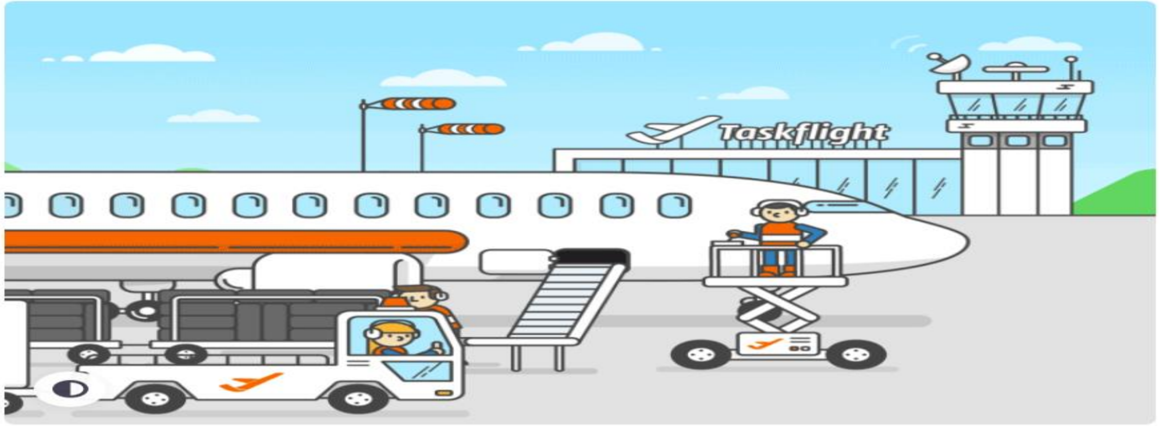
Capstone Project-III

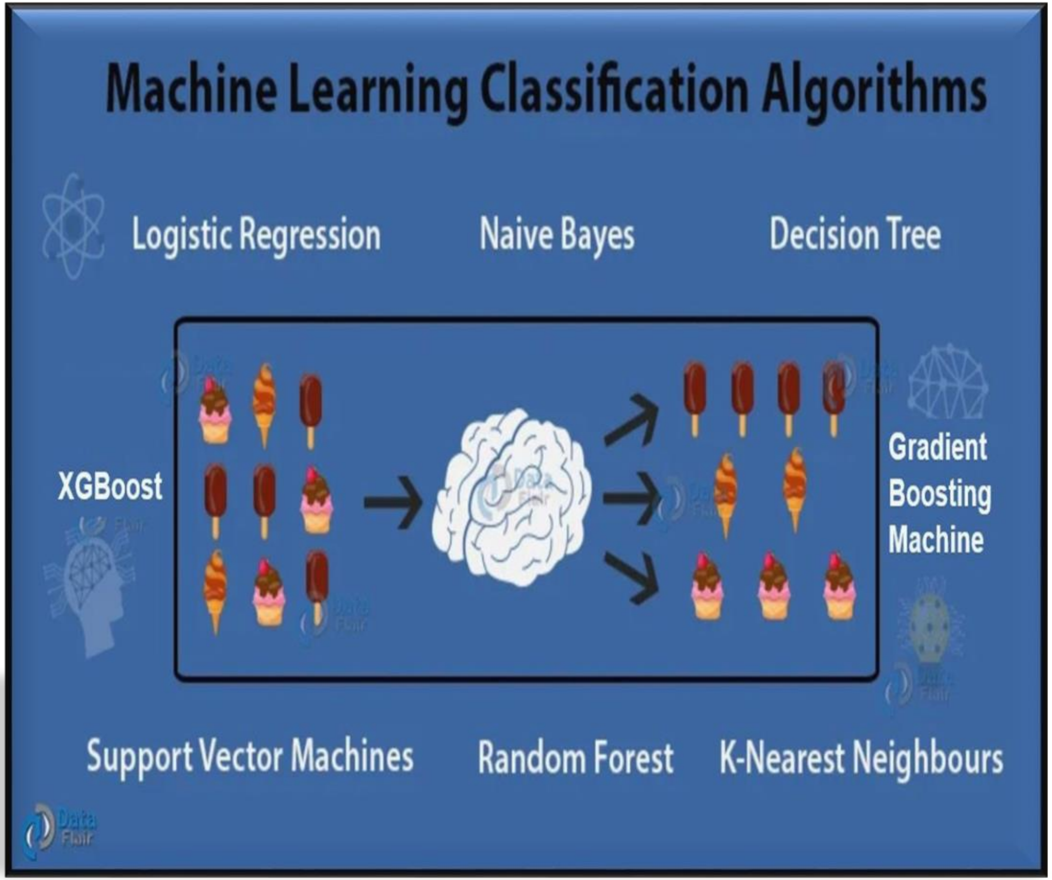
Classification-Airline Passenger Referral Prediction



By- Mahima Phalkey
Sameer Satpute

CONTENTS:

1. Agenda
2. Problem Statement
3. Data Summary
4. Basic Observations
5. EDA
6. One Hot Encoding
7. Data Modelling
8. Model Explainability
9. Conclusion
10. Challenges
11. References



AGENDA

Our main objective is to predict how many passengers will refer the flights they travel by using classification algorithm.

We will also see what are the factors which are affecting the passenger to not recommend the flight.

We will also explain our models by using SHAP and ELI5.

PROBLEM STATEMENT

- Data includes airline reviews from 2006 to 2019 for popular airlines around the world with multiple choice and free text questions.
- Firstly we do EDA to know insights from a business perspective.
- Data were scrapped in Spring 2019.
- The main objective is to predict whether passengers will refer the airline to their friends and others.
- Find out the best model which gives realistic results.

INTRODUCTION

- A century after the first commercial flight, the aviation industry continues to offer a variety of exciting and rewarding career options for qualified professionals.
- “Aviation” is a growing industry with very practical purposes. Worldwide, airlines carry more than 3 billion passengers a year and deliver about one-third of traded goods by value. Aviation sector employment also is seen as strong.
- Airlines employ about 2.5 million workers and expect “to accelerate the pace of hiring over the next year”.
- With the progress in aviation techniques, airlines have paved a way for making travel and tourism better in every way. ‘
- Hence, it plays a major role in the travel and tourism.

DATA SUMMARY

1. Airline: Name of the airline.

2. Overall: Overall point is given to the trip between 1 to 10.

3. Author: Author of the trip

4. Review date: Date of the Review customer review: Review of the customers in free text format

5. Aircraft: Type of the aircraft

6. Traveler type: Type of traveler (e.g. business, leisure)

DATA SUMMARY

7. Cabin: Cabin at the flight date flown: Flight date

8. Seat comfort: Rated between 1-5.

9. Cabin Service: Rated between 1-5.

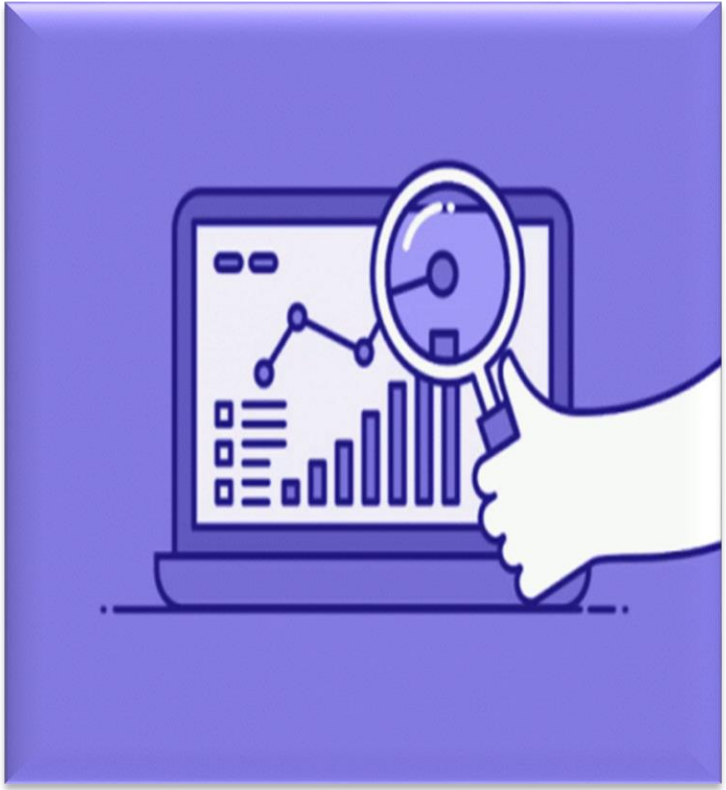
10. Food Bev: Rated between 1-5 Entertainment: Rated between 1-5

11. Ground service: Rated between 1-5

12. Value for money: Rated between 1-5

13. Recommended: Binary, target variable

OBSERVATIONS



- Our dataset has a shape of 131895 rows and 17 columns.
- There are a lot of null values.
- We see that more than 50% of the dataset are not having values.
- We have to drop the aircraft column as it is having nearly 80% of null values which means that column will be of no use in our prediction
- Here we can see that the mean values and the 50 % values are nearly equal which means the variable is normally distributed.
- There were 85121 duplicated values. Removing the duplicated values and keeping only the first values.

EDA (DATA CLEANING)

1) NULL VALUES TREATMENT

- We see that more than 50% of the dataset are not having values.
- We have to drop the aircraft column as it is having nearly 80% of null values.
- For the numerical values we have used KNN Imputer to impute data into null values.
- There are null values in categorical variables and DV we have to drop those rows because even if we use mode to fill the null values It will result in wrong predictions so it's better to drop those.

2) CHECKING DUPLICATE

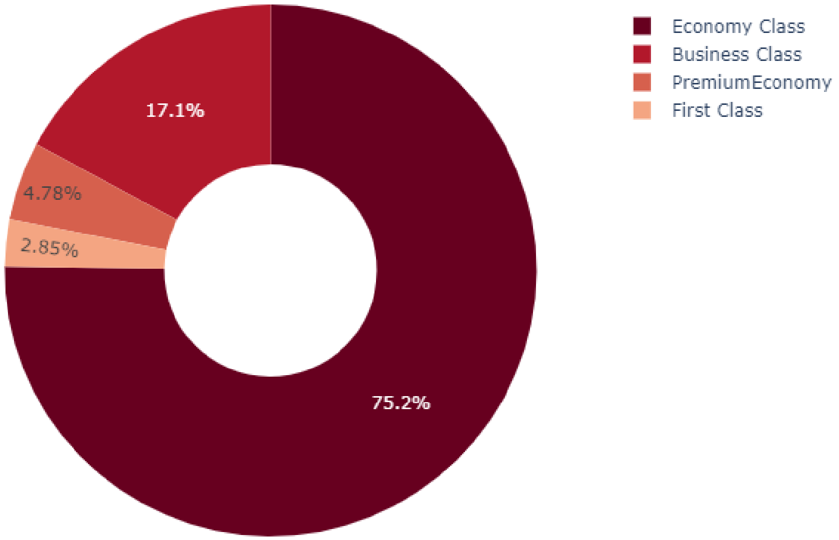
- There were 85121 duplicated values.
- Removing the duplicated values and keeping only first values.

3) OUTLIER DETECTION

- We can see that there are no outliers present in our data.

EDA(DATA VISUALIZATIONS)

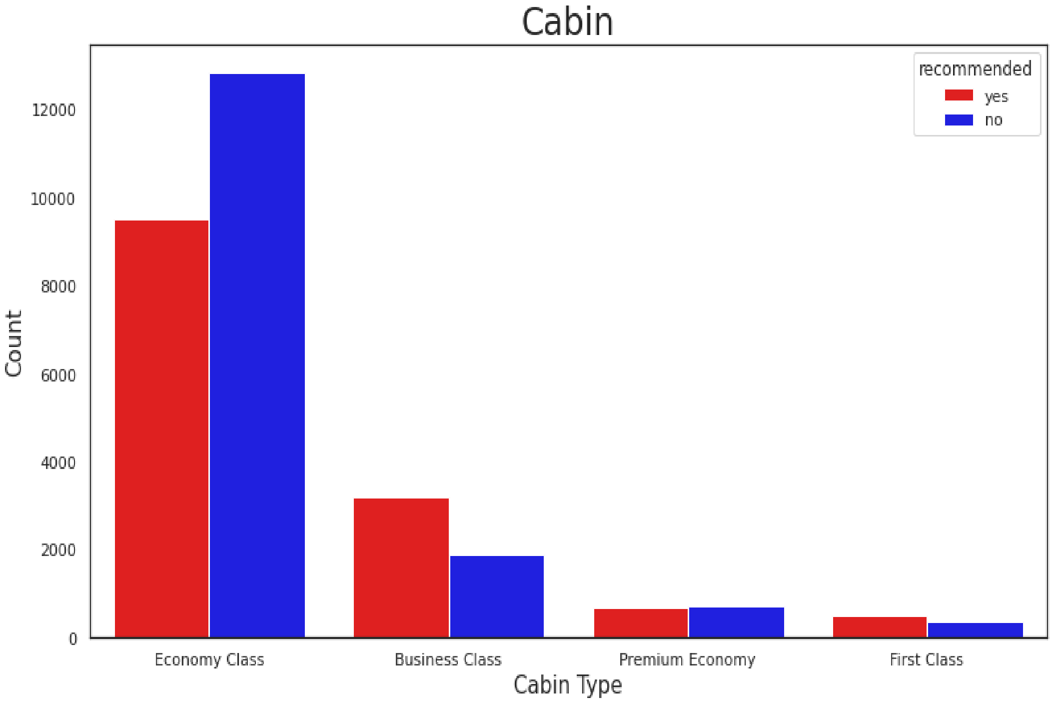
PIE CHART FOR UNIQUE CABIN



From the graph we can clearly see that nearly 76 % of flyers are from Economy Class cabin followed by Business class that is 17 % .

EDA(DATA VISUALIZATIONS)

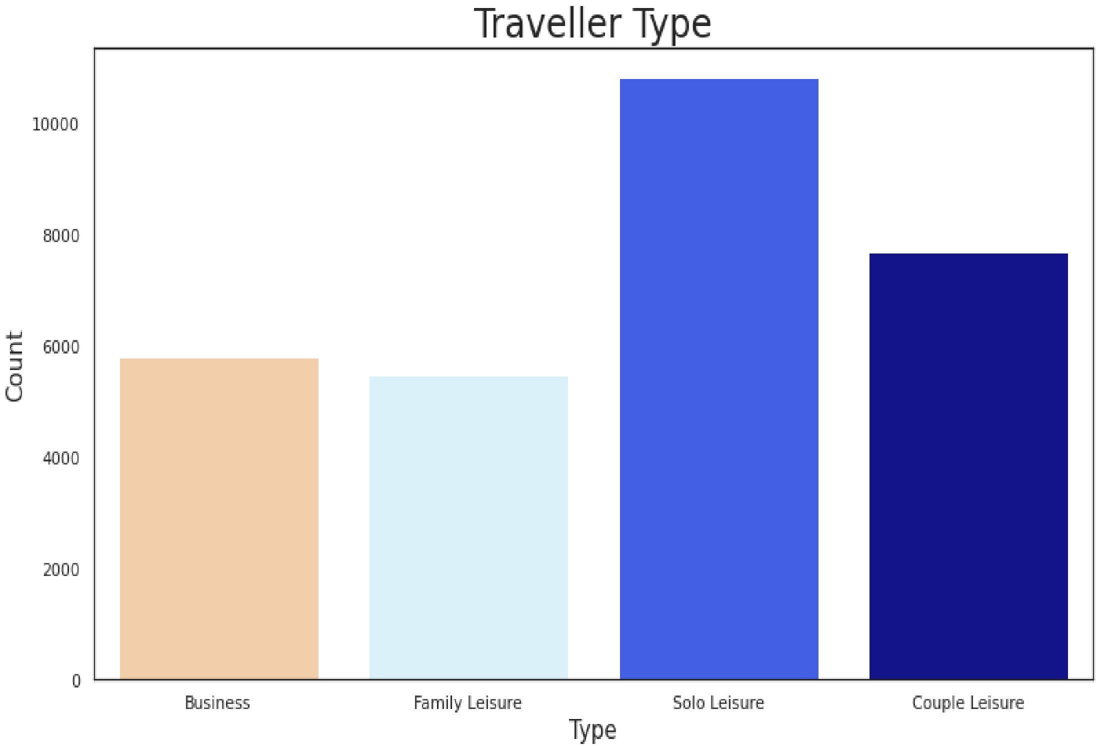
COUNTPLOT FOR CABIN WRT RECOMMENDED



- So, the economy class has the most recommendation whereas the first class has the least recommendation.
- In economy class we can see No is more than yes.

EDA(DATA VISUALIZATIONS)

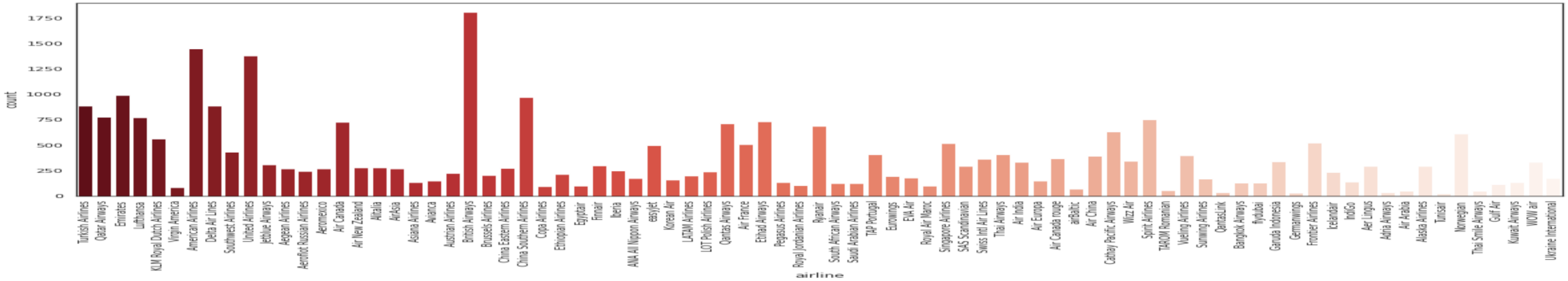
COUNTPLOT FOR TRAVELLER_TYPE WITH MOST RATINGS

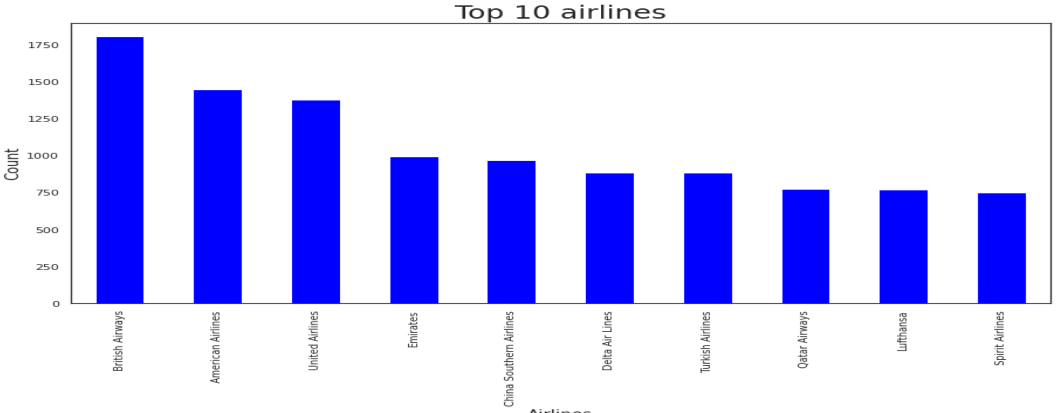


- It's clear from the count plot that 'Solo Leisure' has the highest ratings among all whereas 'Family Leisure' has the least ratings.

EDA(DATA VISUALIZATIONS)

BAR GRAPH TO SEE THE TOP 10 AIRLINES AND COUNTPLOT OF AIRLINES

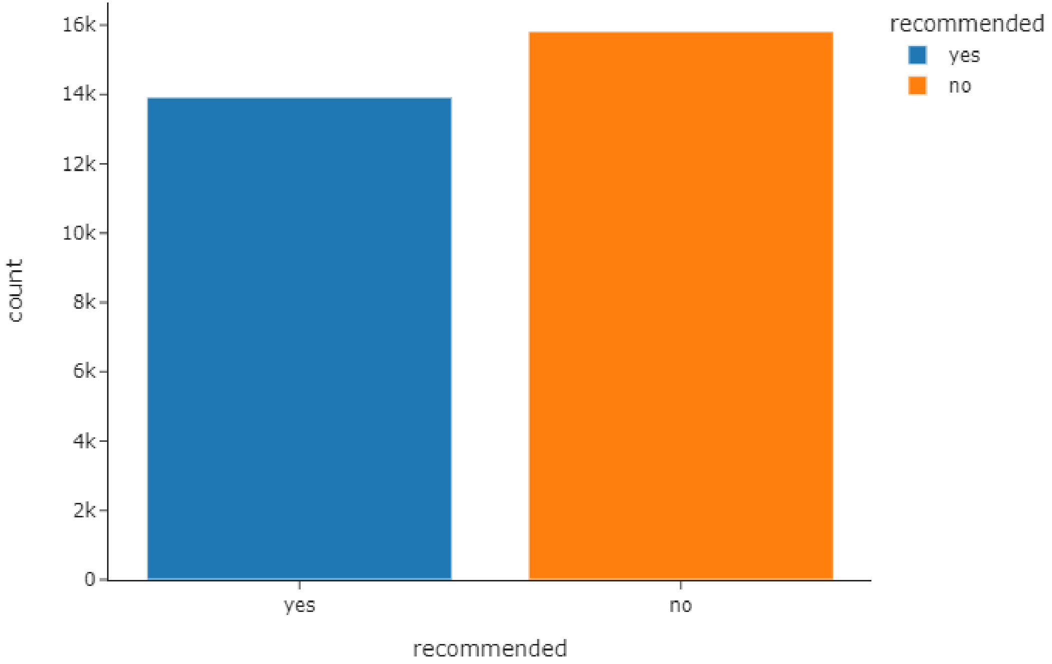




- 'British airways' has the maximum number of trips and this can be attributed to its ultra-low-cost fare compared to other airlines.
- 'Tunisair', 'Germanwings' etc. are the lowest number of trips.

EDA(DATA VISUALIZATIONS)

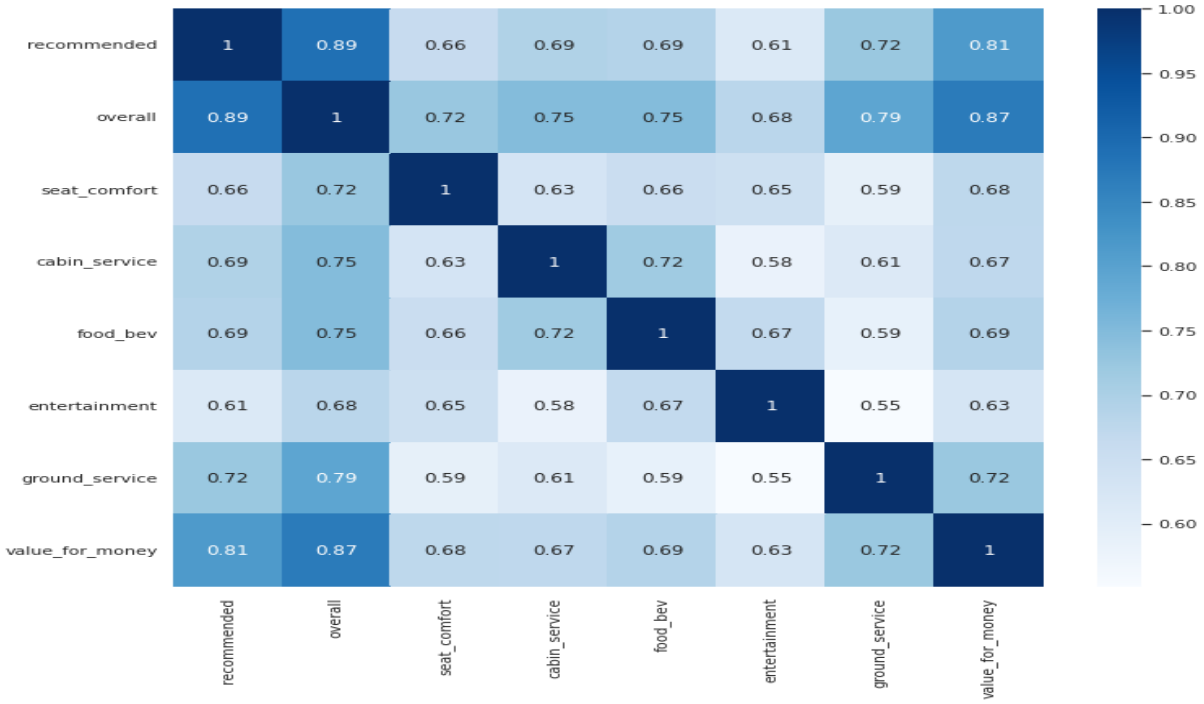
HISTOGRAM FOR RECOMMENDED



- Clearly, 'No' responses are more as compared to 'Yes' responses
- But It seems nearly balanced target variable.

EDA(DATA VISUALIZATIONS)

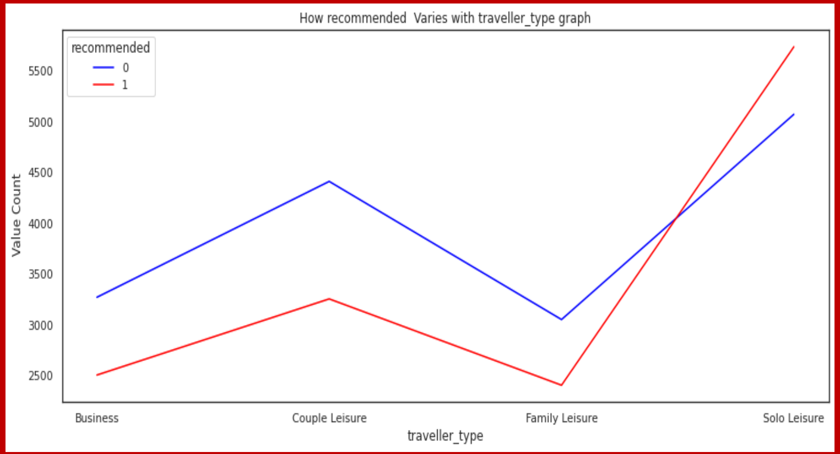
HEATMAP TO SEE CORRELATION

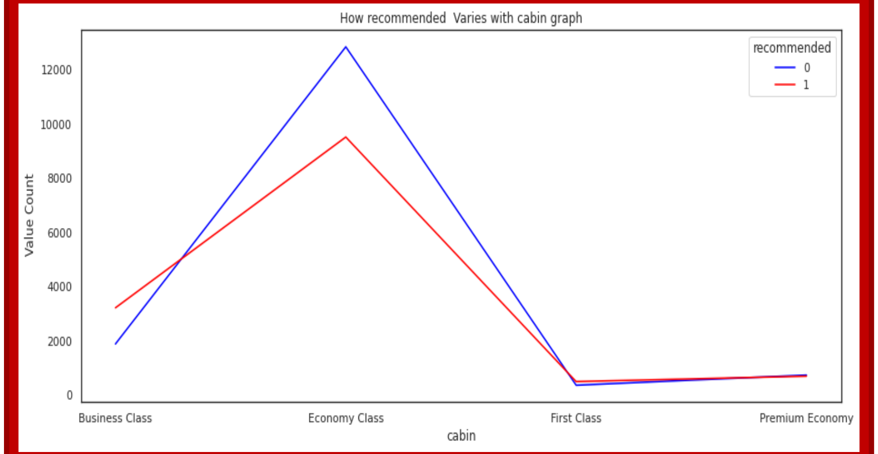


We can see there are some highly correlated values like value_for_money, overall, etc.

EDA(DATA VISUALIZATIONS)

PLOT FOR THE FEATURES WRT TO RECOMMENDED



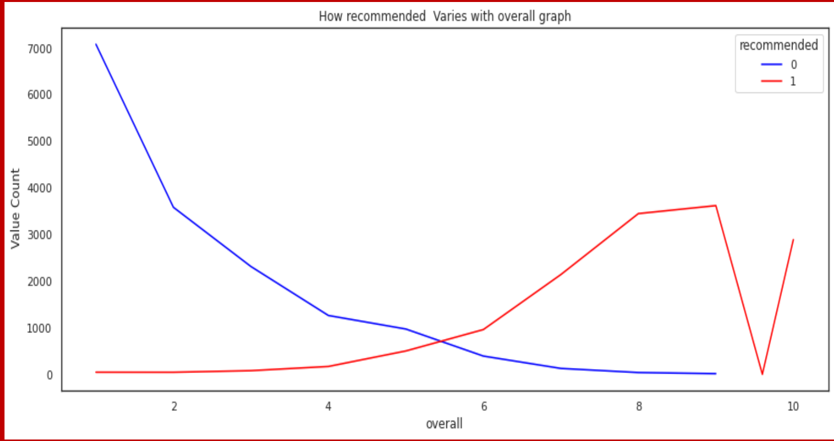


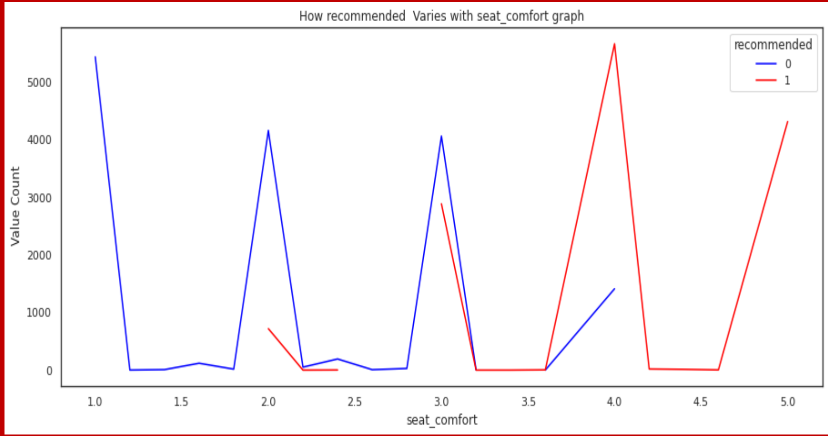
- We can see, in both the business and leisure traveler types, that both the recommendation trend in terms of yes or no increases from business to couple leisure and it decreases to family and again reaches a high level in solo leisure.
- This indicates people prefer solo leisure higher than any of the other leisure.

- With regards to cabin type, it has been determined that both yes and no recommendation trends increase from business class to economy class, then decrease to first class, and again increase slightly in the premium class.
- This indicates most people travel in economy class.

EDA(DATA VISUALIZATIONS)

PLOT FOR THE FEATURES WRT TO RECOMMENDED



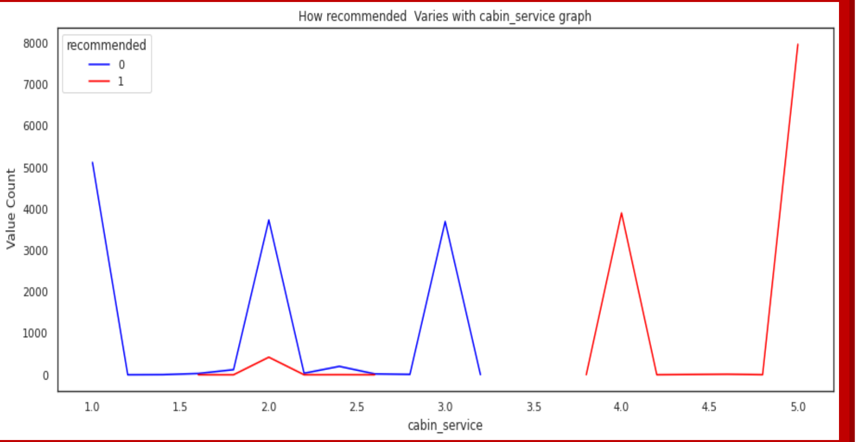


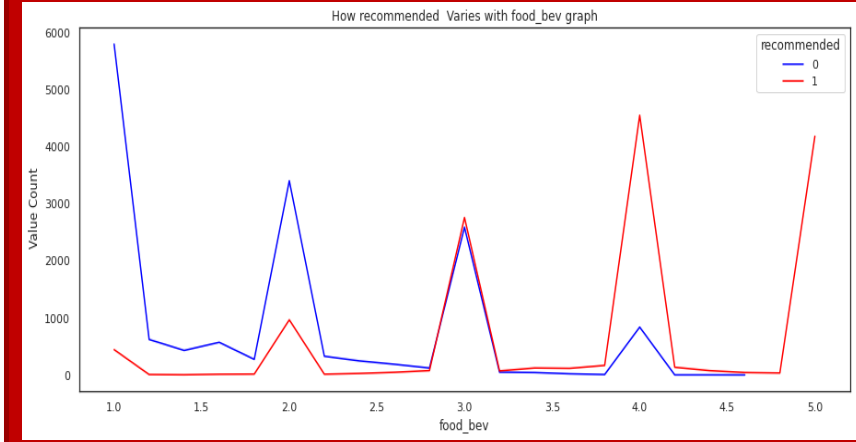
- Generally, we can observe a very good insight which is also regular in the overall rating.
- We can see that positive recommendation increase with the overall rating, while negative recommendations decrease.

- In seat comfort we can see the negative recommendation is there till 4.0 rating but after that, we can see positive recommendation also.

EDA(DATA VISUALIZATIONS)

PLOT FOR THE FEATURES WRT TO RECOMMENDED



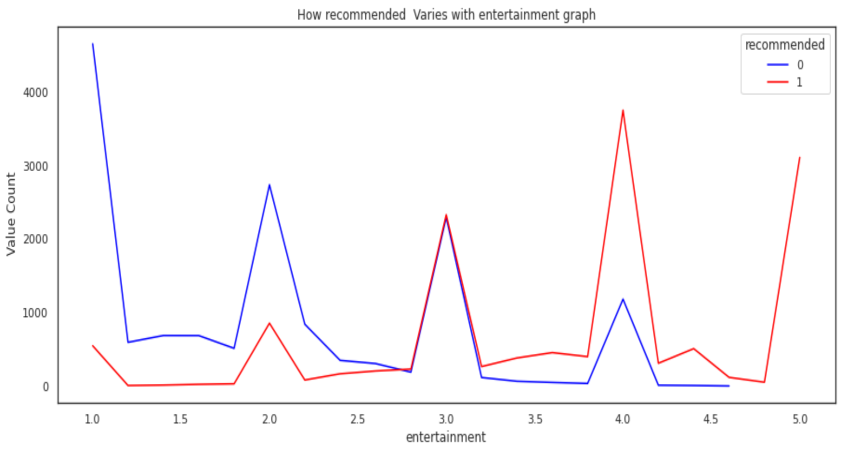


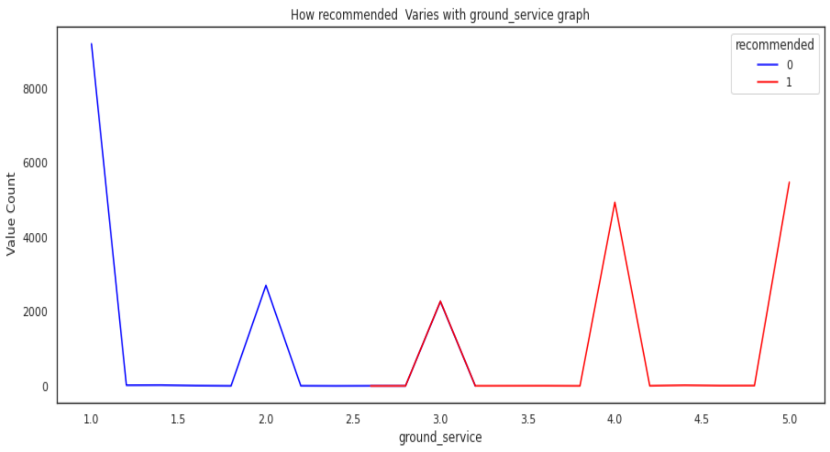
- In cabin service also we can see the similar trend as seat comfort negative recommendation is there till 3.0 rating but after that we can see positive recommendation also.

- In food bev we can see mixed recommendations initially as the negative recommendation decreases positive recommendations are increasing.

EDA(DATA VISUALIZATIONS)

PLOT FOR THE FEATURES WRT TO RECOMMENDED



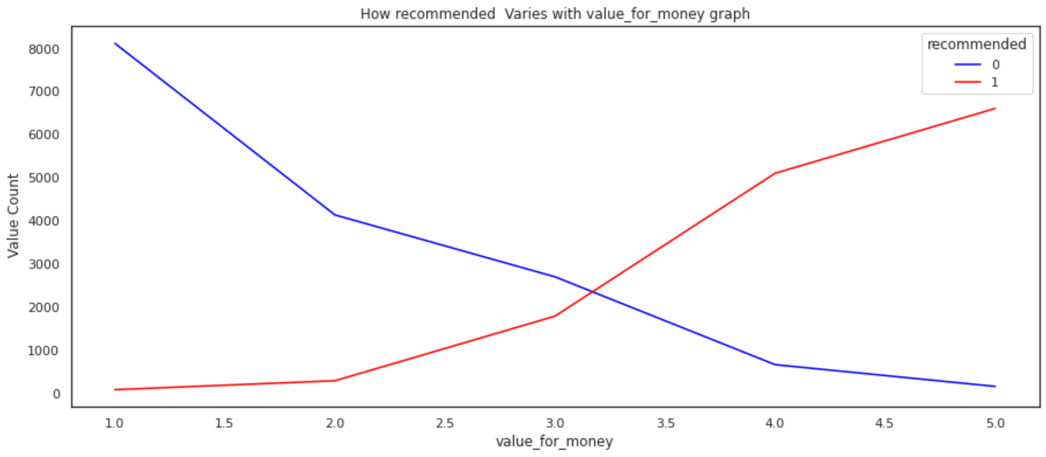


- In entertainment we can see mixed recommendations initially as the negative recommendation decreases positive recommendations are increasing.

- In ground service we can see negative recommendations only at first till 2.5 after that positive recommendations took over

EDA(DATA VISUALIZATIONS)

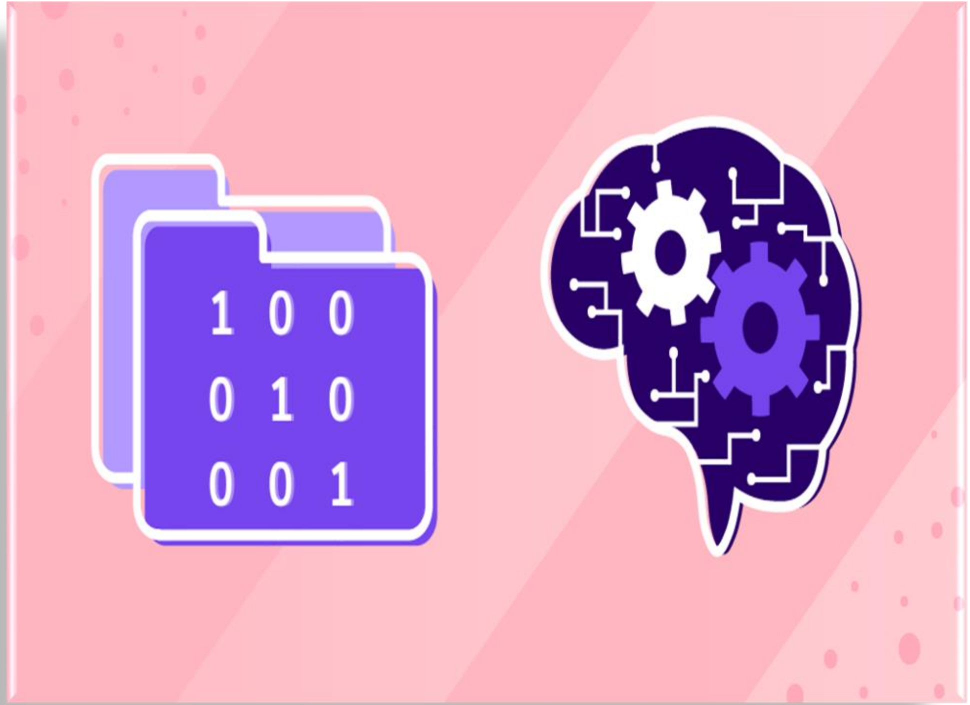
PLOT FOR THE FEATURES WRT TO RECOMMENDED



- Lastly in Value for money rating we can see the 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.

EDA(DATA PREPROCESSING)

ONE HOT ENCODING



- In this technique, the categorical parameters will prepare separate columns for both Male and Female labels.
- So, wherever there is a Male, the value will be 1 in the Male column and 0 in the Female column, and vice-versa.
- We did one hot encoding on traveller type and on the cabin.
- In traveller type columns are made for Solo, Couple, Family leisure, and Business and in Cabin columns are made for Business class, Economy class, Premium class, First class.

MODEL PREPARATION

Splitting data

X = Independent variable

Y = Dependent variable

We have split train-test data with 80-20 data.

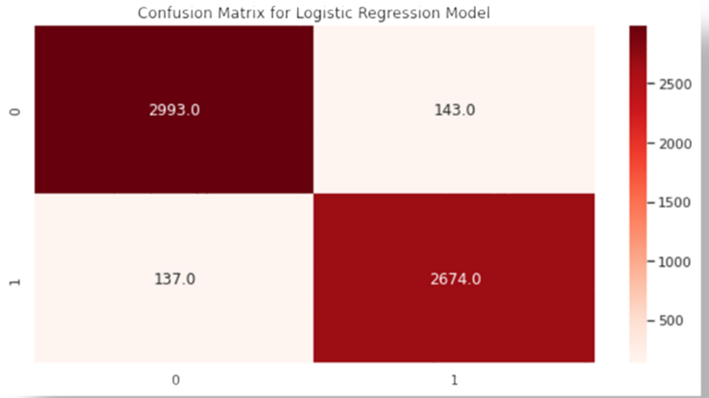
We can see the classes for train and tests are properly scaled. So we do not need to perform under-sampling or oversampling as it is already properly scaled. Thus, all the data features tend to have a similar impact on the modeling portion.

```
Distribution of classes of dependent variable in train :
0    12681
1    11103
Name: recommended, dtype: int64

 Distribution of classes of dependent variable in test :
0    3136
1    2811
Name: recommended, dtype: int64
```

DATA MODELLING

LOGISTIC REGRESSION

- Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.
- Since the outcome is a probability, the dependent variable is bounded between 0 and 1.
- Logistic regression is a robust supervised ML algorithm for binary classification problems (when the target is categorical).
- In Logistic Regression the accuracy is 95.29 % and recall is 95.12%



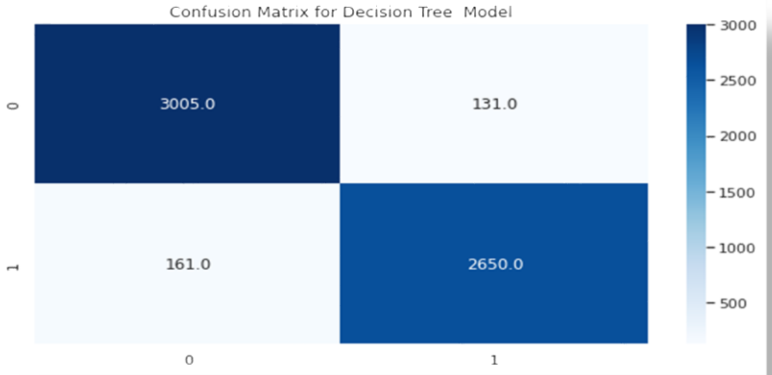
	precision	recall	f1-score	support
0	0.96	0.95	0.96	3136
1	0.95	0.95	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

Accuracy of the Model: 95.29174373633765%

DATA MODELLING

DECISION TREE

- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression.
- The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- A tree can be seen as a piecewise constant approximation.
- The accuracy for the Decision tree is 95.08% and recall is 94.27%.



	precision	recall	f1-score	support
0	0.95	0.96	0.95	3136
1	0.95	0.94	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

Accuracy of the Model: 95.08996132503783%

DATA MODELLING

ENSEMBLE OF DECISION TREE

- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

BAGGING:

- Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree.
- Here idea is to create several subsets of data from training sample chosen randomly with replacement.
- Now, each collection of subset data is used to train their decision trees.
- As a result, we end up with an ensemble of different models.
- Average of all the predictions from different trees are used which is more robust than a single decision tree.
- Algorithms: Random Forest

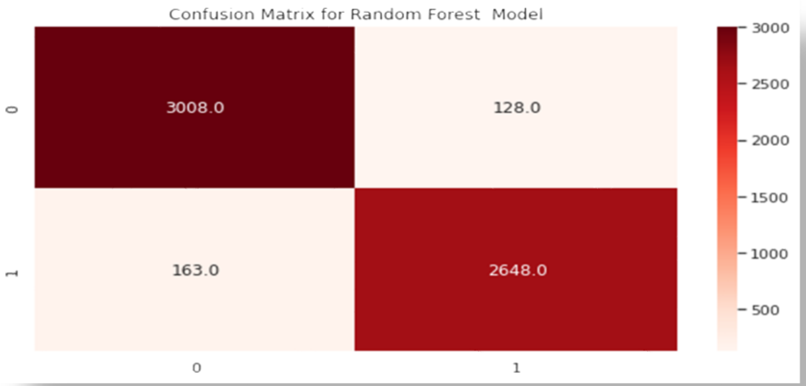
BOOSTING:

- Boosting is another ensemble technique to create a collection of predictors.
- In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors.
- In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.
- Algorithms: 1.XGBoost
  2.Gradient Boosting Machine

DATA MODELLING

RANDOM FOREST

- Random Forest is a powerful and versatile **supervised machine learning algorithm** that grows and combines multiple decision trees to create a “forest.”
- It can be used for both classification and regression problems in R and Python.
- Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not.
- But together, all the trees predict the correct output.
- Accuracy for random forest is 95.10% and recall is 94.20%.



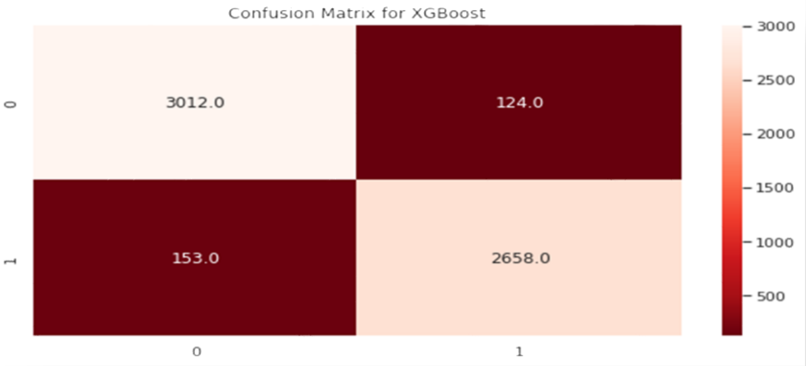
	precision	recall	f1-score	support
0	0.95	0.96	0.95	3136
1	0.95	0.94	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

Accuracy of the Model: 95.10677652597948%

DATA MODELLING

XG-BOOST

- XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.
- In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.
- However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.
- Accuracy for “XGBoost” is 95.34% and recall is 94.55%.



	precision	recall	f1-score	support
0	0.95	0.96	0.96	3136
1	0.96	0.95	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

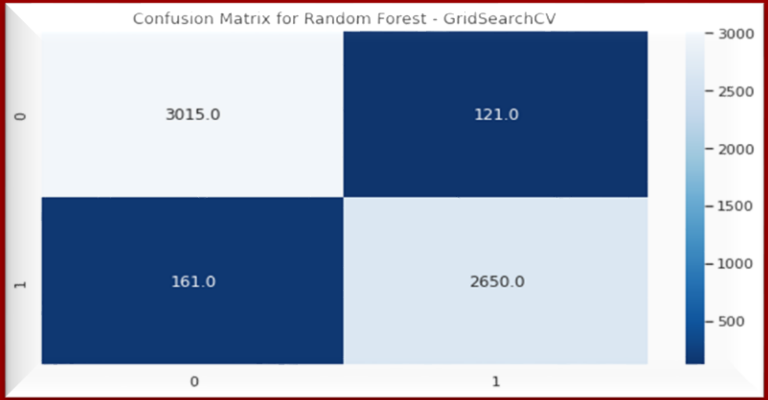
Accuracy of the Model: 95.34218933916262%

DATA MODELLING

CROSS VALIDATION FOR RANDOM FOREST

	precision	recall	f1-score	support
0	0.95	0.96	0.96	3136
1	0.96	0.94	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

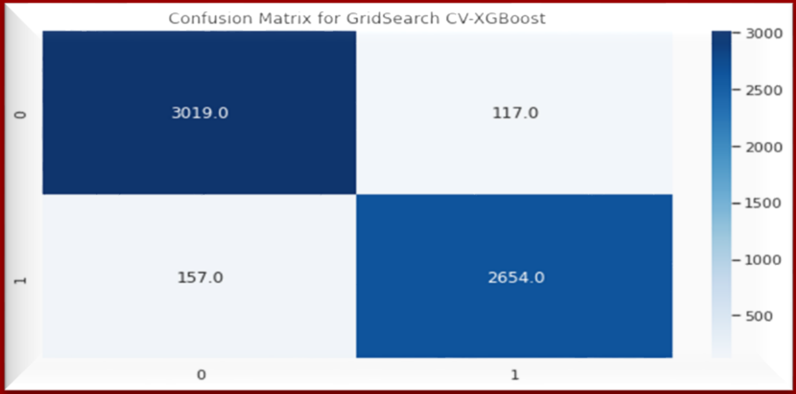
Accuracy of the Model: 95.25811333445434%



CROSS VALIDATION FOR XG-BOOST

	precision	recall	f1-score	support
0	0.95	0.96	0.96	3136
1	0.96	0.94	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

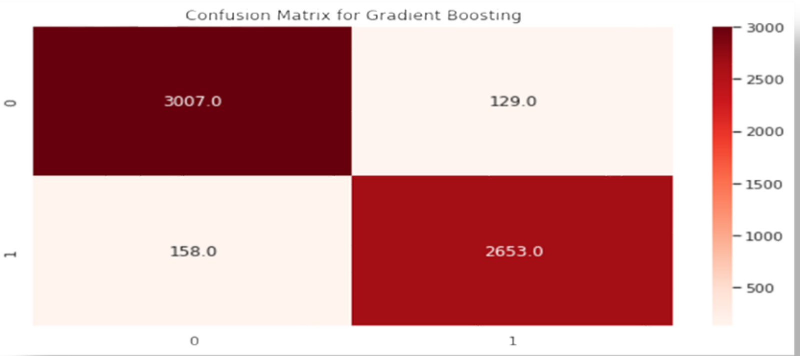
Accuracy of the Model: 95.39263494198755%



DATA MODELLING

GRADIENT BOOSTING MACHINE

- Gradient Boosting is an extension over boosting method.
- Gradient Boosting= Gradient Descent + Boosting.
- It uses gradient descent algorithm which can optimize any differentiable loss function.
- An ensemble of trees are built one by one and individual trees are summed sequentially.
- Next tree tries to recover the loss (difference between actual and predicted values).
- Accuracy for “gradient boosting machine” is 95.17% and recall is 94.37%.



	precision	recall	f1-score	support
0	0.95	0.96	0.95	3136
1	0.95	0.94	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

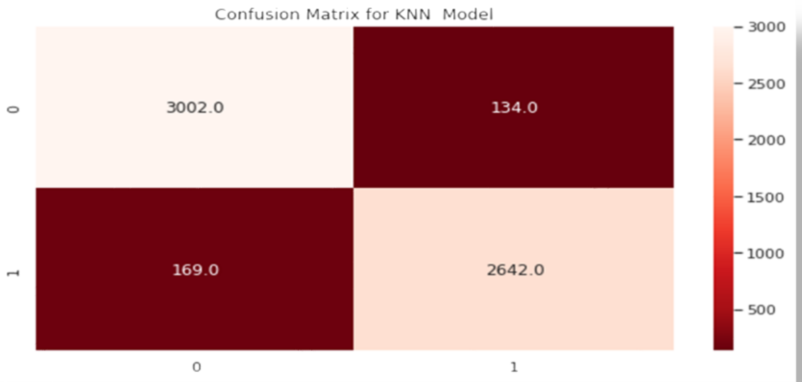
Accuracy of the Model: 95.17403732974608%

DATA MODELLING

AI

K- NEAREST NEIGHBOUR

- K nearest neighbour or KNN Algorithm is a simple algorithm that uses the entire dataset in its training phase.
- Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances, and the data with the most similar instance is finally returned as the prediction.
- Accuracy for "KNN" is 94.90% and recall is 93.98%.



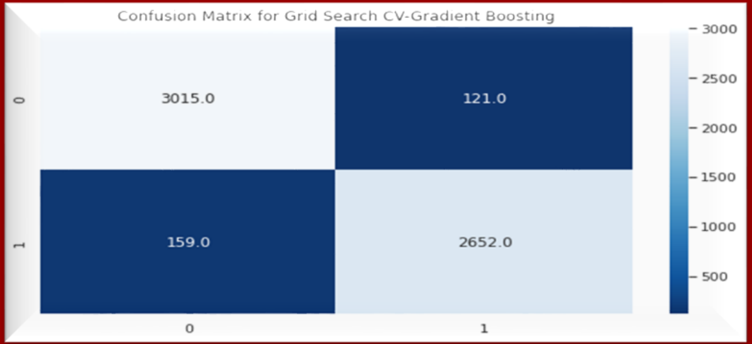
	precision	recall	f1-score	support
0	0.95	0.96	0.95	3136
1	0.95	0.94	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

Accuracy of the Model: 94.90499411467968%

DATA MODELLING

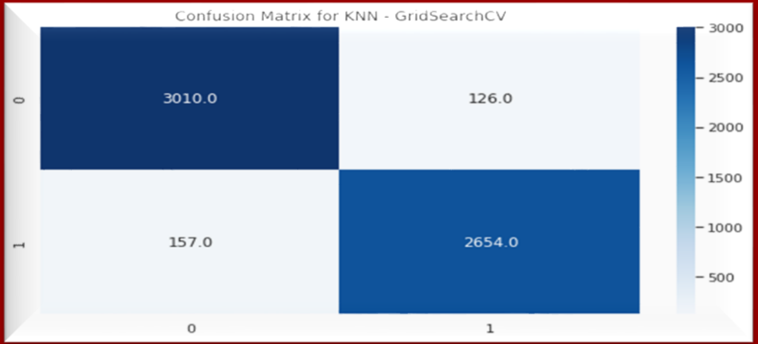
CROSS VALIDATION FOR GRADIENT BOOSTING

```
              precision    recall  f1-score   support

           0       0.95      0.96      0.96      3136
           1       0.96      0.94      0.95      2811

    accuracy                           0.95      5947
   macro avg       0.95      0.95      0.95      5947
weighted avg       0.95      0.95      0.95      5947

Accuracy of the Model: 95.29174373633765%
```



CROSS VALIDATION FOR KNN

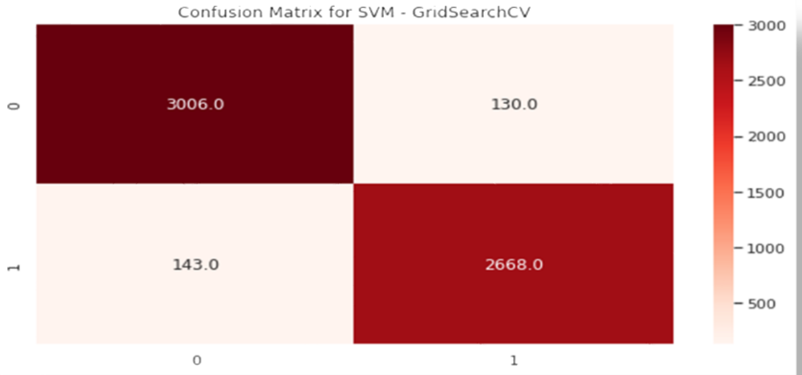
```
              precision    recall  f1-score   support

           0       0.95      0.96      0.96      3136
           1       0.95      0.94      0.95      2811

    accuracy                           0.95      5947
   macro avg       0.95      0.95      0.95      5947
weighted avg       0.95      0.95      0.95      5947

Accuracy of the Model: 95.2412981335127%
```



DATA MODELLING

SUPPORT VECTOR MACHINE

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.
- However, primarily, it is used for Classification problems in Machine Learning.
- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.
- This best decision boundary is called a hyperplane.
- Accuracy for "SVM" is 95.40% and recall is 94.91%.



	precision	recall	f1-score	support
0	0.95	0.96	0.96	3136
1	0.95	0.95	0.95	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

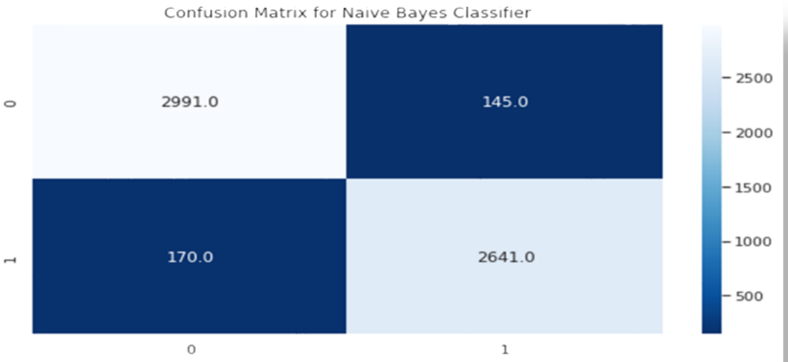
Accuracy of the Model: 95.4094501429292%

DATA MODELLING

AI

NAÏVE BAYES CLASSIFIER

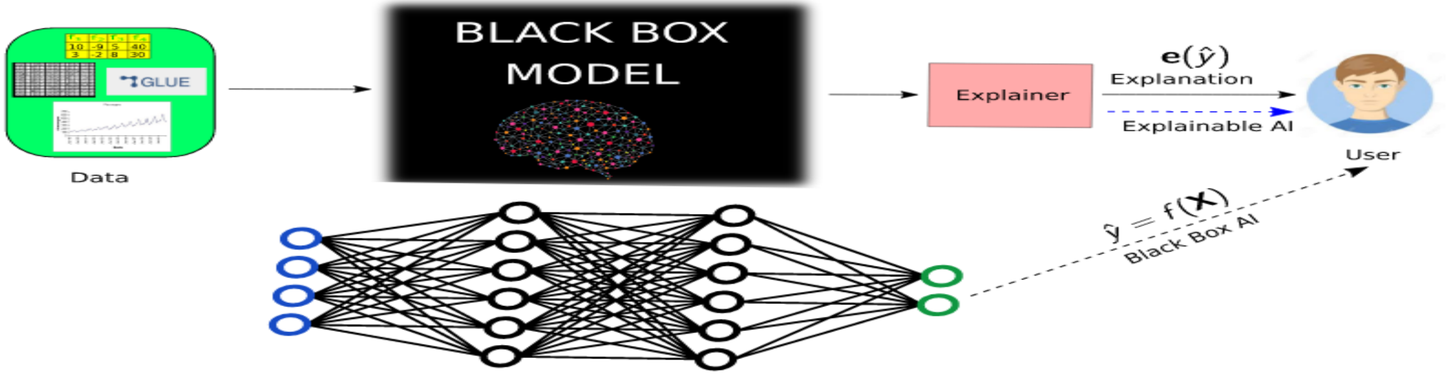
- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Accuracy for "Naïve Bayes Classifier" is 94.70% and recall is 93.95%.



	precision	recall	f1-score	support
0	0.95	0.95	0.95	3136
1	0.95	0.94	0.94	2811
accuracy			0.95	5947
macro avg	0.95	0.95	0.95	5947
weighted avg	0.95	0.95	0.95	5947

Accuracy of the Model: 94.70321170337985%

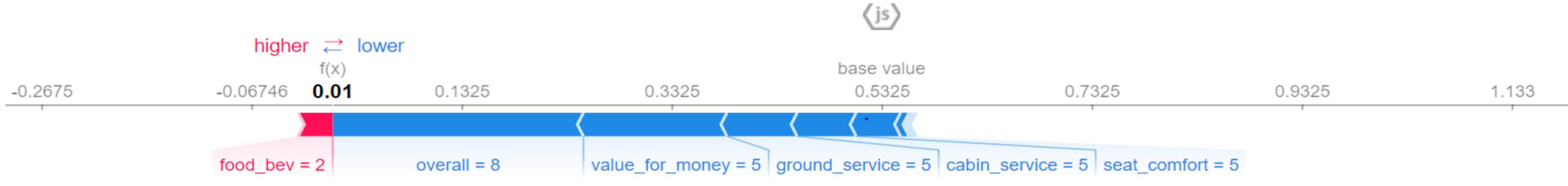
MODEL EXPLAINABILITY



- Interpretability is about the extent to which a cause and effect can be observed within a system.
- Or, to put it another way, it is the extent to which you can predict what is going to happen, given a change in input or algorithmic parameters.
- Explainability, meanwhile, is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
- We have used SHAP and ELI5 to explain our models.

MODEL EXPLAINABILITY(CONTD.)

SHAP



- SHAP Values (an acronym from SHapley Additive exPlanations) break down a prediction to show the impact of each feature.
- SHAP values interpret the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value.
- We have used SHAP to explain random forest.

MODEL EXPLAINABILITY(CONTD.)

ELI5

XG-Boost

Weight	Feature
0.9005	overall
0.0274	value_for_money
0.0190	ground_service
0.0115	cabin_service
0.0094	seat_comfort
0.0081	food_bev
0.0075	Couple Leisure
0.0045	entertainment
0.0044	Economy Class
0.0036	First Class
0.0026	Family Leisure
0.0016	Premium Economy
0	Solo Leisure

y=0 (probability **0.993**, score **-5.018**) top features

Contribution?	Feature	Value
+3.474	overall	1.000
+0.897	value_for_money	1.000
+0.395	food_bev	1.000
+0.232	<BIAS>	1.000
+0.114	seat_comfort	3.000
+0.094	cabin_service	3.000
+0.091	entertainment	1.400
+0.012	Couple Leisure	0.000
+0.010	Economy Class	1.000
-0.004	First Class	0.000
-0.007	Premium Economy	0.000
-0.024	Family Leisure	0.000
-0.263	ground_service	3.000

Gradient Boosting

Weight	Feature
0.4343 ± 0.5066	x0
0.3345 ± 0.4247	x6
0.1401 ± 0.3373	x5
0.0407 ± 0.3245	x3
0.0277 ± 0.3104	x1
0.0170 ± 0.2994	x2
0.0046 ± 0.3080	x4
0.0004 ± 0.1245	x7
0.0002 ± 0.1858	x10
0.0002 ± 0.1248	x11
0.0002 ± 0.1384	x9
0.0001 ± 0.1212	x12
0.0001 ± 0.0811	x8

y=0 (probability **0.995**, score **-5.261**) top features

Contribution?	Feature	Value
+2.064	overall	1.000
+1.096	value_for_money	1.000
+0.514	food_bev	1.000
+0.392	seat_comfort	3.000
+0.374	entertainment	1.400
+0.291	ground_service	3.000
+0.198	Economy Class	1.000
+0.133	<BIAS>	1.000
+0.115	cabin_service	3.000
+0.025	Family Leisure	0.000
+0.024	First Class	0.000
+0.014	Couple Leisure	0.000
+0.013	Premium Economy	0.000
+0.008	Solo Leisure	0.000

- ELI5 is a Python package which helps to that machine learning classifiers and explain their predictions.
- It provides support for the following machine learning frameworks and packages: sci-kit-learn.
- Currently ELI5 allows to explain weights and predictions of sci-kit-learn linear classifiers and regressors, print decision trees as text or as SVG, show feature importance, and explain predictions of decision trees and tree-based ensembles.

CONCLUSION

- It is apparent that people gave a high recommendation to the economic class in the cabin. This tells us that people like to travel in economy class due to the low price, but we can also see that they give the economy class the highest negative ratings because they receive less infrastructure or service. Likewise, the business class has received the highest rating due to the quality service offered there, while the economy class has received the lowest rating due to its price or low attendance.
- 'British airways' has the maximum number of trips and this can be attributed to its ultra-low-cost fare compared to other airlines.
- Clearly, 'No' responses are more than 'Yes' responses in recommended, which means airlines have to focus on some aspects to make their fliers happy.
- 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 for random forest 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, and ground_service has high shap feature value.
- From Eli5 we can see overall and value for money contributed more to giving the positive recommendation and ground service and family leisure contributed to giving a negative recommendation for XGBoost.
- From Eli5 we can see overall and value for money contributed more to giving the positive recommendation and Gradient Boosting model.

CONCLUSION

MODEL NAME	ACCURACY	RECALL	PRECISION	F1-SCORE	ROC AUC SCORE
Support Vector Machine	0.954095	0.949128	0.953538	0.951328	0.953837
Grid Search CV- XGBoost	0.953926	0.944148	0.957777	0.950914	0.953420
XGBoost	0.953422	0.945571	0.955428	0.950474	0.953015
Logistic Regression	0.952917	0.951263	0.949237	0.950249	0.952832
Grid Search CV-Gradient Boosting	0.952917	0.943436	0.956365	0.949857	0.952426
Random Forest - GridSearchCV	0.952581	0.942725	0.956333	0.949480	0.952070
KNN - GridSearchCV	0.952413	0.944148	0.954676	0.949383	0.951985
Gradient Boosting	0.951740	0.943792	0.953630	0.948686	0.951329
Random Forest	0.951068	0.942014	0.953890	0.947915	0.950599
Decision Tree	0.950900	0.942725	0.952895	0.947783	0.950476
KNN Model	0.949050	0.939879	0.951729	0.945767	0.948575
Naive Bayes Classifier	0.947032	0.939523	0.947954	0.943720	0.946643

- According to our business needs, we will give first priority to recall and then to accuracy from a metrics point of view because we need to find how many people will recommend it.
- We can see that our models have performed very well all of the models have given recall greater than 90% which means our models are performing very well.
- Logistic Regression has the highest recall value It gave a recall of 95.12% followed by SVM which gave 94.91%.
- Support Vector Machine has the highest accuracy of the models but others also performed very well SVM gave 95.40% accuracy.
- Even after using Grid Search CV our models are giving similar accuracy.
- Naive Bayes Classifier and Random forest has the lowest recall of 93.95%

CHALLENGES

The close proximity of the evaluation scores of the models.

Analyzing how to fill the null values without losing the data

It was a very large dataset with more than 50% of null values.

References

I. Stack overflow
II. GeeksforGeeks
III. Jovian
IV. Research paper based on Study of Airline Industry
V. Analytics Vidhya
VI. Towards data science

THANK-YOU