ML Classification Project on Airline Passenger Referral Prediction

Ashwani Kumar, Shahbaz khan Data Science trainees, Alma Better, Bangalore

Abstract:

This is the final technical document report of our project title “ML Classification Airline Passenger Referral Prediction” as a part of ourdata science Course at Alma better.

Contents:

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. Coming to the null values and missing values in the dataset, it was observed that there was a mismatch in the non-null counts which clearly stated that a large number of missing and null values were present in the dataset.

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.

Work Flow:

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 Neighbour 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, Accuracy 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.

Feature 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

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

foodbev: Rated between 1-5

entertainment: Rated between 1-5

ground service: Rated between 1-5

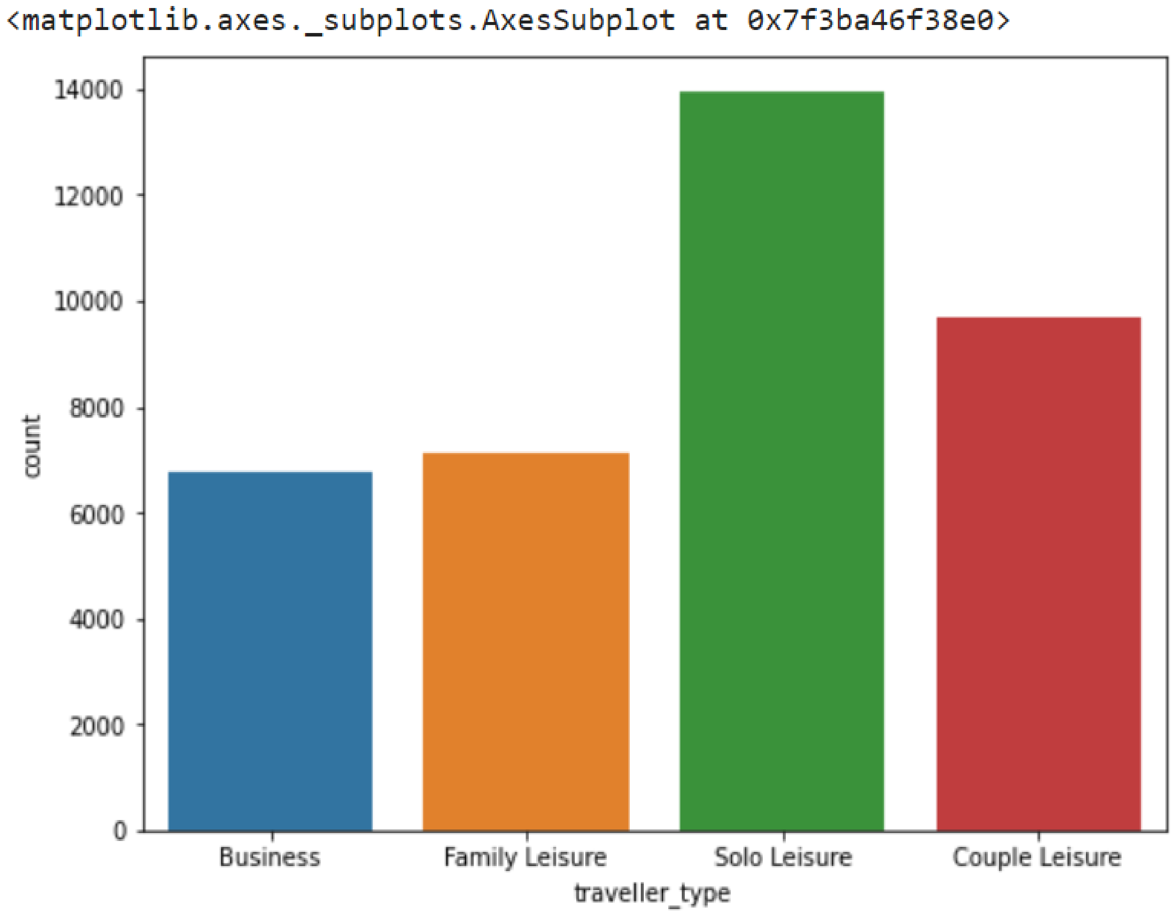
value for money: Rated between 1-5

recommended: Binary, target variable.

Observation 1:

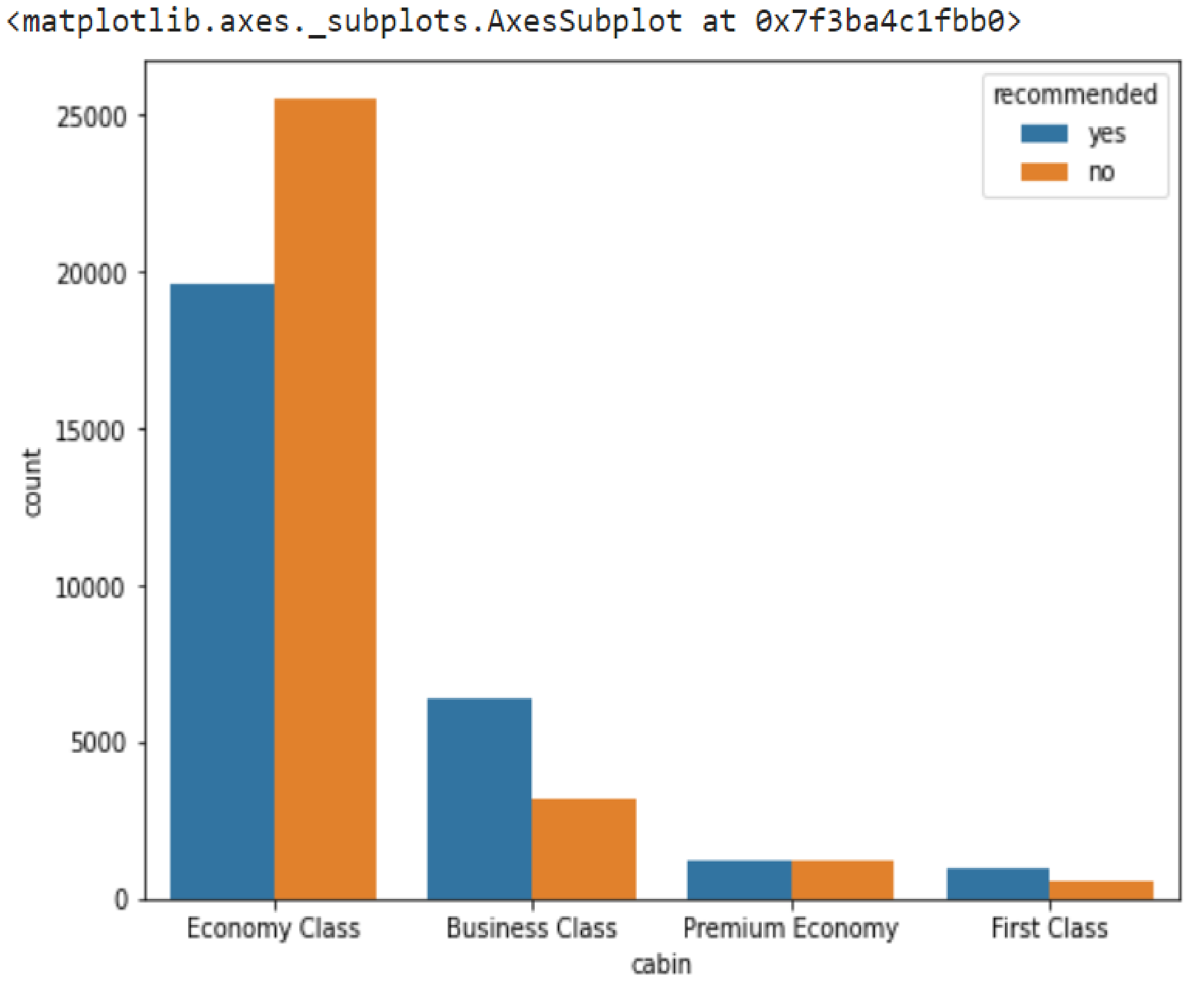
From above plot:

Travelling type of Solo Leisure has more ratings



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3ba46f38e0>
```

Observation 2:



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3ba4c1fbb0>
```

On the basis of graph -

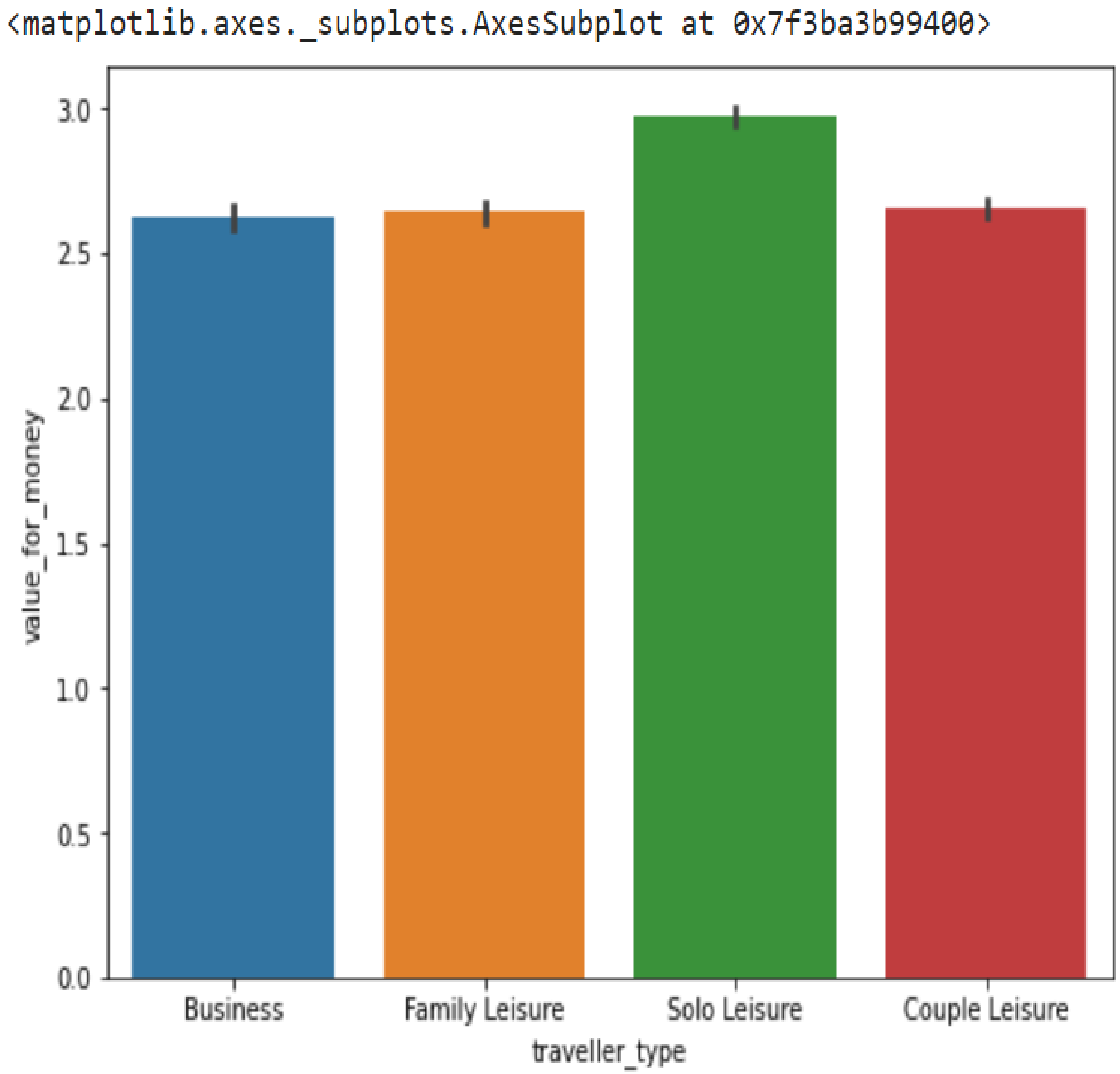
*** Economy class has highest recommendation with bad reviews.**

*** Business class has second most recommended cabin type with good reviews.**

*** Premium economy has equal reviews.**

*** First class is least Recommend cabin type with good reviews.**

Observation 3:

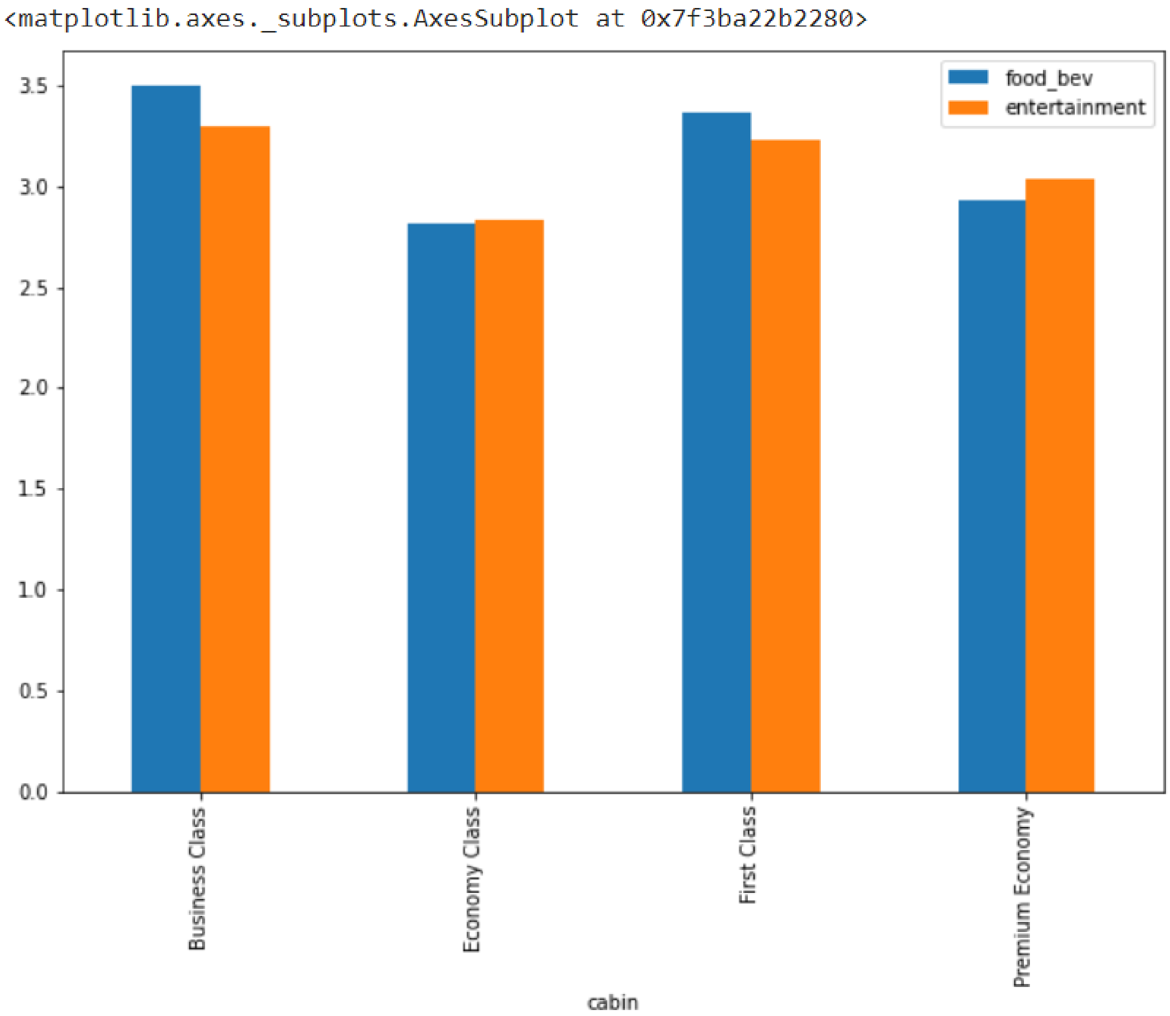


```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3ba3b99400>
```

From above plot:

Yes,Travelling Type of Solo Leisure worth of Money compare to other type of travelling.

Observation 4:



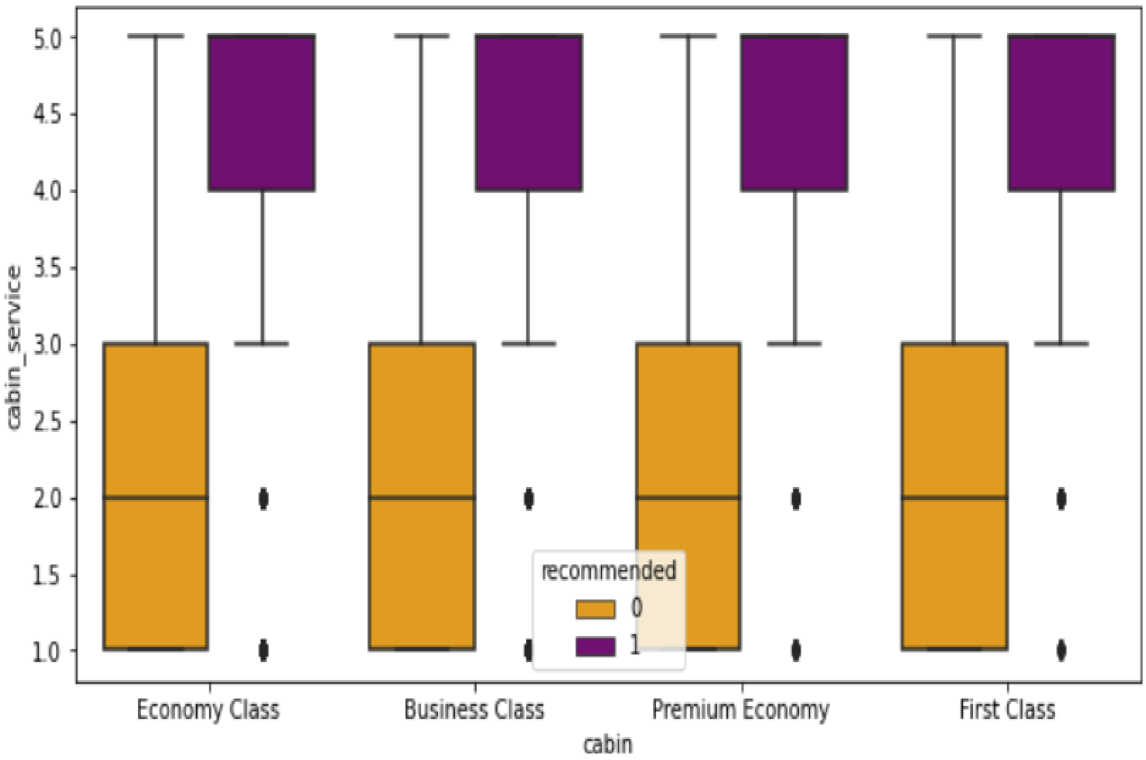
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3ba22b2280>
```

From above plot

In Economy Class the average ratings of Food_bev and entertainment given by passenger is lowest compared to other cabin classes.

Observation 5:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3ba1b86130>
```

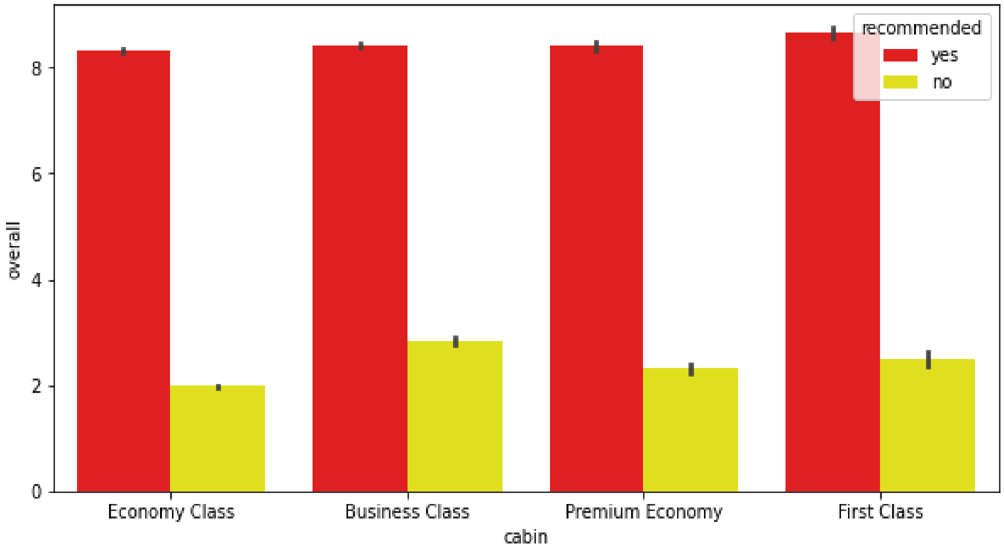


*** First class travelers are least likely to recommend the airlines.**

*** Recommendation is most probable when the cabin service is given full star rating is 5 out of 5 here.**

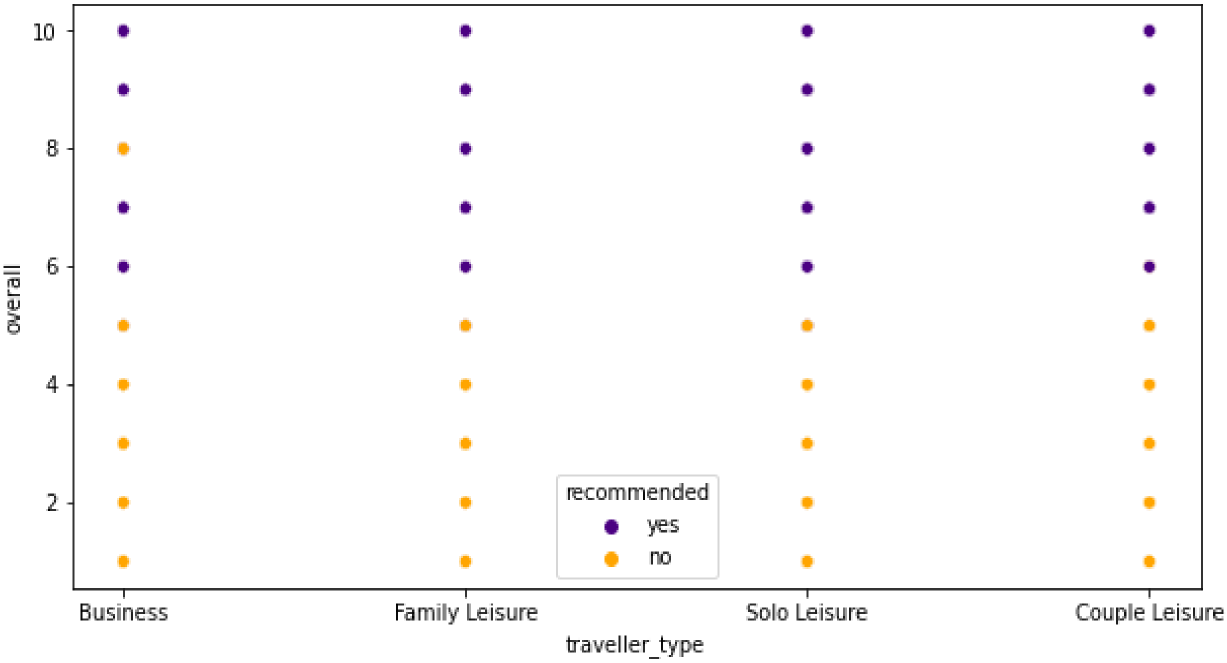
*** In economy class if we got ratings between 4 to 5 that means airlines recommended.**

Observation 6:



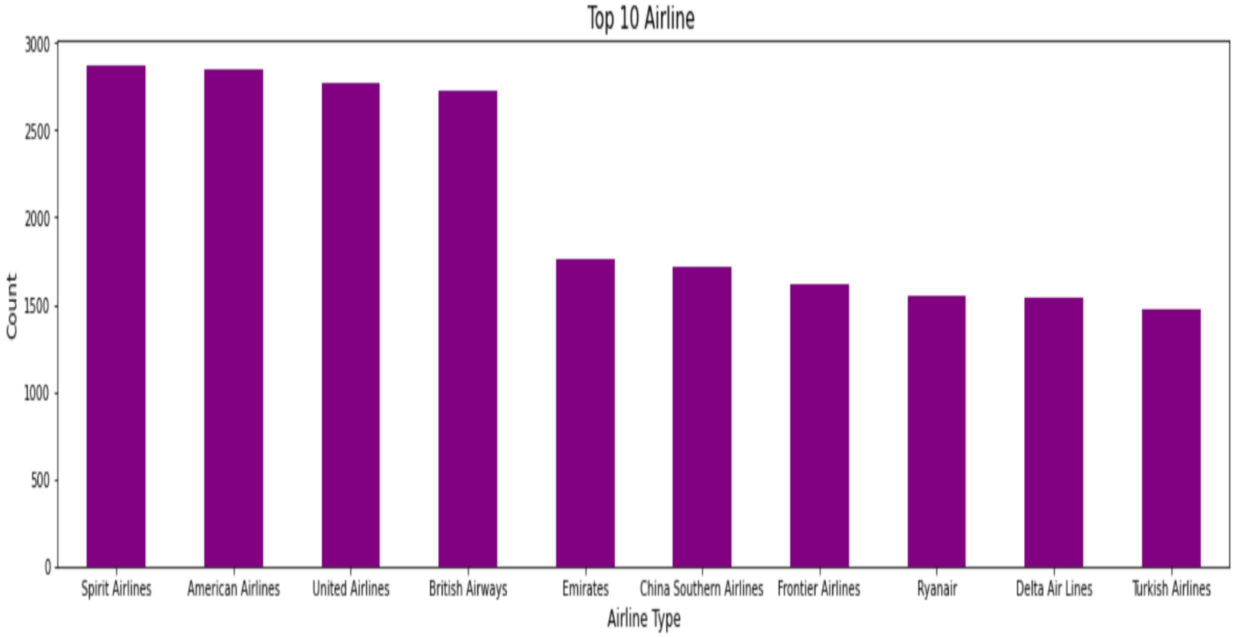
*** If the trip is rated above 8 for overall section, the trip is most likely be recommended by the travellers.**

Observation 7:



`Traveler type and overall service ratings (out of 10)`

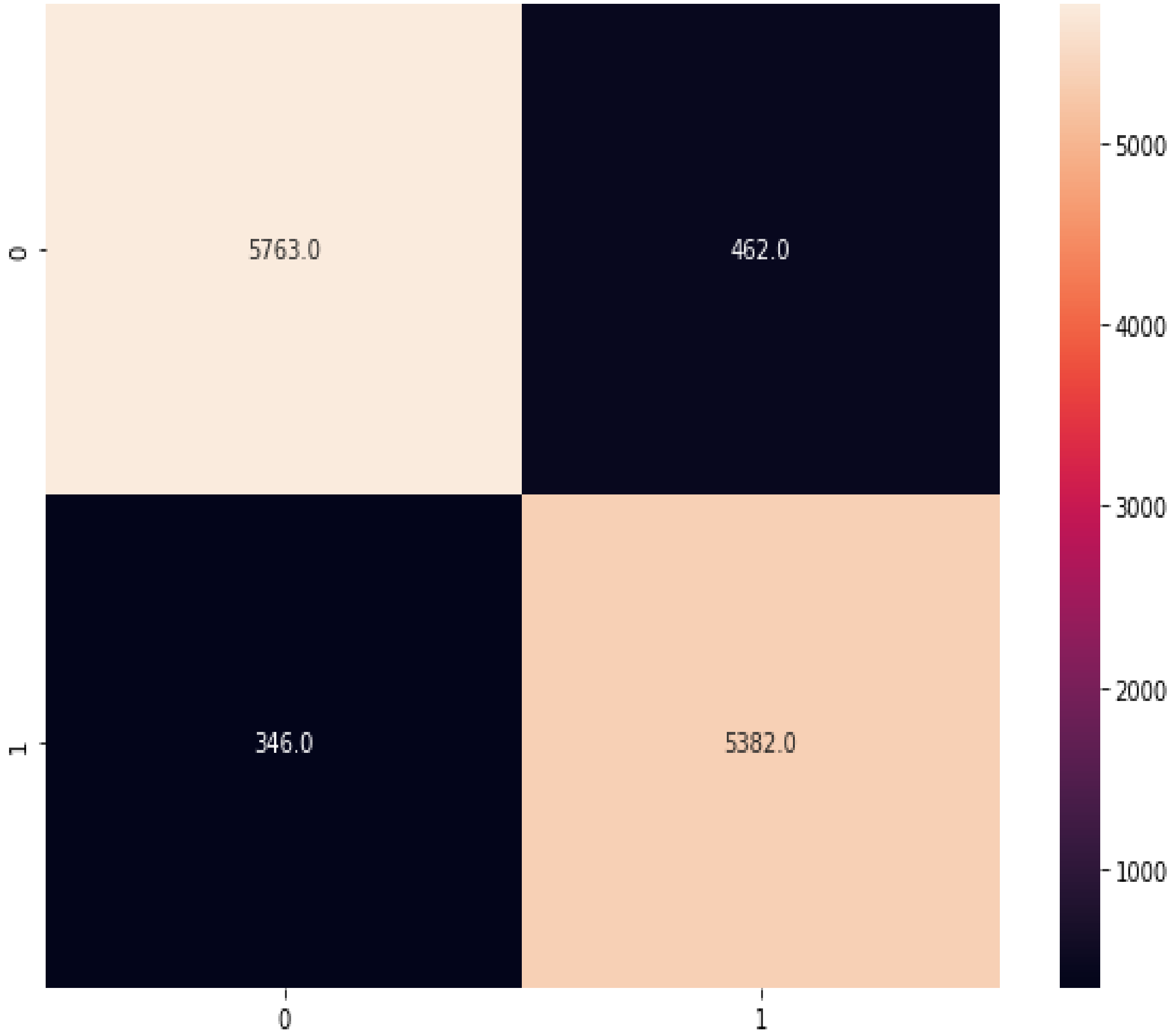
Observation 8



We have observed that the top 10 airlines with most trips are-

*** Spirit Airlines**
*** American Airlines**
*** United Airlines**
*** British Airways**
*** Emirates**
*** China southern airline**
*** Frontier airlines**
*** Ryanair**
*** Delta Airlines**
*** Turkish airlines**

Observation 9:



Confusion matrix of logistic regression:

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. Support Vector Machine Model
6. Naive Bayes

- We performed Hyperparameter 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 Overfitting Criteria, this is done. After that, we finalized the Gradient Boosting model by fine-tuning the hyperparameters.
- 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 Logistic Regression 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.

References

1 https://www.geeksforgeeks.org/

2. Alma better notes

3. some others reference

4. https://www.python.org.com

Thank You