Airline_Passenger_Referral_Prediction

Ajinkya Shingote

Data science trainee,

Almabetter

Abstract:

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. I have tested the data and done some Exploratory data analysis to build machine learning models for the prediction of the dependent factor which is the recommendation of airlines by the passenger to his\her friend.

1. Introduction:

In today's world, a major part of the transportation is done by the airline's mode. Lots Air transport is one of the fastest modes of public transport which connects international boundaries. Air transport allows people from different countries to cross international boundaries and travel to other countries for personal, business, medical, and tourism purposes. Although, air transport provides the fastest means by saving the time of journey, another aspect of air transport is the facilities and comfort level of the passengers. Nowadays, there is a competitive environment among the airline industries. Every company is providing a variety of facilities to attract the passengers. The only motive is to improve their profit. Few years back, it was difficult to identify the needs and desires of passengers. But with the advancement of social media like Facebook, Twitter, etc., passengers are sharing their views on different types of airline facilities during their travel on social media platforms. This sharing of information plays a huge role to increase the competitiveness among the airline industries. It also provides a chance to improve its services and facilities for travellers worldwide.

2. Problem Statement

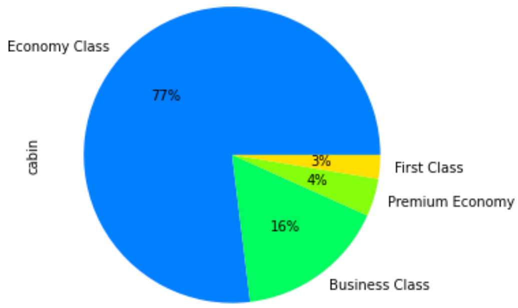
Data is scraped in Spring 2019 from Skytrax website. Data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions. The main objective is to predict whether passengers will refer the airline to their friends or not.

3.ExploratoryData Analysis

The Exploratory Data Analysis (EDA) plays a vital role in the analysis of the data variables which are important from the aspect of feature engineering. It will help me to distribute and relate between dependent and independent variables. I have gone through an analysis of every independent as well as the dependent variable to check which independent factor affects the dependent factor.

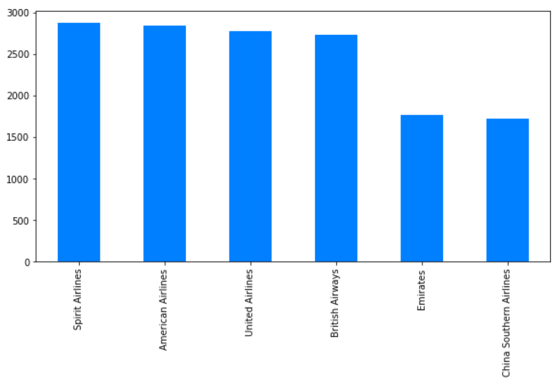
3.1 percentage of class of passenger in cabin

The pie plot on the down tells about the cabin in which most of the passengers travelled. It can be clearly derived from the plot that almost 77% of total passengers were “Economy Class” traveller and only 3% of total passengers were “First Class” traveller.



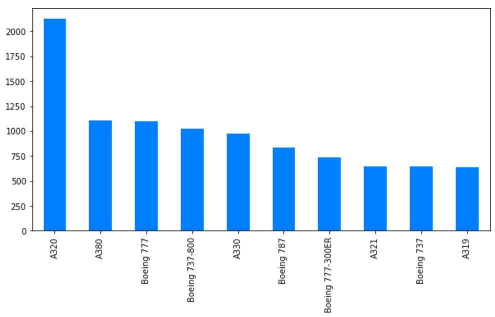
3.2 top 6 most frequently used airlines

These are the top 5 most frequently used airlines in the given data.



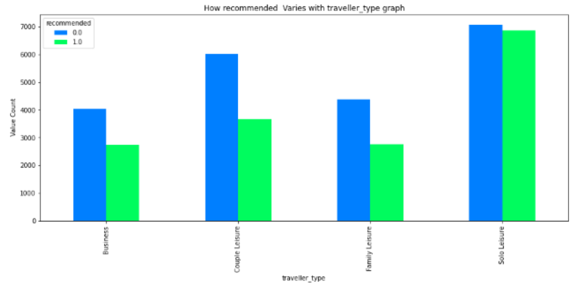
3.3 Top 10 aircrafts used

The below plot shows the top 10 aircrafts used in the given data. The aircraft A320 is most used aircraft in given data.



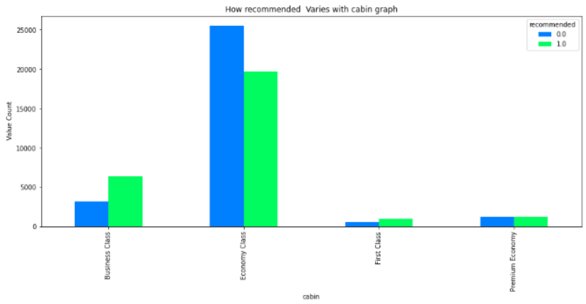
3.4 Types of customer categories and their opinion

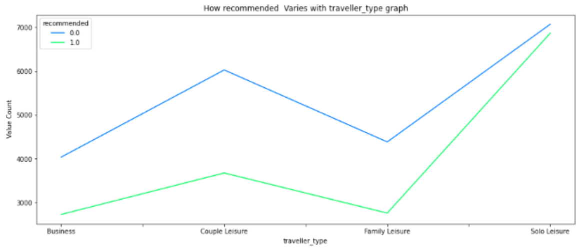
We can see that people have given both 1 or 0 which we will consider from now on as positive and negative recommendation so to interpret it effectively to the solo leisure. This may because of the poor infrastructure or the service received by the people and positive recommendation may be because of 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 leisure's.



3.5 Cabin wise analysis

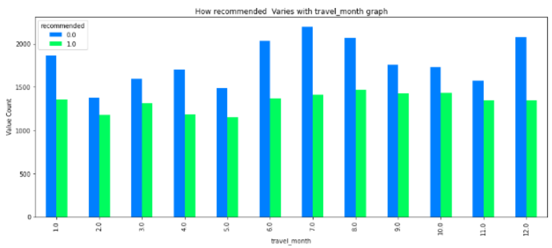
Also, we can see that people gives the high positive recommendation to economic class in cabin. From this we can conclude that people love to travel in economic class as of low price also in same way we can see people give highest negative recommendation to economy class maybe because less infrastructure 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. 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.

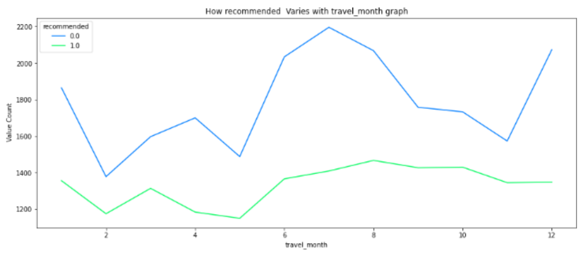




3.6 travel month analysis

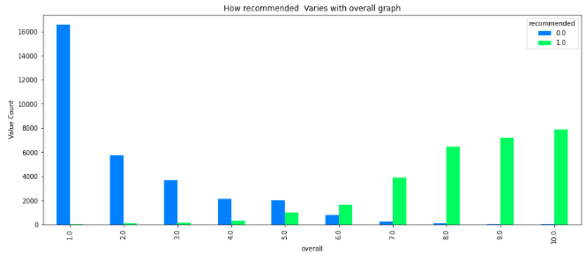
From month vs no. of recommendation. We can see that people tents to travel most in the month of July considering the total of positive and negative recommendation combined. In month we cannot see any preferable trend but here we can conclude people tent to travel highest during the month of July.

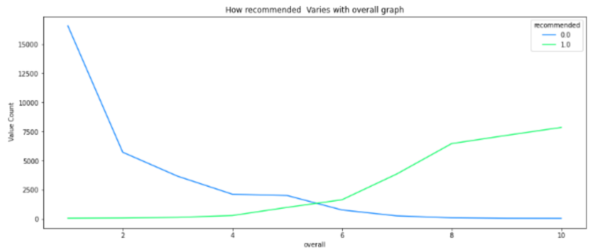




3.7 overall rating

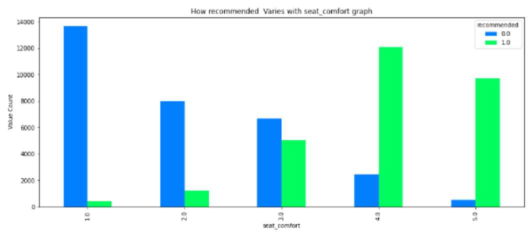
From overall vs recommended graph we can see which is perfectly understandable that negative recommendation has been given to the overall rating of 1.0 and high positive recommendation has been given to the overall rating of 10. But it is very true that highest negative recommendation has been given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

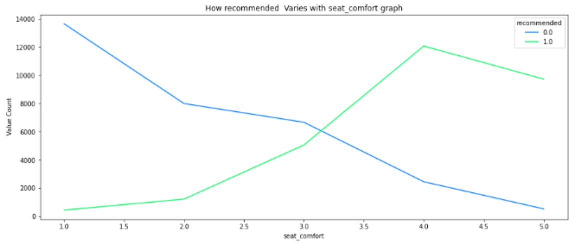




3.8 Seat comfort Analysis

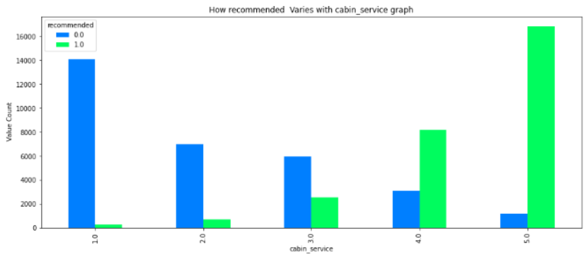
In seat comfort people has given highest positive recommended to the seat of class 5 as compared to very low negative recommendation to the same. Also we can see seat of class 1 have been given highest negative recommendation as compare to its positive recommendation. Here we come to a conclusion it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

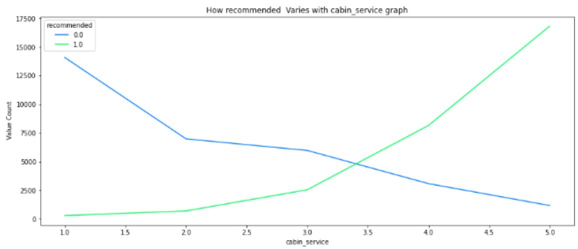




3.9 Cabin service Analysis

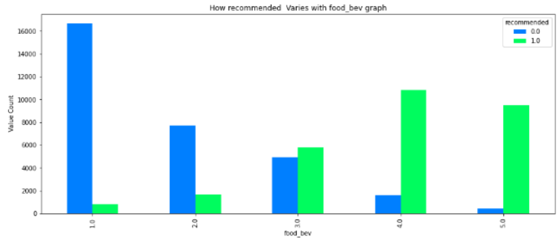
In cabin service rating people has given highest recommendation to rating to cabin service rating 5 as compare to its counterpart. From this we can conclude that cabin service is doing pretty good.In cabin service we can see same as the positive recommendation increases with the overall rating and also 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.

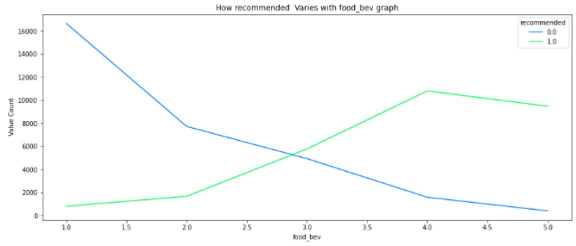




3.10 Food beverages

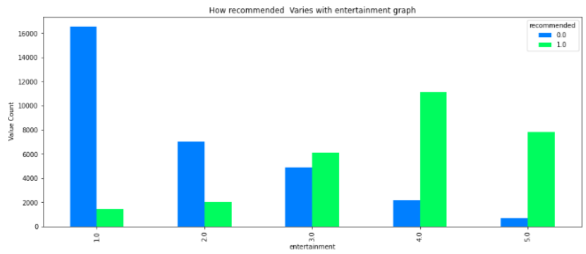
In food and beverage rating people have given highest negative recommendation to rating 1.0 from this we can conclude that airline service has to improve their food delivery and quality service. In food service we can see same as the positive recommendation increases with the overall rating and also 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.

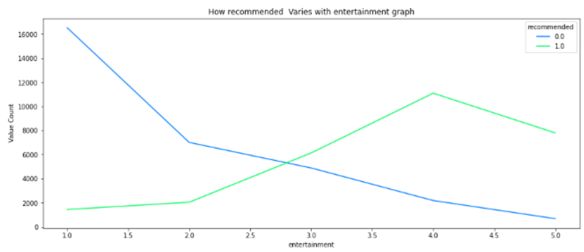




3.11 Entertainment Analysis

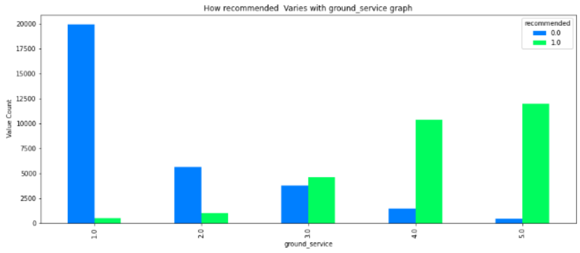
In entertainment also we can see most people has given highest negative recommendation to entertainment rating 1 which shows that airline has to improve their entertainment system as well. In Entertainment service too we can see same as the positive recommendation increases with the overall rating and 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.

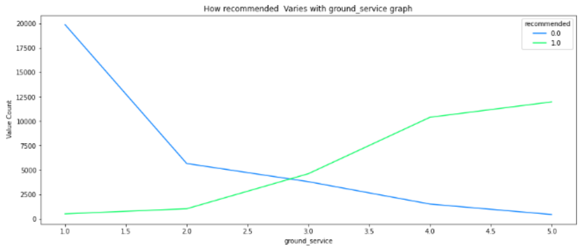




3.12 Ground service Analysis

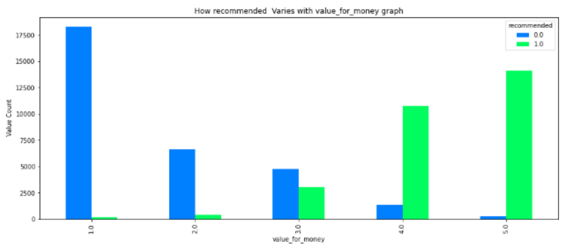
In Ground service also we can see same as the positive recommendation increases with the overall rating and also negative recommendation on the same decreases also, we can an intersection in Ground service rating close 3.0 where we can see similar positive and negative recommendation. In ground service also we can see most people has given highest negative recommendation to ground service rating 1 which shows that airline must improve their ground service.

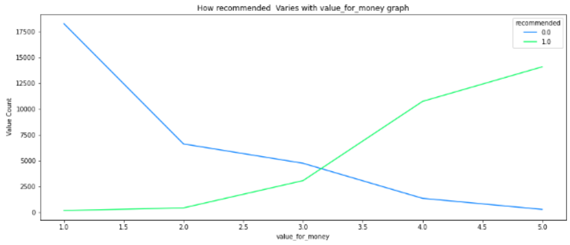




3.13 Value for money

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. Lastly in Value for money rating 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 Value for money rating greater than 3.0 where we can see similar positive and negative recommendation.





4. Correlation Analysis

The correlation analysis has been done to get a better understanding of dependent and independent variables' multicollinearity. Multicollinearity may not affect the accuracy of the model as much but we might lose reliability in determining the effects of individual independent features on the dependent feature in your model and that can be a problem when we want to interpret your model.

4.1 Heatmap

Let's check the heatmap plotted concerning independent variables.

	recommended	travel_month	overall	seat_comfort	cabin_service	food_bev	entertainment	ground_service	value_for_money
recommended	1.000000	-0.004002	0.898390	0.719521	0.756843	0.736565	0.668660	0.797478	0.837220
travel_month	-0.004002	1.000000	-0.004173	0.000088	-0.005573	-0.002793	-0.015751	-0.004096	-0.007617
overall	0.898390	-0.004173	1.000000	0.791971	0.820029	0.803981	0.740649	0.881449	0.896356
seat_comfort	0.719521	0.000088	0.791971	1.000000	0.708728	0.725471	0.708497	0.719685	0.758560
cabin_service	0.756843	-0.005573	0.820029	0.708728	1.000000	0.776758	0.666898	0.747785	0.764541
food_bev	0.736565	-0.002793	0.803981	0.725471	0.776758	1.000000	0.729318	0.716689	0.763086
entertainment	0.668660	-0.015751	0.740649	0.708497	0.666898	0.729318	1.000000	0.671103	0.706957
ground_service	0.797478	-0.004096	0.881449	0.719685	0.747785	0.716689	0.671103	1.000000	0.822223
value_for_money	0.837220	-0.007617	0.896356	0.758560	0.764541	0.763086	0.706957	0.822223	1.000000

Overall and Recommended are highly correlated, Overall and Value for money are highly correlated

5 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

5. Feature Engineering

The given information in its crude structure was not straightforwardly utilized as a contribution to the model. A few components designing was completed where barely any elements were changed, few were dropped, and few were added. The following is a rundown of the element designing completed with the gave informational index

I have Engineered new features based on the existing features which are date of travel, review text, overall rating etc.

I have done imputation of missing values in the target variable, I also did imputation of missing values in the independent variable. I handled categorical variables and date columns. I used NLP for handling the review text feature.

I also did onehot encoding on the categorical features like airline, cabin, traveller_type.

NLP(Natural Language Processing) for reviews

I have used NLP to extract sentiment from customer_review features by using Sentiment analysis and added numeric_review feature so to store the sentiments we have extracted.

6. WORKING WITH DIFFERENT MODELS

6.1 Train/Test Split

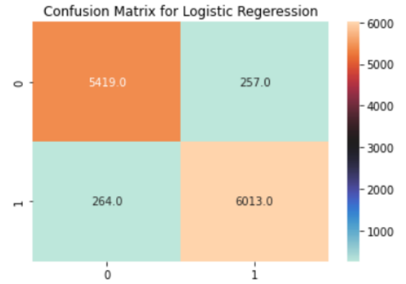
The train/test split was done as 80/20 % of data with a random state of 0. The final dataset was of shape (61183, 17) which was split to (48946, 17) as Train data and (12237,17) as Test data.

6.2 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. The likelihood that each input belongs to a specific category is modelled using logistic regression. Logistic regression is a fantastic tool to have in your toolbox for classification purposes. For classification situations, where the output value we want to predict only takes on a small number of discrete values, logistic regression is an important technique to know. The logistic function offers a number of appealing characteristics. The probability is represented by the y-value, which is always confined between 0 and 1, which is exactly what we wanted for probabilities. A 0.5 probability is obtained for an x value of 0. A higher likelihood is also associated with a higher positive x value, while a lower probability is associated with a greater negative x value. In logistic regression to learn the coefficients of features in order to maximize the probability of correctly classifying the classes. For this maximum likelihood concept is used.

	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%



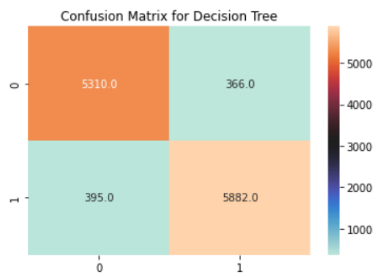
6.3 Decision Tree

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. The decision criteria for classification and regression trees are different. To decide whether to break a node into two or more sub-nodes, decision trees employ a variety of techniques. The homogeneity of the generated sub-nodes improves with the generation of sub-nodes. To put it another way, the purity of the node improves as the target variable grows. The decision tree separates the nodes into sub-nodes based on all available variables, then chooses the split that produces the most homogenous sub-nodes.

	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%

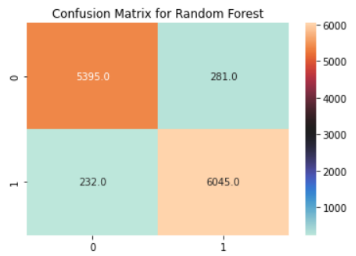


6.4 Random Forest

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. Each tree assigns a categorization to a new object based on attributes, and we say that the tree "votes" for that class. The classification with the highest votes is chosen by the forest.

	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%

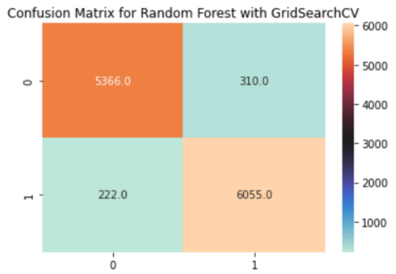


6.5 Random Forest with GridSearchCV

The best parameters for this grid search is max_depth 12, min_sample_leaf 50, min_sample_split 100 and n_estimators as 80. So that the model accuracy we obtained is 95.5%.

	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%

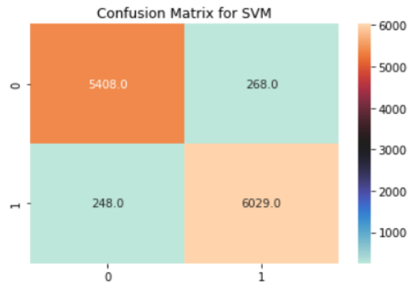


6.6 Support Vector Machine

SVM(Support Vector Machine) take a direct approach to binary classification by attempting to find a hyperplane in a feature space that "best" separates the two classes. In practise, however, finding a hyperplane that completely separates the classes using only the original features is challenging. SVMs get around this by expanding the idea of separating hyperplanes in two different ways. (1)Expand the feature space to the point where perfect separation of classes is (more) likely, and(2) apply the so-called kernel trick to extend the feature space.

	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%

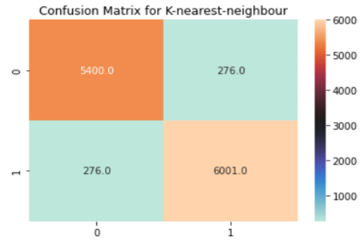


6.7 K_nearest Neighbour Model

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 classifies a data point based on how its neighbours are classified.

	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%

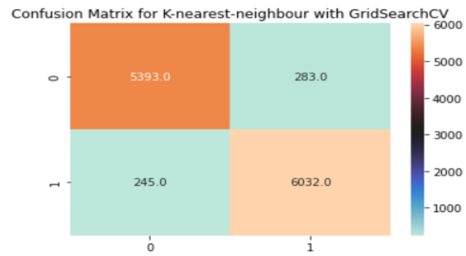


6.8 K_nearest Neighbour Model with GridSearchCV

The best n_neighbors value we get is 42. The accuracy for K_nearest Neighbour model is 95.38% and by applying GridSearchCV on we got the accuracy with 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 95.58%



6.9 XGBoost Model

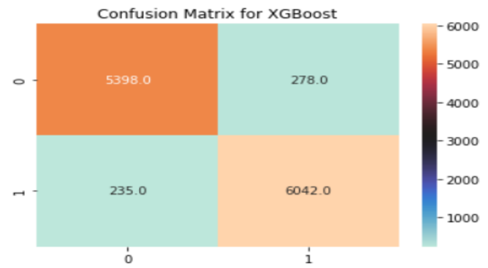
XGBoost is a distributed gradient boosting library that has been optimised for performance, flexibility, and portability. It uses the Gradient Boosting paradigm to implement machine learning algorithms. XGBoost is a parallel tree boosting (also known as GBDT, GBM) algorithm that solves a variety of data science problems quickly and accurately.

```
              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%
```



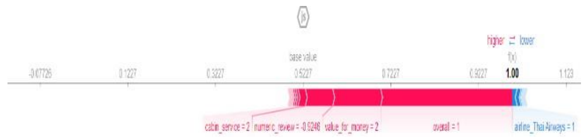
6.10 Different Model Comparison according to their Evaluation metrices

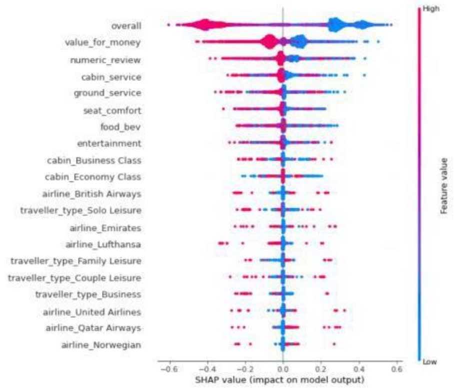
	Model	Accuracy	Recall	Precision	f1-score	roc_auc_score
0	Logistic Regression	0.956413	0.954722	0.953546	0.954133	0.956332
1	Decision Tree	0.936334	0.935518	0.930762	0.933134	0.936295
2	Random Forest	0.957082	0.950493	0.958770	0.954614	0.956766
3	Random Forest with GridSearchCV	0.955492	0.945384	0.960272	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

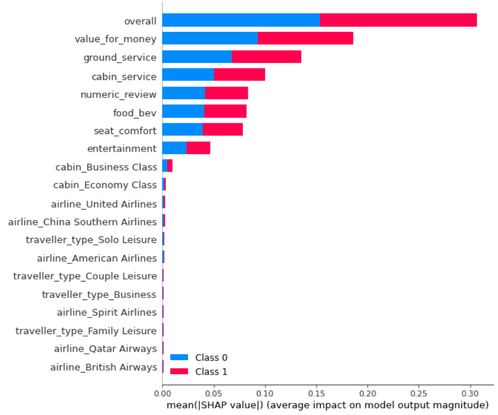
7. Model explainability :

SHAP(Shapley Additive explanations)

1. In Shap JS summary we can see positive features overall, value for money,numeric_review combined red colour block pushes the prediction toward right over base value and causing positive model prediction and it is common for all model.
2. 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.







8. Conclusion

1. 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.
2. 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 similarity negative recommendation because of high price of business class or less travelling percentage.

3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. In Shap JS summary we can see positive features overall, value for money, numeric_review combined red colour block pushes the prediction toward right over base value and causing positive model prediction and it is common for all model.
13. 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.