# CAPSTONE PROJECT - 3

## Airline Passenger Referral Prediction Classification

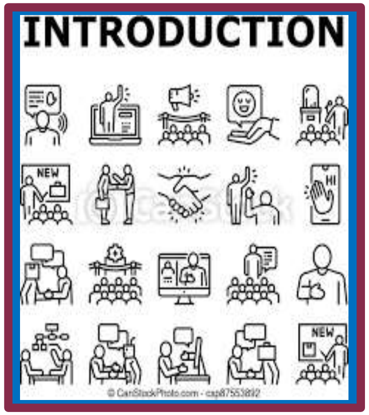
Submitted by
SANJU KHANRA
Data science trainee, Alma better

# CONTENT :

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

# 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.



# PROBLEM STATEMENT

- Which Traveller-type has more ratings?
- Which type of Cabin has more recommendation?
- Which type of traveller type is value for money?
- What is distribution of average ratings of Food-BEV and entertainment given by the passenger in every class?
- Which cabin type has more recommended based on ground service ratings?
- Which cabin type has more recommended based on overall service ratings by customers?
- Which airline made highest trips?
- Which cabin type has more recommended based on overall service ratings by customers?
- Comparison of all independent variable/features?

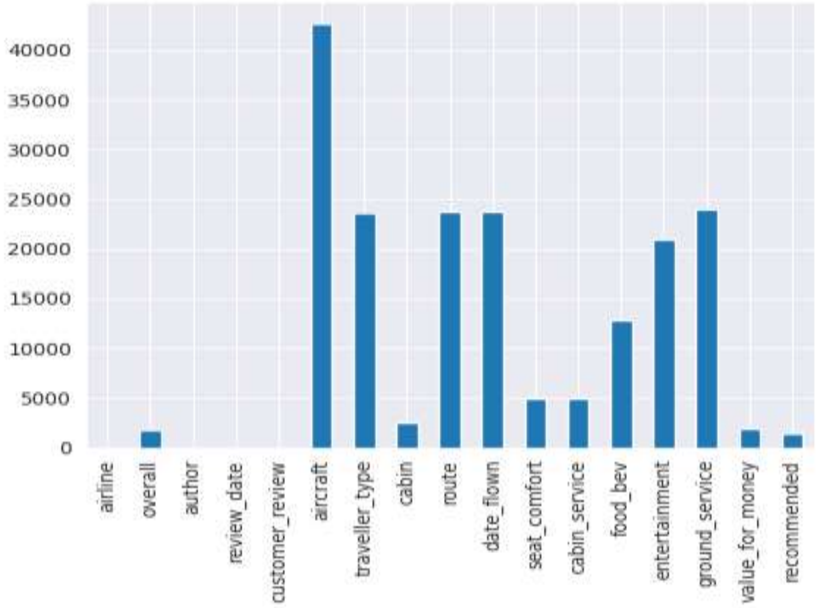
# DATA DESCRIPTION

- **airline:** Name of the airline.
- **overall:** Overall point is given to the trip between 1 to 10.
- **author:** Author of the trip
- **review date:** Date of the Review
- **customer review :** Review of the customers in free text format
- **aircraft:** Type of the aircraft
- **traveller 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.

# DATA CLEANING

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

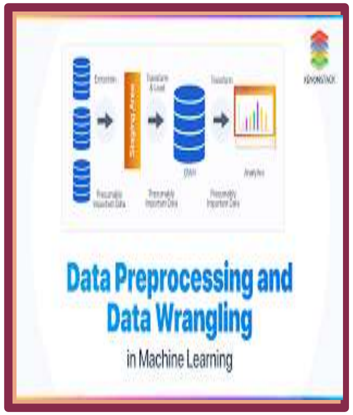
Missing values/Null Values count



| airline | 1 |
|---|---|
| overall | 1783 |
| author | 1 |
| Review-date | 1 |
| Customer-review | 1 |
| aircraft | 42696 |
| Traveller-type | 23644 |
| cabin | 2479 |
| route | 23671 |
| Date-flown | 23750 |
| Seat-comfort | 4973 |
| Cabin-service | 4944 |
| Food-BEV | 12843 |
| entertainment | 20954 |
| Ground-service | 24015 |
| Value-for-money | 1857 |
| recommended | 1423 |

# DATA WRANGLING CODE

- Percentage wise missing values checking after Dropping the aircraft column from data as it have 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 traveller-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



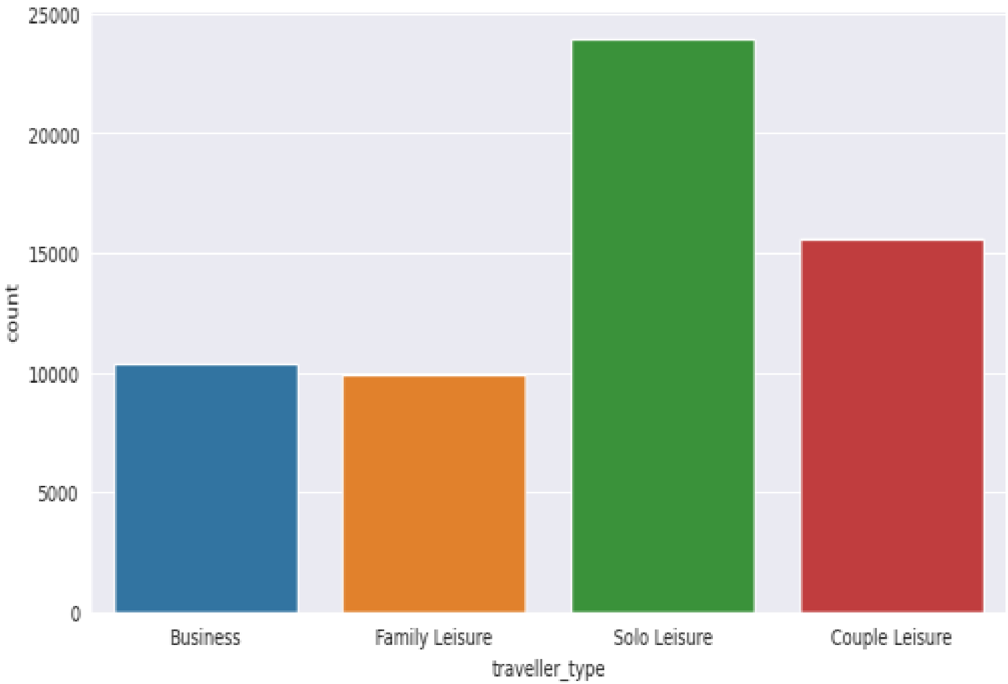
- **Reason of dropping columns–**

- 1. Author - Being the categorical with high Variability not required for prediction.
- 2. Route - Not needed for building a model as it is independent of the Services and Quality of travel.
- 3. 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.
- 4. Review-date - Similar to Date-flown
- 5. 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

# EXPLORATORY DATA ANALYSIS(EDA)

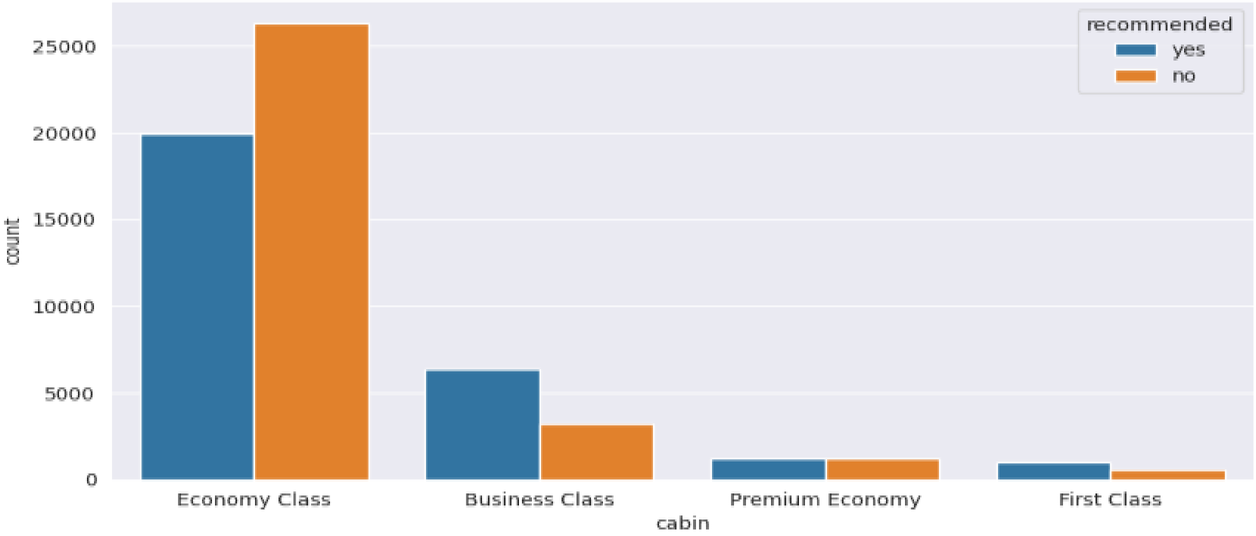
## ❑ Which Traveller-type has more ratings ?

- 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.



# EXPLORATORY DATA ANALYSIS(EDA)

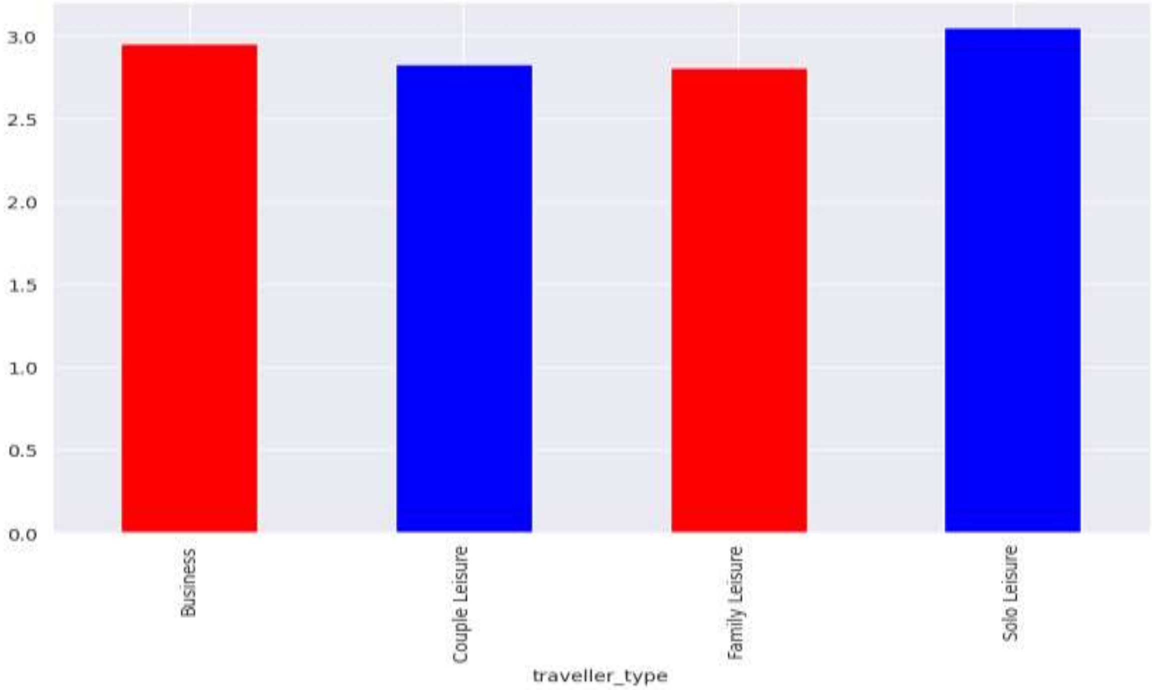
- Which type of Cabin has more recommendation?
- **On the basis of above graph -**
- Economy class has highest recommendation with bad reviews.
- Business class has second most recommended cabin type with good reviews.
- premium economy has equal reviews.
- first class is least recommend cabin type with 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.



# EXPLORATORY DATA ANALYSIS(EDA)

## ❑ Which type of traveller type is value for money?

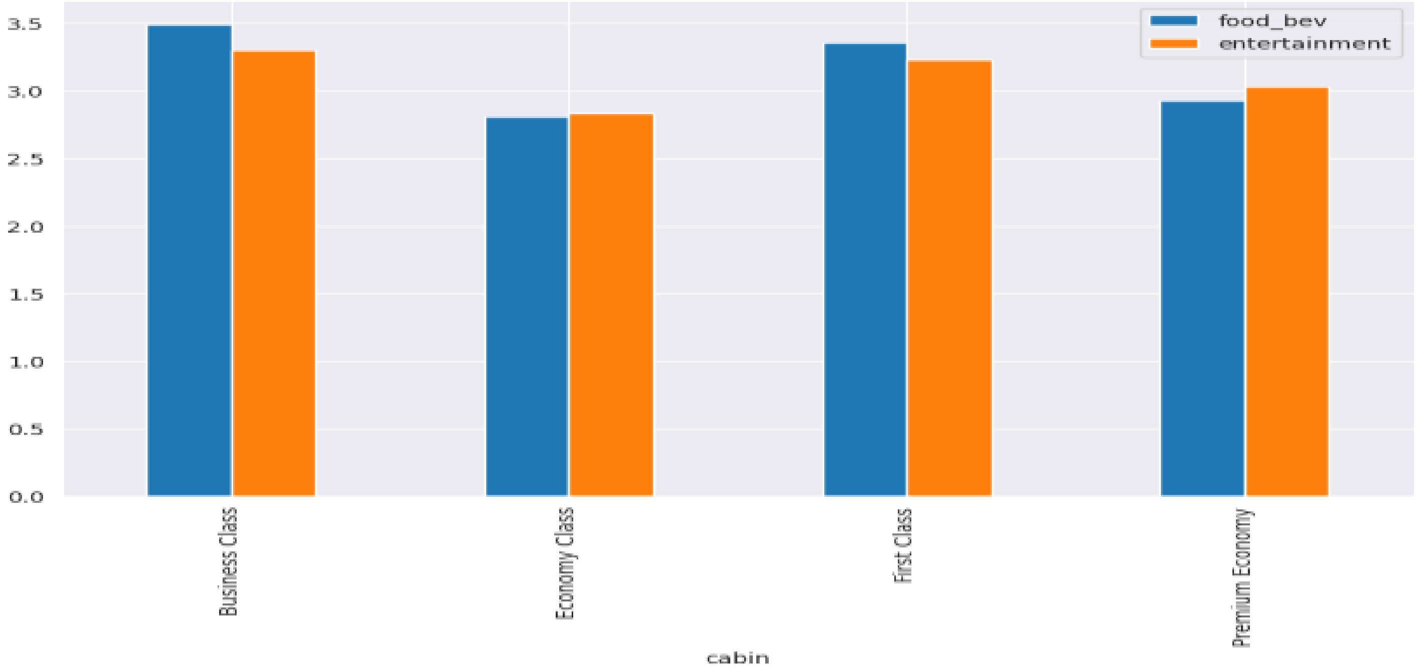
- We can say that travelling Type of Solo Leisure worth of Money compare to other type of travelling.
- Solo Leisure traveling type gives best rating in value for money section compare to other type of traveling. this is a positive sign of business.



# EXPLORATORY DATA ANALYSIS(EDA)

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

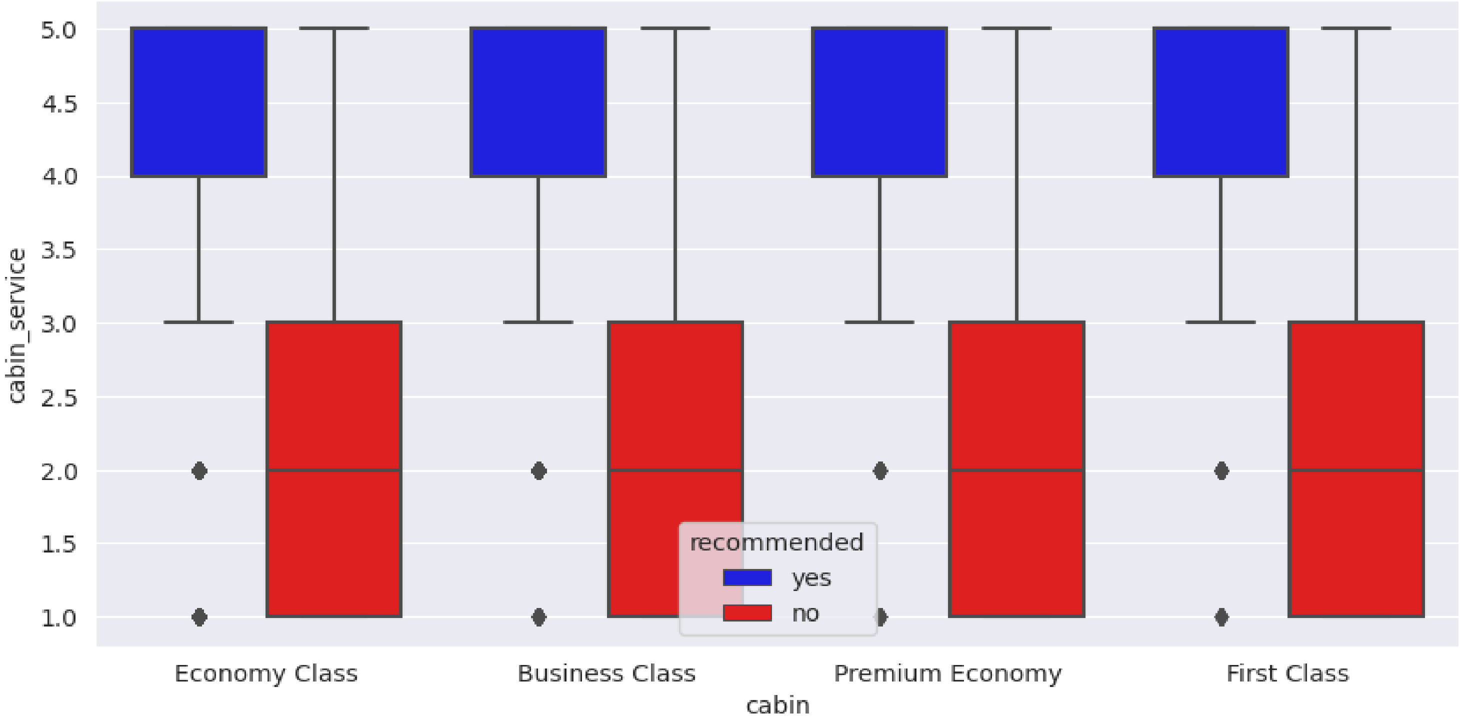
- In Business Class the average ratings of Food-BEV and entertainment given by passenger is highest.
- In Economy Class the average ratings of Food-BEV and entertainment given by passenger is lowest.



# EXPLORATORY DATA ANALYSIS(EDA)

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

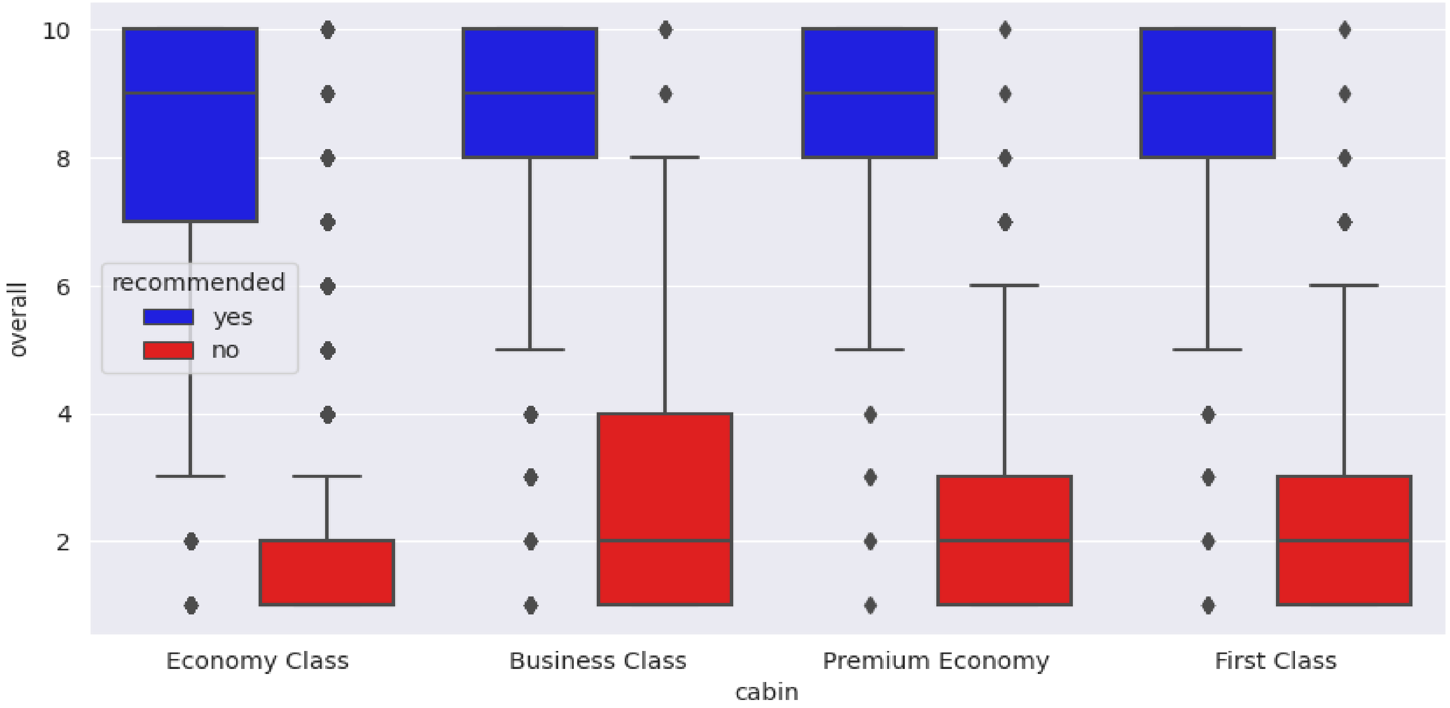
- First class travellers are least likely to recommend the airlines.
- Recommendation is most probable when the cabin service is given star rating greater then 4.
- In economy class if we got ratings between 4 to 5 that means airlines recommended.



# EXPLORATORY DATA ANALYSIS(EDA)

## ❑ 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 travellers.
- If it is below 3 , the unhappy travellers has not referred the airlines to their friends irrespective of their cabin type.

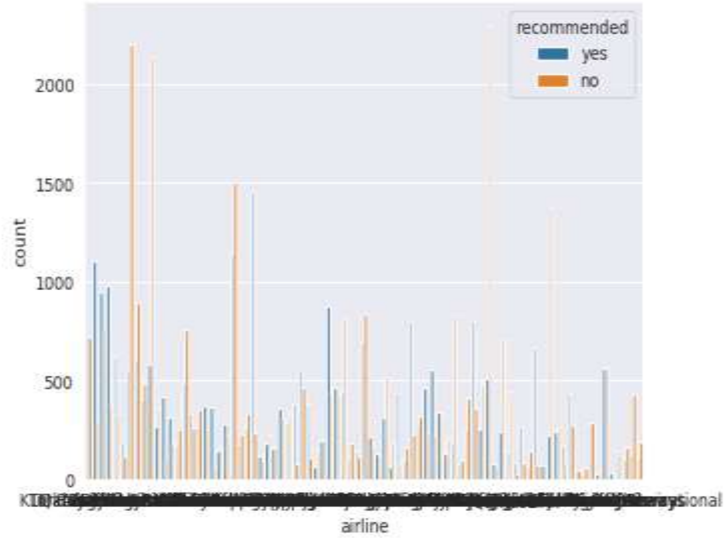


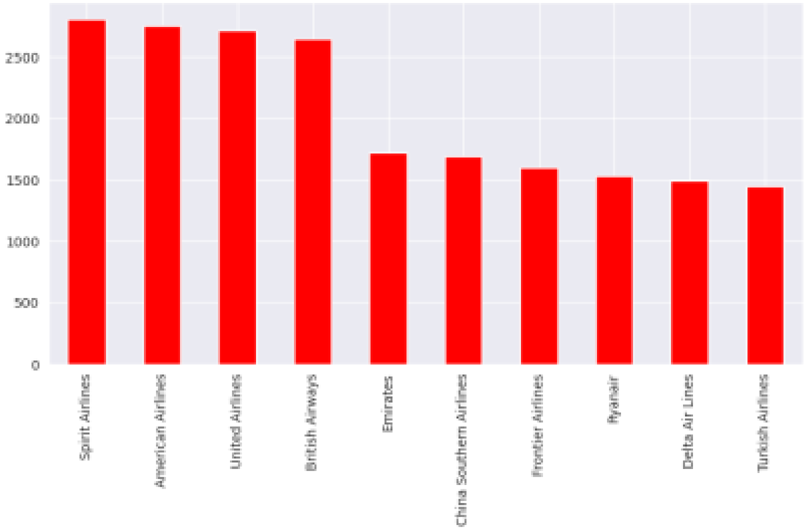
# EXPLORATORY DATA ANALYSIS(EDA)

- 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 air lines 10. Turkish airlines

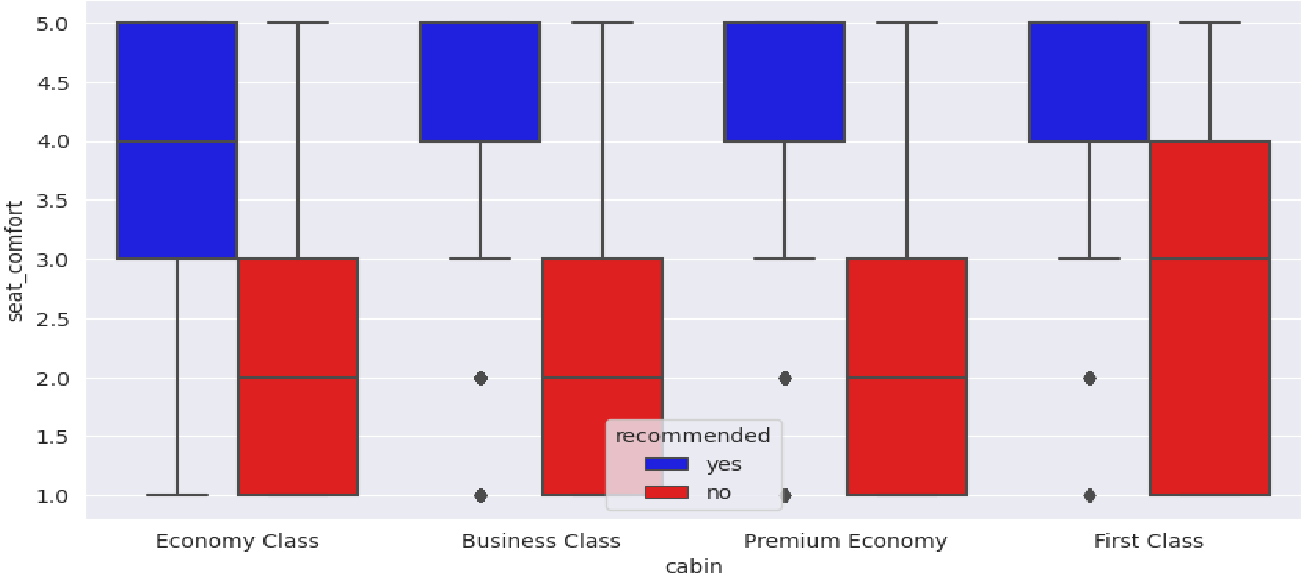




# EXPLORATORY DATA ANALYSIS(EDA)

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

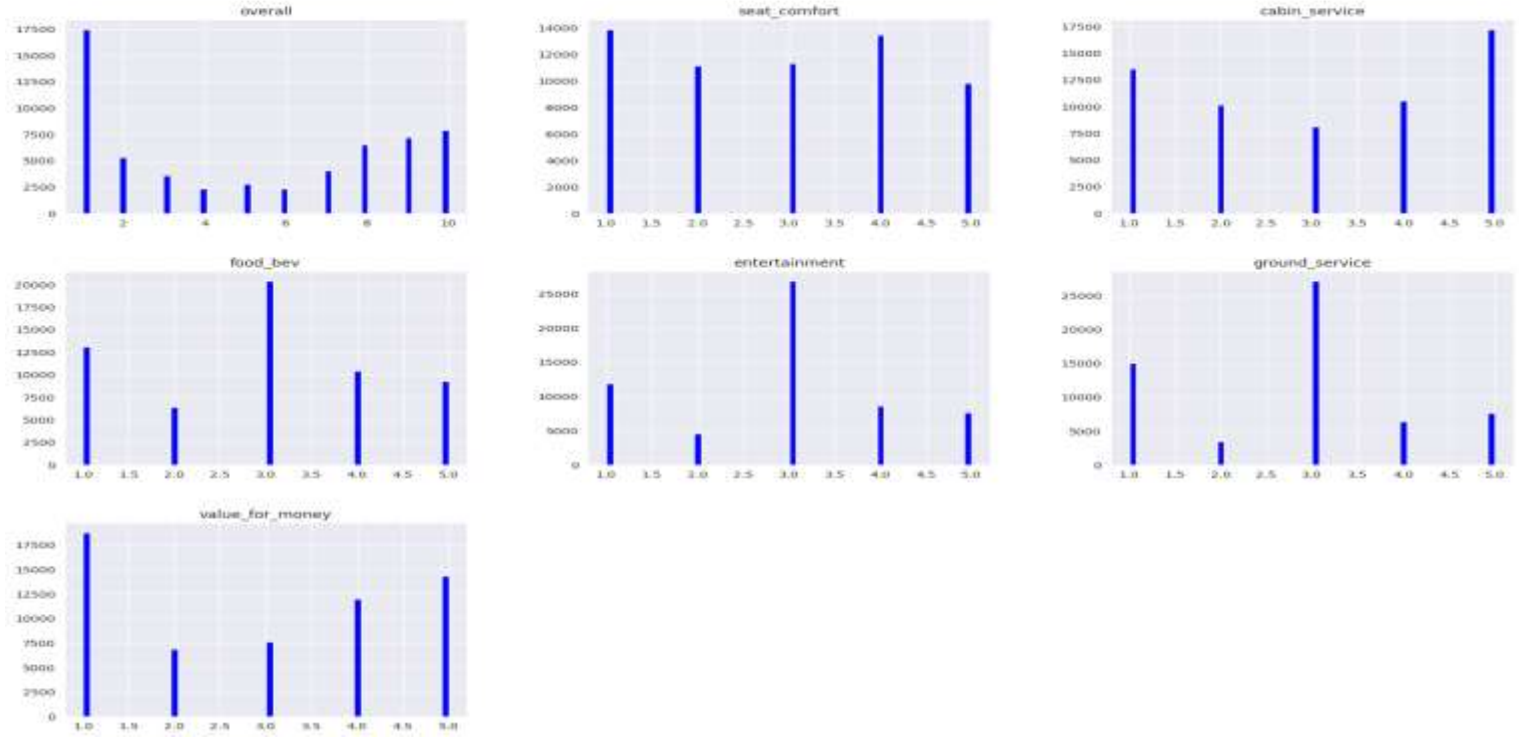
- If the trip is rated above 3 for seat comfort section, the trip is most likely be recommended by the travellers in Economy class but in Business class , Premium Class or First class
- If the trip is rated above 4 for seat comfort section, the trip is most likely be recommended by the travellers in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



# EXPLORATORY DATA ANALYSIS(EDA)

## ❑ Comparison of all independent variable/features?

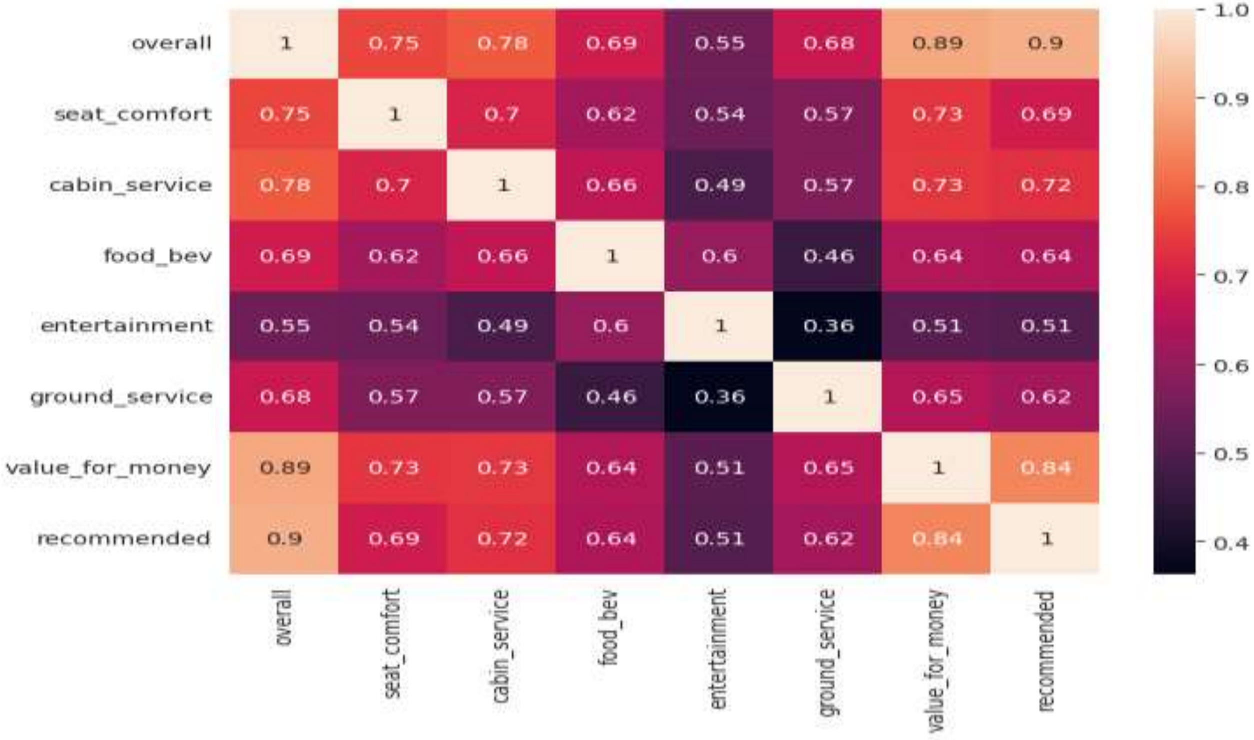
- The overall feature ratings of 1 to 2 occur more frequently.
- From Seat comfort feature, We can say that rating of 1 is highest and rating of 4 is the second highest.
- From cabin service feature, We can say that rating of 5 is highest and rating of 1 is the second highest.
- The food BEV feature 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.



# EXPLORATORY DATA ANALYSIS(EDA)

## ❑ 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



| | overall | seat_comfort | cabin_service | food_bev | entertainment | ground_service | value_for_money | recommended |
|---|---|---|---|---|---|---|---|---|
| overall | 1 | 0.75 | 0.78 | 0.69 | 0.55 | 0.68 | 0.89 | 0.9 |
| seat_comfort | 0.75 | 1 | 0.7 | 0.62 | 0.54 | 0.57 | 0.73 | 0.69 |
| cabin_service | 0.78 | 0.7 | 1 | 0.66 | 0.49 | 0.57 | 0.73 | 0.72 |
| food_bev | 0.69 | 0.62 | 0.66 | 1 | 0.6 | 0.46 | 0.64 | 0.64 |
| entertainment | 0.55 | 0.54 | 0.49 | 0.6 | 1 | 0.36 | 0.51 | 0.51 |
| ground_service | 0.68 | 0.57 | 0.57 | 0.46 | 0.36 | 1 | 0.65 | 0.62 |
| value_for_money | 0.89 | 0.73 | 0.73 | 0.64 | 0.51 | 0.65 | 1 | 0.84 |
| recommended | 0.9 | 0.69 | 0.72 | 0.64 | 0.51 | 0.62 | 0.84 | 1 |

# FEATURE ENGINEERING & DATA PRE-PROCESSING

## ❑ 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.

## ❑ 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

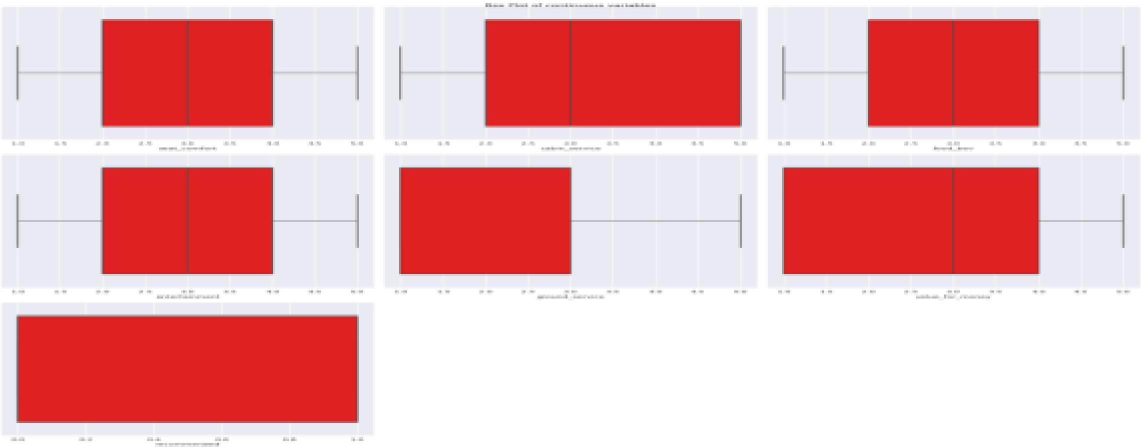
## ❑ Data Splitting :

- 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.

# FEATURE ENGINEERING & DATA PRE-PROCESSING

## Handling Outliers :

- As we can see above there are no outliers presents



## Categorical Encoding :

The Percentage of No labels of Target Variable is 52.0

The Percentage of Yes labels of Target Variable is 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.

# ML MODEL IMPLEMENTATION

- Fitting Logistic Regression :

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

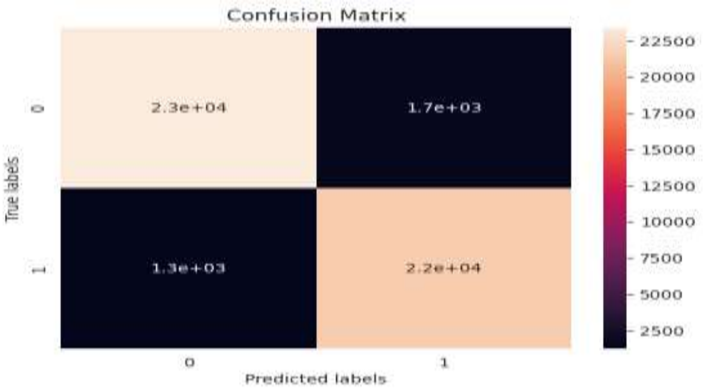
The accuracy on test data is 0.93240

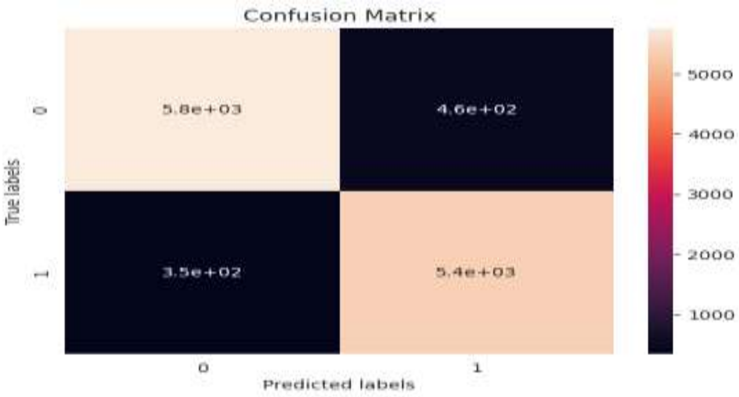
confusion matrix for both y-train and train predicted classes

[[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,

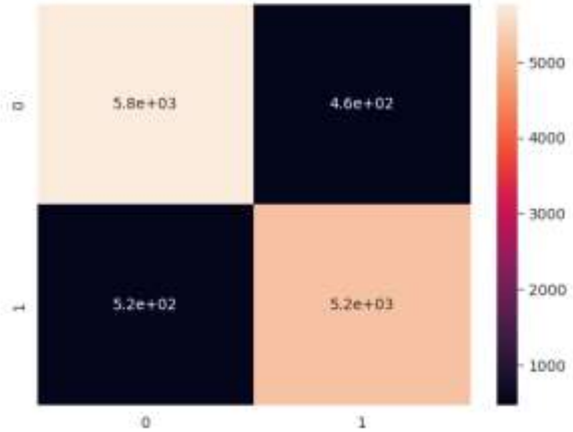
Grid-Search-CV Hyper parameter optimization technique for better accuracy score

# ML MODEL IMPLEMENTATION

## Fitting Decision Tree Classifier

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

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



### Decision Tree Report

| | 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

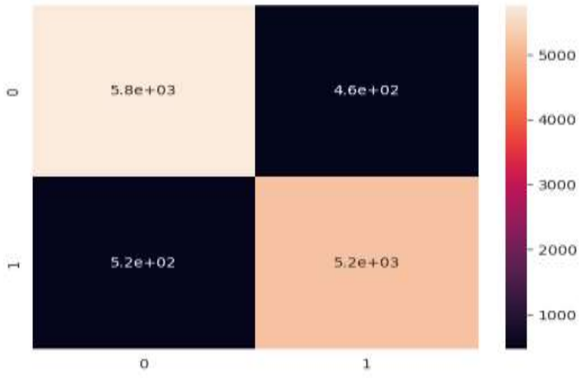
Here our model is Over fitted. So Hyper parameter tuning is done to prune a Decision tree to preserve Generalized Model.

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

# ML MODEL IMPLEMENTATION

## ❑ Fitting Random Forest

- 92% accuracy with Random Forest



| | 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%.

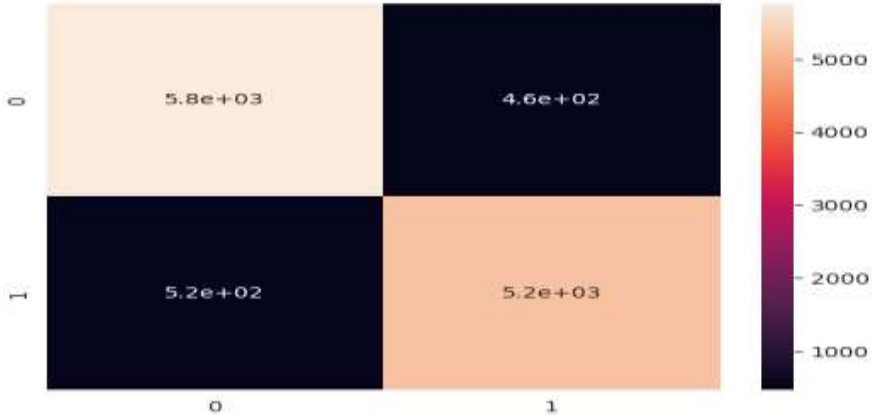
# ML MODEL IMPLEMENTATION

## K-Nearest Neighbor

- 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

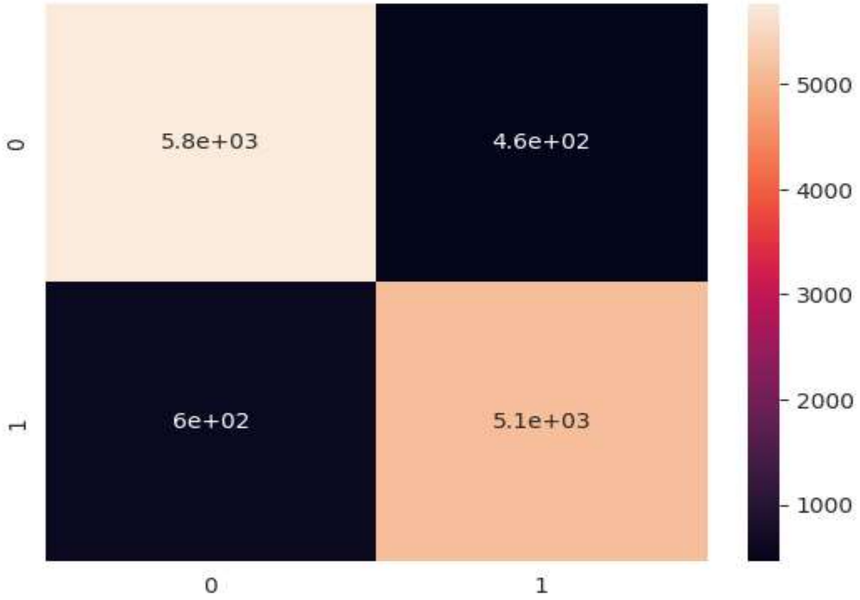
K-Nearest Neighbor Grid-Search-CV uses after best score values 0.93805

We see here Hypermeter tuning we get a better accuracy score and the accuracy is 94%.

# ML MODEL IMPLEMENTATION

## ❑ Naive Bayes Classifier

- Gaussian Naive Bayes model (y-test, y-predict) accuracy score 0.91182
- confusion matrix [[5767 458] [ 596 5132]]
- 91% accuracy with Naive Bayes classifier



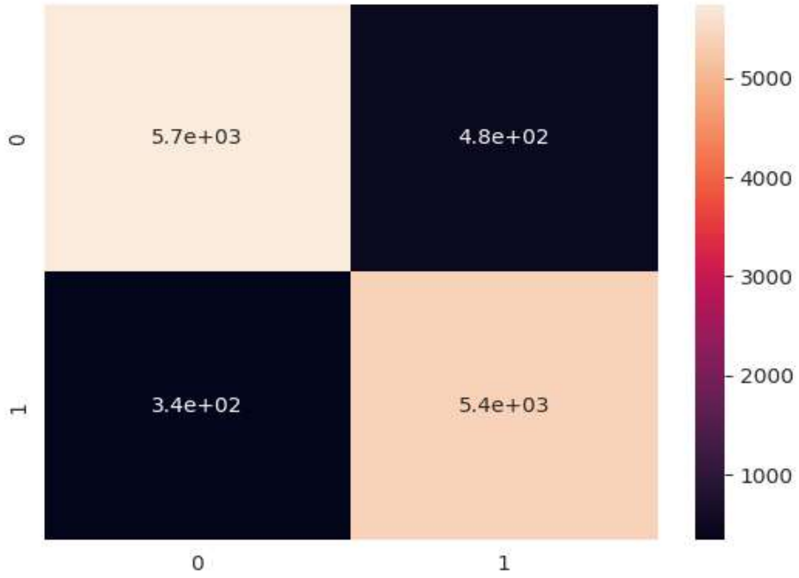
# ML MODEL IMPLEMENTATION

## Support Vector Machine

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



# ML MODEL IMPLEMENTATION

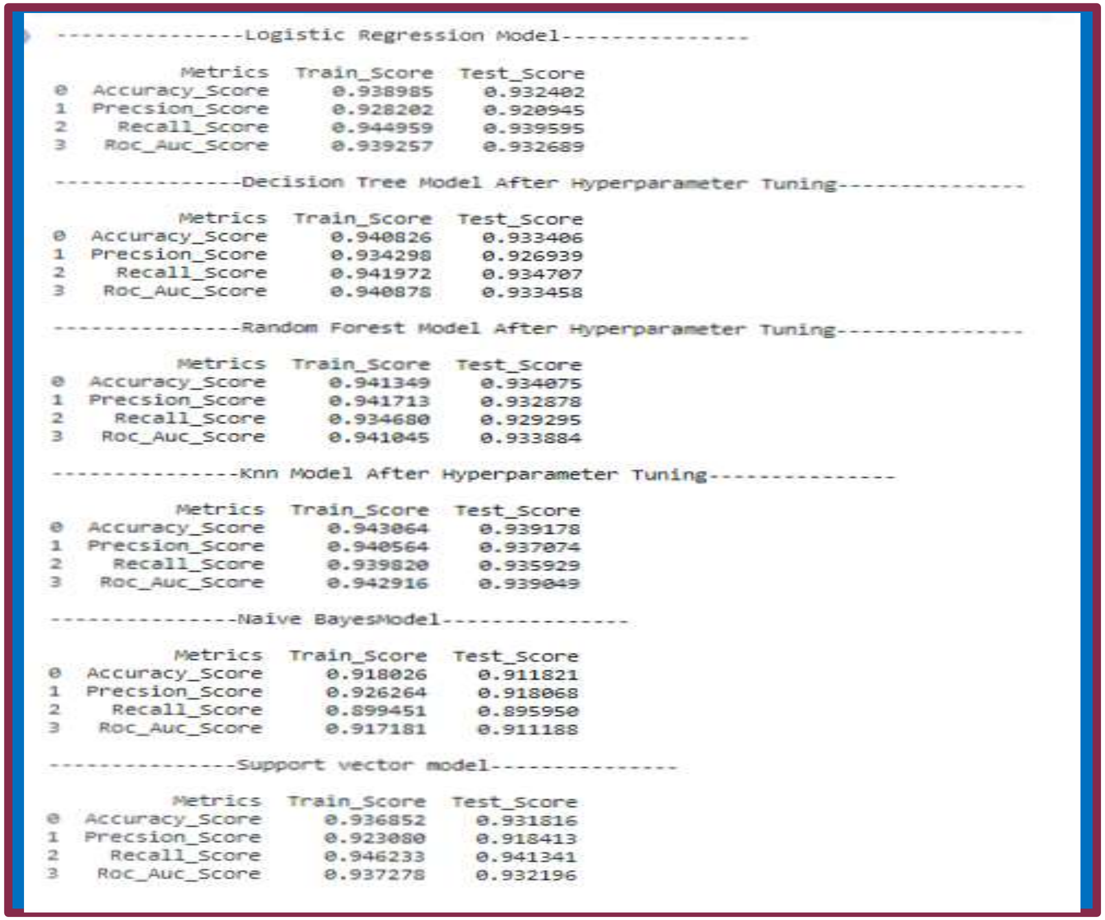
## Which Evaluation metrics did you consider for a positive business impact and why?

- I would like to go with Precision.
- model evaluation metrics comparison, we can see that K-Nearest Machine being the model with highest accuracy rate by a very small margin, works best among the experimented models for the given dataset.
- Accuracy : Accuracy will require two inputs (i) actual class labels (ii)predicted class labels. To get the class labels from probabilities( these probabilities will be probabilities of getting a HIT), you can take a threshold of 0.5. Any probability above 0.5 will be labeled as class 1 and anything less than 0.5 will be labeled as class 0.
- Precision : Precision for a label is defined as the number of true positives divided by the number of predicted positives. Report precision in percentages.
- Recall : Recall for a label is defined as the number of true positives divided by the total number of actual positives. Report recall in percentages.
- F1-Score : This is defined as the harmonic mean of precision and recall

# ML MODEL IMPLEMENTATION

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

- 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 perform well in this dataset but I chose K-Nearest Neighbors model.



```
---------------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
```

# ML MODEL IMPLEMENTATION

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

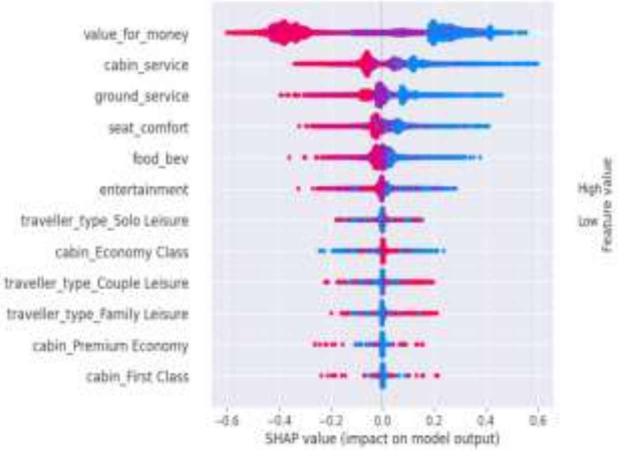
- This plot is made of all the dots in the train data. It demonstrates the following information:

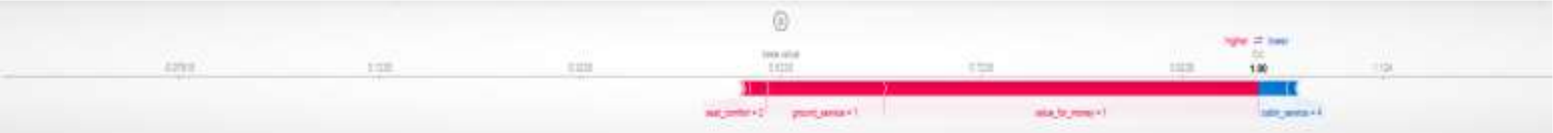
- Feature importance: Variables are ranked in descending order.

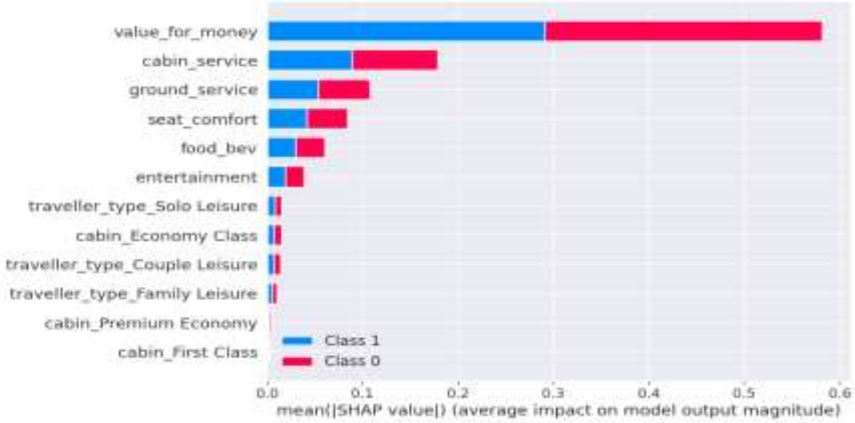
- Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.

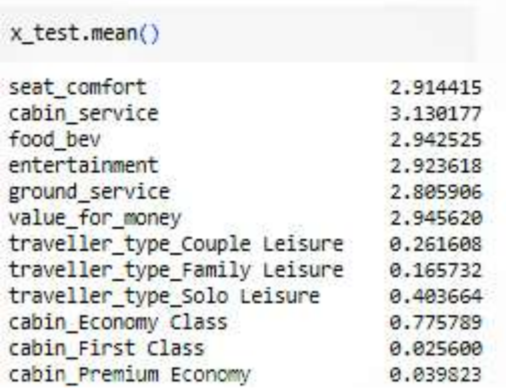
- Original value: Color shows whether that variable is high (in red) or low (in blue) for that observation.

- 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.









```
x_test.mean()

seat_comfort                     2.914415
cabin_service                    3.130177
food_bev                         2.942525
entertainment                    2.923618
ground_service                   2.805906
value_for_money                  2.945620
traveller_type_Couple Leisure    0.261608
traveller_type_Family Leisure    0.165732
traveller_type_Solo Leisure      0.403664
cabin_Economy Class              0.775789
cabin_First Class                0.025600
cabin_Premium Economy            0.039823
```

- 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 recommended a particular airline to his/her friends.

# 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 recommended 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.

THANK
YOU