PREDICTION OF AIRLINE PASSENGER SATISFACTION

Post Graduate Program in Data Science Engineering

Location:
Chennai

Batch:
PGPDSE-FT-Feb22

Submitted by

Annish M

Akash P

Bhuvanesh V

AshwinVikash R

Mentored by

Vibha Santhanam

Table of Contents

1. INTRODUCTION

- In the airline travel industry, high grade customer satisfaction is the key factor to run the business as it is very competitive and customer satisfaction varies with small changes in services therefore companies need to the understand the customers expectations and deliver the best service possible
- As study by 'Think with Google' found that great customer service is the most compelling factor of consideration for high-value travellers. 60% of them say that when choosing a brand to travel with, their customer service matters most to them.
- 89% of travellers would post a photo or video about a travel destination they loved.
- 83% of them would post about a positive hotel or resort experience.
- 67% of them would post about a positive transportation experience.

1.1. Dataset Information

- This dataset contains an airline passenger satisfaction survey with 25 variables describing the 129880 observations of passenger satisfaction data.

1.2. Problem Statement

- Recently, airline companies have realized the importance of satisfied customers to find a place for themselves in this competitive world and initiated many projects to measure service quality and satisfy the customers by improving service quality.
- Airline collects an incredible amount of passenger feedback, it manually calculating the customer satisfaction is time consuming and by the time satisfaction are predicted through by machine learning model, we can easily sort out the issues that led to negative feedbacks.

1.3. Variable Categorization with Description

- The dataset consists of 24 variables. Out of these variables 23 are independent variables and 1 is a target variable. The variables are a mixture of both numerical and categorical type.

1.3.1. Numeric Variables :

Sr No.	Variable	Datatype	Description
1	ID	int64	Travel ID of the passengers
2	Age	int64	The actual age of the passengers
3	Flight distance	int64	The flight distance of this journey
4	Inflight WIFI Service	int64	Satisfaction level of the inflight wifi service Rating: 0 (least) - 5 (highest)
5	Ease of Online Booking	int64	Satisfaction level of online booking Rating: 0 (least) - 5 (highest)
6	Online boarding	int64	Satisfaction level of online boarding Rating: 0 (least) - 5 (highest)
7	Inflight Entertainment	int64	Satisfaction level of inflight entertainment Rating: 0 (least) - 5 (highest)
8	Food and drink	int64	Satisfaction level of Food and drink Rating: 0 (least) - 5 (highest)
9	Seat comfort	int64	Satisfaction level of Seat comfort Rating: 0 (least) - 5 (highest)
10	On-board service	int64	Satisfaction level of On-board service Rating: 0 (least) - 5 (highest)
11	Leg room service	int64	Satisfaction level of Leg room service Rating: 0 (least) - 5 (highest)

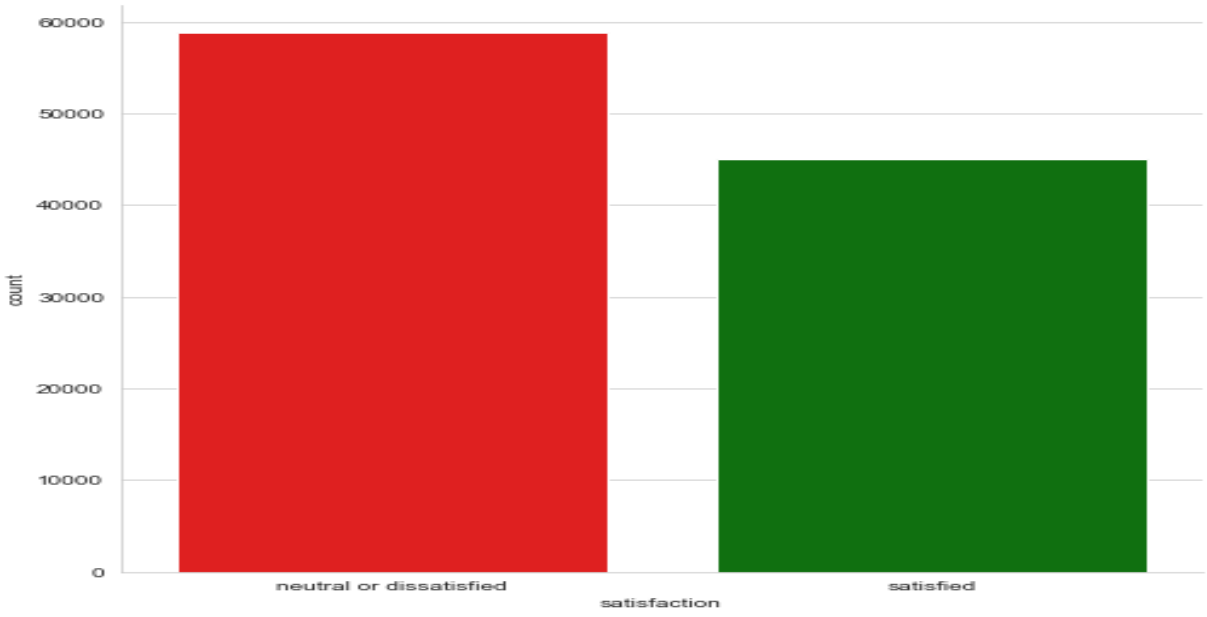
12	Departure/Arrival time convenient	int64	Satisfaction level of Departure/Arrival time convenient Rating: 0 (least) - 5 (highest)
13	Baggage handling	int64	Satisfaction level of baggage handling Rating: 0 (least) - 5 (highest)
14	Gate location	int64	Satisfaction level of Gate location Rating: 0 (least) - 5 (highest)
15	Cleanliness	int64	Satisfaction level of Cleanliness Rating: 0 (least) - 5 (highest)
16	Check-in service	int64	Satisfaction level of Check-in service Rating: 0 (least) - 5 (highest)
17	Departure Delay in Minutes	float64	Minutes delayed when departure
18	Arrival Delay in Minutes	float64	Minutes delayed when Arrival
19	Inflight service	int64	Satisfaction level of Inflight service Rating: 0 (least) - 5 (highest)

1.3.2. Categorical Variable :

Sr No.	Variable	Datatype	Description
1	Gender	object	Gender of the passengers Female, Male Type of Travel Purpose of the flight of the passengers
2	Type of Travel	object	Purpose of the flight of the passengers Personal Travel, Business Travel
3	Class	object	Travel class in the plane of the passengers Business, Eco, Eco Plus
4	Customer Type	object	The customer type Reliable customer, Unreliable Type of Travel customer

1.4. Target Variable

The target variable of the above dataset is satisfaction. We have to predict whether a passenger is satisfied or (neutral or dissatisfied).



In the above dataset, 57% of the passengers have not been satisfied and 43% of the passengers have been satisfied. We observe that there is **presence of moderate amount of class imbalance**.

2. DATA PRE-PROCESSING

- Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data pre-processing task.
- A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.
- The data consists of 129880 rows and 24 columns. Out of these we have 4 categorical columns and the rest as numerical.

2.1. Datatype Verification

We first check the data types of each of the columns of the data.

Variable	**Datatype**
ID	int64
Age	int64
Flight distance	int64
Inflight WIFI Service	int64
Ease of Online Booking	int64
Online boarding	int64
Inflight Entertainment	int64
Food and drink	int64
Seat comfort	int64
On-board service	int64
Leg room service	int64
Departure/Arrivaltime convenient	int64
Baggage handling	int64
Gate location	int64
Cleanliness	int64
Check-in service	int64
Departure Delay in Minutes	float64
Arrival Delay in Minutes	float64
Inflight service	int64
Gender	object

Type of Travel	object
Class	object
Customer Type	object

2.2. Missing Value Treatment

- The next step of data pre-processing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

Attribute	Null Value Percentage
Arrival Delay in Minutes	2.98352

- For Arrival Delay in Minutes the null means passengers do not have minutes delayed when arrival. Hence replace it with 0.

2.3. Redundant Features Removal

- Checking and removal of duplicate rows is important because presence of duplicates can lead us to make incorrect conclusions by leading us to believe that some observations are more common than they really are.
- From the 5-point summary of the data, we observe that the 'id' column is not useful to build the model. Hence, we remove the columns containing 'id' values.

- After performing all these steps, we finally have data which has 129880 rows and 23 columns.

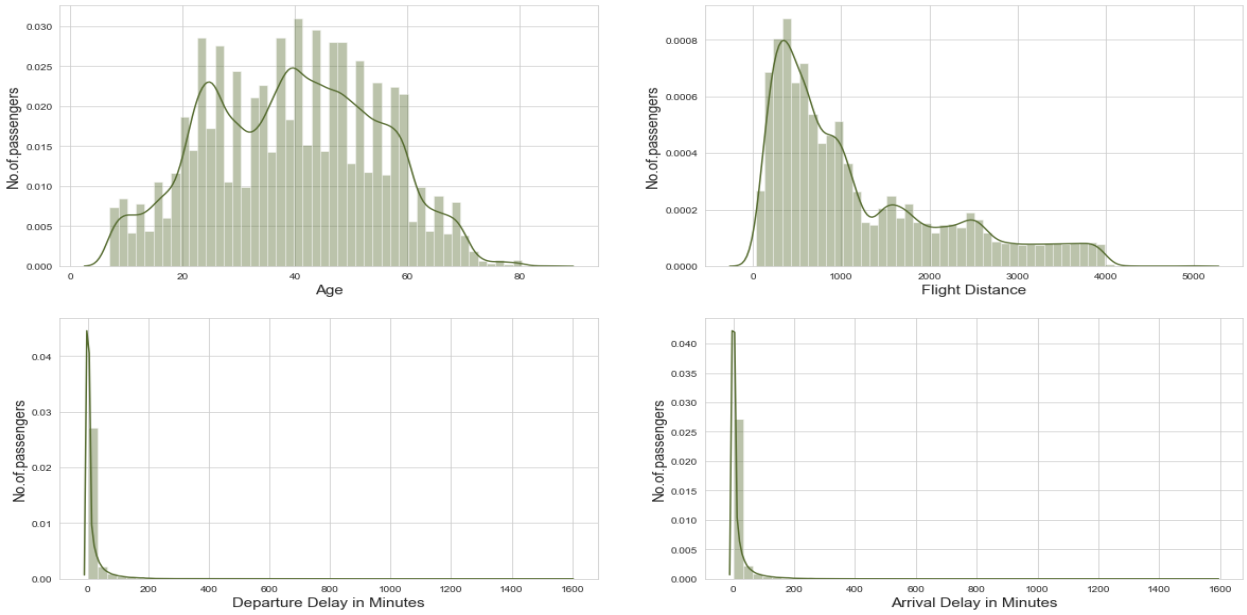
2.4. Check for Outliers

- Data has outliers present in the numerical columns. For making the base model, we do not perform any outlier treatment and retain all the rows present in the data.

3. EXPLORATORY DATA ANALYSIS

- **For Numerical Variables**: - We plot the distribution curve and histogram to study the variation of the numerical data.

3.1. Histogram :

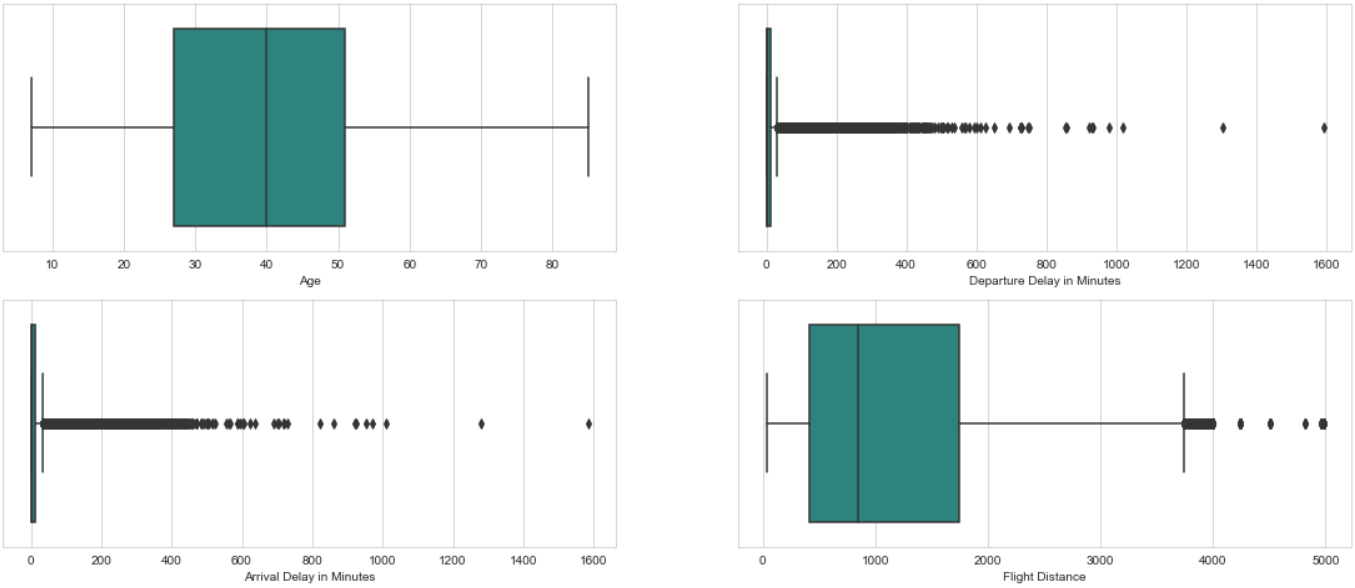


- **Age**
    - Age is not a perfect normal distribution.
    - It is partially normal distribution
    - the major age category that travel ranges between 20 to 60

- **Flight Distance**
    - Flight distance data is right skewed
    - majority of flights travel a distance which ranges between 100 miles to 1000 miles, starting from 31 miles (i.e.) from Alaska to Petersburg.

- **Departure delay in minutes & Arrival delay in minutes**
    - Departure delay in minutes & Arrival delay in minutes data is heavily right skewed,
    - most of the delay in minutes ranges between 10 to 100 which is fairly understandable as the airline companies look to start the flight without any delay.

3.2. Box Plot



➢ Departure Delay in Minutes & Arrival Delay in Minutes

- Huge amount of outliers are present in Departure Delay in Minutes and Arrival Delay in Minutes is huge which is understandable as most of the flights take off within the prescribed time and since it being a continuous variable we can see huge amount of outliers .

➢ Flight Distance

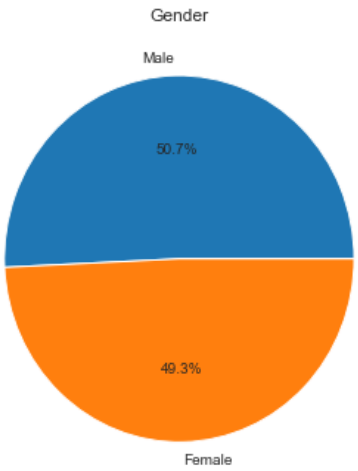
- Majority of flights travel with a distance range of 500 to 1800 miles and few outliers can be found which indicates these are long distance flights.

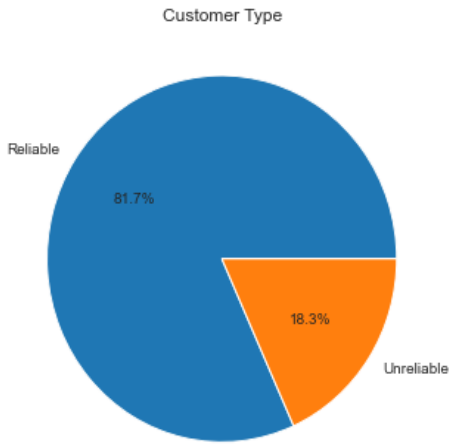
➢ Age

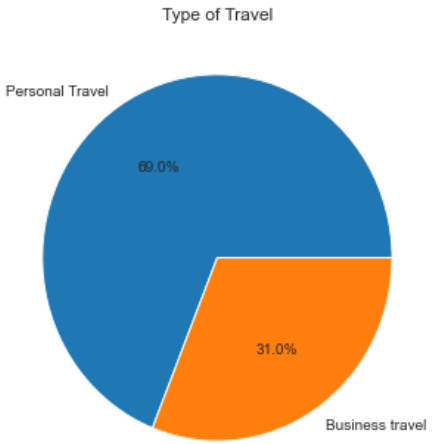
- Majority of passengers vary within the age group of 28 – 52.

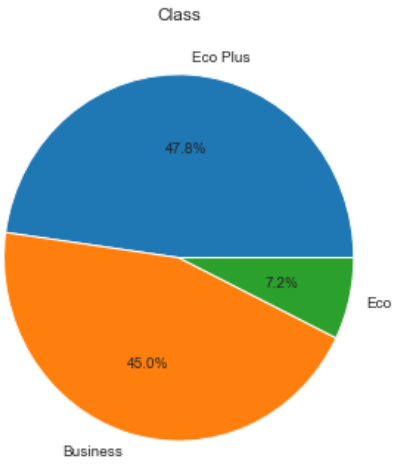
- **For Categorical Variables** – We plot a combination of bar graph and pie chart to understand the distribution of categorical data in the dataset.

3.3. Pie Chart









➢ Gender

- From the pie chart gender we can see that although gender of the male and female is almost equal
- male is 50.7%
- female is 49.3%

➢ Customer Type

- From the plot customer type it can be seen that the data,the majority of passengers from reliable customer (81.7%) and remaining passengers from unreliable (18.3%).

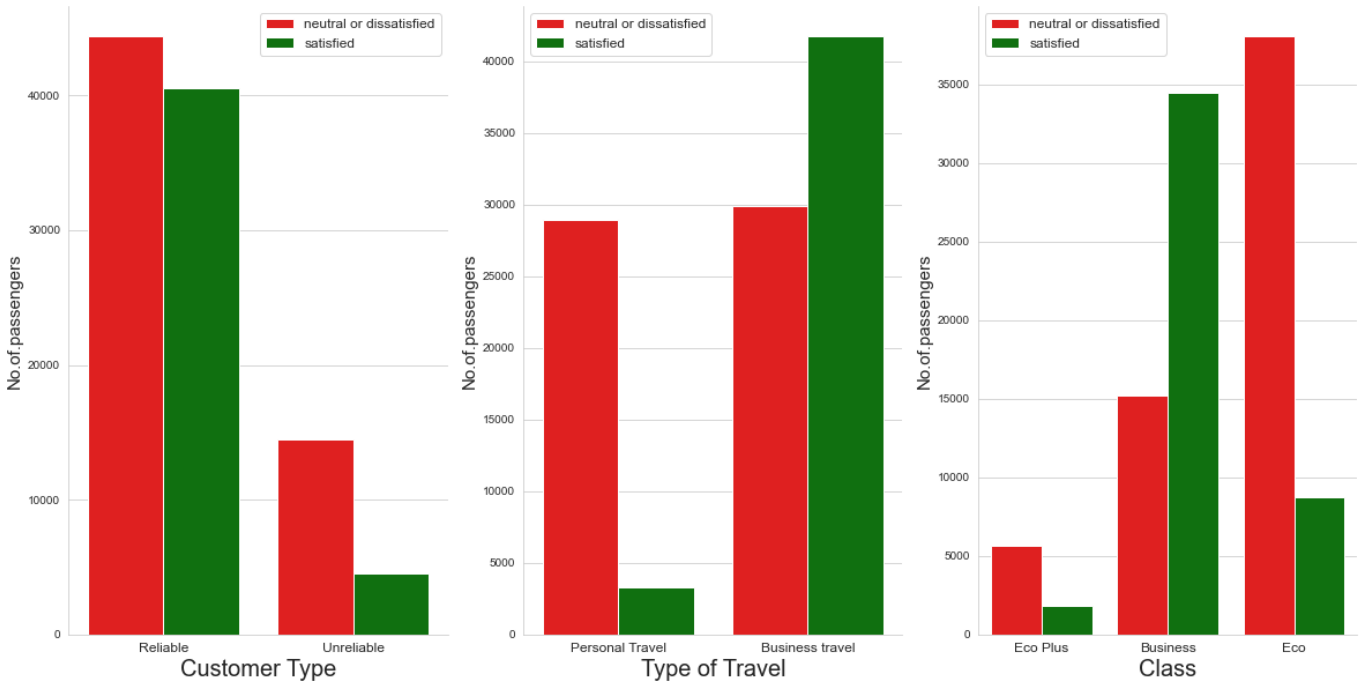
➢ Type of Travel

- From the plot type of travel it can be seen that the data,the majority of passengers travel for personal work (69%) and remaining passengers for offical work(31%).

➢ Class

- from the plot class It can be seen that the data, 47.8% of passengers from Economy plus class,45% of passengers from Business class and remaining 7.2% of passengers from Economy class.

3.4. Bar Chart



➢ Customer Type

- We can be seen that the number of returning/reliable customers are more when compared to the number of non reliable customers.
- The number of dissatisfied customers seem to high within the unreliable customers and there exists a kind of balanced within the reliable customers .

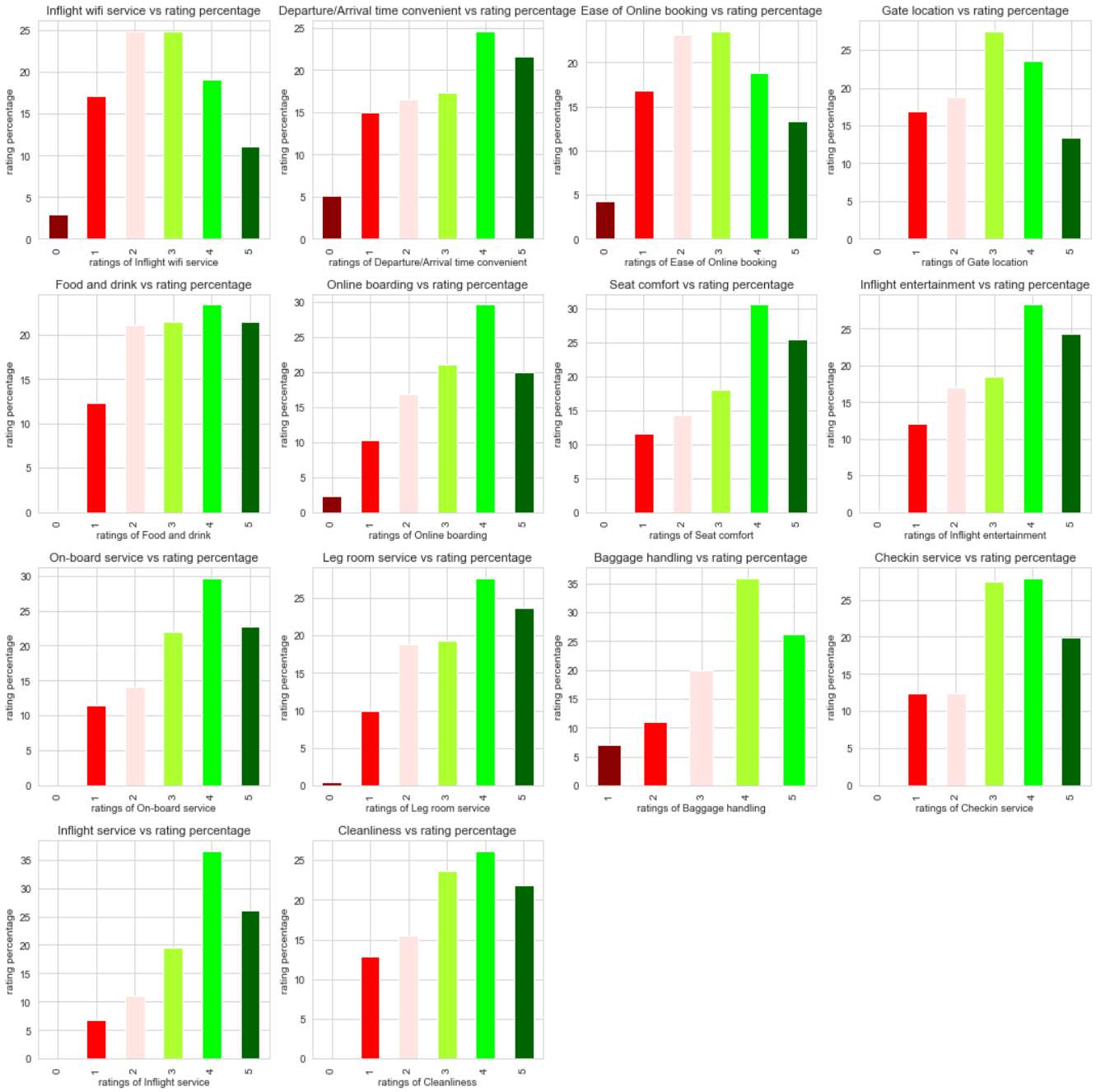
➢ Type of Travel

- We can see that business travel people are highly satisfied with the services, when compared to the personal travel.
- There is a clear disparity when satisfaction is compared between the type of travelers, which clearly depics some underlying issues in services based on the type of travelers.

➢ Class

- Here we can see that when compared within classes , there is a huge difference in satisfaction within the three classes.
- Business class people seem to be the most satisfied when compared to eco and eco plus class passengers, which is undertandable compared to the amount of services offered to the business class people in the airlines, although such high variance within the classes can be balanced rather than being over biased to one particular section of the passengers to a considerable amount.

3.5. Count Plot Chart

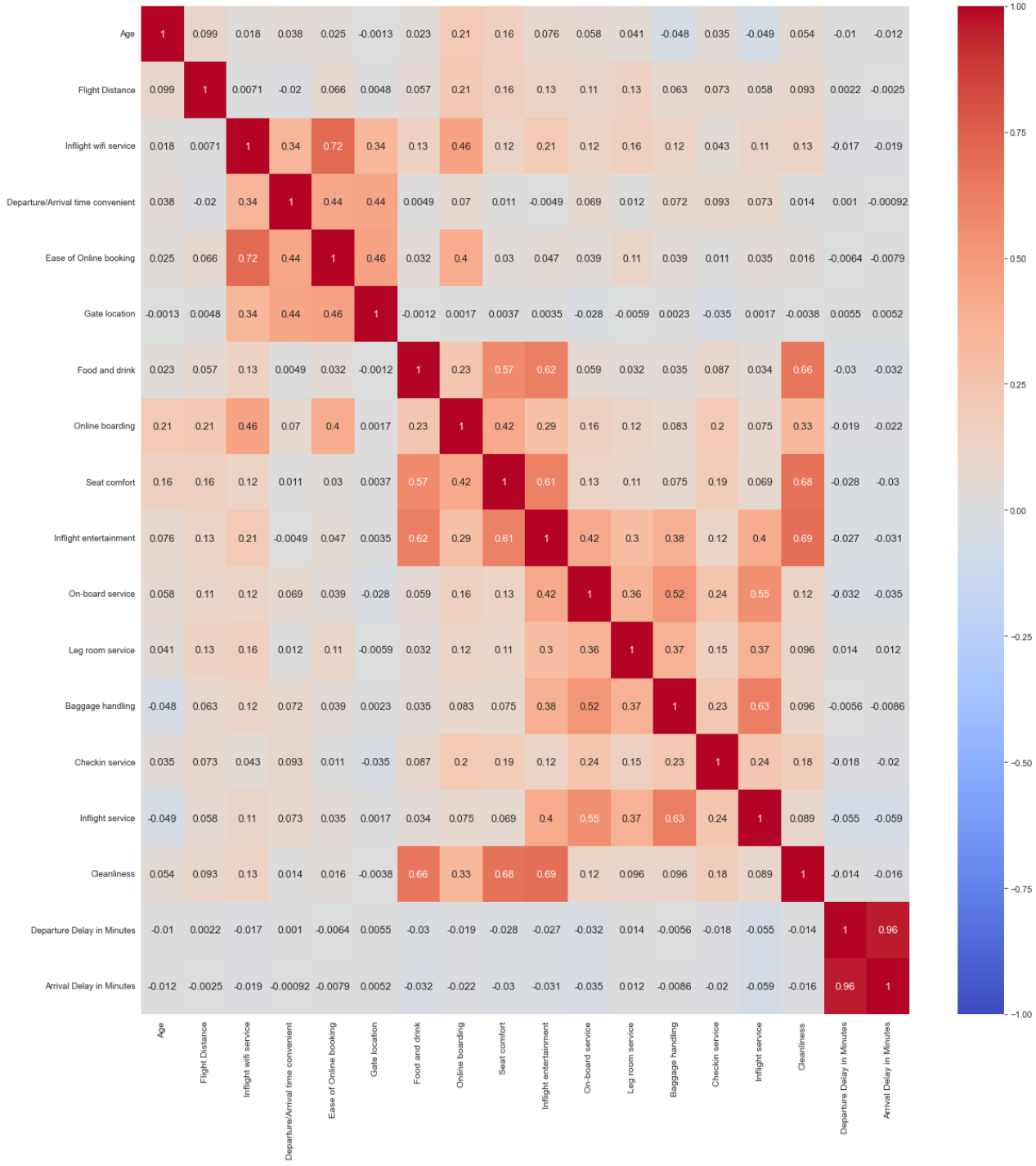


➢ Ratings

- **From the plots above we can say that the airline performs better in:**
- Inflight services
- Baggage handling
- Leg room services
- On board services
- Seat comfort
- Inflight entertainment

- **Services could be improved in these areas:**
- Food and drinks
- Gate location perhaps changes can't be made over here
- Ease of online booking
- In flight Wi-Fi services
- Thus, a simple count plot could give us so much insight about our dataset

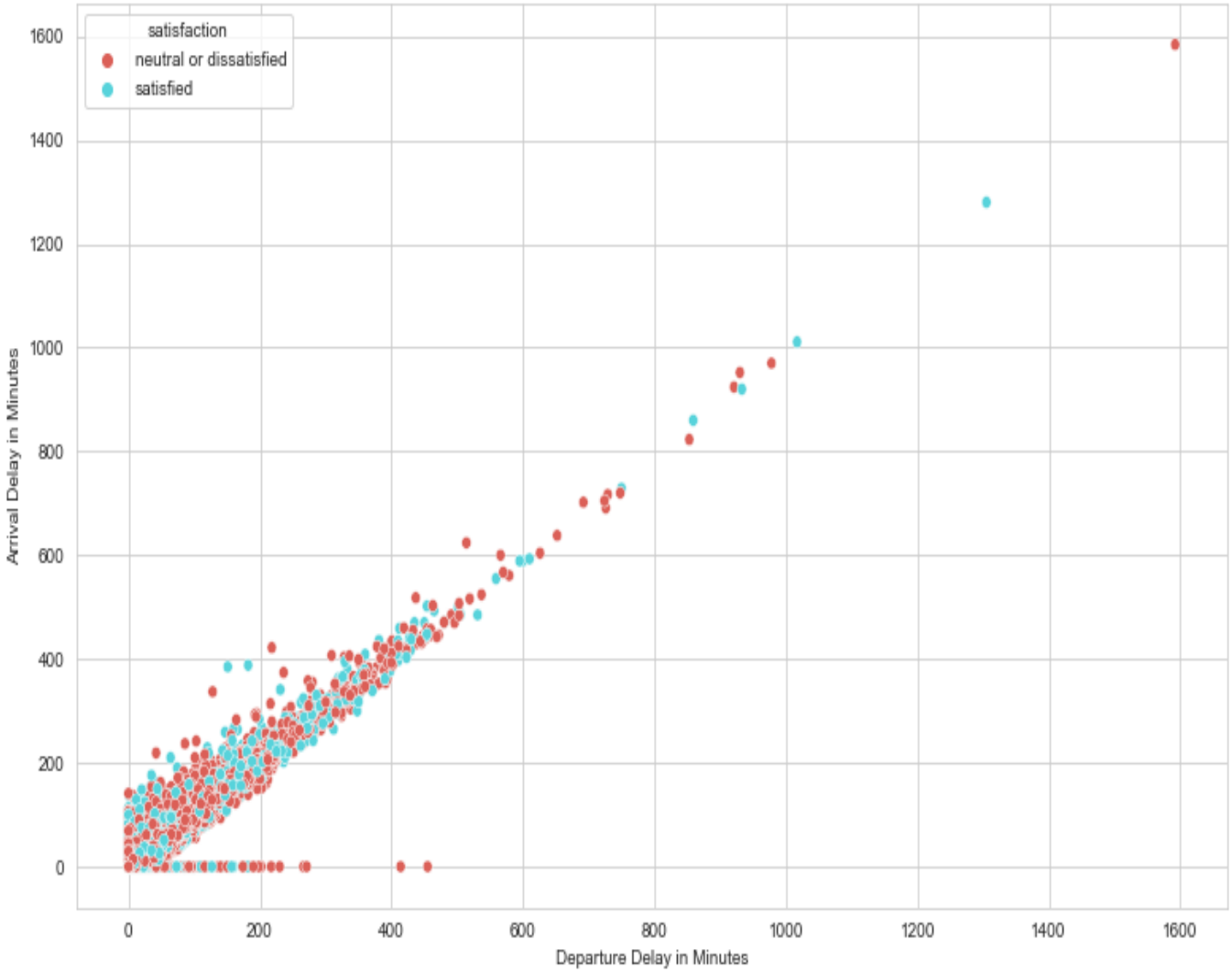
3.6. Heat Map (Correlation plot)



- High amount of correlation is present between 'Departure Delay in Minutes' and 'Arrival Delay in Minutes'.

3.7. Scatterplot

- It can be seen that as there is a delay in departure when there is a delay in arrival. Which shows a strong positive correlation between the departure and arrival, but based on this alone the satisfaction couldn't be predicted as they seem to be clustered all over.



3.8. FEATURE ENIGNEERING:

- Age is continuous variable, from the above code, we split the age groups - (7 to 16) age passengers considered as kids, (17 to 28) age passengers considered as youth, (29 to 55) age passengers considered as adults and remaining (56 to 85) age passengers considered as elders.

3.9. ENCODING TECHNIQUE:

3.9.1. N-1) ENCODING :

- Gender, customer type, age, and type of travel are categorical variables(nominal), from the above code we encoding the categorical variable into numerical variable (0 and 1) because machine doesn't understand the categorical variables.

3.9.2. ORDINAL ENCODING :

- Class is ordinal categorical variable, we convert into numerical variable using map function and Eco (Economy) is simple (third class) and seating capacity is high compared to both Eco plus and Business class. So, we considered as 0. Eco plus (Economy plus) is slightly premium than Eco (Economy).so, Eco plus considered as 1 and Business class is fully premium we considered as 2.

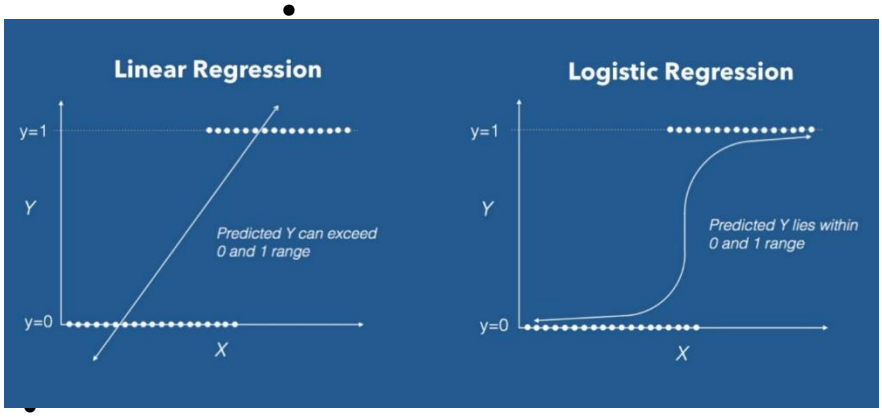
3.9.3. MAP FUNCTION :

- Satisfaction is our target variable ,we convert the class neutral or dissatisfied and satisfied into 0 and 1.

4. MODEL BUILDING:

4.1.1. LOGISTIC REGRESSION

- Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression transforms its output using the logistic sigmoid function to return a probability value.
- **What are the types of logistic regression**
- Binary (eg. Tumor Malignant or Benign)
- Multi-linear functions fail Class (eg. Cats, dogs or Sheep's)
- Logistic Regression is a Machine Learning algorithm which is used for the classification problems; it is a predictive analysis algorithm and based on the concept of probability.



Linear Regression VS Logistic Regression Graph| Image: Data Camp

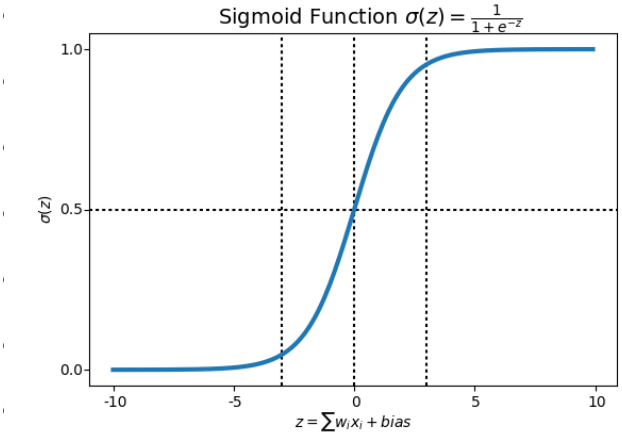
- We can call a Logistic Regression a Linear Regression model but the

Logistic Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function.

- The hypothesis of logistic regression tends it to limit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logistic regression.

$$0 \leq h_\theta(x) \leq 1$$

- Logistic regression hypothesis expectation
- What is the Sigmoid Function?
- In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



Sigmoid Function Graph

$$f(x) = \frac{1}{1+e^{-(x)}}$$

4.2. Base Model:

Logit Regression Results

Dep. Variable:	satisfaction	No. Observations:	103904
Model:	Logit	Df Residuals:	103879
Method:	MLE	Df Model:	24
Date:	Thu, 16 Jun 2022	Pseudo R-squ.:	0.5120
Time:	22:03:05	Log-Likelihood:	-34695.
converged:	True	LL-Null:	-71094.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-7.0795	0.079	-89.747	0.000	-7.234	-6.925
Class	0.3623	0.013	28.237	0.000	0.337	0.387
Flight Distance	2.551e-06	1.12e-05	0.228	0.819	-1.93e-05	2.44e-05
Inflight wifi service	0.3870	0.011	33.870	0.000	0.365	0.409
Departure/Arrival time convenient	-0.1235	0.008	-15.022	0.000	-0.140	-0.107
Ease of Online booking	-0.1416	0.011	-12.493	0.000	-0.164	-0.119
Gate location	0.0294	0.009	3.204	0.001	0.011	0.047
Food and drink	-0.0269	0.011	-2.529	0.011	-0.048	-0.006
Online boarding	0.6161	0.010	60.465	0.000	0.596	0.636
Seat comfort	0.0723	0.011	6.486	0.000	0.050	0.094
Inflight entertainment	0.0603	0.014	4.235	0.000	0.032	0.088
On-board service	0.3066	0.010	30.104	0.000	0.287	0.327
Leg room service	0.2561	0.009	30.012	0.000	0.239	0.273
Baggage handling	0.1357	0.011	11.863	0.000	0.113	0.158
Checkin service	0.3264	0.009	38.142	0.000	0.310	0.343
Inflight service	0.1214	0.012	10.077	0.000	0.098	0.145
Cleanliness	0.2204	0.012	18.290	0.000	0.197	0.244
Departure Delay in Minutes	0.0042	0.001	4.568	0.000	0.002	0.006

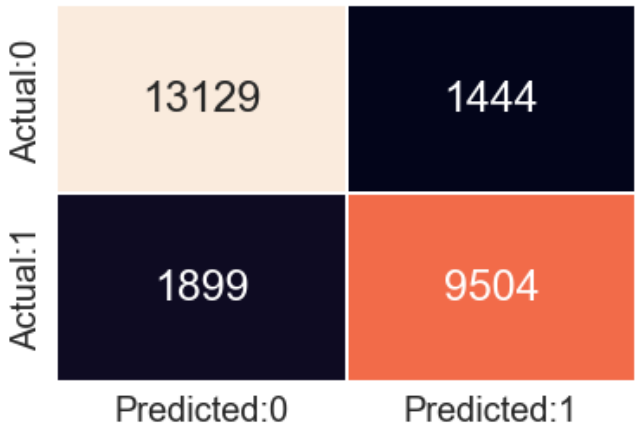
```
Arrival Delay in Minutes             -0.0089      0.001     -9.814      0.000      -0
.011      -0.007
Gender_Male                           0.0462      0.019      2.376      0.018       0
.008       0.084
Customer Type_Unreliable             -2.0631      0.030    -68.048      0.000      -2
.123      -2.004
Age_youth                             0.3887      0.048      8.027      0.000       0
.294       0.484
Age_adults                            0.0524      0.046      1.139      0.255      -0
.038       0.143
Age_elders                           -0.1742      0.050     -3.454      0.001      -0
.273      -0.075
Type of Travel_Personal Travel       -2.7067      0.032    -85.292      0.000      -2
.769      -2.645
==============================================================================
```

- We train the model from our training dataset, fit is normally train the data for given dataset and summary gives the explanation about our model such as PsuedoR2, Log likelihood and p-value etc.
- In our base model gives Pseudo R-squ.: 0.5120, Log-Likelihood: -34695, LL-Null: -71094.

4.3. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 87%.

4.4. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

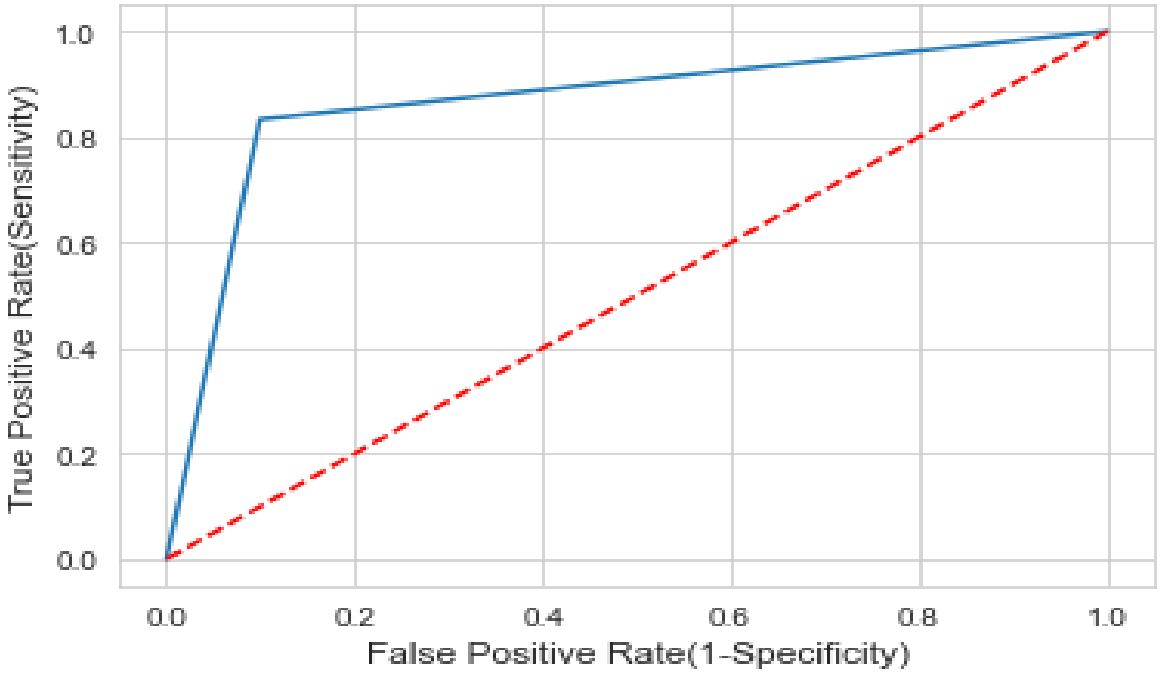
- TN -13129
- TP-9504
- FP-1444
- FN-1899

4.5. Classification Report :

	precision	recall	f1-score	support
0	0.87	0.90	0.89	14573
1	0.87	0.83	0.85	11403
accuracy			0.87	25976
macro avg	0.87	0.87	0.87	25976
weighted avg	0.87	0.87	0.87	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.87 and 0.87
- recall score for class 0 and class 1 are 0.90 and 0.83
- f1-score for class 0 and class 1 are 0.89 and 0.85.

4.6. Roc – Curve :



- This graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters :
  - True Position Rate
  - False Positive Rate
- AUC provides an aggregate measure of performance across all possible classification thresholds.
- Here our AUC score is 86.7%

5. NAIVE BAYES

- Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
- The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

  contribution to the outcome.

5.1. Bayes' Theorem

- Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

A, B = events
$P(A|B)$ = probability of A given B is true
$P(B|A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B

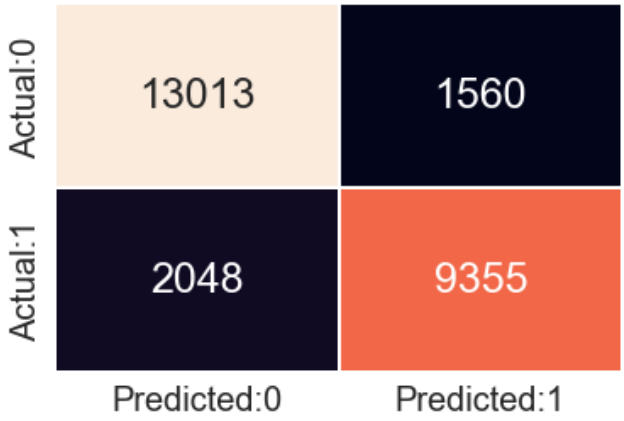
- where A and B are events and P(B) ≠ 0.
- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.
- P(A) is the priori of A (the prior probability, i.e., Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
- P(A|B) is a posteriori probability of B, i.e., probability of event after evidence is seen.

5.2. All Features:

5.2.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 86%.

5.2.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -13013
- TP-9355
- FP-1560
- FN-2048

5.2.3. Classification Report :

	precision	recall	f1-score	support
0	0.86	0.89	0.88	14573
1	0.86	0.82	0.84	11403
accuracy			0.86	25976
macro avg	0.86	0.86	0.86	25976
weighted avg	0.86	0.86	0.86	25976

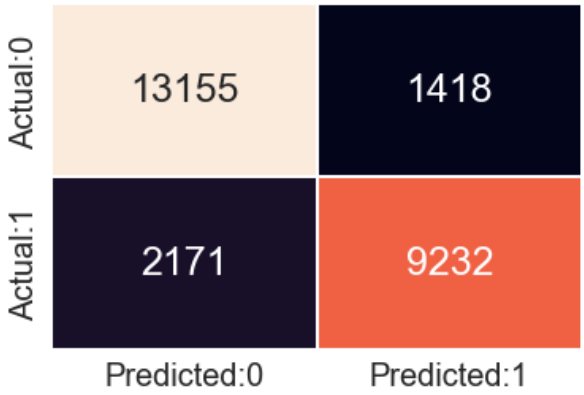
- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.86 and 0.86
- recall score for class 0 and class 1 are 0.89 and 0.82
- F1-score for class 0 and class 1 are 0.88 and 0.84.

5.3. Best Features:

5.3.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 86%.

5.3.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -13155
- TP-9232

- FP-1418
- FN-2171

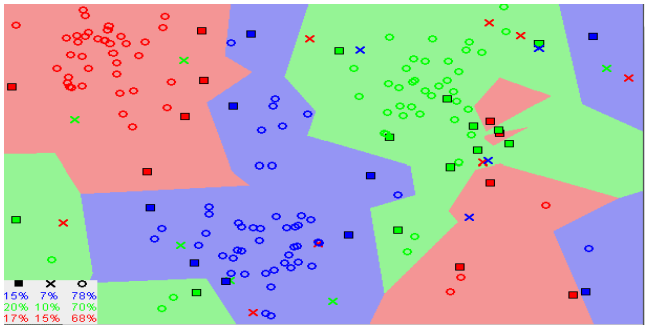
5.3.3. Classification Report :

	precision	recall	f1-score	support
0	0.86	0.90	0.88	14573
1	0.87	0.81	0.84	11403
accuracy			0.86	25976
macro avg	0.86	0.86	0.86	25976
weighted avg	0.86	0.86	0.86	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.86 and 0.87
- recall score for class 0 and class 1 are 0.90 and 0.81
- F1-score for class 0 and class 1 are 0.88 and 0.84.

6. K-NEAREST NEIGHBOURS

- The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.
- "Birds of a feather flock together.



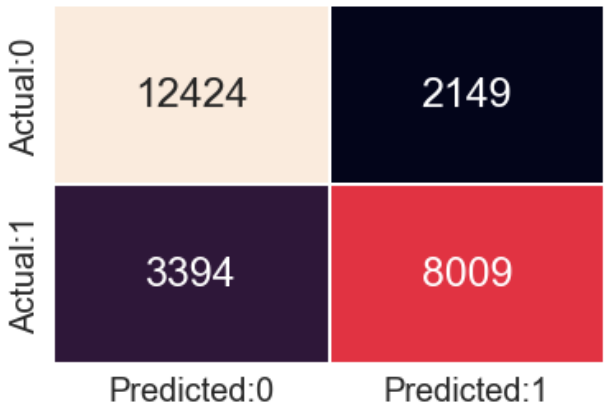
- Image showing how similar data points typically exist close to each other
- Notice in the image above that most of the time, similar data points are close to each other. The KNN algorithm hinges on this assumption being true enough for the algorithm to be useful. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph.
- KNN's main disadvantage of becoming significantly slower as the volume of data increases makes it an impractical choice in environments where predictions need to be made rapidly. Moreover, there are faster algorithms that can produce more accurate classification and regression results.

6.1. All Features:

6.1.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 79%.

6.1.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -12424
- TP-8009
- FP-2149
- FN-3394

6.1.3. Classification Report :

	precision	recall	f1-score	support
0	0.79	0.85	0.82	14573
1	0.79	0.70	0.74	11403
accuracy			0.79	25976
macro avg	0.79	0.78	0.78	25976
weighted avg	0.79	0.79	0.78	25976

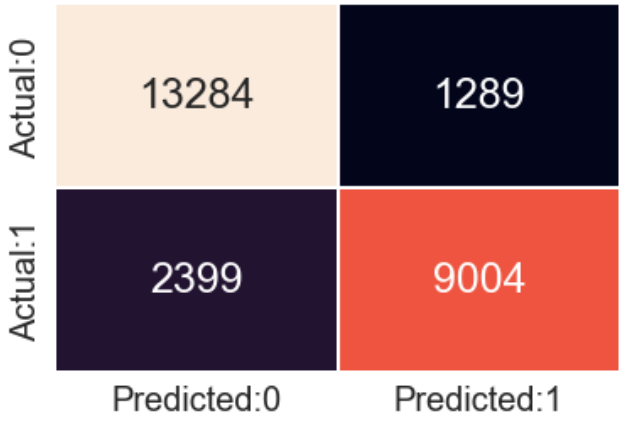
- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.79 and 0.79
- recall score for class 0 and class 1 are 0.85 and 0.70
- F1-score for class 0 and class 1 are 0.82 and 0.74.

6.2. Best Features:

6.2.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 86%.

6.2.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -13284
- TP-9004
- FP-1289
- FN-2399

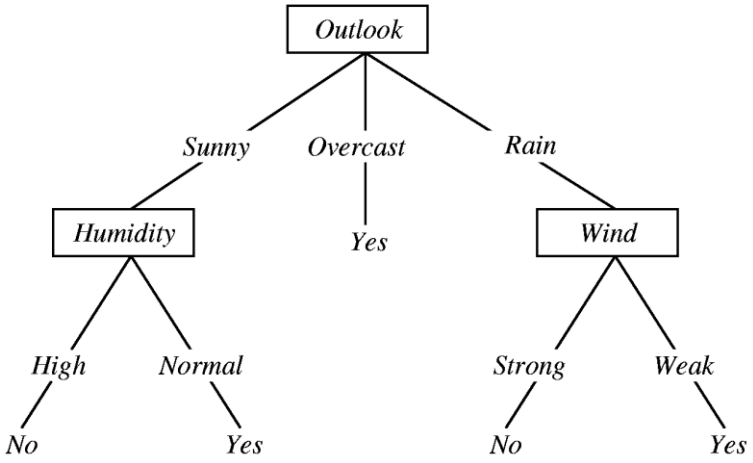
6.2.3. Classification Report :

	precision	recall	f1-score	support
0	0.85	0.91	0.88	14573
1	0.87	0.79	0.83	11403
accuracy			0.86	25976
macro avg	0.86	0.85	0.85	25976
weighted avg	0.86	0.86	0.86	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.85 and 0.87
- recall score for class 0 and class 1 are 0.91 and 0.79
- f1-score for class 0 and class 1 are 0.88 and 0.83.

7. DECISION TREE ALGORITHM

- A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules. Below diagram illustrate the basic flow of decision tree for decision making with labels (Rain(Yes), No Rain(No)).



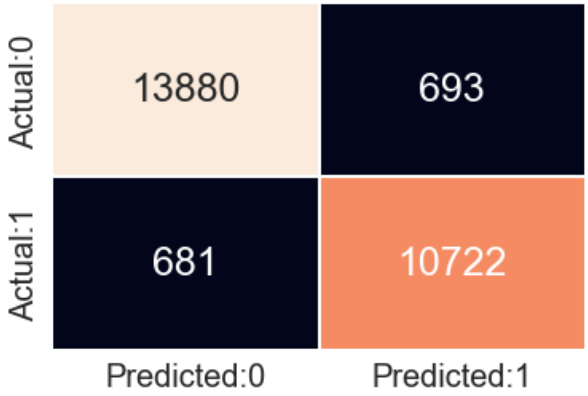
- Decision tree is one of the predictive modelling approaches used in statistics, data mining and machine learning.
- Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric **supervised learning** method used for both **classification** and **regression** tasks.
- Tree models where the target variable can take a discrete set of values are called **classification trees**. Decision trees where the target variable can take continuous values (typically real numbers) are called **regression trees**. Classification And Regression Tree (CART) is general term for this.

7.1. All Features:

7.1.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 95%.

7.1.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -13880
- TP-10722
- FP-693
- FN-681

7.1.3. Classification Report :

	precision	recall	f1-score	support
0	0.95	0.95	0.95	14573
1	0.94	0.94	0.94	11403
accuracy			0.95	25976
macro avg	0.95	0.95	0.95	25976
weighted avg	0.95	0.95	0.95	25976

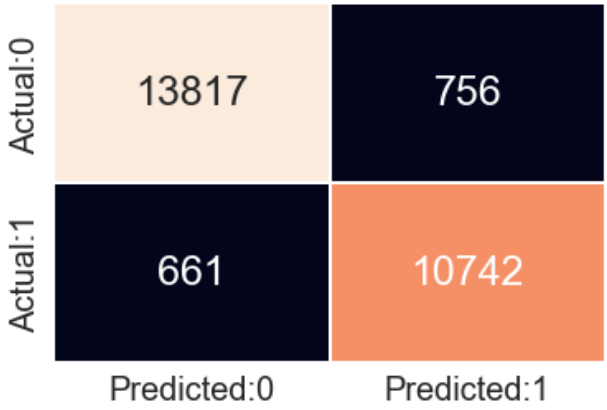
- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.95 and 0.94
- recall score for class 0 and class 1 are 0.95 and 0.94
- f1-score for class 0 and class 1 are 0.95 and 0.94.

7.2. Best Features:

7.2.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 95%.

7.2.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -13817

- TP-10742
- FP-756
- FN-661

7.2.3. Classification Report :

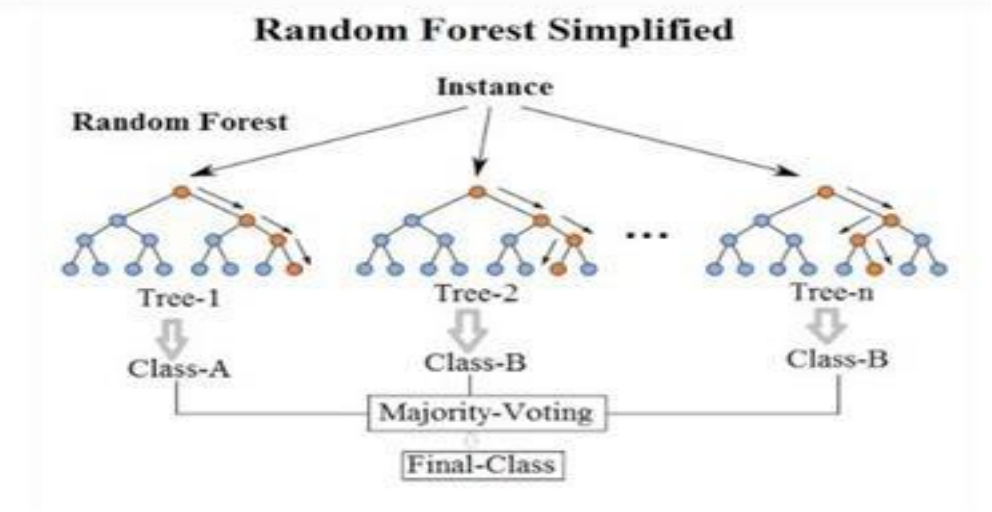
	precision	recall	f1-score	support
0	0.95	0.95	0.95	14573
1	0.93	0.94	0.94	11403
accuracy			0.95	25976
macro avg	0.94	0.95	0.94	25976
weighted avg	0.95	0.95	0.95	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.95 and 0.93
- recall score for class 0 and class 1 are 0.95 and 0.94
- F1-score for class 0 and class 1 are 0.95 and 0.94.

8. RANDOM FOREST CLASSIFIER

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
- To say it in simple words: Random forest builds multiple decision trees and merges them

- Together to get a more accurate and stable prediction.
- One big advantage of random forest is, that it can be used for both classification and regression problems.
- Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier.
- Fortunately, we don't have to combine a decision tree with a bagging classifier and can just
- Easily use the classifier-class of Random Forest. Like I already said, with Random Forest, you can also deal with Regression tasks by using the Random Forest repressor.



- Random Forest adds additional randomness to the model, while growing the trees. Instead of
- searching for the most important feature while splitting a node, it searches for the best feature
- among a random subset of features. This results in a wide diversity that generally results in a better model.

Advantages of Random Forest:

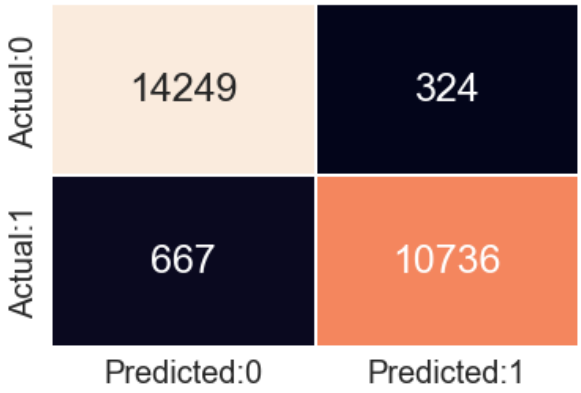
- There is no need for feature normalization
- Individual decision trees can be trained in parallel
- Reduced over fitting
- Require almost no input preparation
- Performs implicit feature selection
- It's very quick to train
- Modeling and Predicting Online Purchasing Intention of Shopper
- Disadvantages of Random Forest:
- No interpretability

8.1. All Features:

8.1.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 96%.

8.1.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.

- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -14249
- TP-10736
- FP-324
- FN-667

8.1.3. Classification Report :

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14573
1	0.97	0.94	0.96	11403
accuracy			0.96	25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

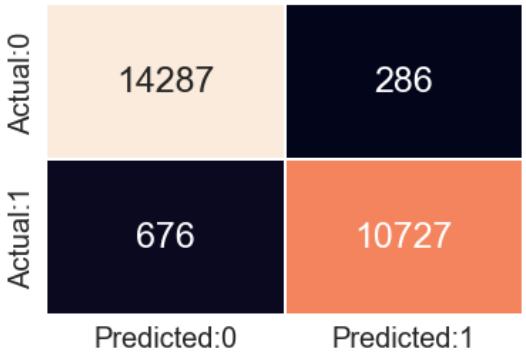
- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.96 and 0.97
- recall score for class 0 and class 1 are 0.98 and 0.94
- F1-score for class 0 and class 1 are 0.97 and 0.96.

8.2. Best Features:

8.2.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 96%.

8.2.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -14287
- TP-10727
- FP-286
- FN-676

8.2.3. Classification Report :

	precision	recall	f1-score	support
0	0.95	0.98	0.97	14573
1	0.97	0.94	0.96	11403
accuracy			0.96	25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

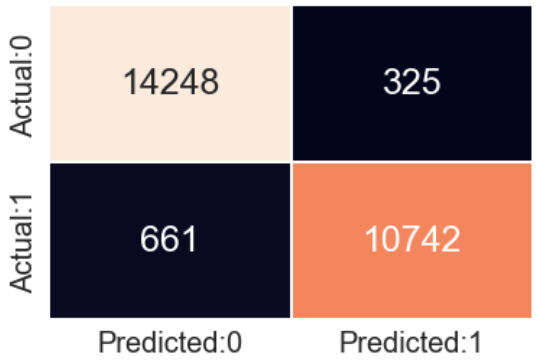
- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.95 and 0.97
- recall score for class 0 and class 1 are 0.98 and 0.94
- F1-score for class 0 and class 1 are 0.97 and 0.96.

8.3. Randomized SearchCV of Random Forest:

8.3.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 96%.

8.3.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -14248
- TP-10742
- FP-325
- FN-661

8.3.3. Classification Report :

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14573
1	0.97	0.94	0.96	11403
accuracy			0.96	25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

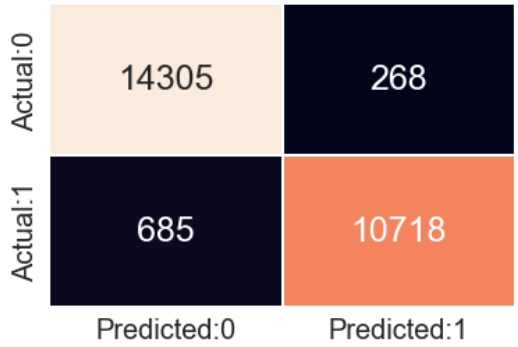
- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.96 and 0.97
- recall score for class 0 and class 1 are 0.98 and 0.94
- F1-score for class 0 and class 1 are 0.97 and 0.96.

8.4. Randomized SearchCV of Random Forest Using Best Features:

8.4.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 96%.

8.4.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -14305
- TP-10718
- FP-268
- FN-685

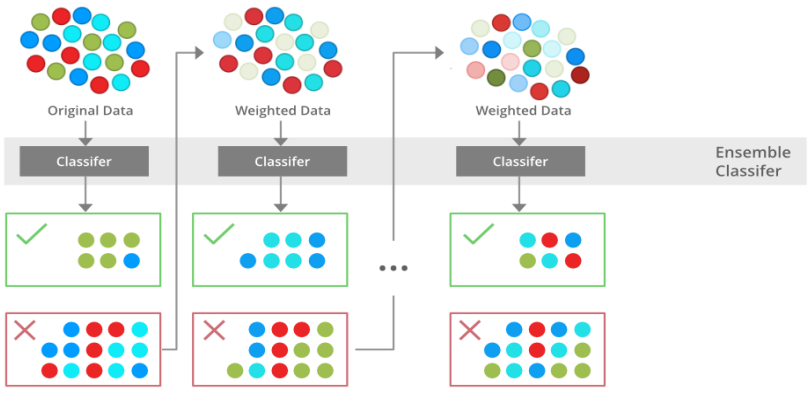
8.4.3. Classification Report :

	precision	recall	f1-score	support
0	0.95	0.98	0.97	14573
1	0.98	0.94	0.96	11403
accuracy			0.96	25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.95 and 0.98
- recall score for class 0 and class 1 are 0.98 and 0.94
- F1-score for class 0 and class 1 are 0.97 and 0.96.

9. XG BOOST

- Boosting is an ensemble modeling, technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built



from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

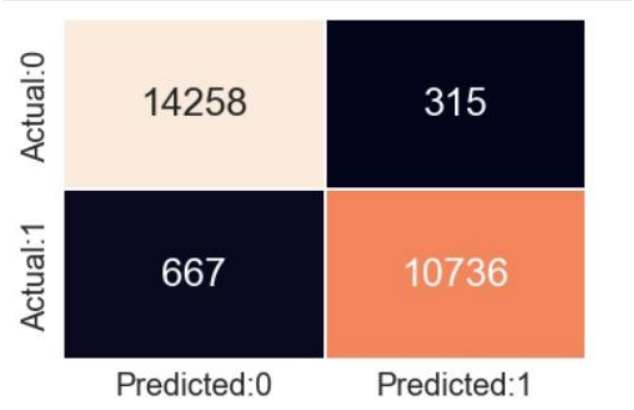
- In XGBoost, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

9.1. All Features:

9.1.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 96%.

9.1.2. Confusion Matrix:



- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -14258
- TP-10736
- FP-315
- FN-667

9.1.3. Classification Report :

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14573
1	0.97	0.94	0.96	11403
accuracy			0.96	25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.96 and 0.97
- recall score for class 0 and class 1 are 0.98 and 0.94
- F1-score for class 0 and class 1 are 0.97 and 0.96.

9.2. Best Features:

9.2.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 96%.

9.2.2. Confusion Matrix:

- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -14268
- TP-10745
- FP-305
- FN-658

9.2.3. Classification Report :

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14573
1	0.97	0.94	0.96	11403
accuracy			0.96	25976
macro avg	0.96	0.96	0.96	25976
weighted avg	0.96	0.96	0.96	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.96 and 0.97
- recall score for class 0 and class 1 are 0.98 and 0.94
- F1-score for class 0 and class 1 are 0.97 and 0.96.

9.3. XGBoost with PCA Features:

9.3.1. Accuracy :

- Accuracy is the ratio of correct predictions (i.e. TN+TP) to the total observations. Here, we get our accuracy is 82%.

9.3.2. Confusion Matrix:

- Confusion matrix gives True positive, True negative, False positive and False negative.
- In model classified correctly are TN (class 0), TP (class 1) and misclassified are FN (Actual '1' values which are classified wrongly as '0'), FP (Actual '0' values which are classified wrongly as '1').

Here, output of

- TN -12625
- TP-8724
- FP-1948
- FN-2679

9.3.3. Classification Report :

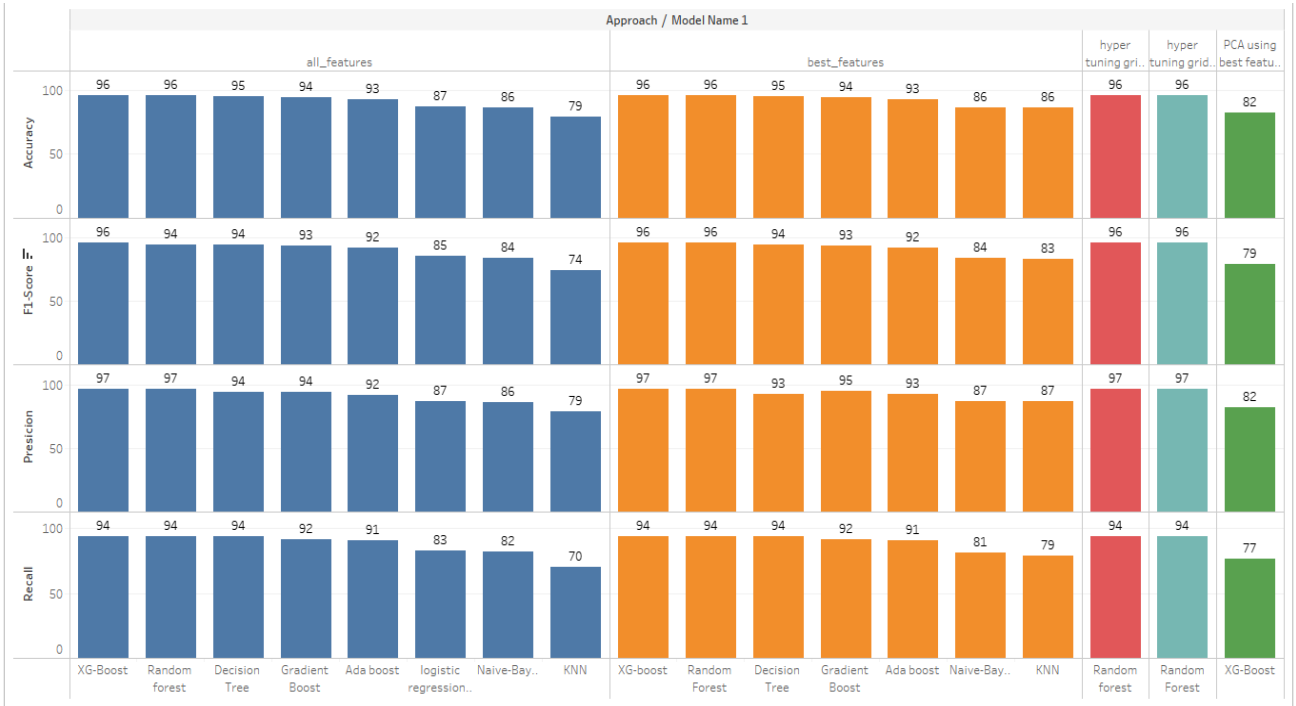
	precision	recall	f1-score	support
0	0.82	0.87	0.85	14573
1	0.82	0.77	0.79	11403
accuracy			0.82	25976
macro avg	0.82	0.82	0.82	25976
weighted avg	0.82	0.82	0.82	25976

- In classification report, we get various performance measures such as precision, recall, f1-score and accuracy.
- Precision score for class 0 and class 1 are 0.82 and 0.82
- recall score for class 0 and class 1 are 0.87 and 0.77
- F1-score for class 0 and class 1 are 0.85 and 0.79.

10. SUMMARY OF THE FINDINGS:

	model_name	Approach	Accuracy	Presicion	Recall	F1-Score
0	logistic regression -base_model	all_features	87	87	83	85
1	Naive-Bayes	all_features	86	86	82	84
2	KNN	all_features	79	79	70	74
3	Decision Tree	all_features	95	94	94	94
4	Random forest	all_features	96	97	94	94
5	Ada boost	all_features	93	92	91	92
6	Gradient Boost	all_features	94	94	92	93
7	XG-Boost	all_features	96	97	94	96
8	Random Forest	best_features	96	97	94	96
9	Naive-Bayes	best_features	86	87	81	84
10	KNN	best_features	86	87	79	83
11	Decision Tree	best_features	95	93	94	94
12	Ada boost	best_features	93	93	91	92
13	Gradient Boost	best_features	94	95	92	93
14	XG-boost	best_features	96	97	94	96
15	Random forest	hyper tuning gridsearchCV	96	97	94	96
16	Random Forest	hyper tuning gridsearchCV with best features	96	97	94	96
17	XG-Boost	PCA using best features	82	82	77	79

- From the above we can find that XGBoost (with all features) and Random Forest (using randomizer search cv) features seems to be performing good.



11. SUGGESTIONS FOR AIRLINE PASSENGER SATISFACTION:

Based on EDA observations the following suggestions have been made:

- Seat comfort, inflight experience and ease online level significantly affect the customer experience along with the several other variables considered.
- Airlines should highly focus on inflight Wi-Fi experience.
- Ease of online booking is important for business customers.
- The airline service companies must ensure the high quality of service in these parameters to ensure a high level of customer satisfaction.

12. REFERENCES:

- Hiver - https://hiverhq.com/blog/customer-service-travel-hospitality
- Kaggle - https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
- Towards Data science - https://towardsdatascience.com/predicting-satisfaction-of-airline-passengers-with-classification-76f1516e1d16

12.1. Notes For Project Team

Sample Reference for Datasets (to be filled by team and mentor)

Original owner of data	John
Data set information	Airline Passenger Satisfaction
Any past relevant articles using the dataset	-
Reference	Kaggle
Link to web page	https://www.kaggle.com/datasets/tejmahal20/airline-passenger-satisfaction