# AIRLINE CUSTOMER SATISFACTION

**Group 7**

Antrika Bachloo, Akhil Datta Bhimana, Abhilash Gattu,
Sai Suma Maguluri, Sri Niharika Singamsetty

## Table of Contents

# AIRLINE CUSTOMER SATISFACTION

## Business Problem:

Although there is a lot of airline customer evaluation data, they are not enough to accurately predict consumer happiness in daily life. They often fall short when it comes to determining the data's worth and creating a customer satisfaction prediction model using customer assessment data. In this study, an experiment of correlation analysis between customer assessment data is used to analyze airline customer happiness.

## Introduction:

The primary goal is to forecast if a potential customer will be satisfied with the service, given the specifics of the parameter values. This project aims to guide an airline company to determine the important factors that influence customer or passenger satisfaction.

Customer satisfaction plays a major role in affecting the business of a company therefore analyzing and improving the factors that are closely related to customer satisfaction is important for the growth and reputation of a company

Using feature selection, we have understood the important factors that are affecting the satisfaction rate.

Using Random Forest, KNN, Decision Tree, and AdaBoost, an airline customer satisfaction prediction model is constructed, and the accuracy of the model is compared.

**Data Description:**

❖ This data set contains a survey on air passenger satisfaction. The following classification problem is set:

❖ It is necessary to predict which of the two levels of satisfaction with the airline the passenger belongs to

❑ Satisfied

❑ Dissatisfied

❖ We have 1,29,880 rows and 23 columns in our dataset.

❖ Dependent variable/Target variable: satisfaction

❖ Independent variables: Attributes like Gender, Customer Type, Age, etc.

```
: air_df.columns

: Index(['satisfaction', 'Gender', 'Customer Type', 'Age', 'Type of Travel',
       'Class', 'Flight Distance', 'Seat comfort',
       'Departure/Arrival time convenient', 'Food and drink', 'Gate location',
       'Inflight wifi service', 'Inflight entertainment', 'Online support',
       'Ease of Online booking', 'On-board service', 'Leg room service',
       'Baggage handling', 'Checkin service', 'Cleanliness', 'Online boarding',
       'Departure Delay in Minutes', 'Arrival Delay in Minutes'],
      dtype='object')

: air_df.dtypes

: satisfaction                        object
  Gender                              object
  Customer Type                       object
  Age                                  int64
  Type of Travel                      object
  Class                               object
  Flight Distance                      int64
  Seat comfort                         int64
  Departure/Arrival time convenient    int64
  Food and drink                       int64
  Gate location                        int64
  Inflight wifi service                int64
  Inflight entertainment               int64
  Online support                       int64
  Ease of Online booking               int64
  On-board service                     int64
  Leg room service                     int64
  Baggage handling                     int64
  Checkin service                      int64
  Cleanliness                          int64
  Online boarding                      int64
  Departure Delay in Minutes           int64
  Arrival Delay in Minutes           float64
  dtype: object
```

## Data Cleaning:

1. **Removing null values:**

- Finding missing values is one of the most important steps in data cleaning.
- We used the command. isnull(). sum() to check for null values in our dataset/columns. The column corresponding to the Arrival Delay in Minutes feature has 393 missing values. Many columns contain categorical values but are of type 'object' or 'int64'. Let's replace this type with a special one designed for storing categorical values.
- To handle this, the column was dropped since it had many null values.

```
In [11]: df_air.isnull().sum()

Out[11]: satisfaction                        0
         Gender                              0
         Customer Type                       0
         Age                                 0
         Type of Travel                      0
         Class                               0
         Flight Distance                     0
         Seat comfort                        0
         Departure/Arrival time convenient   0
         Food and drink                      0
         Gate location                       0
         Inflight wifi service               0
         Inflight entertainment              0
         Online support                      0
         Ease of Online booking              0
         On-board service                    0
         Leg room service                    0
         Baggage handling                    0
         Checkin service                     0
         Cleanliness                         0
         Online boarding                     0
         Departure Delay in Minutes          0
         Arrival Delay in Minutes          393
         dtype: int64
```

2. **Removing duplicate values:**

- Our second step was removing the duplicate values.
- Using. duplicated().sum() command, there were no duplicate values.

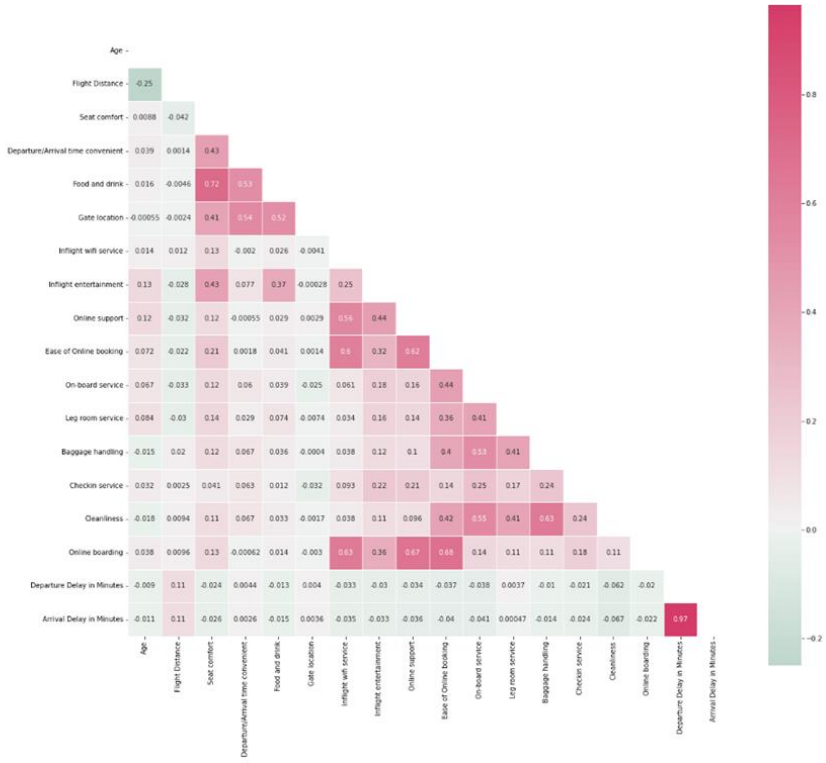
### Check for Duplicates

```
In [13]: df_air.duplicated().sum()

Out[13]: 0
```

## 3. Correlation:

- Correlation between columns can lead to overfitting and hence removing the highly correlated values is a very crucial step.
- There is a significant correlation between Departure Delay in Minutes and Arrival Delay in Minutes so we will drop one of 2 features to avoid the multicollinearity problem.




## 4. Standardization:

- Standardization is performed to bring down all the features to a common scale without distorting the range of the values.
- Here we used MinMaxScaler() to scale the range of selected variables.

Scaling

In [22]:
```
from sklearn import preprocessing
r_scaler = preprocessing.MinMaxScaler()
r_scaler.fit(df_air)
modified_data = pd.DataFrame(r_scaler.transform(df_air), columns=df_air.columns)
modified_data.head()
```

Out[22]:

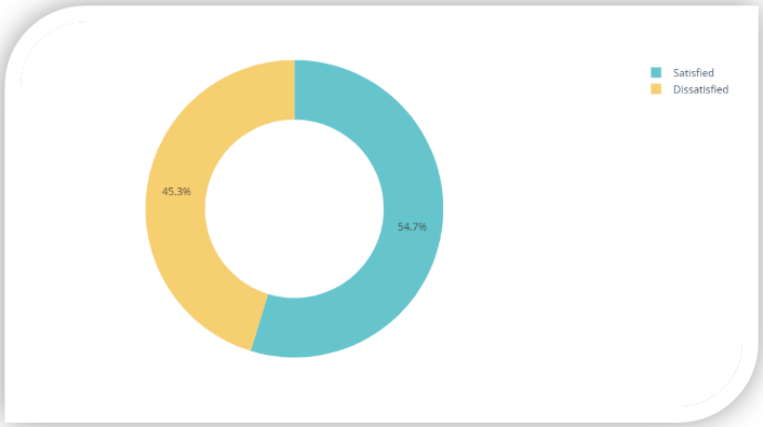
| | satisfaction | Gender | Customer Type | Age | Type of Travel | Class | Flight Distance | Seat comfort | Departure/Arrival time convenient | Food and drink | ... | Inflight entertainment | Online support | Ease of Online booking | On-board service | Leg room service |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 0.0 | 0.0 | 0.743590 | 1.0 | 0.5 | 0.031155 | 0.0 | 0.0 | 0.0 | ... | 0.8 | 0.4 | 0.6 | 0.6 | 0.0 |
| 1 | 1.0 | 1.0 | 0.0 | 0.512821 | 1.0 | 0.0 | 0.349804 | 0.0 | 0.0 | 0.0 | ... | 0.4 | 0.4 | 0.6 | 0.8 | 0.8 |
| 2 | 1.0 | 0.0 | 0.0 | 0.102564 | 1.0 | 0.5 | 0.302565 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.4 | 0.4 | 0.6 | 0.6 |
| 3 | 1.0 | 0.0 | 0.0 | 0.679487 | 1.0 | 0.5 | 0.083031 | 0.0 | 0.0 | 0.0 | ... | 0.8 | 0.6 | 0.2 | 0.2 | 0.0 |
| 4 | 1.0 | 0.0 | 0.0 | 0.807692 | 1.0 | 0.5 | 0.044052 | 0.0 | 0.0 | 0.0 | ... | 0.6 | 0.8 | 0.4 | 0.4 | 0.0 |

5 rows × 22 columns

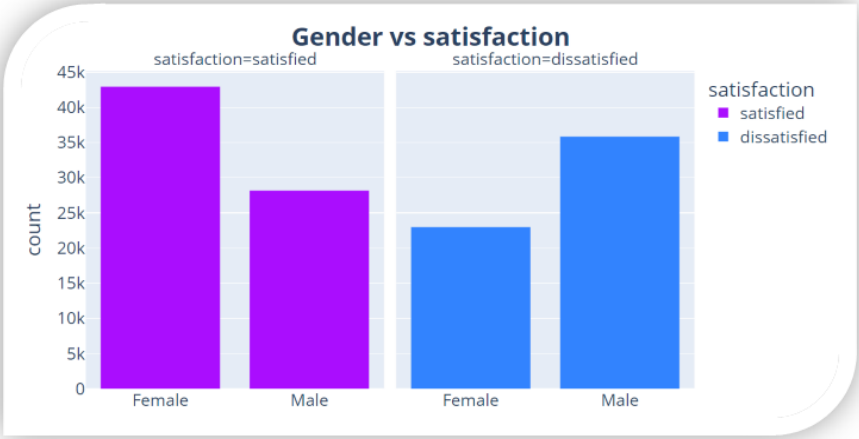
5. **<u>Encoding for categorical features:</u>**

- Encoding is performed to convert categorical variables like Gender, Type of Travel, Customer Type and Class into numerical values for this dataset.
- Here we used Label Encoder () to convert categorical variables to numerical values.

**Data Visualization:**

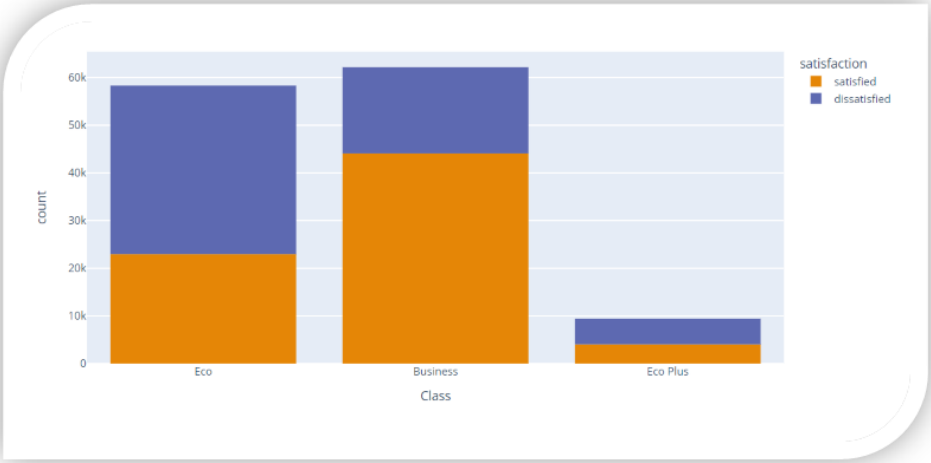



In the above plot, we see that survey results have 54.7% people satisfied and 45.3% dissatisfied.

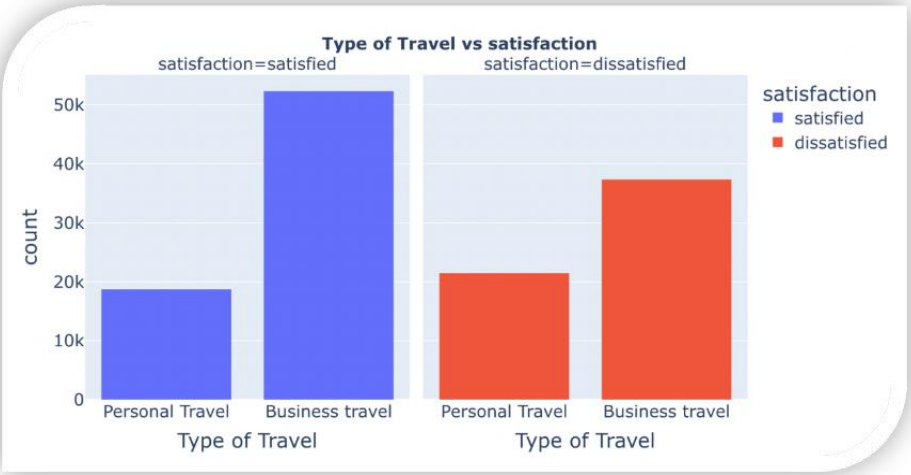


Based on the above gender level plot,

- The number of male and female are almost the same, there is a just a slight difference between the gender count
- Therefore, we are safe to say that percentage of male is dissatisfied more than the percentage of female




The above plot shows the satisfied and dissatisfied count based on class of booking. We see that the Economy class has more dissatisfied people than the rest.



Based on the above type of travel level plot, we see that

- We have two types of Travel which is ‘Personal Travel’ and ‘Business Travel’
- Business travel customers are more than Personal travel customers, so we are dealing with Customers in Business Layer more.

## Machine Learning Models (Performance Evaluation)

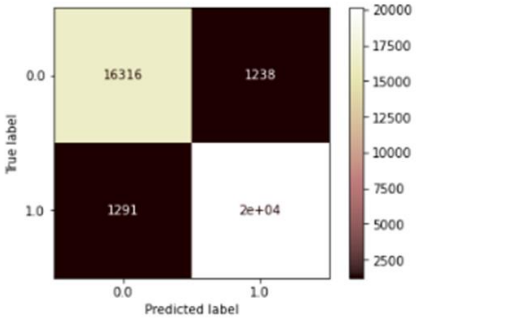
### ->Ensemble Learning Models

### Random Forest

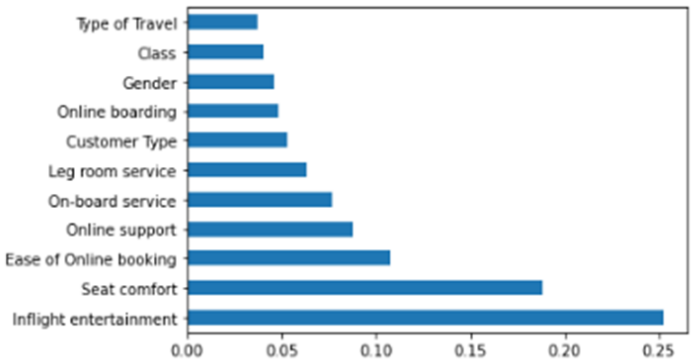
The sum of the boxes in the confusion matrix is the test data i.e., 38771

- Individual tree decisions are combined from multiple unpruned trees (Voting Classifiers + Bagging + Decision Tree)
- For each tree, the training sample will be a random sample with replacement and for each node, m attributes are chosen, and the best split will be found.
- Output is the class selected by most trees
- FN is lowest in Random Forest so accuracy is highest.
- Accuracy is 93.50%

```
Using model: Random Forest
Trainning Score: 0.9613709358088786
Test Score: 0.9350939328611025
Acc Train: 0.9613709358088786
Acc Test: 0.9350939328611025
======================================================
```



## Feature Importance

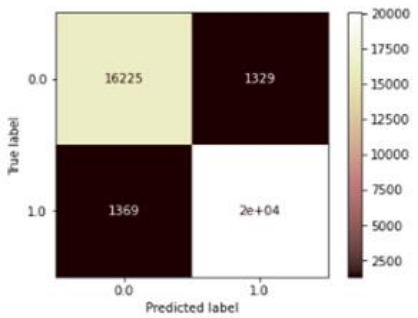



---

## Adaboost

- **Boosting** - Improving a single weak model by combining it with several other weak models to generate a collectively strong model
- It is called adaptive boosting as the weights are reassigned to each instance, with higher weights assigned to incorrectly classified instances.
- Here, each new tree considers the errors or mistakes made by the previous trees. Hence, every successive decision tree is built on the errors of the previous trees.
- Here our base model is a decision tree with best k = 15 for max depth obtained using hyperparameter tuning. We have used learning rate of 0.7 for this model
- Accuracy is 93%

```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1fea45d1f10>
```

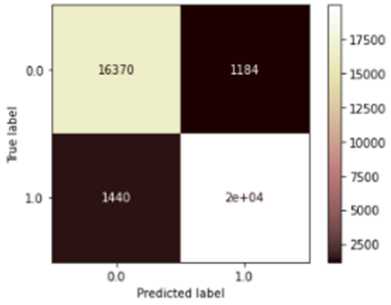


## ->Simple Classification Models

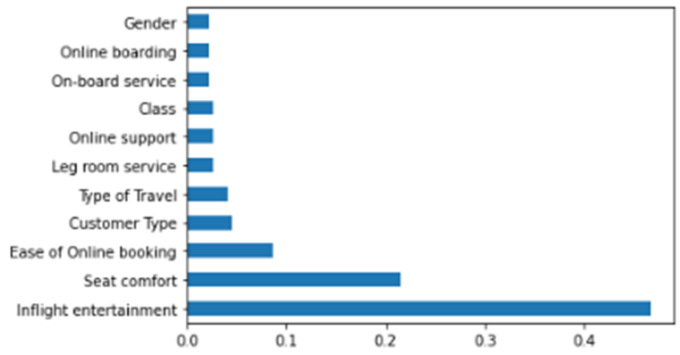
## Decision Tree with Hyperparameter Tuning

- Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter
- We did hyperparameter tuning by giving the depth of the tree to be considered between 1 and 20.
- Best parameters obtained to be with max_depth of 15.
- Accuracy of the model is obtained to be 93.26%.

```
Training Accuracy using grid search: 0.9479849531435611
Testing Accuracy using grid search: 0.9326557848270198
7]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1fea3c765e0>
```




## Feature Importance

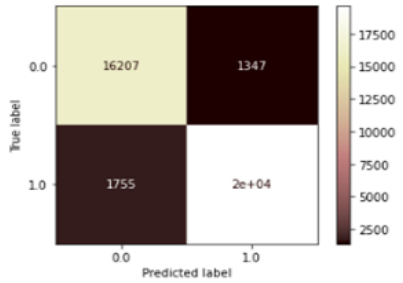


## KNN with Hyperparameter Tuning

- One of the simple supervised machine learning algorithms.
- For choosing the right value of k we did hyperparameter tuning with a range of (3 to 26) and grid search cross-validation to find the best model. Best k =7.
- Obtained an accuracy of 92.03%

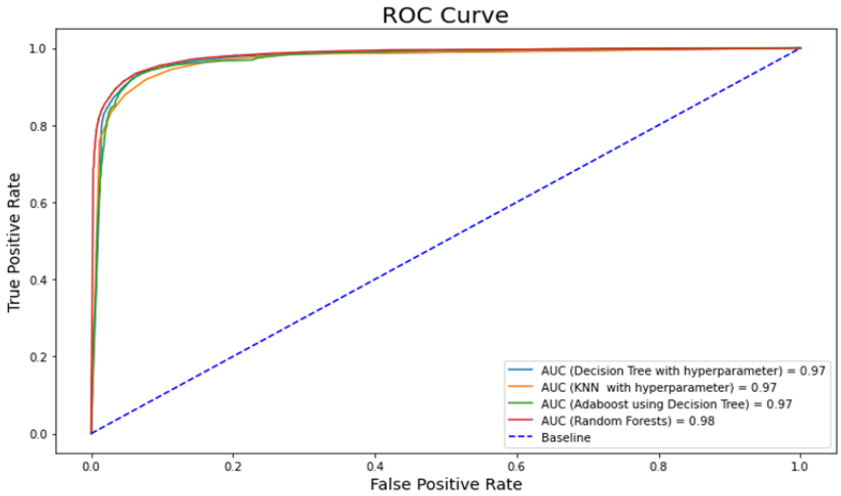
```
Best k is: {'n_neighbors': 7}
Mean validation score is: 0.9192111471729261
Training accuracy: 0.9203880505081614
Testing accuracy: 0.9203880505081614
```

```
]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x1fe8bba3b50>
```




## Model Comparison:

| Model | Accuracy |
|---|---|
| Random Forest | 0.93 |
| Decision Tree (with hyperparameter) | 0.93 |
| KNN with hyperparameter | 0.92 |
| Adaboost | 0.93 |




By comparing all four models and observing AUC and accuracy, Random Forest has the highest scores among the other models. Regression and classification can both be accomplished using Random Forest, a powerful machine-learning technique. Being an ensemble method, a random forest model is constructed from a variety of little estimators, or decision trees, each of which generates its predictions. The estimators' estimates are combined with the random forest model to get a more accurate prediction.

**Conclusion:**

- Based on accuracy and roc curve, we conclude that Random Forest yields best results for this classification data set with accuracy 0.93 and AUC value 0.98.
- Also, False negative values are lowest for Random Forest when compared with other models.
- Based on feature importance generated from random forest, we see that Inflight entertainment, Seat Comfort, Ease of online booking and online support are four important factors that affect customer satisfaction.
- Airlines can try and focus on these four important factors to improve customer satisfaction, mainly for economy class since they have high customer dissatisfaction.