Capstone Project

Airline Passenger Referral Prediction

By Sayan Bandopadhyay

Synopsis

Air transport or aviation plays a very important role in present transport structure of the world and surely it is considered as 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. 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 on services provided by different airlines to the customers. Main objective of this project is to understand how likely will the passengers recommend the airlines to others.

Data Briefing

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 :

- 7 columns with ‘float64’ dtypes.
- 10 columns with ‘object’ dtypes.

Coming to the null 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. 1326305 were the total count of null or missing values present in the dataset.

Data Scrubbing

Data cleaning steps performed:

- Dropped some irrelevant columns from the dataset which were not at all necessary for prediction purpose.

- Dropping all rows which had null values for all columns.

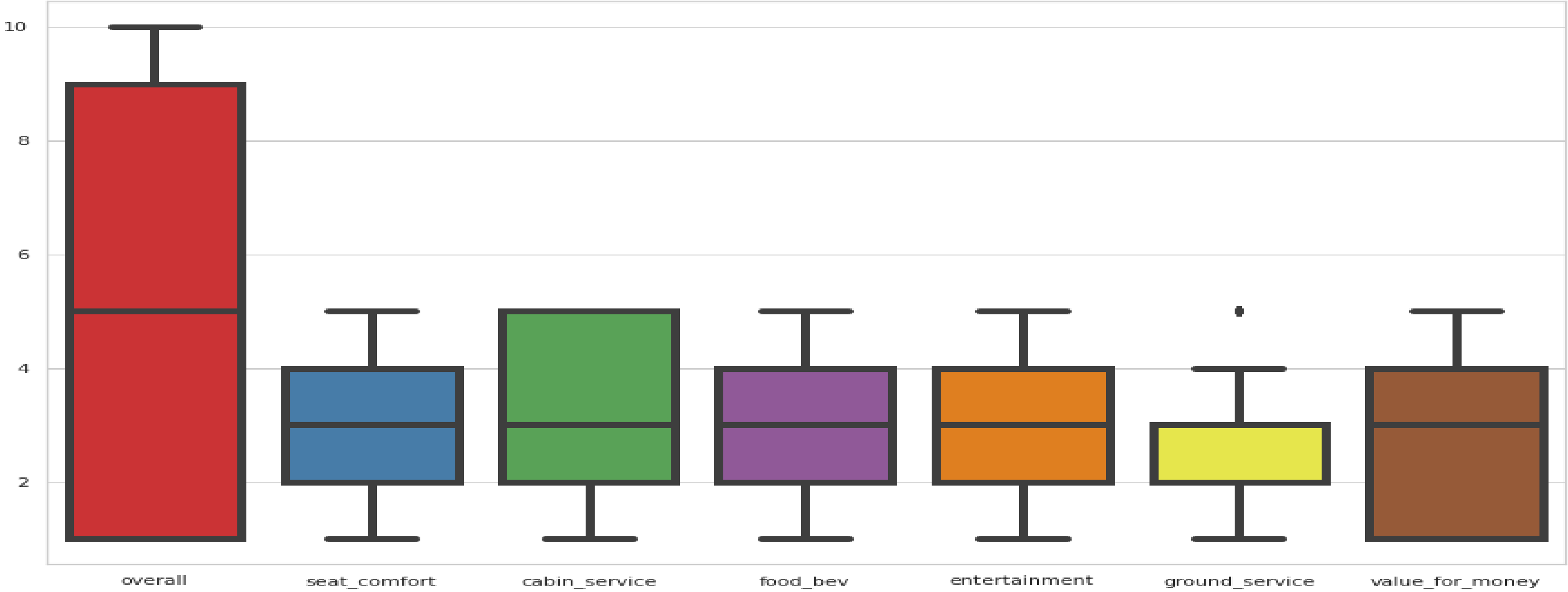
- Dropping rows which had null values in target variable column.

- Replaced null values in numerical columns with the median value of that particular column.

- Replaced null values in textual data columns with mode value of that particular columns.

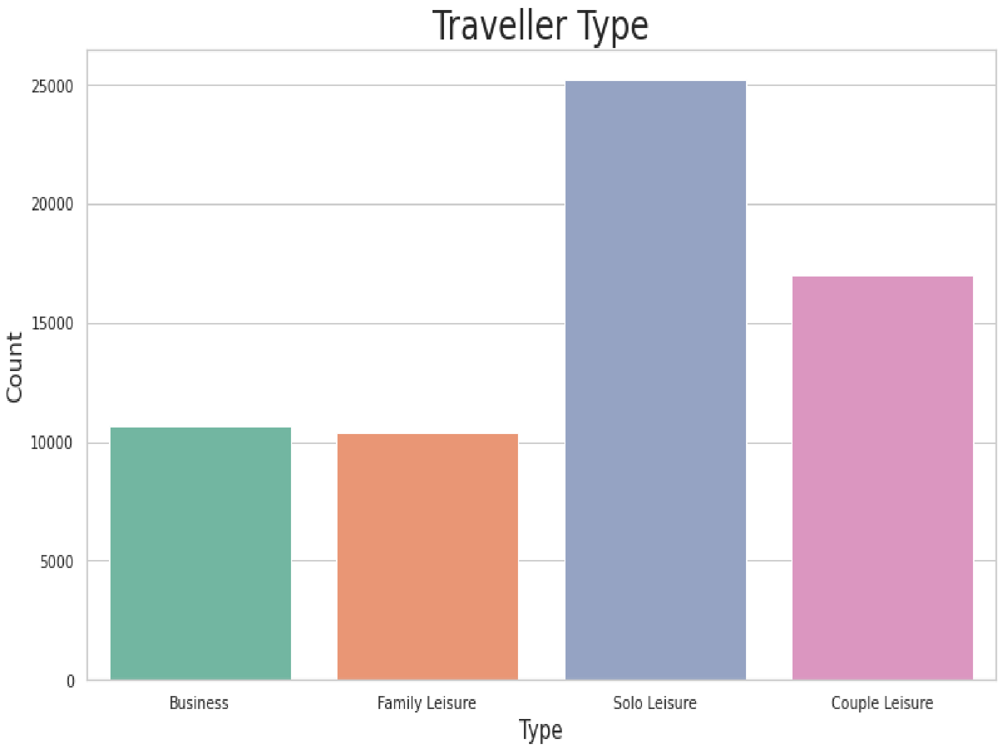
ANALYSIS REPORT

Outlier Detection



Plotted a box plot using ‘seaborn’ library with the selected features of cleaned dataset in order to check for any outlier. However, found no outlier in the dataset.

Exploratory Data Analysis (EDA)



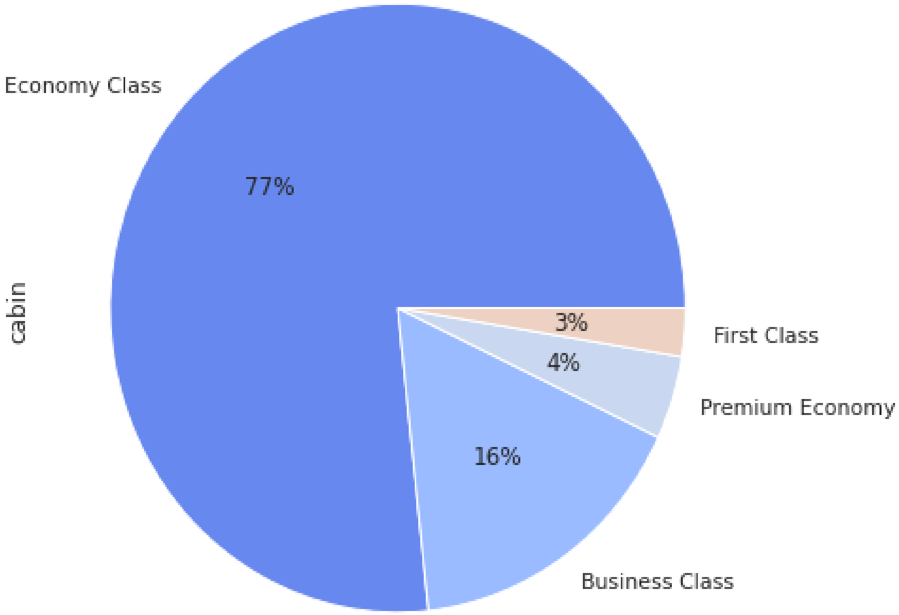
Plotted a countplot using the 'seaborn' library to get some visualisation on different traveller types in the dataset.

The plot on the left clearly shows that 'Solo Leisure' has the most ratings whereas 'Family Leisure' has the least.

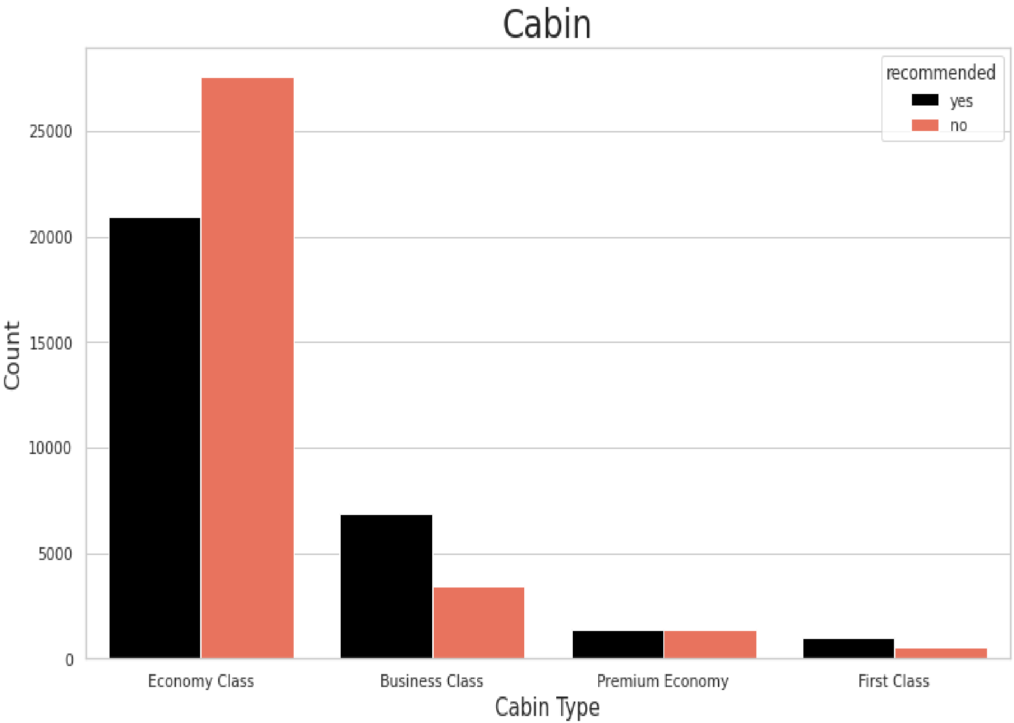
Continued…….

The pie plot on the right hand side tells about the cabin in which most of the passengers travelled.

It can be clearly derived from the plot that almost 77% of total passengers were ‘Economy Class’ traveller and only 3% of total passengers were ‘First Class’ traveller.



Continued…….



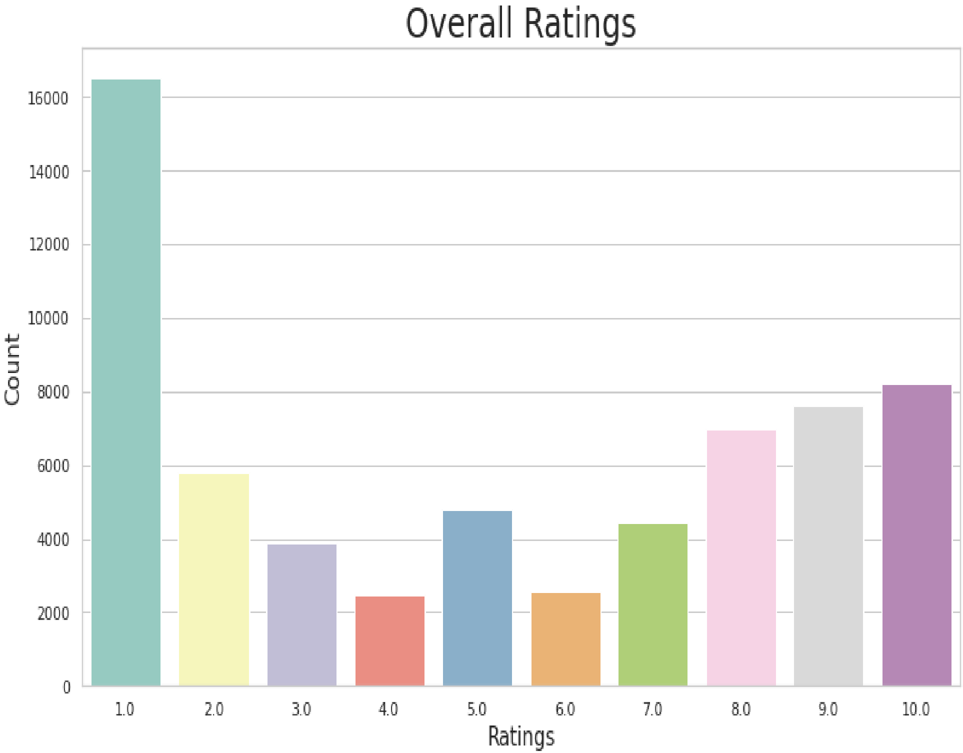
This depicts the recommendations on different cabin types in the dataset.

- For ‘Economy Class’ negative recommendations are more than the positive ones.
- For ‘Business Class’ positive recommendations are more than the negative ones.
- For ‘Premium Economy’ both negative and positive recommendations have equal weightage
- For ‘First Class’ positive recommendations are more than the negative ones.

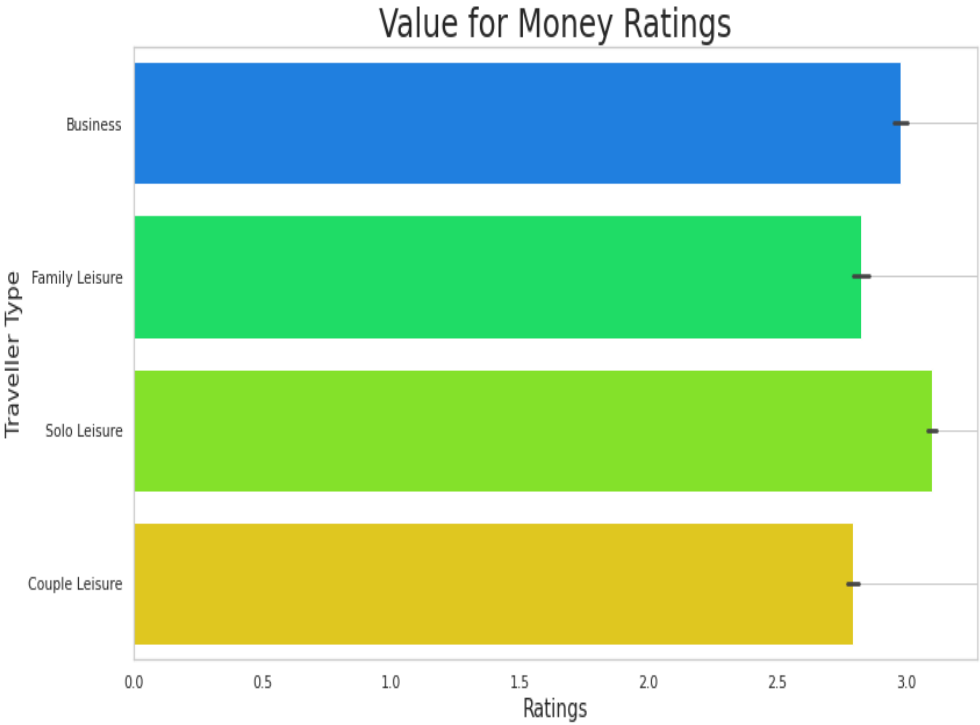
Continued.......

The plot on the right shows count of overall ratings ranging from 1 to 10.

Passengers seems dissatisfied with the services offered by the airlines as most of the ratings given is 1 out of 10.



Continued…….



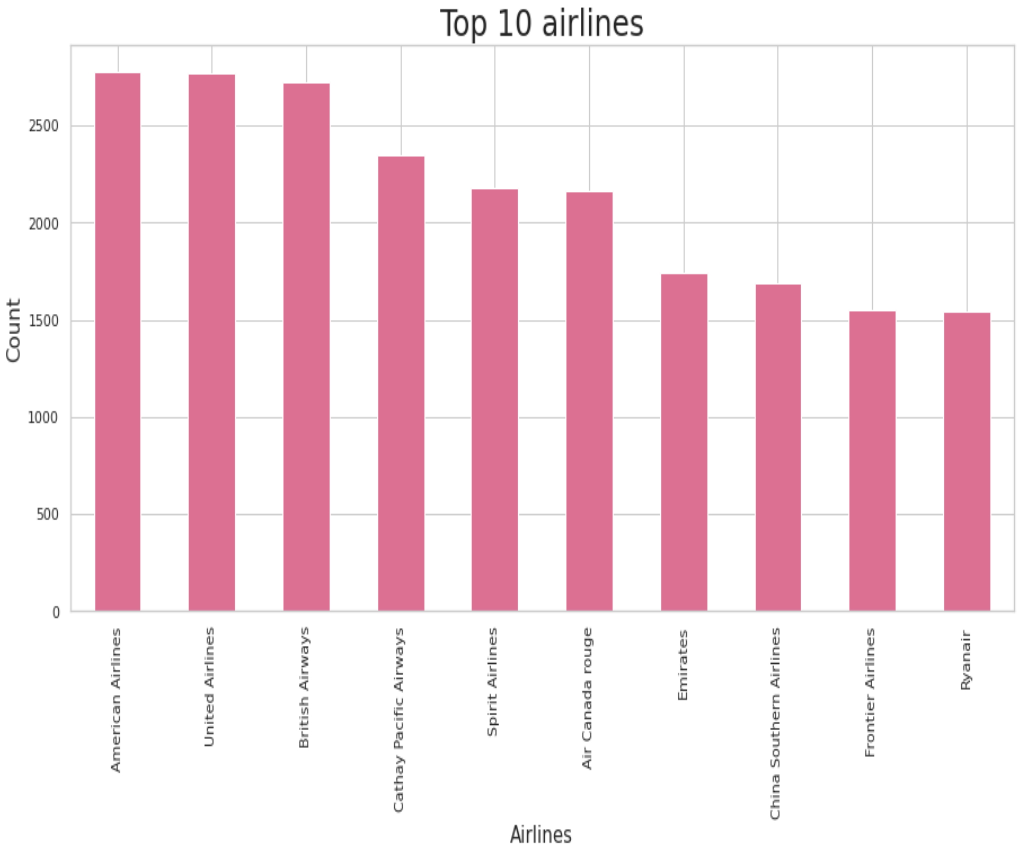
This plot summarizes the most value for money traveller type.

‘Solo Leisure’ has been the most value for money travelling type whereas ‘Couple Leisure’ is not so value for money.

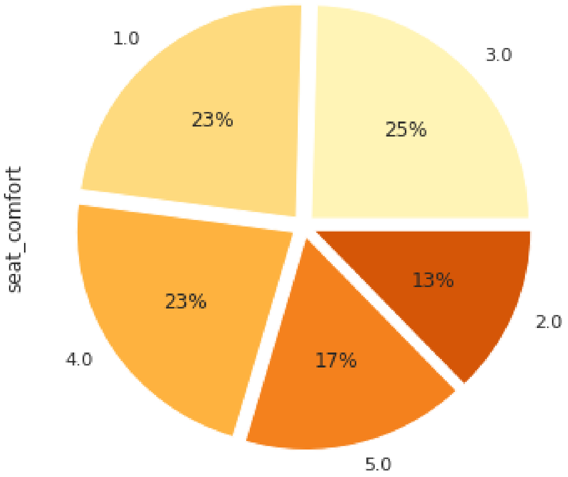
Continued…….

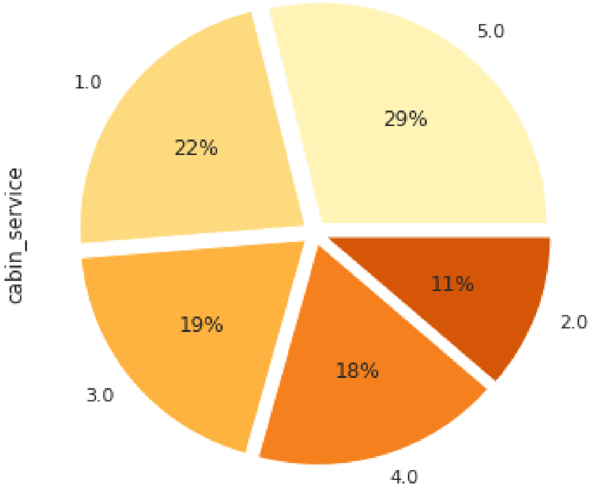
There are too many airlines data in the dataset so the plot only shows the top 10 airlines in the dataset.

‘American Airlines’ has the maximum number of trips and this can be attributed to its ultra low cost fare compared to other airlines.



Continued…….

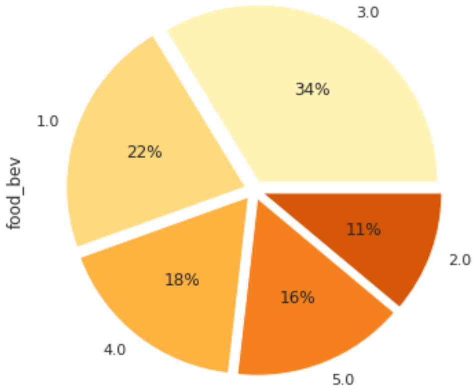


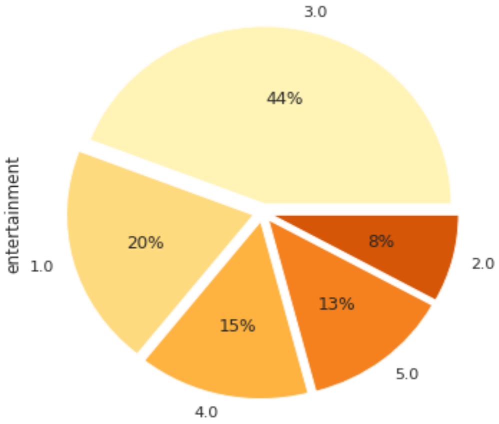


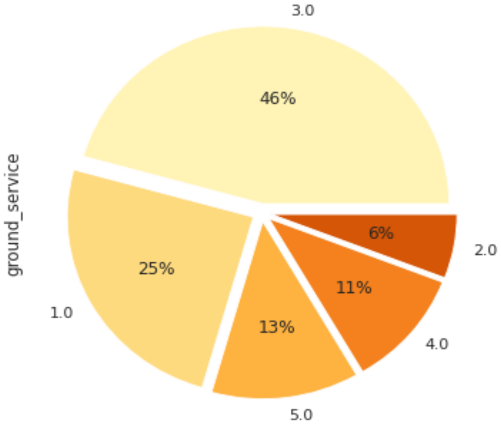
Left pie chart shows share of ratings for ‘seat_comfort’ and the right one shows for ‘cabin_service’ ranging from 1 to 5.

- For ‘seat_comfort’ maximum passengers has rated it 3 out of 5.
- For ‘cabin_service’ maximum passengers has rated it 5 out of 5.

Continued…….



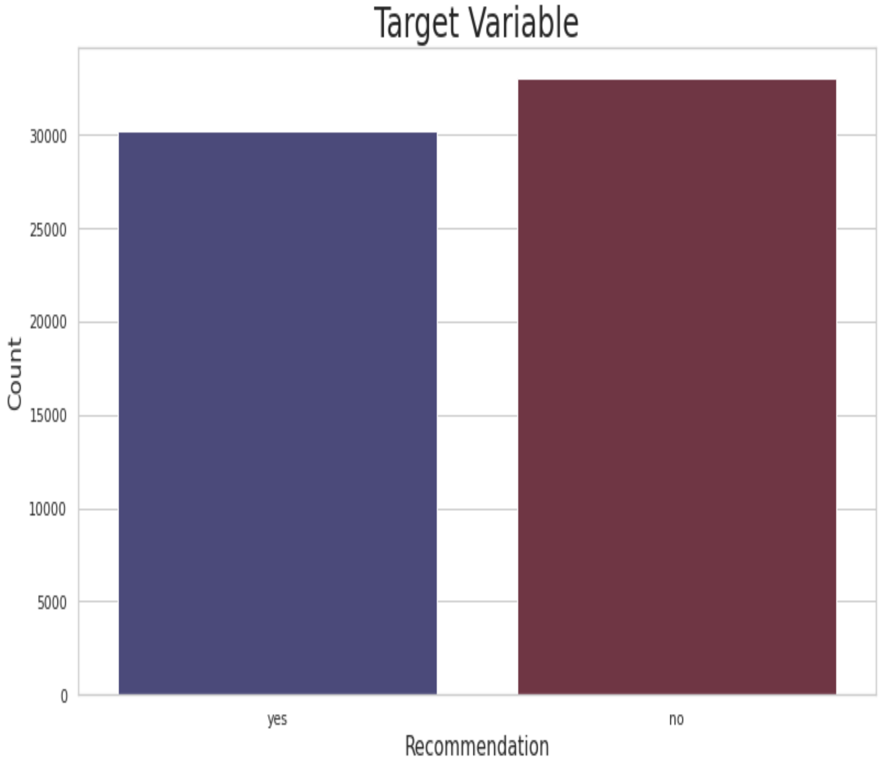




Left pie chart shows share of ratings for ‘food_bev’ , the right one shows for ‘ground_service’ and middle one shows for ‘entertainment’ ranging from 1 to 5.

- For ‘food_bev’ maximum passengers has rated it 3 out of 5.
- For ‘entertainment’ maximum passengers has rated it 3 out of 5.
- For ‘ground_service’ maximum passengers has rated it 3 out of 5.

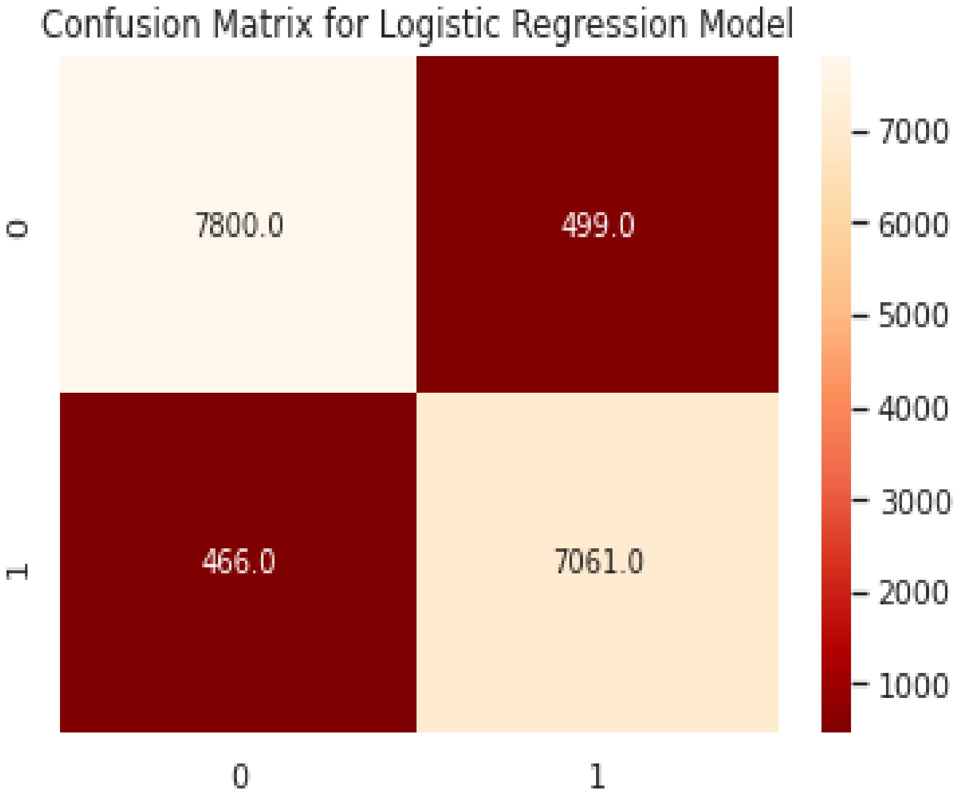
Continued…….



Finally, its time to plot the target variable.
It's a binary classification Machine learning problem so, target variable has only two classes : Yes and No.
After converting it to numerical data we get 1 and 0 where 1 means Yes and 0 means No.

From the plot its clear that negative responses are more as compared to positive responses though by a minute margin.

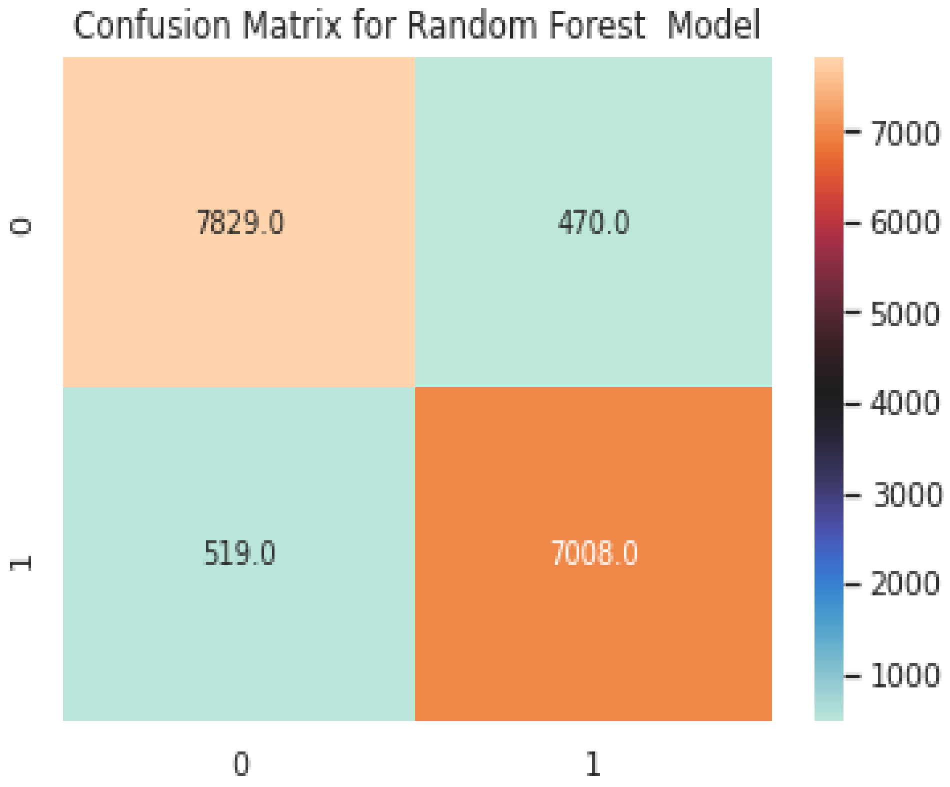
Logistic Regression



Confusion Matrix for Logistic Regression Model

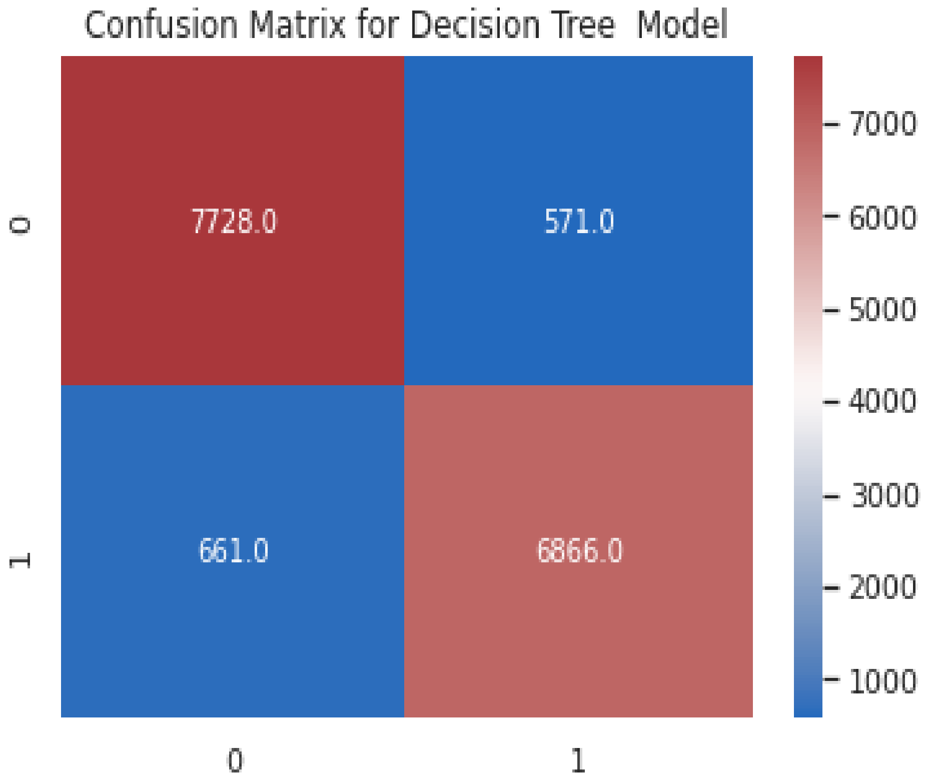
•	**Accuracy**	: 0.939024
•	**Recall**	: 0.938090
•	**Precision**	: 0.933995
•	**F1-Score**	: 0.936038
•	**ROC AUC Score**	: 0.938981

Random Forest



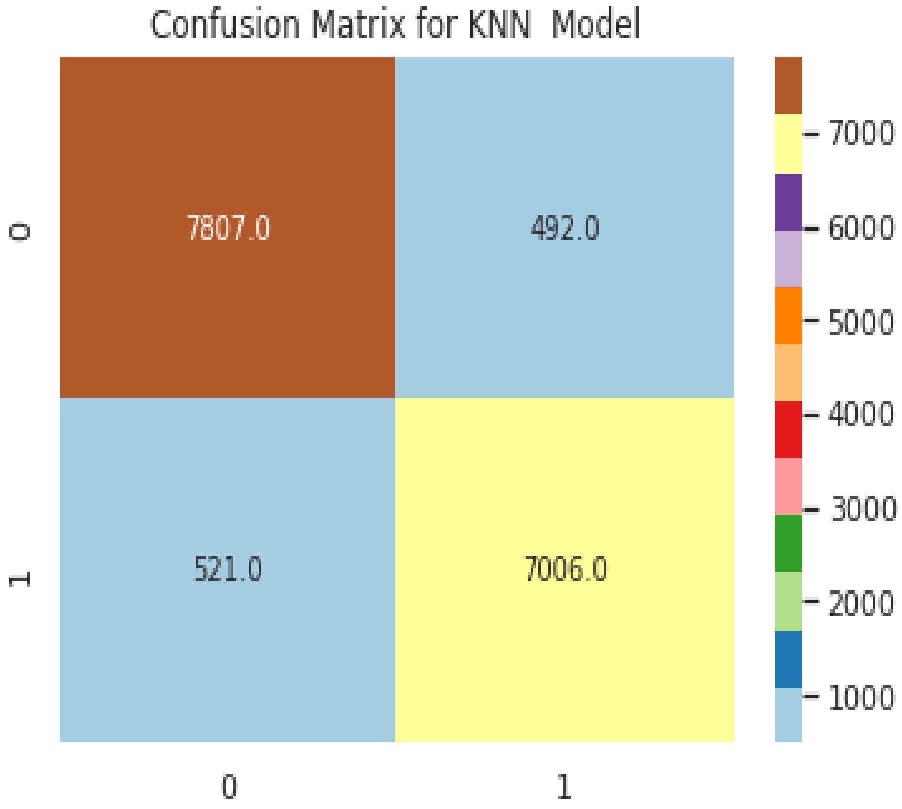
- **Accuracy** : 0.937508
- **Recall** : 0.931048
- **Precision** : 0.937149
- **F1-Score** : 0.934089
- **ROC AUC Score** : 0.937207

Decision Tree



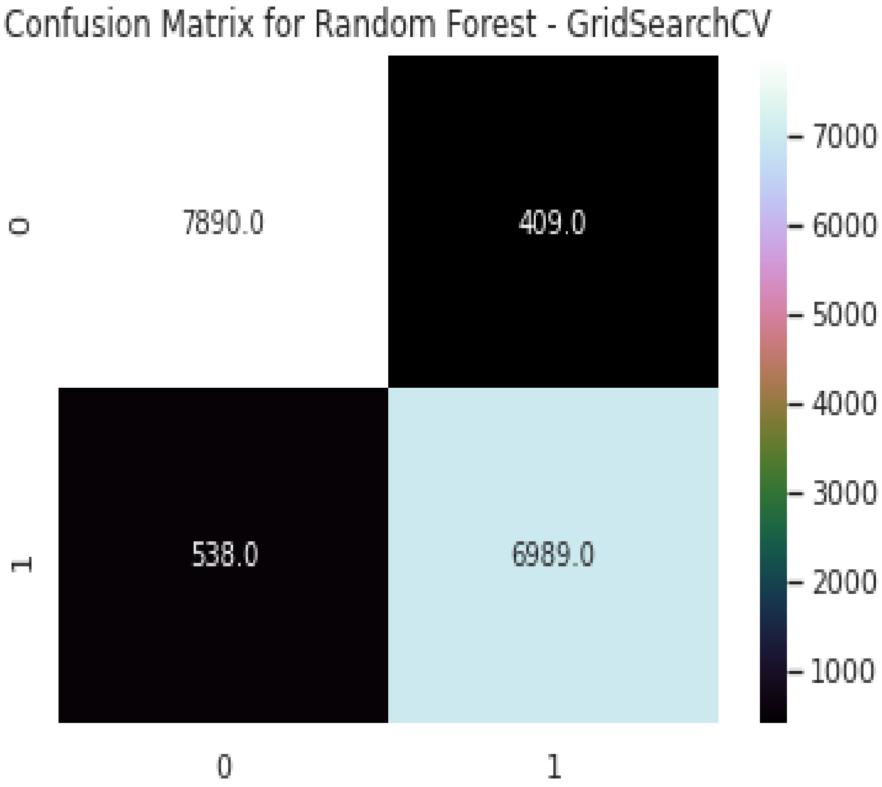
• **Accuracy**	: 0.922153
• **Recall**	: 0.912183
• **Precision**	: 0.923222
• **F1-Score**	: 0.917669
• **ROC AUC Score**	: 0.921690

KNN



- **Accuracy** : 0.935991
- **Recall** : 0.930783
- **Precision** : 0.934383
- **F1-Score** : 0.932579
- **ROC AUC Score** : 0.935749

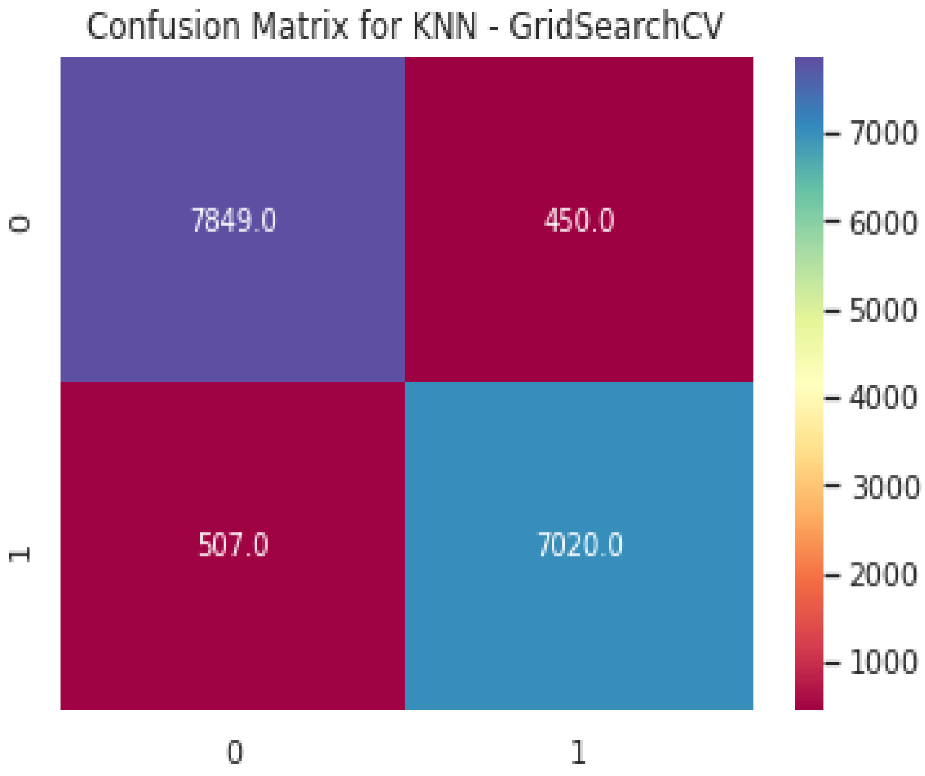
Random Forest (GridSearchCV)



Confusion Matrix for Random Forest - GridSearchCV

- **Accuracy** : 0.940162
- **Recall** : 0.928524
- **Precision** : 0.944715
- **F1-Score** : 0.936549
- **ROC AUC Score** : 0.939620

KNN (GridSearchCV)



Accuracy	: 0.939530
Recall	: 0.932642
Precision	: 0.939759
F1-Score	: 0.936187
ROC AUC Score	: 0.939210

Model Comparison

MODEL NAME	ACCURACY	RECALL	PRECISION	F1-SCORE	ROC AUC SCORE
Random Forest (GridSearchCV)	0.940162	0.928524	0.944715	0.936549	0.939620
KNN (GridSearchCV)	0.939530	0.932642	0.939759	0.936187	0.939210
Logistic Regression	0.939024	0.938090	0.933995	0.936038	0.938981
Random Forest	0.937508	0.931048	0.937149	0.934089	0.937207
KNN	0.935991	0.930783	0.934383	0.932579	0.935749
Decision Tree	0.922153	0.912183	0.923222	0.917669	0.921690

Conclusion

- "Solo Leisure" has the highest ratings among all traveler type category.
- "Solo Leisure" has been the most value for money.
- 77% of the passengers are economy class traveler.
- Economy class has most recommendations.
- Passengers seems dissatisfied as most of them rated 1 out of 10 in overall rating.
- 'American Airlines' has maximum number of trips.
- 'NO' responses are more compared to 'YES' responses.
- Four machine learning models have been implemented (KNN, Random Forest, Decision Tree, Logistic Regression).
- GridSearchCV used for the purpose of hyperparameter tuning.
- 'Accuracy', 'Recall', 'Precision', 'F1-score' & 'ROC-AUC-SCORE' are the evaluation metrics taken into consideration.
- Random Forest after hyperparameter tuning performed the best out of all models.
- Decision Tree performed the worst.

THANK YOU