CAPSTONE 4

AIRLINE PASSENGER SATISFACTION

CONTENT:

INTRODUCTION

- **Target Stakeholder:**
  - The CEO of the airlines from the USA
  - The management of the **customer services** department
    - Ground Services
    - Flight/On-Board Services
    - Online/Internet/IT
  - Airlines Stock/Shareholders

INTRODUCTION

- **Problem statement:**
  - It is well-known for the Airlines in USA to provide MEMORABLE services
  - Hence we would like to analyst the factors that brings satisfactory or unsatisfactory ratings
  - Therefore hopefully creating a better airline branding that inspire confidence to all its stakeholders

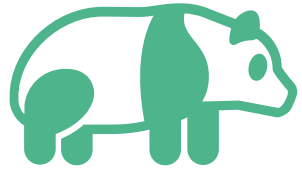
METHODOLOGY

- Source: **Kaggle**
  - https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

METHODOLOGY

Model: Decision tree, Random Forest

Metric: F1-Score



Tools: Pandas, Matplotlib, scikit-learn, etc

DATA PREPARATION

- Dataset overview

```
Data columns (total 25 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Unnamed: 0                        129880 non-null  int64
 1   id                                129880 non-null  int64
 2   Gender                            129880 non-null  object
 3   Customer Type                     129880 non-null  object
 4   Age                               129880 non-null  int64
 5   Type of Travel                    129880 non-null  object
 6   Class                             129880 non-null  object
 7   Flight Distance                   129880 non-null  int64
 8   Inflight wifi service             129880 non-null  int64
 9   Departure/Arrival time convenient 129880 non-null  int64
 10  Ease of Online booking            129880 non-null  int64
 11  Gate location                     129880 non-null  int64
 12  Food and drink                    129880 non-null  int64
 13  Online boarding                   129880 non-null  int64
 14  Seat comfort                      129880 non-null  int64
 15  Inflight entertainment            129880 non-null  int64
 16  On-board service                  129880 non-null  int64
 17  Leg room service                  129880 non-null  int64
 18  Baggage handling                  129880 non-null  int64
 19  Checkin service                   129880 non-null  int64
 20  Inflight service                  129880 non-null  int64
 21  Cleanliness                       129880 non-null  int64
 22  Departure Delay in Minutes        129880 non-null  int64
 23  Arrival Delay in Minutes          129487 non-null  float64
 24  satisfaction                      129880 non-null  object
```

DATA PREPARATION

- **Data Cleaning**
  - Remove unnecessary data
  - Fill in and remove missing data
  - Convert Categorical / Object / String Variables to Numeric Data

In [19]:
```
# Drop unnecessary columns

Air_df = Air_df.drop(['id', 'Unnamed: 0'], axis=1)
```

In [18]:
```
# this will show only features that have nonzero missing values
Air_df_na[Air_df_na!=0]
```

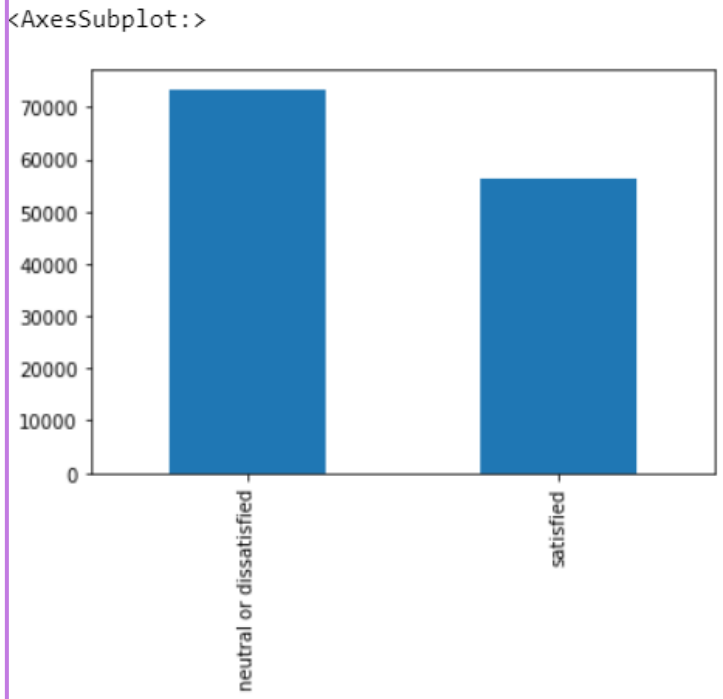
Out[18]:
```
Arrival Delay in Minutes    393
dtype: int64
```

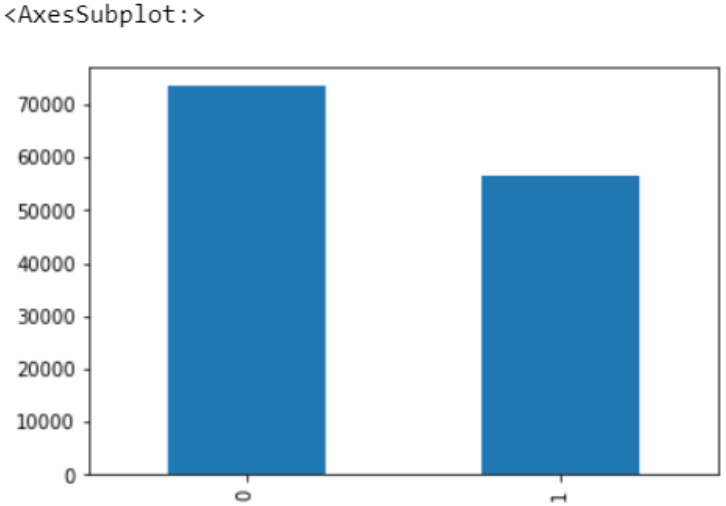
In [21]:
```
# Imputing missing value with mean (Alternative is just to drop the column)
# Arrival_Delay_in_Minutes has missing value
# fill the missing values with the average flight delay time. Because don't want model to be affected by this parameter.

Air_df['Arrival_Delay_in_Minutes'] = Air_df['Arrival_Delay_in_Minutes'].fillna(Air_df['Arrival_Delay_in_Minutes'].mean())
```

In [23]:
```
# Air_df = pd.get_dummies(Air_df['Gender'], dtype=float)

Air_df = pd.get_dummies(data=Air_df, columns=['Gender', 'Customer_Type', 'Type_of_Travel', 'Class'])
```

EDA & DATA ANALYSIS

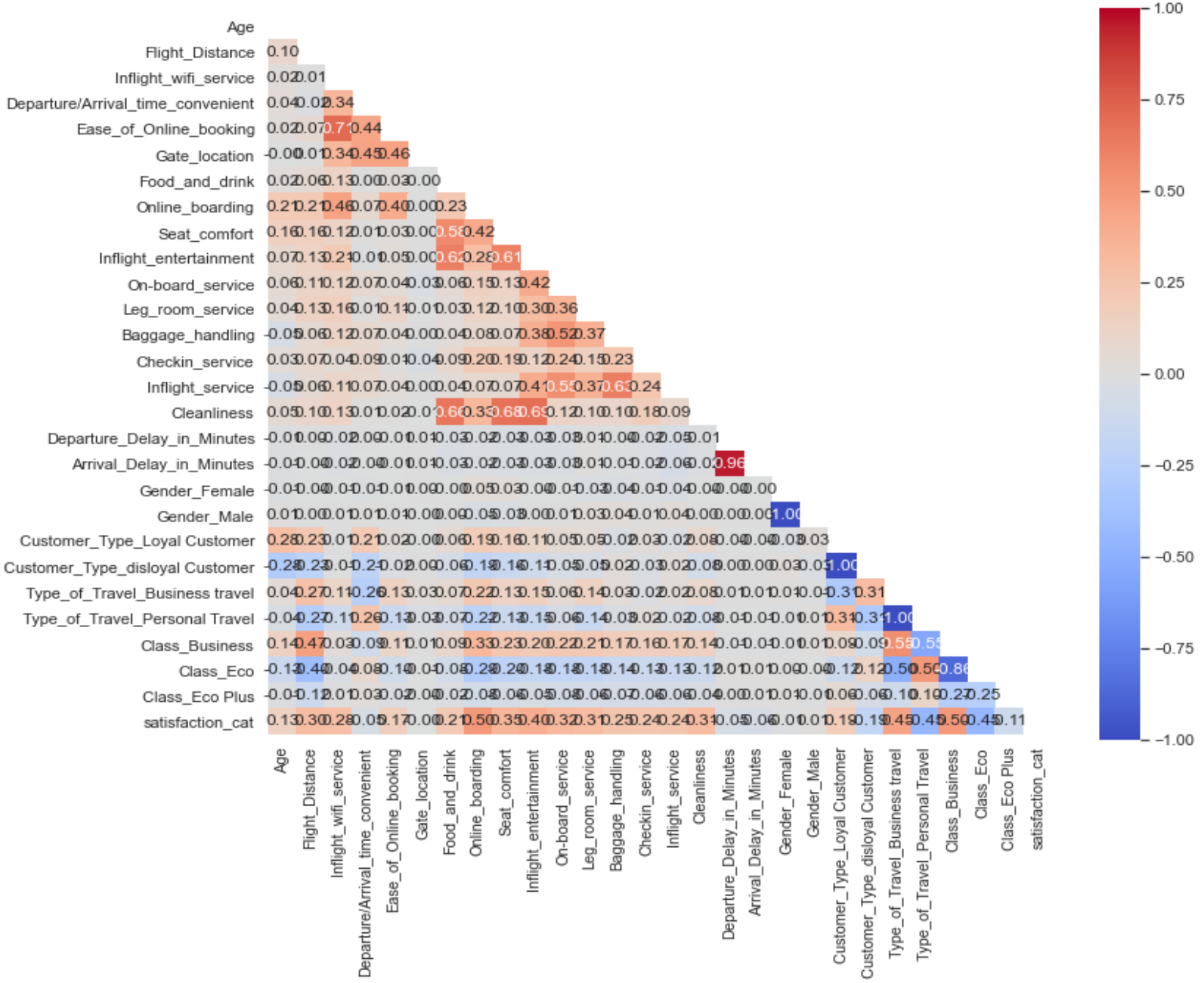
- Visualize how well balanced is the target (dependent variable)
- Target (dependent variable) is **Satisfaction**





EDA & DATA ANALYSIS

- Heat Map

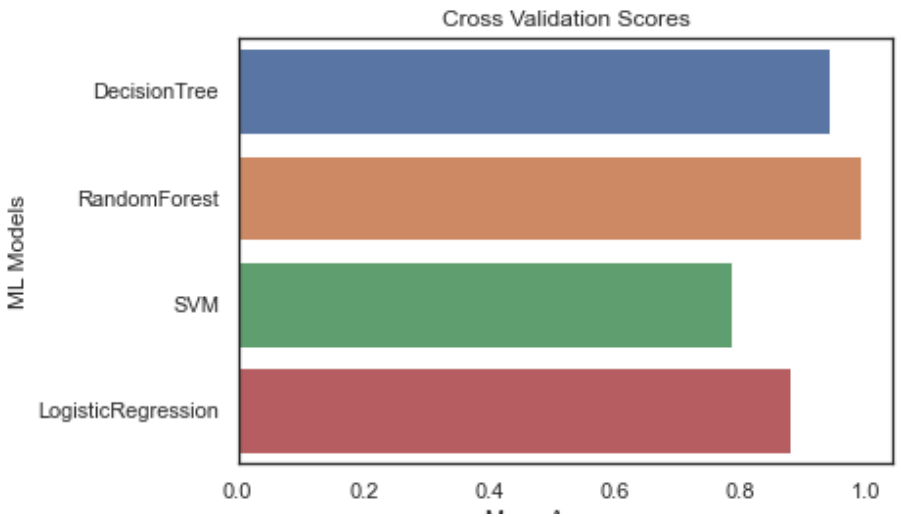


ML MODEL TRAINING

- Main Training Model
  1. Decision Tree
  2. Random Forest
- Alternative (“Trying Out”) Model
  1. SVM
  2. Logistic Regression

ML MODEL TRAINING

```
Text(0.5, 1.0, 'Cross Validation Scores')
```

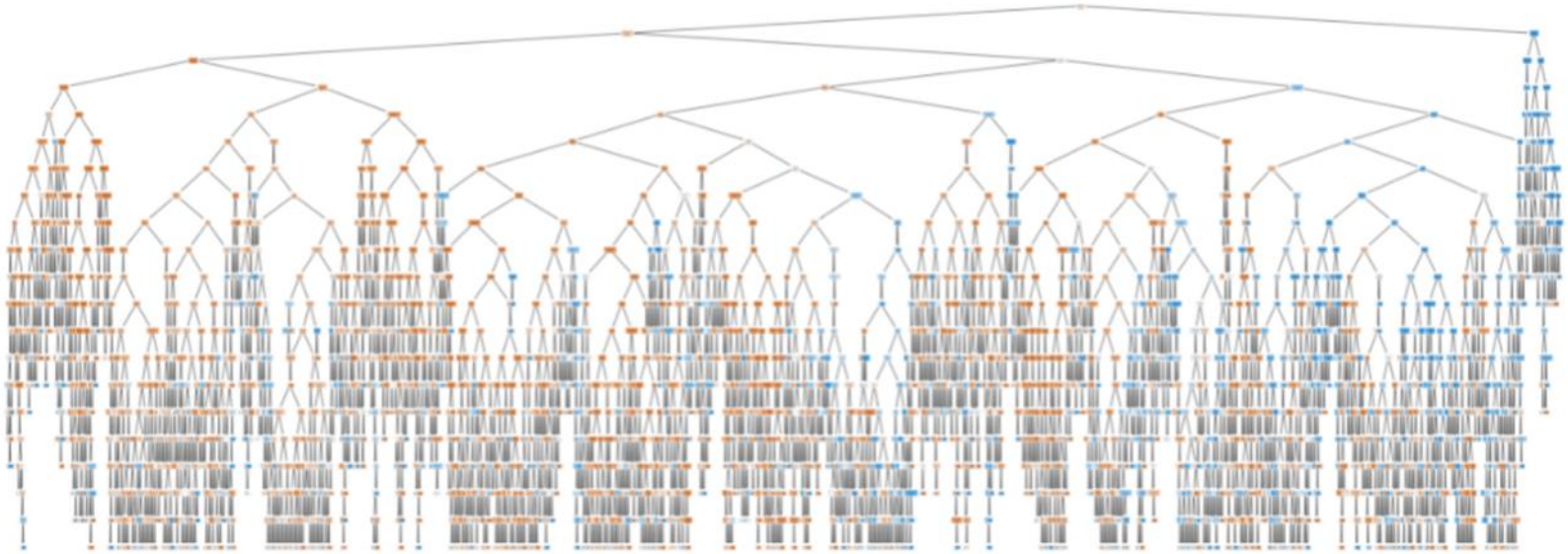


	Score
DecisionTree	0.942912
RandomForest	0.994264
SVM	0.788199
LogisticRegression	0.881849

ML MODEL TRAINING

- **Decision Tree Visualisation**

```
# Visualize the individual tree
from sklearn import tree

fn = X.columns
cn = ["Class_0","Class_1", "Class_2"]

fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (40,15))

tree.plot_tree(classifier.estimators_[1],
               feature_names = fn,
               class_names=cn,
               filled = True);
```



ML MODEL TRAINING

- **Testing the ML Model**

In [64]:
```
# Kept aside some data to test - X_test
y_pred = classifier.predict(X_test)

compare_df = pd.DataFrame({"Desired Output (Actuals)": y_test,
                           "Predicted Output": y_pred})
```

In [65]:
```
compare_df[:10]
```

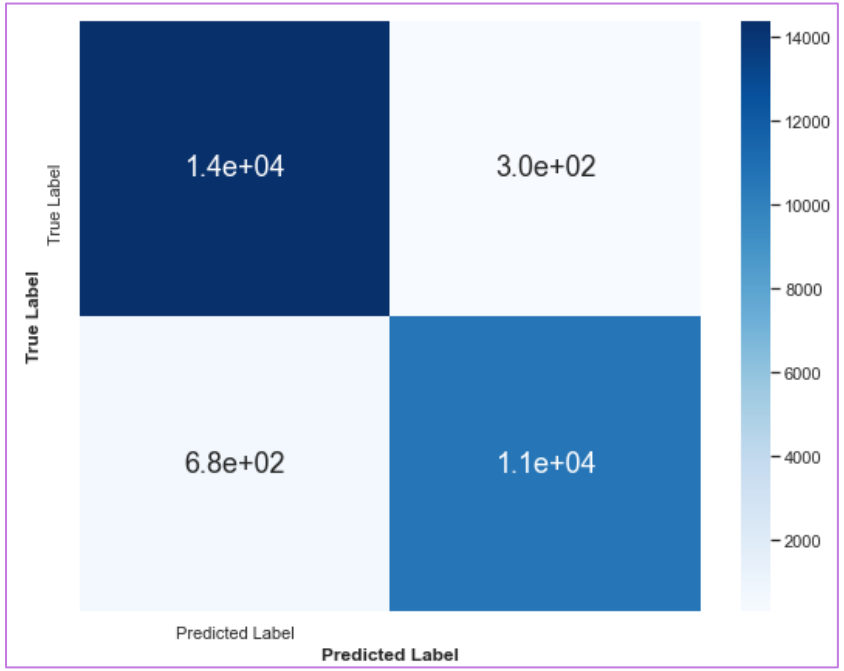
Out[65]:

	Desired Output (Actuals)	Predicted Output
22682	0	0
12418	0	0
24993	1	1
2429	0	0
43539	1	1
42104	0	0
29518	1	1
92724	0	0
6131	1	1
52580	0	0

ML MODEL TRAINING

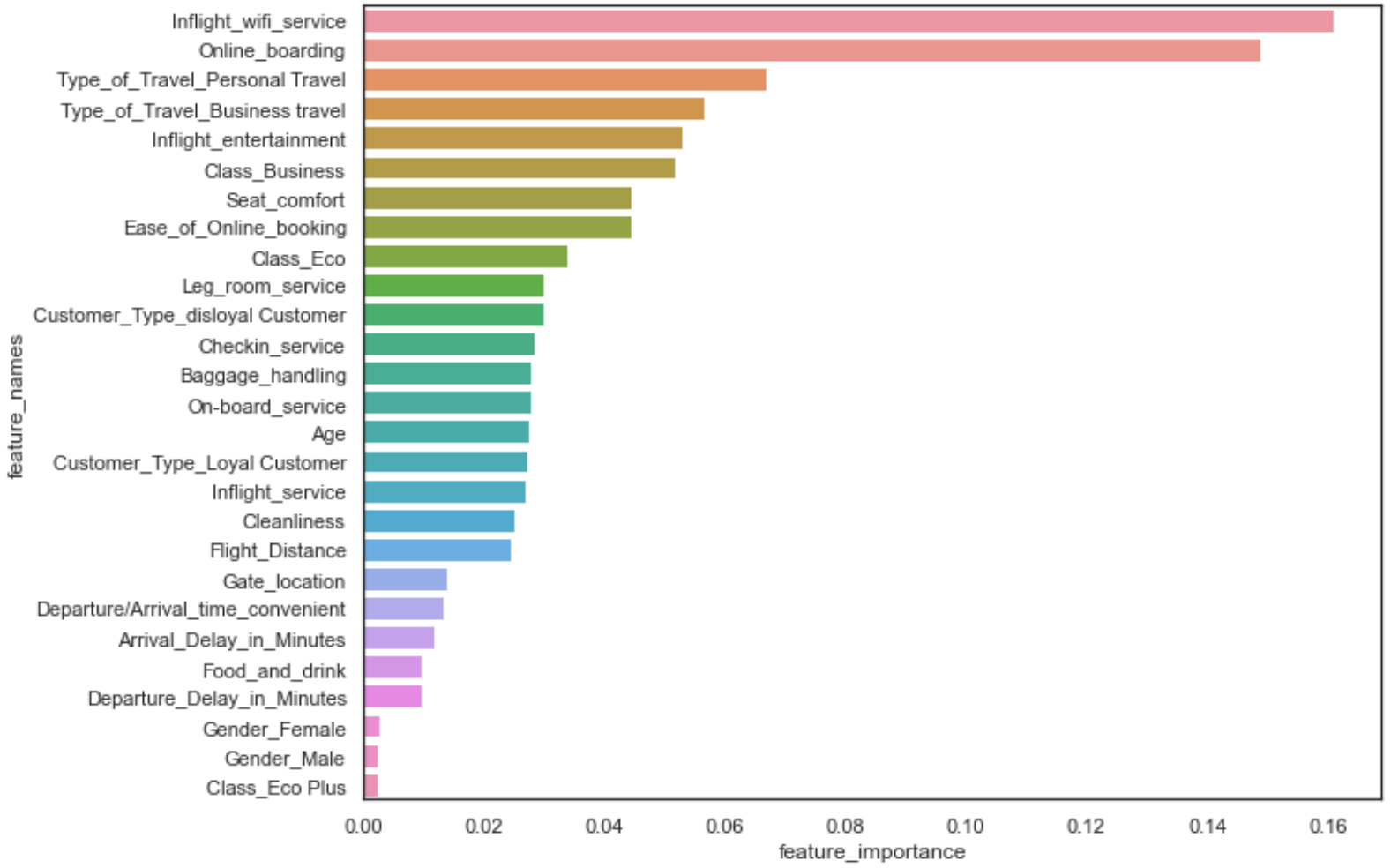
- **EVALUATE THE MODEL**

```
Classification report:

              precision    recall  f1-score   support

           0       0.95      0.98      0.97     14690
           1       0.97      0.94      0.96     11286

    accuracy                           0.96     25976
   macro avg       0.96      0.96      0.96     25976
weighted avg       0.96      0.96      0.96     25976


Confusion Matrix:

array([[14385,   305],
       [  679, 10607]], dtype=int64)
```



EVALUATION

- **Insights**
  - **Feature Importance**



EVALUATION

- Limitations of the Airline Dataset:
  - The Dataset is created in 2019, it does not reflect the changes in the industry due to Covid-19
  - Hence, the result of this dataset is very limited for the current year as of 2021
- Limitations for using Random Forest:
  - It requires a lot of computational power as it builds numerous trees to combine their outputs
  - Large number of trees can make the algorithm too slow and ineffective for real-time predictions

CONCLUSION

Random Forest is the best ML Model for this Dataset

In-Flight Wifi and Online Boarding are very important satisfaction factors

Passenger on personal travel have higher satisfaction compared to business travel

Business class passengers have higher satisfaction compared to both economy classes