Airline Passenger Referral Prediction

Snehal Dapke

Data science trainee, AlmaBetter

Abstract:

The airline industry exists in an intensely competitive market. Observing a growth of around 18% over the year and with the recent development, there has been a significant increase in the airline opportunities. We have data that contain rating on service provided by airline to customer. We do data extraction, data wrangling, data analysis, visualization. Our main mission to develop prediction model classification problem. Objective of this project to know how customers recommend the airlines to other people.using machine learning algorithm and ,hyperparameter tunning find best model that gives effective results.

Keywords: Airline Industry, Algorithm, wrangling

1. 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 business perspective.

Data is scraped in Spring 2019. The main objective is to predict whether passengers will refer the airline to their friends and others. Find out best model which gives realistic result.

2. 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.

2.1 Airline Referral Dataset

We have been provided with airline referral dataset. The data has been provided of reviews from 2006 to 2019. We analysed the dataset and worked on predicting whether people refer to others with using model parameter .

2.3 Python

Most of the info scientists use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is a simple programing language for select compared other languages. The most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is easy to use. That's one among the most reasons to use python. Specifically, for data scientists, the foremost popular data inbuilt open-source library is named panda. As we've seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, and 3-dimensional data python built-in library comes very helpful choose to wait a few minutes to see if the rates go back down.

3. Steps involved:

- **Exploratory Data Analysis**

  - **Data Exploration**

    After loading the dataset we started exploring about the data what we have at first I explored that our dataset is of shape 131895 rows and 17 columns. After knowing the shape we saw the datatypes of our the field we have and some info about our data where I got to know about null values which I need to treat. We also removed unnecessary columns which were not required for our prediction.

- **Null values Treatment / Duplicate values Treatment**

  In the dataset of airlines we saw that more than 50% of the data was null. Aircraft column was having 80% of there value as null there is no point in keeping that column so we dropped that column. There were duplicates as well around 85121 values we need to remove those duplicates by keeping only first row.

  Then comes the null values treatment part we used **"KNN Imputer"** to fill the null values for numerical features. For categorical features we decided to drop those values as they can predict false values after filling those values because we have more than 50% of null values in our dataset. After treating null values our dataset became of shape 29731 rows and 11 columns .

- **Outlier Treatment**

  To see if the outlier is present or not we need to see the Box Plot. We plotted Boxplot there were no outliers present in our data.

- **Data Visualisations**

  Data visualisation is used to see huge amount of data visually to get better insights of the data. In our case we have a huge dataset with the number of features of 11 so we tried visualizing the number of unique cabins to see passenger prefers which cabin more. Then

we also visualized the cabins with respect to recommended. After that we plotted a bar graph to see the type of traveller and the top 10 airlines which passenger tend to fly-in more and saw the value count for our dependent variable i.e. Recommended. Lastly we saw the correlation present in our data and all the features with respect to recommended to get the relation between independent and dependent variables.

- **One Hot Encoding**

  In this technique, the categorical parameters will prepare separate columns for both Male and Female labels. So, wherever there is Male, the value will be 1 in Male column and 0 in Female column and vice-versa. We did one hot encoding on traveller type and on cabin.

- **Fitting into different models**

  For modelling, we tried various classification algorithms like:

  - Logistic Regression
  - Decision Tree
  - Random Forest
  - XGBoost
  - Gradient Boosting Machine
  - KNN
  - SVM
  - Naïve Bayes Classifier

- **Cross-Validation on model**

  Cross-Validation is a resampling technique that helps to make our model sure about its efficiency and accuracy on the unseen data. It is a method for evaluating Machine Learning models by training several other Machine learning models in subsets of the available input data set and evaluating them on the subset of the data set.

- **Model Explainability / Interpretability**

  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.

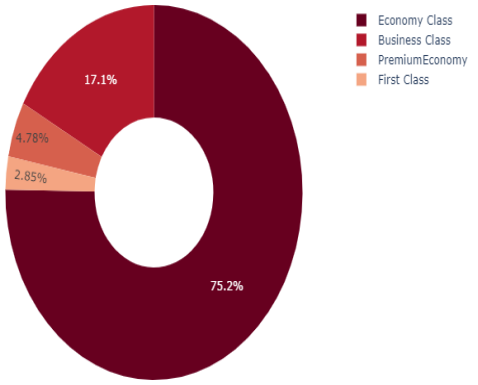
  We used SHAP and ELI5 to explain our model.

3.1 EDA

Exploratory data analysis (EDA) plays vital role in analysis of data and gives idea of feature engineering. EDA help us to determine dependent and independent variables.

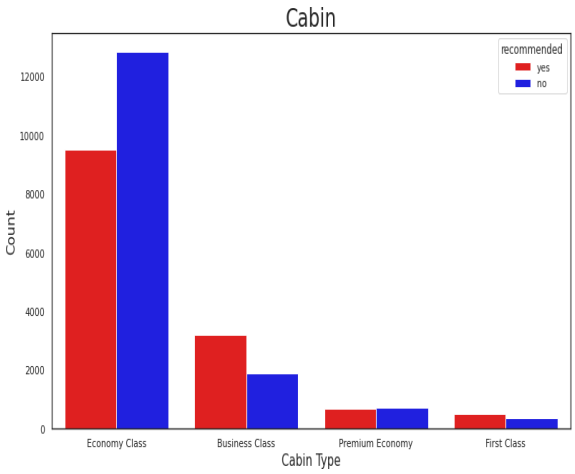
3.1.1 Passenger class percentage

Pie chart clears pictures about cabin in which most number of passenger travelled.From pie chart 75% of total passenger use “Economy class” for travel and only around 3% used “First class”.

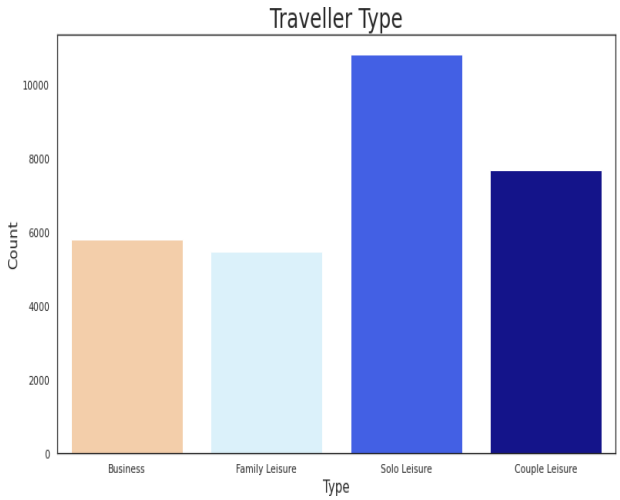


3.1.2 Countplot for cabin with recommended

Feature cabin and recommended combine countplot for insights shows economy class has the most recommendation whereas first class has the least recommendation.As per above main conclusion we can derive that many people say “NO” to business class to recommend to others.



3.1.4 Countplot for traveller type with Ratings

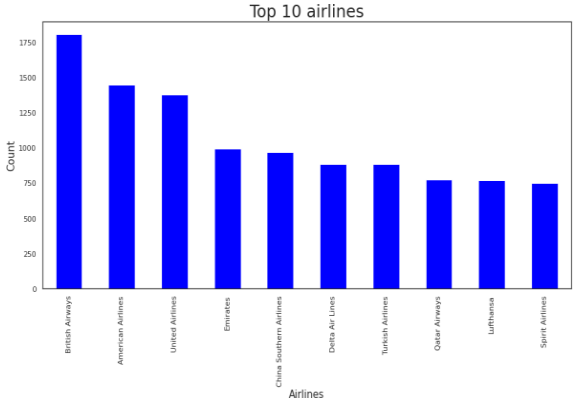


It’s clear from the countplot that 'Solo Leisure' has highest ratings among all whereas 'Family Leisure' has the least ratings.

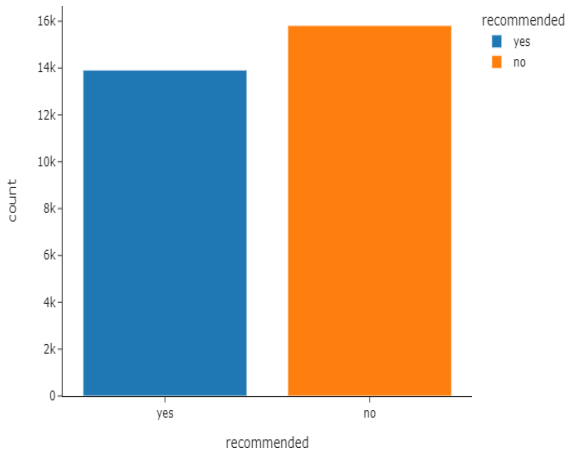
3.1.5 Top 10 airlines distribution

There multiple companies airlines we seen in dataset but for easy mind we have sorted top airlines companies.

British airways' has the maximum number of trips and this can be attributed to its ultra low cost fare compared to other airlines.

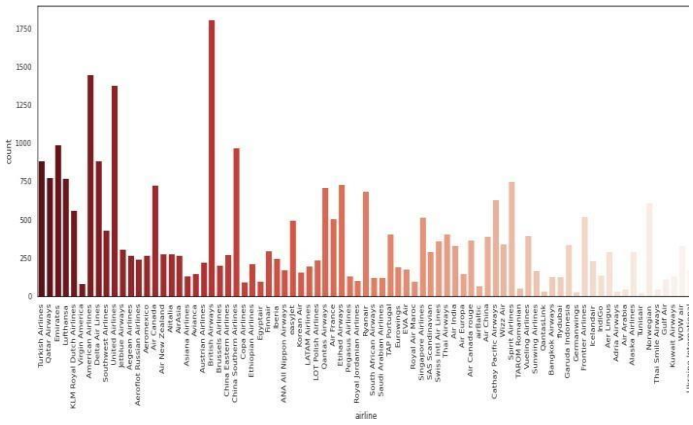


3.1.6 Histogram for Recommended



No = 15817 Yes = 13914
Clearly, 'No' responses are more as compared to 'Yes' responses

3.1.7 Countplot for Airlines

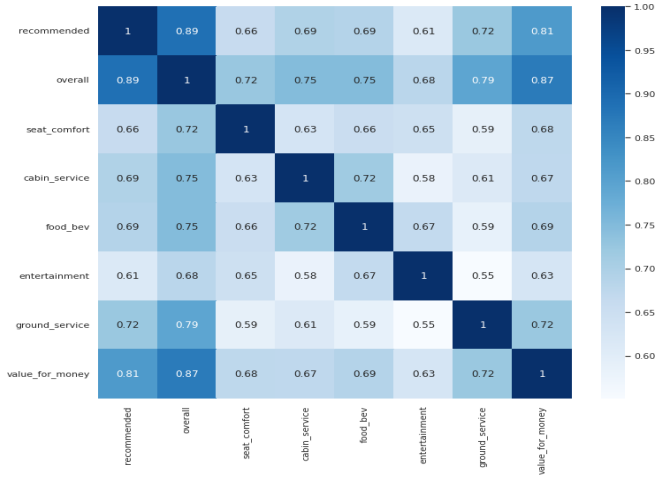


From above count plot we know that, As we have seen earlier 'British Airways' is the topmost Airline. 'Tunisair','Germanwings' etc are the lowest number of trips.

3.1.8 Correlation

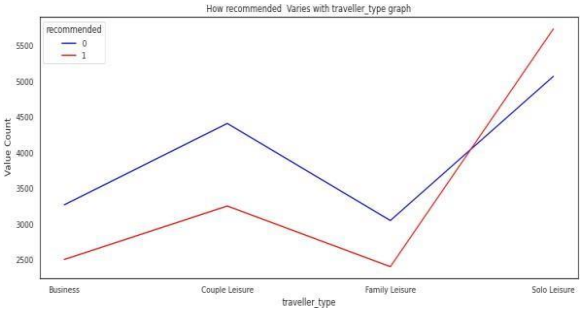
Correlation refers to a process for establishing the relationships between two variables.
Positive correlation: A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.

Heatmap



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

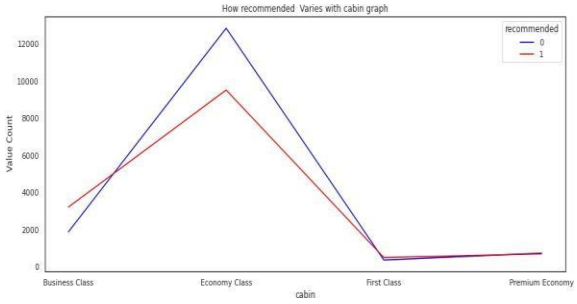
3.1.9 PLOT FOR THE FEATURES WRT TO RECOMMENDED



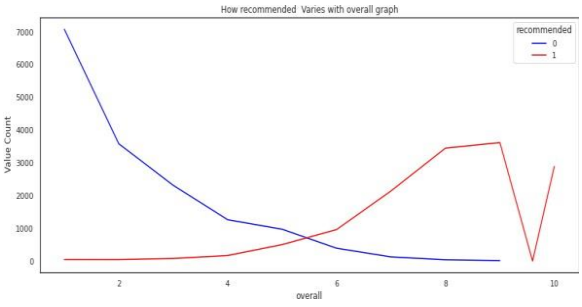
- We can see, in both the business and leisure traveller types, that both the recommendation trend in terms of yes or no increases from business to couple leisure and it decreases to

family and again reaches a high level in solo leisure.

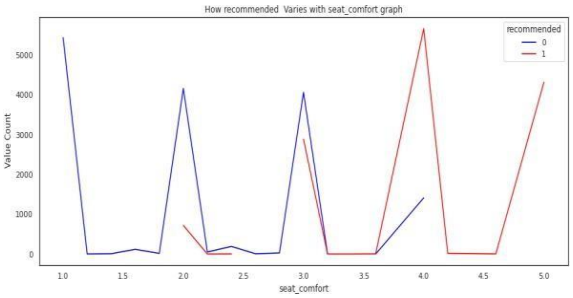
- This indicates people prefer solo leisure higher than any of the other leisure.



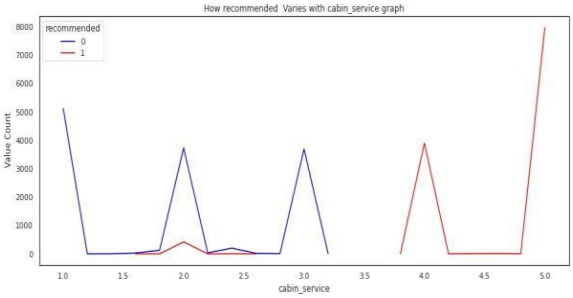
- With regards to cabin type, it has been determined that both yes and no recommendation trends increase from business class to economy class, then decrease to first class, and again increase slightly in the premium class.
- This indicates most people travel in economy class.



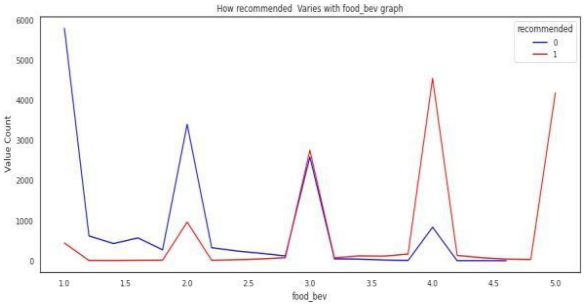
- Generally, we can observe a very good insight which is also regular in the overall rating.
- We can see that positive recommendation increase with the overall rating, while negative recommendations decrease.



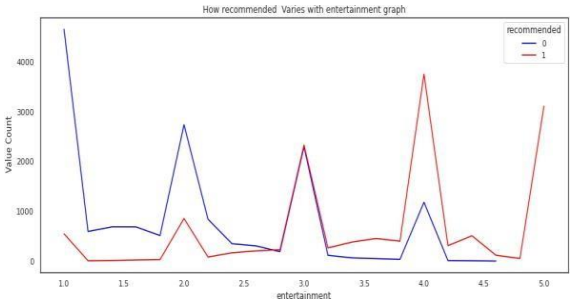
- In seat comfort we can see the negative recommendation is there till 4.0 rating but after that, we can see positive recommendation also.



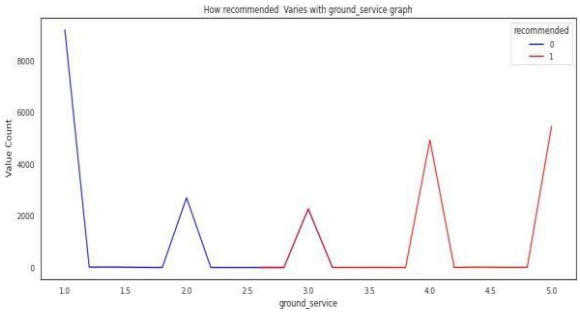
- In cabin service also we can see the similar trend as seat comfort negative recommendation is there till 3.0 rating but after that we can see positive recommendation also.



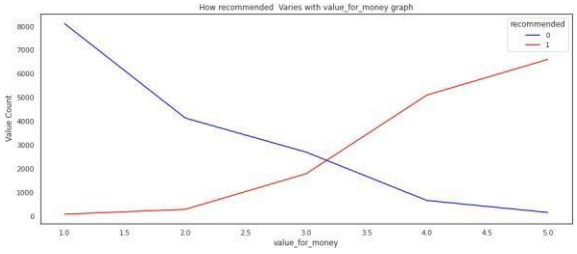
- In food bev we can see mixed recommendations initially as the negative recommendation decreases positive recommendations are increasing.



- 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



- 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.

3.2 Feature Description

1. Airline: Name of the Airline
2. Overall: Overall rating is given to the trip between 1 to 10.
3. Author: Name of the Author to provide reviews about the trip
4. Review Date: Date of the Review
5. Customer review: Review of the customers in a text format
6. Aircraft: Type of the Aircraft
7. Traveler type: Type of traveler (Labels: Business, Family Leisure,Solo Leisure & Couple Leisure)
8. Cabin: Cabin at the flight (Labels: Economy Class,Business Class,Premium Class & First Class)
9. Date flown: Flight date
10. Seat comfort: Rating between 1-5
11. Cabin Service: Rating between 1-5
12. Foodbev: Rating between 1-5
13. Entertainment: Rating between 1-5
14. Groundservice: Rating between 1-5
15. Value For Money: Rating between 1-5
16. Recommended: Binary Labels(Target Variable)

3.2 Data Modelling

Splitting data

X = Independent variable

Y = Dependent variable

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

```
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
```

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 modelling portion.

4.1 Algorithms

Its time to apply different models on given dataset as follows.

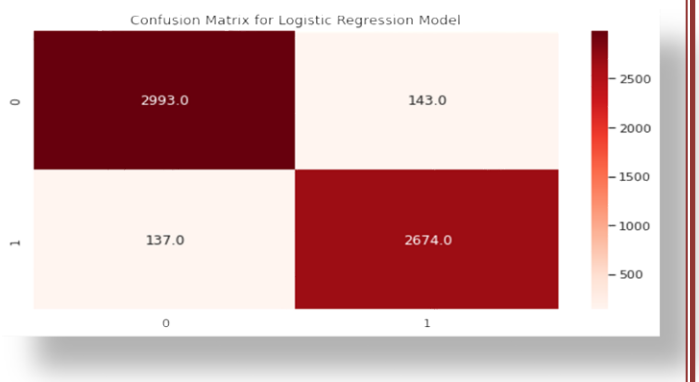
1) Logistic regression

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 powerful supervised ML algorithm used for binary classification problems (when target is categorical).

	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%



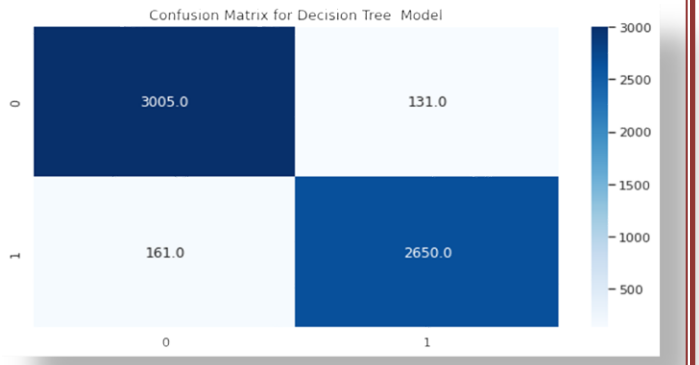
2)Decision Tree

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%



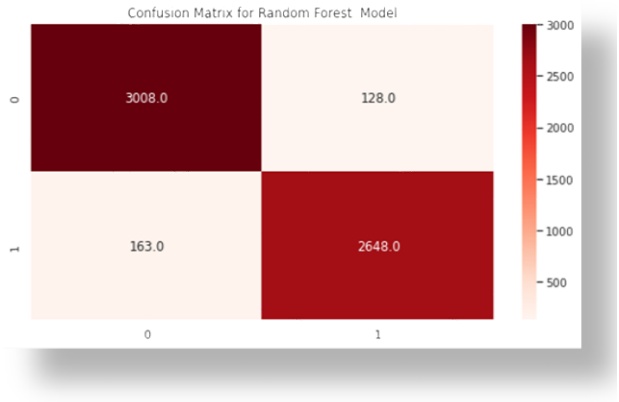
3) Random forest

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.

	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%



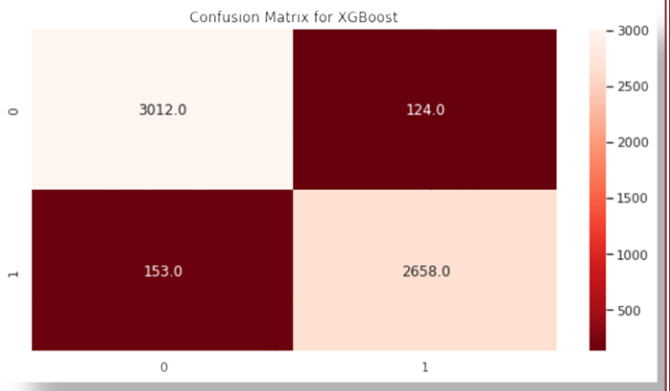
4) XGBoost

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%



5) Gradient Boosting Machine

Gradient Boosting is an extension over boosting method.

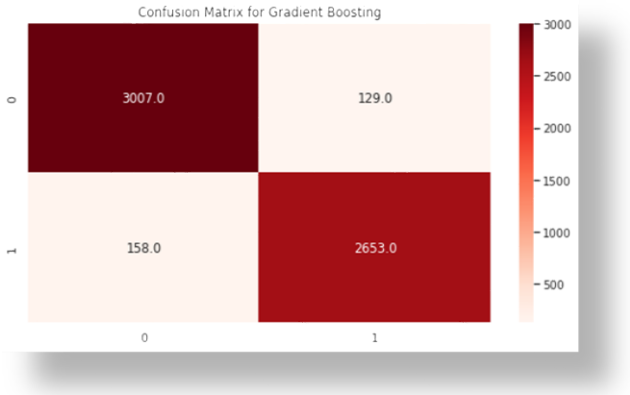
Gradient Boosting=Gradient Descent + Boosting.

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%



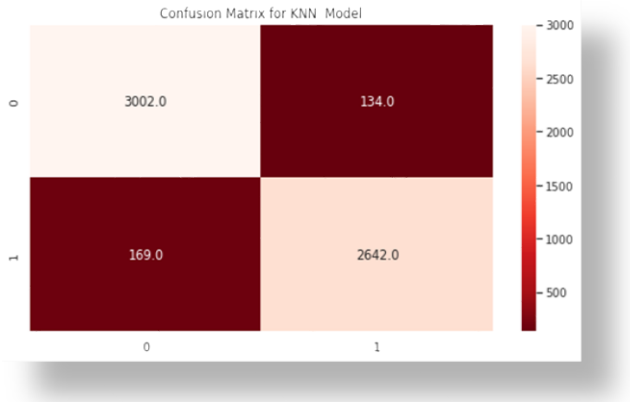
6) K-Nearest Neighbour

K nearest neighbour or KNN Algorithm is a simple algorithm which 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.

The k-nearest neighbors algorithm uses a very simple approach to perform classification. When tested with a new example, it looks through the training data and finds the k training examples that are closest to the new example. It then assigns the most common class label (among those k-training examples) to the test example.

	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%

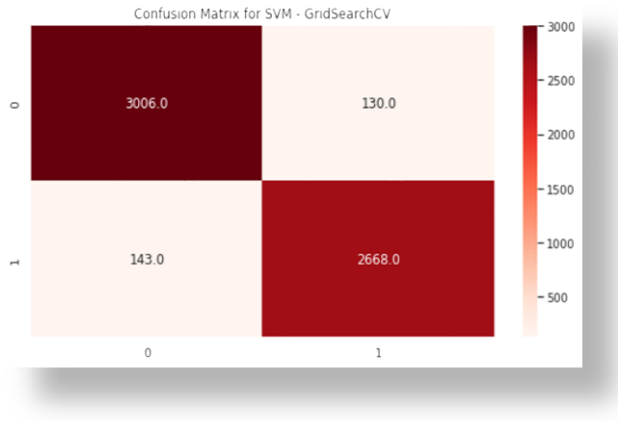


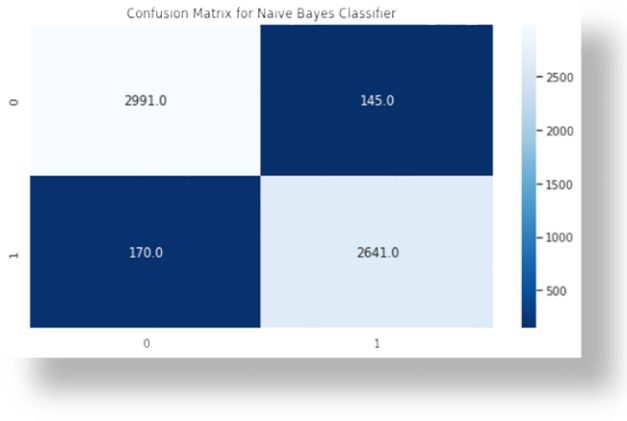
7) Support vector machine

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. he 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.

	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%





8) Naive bayes

The naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. It implements the Bayes theorem for the computation and used class levels represented as feature values or vectors of predictors for classification. Naive Bayes Algorithm is a fast algorithm for classification problems. This algorithm is a good fit for real-time prediction, multi-class prediction, recommendation system, text classification, and sentiment analysis use cases. Naive Bayes Algorithm can be built using Gaussian, Multinomial and Bernoulli distribution. This algorithm is scalable and easy to implement for a large data set.

	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%

4.2 Hyper parameter tuning

Parameter- A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

Hyper-prameter- A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data.

Grid Searching of hyperparameter-

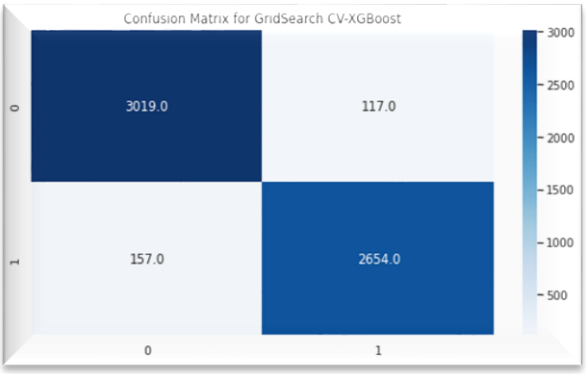
Grid search is an approach to hyper parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

Grid Search combines a selection of hyper parameters established by the scientist and runs through all of them to evaluate the model's performance. This is a simple technique that will go through all the programmed combinations. The biggest disadvantage is traverses a specific region of the parameter space and not understand which movement or which region space is important to optimize the model.

Grid search CV XGBOOST:

	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

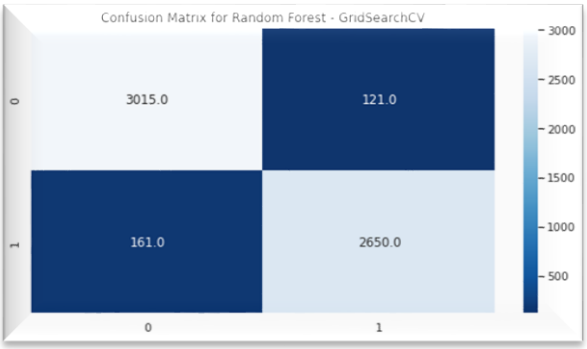
Accuracy of the Model: 95.39263494198755%



Grid search CV 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

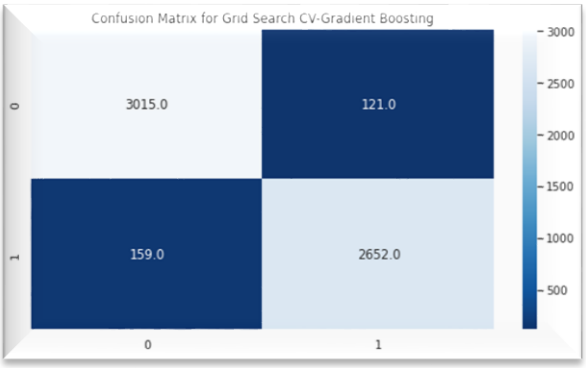
Accuracy of the Model: 95.25811333445434%



Grid search CV 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

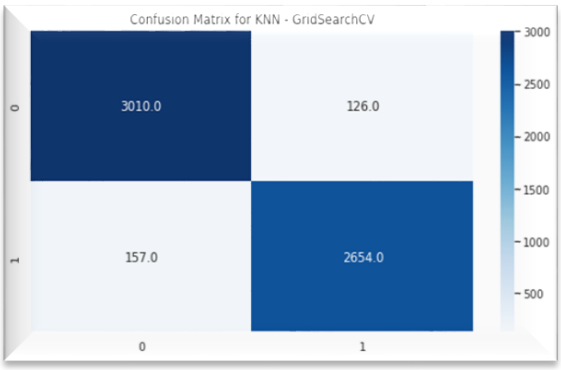
Accuracy of the Model: 95.29174373633765%



Grid search CV 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%

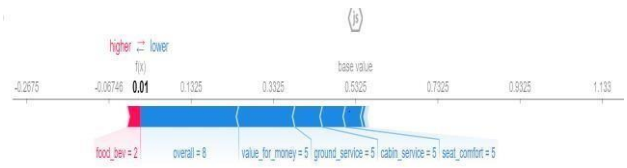


5. Model Explanability or Interpretability

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.

5.1 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.



5.2 ELI5

ELI5 is a Python package which helps to debug machine learning classifiers and explain their predictions. It provides support for the following machine learning frameworks and packages:

scikit-learn. Currently ELI5 allows to explain weights and predictions of scikit-learn linear classifiers and regressors, print decision trees as text or as SVG, show feature importances and explain predictions of decision trees and tree-based ensembles.

XGBOOST:

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

6.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

- Most of the people recommendation on economic class cabin people enjoy journey with this class.Apart from this high negative ratings for this class.
- ‘British airways' has the maximum number of trips and this happened due to its ultra low cost fare compared to other airlines.
- 'No' responses are more as compared to 'Yes' responses in recommended that means airlines have to focus on some aspects to make there fliers happy.
- food and beverage rating get highest negative recommendation to rating 1.0 due to this we can conclude that airline service has to improve their food delivery and other services.
- Ground service, value for money, entertainment has negative rating 1 it shows need to improve these services.
- Logistic Regression has the highest recall value It gave the recall of 95.12% followed by SVM which gave 94.91%.
- Support Vector Machine has the highest accuracy from the models but others are also performed very well SVM gave 95.40% accuracy. Even after using Grid Search CV our models are giving similar accuracy.
- 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.
- From Eli5 we can see overall and value for money contributed more to give the positive recommendation and ground service and family leisure contributed to give negative recommendation for XGBoost .

References-

5. Stackoverflow
6. GeeksforGeeks
7. Jovian
8. Research paper based on Study of airline industry.