CAPSTONE 4

AIRLINE PASSENGER SATISFACTION

CONTENT:

INTRODUCTION

- **Target Stakeholders:**
  - 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

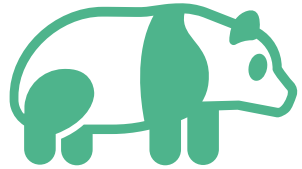
METHODOLOGY

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

METHODOLOGY

Model: Decision tree, Random Forest

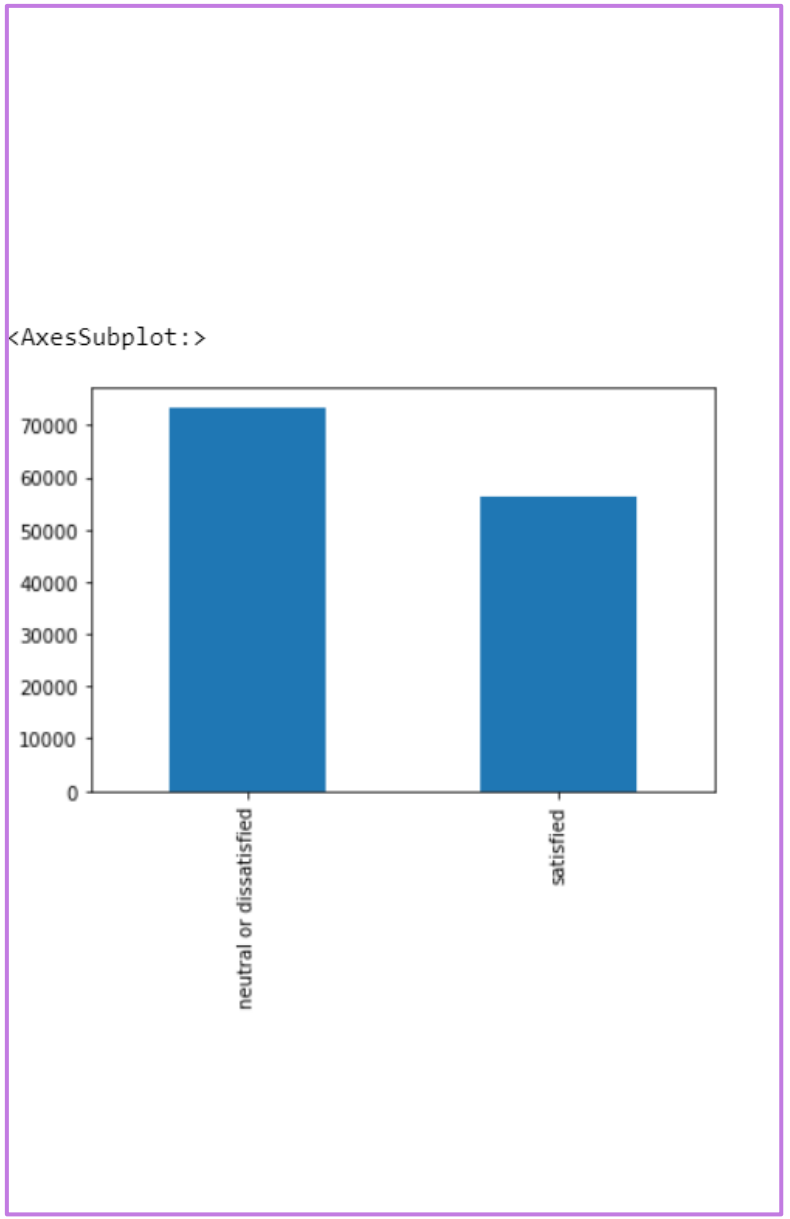
Metric: F1-Score

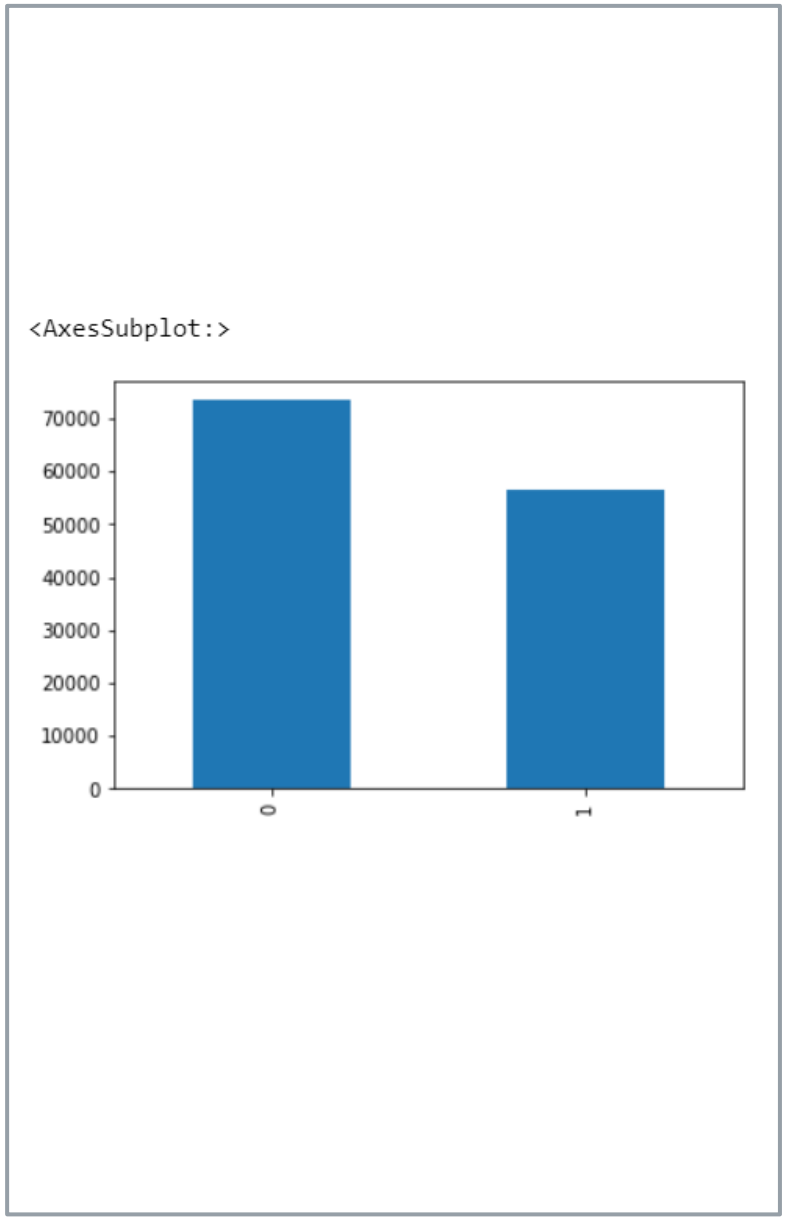


Tools: Pandas, Matplotlib, scikit-learn, etc

EXPLORATORY DATA ANALYSIS (EDA)

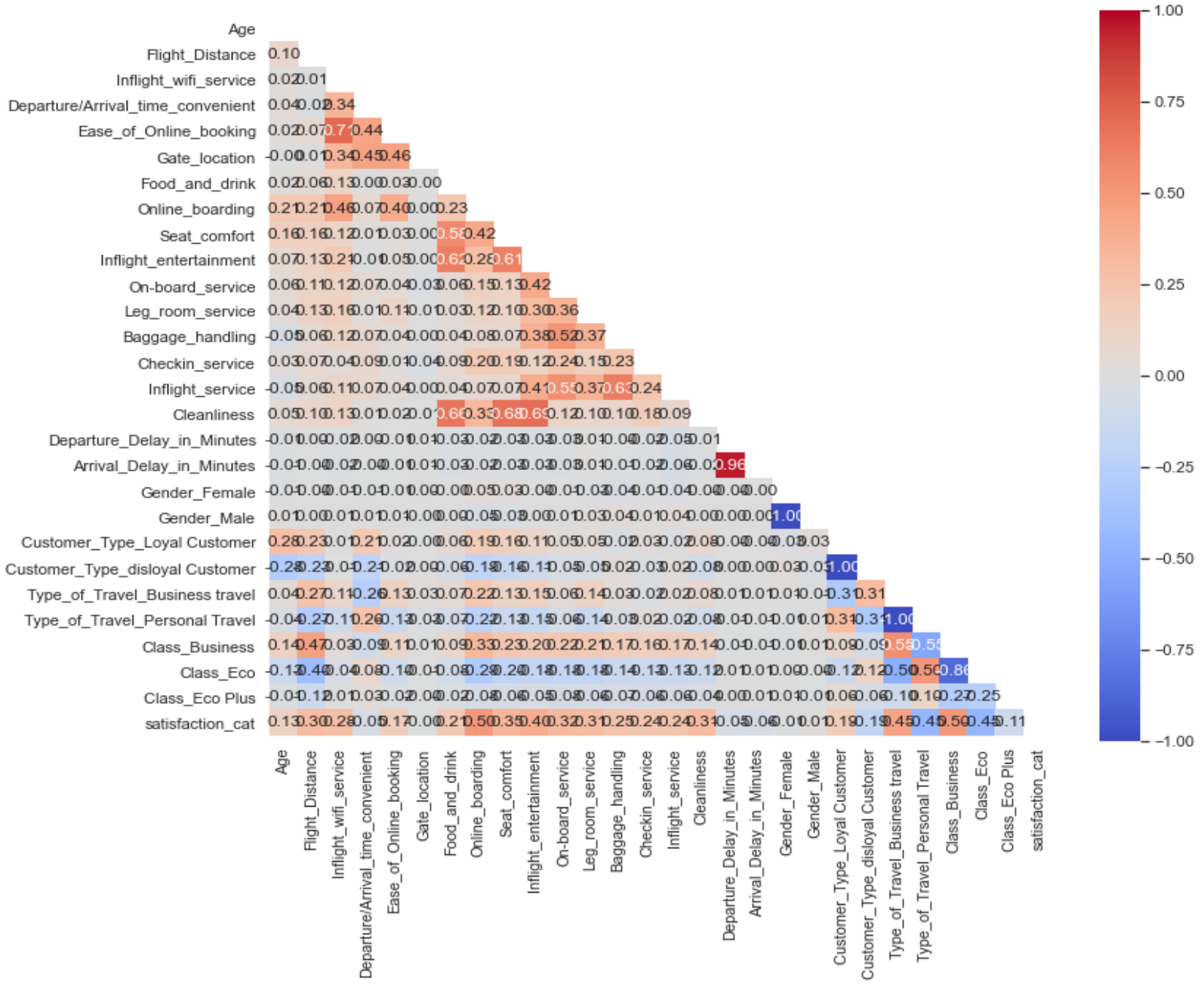
- Visualization of how well is the target (dependent variable) balanced
- Target (dependent variable) for this project is Satisfaction





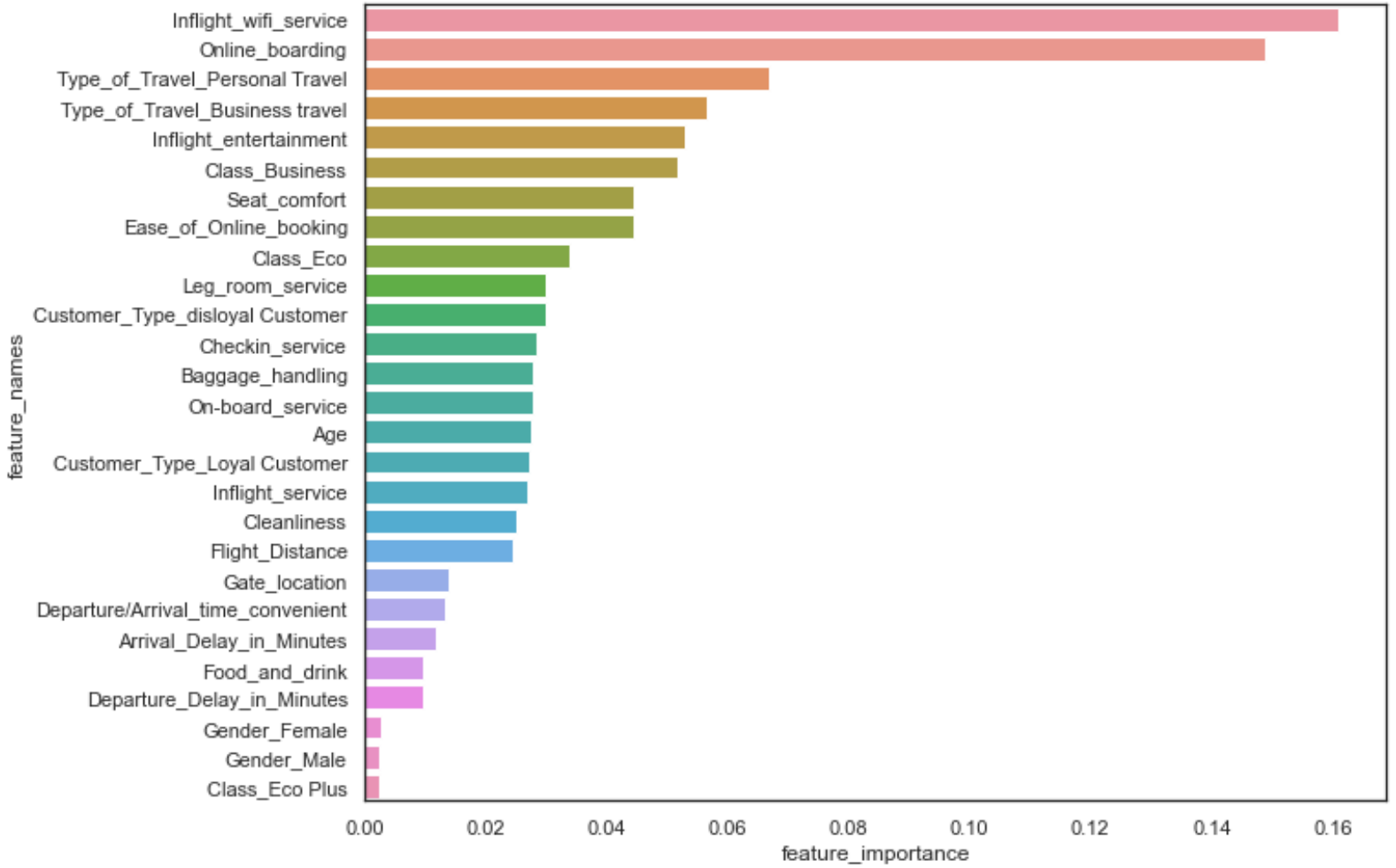
DATA ANALYSIS

- **Correlation Matrix Heatmap Visualization**



DATA ANALYSIS

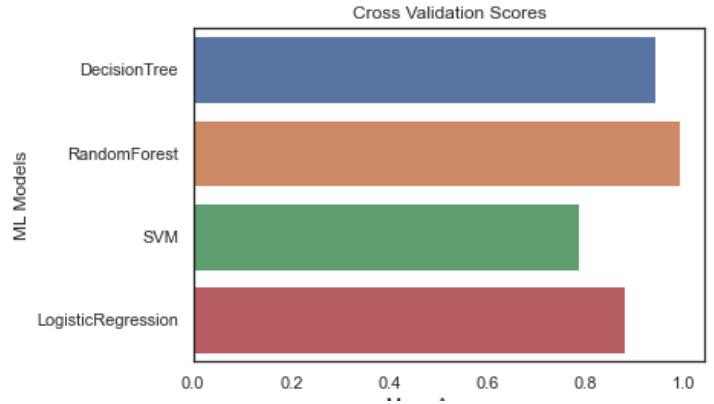
- **Bar Plot Visualization on Feature Importance**



MACHINE LEARNING MODEL TRAINING

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

MACHINE LEARNING MODEL TRAINING



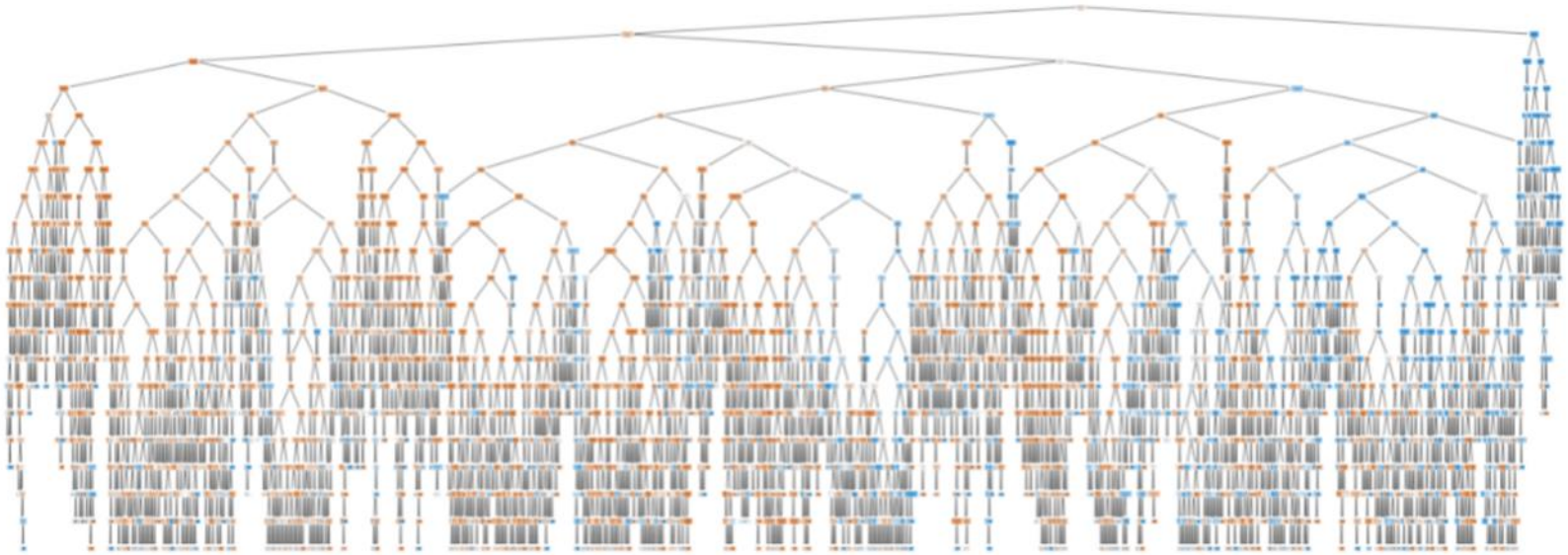
	Score
DecisionTree	0.942912
RandomForest	0.994264
SVM	0.788199
LogisticRegression	0.881849

- **Cross validation scores between each model**

MACHINE LEARNING 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);
```



MACHINE LEARNING MODEL TRAINING

- **Testing the Machine Learning 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]
```

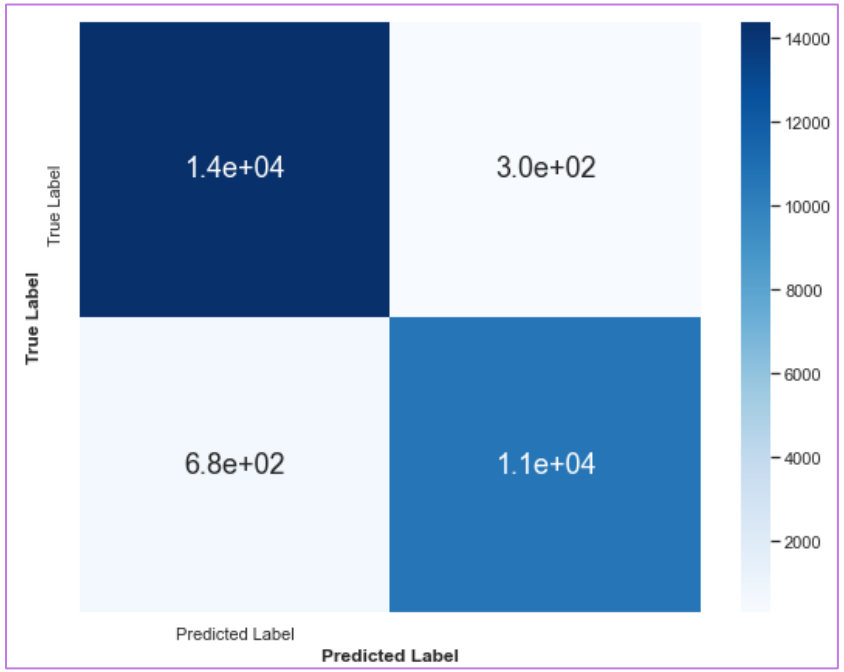
Out[65]:

	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

MACHINE LEARNING MODEL TRAINING

- **Evaluate the model with confusion matrix**

```
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

- 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