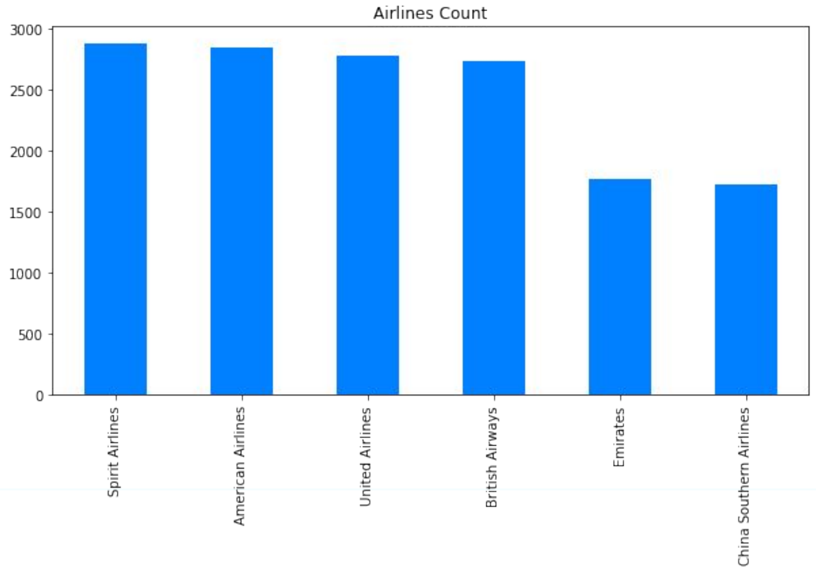
Feature Description:-

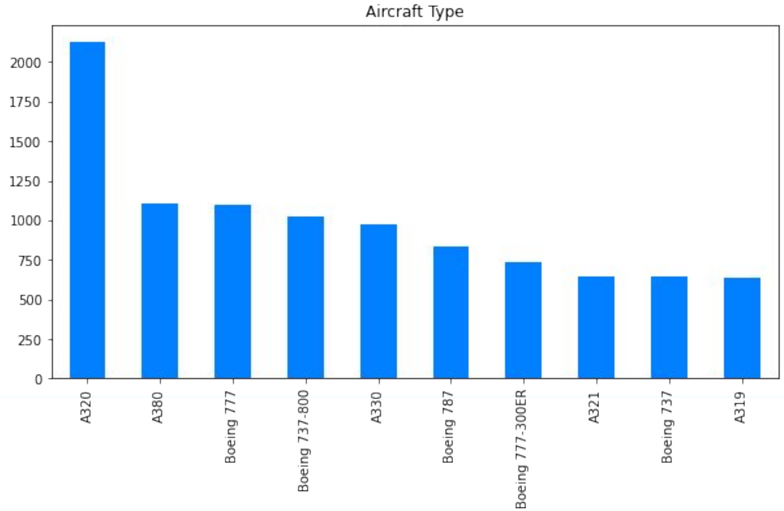
- **airline**: Name of the airline in str format.
- **overall**: Overall point is given to the trip between 1 to 10 in float format.
- **author**: Author of the trip in str format.
- **reviewdate**: Date of the Review customer review: Review of the customers in free text format in str need to be converted into DateTime Format.
- **aircraft**: Type of the aircraft in str format.
- **travellertype**: Type of traveler (e.g. business, leisure) consist of four class in str format.
- **cabin**: Cabin at the flight date flown: Flight date in str format consist of 4 class.
- **seatcomfort**: Rated between 1-5 in float format.
- **cabin service**: Rated between 1-5 float format.
- **foodbev**: Rated between 1-5 entertainment: Rated between 1-5 in float format.
- **groundservice**: Rated between 1-5 in float format.
- **valueformoney**: Rated between 1-5 in float format.

Exploratory Data Analysis

EDA for Cabin, Airlines Company and Aircraft Carrier has been done which showed the following output.

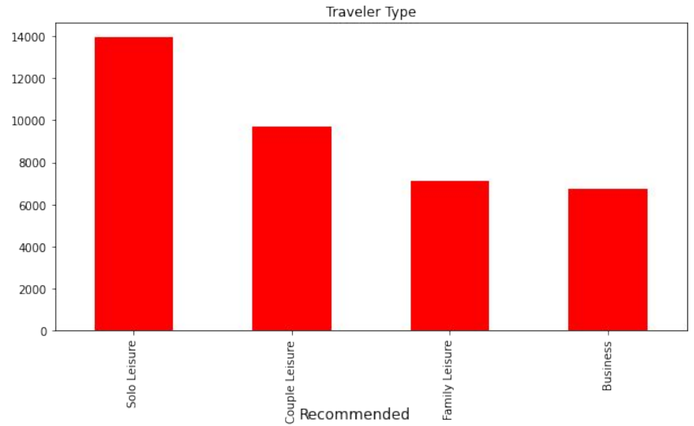


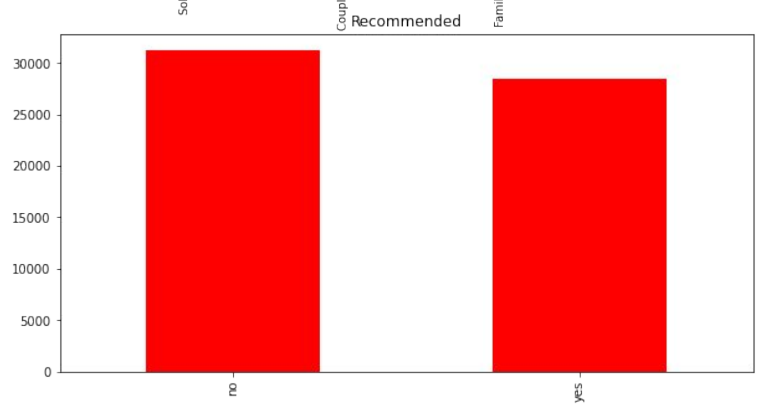




Exploratory Data Analysis

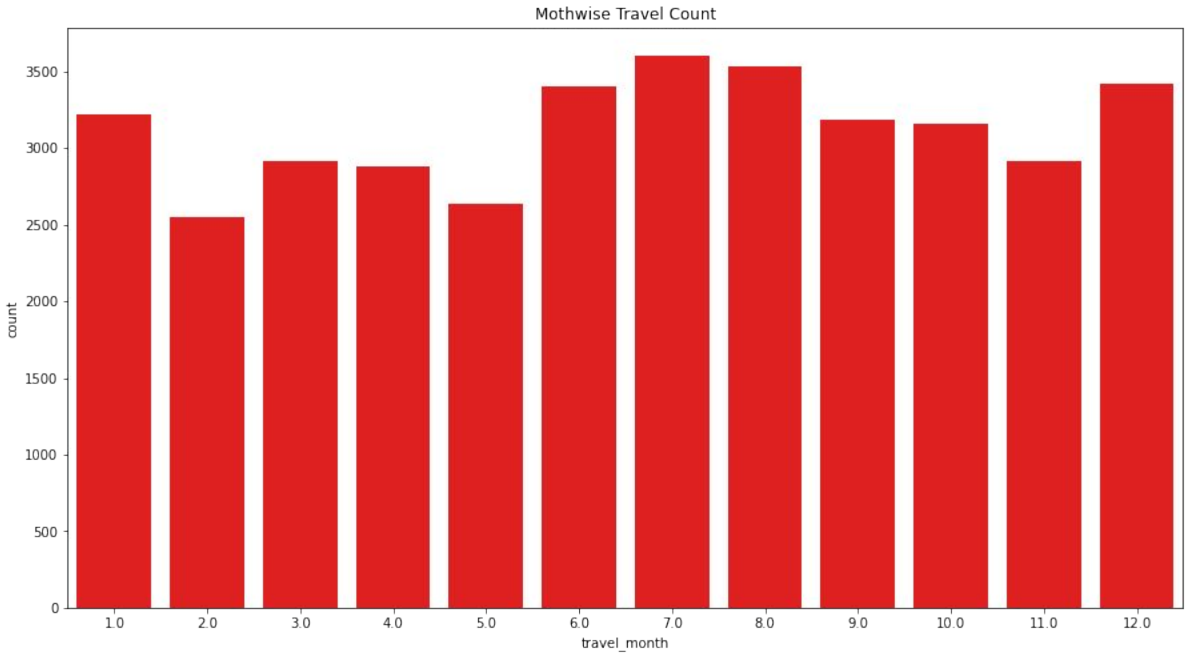
- We can see there are 4 classes present in the Traveler type feature. Also, we can notice that Solo Leisure has the highest value count. From this, we can conclude that most people who travel by airline travel in solo. Followed by College then Family. A very small percentage of people prefer flying for business.
- In recommended plot we can see that the Dependent feature ‘recommended’ has balanced data in its classes Yes and No.





Exploratory Data Analysis

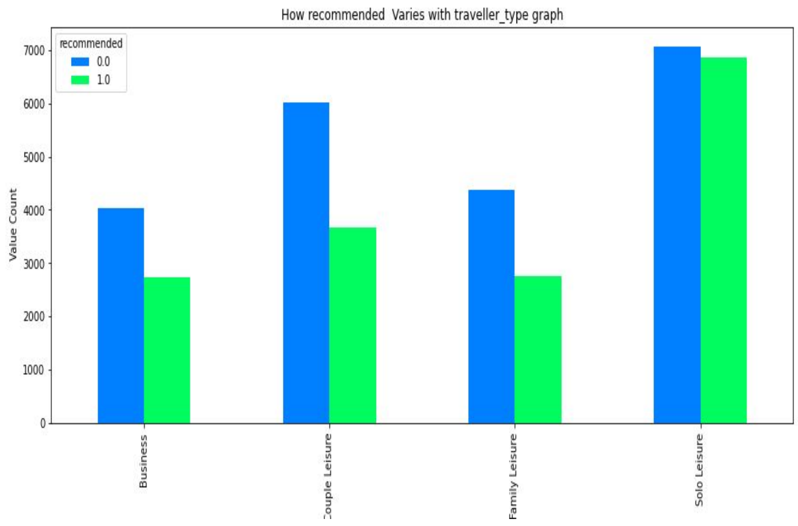
Here we can see that people have flown most frequently in the month of July and least frequently in the month of February.



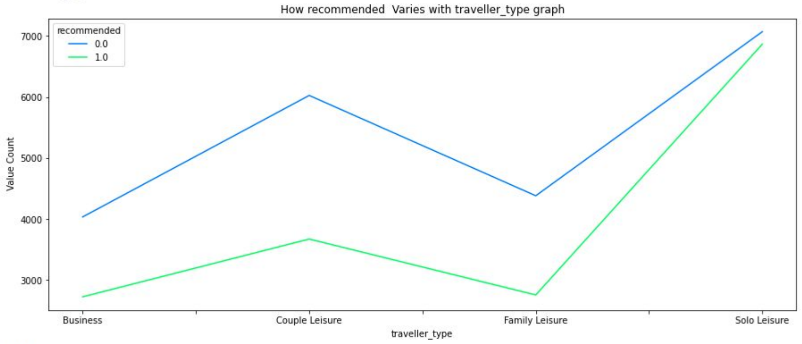
Exploratory Data Analysis

Variation of recommendation feature Traveller type:

- We can see that people have given both 1 or 0 which we will consider from now on as positive and negative recommendation so to interpret it effectively to the solo leisure. This may because of the poor infrastructure or the service received by the people and positive recommendation may be because of low price for solo. But this is approximate analysis based on the data provided.
- In Traveller type we can see that both the recommendation trend as of yes or no increases from business to couple leisure and decreases to family then again increases high in solo leisure. Which indicate people prefer solo leisure higher than any of the other leisures.



traveller_type

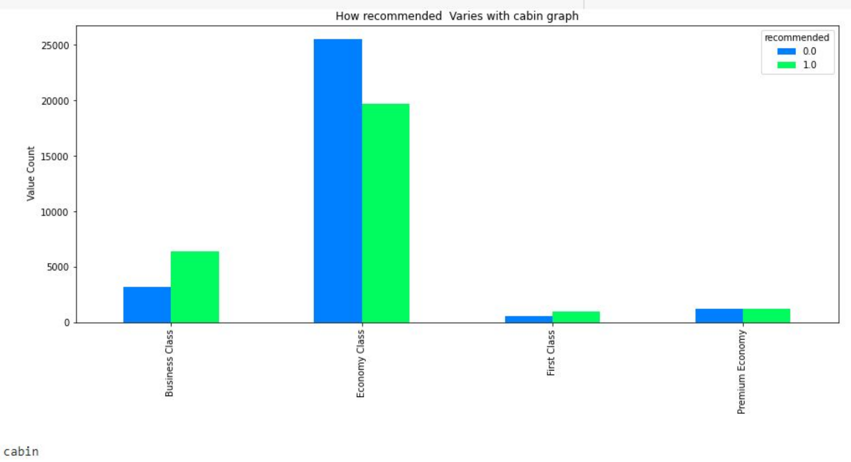


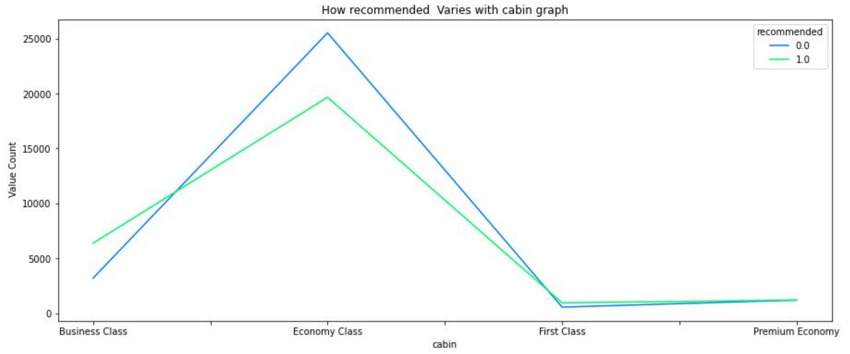
cabin

Exploratory Data Analysis

Variation of Recommendation with Cabin Type:

- we can see that people gives the high positive recommendation to economic class in cabin. From this we can conclude that people love to travel in economic class as of low price also in same way we can see people give highest negative recommendation to economy class maybe because less infrastructure or service provided to them. Also we can see people have given higest positive recommendation to Business class it may be because of the quality of service provided to them in Business class and similarly negative recommendation because of high price of business class or less travelling percentage.
- In Cabin type we can see that both the recommendation trend as of yes or no increases from business to Economy class and decreases to First class then again increases slightly in Premium class. Which indicate most people travel on economy class.

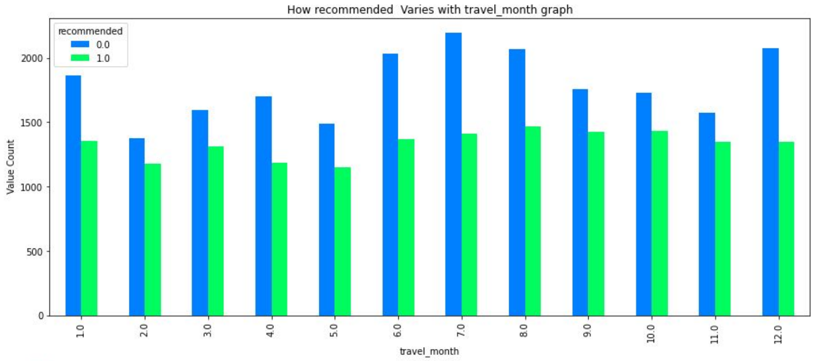




Exploratory Data Analysis

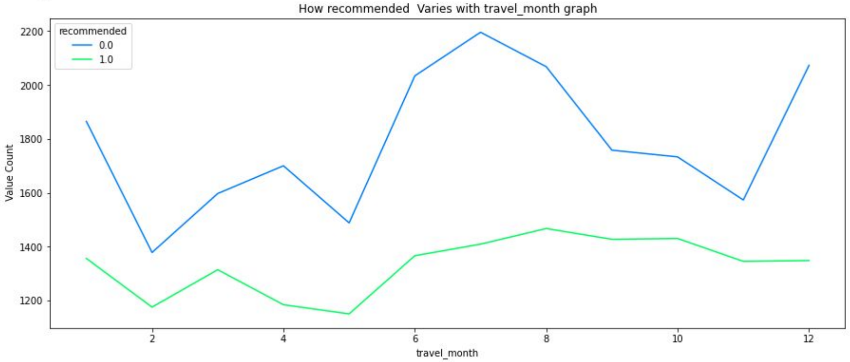
Variation of Recommendation feature with Travel Month:

- From month vs no. of recommendation. We can see that people tents to travel most in the month of July considering the total of positive and negative recommendation combined.



- In month we cannot see any preferable trend but here we can conclude people tent to travel highest during the month of July.

travel_month

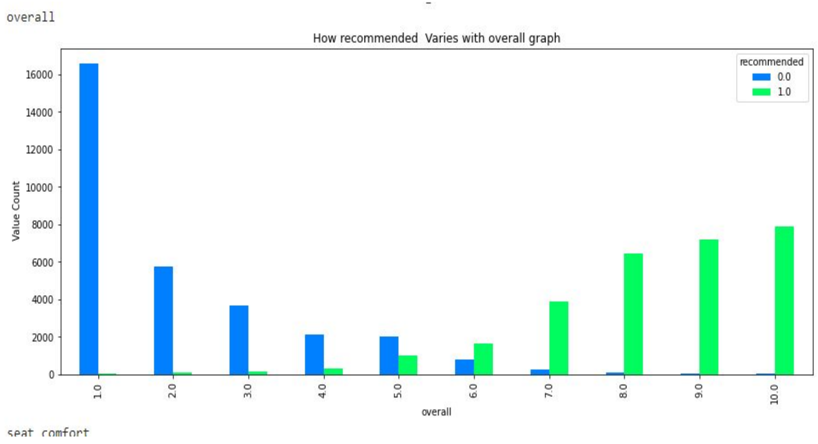


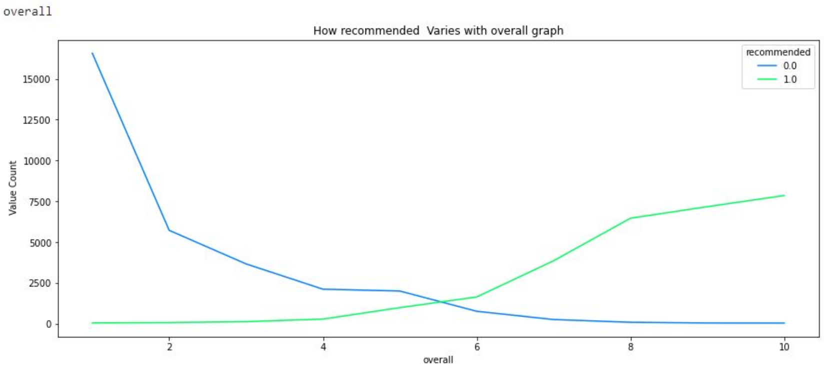
Exploratory Data Analysis

Variation of Recommendation feature with overall rating:

- From overall rating vs recommended graph we can see which is perfectly understandable that negative recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10. But it is very true that highest negative recommendation has been given to overall rating of 1.0 which is really a matter of concern.

- In overall rating we can experience a very good insights which is also regular. We can see as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases.





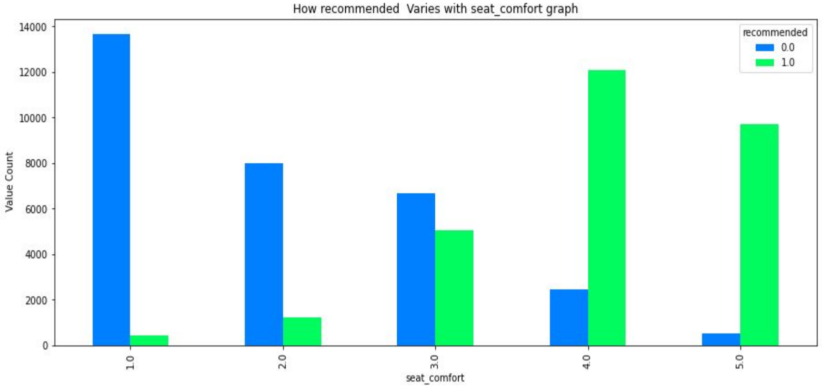
Exploratory Data Analysis

Variation of Recommendation feature with seat comfort :

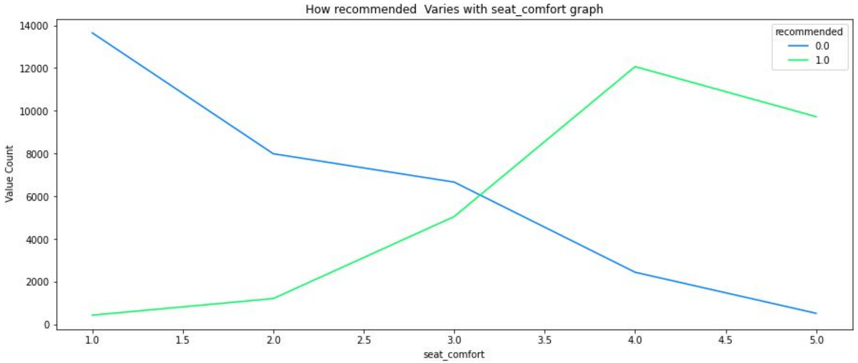
- In seat comfort people has given highest positive recommended to the seat of class 5 as compared to very low negative recommendation to the same. Also we can see seat of class 1 have been given highest negative recommendation as compare to its positive recommendation. Here we come to a conclusion it must be removed as early as possible.

- 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.

seat_comfort



seat_comfort



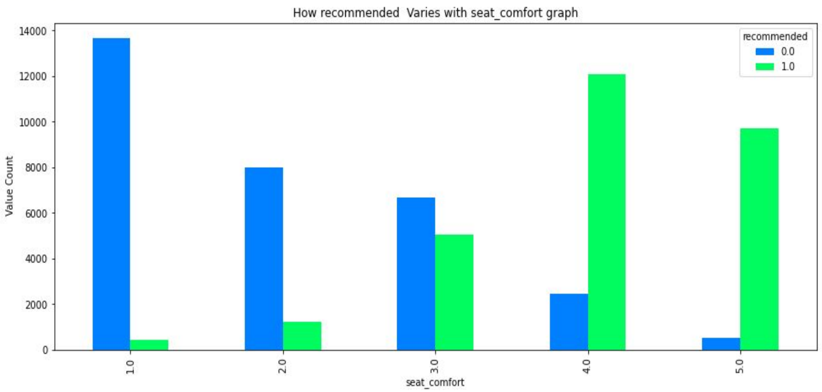
Exploratory Data Analysis

Variation of Recommendation feature with seat comfort :

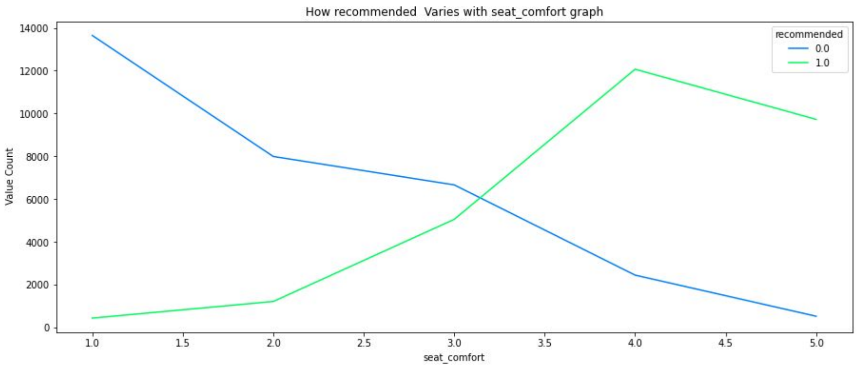
- In seat comfort people has given highest positive recommended to the seat of class 5 as compared to very low negative recommendation to the same. Also we can see seat of class 1 have been given highest negative recommendation as compare to its positive recommendation. Here we come to a conclusion it must be removed as early as possible.

- 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.

seat_comfort



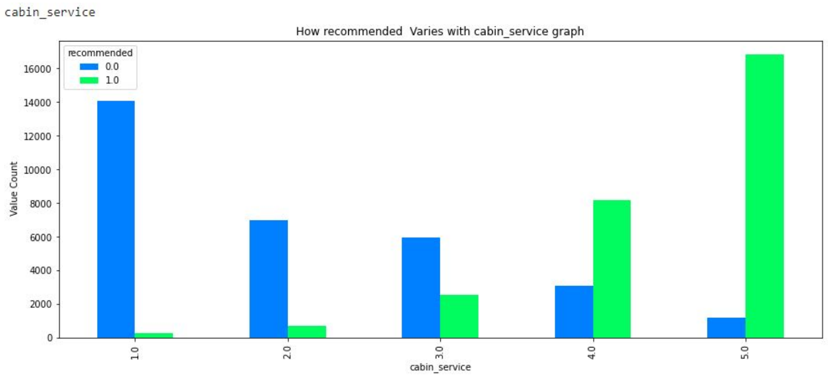
seat_comfort



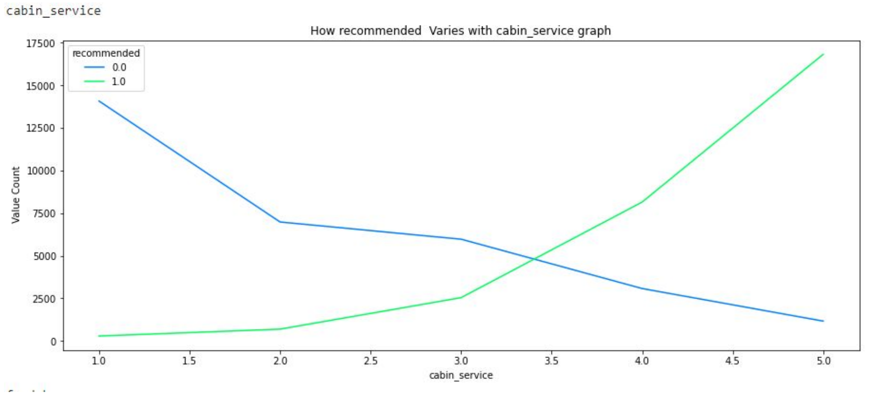
Exploratory Data Analysis

Variation of Recommendation feature with Cabin Service :

- In cabin service rating people has given highest recommendation to rating to cabin service rating 5 as compare to its counterpart. From this we can conclude that cabin service is doing pretty good.



- 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

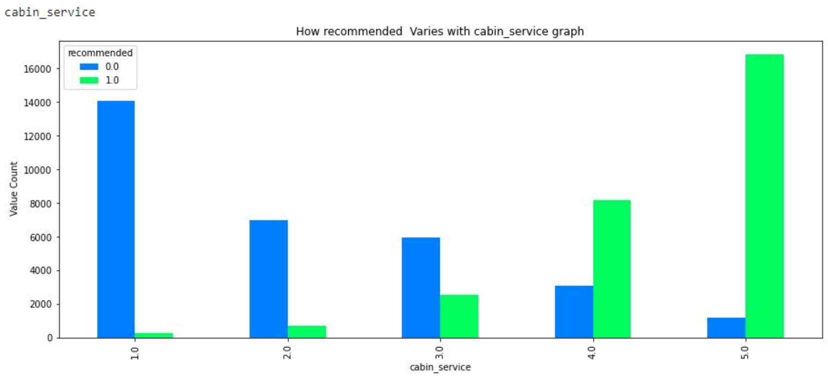


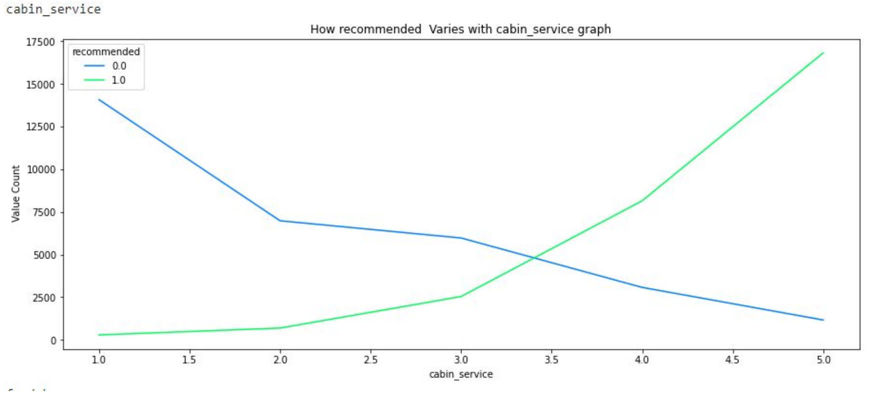
Exploratory Data Analysis

Variation of Recommendation feature with Food Bev :

- In food and beverage rating people have given highest negative recommendation to rating 1.0 from this we can conclude that airline service has to improve their food delivery and quality service.

- 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.



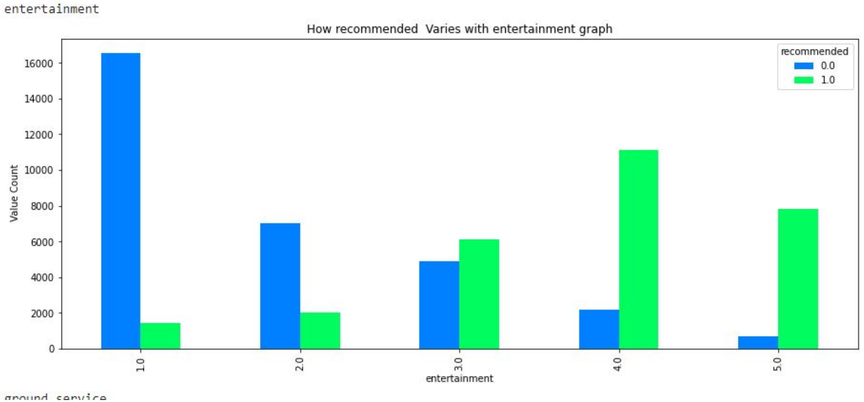


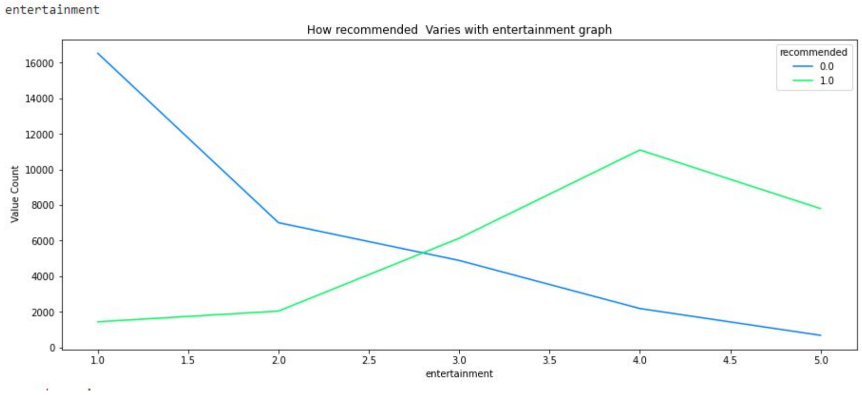
Exploratory Data Analysis

Variation of Recommendation feature with Entertainment:

- In entertainment also we can see most people has given highest negative recommendation to entertainment rating 1 which shows that airline has to improve their entertainment system as well.

- In Entertainment service too 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 Entertainment service rating between 2.5 and 3.0 where we can see similar positive and negative recommendation.



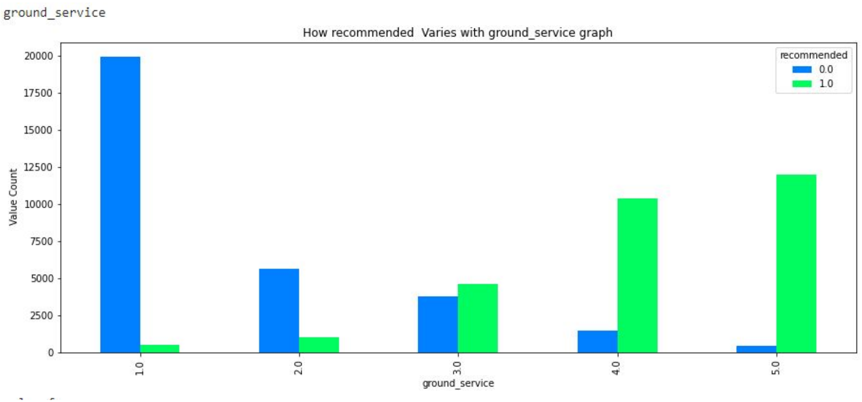


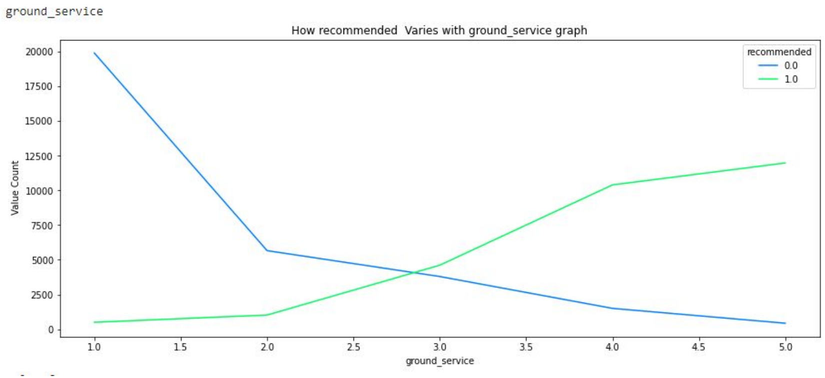
Exploratory Data Analysis

Variation of Recommendation feature with Ground Service:

- In Ground Service also we can see most people has given highest negative recommendation to entertainment rating 1 which shows that airline has to improve their entertainment system as well.

- In Ground Service too 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 Entertainment service rating between 2.5 and 3.0 where we can see similar positive and negative recommendation.





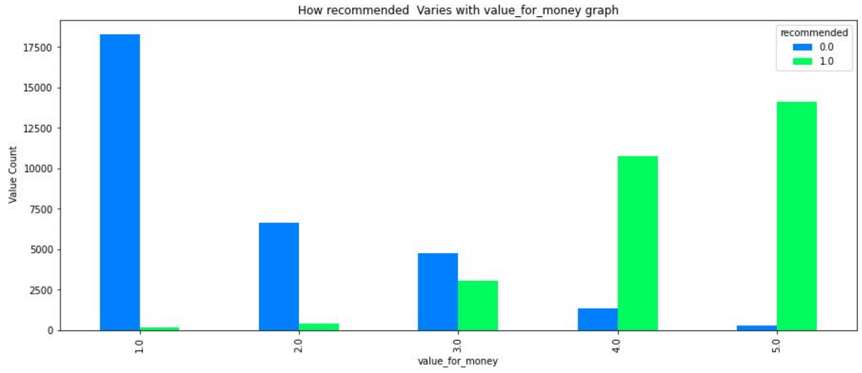
Exploratory Data Analysis

Variation of Recommendation feature with Value for Money:

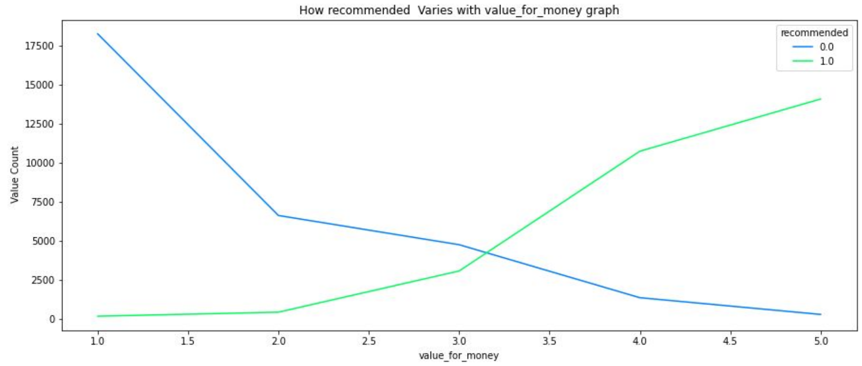
- In ground service also we can see most people has given highest negative recommendation to ground service rating 1 which shows that airline has to improve their ground service.

- In Ground service also 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 Ground service rating close 3.0 where we can see similar positive and negative recommendation.

value_for_money

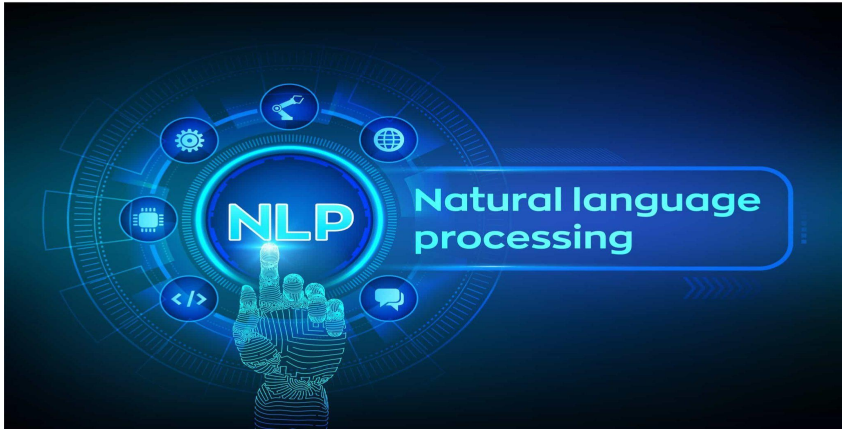


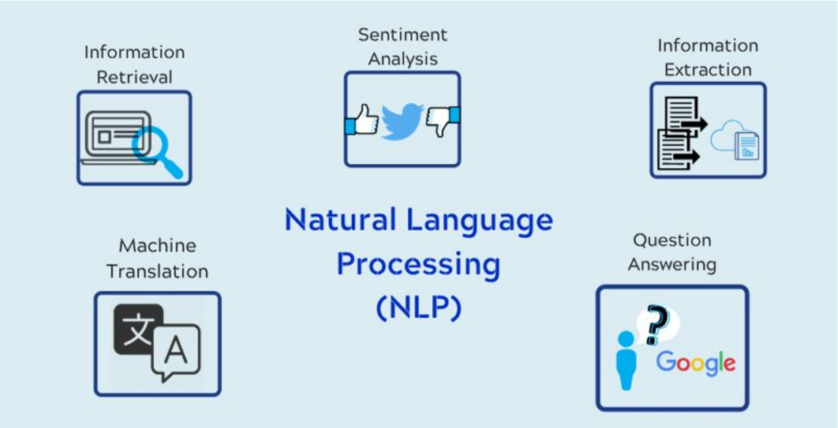
value_for_money



NLP(Natural Language Processing):

- We have used vander sentiment in NLP so to convert sentiments in customer review into score so to have our model prediction.
- We have also created new feature numeric review so to store sentiment score we have retrieved using sentiment analysis from customer review feature.

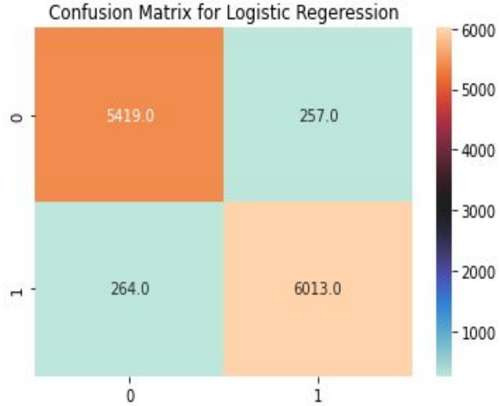




Model Building:

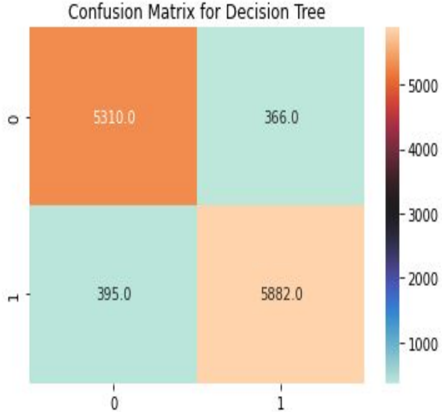
	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.95	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.64%



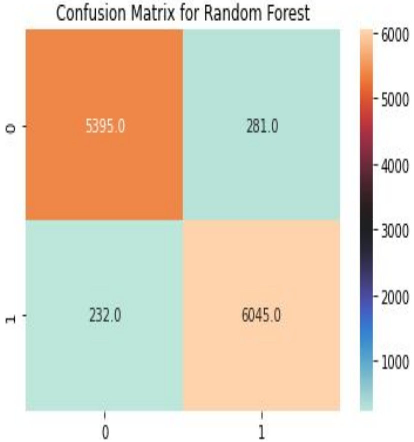
	precision	recall	f1-score	support
0.0	0.94	0.94	0.94	6277
1.0	0.93	0.94	0.93	5676
accuracy			0.94	11953
macro avg	0.94	0.94	0.94	11953
weighted avg	0.94	0.94	0.94	11953

Accuracy score % of the model is 93.63%



	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

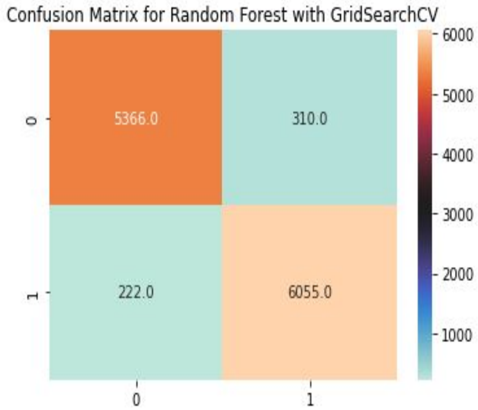
Accuracy score % of the model is 95.71%



Model Building(Continued….)

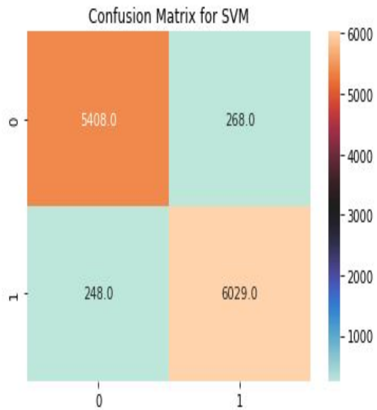
	precision	recall	f1-score	support
0.0	0.95	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.55%



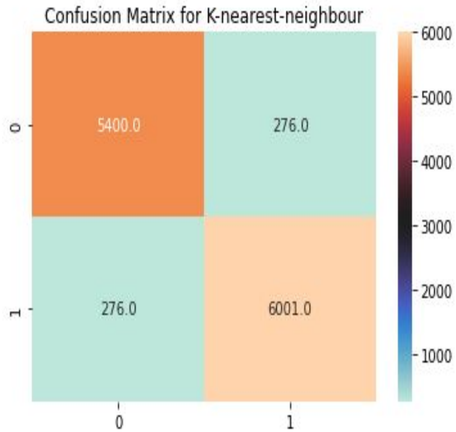
	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.68%



	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.95	0.95	0.95	5676
accuracy			0.95	11953
macro avg	0.95	0.95	0.95	11953
weighted avg	0.95	0.95	0.95	11953

Accuracy score % of the model is 95.38%



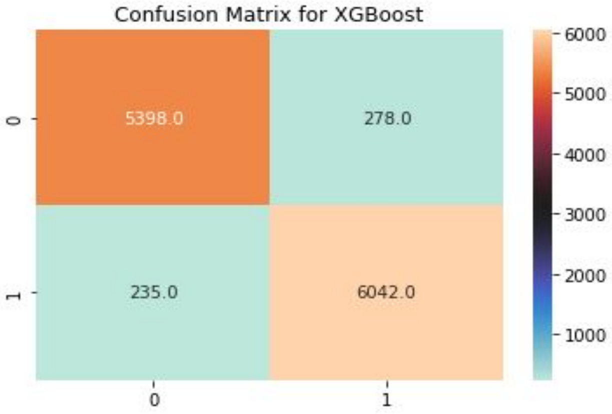
Model Building(Continued….)

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.58%

	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	6277
1.0	0.96	0.95	0.95	5676
accuracy			0.96	11953
macro avg	0.96	0.96	0.96	11953
weighted avg	0.96	0.96	0.96	11953

Accuracy score % of the model is 95.71%



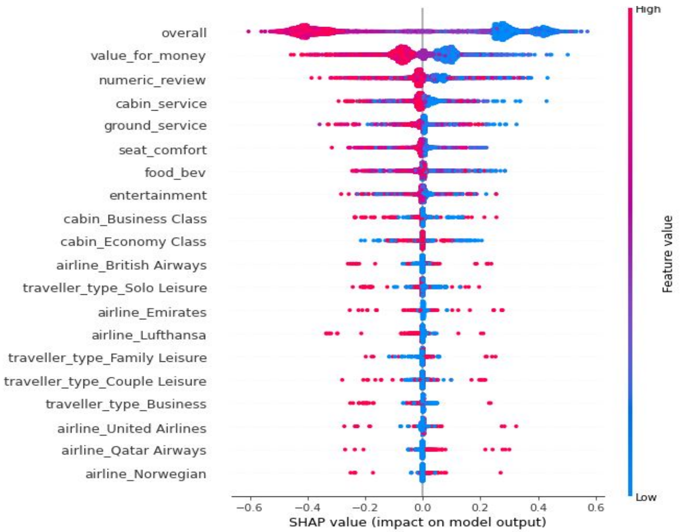
Model Building(Continued….)

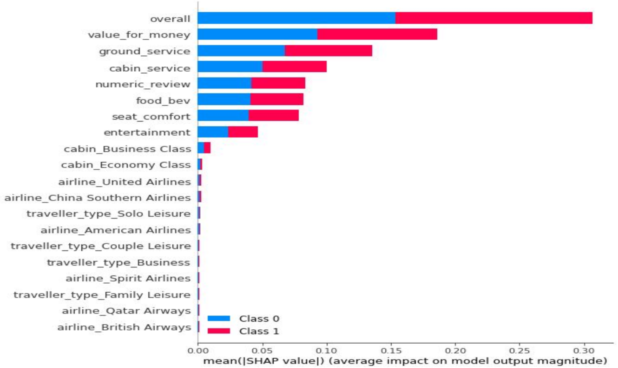
1. In model Selection we can see that Random Forest and XGBoost Model is having the same high Model Accuracy with a score 0.957082 but we can also see that recall, precision, f1-score and roc_auc_score of XGBoost model combined is giving higher score than Random Forest from which we have chosen XGBoost Model for further prediction.

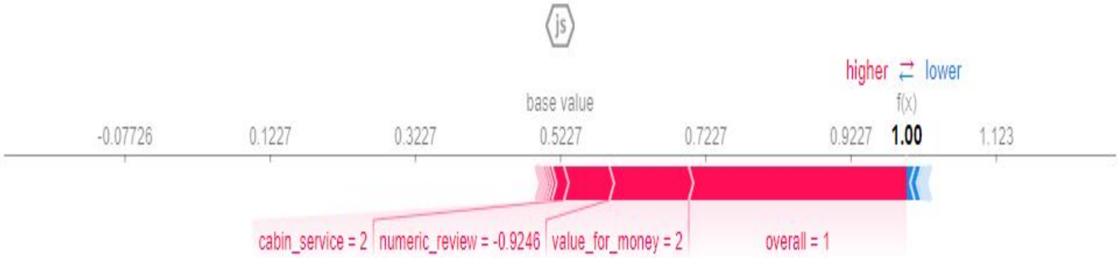
	Model	Accuracy	Recall	Precision	f1-score	roc_auc_score
0	Logistic Regression	0.956413	0.954722	0.953546	0.954133	0.956332
1	Decision Tree	0.936334	0.935518	0.930762	0.933134	0.936295
2	Random Forest	0.957082	0.950493	0.958770	0.954614	0.956766
3	Random Forest with GridSearchCV	0.955492	0.945384	0.960272	0.952770	0.955008
4	SVM	0.956831	0.952784	0.956153	0.954465	0.956637
5	K-nearest-neighbour	0.953819	0.951374	0.951374	0.951374	0.953702
6	K-nearest-neighbour	0.955827	0.950141	0.956545	0.953332	0.955555
7	XGBoost	0.957082	0.951022	0.958282	0.954638	0.956792
8	K-nearest-neighbour with GridSearchCV	0.955827	0.950141	0.956545	0.953332	0.955555

Model Explainability: SHAP:

- In Shap JS summary we can see positive features overall, value for money,numeric_review combined red color block pushes the prediction toward right over base value and causing positive model prediction and it is common for all model.
- In Shap summary scatter plot we can see in scatter plot high overall,value for money,numeric_review,cabin service,ground_service positive features and low airline_British_airways is increasing positive prediction and it is common for all models. Also we can see that overall,value for money,numeric_review,cabin service,ground_service has high shap feature value.







Conclusion:

- We can see that people have given both 1 or 0 which we will consider from now on as positive and negative recommendation so to interpret it effectively to the solo leisure. This may because of the poor infrastructure or the service received by the people and positive recommendation may be because of low price for solo. But this is approximate analysis based on the data provided.
- Also we can see that people gives the high positive recommendation to economic class in cabin. From this we can conclude that people love to travel in economic class as of low price also in same way we can see people give highest negative recommendation to economy class maybe because less infrastructure or service provided to them. Also we can see people have given highest positive recommendation to Business class it may be because of the quality of service provided to them in Business class and similarly negative recommendation because of high price of business class or less travelling percentage.
- From month vs no. of recommendation. We can see that people tents to travel most in the month of July considering the total of positive and negative recommendation combined.
- From overall vs recommended graph we can see which is perfectly understandable that negative recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10. But it is very true that highest negative recommendation has been given to overall rating of 1.0 which is really a matter of concern.
- In seat comfort people has given highest positive recommended to the seat of class 5 as compared to very low negative recommendation to the same. Also we can see seat of class 1 have been given highest negative recommendation as compare to its positive recommendation. Here we come to a conclusion it must be removed as early as possible.

Conclusion:

- In cabin service rating people has given highest recommendation to rating to cabin service rating 5 as compare to its counterpart. From this we can conclude that cabin service is doing pretty good.
- In food and beverage rating people have given highest negative recommendation to rating 1.0 from this we can conclude that airline service has to improve their food delivery and quality service.
- In entertainment also we can see most people has given highest negative recommendation to entertainment rating 1 which shows that airline has to improve their entertainment system as well.
- In ground service also we can see most people has given highest negative recommendation to ground service rating 1 which shows that airline has to improve their ground service.
- In value for money also we can see most people has given highest negative recommendation to value for money rating 1 which shows that airline has to make their flight service more cost effective.
- In model Selection we can see that Random Forest and XGBoost Model is having the same high Model Accuracy with a score 0.957082 but we can also see that recall, precision, f1-score and roc_auc_score of XGBoost model combined is giving higher score than Random Forest from which we have chosen XGBoost Model for further prediction.
- In Shap JS summary we can see positive features overall, value for money,numeric_review combined red color block pushes the prediction toward right over base value and causing positive model prediction and it is common for all model.
- In Shap summary scatter plot we can see in scatter plot high overall,value for money,numeric_review,cabin service,ground_service positive features and low airline_British_airways is increasing positive prediction and it is common for all models. Also we can see that overall,value for money,numeric_review,cabin service,ground_service has high shap feature value.

Thank you