Capstone Project

Airline Passenger Referral Prediction

By

Ankit Patil
Naga Sai Kiran
Saugata Deb
Shreyash Movale

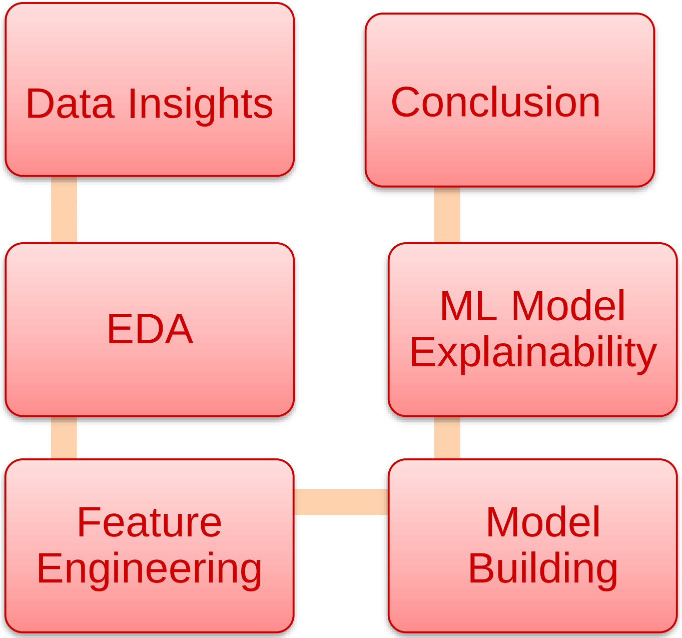
Table of Contents

Objective

- The 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.
- to predict whether passengers will refer the airline to their friends

Process Flow

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



Data Insights

- The data set has 16 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.

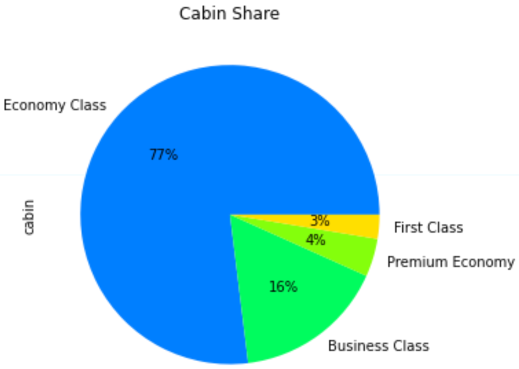
```
df.info()
```

```
<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)
```

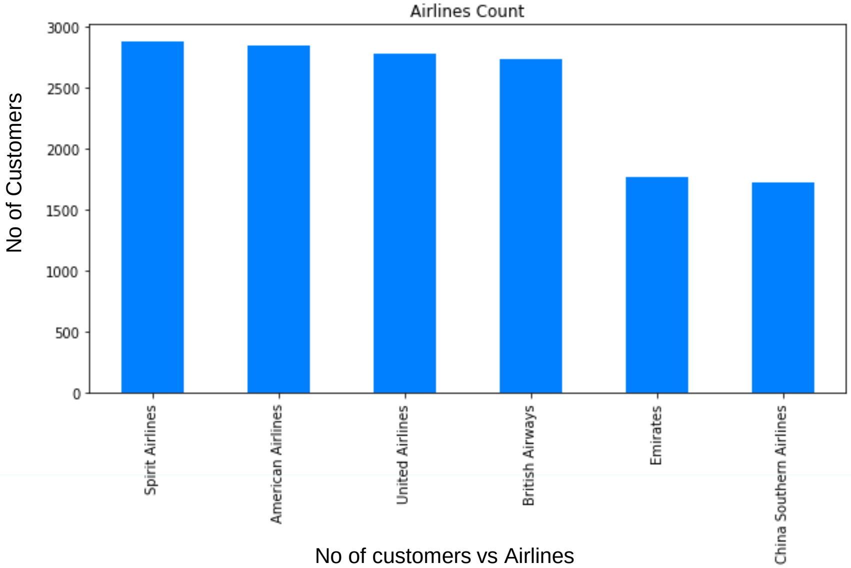
Feature Description:-

Airline: Name of the airline.
overall: Overall point is given to the trip between 1 to 10.
author: Author of the trip
Review date: Date of the Review customer review: Review of the customers in free text format
Customer Review: Feedback shared by the customers
Aircraft: Type of the aircraft
Traveler Type: Type of traveler (e.g. business, leisure)
Cabin: Cabin
Flight date: Date on which The flight has flown
Route: Route taken by flight
Seat comfort: Rated between 1-5
cabin service: Rated between 1-5
Food-Bev: Rated between 1-5
entertainment: Rated between 1-5
Ground service: Rated between 1-5
Value for money: Rated between 1-5
Recommended: The passenger has referred his friend or not.

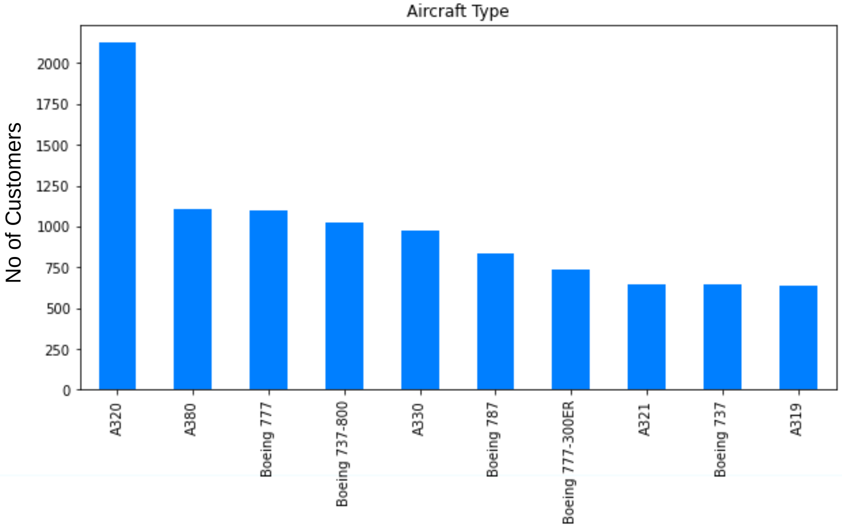
Exploratory Data Analysis



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



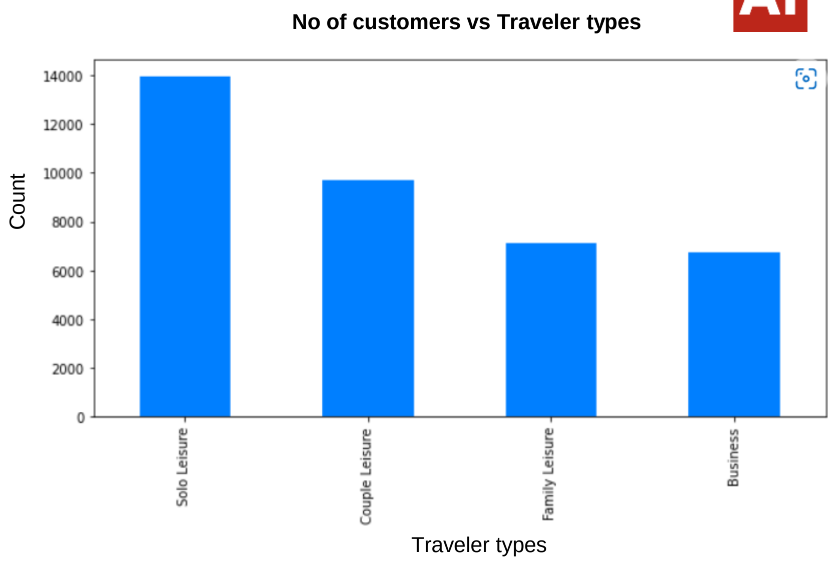
No of customers vs Airlines

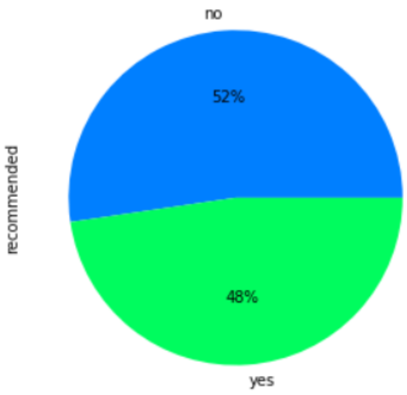


No of customers vs Aircrafts

Exploratory Data Analysis

- We can see there are 4 classes present in the Traveller 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.

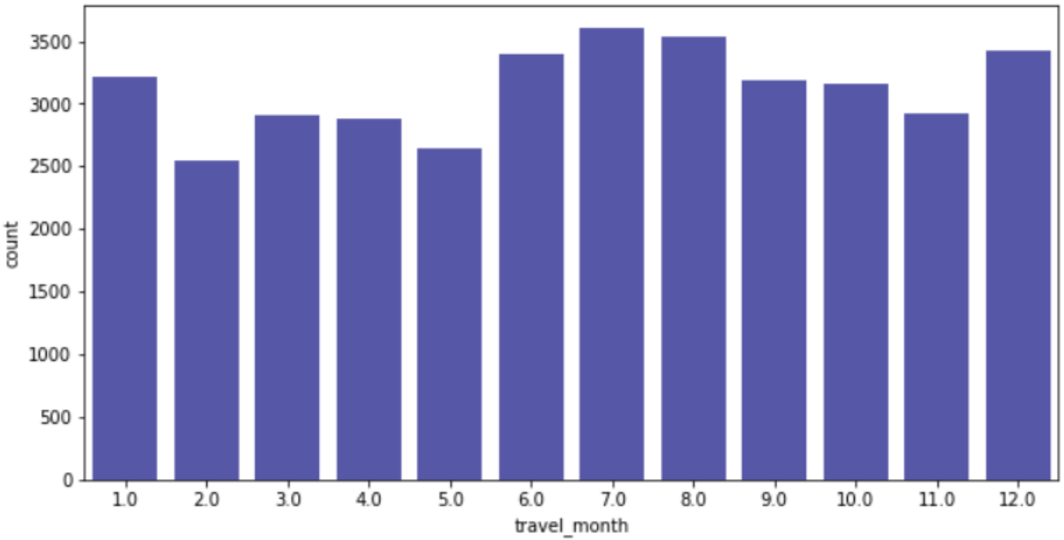




Exploratory Data Analysis

- Here we can see that people have flown most frequently in the months June to August and in December and January.
- Least frequently in the month of February.

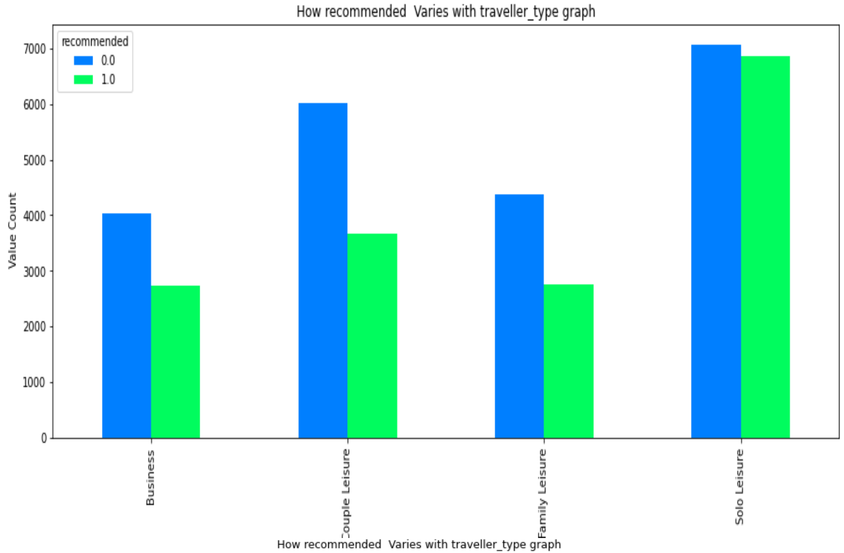
Month wise travel count

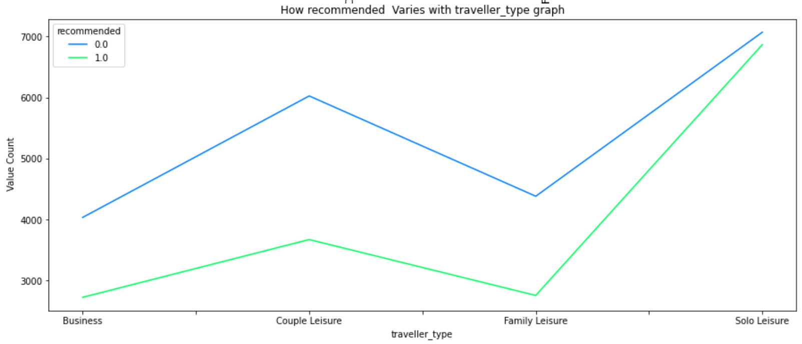


Exploratory Data Analysis

Variation of Traveller type feature with recommendation:

- 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
- 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 leisure.



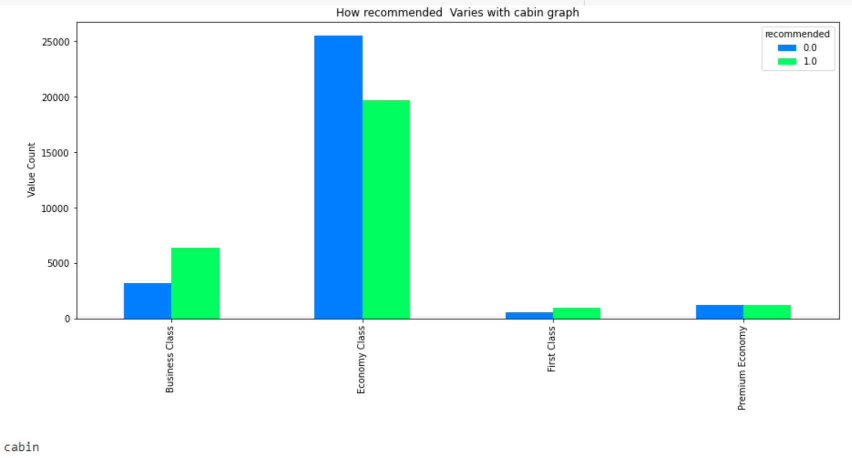


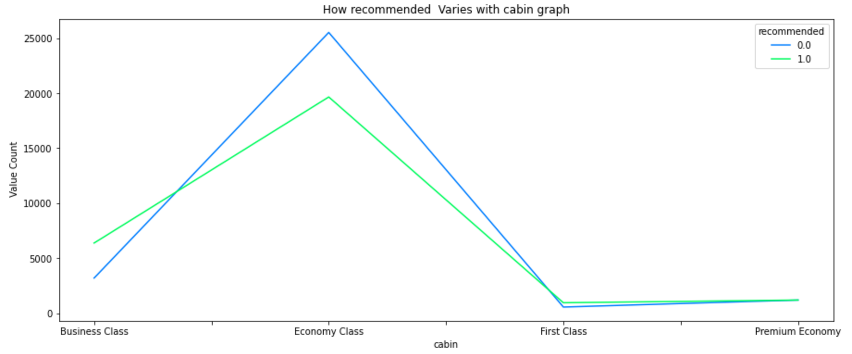
cabin

Exploratory Data Analysis

Variation of Traveller type feature with Cabin:

- For Business class, more number of customers have recommended airlines as compared to non recommendations.
- Whereas this scenario reverses for Economy class i.e. more customers have not recommended
- We can say that customers who have travelled from business class are satisfied with airline's services whereas those who travelled from economy class are less satisfied with services hence the less recommendation

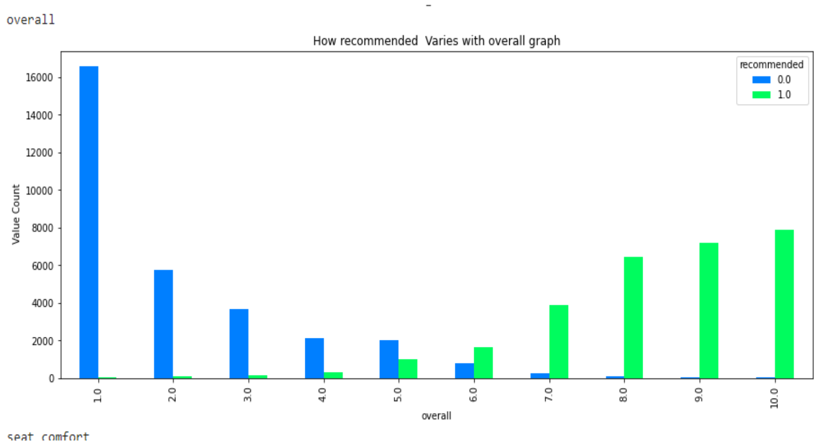


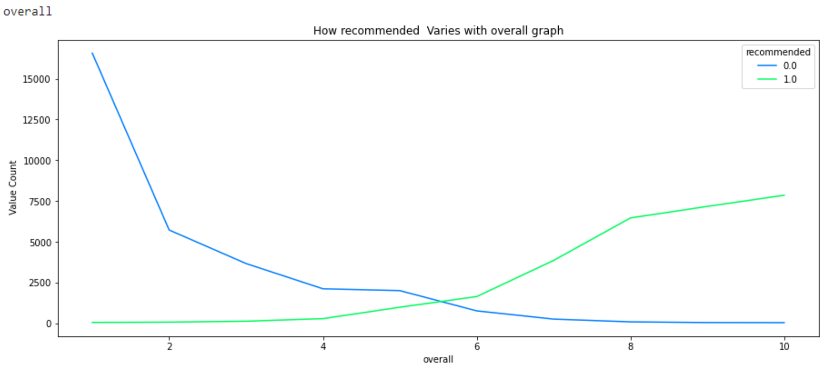


Exploratory Data Analysis

Variation of Traveller type feature with overall rating:

- From overall rating vs recommended graph we can see which is perfectly understandable that non recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10
- In overall rating we can experience the positive recommendation increases with the overall rating and non recommendation on the same decreases.

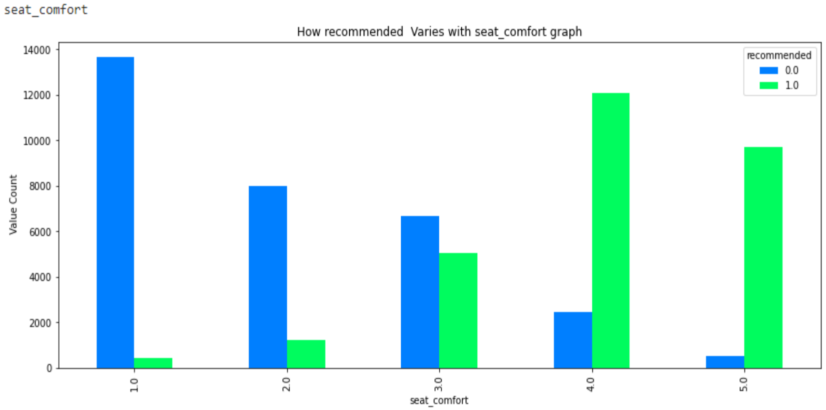


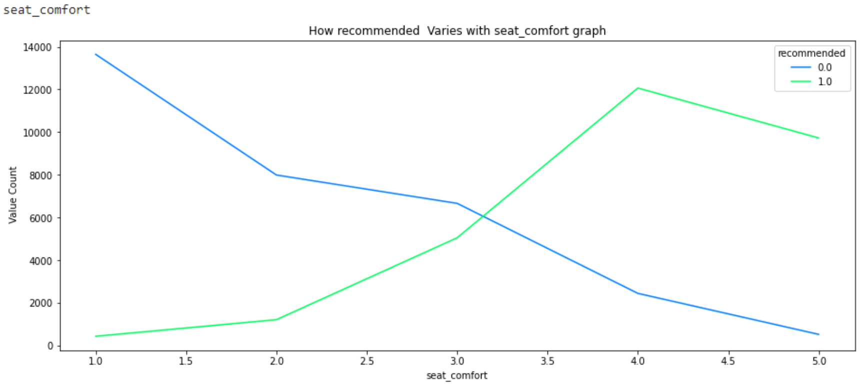


Exploratory Data Analysis

Variation of Traveller type 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
- 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.



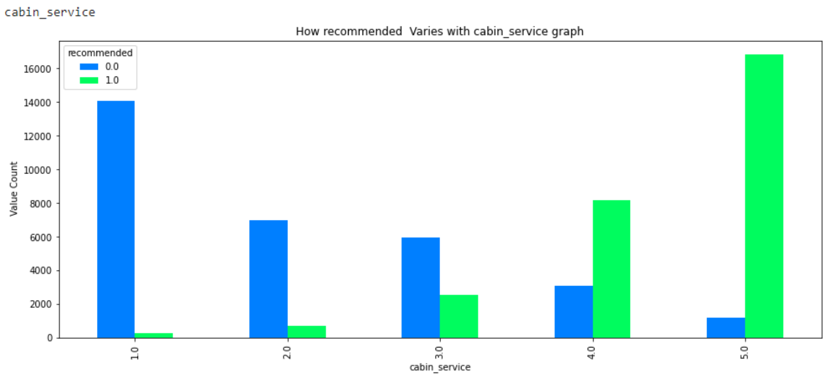


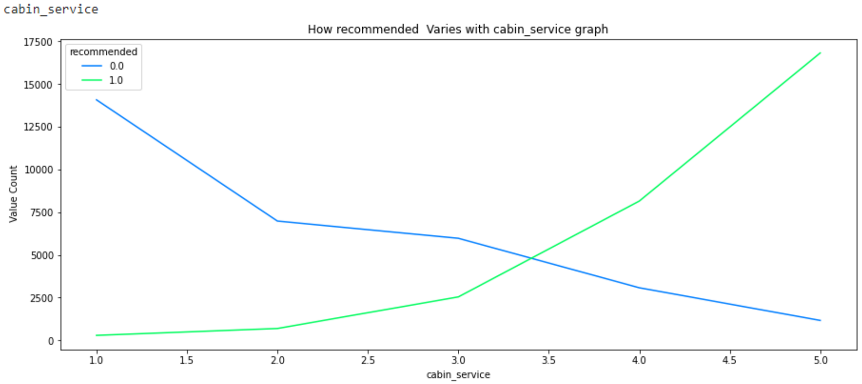
Exploratory Data Analysis

Variation of Traveller type 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



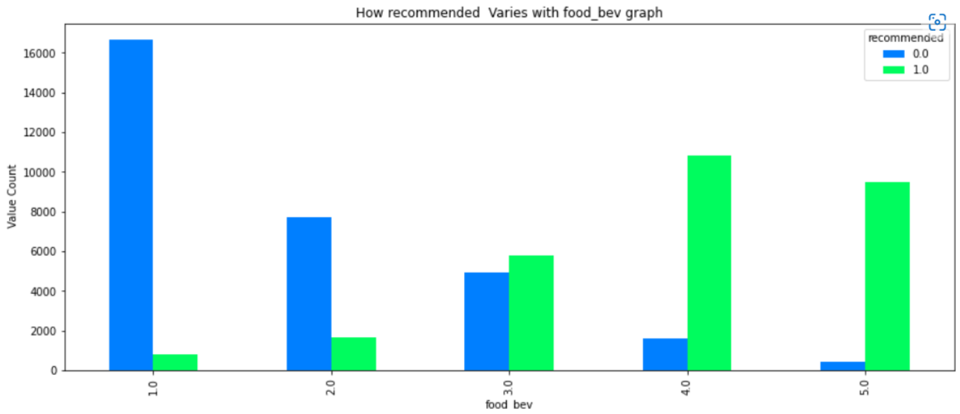


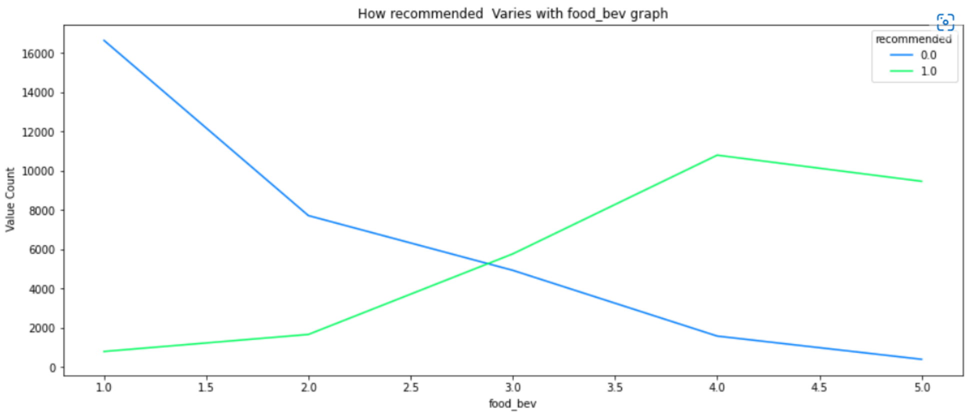
Exploratory Data Analysis

Variation of Traveller type 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.

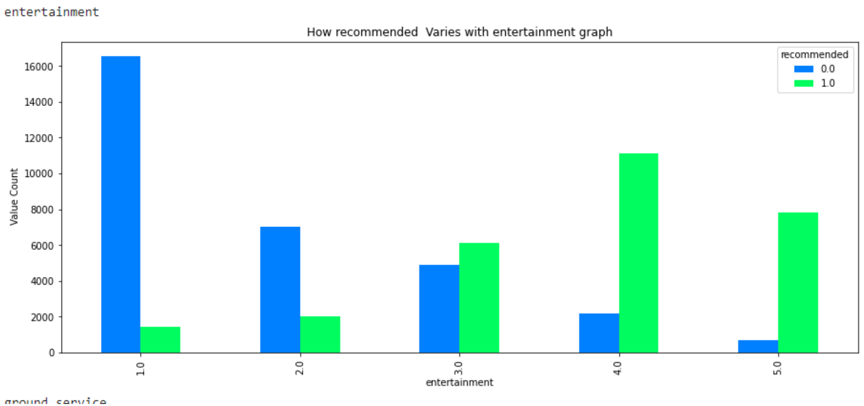




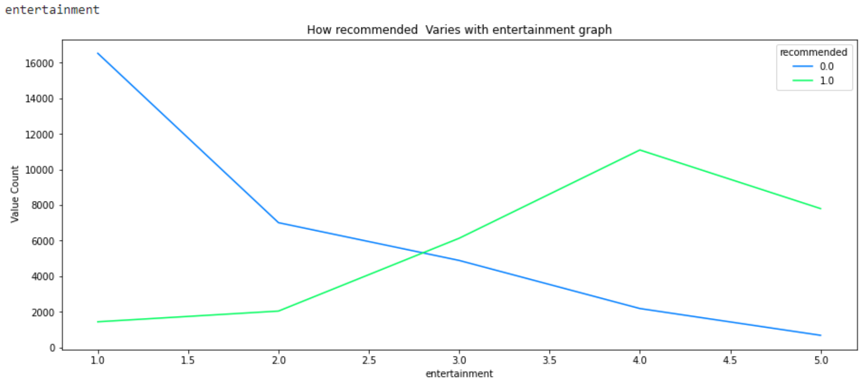
Exploratory Data Analysis

Variation of Traveller type 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.



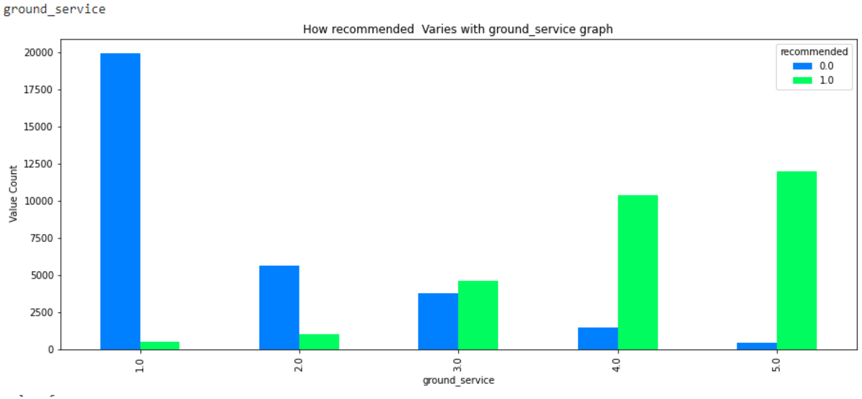
- In Entertainment service too we can see same as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also we can an intersection in Entertainment service rating between 2.5 and 3.0 where we can see similar positive and negative recommendation.



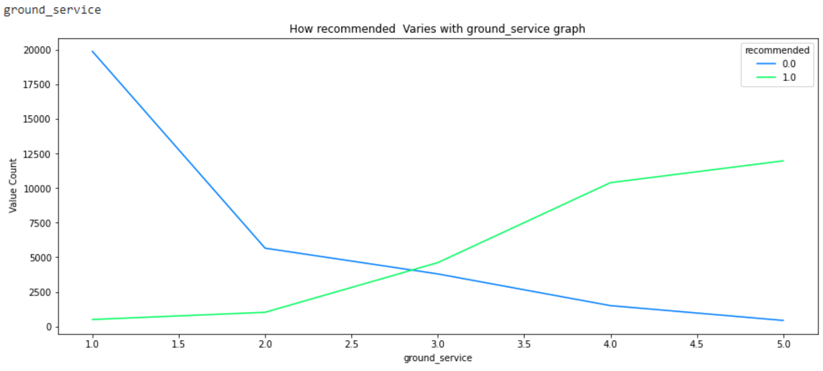
Exploratory Data Analysis

Variation of Traveller type feature with Ground Service:

- 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.

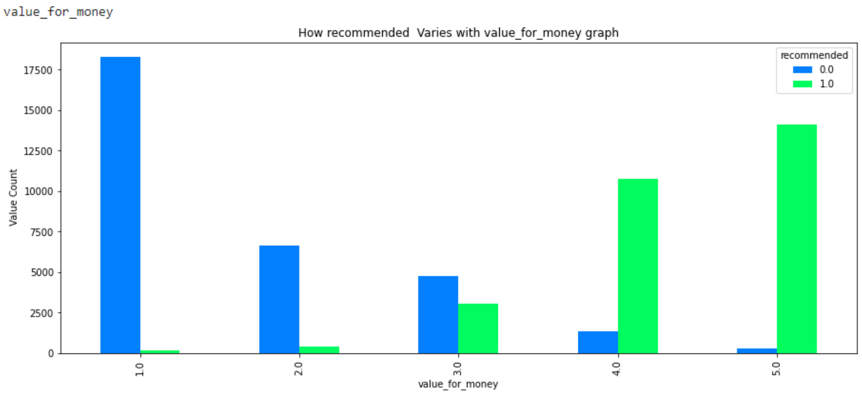


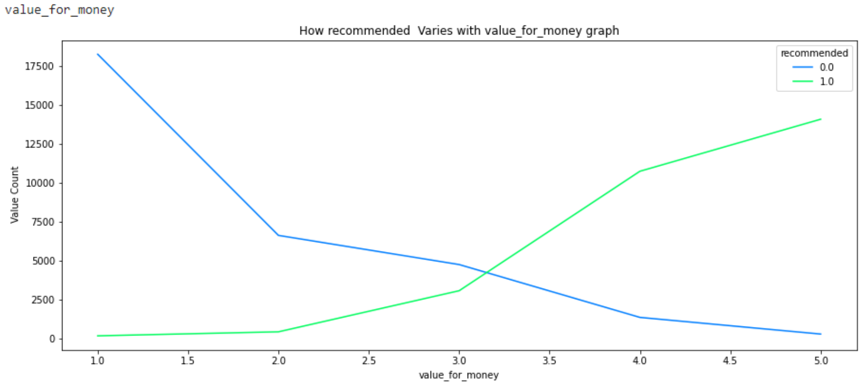
Exploratory Data Analysis

Variation of Traveller type 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.



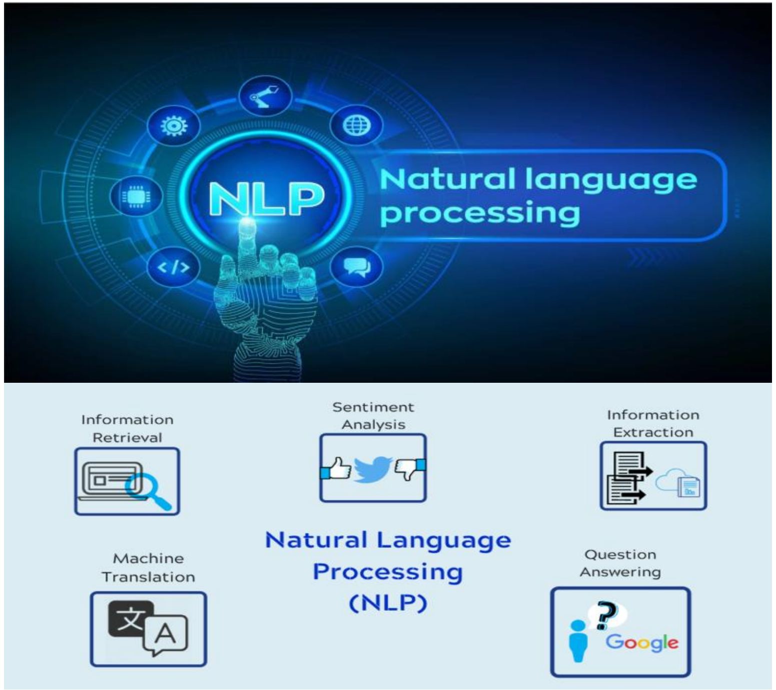


NLP(Natural Language Processing):

- We have used vader sentiment in NLP so to convert sentiments in customer review into score so to have our model prediction.

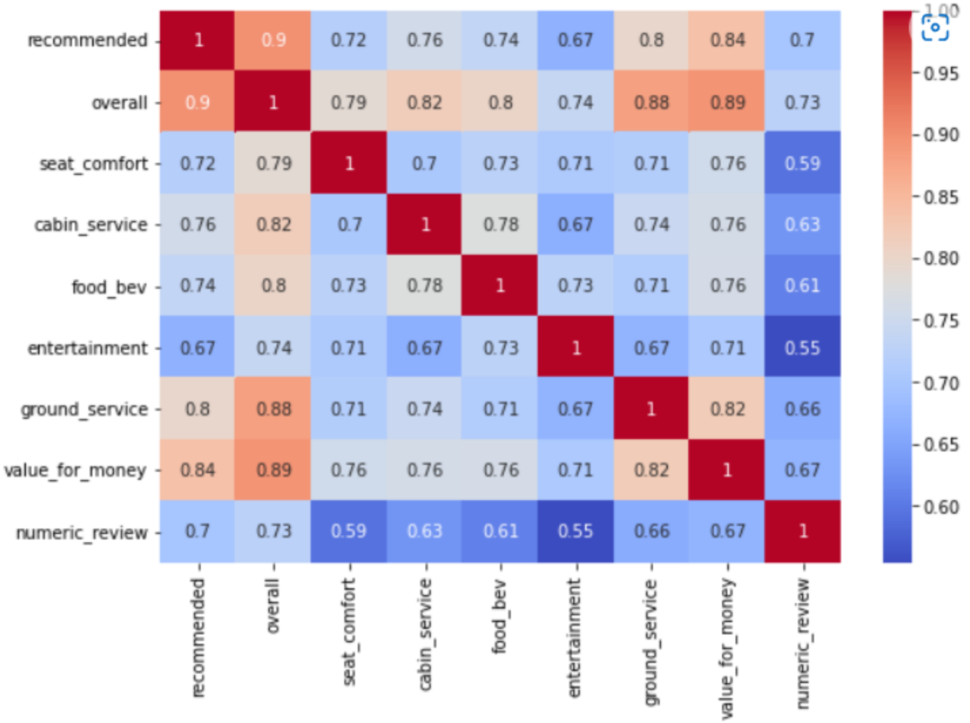
- We have also created new feature numeric review so to store sentiment score we have retrieved using sentiment analysis from customer review feature.

- After storing numeric value in range of -1 to 1 in column numeric review, we have dropped original review column



Correlation Plot

- Target Variable is highly positively correlated with Overall rating, Value for money and ground service
- These 3 features are most important factor for Recommendation therefore more focus should be given for improving those.
- Other ratings such as Cabin service, food and beverages, entertainment are also highly correlated with the recommendations

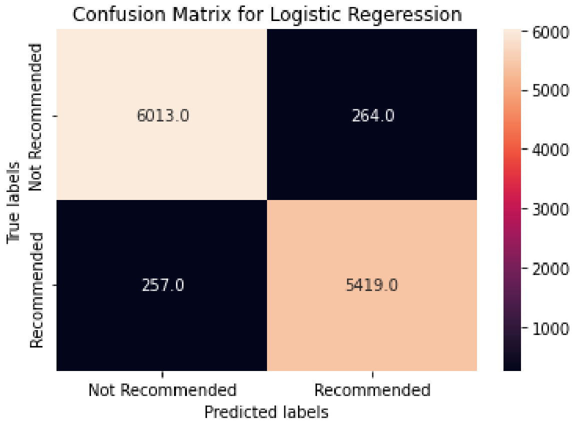


Feature Engineering

- Mapped recommend column for not recommended to 0, for recommended to 1
- Dropped author, aircraft , review_date, route ,travel_month
- Dummy columns for categorical features which are airline, cabin, traveler type

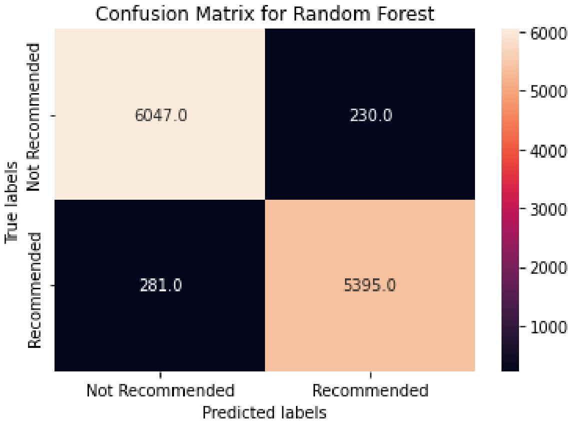
Model Building

Logistic Regression



Accuracy	0.9564
Precision	0.9547
Recall	0.9535
F1	0.9541
ROC AUC	0.9563

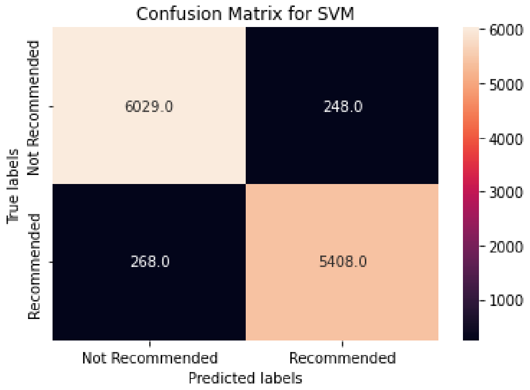
Random Forest



Accuracy	0.9572
Precision	0.9504
Recall	0.9591
F1	0.9547
ROC AUC	0.9569

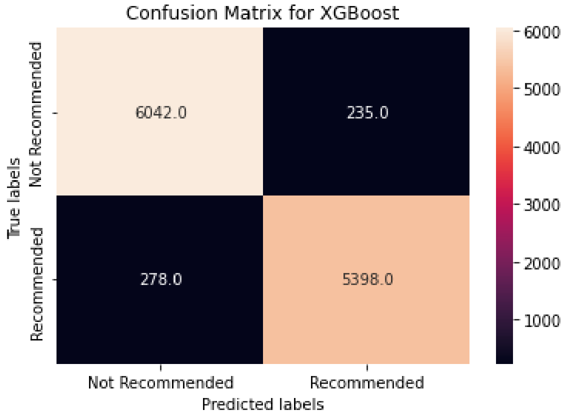
Model Building

Support Vector Classifier



Accuracy	0.9568
Precision	0.9527
Recall	0.9561
F1	0.9544
ROC AUC	0.9566

XGBoost

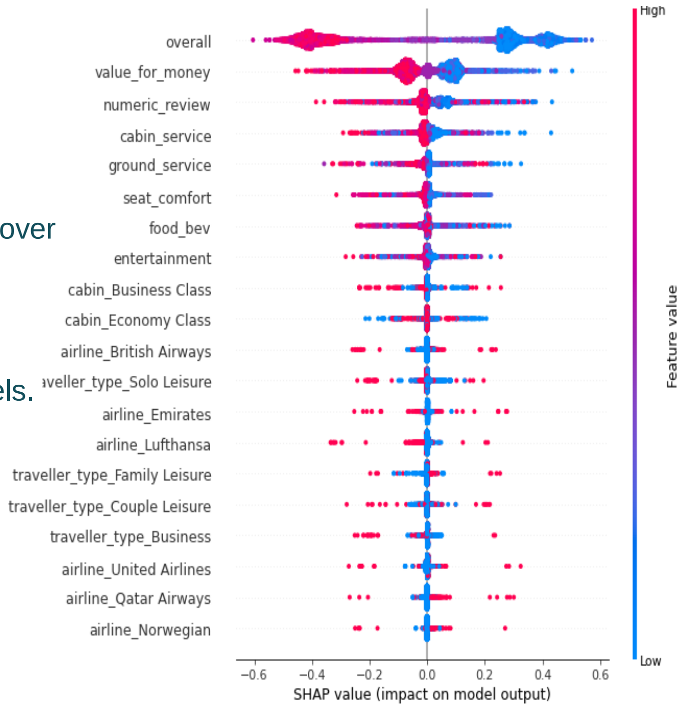


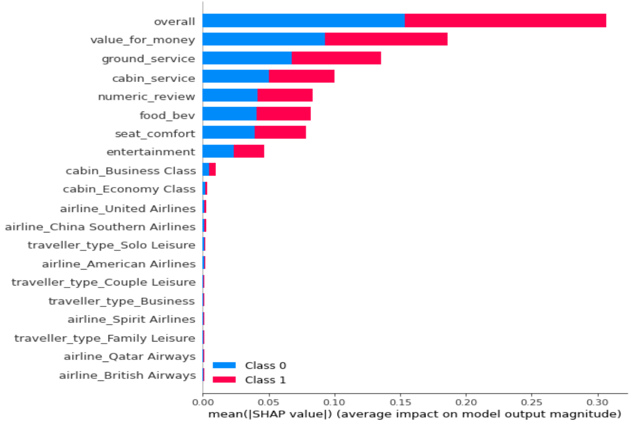
Accuracy	0.9570
Precision	0.9510
Recall	0.9582
F1	0.9546
ROC AUC	0.9567

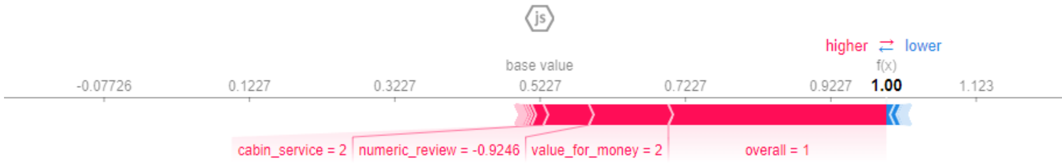
Model Explainability:

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 observed that people gave high positive recommendation to economic class in cabin. From this we can conclude that people are satisfied for services in economy class, this exactly opposite for business class, since customers from business class have least recommended airlines
- From month vs no. of recommendation. We can see that people tend to travel most in the month of June-August and December January this may be because of Christmas holidays.
- From overall rating vs recommended graph we can see which is perfectly understandable that non recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10 which is similar for all the other types of ratings

Conclusion

1. From the table with accuracy values sorted, we can see that XGBoost followed by Random Forest is giving highest accuracy.
2. Since our target variable had balanced classes, we have good recall as well as precision for all the models, still to verify we calculated F1 scores
3. Support Vector Model requires highest amount of time to build the model, which is almost 10 times more than the time required for XGBoost

Model	Accuracy	Recall	Precision	f1-score	roc_auc_score	Time
XGBoost	0.957082	0.951022	0.958282	0.954638	0.956792	10.375130
Random Forest	0.956998	0.950846	0.958274	0.954545	0.956704	22.405438
SVM	0.956831	0.952784	0.956153	0.954465	0.956637	98.100118
Logistic Regression	0.956413	0.954722	0.953546	0.954133	0.956332	12.586273

4. In Shap summary scatter plot we can see that shap value is high for overall, value for money, numeric_review, cabin service, ground_service which is increasing positive prediction and it is common for all models

Thank you