Capstone Project-01

Predicting Customer Behaviour in Flight Ticket Booking

Problem Statement

- Understanding customer behaviour is crucial to improve sales and customer satisfaction.
- Predicting customer behaviour in flight ticket booking is a challenging task due to the dynamic nature of the travel industry and the wide range of factors that influence customer purchasing decisions.
- Manipulate and prepare the provided customer booking data so that you can build a high-quality predictive model.

Objective

- The aim of this project is to analyze customer behaviour in flight ticket booking by analyzing historical data on flight ticket purchases and customer demographics.
- Model will be able to predict which customers are most likely to purchase flight tickets in the future, which will allow airlines and travel companies to target their marketing efforts more effectively and increase their revenue.
- Also , analyze which factors are important in predicting the customer buying behaviour accurately.

Table of Content

1. Overview of the Data
2. Data Pre-processing
3. EDA
4. Model Building
5. Comparison of Models Performance
6. Conclusion

Overview of the Data

- Shape of the dataset is (50,000 , 14).
- Data types of features are object, int and float.
- Total columns with "object" Dtype = 5
- Total columns with "int" Dtype = 8
- Total columns with "float" Dtype = 1
- No null values are present in the dataset.
- Duplicate entries are 719.

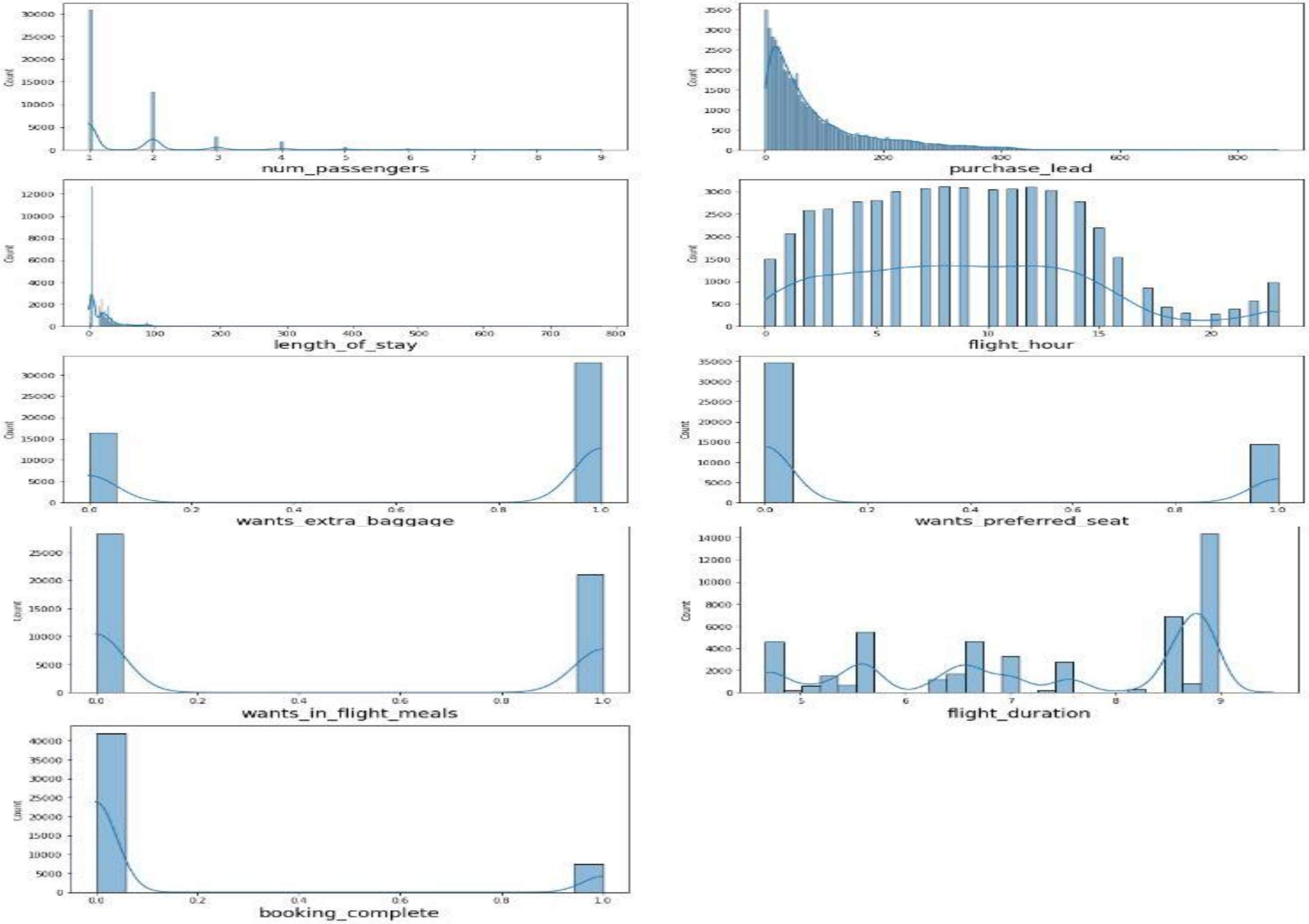
2. Data Pre-processing

- Data Cleaning:-
  - Removing the duplicates.
  - Checking for Null values
  - Feature Encoding
  - Outlier Handling

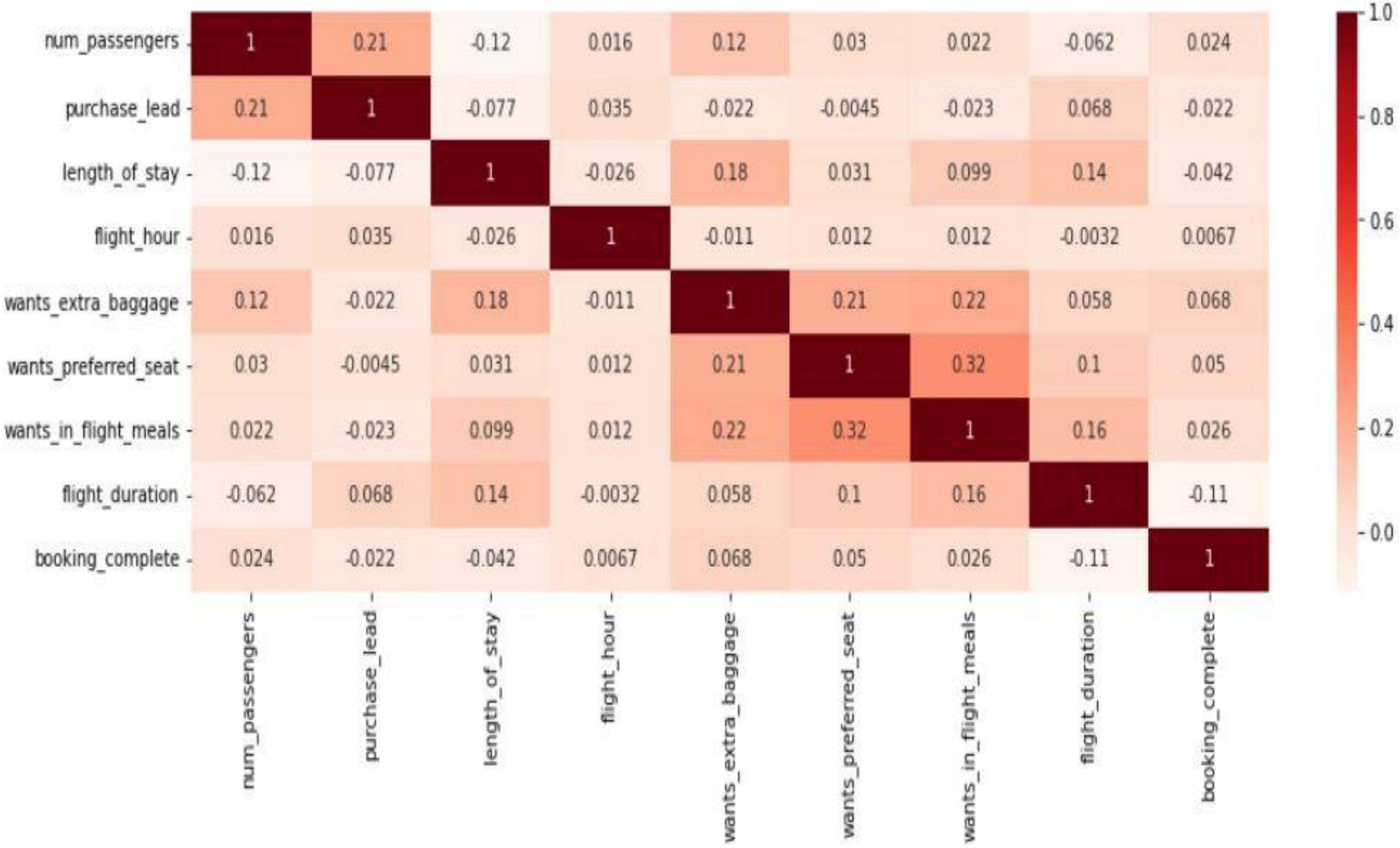
3.EDA

Plotting various plots to get data insights

Checking Distribution of data

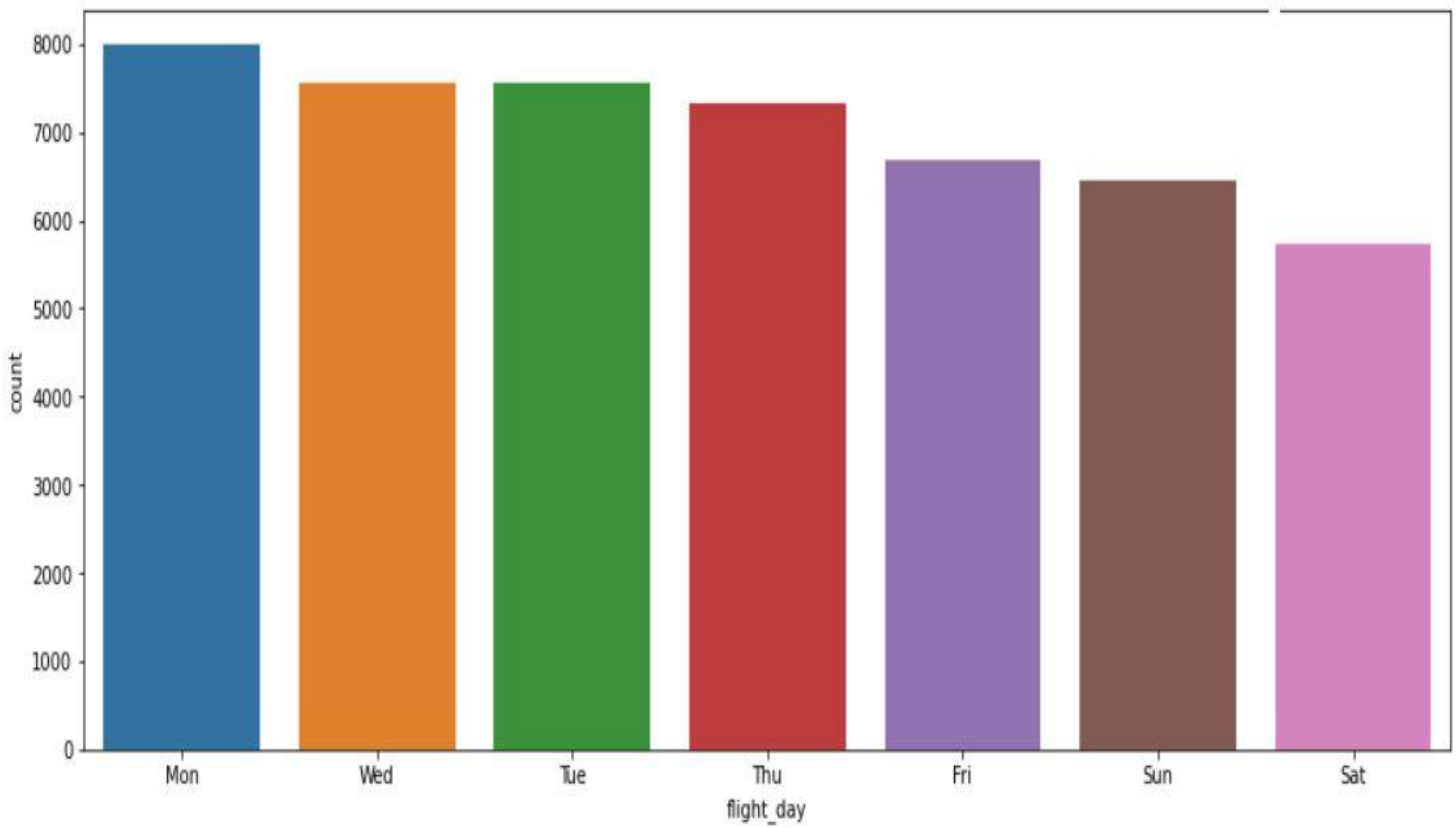


Heat map- to check correlation



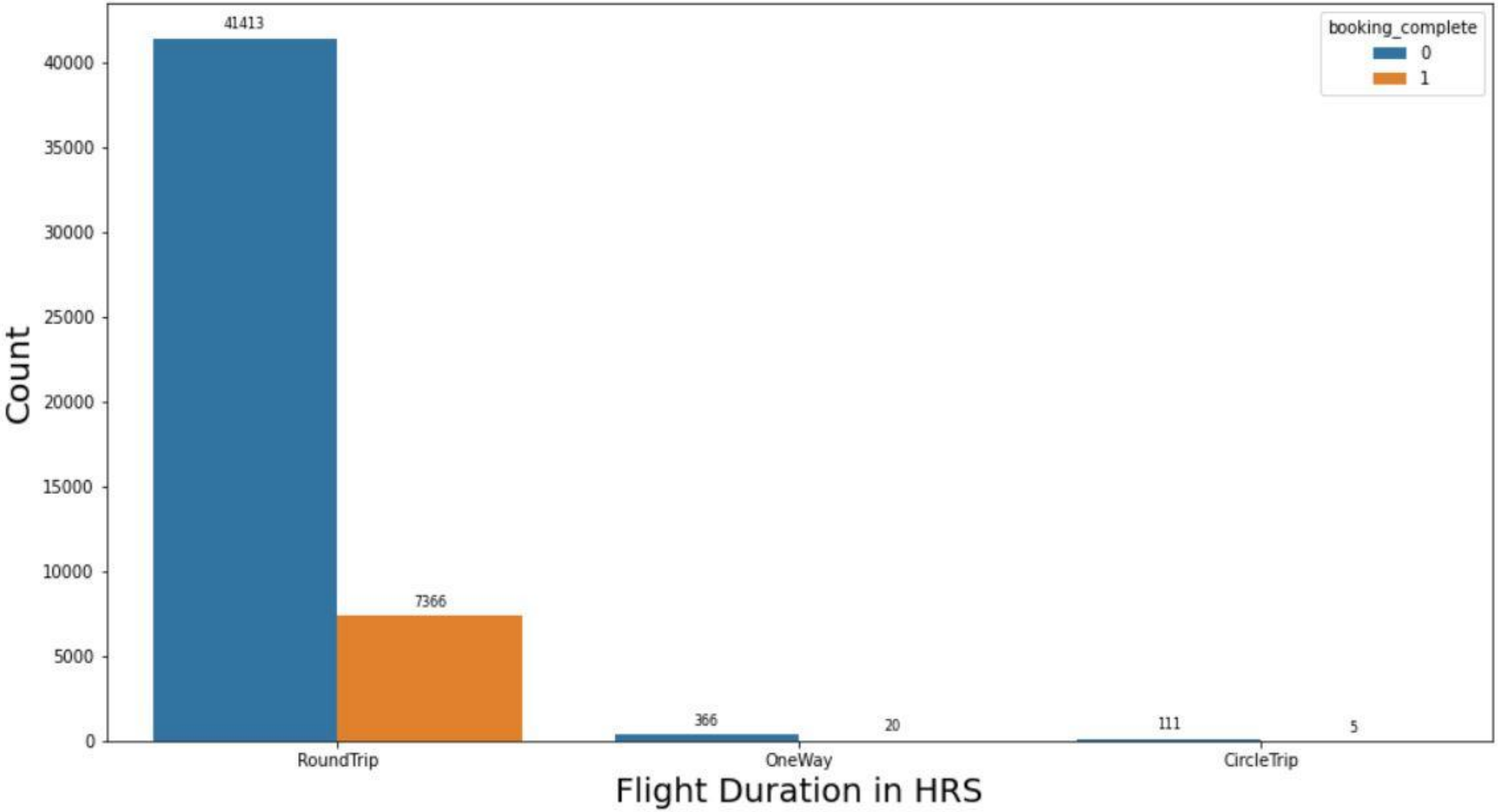
There is no significant high or negative correlation among the features or with the target variable.

No. of flights on each day of the week



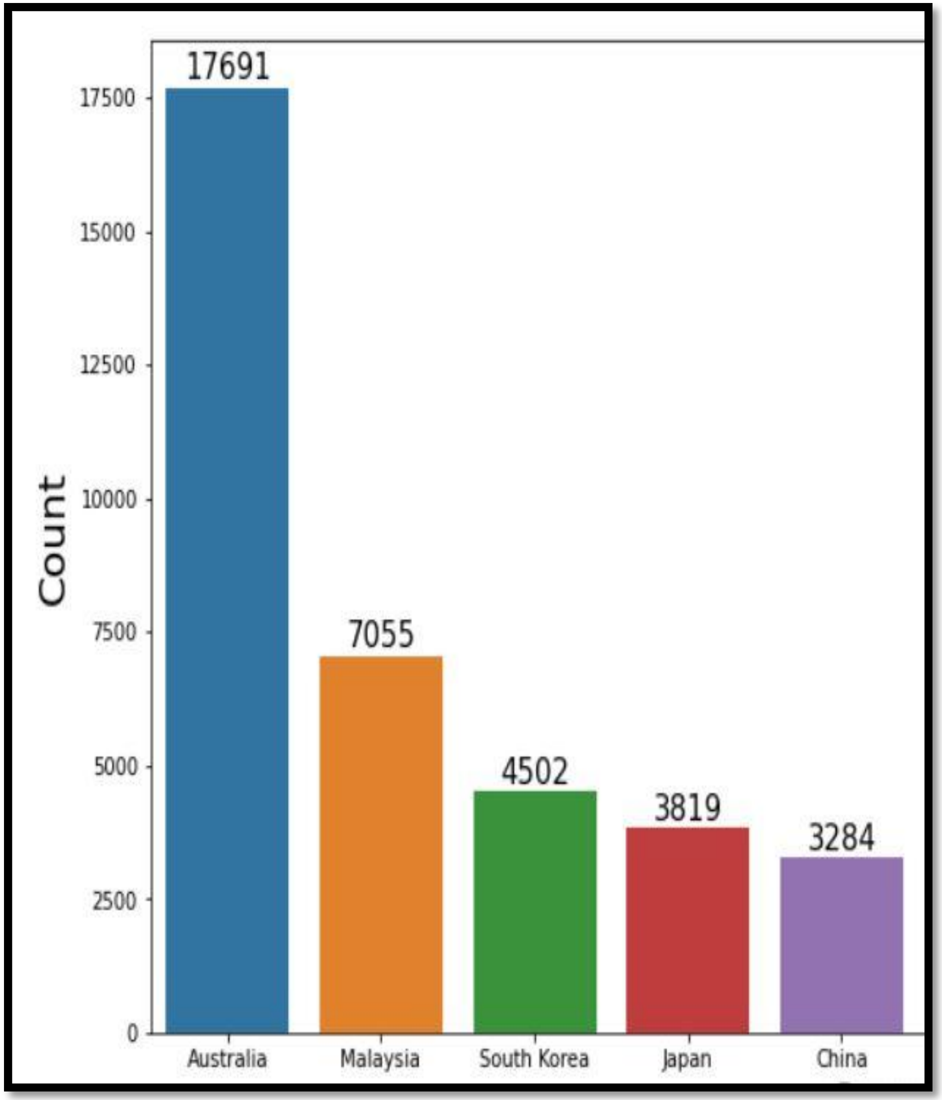
No of flights is max on Monday and lowest on Saturday

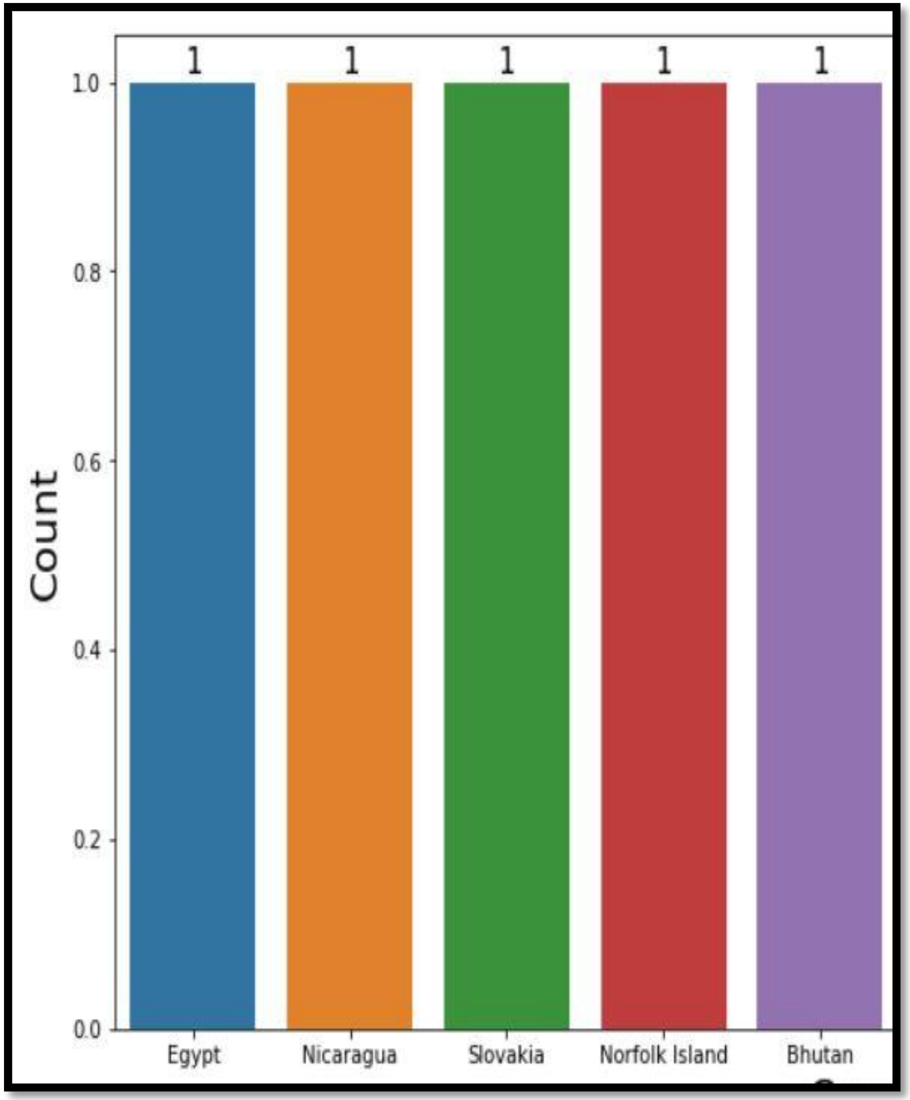
Most frequently opted trip type and booking status



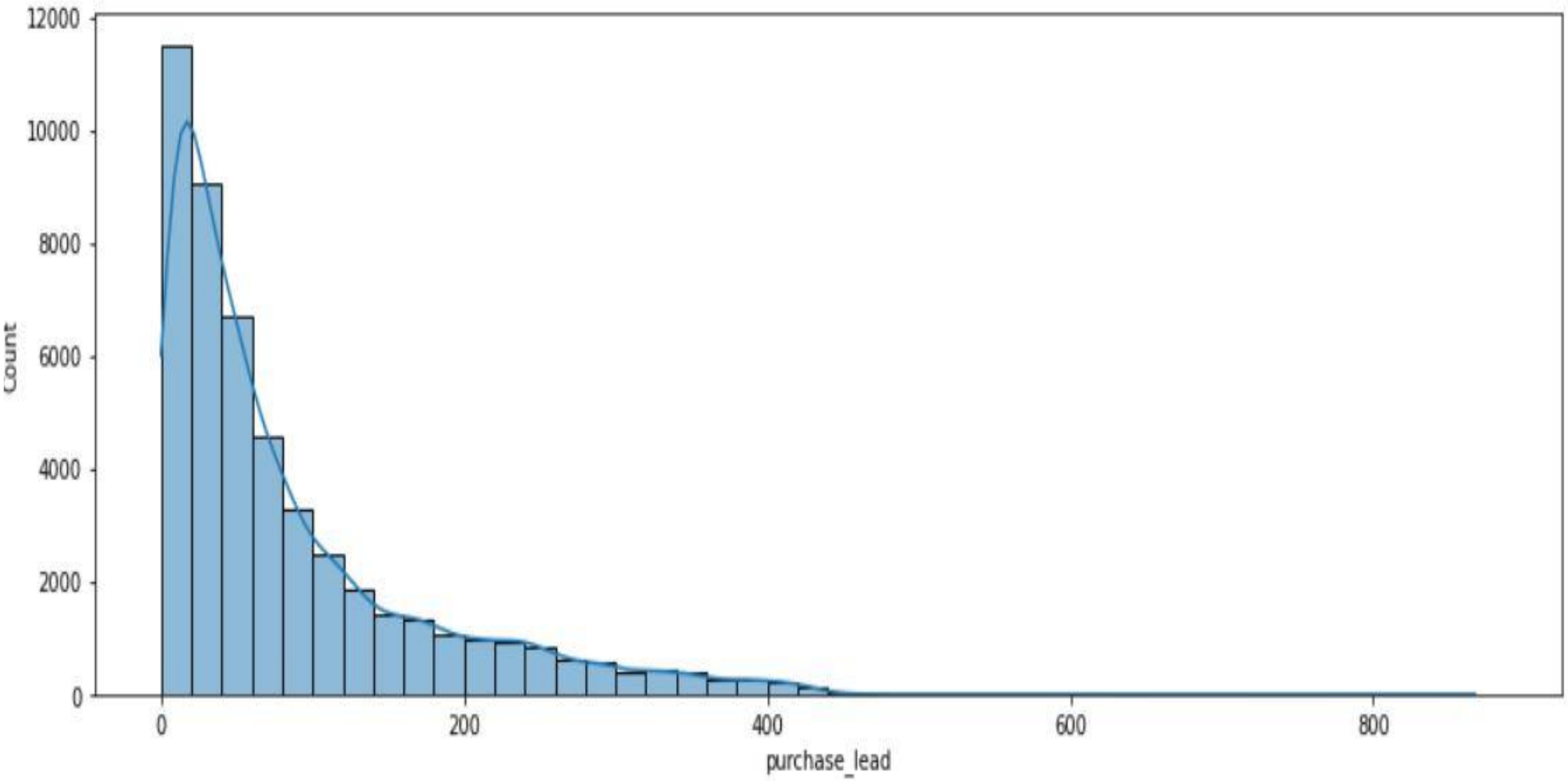
Most of the traffic is for “Round Trip” searches. Airline should provide more offers or schemes

Top and bottom 5 Countries in flight bookings



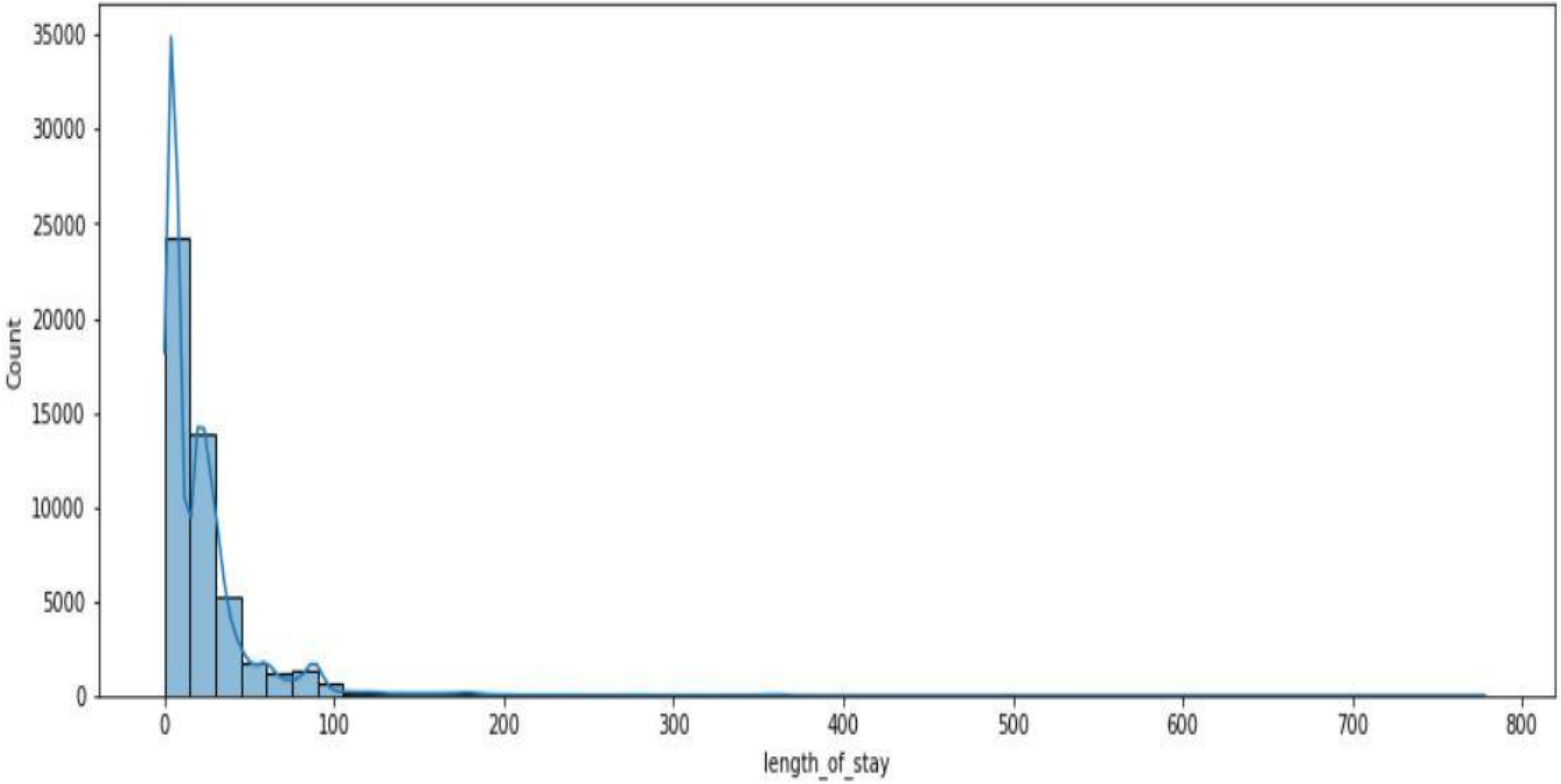


Purchase Lead

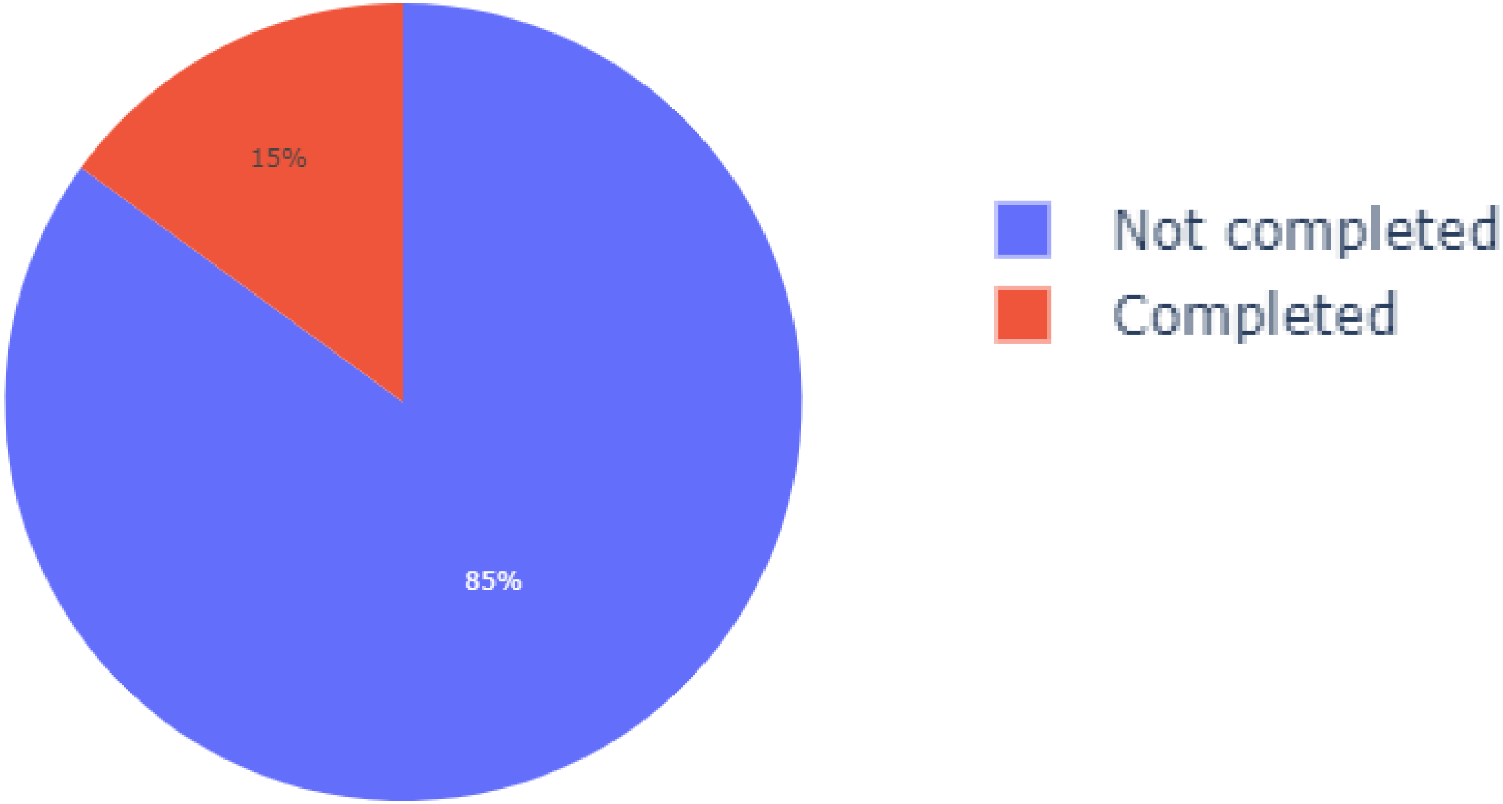


There are few bookings that were done more than 2 years before the travel date and it seems very unlikely that book that in advance. However, it might also be because of the cancellation and rebooking in a period of 6 months for twice. But at this point we will consider them as outliers which will effect the results of predictive model in a huge way.

Length Of Stay

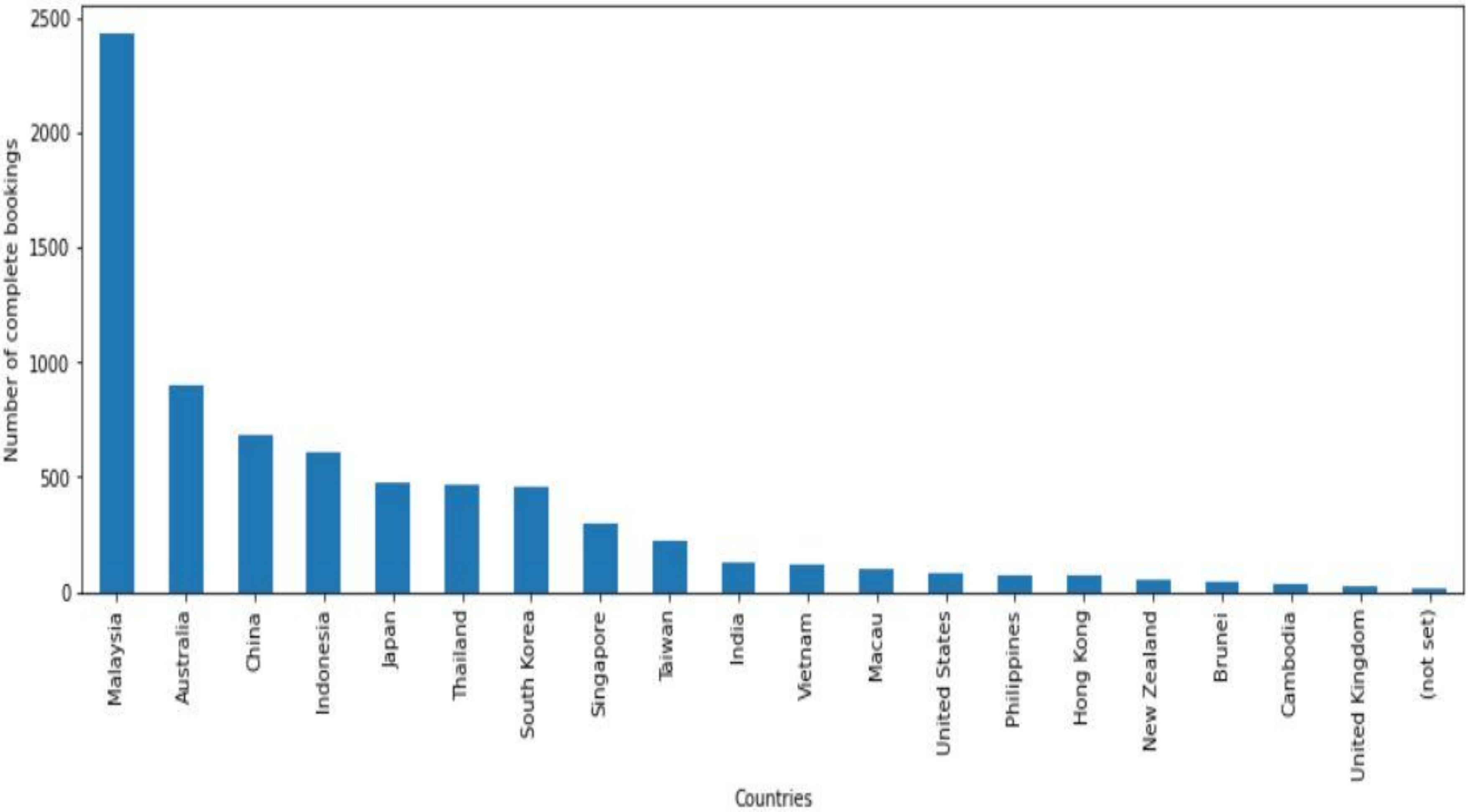


Booking complete



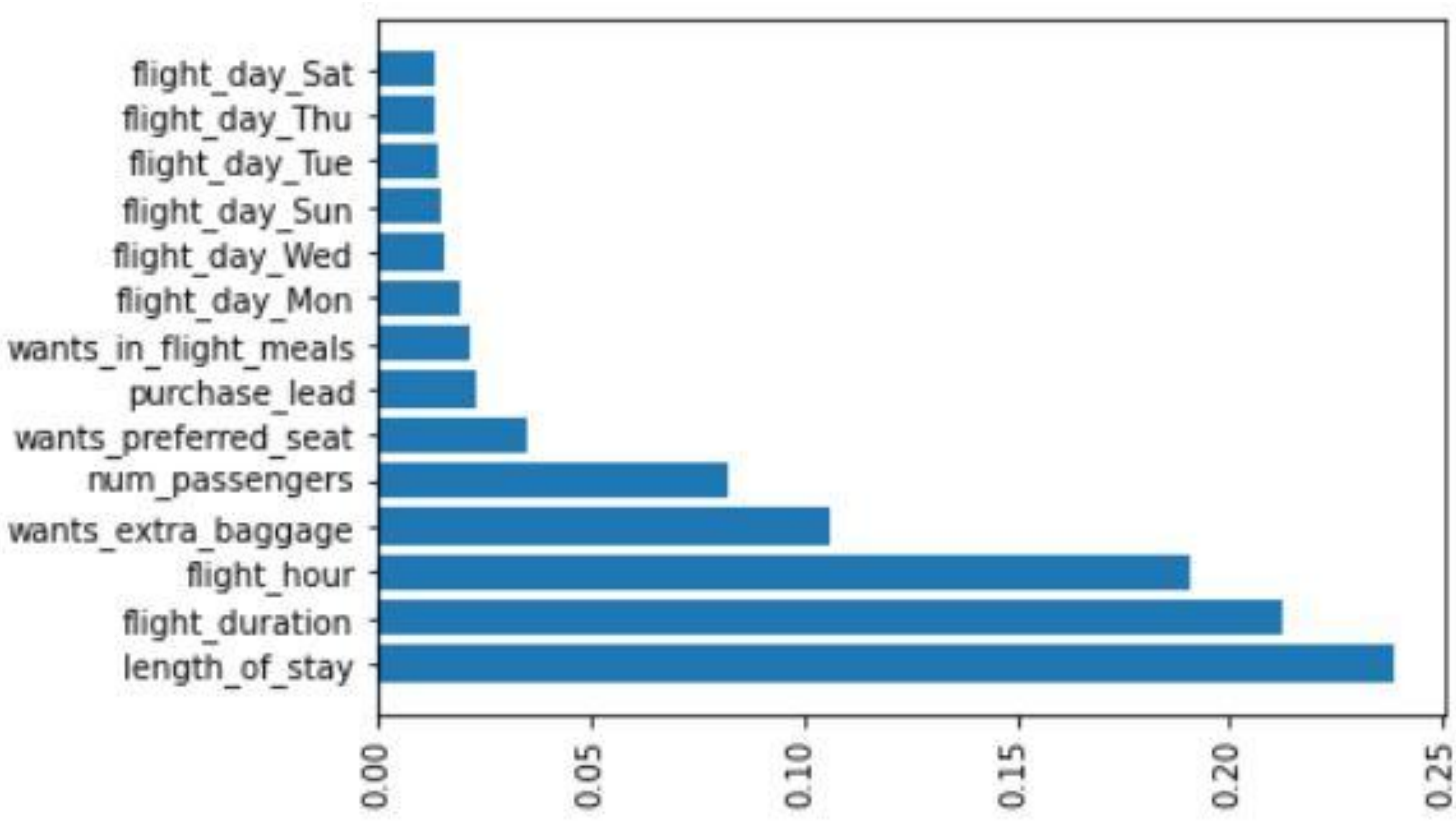
Out of 50000 booking entries only 15.0 % bookings were successfull or complete.

Travellers from which country had their booking complete.



This plot can be used by airline to provide more offers to customers from these countris.

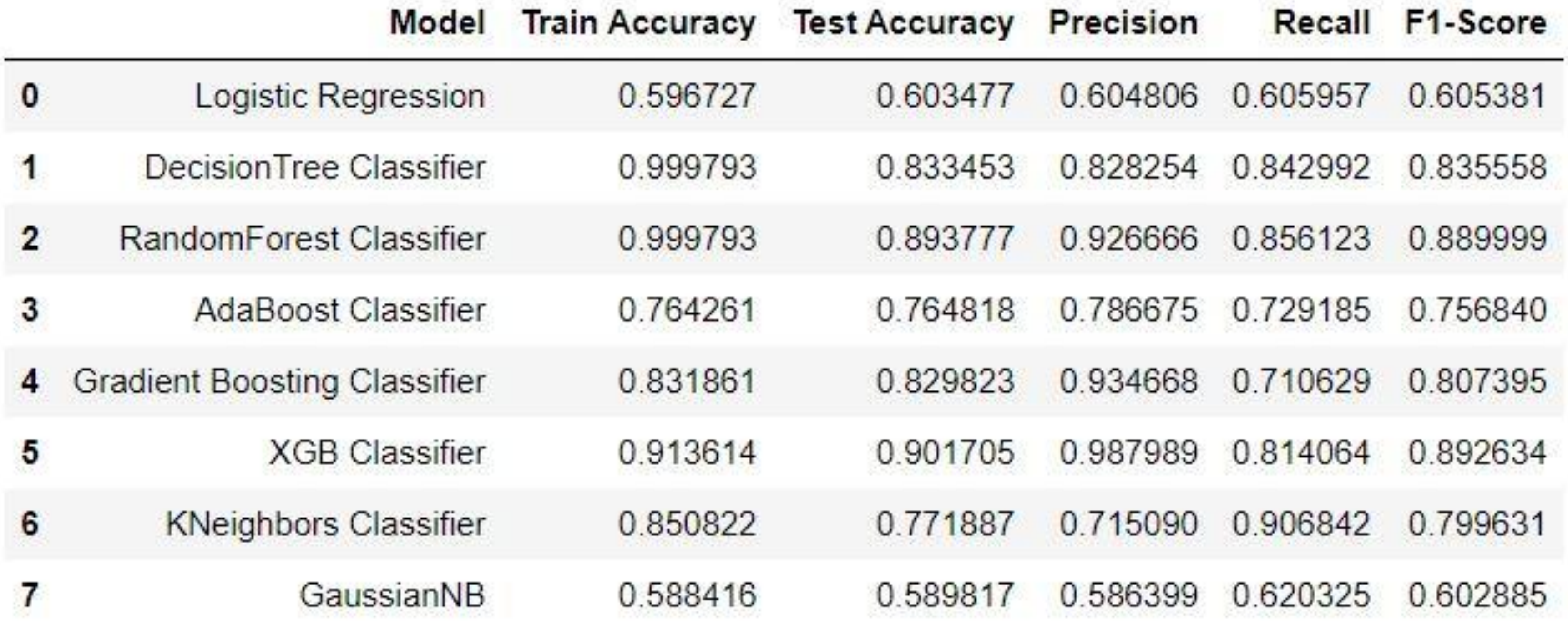
Important Features in our model



5. Model Building

- Separating the independent and dependent features.
- Splitting the data into train and test.
- Training various models.

6.Model Comparison



	Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.596727	0.603477	0.604806	0.605957	0.605381
1	DecisionTree Classifier	0.999793	0.833453	0.828254	0.842992	0.835558
2	RandomForest Classifier	0.999793	0.893777	0.926666	0.856123	0.889999
3	AdaBoost Classifier	0.764261	0.764818	0.786675	0.729185	0.756840
4	Gradient Boosting Classifier	0.831861	0.829823	0.934668	0.710629	0.807395
5	XGB Classifier	0.913614	0.901705	0.987989	0.814064	0.892634
6	KNeighbors Classifier	0.850822	0.771887	0.715090	0.906842	0.799631
7	GaussianNB	0.588416	0.589817	0.586399	0.620325	0.602885

Best Model

- Models Logistic Regression and GaussianNB are rejected because they were not able to generate high accuracy on both Train and test data.
- Models AdaBoost Classifier,Gradient Boosting show less variation but accuracy is low if compared with other models. KNeighbors Classifier shows high variation between train and tst accuracy.
- Models with highest accuracy on training data is RandomForest & Decision tree. But accuracy of both models on Test data is very low as compared with train accuracy. Both models shows a variation on train and test data.
- Hence XGB Model is chosen as the best mdoel because of low variation between train and test accuracy. Also, we can improve the accuracy of the model using cross validation.

Key insights

XGB Classifier

Recall
81%

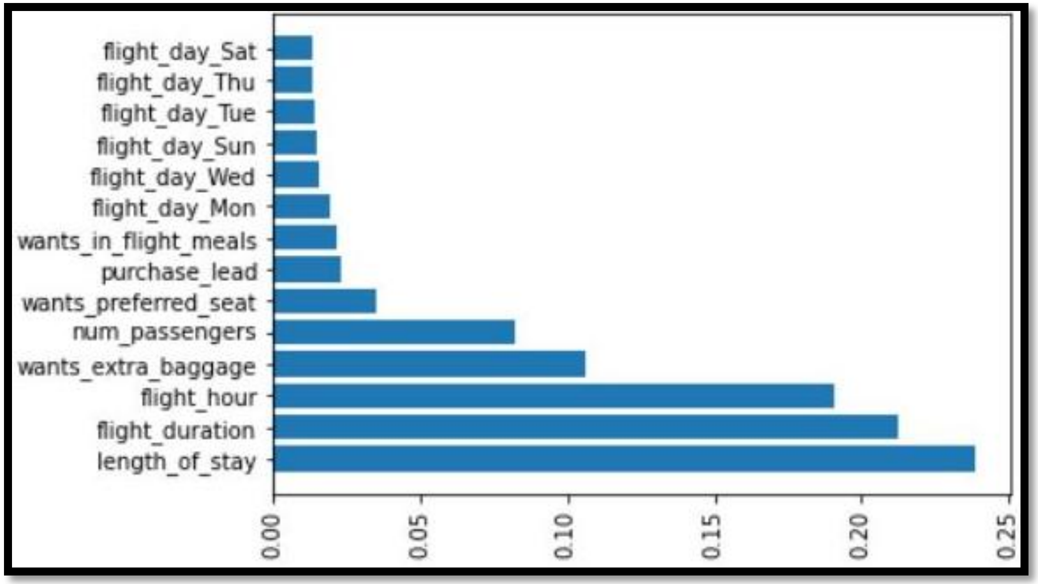
Chance of predicting true successful bookings.

Precision
98%

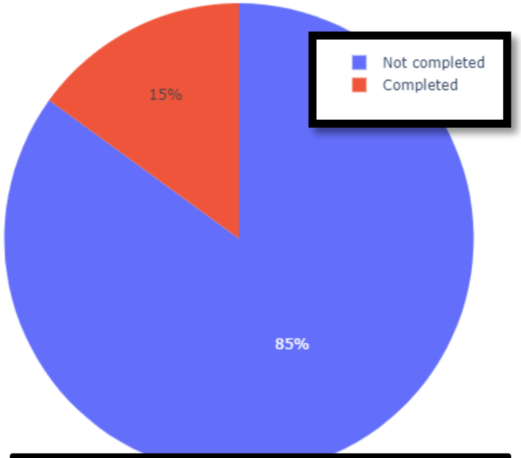
actually completed bookings out of all successfully completed bookings.

Accuracy
91%

Accuracy of the model predicting successful or incomplete booking



Dataset contains more negative instances than positive instances. In this case, the model will likely have a higher precision score, but a lower recall score.



Only 15.0 % bookings were completed

- Highest enquires are from “Australia” but maximum successful booking s are from Malaysia.
- “Round Trip” is more preferred.
- Highest No of flights is on Monday and lowest on Saturday.