Supervised ML: - Classification

Airline Passenger Referral Prediction

Content:

- **Introduction**
- **Problem Statement**
- **Data Description**
- **Data cleaning**
- **Data Wrangling Code**
- **Exploratory Data Analysis**
- **Feature Engineering & Data Pre-processing**
- **ML-Model Implementation**
- **Conclusion**

1 INTRODUCTION

- Air transport or aviation plays a very important role in the present transport structure of the world and surely it is considered the gift of the twentieth century to the world. In today's fast-paced world, air transport has been a blessing to all because of its speed of transportation. This mode of transport is very useful to get the products with short delivery times quickly and safely to those who require it also allows the tourism industry in each country to have stable growth by shortening the distance among all the people who inhabit the world. Here, I have a dataset regarding the ratings of services provided by different airlines to customers. The main objective of this project is to understand how likely the passengers will recommend the airlines to others.

- The dataset here is quite large which initially had 131895 rows and 17 columns. On checking the data information, it was derived that there were basically two different types of data in the dataset there are 7 columns of floats64, data types 10 columns with object types.

- Data is scrapped in the Spring of 2019 from the 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 them or not.

2 Data Summary

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

We have Started with data loading and we have done EDA, feature engineering, data cleaning, target encoding feature selection and then model building. So, we have used this model:

- Logistic Regression Model
- Decision Tree Model
- Random Forest Model
- Support Vector Machine Model
- K-Nearest-Neighbor Model
- Naïve Bayes

We performed hyperparameter tuning for decision tree models, random forest models, K-Nearest Neighbors, Support Vector Machines, and Naive Bayes using the Grid Search CV method. This is done to improve accuracy and avoid overfitting criteria. After that, we completed the Gradient Boosting model by fine-tuning the hyperparameters.

Based on an understanding of the business and problem use cases. The classification metrics for Recall are given first priority, Accuray second priority, and ROC AUC third priority.

We created classifier models using 6 different classifier types, all of which provided over 90% accuracy. We can conclude that Logistic Regression gives the best model.

Comparing the model evaluation metrics, we can see that the SVM is the most accurate model with a very small margin, and performs the best among the experimental models on the given dataset.

The most important features were overall rating and value for money, which helped the model predict whether passengers would recommend a particular airline to their friends.

The developed classifier models can be used to predict passenger recommendations as they enable airlines to identify influential passengers who can help generate more revenue.

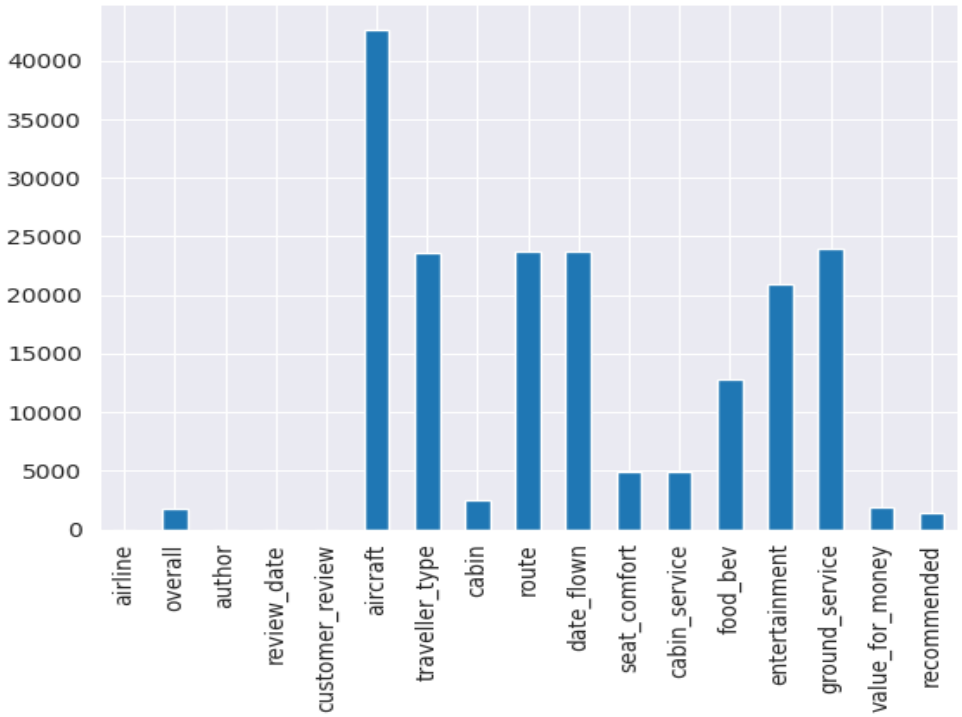
Therefore, in order to grow or grow their business, our customers must provide outstanding cabin service, ground handling, entertaining food, beverages and comfortable seating.

3 DATA OVERVIEW

- airline: Name of the airline.
- overall: Overall point is given to the trip between 1 to 10.
- author: Author of the trip
- review date: Date of the Review
- customer review: Review of the customers in free text format
- aircraft: Type of the aircraft
- traveler type: Type of traveler (e.g. business, leisure)
- cabin: Cabin at the flight date flown: Flight date
- seat comfort: Rated between 1-5
- cabin service: Rated between 1-5
- Food-BEV: Rated between 1-5
- entertainment: Rated between 1-5
- ground service: Rated between 1-5
- value for money: Rated between 1-5
- recommended: Binary, target variable.

4 Data cleaning

- This dataset has 70711 duplicate values that's why drop duplicate values.
- Missing Values/Null Values present here.



5 Data Wrangling

- Percentage wise missing values checking after Dropping the aircraft column from data as it has highest null values.
- Imputed null values by Quantile-1 for the columns have low null value percentage.
- Imputed null values by Median Imputation for the columns have high percentage.
- Filling traveler-type column with Mode Imputation
- cabin column with Forward fill method.
- It is better to work with clean data for prediction rather than huge corrupt data

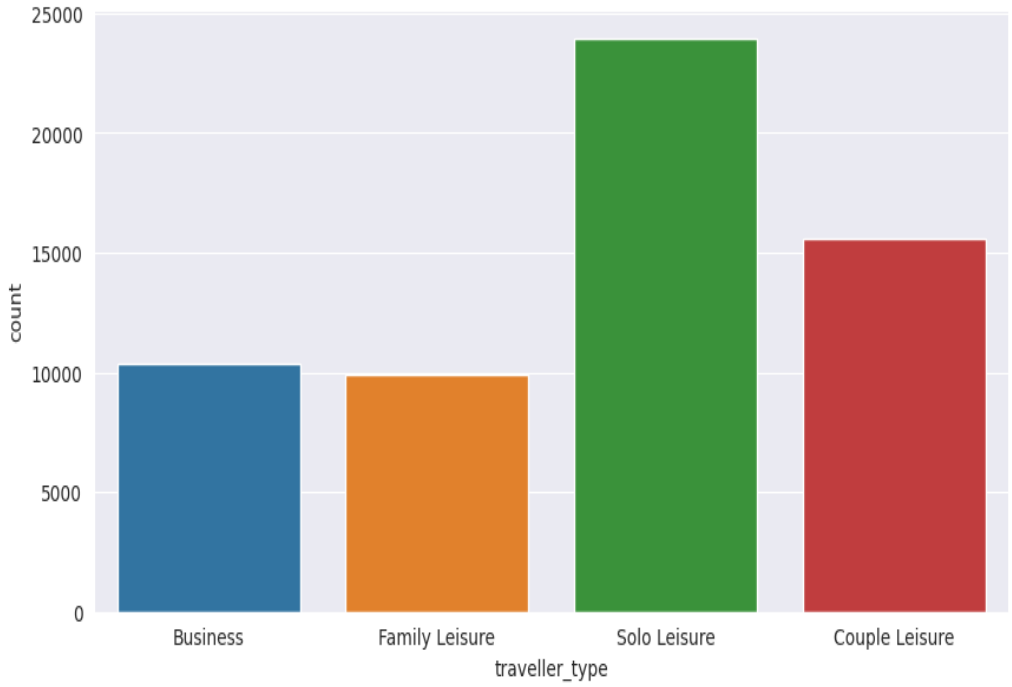
REASON OF DROPPING COLUMNS—

- Author - Being the categorical with high Variability not required for prediction.
- Route - Not needed for building a model as it is independent of the Services and Quality of travel.
- Date-flown - Not needed for building a model as it is not a time series data, also some common time period is there between 2 dates.
- Review-date - Similar to Date-flown
- Customer-review - As it is related to overall review feature of the datasets. On the basis of null value percentage, we divide our data in two parts-
- High-null = columns which have high percentage of null values.
- Low-null = columns which have low percentage of null values

6 EXPLORATORY DATA ANALYSIS

6.1 Which Traveler-type has more ratings?

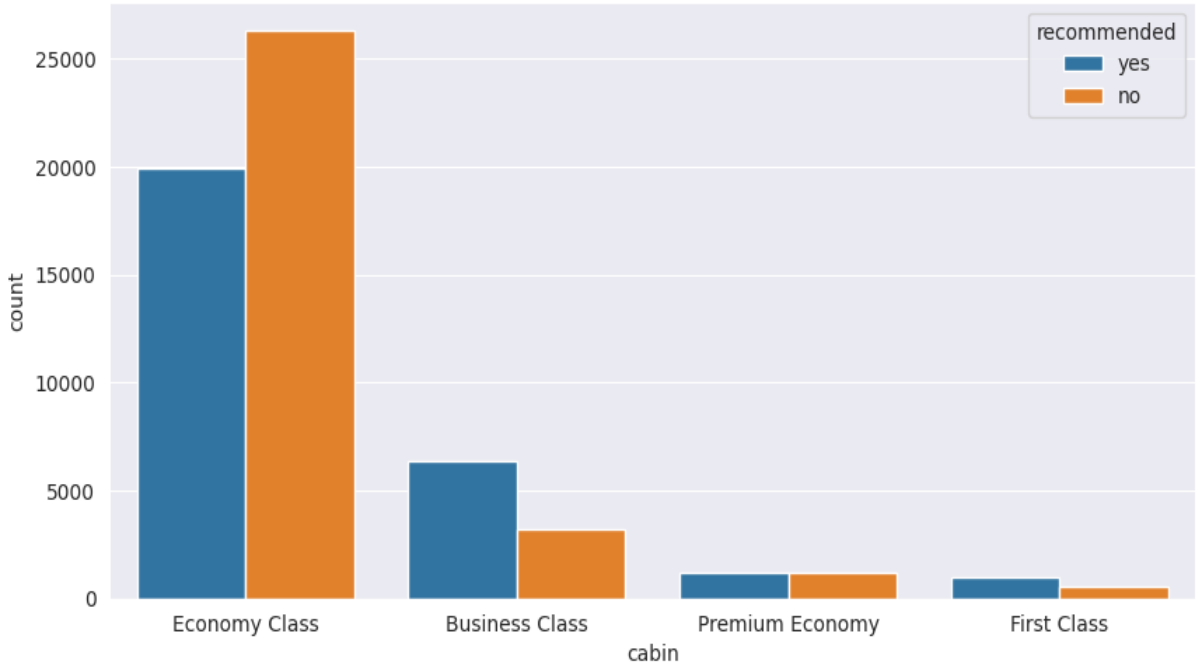
Travelling type of Solo Leisure has more ratings, In the airplane most of the traveler are solo Leisure followed by Couple Leisure, Business and Family Leisure. We should focus on Family Leisure so that more family member travel together in the flight.



6.2 Which type of Cabin has more recommendation?

On the basis of graph –

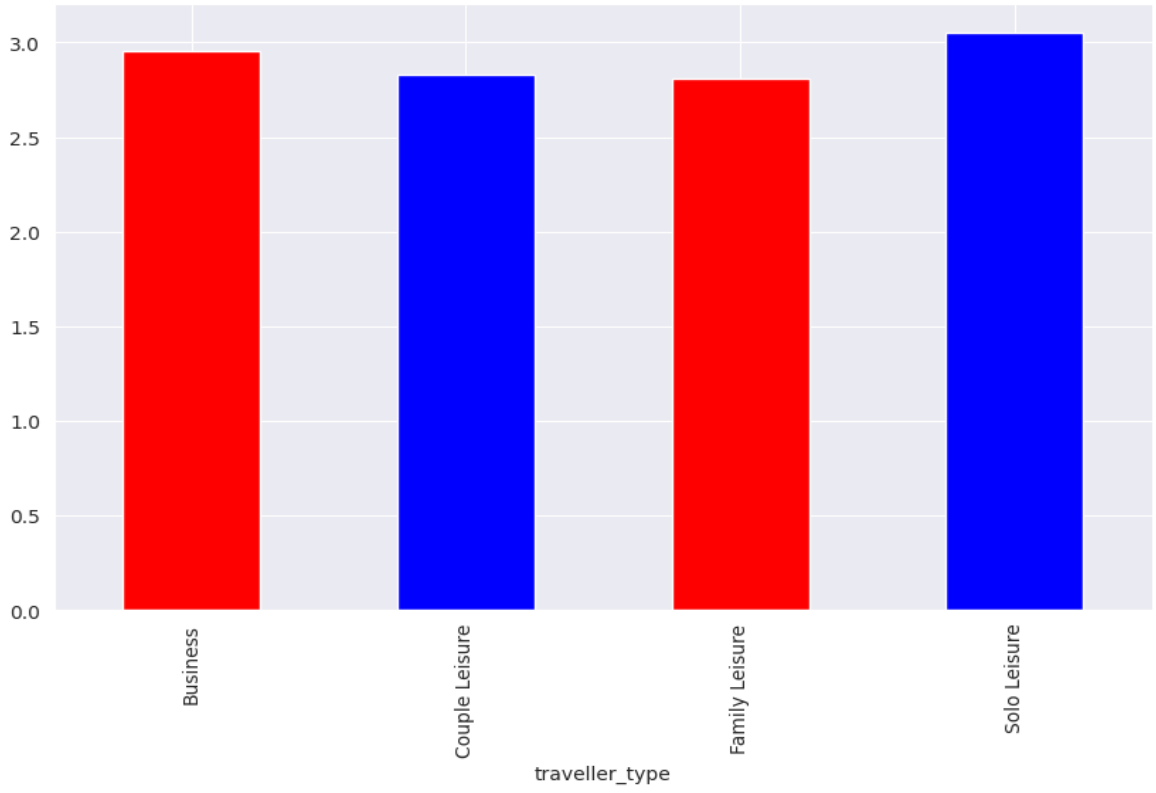
- Economy class has highest recommendation with bad reviews.
- Business class has second most recommended cabin type with good reviews.
- premium economy has equal reviews.
- first class is least recommended cabin type with good reviews.
- Here we can see that most of the traveler travel in the Economy Class and very less traveler travel in the Premium and First class.
- In the Business class about 50% of traveler are not recommended to the other traveler and also same as First class for that reason we should make some effort in business for more recommendation by traveler.
- In the Economy class about 30% passengers are not recommended the other so this is also a drawback of business.



6.3 Which type of traveler type is value for money?

- We can say that travelling Type of Solo Leisure worth of Money compare to other type of travelling.

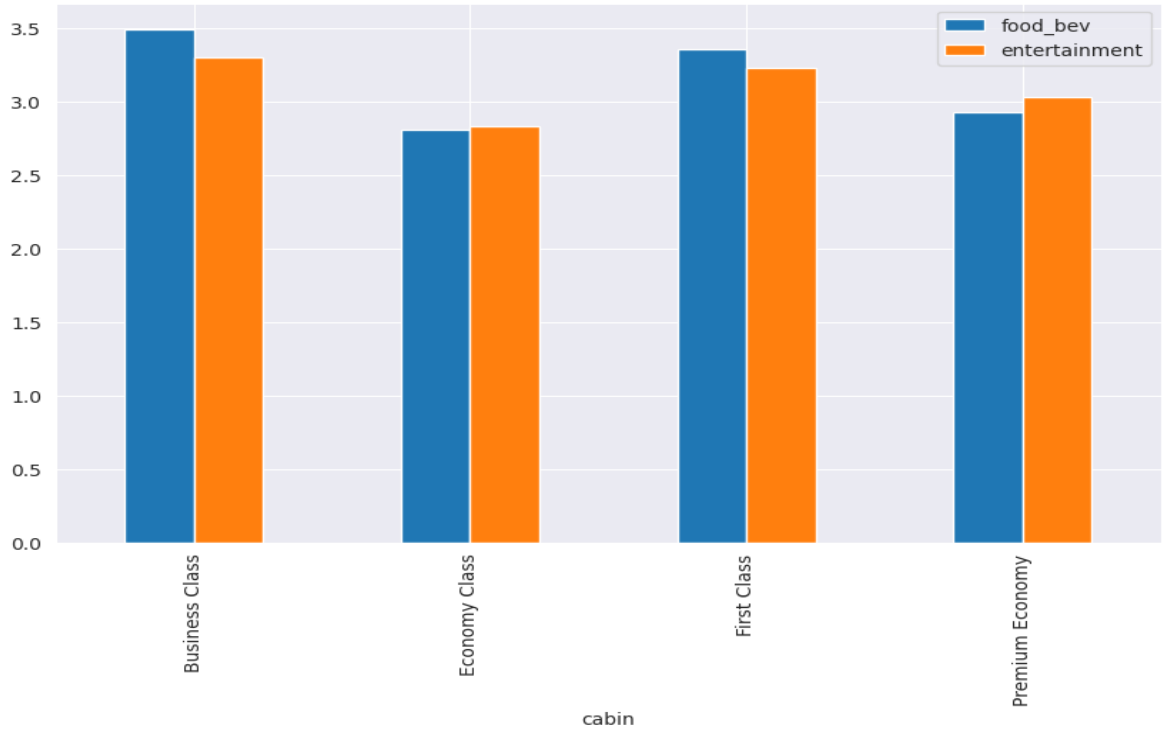
- Solo Leisure traveling type gives best rating in value for money section compare to other type of traveling. this is a positive sign of business.



6.4 What is distribution of average ratings of Food-BEV and entertainment given by the passenger in every class?

- In Business Class the average ratings of Food-BEV and entertainment given by passenger is highest.

- In Economy Class the average ratings of Food-BEV and entertainment given by passenger is lowest.

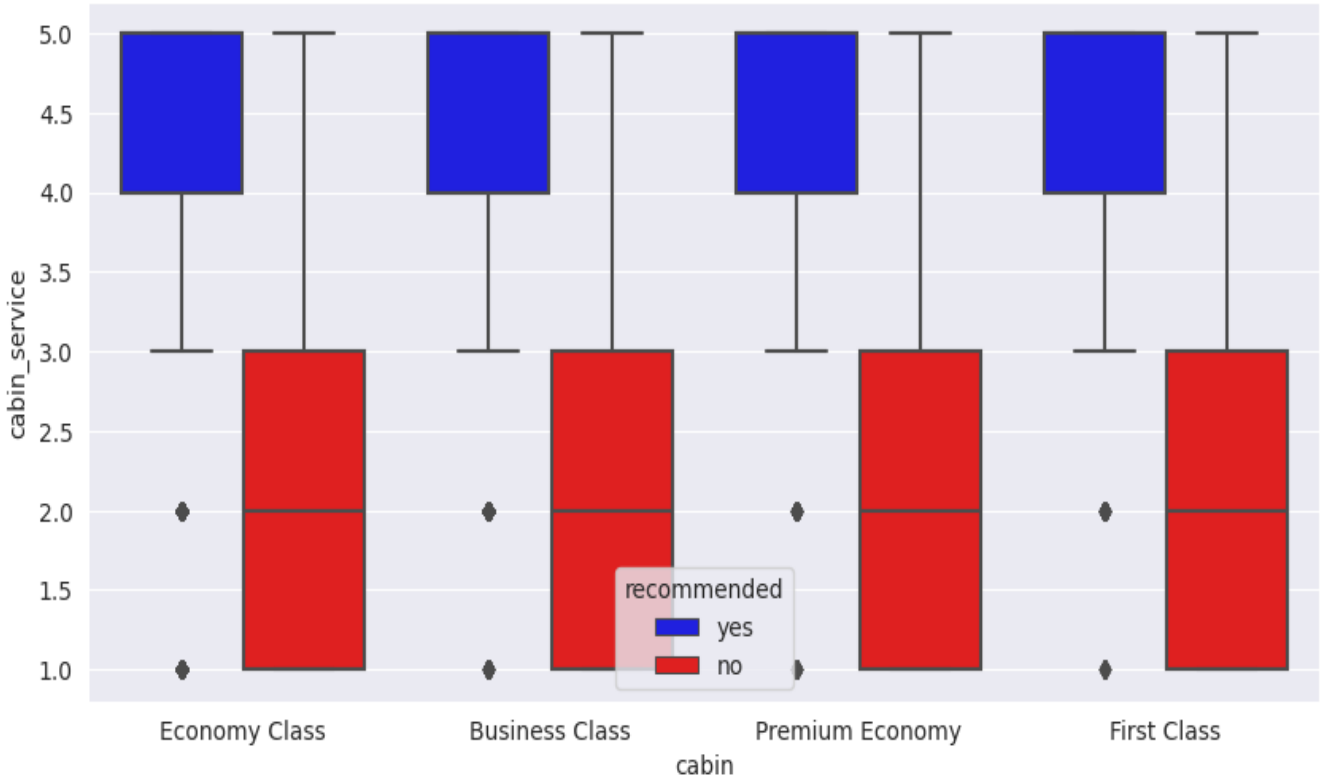


6.5 Which cabin type has more recommended based on ground service ratings?

- First class travelers are least likely to recommend the airlines.

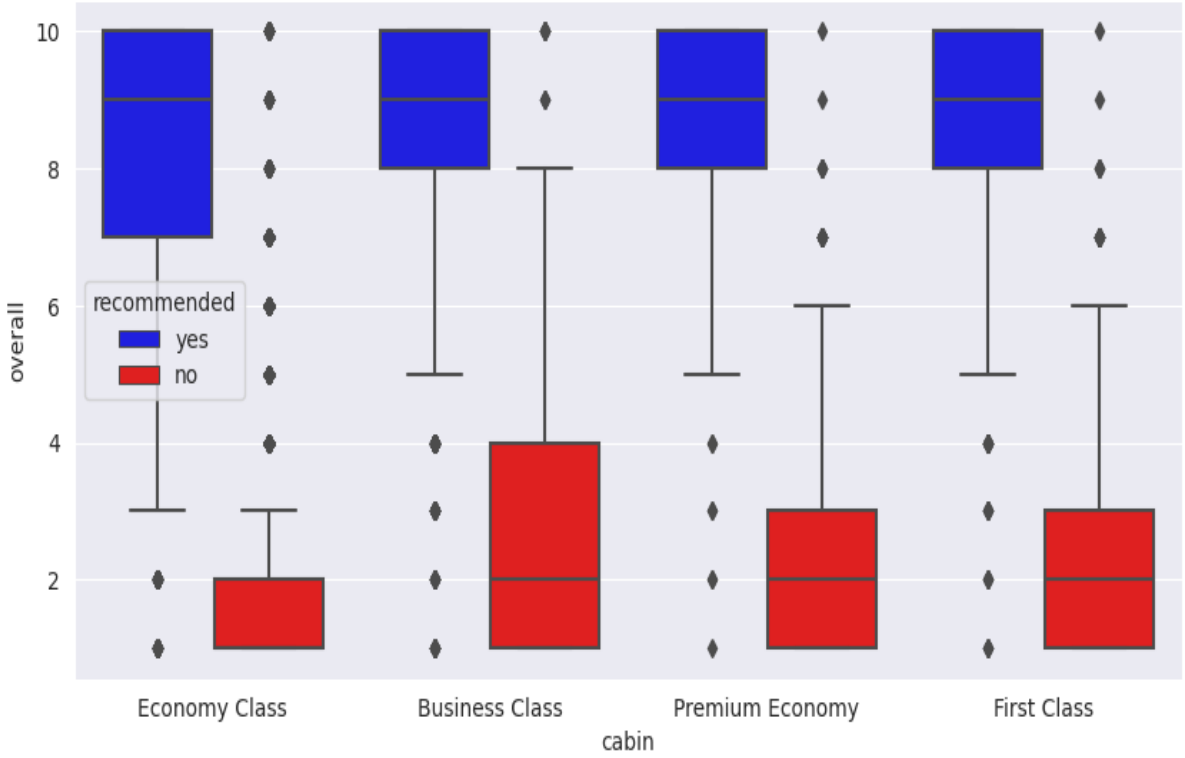
- Recommendation is most probable when the cabin service is given star rating greater than 4.

- In economy class if we got ratings between 4 to 5 that means airlines recommended.



6.6 Which cabin type has more recommended based on overall service ratings by customers?

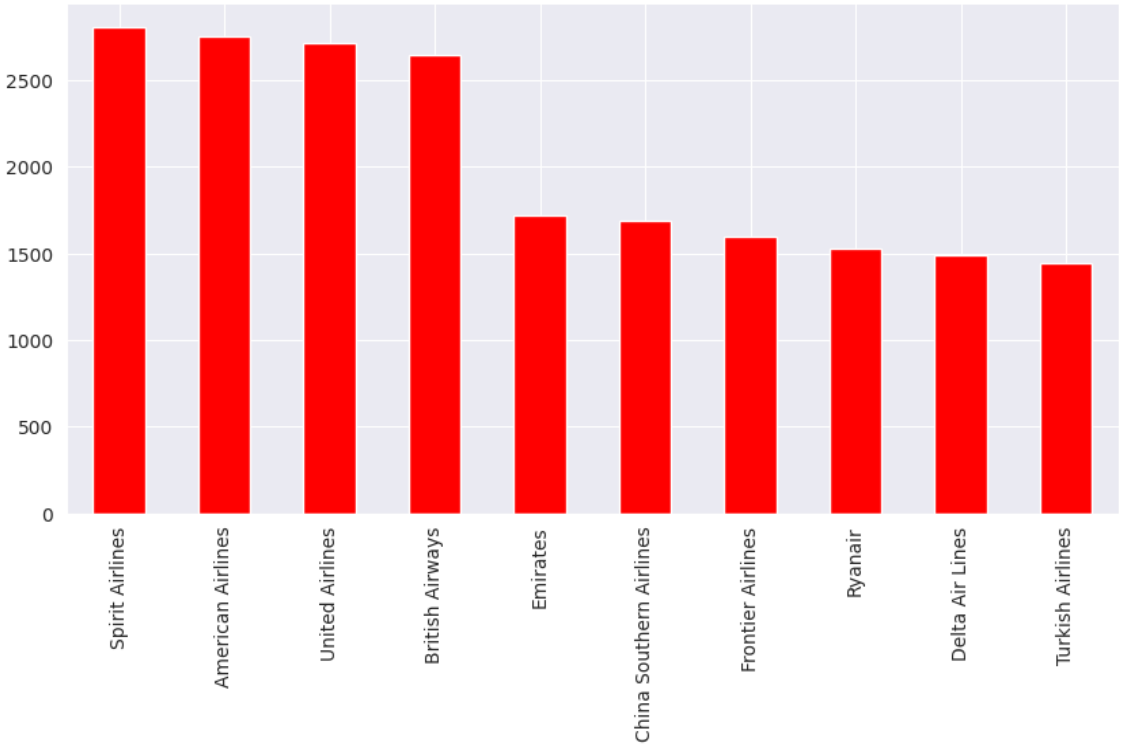
- If the trip is rated above 8 for overall section, the trip is most likely be recommended by the travelers.
- If it is below 3 , the unhappy travelers has not referred the airlines to their friends irrespective of their cabin type.



6.7 Which airline made highest trips?

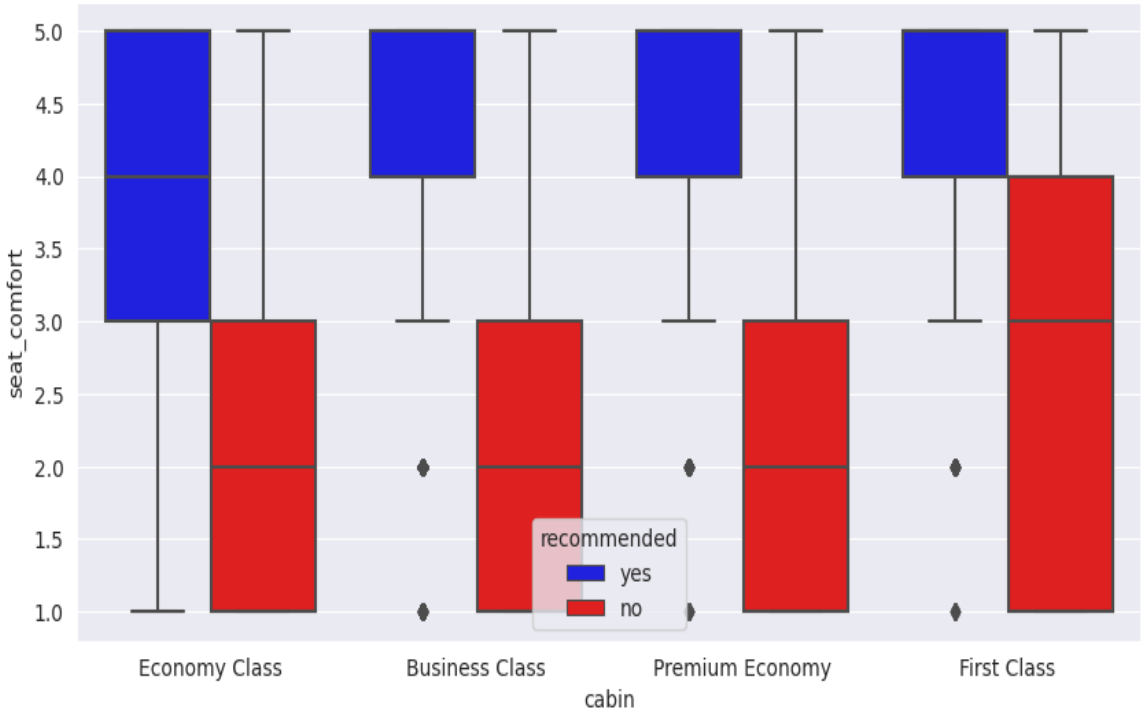
From the above plot we have observed that the top 10 airlines with most trips are-

1. Spirit Airlines
2. American Airlines
3. United Airlines
4. British Airways
5. Emirates
6. China southern airline
7. frontier airlines
8. RYANAIR
9. Delta Airlines
10. Turkish airlines



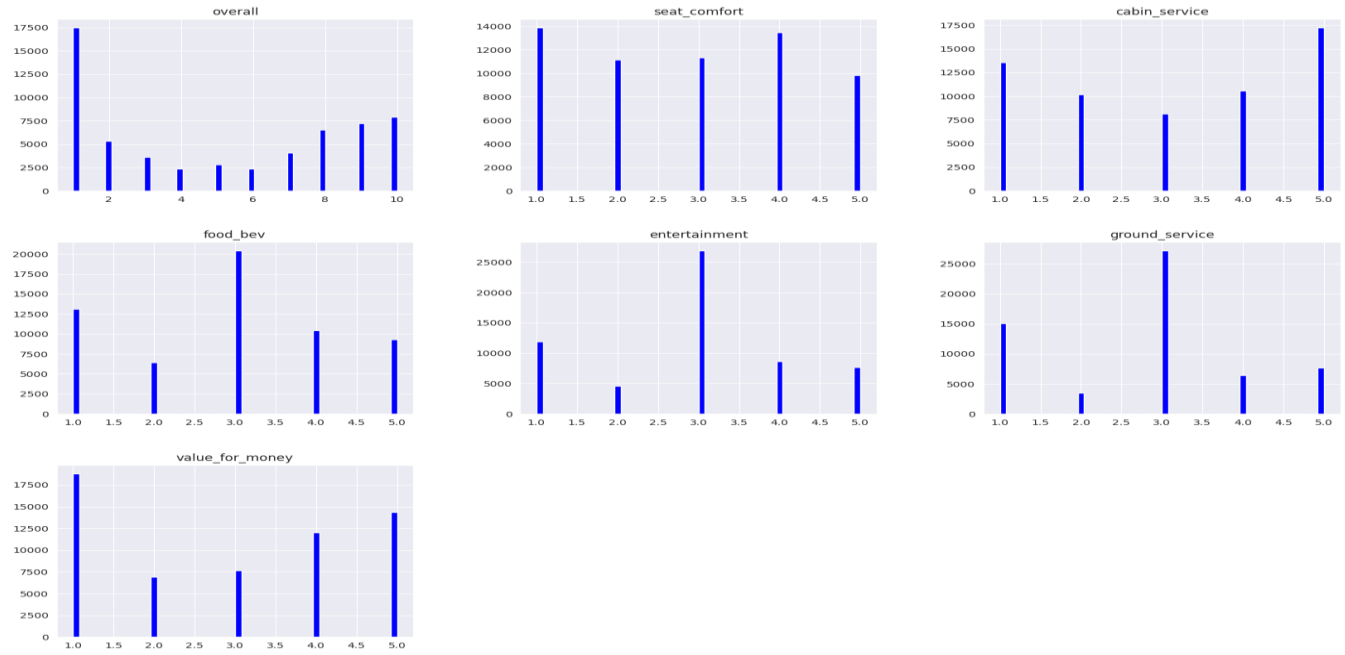
6.8 Which cabin type has more recommended based on overall service ratings by customers?

- If the trip is rated above 3 for seat comfort section, the trip is most likely be recommended by the travelers in Economy class but in Business class, Premium Class or First class
- If the trip is rated above 4 for seat comfort section, the trip is most likely be recommended by the travelers in Business class, Premium Class and the First class.
- If it is below 3, the unhappy travelers has not referred the airlines to their friends irrespective of their cabin type and in the First class section it is below 4.



6.9 Comparison of all independent variable/features?

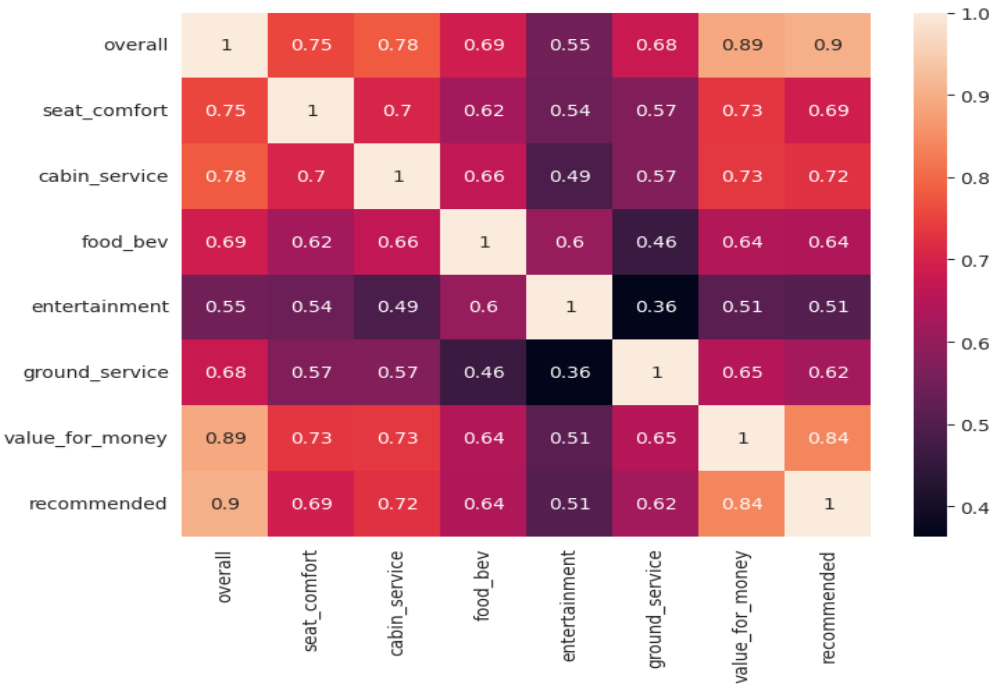
- The overall feature ratings of 1 to 2 occur more frequently.
- From Seat comfort feature, we can say that rating of 1 is highest and rating of 4 is the second highest.
- From cabin service feature, we can say that rating of 5 is highest and rating of 1 is the second highest.
- The food BEV feature ratings of 2,4 and 5 are varies equally which means their frequency are approximately equal.
- The features of both the entertainment & ground service, we can say that ratings of 3 is highest and ratings of 1 is the second highest.
- From value for money feature, it clearly shows that most of the passenger gives ratings of 1 as highest. From this we can say that most of the airline does not provide good service to passenger



7 CORRELATION ANALYSIS

Correlation Heat-map-

- Value for money and overall are highly correlated to each other.
- overall and recommended are highly correlated to each other.
- cabin service and value for money is highly correlated with recommended.
- seat comfort and cabin service are very correlated to value for money.
- value for money and recommended are highly correlated to each other.
- Entertainment and ground service are very low correlated to each other.
- Entertainment and recommended are low correlated to each other



Heatmap of the correlation between all the numerical features

8 Feature Engineering & Data Pre-processing

8.1 Multi-col-linearity techniques:

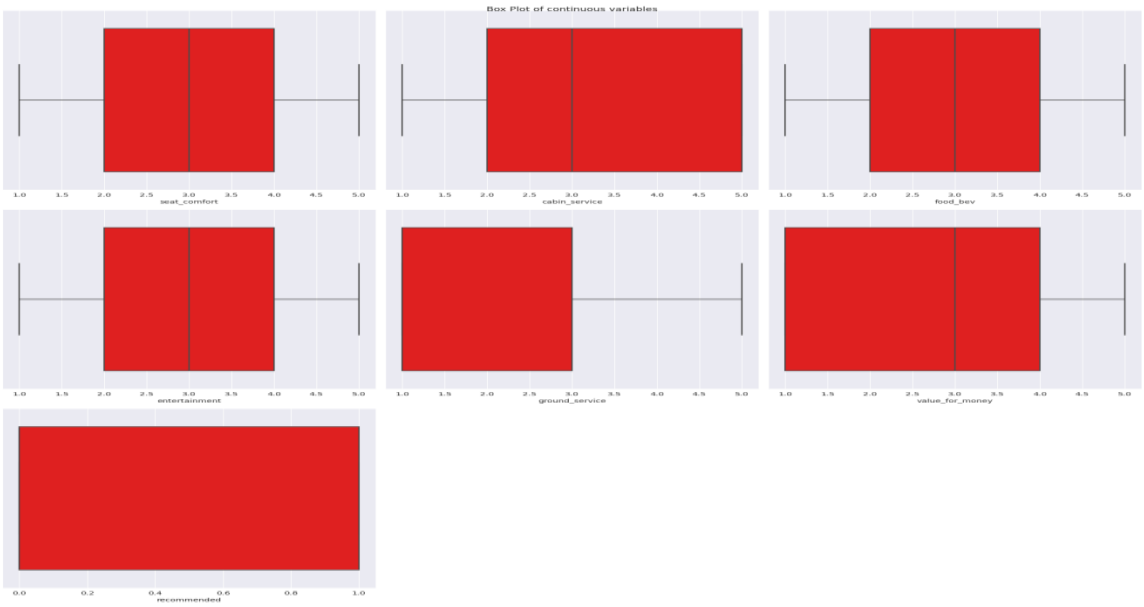
- Here I use VIF for detect multi-col-linearity between variable.
- The variance inflation factor (VIF) directly measures the ratio of the variance of the entire model to the variance of a model with only the feature in question.
- In layman's terms, it gauges how much a feature's inclusion contributes to the overall variance of the coefficients of the features in the model.
- A VIF of 1 indicates that the feature has no correlation with any of the other features.
- Typically, a VIF value exceeding 5 or 10 is deemed to be too high. Any feature with such VIF values is likely to be contributing to multi-col-linearity.

8.2 Feature Selection:

- Here we Drop overall column as it has highest correlation value than others.
- Here we are dropping airline column from our data as it is no use case further

8.3 Handling Outliers:

- As we can see above there are no outliers presents.



8.4 Categorical Encoding:

- The Percentage of No labels of Target Variable are 52.0
- The Percentage of Yes labels of Target Variable are 48.0
- The Percentage of both labels (‘yes’, 'no’) is approximately equal. So no need of Handling Class Imbalance technique
- In here I used Ordinal Encoder on the dataset
- Ordinal Encoder is used when the variables in the data are ordinal, ordinal encoding converts each label into integer values and the encoded data represents the sequence of labels.

- In One-Hot Encoding, each category of any categorical variable gets a new variable. It maps each category with binary numbers (0 or 1). This type of encoding is used when the data is nominal. Newly created binary features can be considered dummy variables. After one hot encoding, the number of dummy variables depends on the number of categories presented in the data.

8.5 Data Splitting:

- X-train and x-test data & y-train and y-test data (47808, 12) (11953, 12)
- The foregoing data splitting methods can be implemented once we specify a splitting ratio. A commonly used ratio is 80:20, which means 80% of the data is for training and 20% for testing which I did in here. Other ratios such as 70:30, 60:40, and even 50:50 are also used in practice. There does not seem to be clear guidance on what ratio is best or optimal for a given dataset. The 80:20 split draws its justification from the well-known Pareto principle, but that is again just a thumb rule used by practitioners.

9 ML Model Implementation

9.1 Fitting Logistic Regression:

Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.

Logistic Regression coefficient values

([[0.29474538, 0.54969246, 0.43542629, 0.2487443, 0.72109819, 1.63391866, -0.19933683, -0.30849387, -0.08268755, -0.21336182, -0.0314013, -0.33690495]])

Logistic Regression intercept values -11.54687704

Logistic Regression score (x-train, y-train) values 0.93898

Logistic Regression score (x-test, y-test) values 0.93240

Evaluation metric:

The accuracy on train data is 0.93898

The accuracy on test data is 0.93240

confusion matrix for both y-train and train predicted classes

[[23379 1664] [1253 21512]]

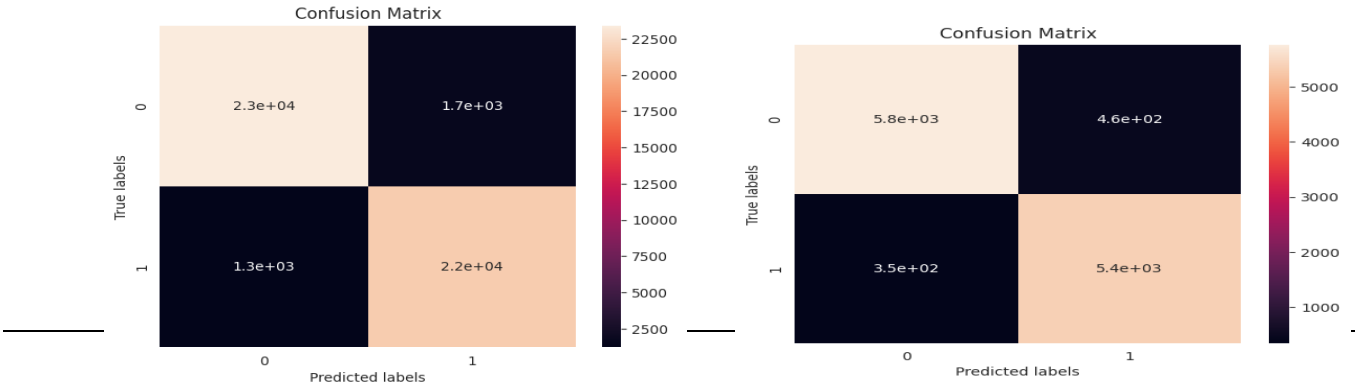
confusion matrix for both y-test and test predicted classes

[[5763 462] [346 5382]]

Cross- Validation & Hyper parameter Tuning

Fitting 3 folds for each of 800 candidates, totaling 2400 fits and Accuracy is 0.938,

Grid-Search-CV Hyper parameter optimization technique for better accuracy score



9.2 Fitting Decision Tree Classifier

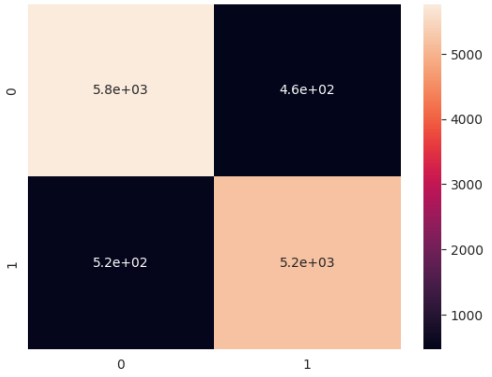
A decision tree is a hierarchical model used in decision support that depicts decisions and their potential outcomes, incorporating chance events, resource expenses, and utility. This algorithmic model utilizes conditional control statements and is non-parametric, supervised learning, useful for both classification and regression tasks. The tree structure is comprised of a root node, branches, internal nodes, and leaf nodes, forming a hierarchical, tree-like structure.

It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

- Root Nodes – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.
- Decision Nodes – the nodes we get after splitting the root nodes are called Decision Node
- Leaf Nodes – the nodes where further splitting is not possible are called leaf nodes or terminal nodes
- Sub-tree – just like a small portion of a graph is called sub-graph similarly a sub-section of this decision tree is called sub-tree.
- Pruning – is nothing but cutting down some nodes to stop overfitting.

Decision classifies score (x-train, y-train) values 0.97454

Decision classifies score (x-test, y-test) values 0.91734



	precision	recall	f1-score	support
0	0.92	0.93	0.92	6225
1	0.92	0.91	0.91	5728
accuracy			0.92	11953
macro average	0.92	0.92	0.92	11953
weighted average	0.92	0.92	0.92	11953

Cross- Validation & Hyper parameter Tuning-

Decision-Tree-Classifier best parameters

{'criterion': GINI, 'max-depth': 7, 'min-samples-leaf': 3, 'min-samples-split': 5}

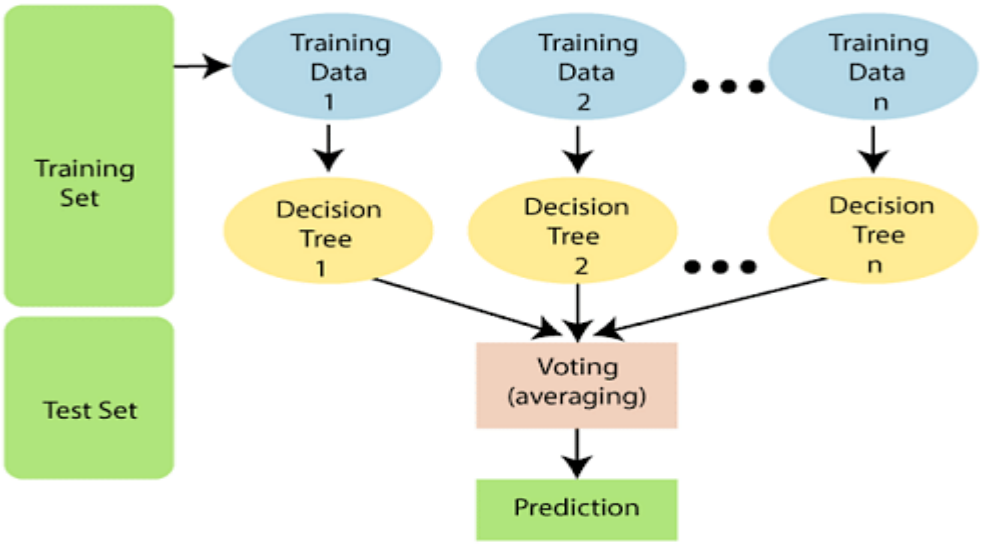
Decision-Tree-Classifier best scoring 0.93716

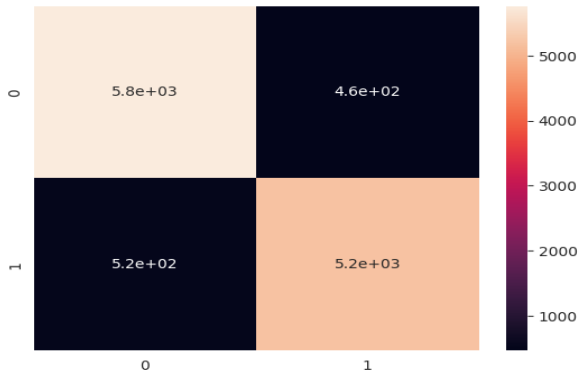
Here our model is Over fitted. So Hyper parameter tuning is done to prune a Decision tree to preserve Generalized Model.

94% accuracy of Decision Tree with the help of hyper parameter tuning.

9.3 Fitting Random Forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.





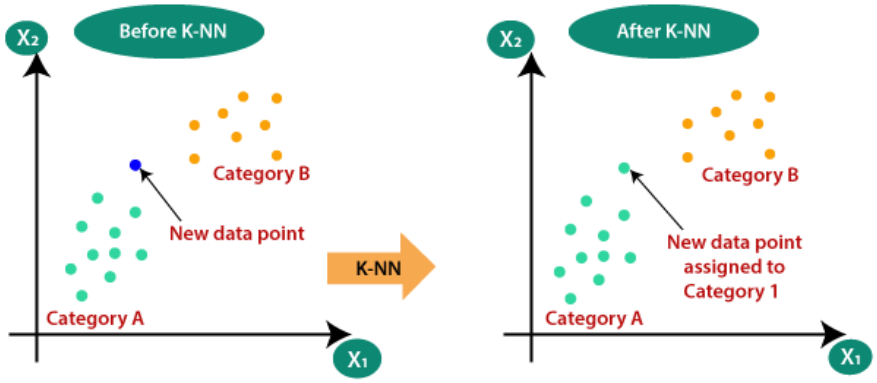
	precision	recall	f1-score	support
0	0.92	0.93	0.92	6225
1	0.92	0.91	0.91	5728
accuracy			0.92	11953
macro average	0.92	0.92	0.92	11953
weighted average	0.92	0.92	0.92	11953

Cross- Validation & Hyper parameter Tuning

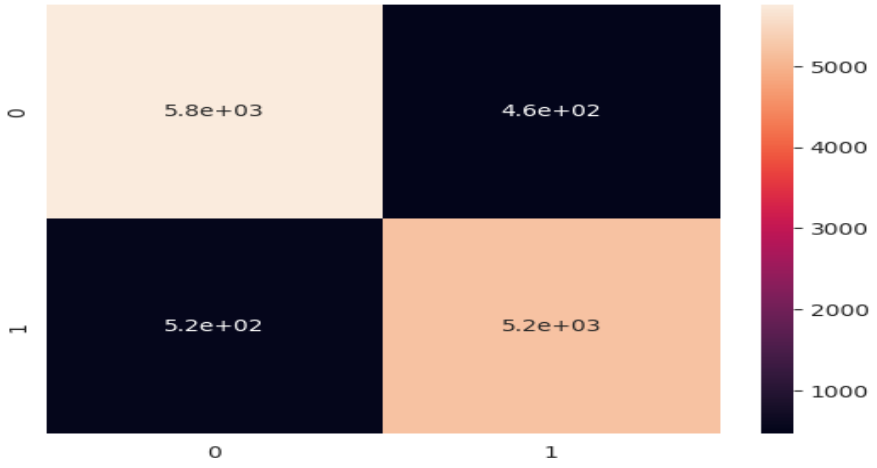
- Random Forest Grid-Search-CV best score values 0.94019
- used Grid-Search-CV for improve the accuracy, After Hypermeter tuning we get a better score and the accuracy is 94%.

9.4 K-Nearest Neighbor

The k-nearest neighbors' algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.



- **Train accuracy**- 95%
- **Test accuracy**- 93%
- **confusion matrix of k-neighbor y-test or y-predict values** [[5762 463] [525 5203]]
- **Area under ROC curve score y-test or y-predict values** 0.91698



- **Hyper parameter Tuning**
- K-Nearest Neighbor Grid-Search-CV uses after best score values 0.93805
- We see here Hypermeter tuning we get a better accuracy score and the accuracy is 94%.

9.5 Naive Bayes Classifier

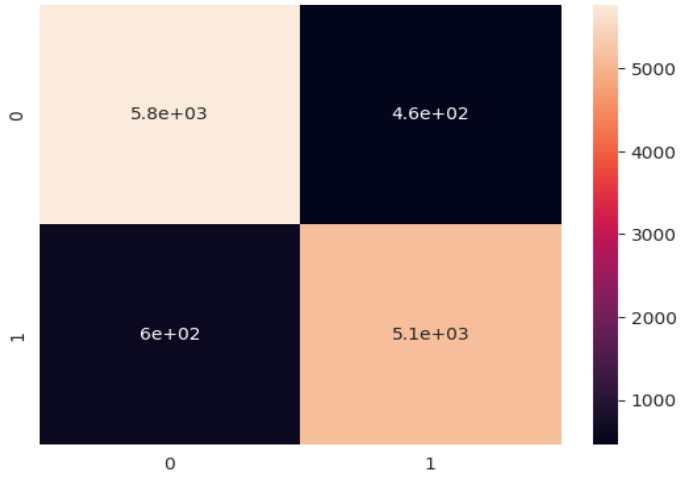
- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the 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.
- Some popular examples of Naive Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.
- Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:
-

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Where

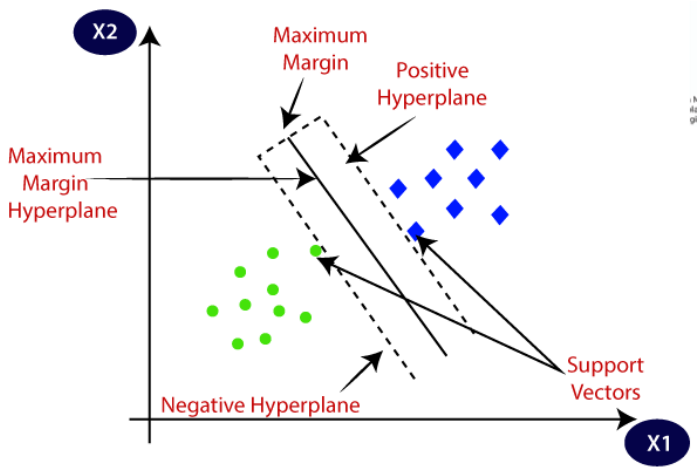
- P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
- P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
- P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
- P(B) is Marginal Probability: Probability of Evidence.

- **Gaussian Naïve Bayes Model (y-test, y-predict) accuracy score** 0.91182
- **Confusion Matrix** [[5767 458] [596 5132]]
- 91% accuracy with Naïve Bayes Classifies

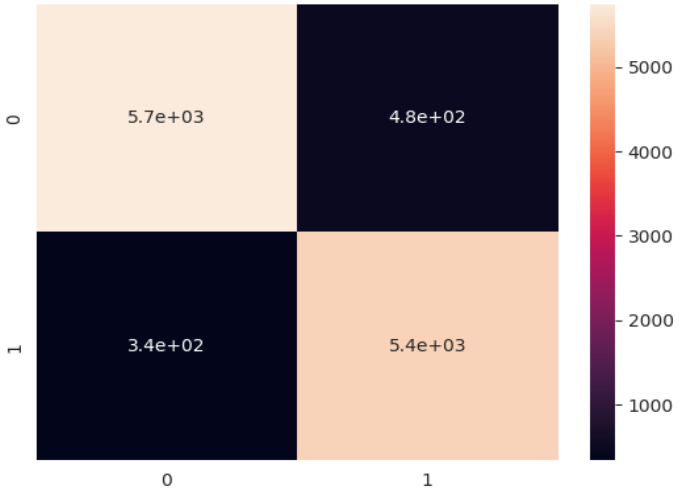


9.6 Support Vector Machine

- 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.
- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



- **Support vector (x-test, y-test) score values** 0.93181
- **Support vector confusion matrix values** [[5746 479] [336 5392]]
- 93% accuracy with support vector machine

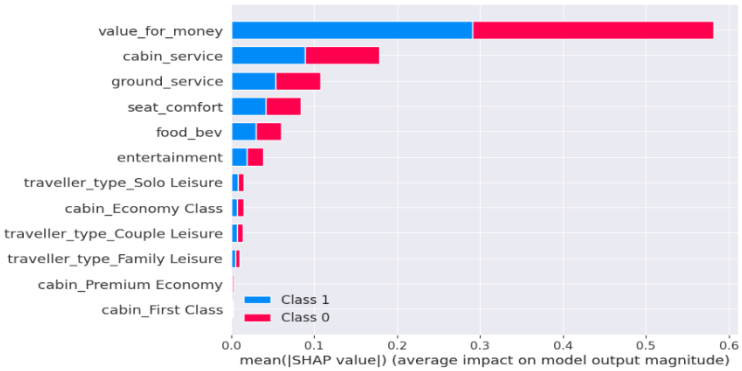


Which ML model did you choose from the above created models as your final prediction model and why?

```
----------Logistic Regression Model----------

          Metrics  Train_Score  Test_Score
0  Accuracy_Score     0.938985    0.932402
1  Precsion_Score     0.928202    0.920945
2    Recall_Score     0.944959    0.939595
3   Roc_Auc_Score     0.939257    0.932689

----------Decision Tree Model After Hyperparameter Tuning----------

          Metrics  Train_Score  Test_Score
0  Accuracy_Score     0.940826    0.933406
1  Precsion_Score     0.934298    0.926939
2    Recall_Score     0.941972    0.934707
3   Roc_Auc_Score     0.940878    0.933458

----------Random Forest Model After Hyperparameter Tuning----------

          Metrics  Train_Score  Test_Score
0  Accuracy_Score     0.941349    0.934075
1  Precsion_Score     0.941713    0.932878
2    Recall_Score     0.934680    0.929295
3   Roc_Auc_Score     0.941045    0.933884

----------Knn Model After Hyperparameter Tuning----------

          Metrics  Train_Score  Test_Score
0  Accuracy_Score     0.943064    0.939178
1  Precsion_Score     0.940564    0.937074
2    Recall_Score     0.939820    0.935929
3   Roc_Auc_Score     0.942916    0.939049

----------Naive BayesModel----------

          Metrics  Train_Score  Test_Score
0  Accuracy_Score     0.918026    0.911821
1  Precsion_Score     0.926264    0.918068
2    Recall_Score     0.899451    0.895950
3   Roc_Auc_Score     0.917181    0.911188

----------Support vector model----------

          Metrics  Train_Score  Test_Score
0  Accuracy_Score     0.936852    0.931816
1  Precsion_Score     0.923080    0.918413
2    Recall_Score     0.946233    0.941341
3   Roc_Auc_Score     0.937278    0.932196
```

- From the above snap shot, we can clearly see that for the accuracy and roc AUC score is improved for K-Nearest Neighbors. So, I have chosen K-Nearest Neighbors as the final prediction model which should be deployed for real user interaction.
- All the others model are also performed well in this dataset but I chose K-Nearest Neighbors model.

Explain the model which you have used and the feature importance using any model explain-ability tool?



- This plot is made of all the dots in the train data. It demonstrates the following information:
- Feature importance: Variables are ranked in descending order.
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.
- Red indicates a higher and blue indicates a lower. From the X-axis we can verify the impact (Positive or Negative) for that specific data.

- A variable importance plot lists the most significant variables in descending order. The top variables contribute more to the model than the bottom ones and thus have high predictive power.
- Here, we can see the feature importance for respective classes in a descending order
- The most important feature are overall rating and Value for money that contribute to a model's prediction whether a passenger will recommend a particular airline to his/her friends.

10 Conclusion

- The Models used for this Classification problem are
    1. Logistic Regression Model
    2. Decision Tree Model
    3. Random Forest Model
    4. K-Nearest Neighbor Model
    5. Naive Bayes
    6. Support vector Machine Model
- We performed Hyper parameter tuning using Grid-search CV method for Decision Tree Model, Random Forest Model, K-Nearest Neighbor, Support Vector Machine and Naive Bayes. To increase accuracy and avoid Over fitting Criteria.
- Based on the knowledge of the business and the problem use case. The Classification metrics of Recall is given first priority, Accuracy is given second priority, and ROC AUC is given third priority.
- We have built classifier models using 6 different types of classifiers and all these are able to give accuracy of more than 90%. We can conclude that Decision Tree gives the best model.
- model evaluation metrics comparison, we can see that Support Vector Machine being the model with highest accuracy rate by a very small margin, works best among the experimented models for the given dataset.
- The most important feature are overall rating and Value for money that contribute to a model's prediction whether a passenger will recommend a particular airline to his/her friends.
- The classifier models developed can be used to predict passenger referral as it will give airlines ability to identify impactful passengers who can help in bringing more revenues.
- As a result, in order to increase their business or grow, our client must provide excellent cabin service, ground service, food beverage entertainment, and seat comfort.