Abstract

The customer review is very crucial and plays an important role in the framing the future business plan for a service providing firm. The performance of Airlines company can be evaluated based on the degree of passenger satisfaction. Here, in order to evaluate the performance of the company, we are considering the dataset containing the information of their customers such as Age, Gender, Travel class and details of the journey like Arrival and Departure delays and also other features that influences customer satisfaction level such as On-board service, Cleanliness, Seat comfort, Baggage handling etc. The dataset considered here contains 25975 rows and 25 columns.

The aim of this project is to evaluate the performance of the Airline company by predicting the passenger satisfaction with the airline services provided by the company. The factors which are highly influential on the customer satisfaction is needed to be find out which can help the company to find the areas which needs further improvement.

The methodology contains data preprocessing techniques such as data mining, data cleaning, data exploration and feature engineering. Then, predictive modelling will be done using various supervised learning techniques and best algorithm with highest accuracy will be selected among them. Then model fine tuning and finally web hosting will be performed.

1.2 Problem Statement

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:

1. *Satisfaction*
2. *Neutral or dissatisfied*

2. Introduction

The growth of the industry provides opportunities as well as challenges to the airlines companies. The opportunities arise because of increasing demands. Whereas the challenges arise from the other airlines and also in making long-term relationship with the customers. To overcome these challenges, Airlines have to be remained on toes. Airlines try best to make passengers have a rich experience every time they travel. Factors such as departure and arrival time, inflight-entertainment, seat comfort could be very crucial in enhancing the customer experience. Also, the factors vary between different age groups. This study seeks to explain the paramount factors which impact the passenger satisfaction in the Airline industry and change in those factors across different age groups.

The customer review is very crucial and plays an important role in the framing the future business plan for a service providing firm. The performance of Airlines company can be evaluated based on the degree of passenger satisfaction.

4. Data Understanding

These are the following information about the passengers of some airline:

1. **Gender:** male or female
2. **Customer type:** regular or non-regular airline customer
3. **Age:** the actual age of the passenger
4. **Type of travel:** the purpose of the passenger's flight (personal or business travel)
5. **Class:** business, economy, economy plus
6. **Flight distance**
7. **Inflight wifi service:** satisfaction level with Wi-Fi service on board (0: not rated; 1-5)
8. **Departure/Arrival time convenient:** departure/arrival time satisfaction level (0: not rated; 1-5)
9. **Ease of Online booking:** online booking satisfaction rate (0: not rated; 1-5)
10. **Gate location:** level of satisfaction with the gate location (0: not rated; 1-5)
11. **Food and drink:** food and drink satisfaction level (0: not rated; 1-5)
12. **Online boarding:** satisfaction level with online boarding (0: not rated; 1-5)
13. **Seat comfort:** seat satisfaction level (0: not rated; 1-5)
14. **Inflight entertainment:** satisfaction with inflight entertainment (0: not rated; 1-5)
15. **On-board service:** level of satisfaction with on-board service (0: not rated; 1-5)
16. **Leg room service**: level of satisfaction with leg room service (0: not rated; 1-5)
17. **Baggage handling:** level of satisfaction with baggage handling (0: not rated; 1-5)
18. **Checkin service:** level of satisfaction with checkin service (0: not rated; 1-5)
19. **Inflight service:** level of satisfaction with inflight service (0: not rated; 1-5)
20. **Cleanliness:** level of satisfaction with cleanliness (0: not rated; 1-5)
21. **Departure delay in minutes**
22. **Arrival delay in minutes**

5. Data Visualisation

5.1 Univariate Analysis

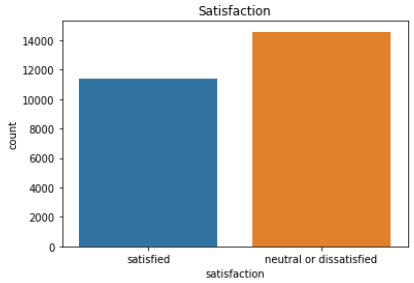


Figure 1: Count plot of satisfaction rate

The no. of 'neutral or dissatisfied' customers and 'satisfied' customers are 14573 and 11403 respectively. This plot shows a distribution of around 56% and 44% between 'neutral or dissatisfied' customers and 'satisfied' customers respectively.

So, the data is quite balanced, and it does not require any special treatment or re-sampling.

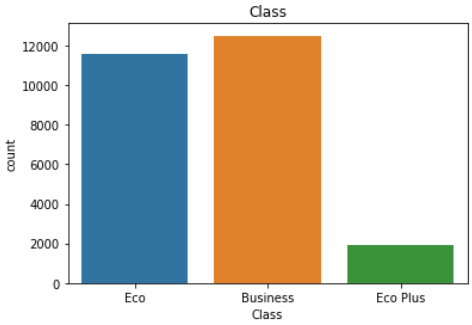


Figure 2: Count plot of 'Class' of passengers

Most of the customers are from 'Business' class. And 'Eco Plus' class customers are least among all.

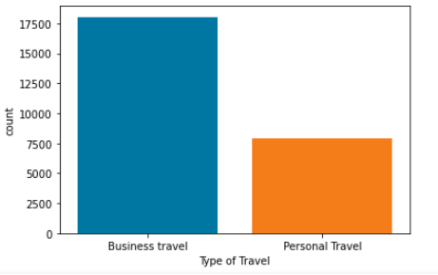


Figure 3: Count plot of 'Type of Travel' of passengers

Most of the customers are travelling for 'Business' purpose.

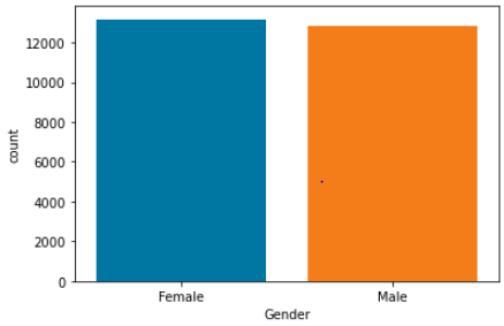


Figure 3: Count plot of 'Gender' of passengers

The gender ratio is almost equal among the customers.

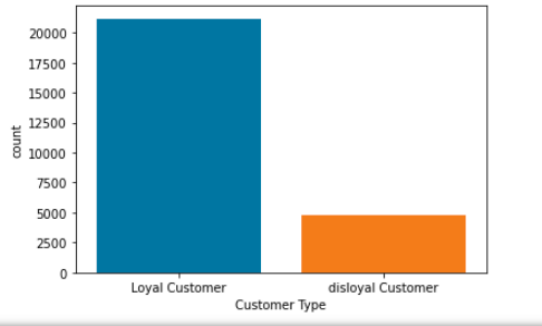


Figure 3: Count plot of 'Customer Type' of passengers

The 'Loyal Customers' are more compared to 'Disloyal Customers'. This shows that, the company is good at providing the services in overall.

5.2 Multivariate Analysis

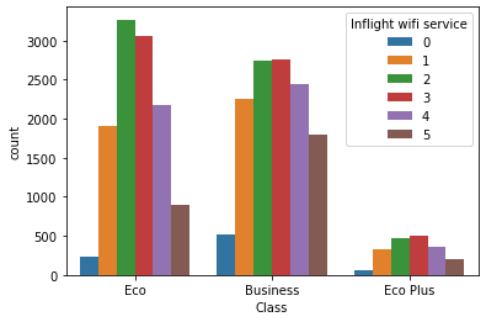


Figure 4: Countplot of rating of 'Inflight wifi service' given by different 'Class'

Wifi service is poorly rated by all 3 classes. So, the Airline company should improve this area of service.

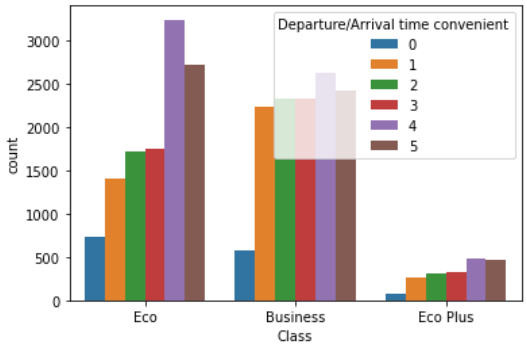


Figure 5: Countplot of rating of 'Departure/Arrival time convenient' given by different 'Class'

Departure/Arrival time convenient is highly rated by Eco class.

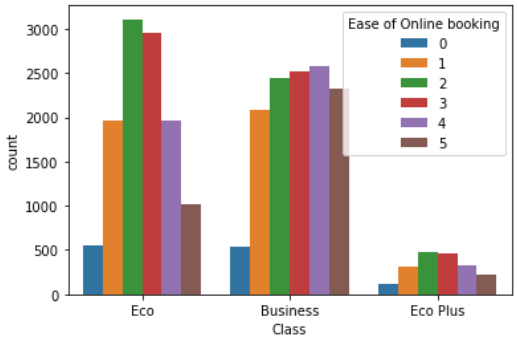


Figure 6: Count plot of rating of 'Ease of Online booking' given by different 'Class'

Ease of Online booking is poorly rated by Eco class. This shows that, the online booking service is not much appropriate for 'Eco' class and needs improvement.

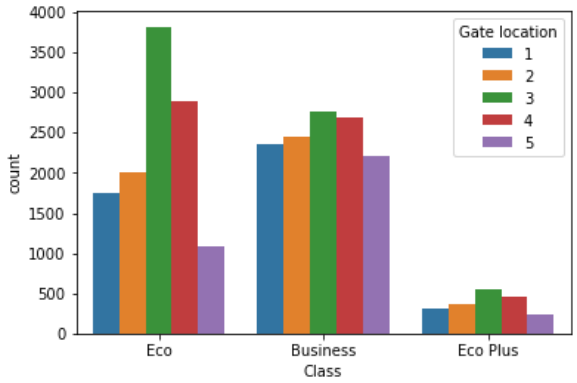


Figure 7: Count plot of rating of 'Gate location' given by different 'Class'

Rating of 'Gate location' is only average among all the 'Class' of customers. The company should really focus on this area of service

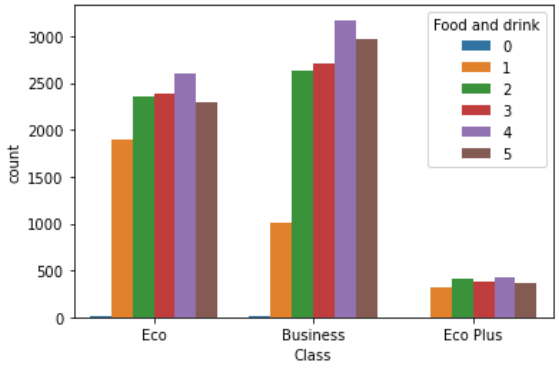


Figure 8: Count plot of rating of 'Food and drink' given by different 'Class'

For 'Food and Drink', all 'Class' of customers has given higher rating. Hence, company is doing well on this service.

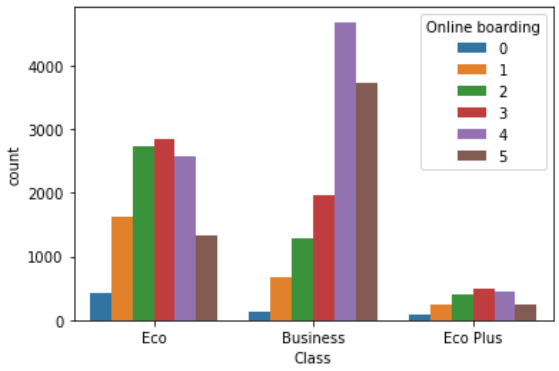


Figure 9: Count plot of rating of 'Online boarding' given by different 'Class'

For 'Online Boarding', Business Class has given higher rating than other classes. 'Eco' and 'Eco Plus' Classes have given average rating.

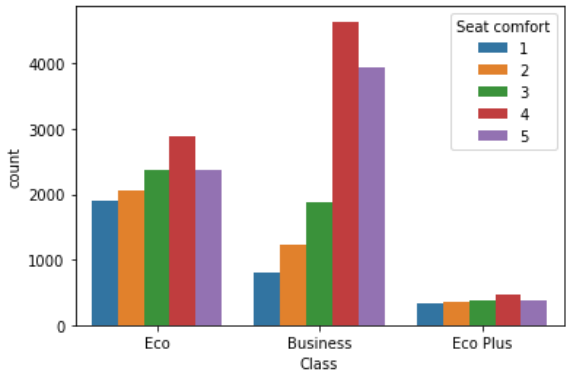


Figure 10: Count plot of rating of 'Seat comfort' given by different 'Class'

For ‘Seat comfort’, all ‘Class’ of customers has given higher rating. Hence, company is doing well on this service.

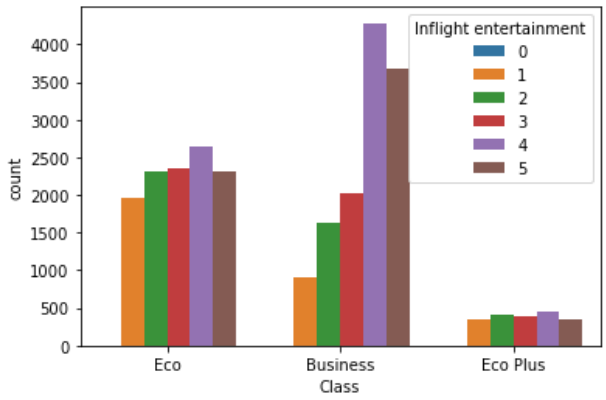


Figure 11: Count plot of rating of ‘Inflight entertainment’ given by different ‘Class’

For ‘Inflight entertainment’, all ‘Class’ of customers has given higher rating. Hence, company is doing well on this service.

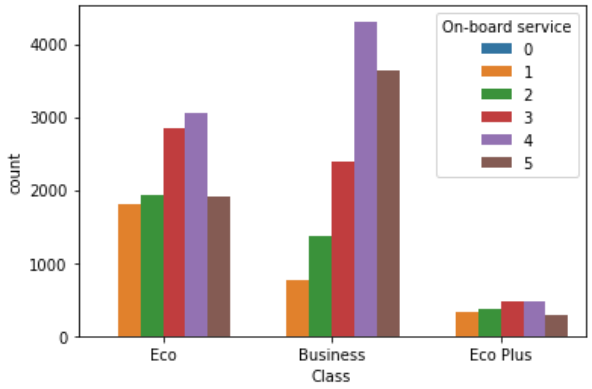


Figure 12: Count plot of rating of ‘On-board service’ given by different ‘Class’

For ‘On-board service’, all ‘Class’ of customers has given higher rating. Hence, company is performing well on this service.

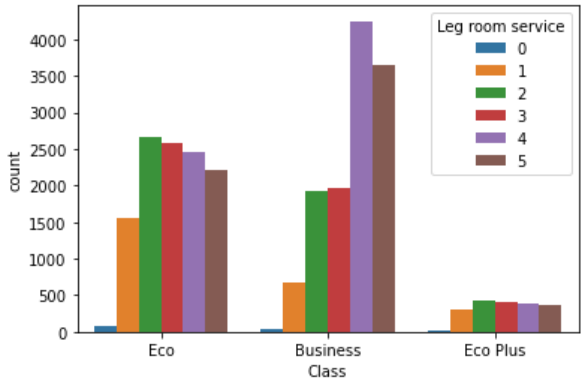


Figure 13: Count plot of rating of ‘Leg room service’ given by different ‘Class’

For ‘Leg room service’, ‘Business’ Class of customers has given higher rating, but ‘Eco’ Class customers are not satisfied.

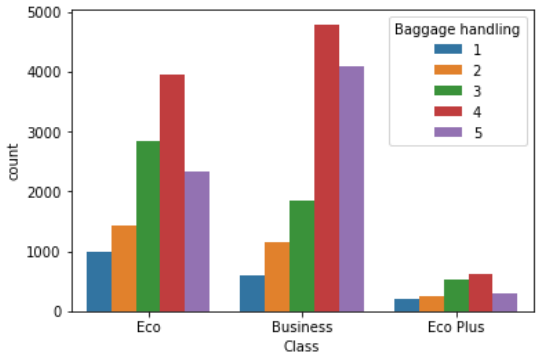


Figure 14: Count plot of rating of ‘Baggage handling’ given by different ‘Class’

For ‘*Baggage handling*’, all Class of customers has given higher rating, and shows the customers are satisfied with the service.

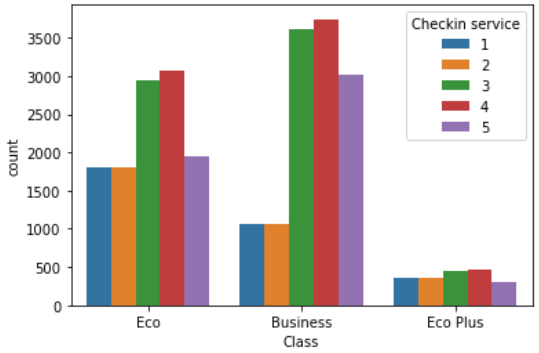


Figure 15: Count plot of rating of ‘Check-in Service’ given by different ‘Class’

For ‘*Check-in Service*’, all Class of customers has given higher rating, and shows the customers are satisfied with the service.

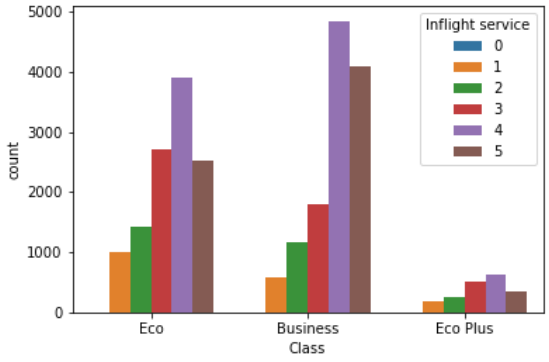


Figure 16: Count plot of rating of ‘Inflight Service’ given by different ‘Class’

For ‘*Check-in Service*’, all Class of customers has given higher rating, and shows the customers are satisfied with the service.

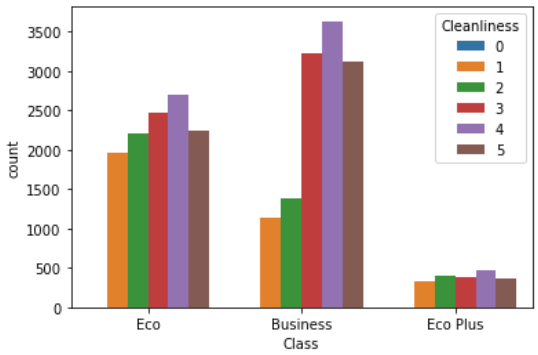


Figure 17: Count plot of rating of 'Cleanliness' given by different 'Class'

For '*Cleanliness*', all Class of customers has given higher rating, and shows the customers are satisfied with the service.

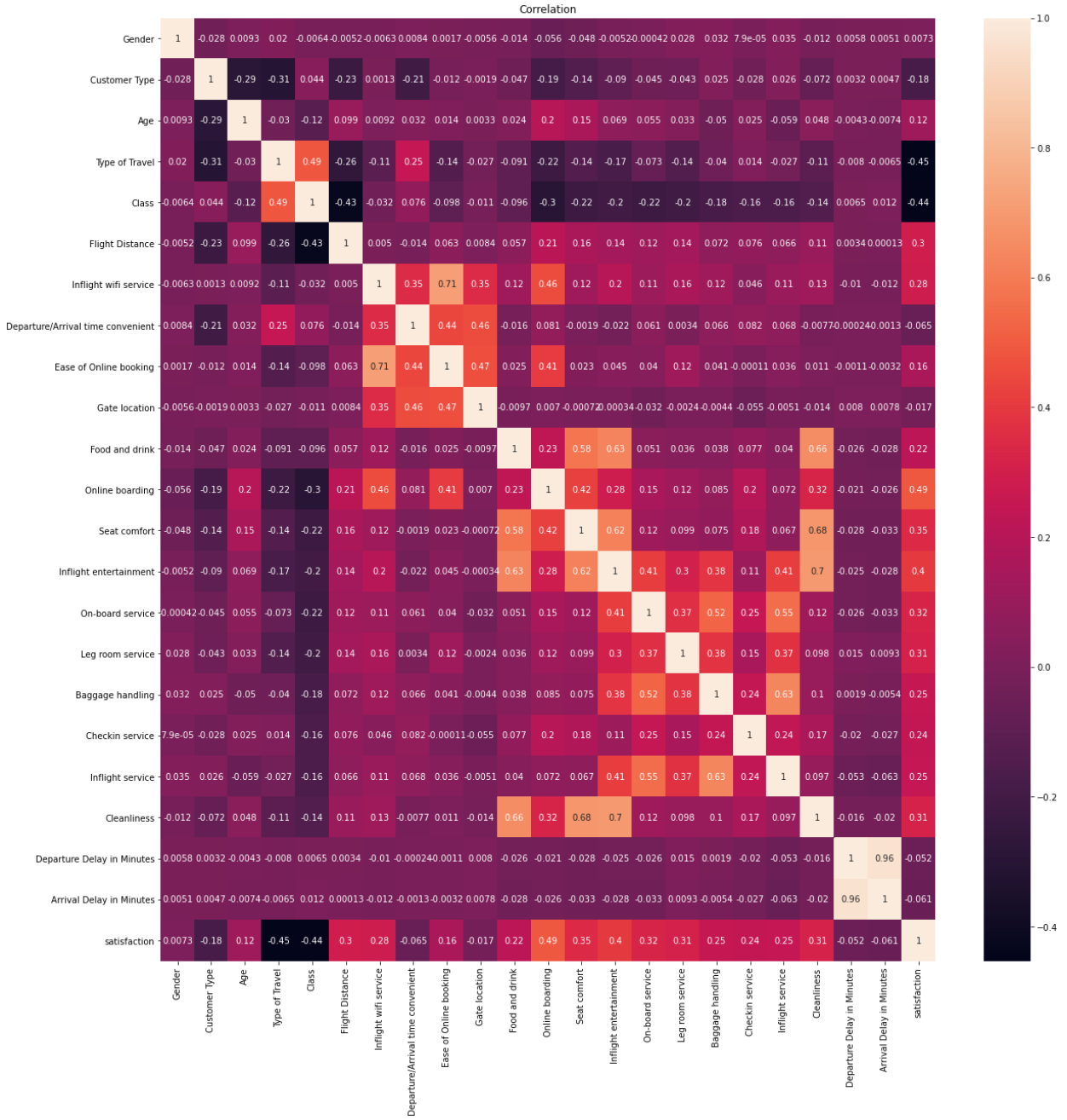


Figure 18: Correlation matrix

We get evidences that some of the more important variables for predicting satisfaction are online boarding experience, in-flight entertainment, the cabin class and whether a passenger travelling for business reason instead or for personal reasons.

6.Data Pre-Processing

Preprocessing is the process of making a data model ready. It involves several stages such as missing value removal, handling white spaces, making calculations, and splitting of data into train and test. Pre-processing was the most time-consuming part of this project as the data needed cleaning and transformation to get its model ready. The data we get from different sources may contain inconsistent data, missing values and repeated data. To get proper prediction results, the dataset must be cleaned, missing values must be taken care of either by deleting or by filling with mean values or some other method. Also, redundant data must be removed or eliminated to avoid biasing of the results. Some datasets may have some outlier or extreme values which also must be removed to get good prediction accuracy.

6.1 DATA CLEANING

Data cleaning is the process of **fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset**. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. In short, Machine Learning is data-driven. With data cleaning in place, **your Machine Learning model will perform better**. So, it is important to process data before use. Without quality data, it is foolish to expect a correct output.

Here we have done the following processes :

• “Unnamed:0” column in the dataset is set as the Index column.

• Column “id” is dropped since it has no significance in the analysis

6.2 CHECKING FOR UNIQUE VALUES

Unique values are **the distinct values that occur only once in the dataset** or the first occurrences of duplicate values counted as unique values.

Here the dataset is checked for unique values in each column.

- Majority of columns(14) have values range from 0 to 5.
- Four columns are with distribution of two values.

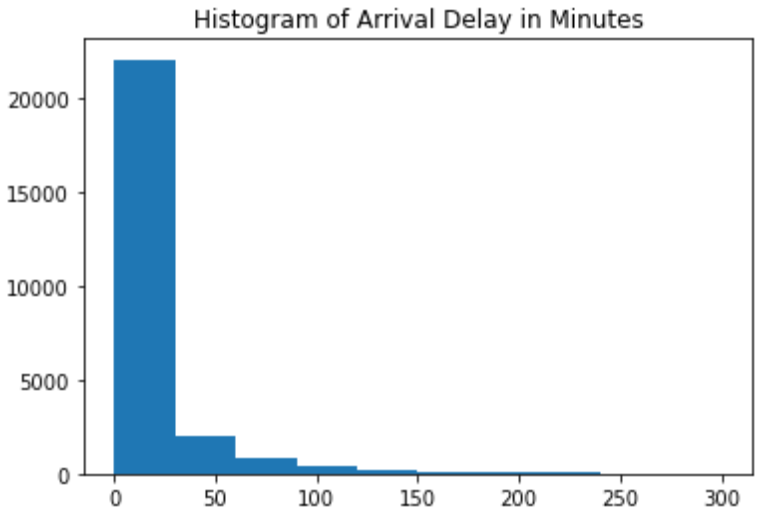
- Target column also have two vales.
- Other columns have unique values of
    - Age – 75
    - Class - 3
    - Flight Distance – 3281
    - Departure Delay in Minutes – 313
    - Arrival Delay in Minutes – 320

6.3 Handling Null Values

Data Cleaning is the process of finding and correcting the inaccurate/incorrect data that are present in the dataset. One such process needed is to do something about the values that are missing in the dataset. In real life, many datasets will have many missing values, so dealing with them is an important step. Why do you need to fill in the missing data? Because most of the machine learning models that you want to use will provide an error if you pass NaN values into it. The easiest way is to just fill them up with 0, but this can reduce your model accuracy significantly. For filling missing values, there are many methods available. For choosing the best method, you need to understand the type of missing value and its significance, before you start filling/deleting the data.Missing values are usually represented in the form of Nan or null or None in the dataset.df.info() the function can be used to give information about the dataset. This will provide you with the column names along with the number of non – null values in each column.

Here,

- Null values are only present in the column “Arrival Delay in Minutes”.



We have plotted the graph for Arrival Delay in Minutes. And as the graph is right skewed, we are using median to fill the missing data.

6.4 Encoding

- 'Gender', 'Customer Type', 'Type of Travel', 'Class', 'satisfaction' are the categorical columns.
- Performed label encoding to the categorical variables.

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

For efficient storage of the strings, **the sequence of code points is converted into a set of bytes**. The process is known as encoding.The main aim of encoding is **to transform data into a form that is readable by most of the systems or that can be used by any external process**. It can't be used for securing data, various publicly available algorithms are used for encoding. Encoding can be used for reducing the size of audio and video files.

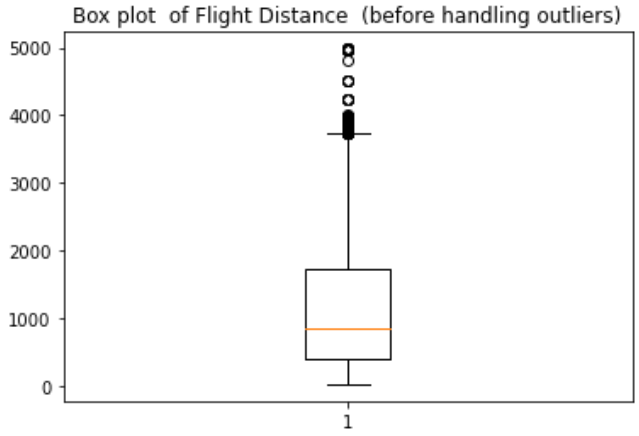
6.5 Handling Outliers

A value that "lies outside" (is much smaller or larger than) most of the other values in a set of data is called an outlier. Identification of potential outliers is important for the following reasons. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly. Outliers may be due to random variation or may indicate something scientifically

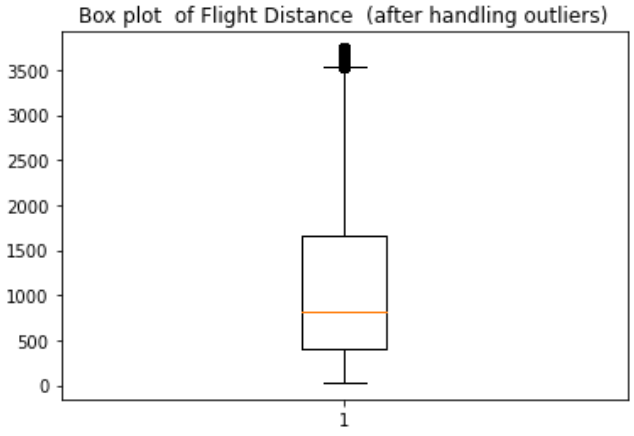
interesting. Outlier analysis tries to find unusual patterns in any dataset. We can detect the outliers by plotting box plots and handle them through the method of inter quartile range.

- Here, the dataset has outliers in four columns namely Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes are non-categorical data.

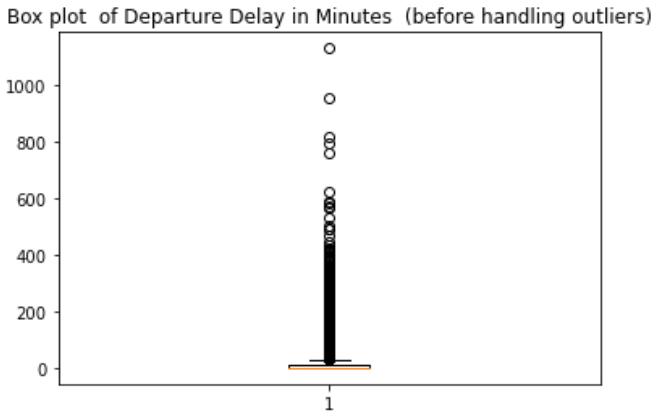
1. Handling outliers in 'Flight Distance'



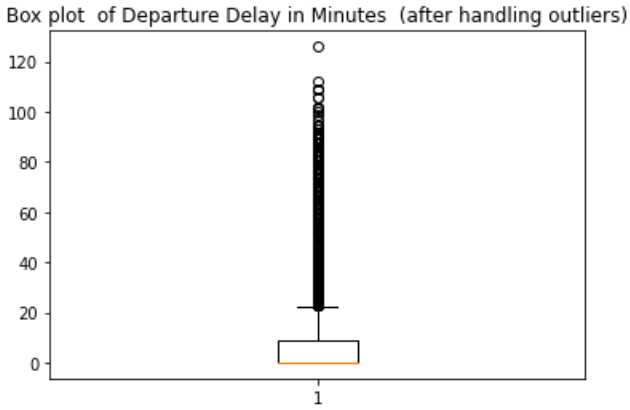
- Dropped observations with Flight Distance greater than upper limit.



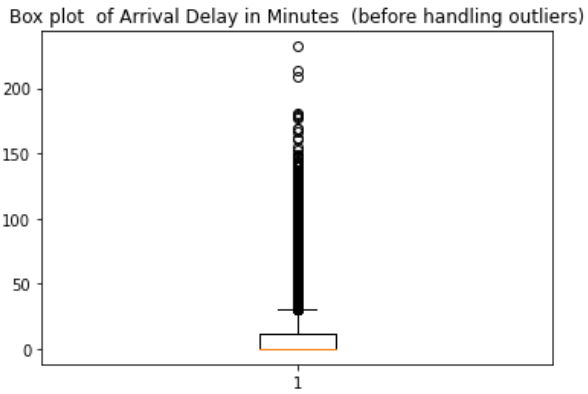
2. Handling outliers in 'Departure Delay in Minutes'



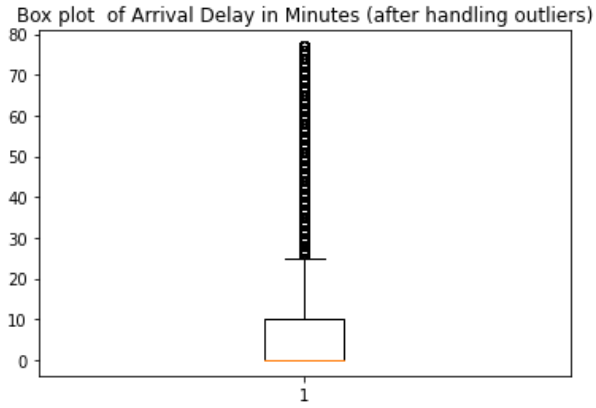
- Dropped observations with Departure Delay by finding the z-score value and greater than eliminating outliers greater than 3.



3. Handling outliers in 'Arrival Delay in Minutes'



- Dropped observations with Arrival Delay by finding the z-score value and greater than eliminating outliers greater than 3.



6.6 Standardisation

Standard Scaler: It transforms the data in such a manner that it has mean as 0 and standard deviation as 1. In short, it standardizes the data. Standardization is useful for data which has negative values. It arranges the data in a standard normal distribution.

- 'Age', 'Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes' columns are applied with standard scaling.

7.PREDICTIVE MODELLING

Predictive modeling is **a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data**. It works by analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.

Here, Predictive modelling can be used to predict customer satisfaction rate.

7.1 Logistic Regression

Logistic regression is **a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set**. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Logistic regression is a supervised learning algorithm used to predict a dependent categorical target variable. In essence, if you have a large set of data that you want to categorize, logistic regression may be able to help.

There are three main types of logistic regression: binary, multinomial and ordinal. They differ in execution and theory. Binary regression deals with two possible values, essentially: yes or no. Multinomial logistic regression deals with three or more values. And ordinal logistic regression deals with three or more classes in a predetermined order.

Performance Indices

The statistical criteria such as **Accuracy**, **Precision**, **Recall** and **f1_score** is used to evaluate each developed model's performance measure.

ACCURACY

Accuracy is one metric for evaluating classification models. Informally, accuracy is **the fraction of predictions our model got right**. Formally, accuracy has the following definition: Accuracy = Number of correct predictions Total number of predictions.

PRECISION

Precision is one indicator of a machine learning model's performance – **the quality of a positive prediction made by the model**. Precision refers to the number of true positives divided by the total number of positive predictions (i.e., the number of true positives plus the number of false positives).

RECALL

By definition recall means **the percentage of a certain class correctly identified** (from all of the given examples of that class). It is **the ability of a model to find all the relevant cases within a data set**. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives.

F1_SCORE

F1 score is defined as **the harmonic mean between precision and recall**. It is used as a statistical measure to rate performance. In other words, an F1-score (from 0 to 9, 0 being lowest and 9 being the highest) is a mean of an individual's performance, based on two factors i.e. precision and recall.

Performace Indices of Logistic Regression Model

Performace Indices	**Values**
Accuracy	0.87321
Precision	0.87153
Recall	0.82854
F1_Score	0.84949

7.2 KNN

The abbreviation KNN stands for "**K-Nearest Neighbour**". It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number

of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

The k-nearest neighbors (KNN) algorithm is **a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems**. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

The KNN algorithm can compete with the most accurate models because **it makes highly accurate predictions**. Therefore, you can use the KNN algorithm for applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.

Performance Indices of KNN Model

Performace Indices	**Values**
Accuracy	0.90444
Precision	0.91060
Recall	0.86350
F1_Score	0.88643

7.3 DECISION TREE

Decision tree build regression in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches , each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

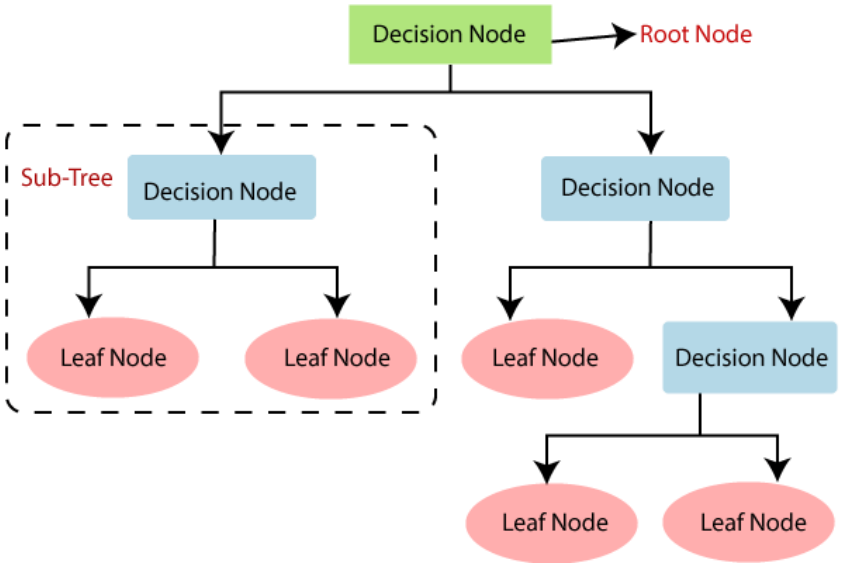


Fig: Decision Tree Model

Performace Indices of Decision Tree Model

Performace Indices	**Values**
Accuracy	0.93547
Precision	0.92774
Recall	0.92241
F1_Score	0.92507

7.4 SVM

7.4.1 Linear Classifier

SVM or Support Vector Machine is **a linear model for classification and regression problems**. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

Linear SVM is **used for linearly separable data**, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Performace Indices of Linear Classifier Model

Performace Indices	Values
Accuracy	0.87445
Precision	0.87684
Recall	0.82519
F1_Score	0.85023

7.4.2 Polynomial SVM

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

Performace Indices of Polynomial SVM Model

Performace Indices	Values
Accuracy	0.92616
Precision	0.94045
Recall	0.88505
F1_Score	0.91191

7.4.3 RBF SVM

In machine learning, the **radial basis function kernel**, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

RBF Kernel is popular **because of its similarity to K-Nearest Neighborhood Algorithm**. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset.

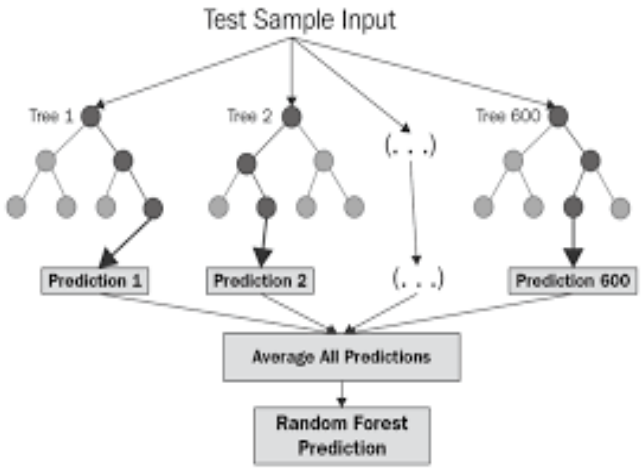
Performace Indices of RBF SVM Model

Performace Indices	Values
Accuracy	0.93795

Precision	0.94389
Recall	0.91044
F1_Score	0.92686

7.5 RANDOM FOREST

Random Forest Classifier is a supervised learning algorithm that uses ensemble learning method for classification. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



The diagram above shows the structure of a Random Forest. You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the predictions as the prediction of all the trees.

Performace Indices of Random Forest Model

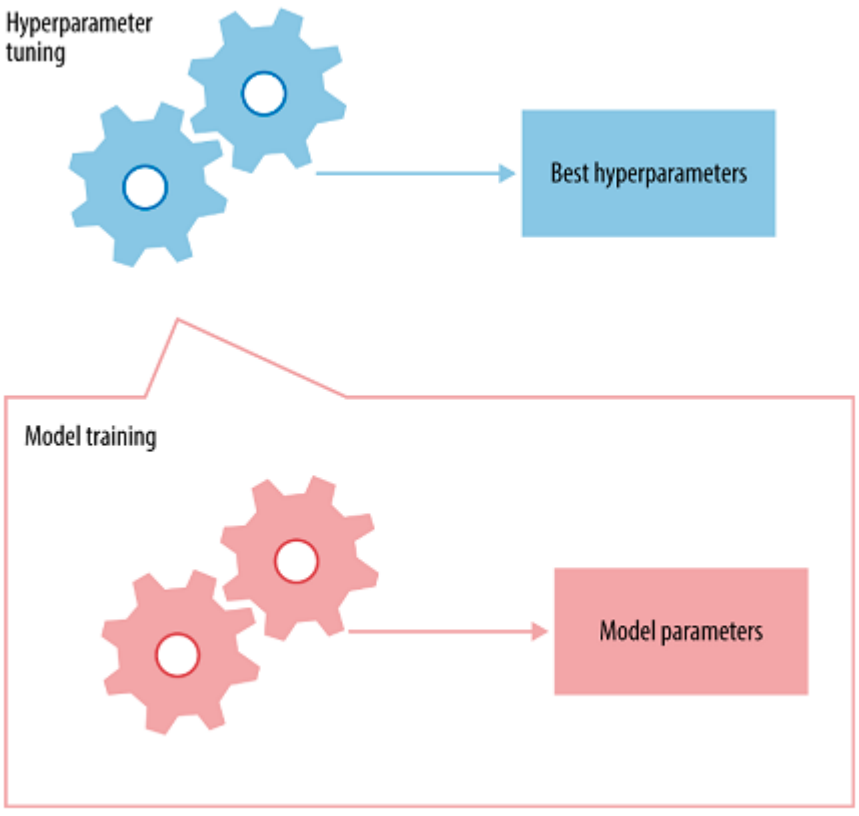
Performace Indices	**Values**
Accuracy	0.95553
Precision	0.96430
Recall	0.93151
F1_Score	0.94762

For our Satisfaction model, Random Forest performs better than Logistic Regression and Decision Tree Regression.

8. Hyperparameter Tuning

Improving the Random Forest Model

The best way to think about hyperparameters is like the settings of an algorithm that can be adjusted to optimize performance. While model parameters are learned during training — such as the slope and intercept in a linear regression — hyperparameters must be set by the data scientist before training. In the case of a random forest, hyperparameters include the number of decision trees in the forest and the number of features considered by each tree when splitting a node. (The parameters of a random forest are the variables and thresholds used to split each node learned during training). Scikit-Learn implements a set of sensible default hyperparameters for all models, but these are not guaranteed to be optimal for a problem. The best hyperparameters are usually impossible to determine ahead of time, and tuning a model is where machine learning turns from a science into trial-and-error based engineering.



Hyperparameters and Parameters

Hyperparameter tuning relies more on experimental results than theory, and thus the best method to determine the optimal settings is to try many different combinations evaluate the performance of each model. However, evaluating each model only on the training set can lead to one of the most fundamental problems in machine learning: overfitting.

Random Search Cross Validation in Scikit-Learn

Using Scikit-Learn's RandomizedSearchCV method, we can define a grid of hyperparameter ranges, and randomly sample from the grid, performing K-Fold CV with each combination of values. The following set of hyperparameters are considered for this:

- n_estimators = number of trees in the foreset
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node
- bootstrap = method for sampling data points (with or without replacement)

Grid Search with Cross Validation

Random search allowed us to narrow down the range for each hyperparameter. Now that we know where to concentrate our search, we can explicitly specify every combination of settings to try. We do this with GridSearchCV, a method that, instead of sampling randomly from a distribution, evaluates all combinations we define. To use Grid Search, we make another grid based on the best values provided by random search:

9. RESULT

Performace Indices	**Values**
Accuracy	0.95553
Precision	0.96904
Recall	0.92959
F1_Score	0.94891

Hence,we can find the performance of Random Forest Algorithm is improved after Hyper parameter tuning.(Random Search CV and Grid Search CV.)

11. CONCLUSION

- **The data is quite balanced, and it does not require any special treatment or re-sampling.**
- **The gender ratio of men and women approximately the same.**
- **The vast majority of the airline's customers are loyal customers.**
- **Most of the customers are for business purpose rather than personal reasons.**
- **About half of the passengers were in business class.**
- **Wifi service is poorly rated by all 3 classes. So, the Airline company should improve this area of service.**
- **Rating of 'Gate location' is only average among all the 'Class' of customers. The company should really focus on this area of service.**
- **For 'Food and Drink', all 'Class' of customers has given higher rating. Hence, company is doing well on this service.**
- **Most of the passengers who flew in Economy Plus or Economy Class were dissatisfied with the flight, and those who were lucky enough to fly in Business Class were satisfied.**
- **Among the various ML algorithms tried, RANDOM FOREST Algorithm has given the maximum accuracy in the predictive modelling. Accuracy is 95.5 % and f1_score is 0.947.**
- **After implementing RandomizedSearchCV, the f1_score is improved to 0.948.**