Department of Economics, Management and Quantitative Methods
Università degli Studi di Milano
Data science and economics

Machine Learning Project

Airline Passenger Satisfaction

Author: Ulan Shaikyp
944462

Academic year 2021-2022

Abstract.

In this paper we are going to present the project for Machine Learning module exam. In particular, we will compare some fundamental classification methods and their results on the Airline Passenger Satisfaction dataset.

Introducing data

The dataset for this project consists of airline passengers, their experience of the flight and their satisfaction will be used. Whilst the dataset has a pre-existing test dataset, the bulk of the focus will be on the train dataset. It should be noted that the test dataset and the train dataset have the same columns and the test dataset is around 20% of the whole dataset (test and train dataset). The data set has 25,975 data points and 25 columns.

- *Gender:* Gender of the passengers (Female, Male)
- *Customer Type:* The customer type (Loyal customer, disloyal customer)
- *Age:* The actual age of the passengers
- *Type of Travel:* Purpose of the flight of the passengers (Personal Travel, Business Travel)
- *Class:* Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- *Flight distance:* The flight distance of this journey
- *Inflight wifi service:* Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
- *Departure/Arrival time convenient:* Satisfaction level of Departure/Arrival time convenient
- *Ease of Online booking:* Satisfaction level of online booking
- *Gate location:* Satisfaction level of Gate location
- *Food and drink:* Satisfaction level of Food and drink
- *Online boarding:* Satisfaction level of online boarding
- *Seat comfort:* Satisfaction level of Seat comfort
- *Inflight entertainment:* Satisfaction level of inflight entertainment
- *On-board service:* Satisfaction level of On-board service
- *Leg room service:* Satisfaction level of Leg room service
- *Baggage handling:* Satisfaction level of baggage handling
- *Check-in service:* Satisfaction level of Check-in service
- *Inflight service:* Satisfaction level of inflight service
- *Cleanliness:* Satisfaction level of Cleanliness
- *Departure Delay in Minutes:* Minutes delayed when departure
- *Arrival Delay in Minutes:* Minutes delayed when Arrival
- *Satisfaction:* Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Note that this data set was modified from this dataset by John D . It has been cleaned up for the purposes of classification.

	Unnamed: 0	id	Age	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	In
count	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	
mean	12987.500000	65005.657992	39.620958	1193.788459	2.724746	3.046812	2.756775	2.977094	3.215353	3.261665	3.449222	
std	7498.769632	37611.526647	15.135685	998.683999	1.335384	1.533371	1.412951	1.282133	1.331506	1.355536	1.320090	
min	0.000000	17.000000	7.000000	31.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000	
25%	6493.750000	32170.500000	27.000000	414.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	
50%	12987.500000	65319.500000	40.000000	849.000000	3.000000	3.000000	3.000000	3.000000	3.000000	4.000000	4.000000	
75%	19481.250000	97584.250000	51.000000	1744.000000	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	5.000000	
max	25975.000000	129877.000000	85.000000	4983.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	

Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes
25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.00000	25893.000000
3.357753	3.385664	3.350169	3.633238	3.314175	3.649253	3.286226	14.30609	14.740857
1.338299	1.282088	1.318862	1.176525	1.269332	1.180681	1.319330	37.42316	37.517539
0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.00000	0.000000
2.000000	2.000000	2.000000	3.000000	3.000000	3.000000	2.000000	0.00000	0.000000
4.000000	4.000000	4.000000	4.000000	3.000000	4.000000	3.000000	0.00000	0.000000
4.000000	4.000000	4.000000	5.000000	4.000000	5.000000	4.000000	12.00000	13.000000
5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	1128.00000	1115.000000

1.1 Basic description of the original dataset.

Before moving to further explorative analysis, we checked for potential missing errors.

As a matter of fact, we found that the missing values are present in the Arrival Delay in Minute's variable. Below is a screenshot showing number missing values in each column.

```
In [7]:
1
2 # Checking missing values
3 df.isnull().sum()

Out[7]: id                          0
Gender                      0
Customer_Type               0
Age                         0
Travel_Type                 0
Class                       0
Flight_Distance             0
Wifi_Service                0
Departure_Arrival_Time      0
Online_Booking              0
Gate_Location               0
Food_Drink                  0
Online_Boarding             0
Seat_Comfort                0
Inflight_Entertainment      0
Onboard_Service             0
LegRoom_Service             0
Baggage_Handling            0
Checking_Service            0
Inflight_Service            0
Cleanliness                 0
Departure_Delay_Time        0
Arrival_Delay_Time         83
satisfaction                0
dtype: int64
```

Figure 1.2 Missing Values

The percentage of missing values in Arrival Delay time is 0.31%. We can either ignore or we can impute values by using mean, median, and KNN imputation method. Now, we can fill the missing values with 0 and check for any duplicates in the dataset. And we didn’t find any duplicate in our dataset.

Since the data is cleaned and processed now, we can start the Exploratory Data Analysis. We can see summary statistics like mean, median, mode, count etc

Correlation Matrix is displayed to find the correlation between the variables. It can be observed that there is less correlation between the variables.

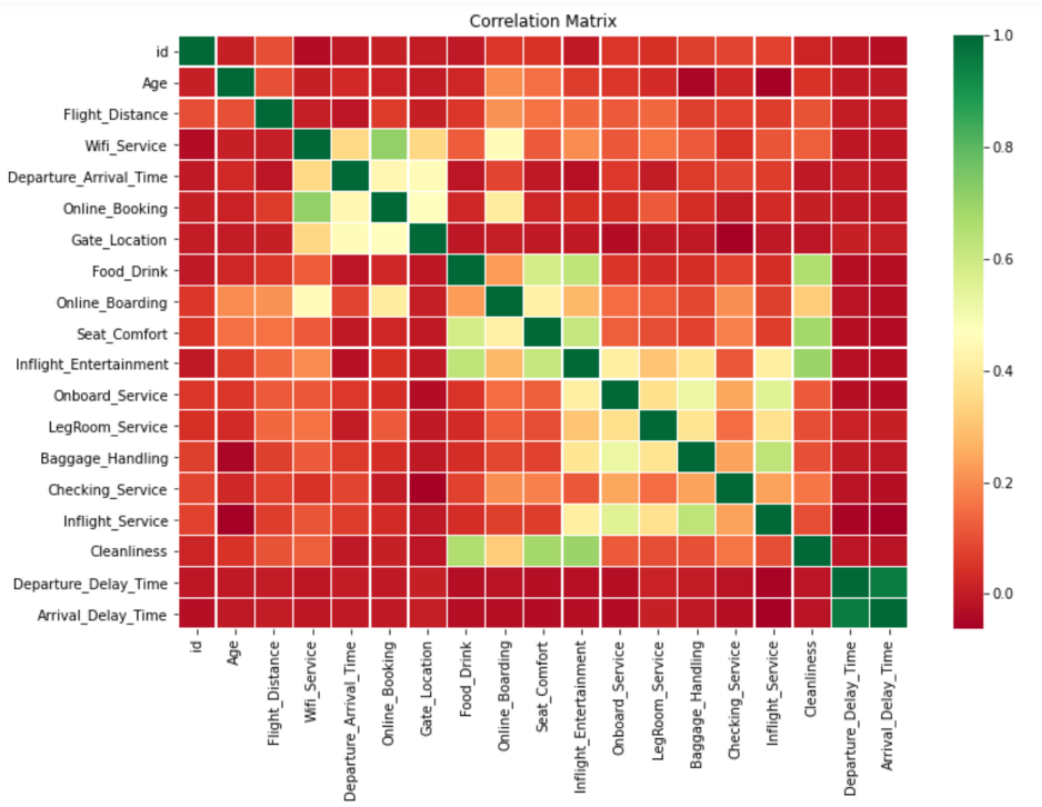


Figure 1.3 Correlation Matrix

Univariate Analysis is used to know the count of the Gender and is displayed below in a bar chart. From the graph we know that number of Female passengers are more than that of Male passengers.

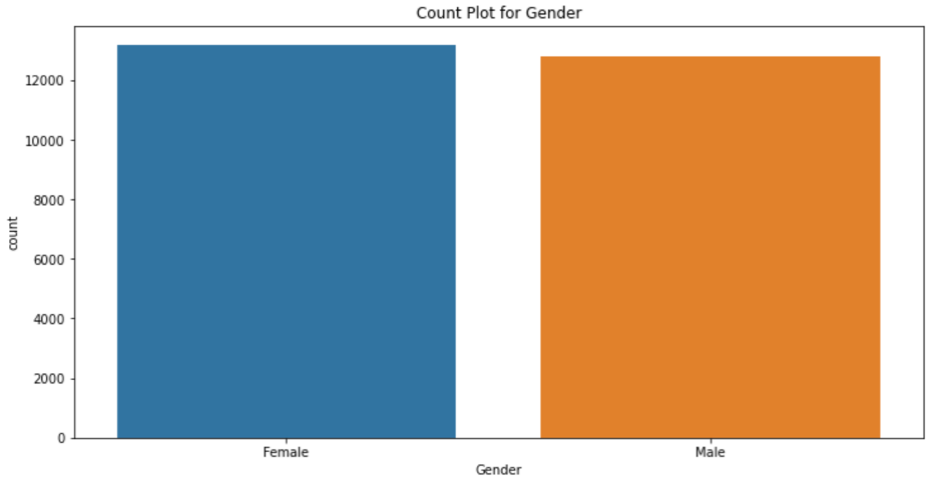


Figure 1.4 Bar Graph to know the count of Gender

We can also count the number of loyal customers and disloyal customers using the bar chart. From the graph shown below we can find that loyal customers are more than disloyal customers.

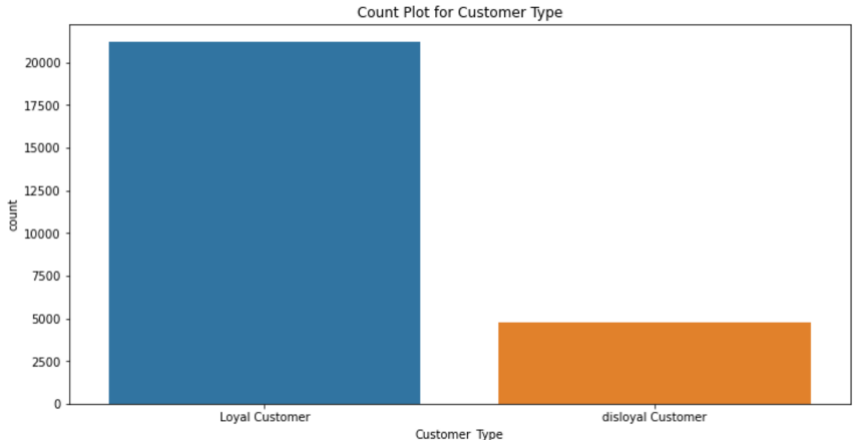


Figure 1.5 Bar Graph to know the count of Customer Type

Let us find use travel type to know the count business travel and personal travel. From the figure 1.13 we see that business travel passengers are more than personal travel passengers.

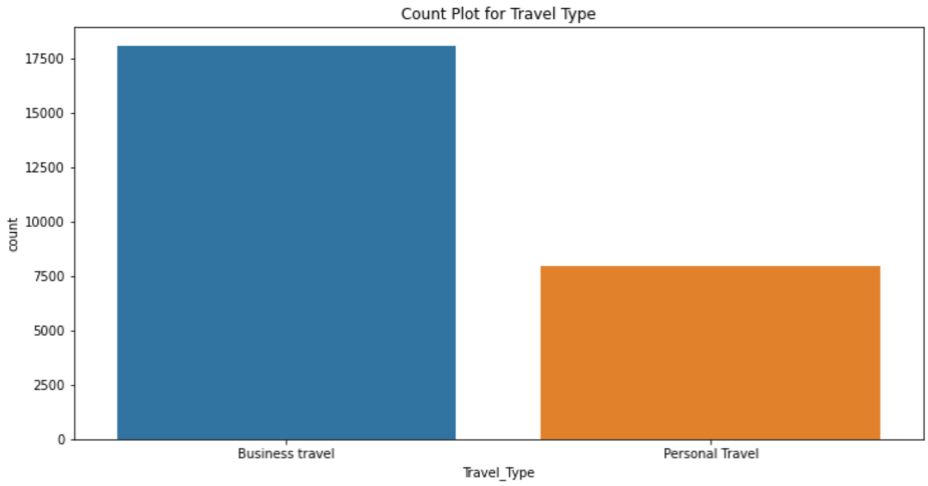


Figure 1.6 Bar Graph of Travel Type

Now, we can find the class which maximum and lowest number of passenger's travel. There are 3 different classes they are Economy, Business class, and Economy Plus. We can find that the count of Business class passengers is the highest and Economy Plus passengers are the lowest.

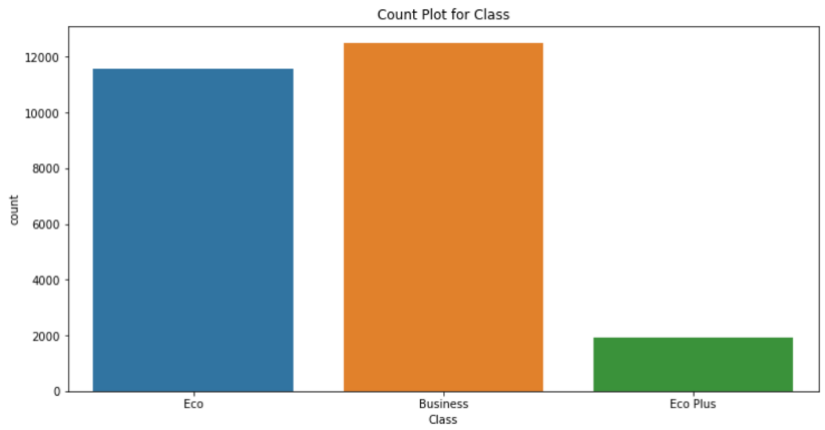


Figure 1.7 Bar Graph of Class

Figure 1.15 shows the customer satisfaction among the airline passengers. From this we can conclude that neutral or dissatisfied customers are higher than the satisfied customers.

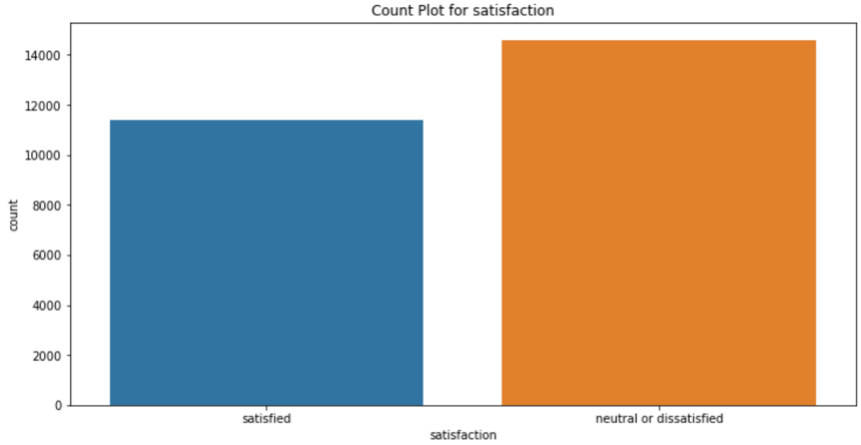


Figure 1.8 Bar Graph of Customer Satisfaction

Bivariate Analysis is used to find the average of customer satisfaction, average of gender, average of customer type, average of travel type, and average of class. The below chart represents the average age of customer satisfaction. The average age of satisfied passengers is around 42 and the average age of neutral or dissatisfied passengers is around 37.

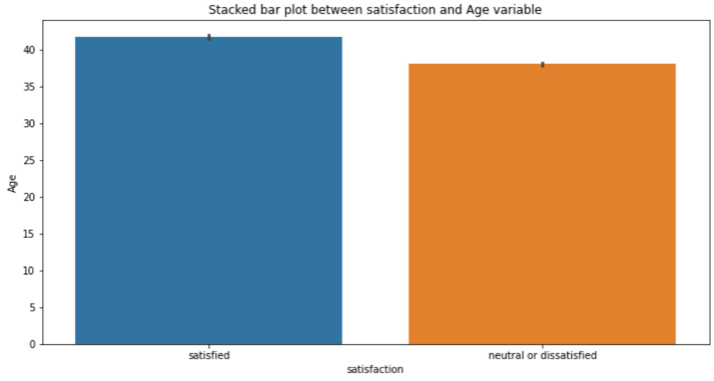


Figure 1.9 Average age of Customer Satisfaction

Next, we can find the average age of Male and Female passengers using bar chart. The average age of Male and Female passengers is around 39.

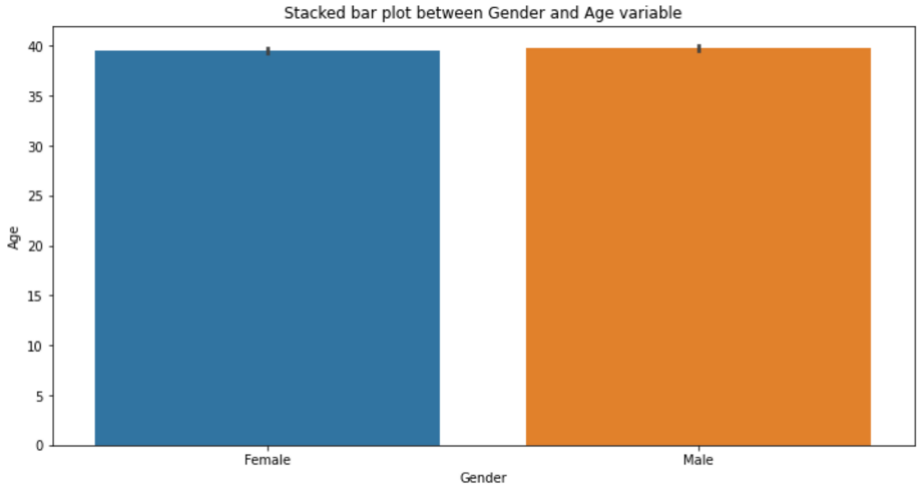


Figure 1.10 Average Age of Gender

Average age of loyal customer is around 42 and average of disloyal customer is around 29.

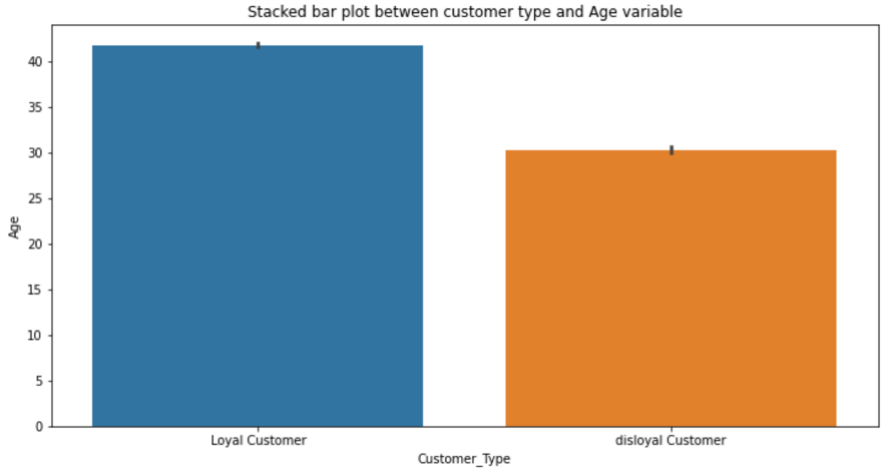


Figure 1.11 Average Age of Loyal & Disloyal Customer

Now, we will find the average age of business travel and personal travel. The average age of business travel passengers is around 39 and the personal travel passengers is around 38.

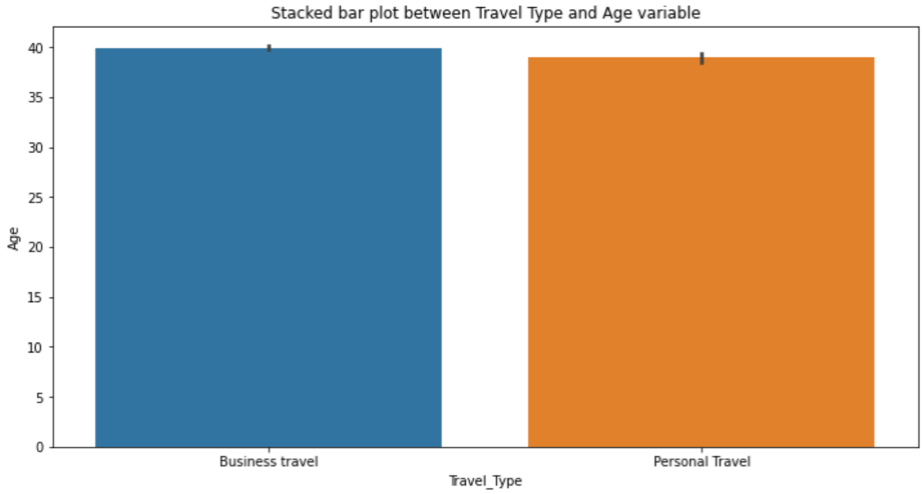


Figure 1.12 Average Age of Business Travel

The average age of all class passengers is visualized. From the figure 1.19 we can say that the average age of Eco class passengers is around 37, the average age of business class passengers is around 43 and the average age of economy plus is around 38.

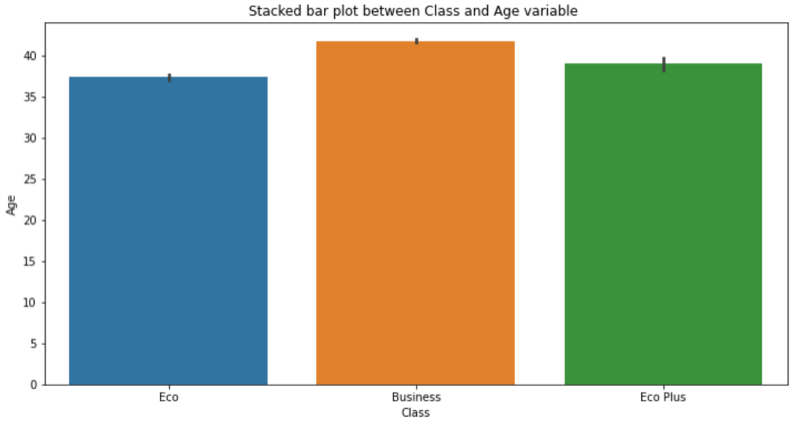


Figure 1.3Average Age of Class

BUILDING MODELS FOR CUSTOMER SATISFACTION

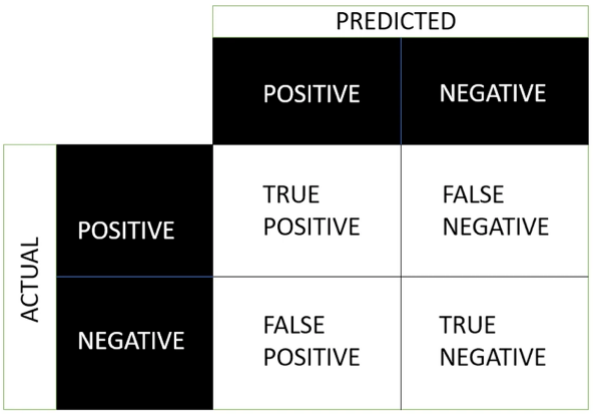
In this section, we will be building various models to predict the customer satisfaction and their ROC and accuracy for performance. Let's start by encoding the variables, since we are going to use this dataset to predict customer satisfaction we will be encoding the 'satisfied' customer as 1 and 'neutral or dissatisfied' customer as 0 and we have encoded the other variables too that are Gender, Customer_Type, Travel_type, and Class. If we need to build models and predict based on these variables the encoded variables can be used. We split the dataset into train and test, we have split the model as 80% as training data and 20% as test data to use in our models.

Decision Tree Classifier: Decision Tree Classifier is a simple machine learning model for classification problems. This is a type of supervised machine

learning where we build a model and feed data with correct outputs and then we let the model learn from these patterns. Then we feed our model new data that it hasn't observed before to see how it performs. As we know, decision tree has three nodes which are root node, decision node and terminal node. Hence, predicting a value means asking question from top node to the terminal node where we get a decision.

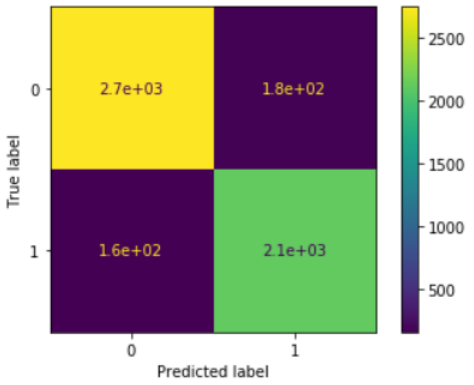
We have built a decision tree classifier model where it shows the CPU time, system time, wall time and total time in milliseconds and model performance is executed. Satisfied and neutral or dissatisfied passengers are represented and the accuracy score of decision tree classifier is found out to be 0.9351424172440339.

We have created a confusion matrix and performance report. To better understand about confusion matrix.

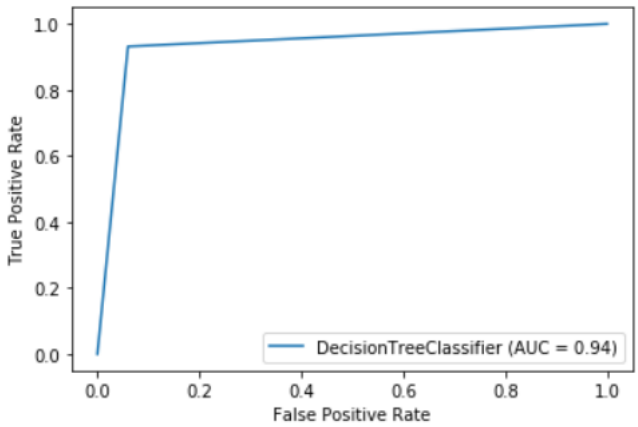


Confusion Matrix

Let's us discuss about the terms mentioned in the confusion matrix, True Positives: Actual positive values and predicted positive values are same. True Negatives: Actual Negative values and predicted negative values are same. False Positives: Where we have actual negative values and predicted positive values. False Negatives: Where we have actual positive values and predicted negative values. So, in our case, we have True Positives: Number of correctly predicted neutral or dissatisfied passengers is 2,123 True Negatives: Number of correctly predicted satisfied passengers is 2,7362 False Positives: Number of incorrectly predicted dissatisfied or neutral passengers is 175 False Negatives: Number of incorrectly predicted satisfied passengers is 162 F1 score is 0.94 and area under curve is 0.93 meaning the model is performing well.

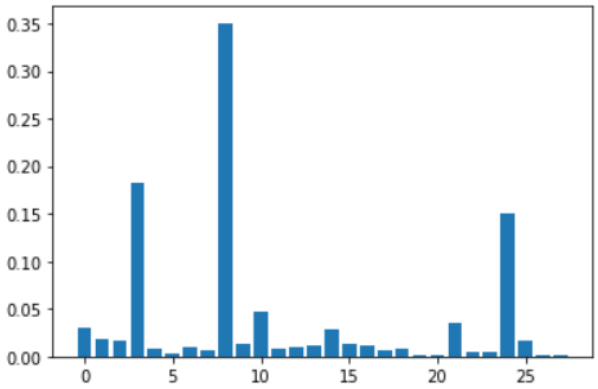


	precision	recall	f1-score	support
0	0.95	0.94	0.94	2918
1	0.92	0.93	0.93	2278
accuracy			0.94	5196
macro avg	0.94	0.94	0.94	5196
weighted avg	0.94	0.94	0.94	5196



The AUC values represent the model performance and their range is mentioned in the below table.

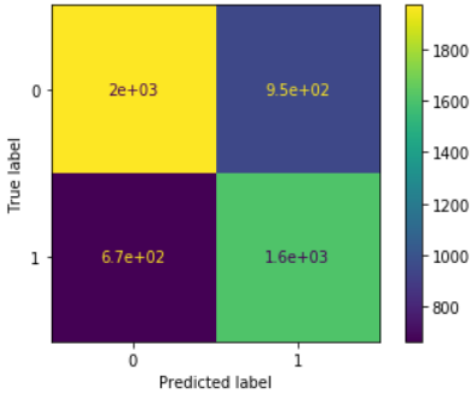
AUC VALUES TEST QUALITY 0.9 – 1.0 Excellent 0.8 - 0.9 Very Good 0.7 – 0.8 Good 0.6 – 0.7 Satisfactory 0.5 – 0.6 Unsatisfactory



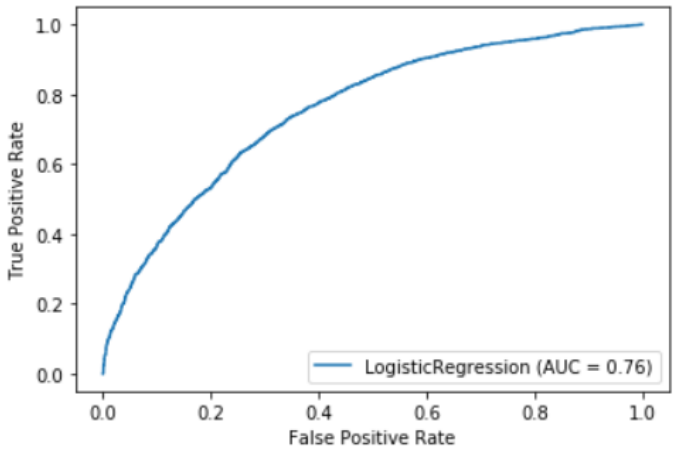
Here represents the important features for decision tree model and the bar plot.

Logistic Regression Classifier: Logistic Regression is a machine learning algorithm to predict the probability of a categorical dependent variable. In logistic regression, dependent variable is a binary value where it contains data as either 1 or 0. Logistic regression is a special type of linear regression in which the target

variable is categorical. Logistic regression can predict the probability of an occurrence of a binary event.

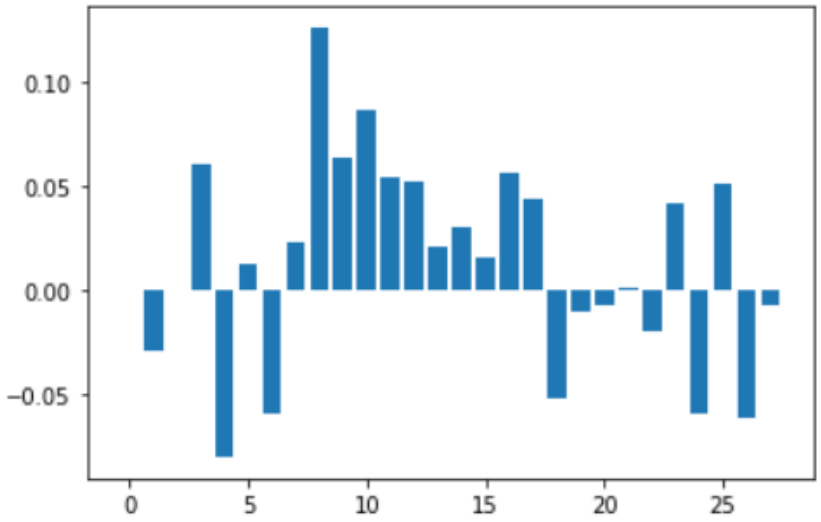


	precision	recall	f1-score	support
0	0.75	0.68	0.71	2918
1	0.63	0.71	0.67	2278
accuracy			0.69	5196
macro avg	0.69	0.69	0.69	5196
weighted avg	0.70	0.69	0.69	5196



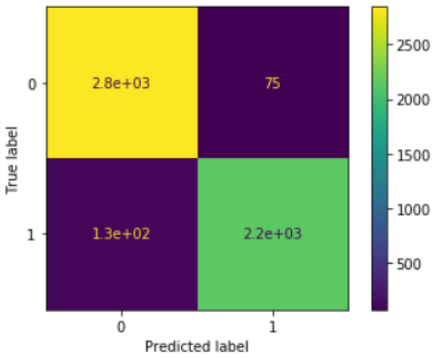
Here represents the logistic regression classifier confusion matrix, performance report, ROC curve, and importance features.

And it represents the logistic regression model and the model accuracy is found to be 0.6774441878367975 . F1-score obtained from performance report as 0.68 with area under curve value as 0.75. The AUC value of 0.75 means that the model is performing good but it is not excellent. In confusion matrix we can say that, True Positives: Number of correctly predicted neutral or dissatisfied passengers is 1,675 True Negatives: Number of correctly predicted satisfied passengers is 1,845 False Positives: Number of incorrectly predicted dissatisfied or neutral passengers is 623 False Negatives: Number of incorrectly predicted satisfied passengers is 1053.

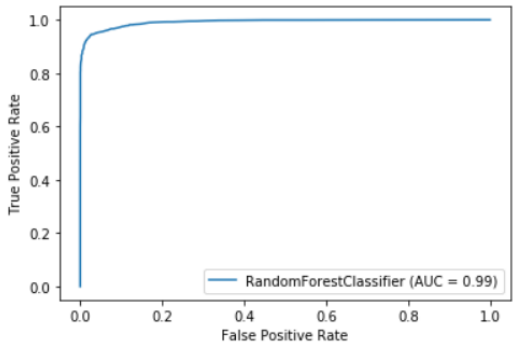


Here represents the importance feature and the scores and a bar plot of importance features.

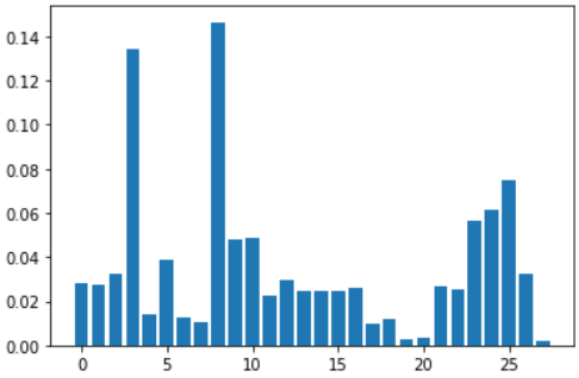
Random Forest Classifier: Random Forest is a supervised learning algorithm that can be used for both regression and classification. Random Forests creates decision trees on randomly selected data samples then gets prediction from each tree to select the best solution by voting. Prediction result that we get with the most votes is the final prediction. Random Forest is regarded as a highly accurate and strong method because of the number of decision trees involved in the process.



	precision	recall	f1-score	support
0	0.96	0.97	0.97	2918
1	0.97	0.94	0.95	2278
accuracy			0.96	5196
macro avg	0.96	0.96	0.96	5196
weighted avg	0.96	0.96	0.96	5196



This represents random forest classifier model where it shows the CPU time, system time, wall time and total time in milliseconds. And Satisfied and neutral or dissatisfied passengers are represented and the accuracy score of random forest classifier is found out to be 0.9586220169361047.



It represents the importance feature for random forest model and its score with the bar plot of importance feature. The table below concludes all the finding from the models.

	Model	Accuracy	f1_score	auc
0	Decision_Tree	0.93	0.94	0.93
1	Logistic_Regression	0.67	0.68	0.75
2	Random_Forest	0.95	0.96	0.99

Model Accuracy F1-SCORE AUC Decision Tree Classifier 0.93 0.94 0.93 Logistic Regression Classifier 0.67 0.68 0.75 Random Forest Classifier 0.95 0.96 0.99.

CONCLUSION

The selected data set was checked for duplicates, renaming the column, checked for duplicates before starting the Exploratory Data Analysis. Anomalies and Outliers was found on flight distance. Various different machine learning models were used to predict the customer satisfaction in which we can find the best performing model for prediction. The Accuracy, f1- score, AUC values and importance features were found for each model.

Random Forest Classifier model has the highest accuracy among all the models with highest value for both f1- score and Area Under Curve value.