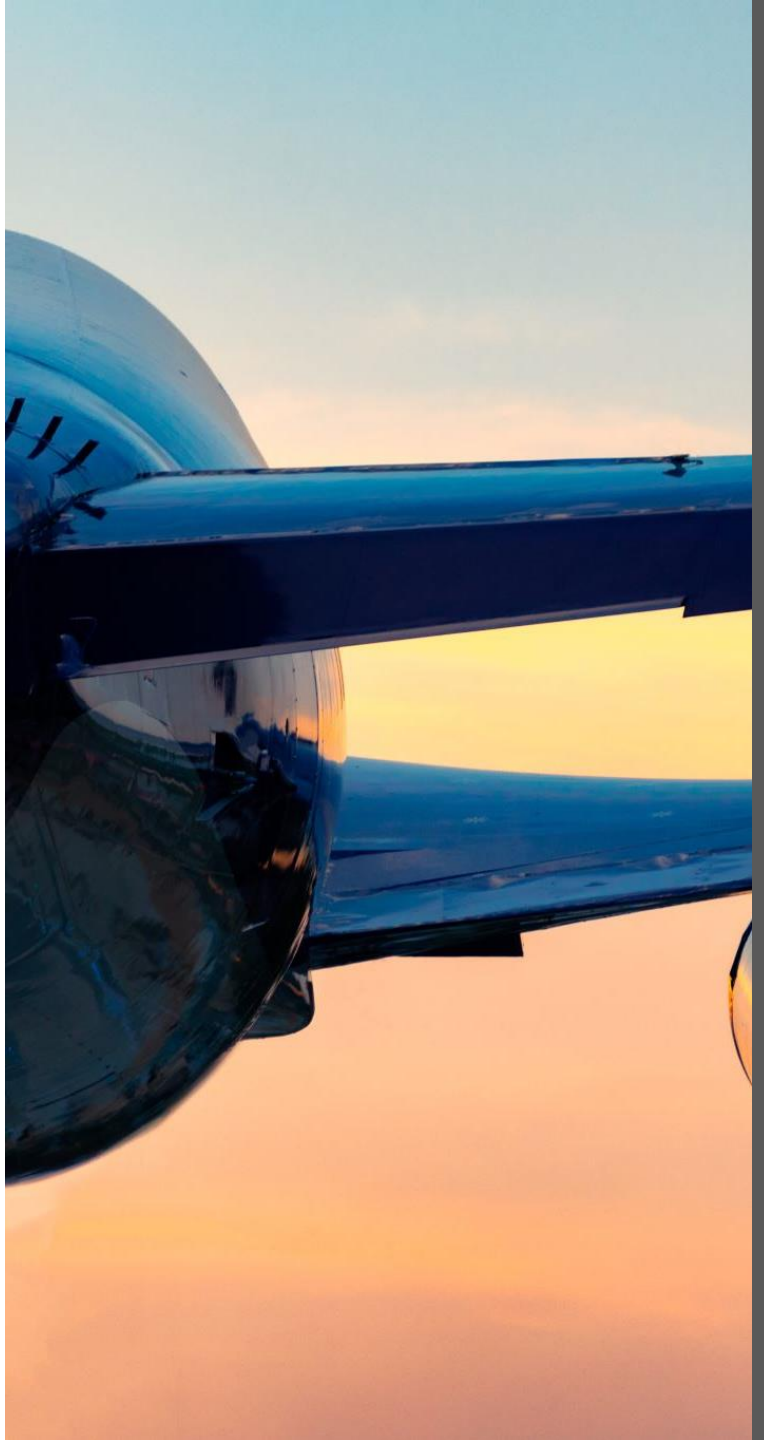
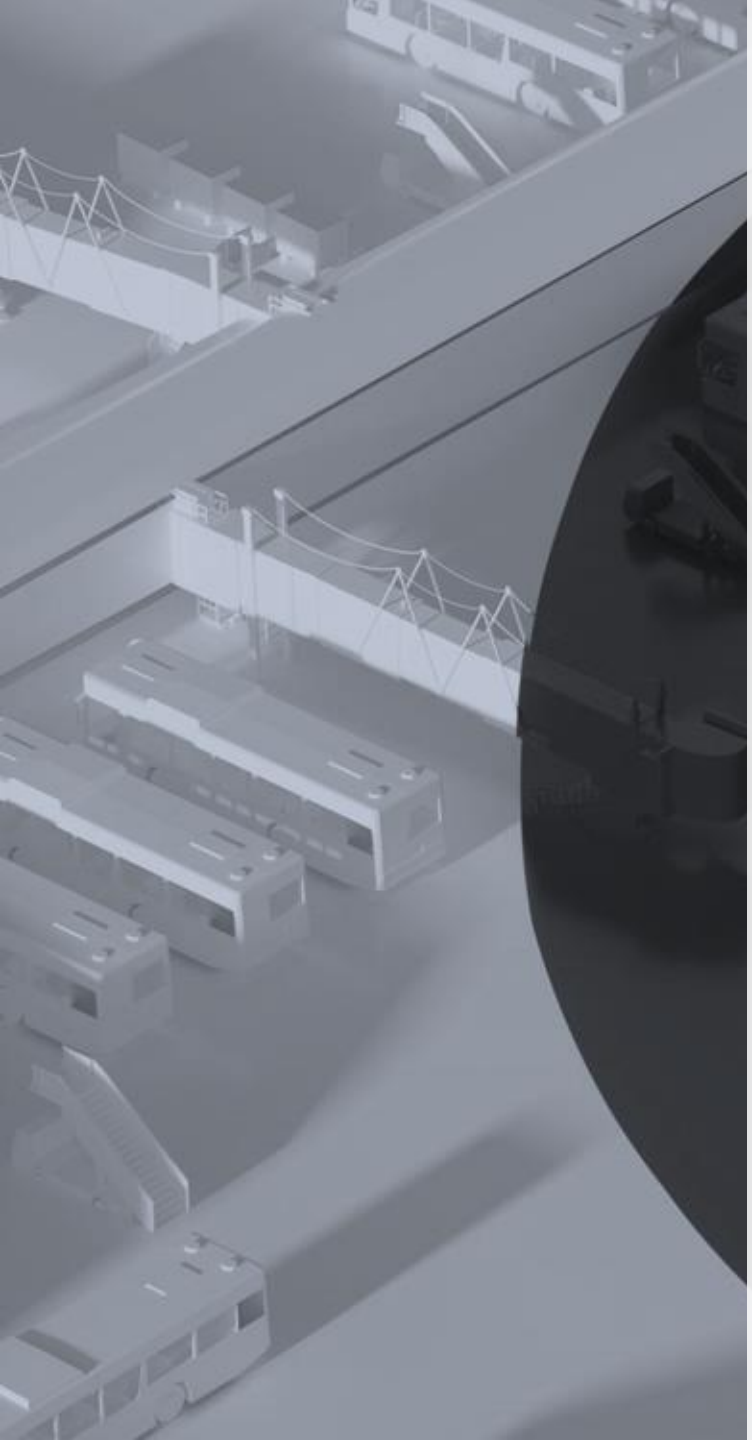
Predicting Customer Satisfaction for Airlines

Introduction

- Customer satisfaction is always top of mind for the aviation industry.
- Unhappy or disengaged customers naturally mean fewer passengers and less revenue.
- It's important that customers have an excellent experience every time they travel.
- The objective of this project is to predict whether a future customer would be satisfied with their service given the details of the other parameters values.

- ***What factors affect customer satisfaction with the Airlines?***
- ***Can we predict the likelihood of a customer being satisfied with the Airline service?***





Data

- The data for this project is provided by Kaggle dataset.
- This data is given by an airline organization. The actual name of the company is not given due to various purposes that's why the name Invistico Airlines.
- The dataset consists of the details of customers who have already flown with them.
- This data shows whether a customer is satisfied with the airlines or not after travelling with them.
- There are several other measurement or to say feedback taken from the customers as well as their demographic data is also recorded.
- The variable *'Satisfaction'* is the target variable where satisfaction=1 means the customer is satisfied with the airline service and satisfaction=0 means the customer is not satisfied with the airline service.
- The exercise is focused on implementing classification Machine learning algorithms to perform the analysis.

Data Wrangling

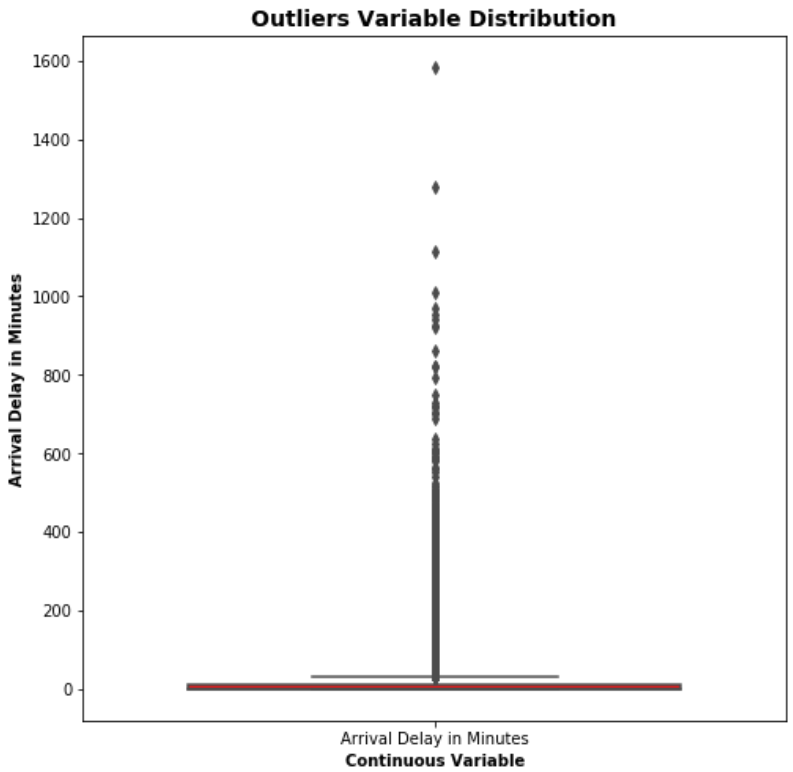
In the data wrangling section, I have cleaned and prepared the data.

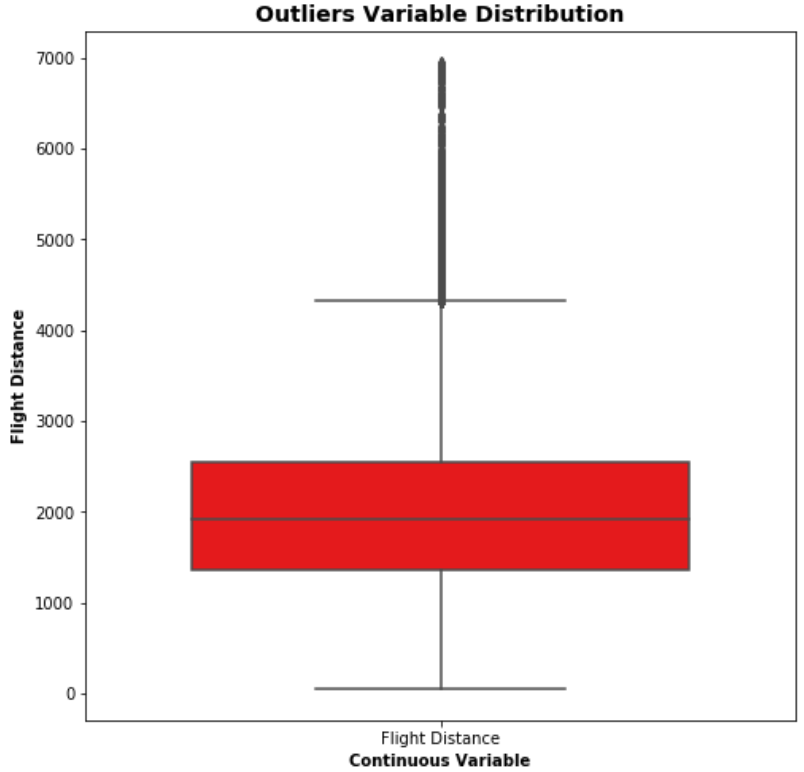
Data Cleaning involved the process of dealing with missing values, looking for outliers in the data, checking class imbalance and labelling accordingly.

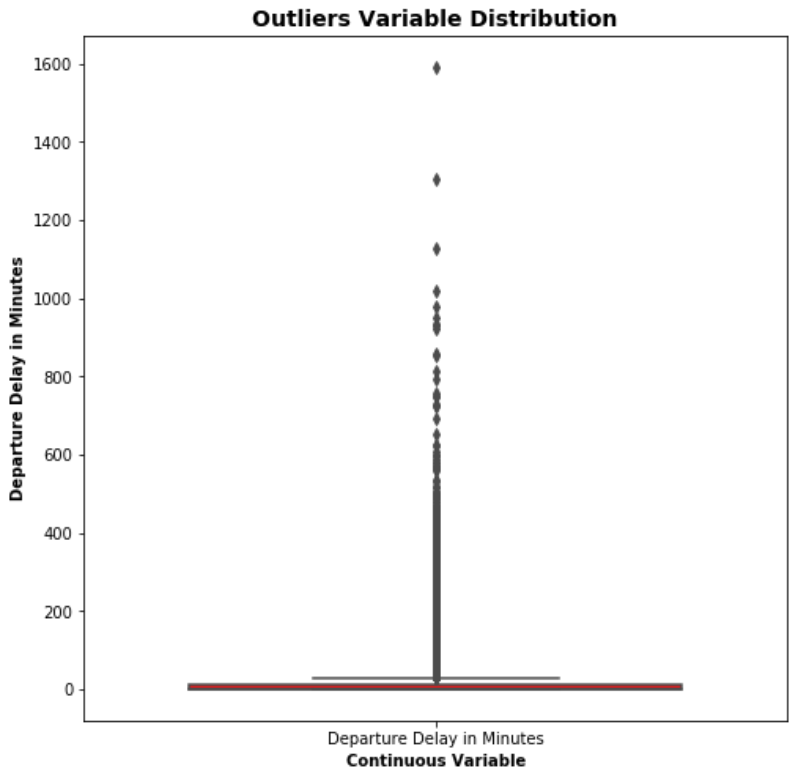
Since, only one attribute contained 3% missing values , those missing records were dropped from the dataset.

The outliers in the variables 'Arrival Delay in Minutes', 'Departure Delay in Minutes' and 'Flight Distance' contains important information about the airlines. Hence, I didn't drop these outliers from the data.

In addition, while checking for any class imbalance, I found that the data is almost class balanced.







Outliers in the data

Missing Values

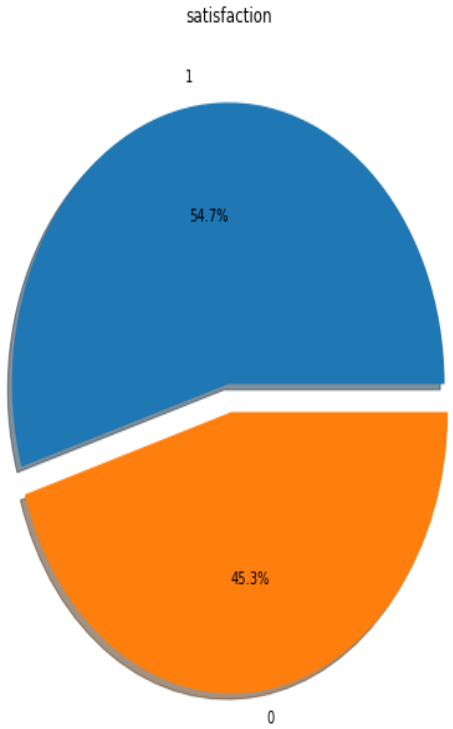
	column_name	percent_missing
satisfaction	satisfaction	0.000000
Gender	Gender	0.000000
Customer Type	Customer Type	0.000000
Age	Age	0.000000
Type of Travel	Type of Travel	0.000000
Class	Class	0.000000
Flight Distance	Flight Distance	0.000000
Seat comfort	Seat comfort	0.000000
Departure/Arrival time convenient	Departure/Arrival time convenient	0.000000
Food and drink	Food and drink	0.000000
Gate location	Gate location	0.000000
Inflight wifi service	Inflight wifi service	0.000000
Inflight entertainment	Inflight entertainment	0.000000
Online support	Online support	0.000000
Ease of Online booking	Ease of Online booking	0.000000
On-board service	On-board service	0.000000
Leg room service	Leg room service	0.000000
Baggage handling	Baggage handling	0.000000
Checkin service	Checkin service	0.000000
Cleanliness	Cleanliness	0.000000
Online boarding	Online boarding	0.000000
Departure Delay in Minutes	Departure Delay in Minutes	0.000000
Arrival Delay in Minutes	Arrival Delay in Minutes	0.302587

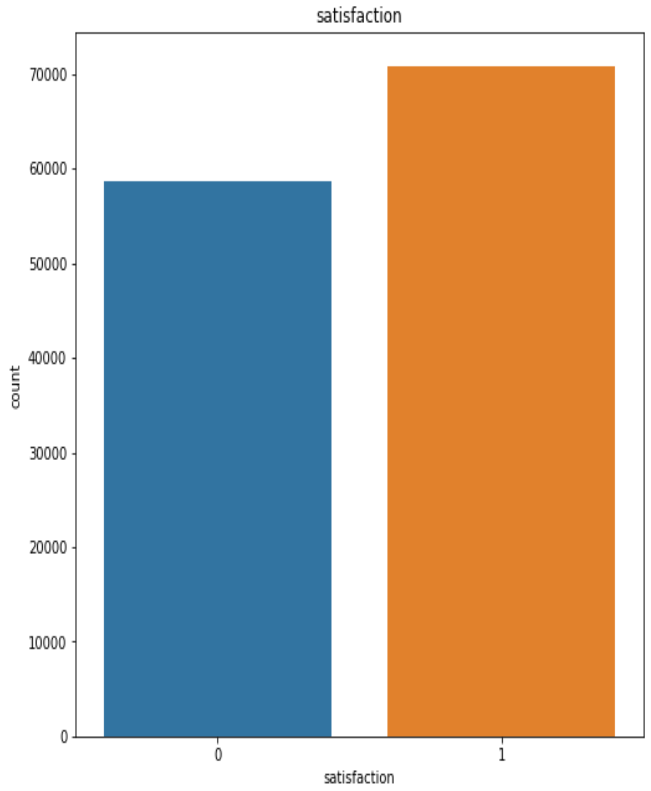
Correlations

Departure_Delay_in_Minutes	Arrival_Delay_in_Minutes	0.740284
Seat_comfort	Food_and_drink	0.705452
Ease_of_Online_booking	Online_boarding	0.662621
Online_support	Online_boarding	0.650395
Baggage_handling	Cleanliness	0.641254
Inflight_wifi_service	Online_boarding	0.616602
Online_support	Ease_of_Online_booking	0.603637
Inflight_wifi_service	Ease_of_Online_booking	0.579588
On_board_service	Cleanliness	0.578680
satisfaction	Inflight_entertainment	0.577601
On_board_service	Baggage_handling	0.555065
Departure/Arrival_time_convenient	Gate_location	0.554920
	Food_and_drink	0.538209
Inflight_wifi_service	Online_support	0.534364
Food_and_drink	Gate_location	0.533950
Eco	Personal_Travel	0.501048
Ease_of_Online_booking	On_board_service	0.464856
Inflight_entertainment	Online_support	0.460996
Ease_of_Online_booking	Cleanliness	0.452210
Seat_comfort	Departure/Arrival_time_convenient	0.438744
Ease_of_Online_booking	Baggage_handling	0.431395
satisfaction	Ease_of_Online_booking	0.429579
On_board_service	Leg_room_service	0.426280
Leg_room_service	Cleanliness	0.423042
	Baggage_handling	0.420031
Seat_comfort	Gate_location	0.410114
satisfaction	Online_support	0.402107
Seat_comfort	Inflight_entertainment	0.399236
Ease_of_Online_booking	Leg_room_service	0.377119
Inflight_entertainment	Online_boarding	0.369503

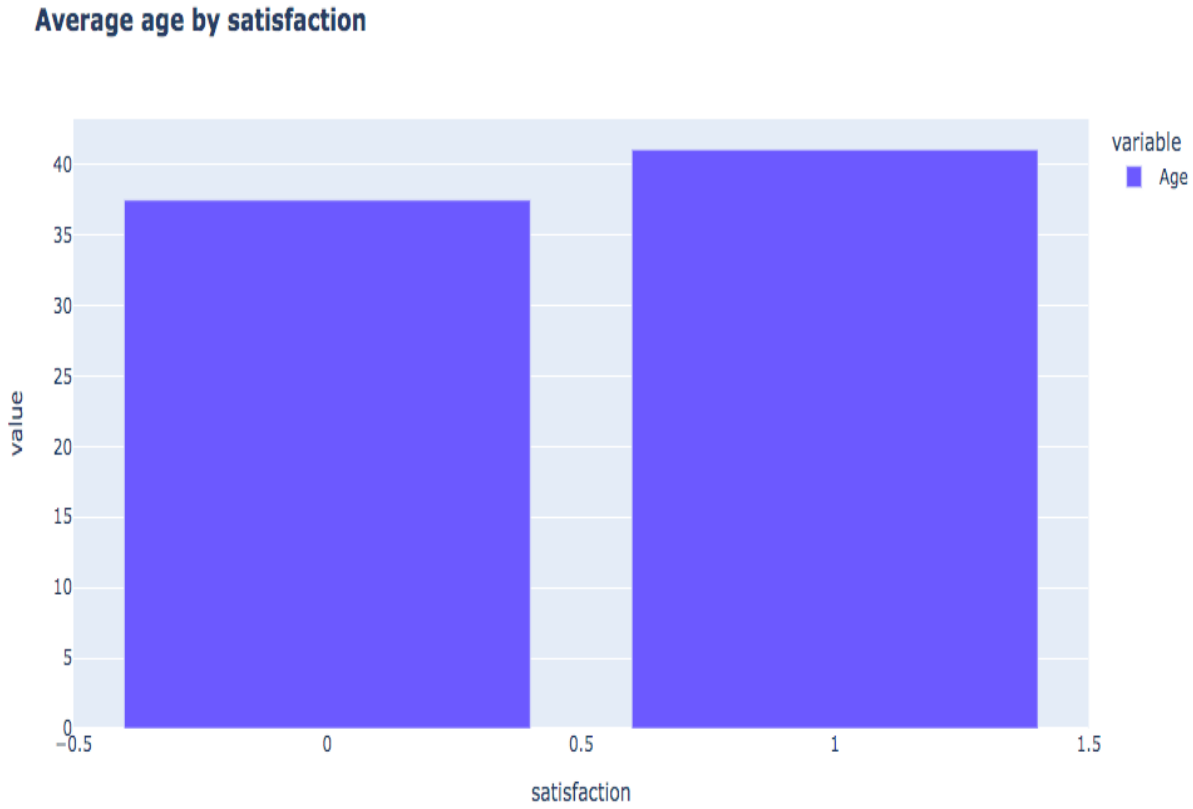
EDA (Exploratory Data Analysis)

54.7% people are satisfied with the Airlines and 45.3% are not satisfied with the Airlines.



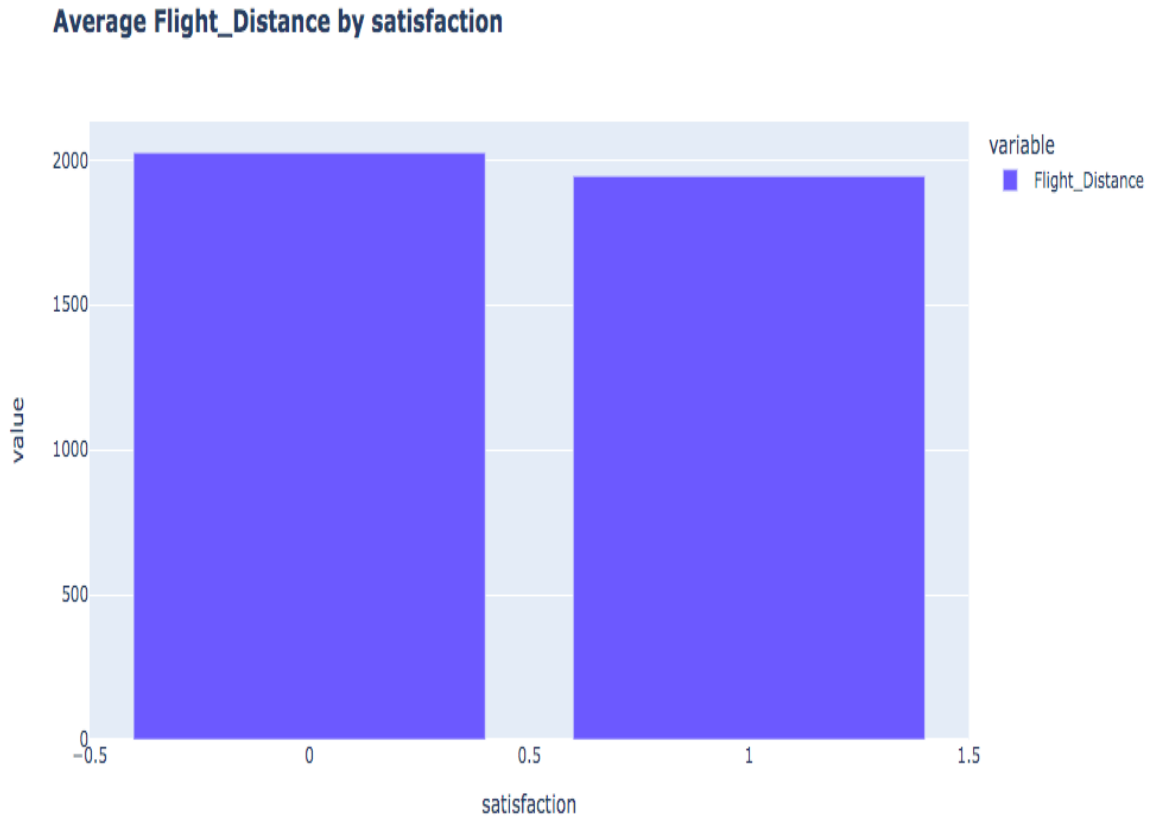


Most satisfied customers lie in the age group of 41 years approx. and most dissatisfied customers lie in the age group of 38 years (approx).

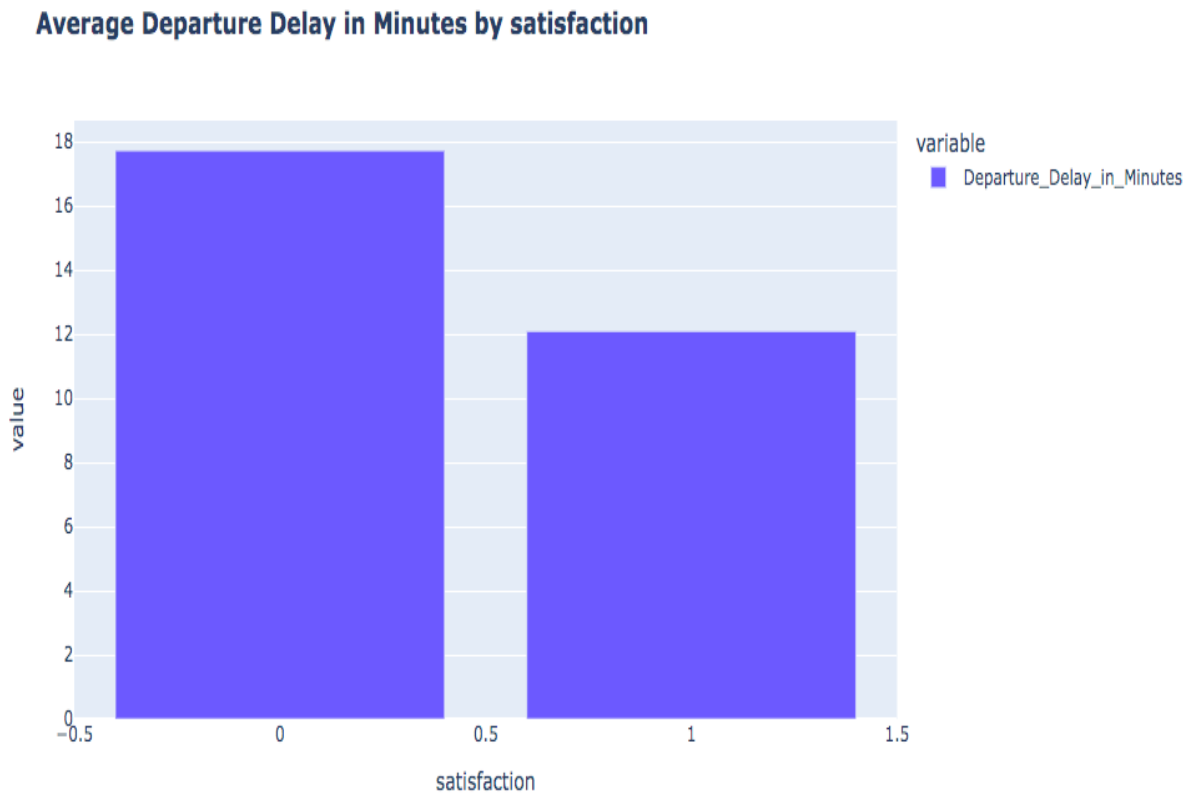


EDA (Exploratory Data Analysis)

Most satisfied customers flew a flight distance of 1944.47 miles with the airlines and dissatisfied customers flew around 2025.203 miles with the airlines.



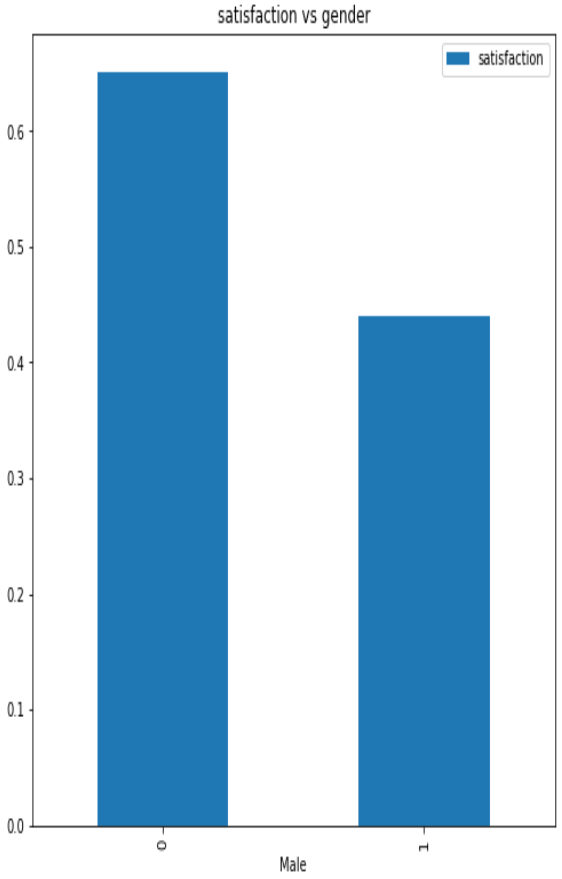
Most dissatisfied customers had an average delay of 18 minutes (approx) and satisfied customers had an average delay of 12 minutes (approx)with the airlines.

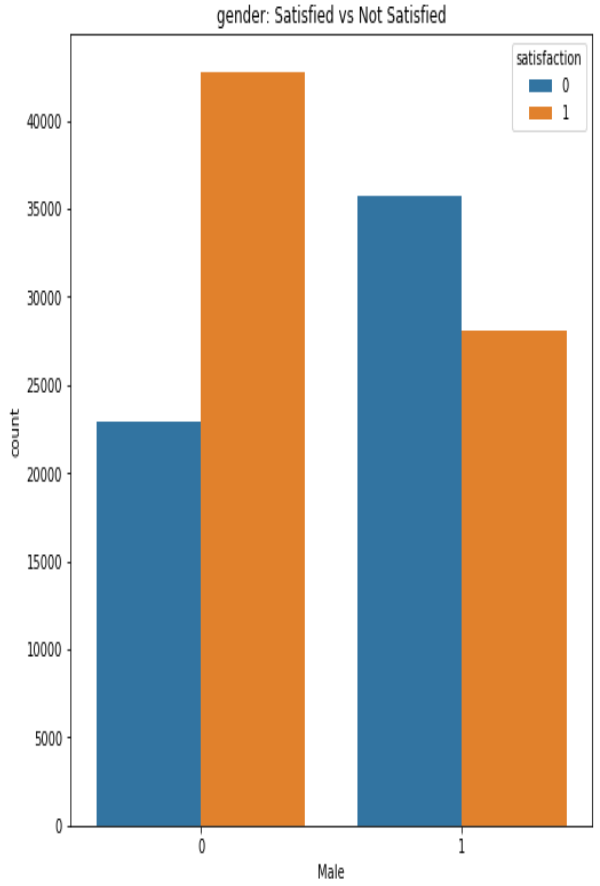


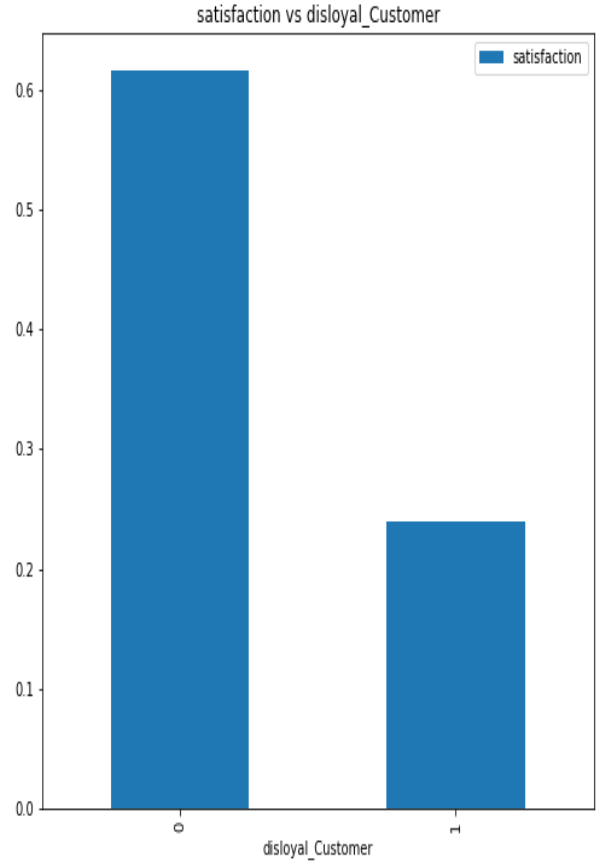
EDA (Exploratory Data Analysis)

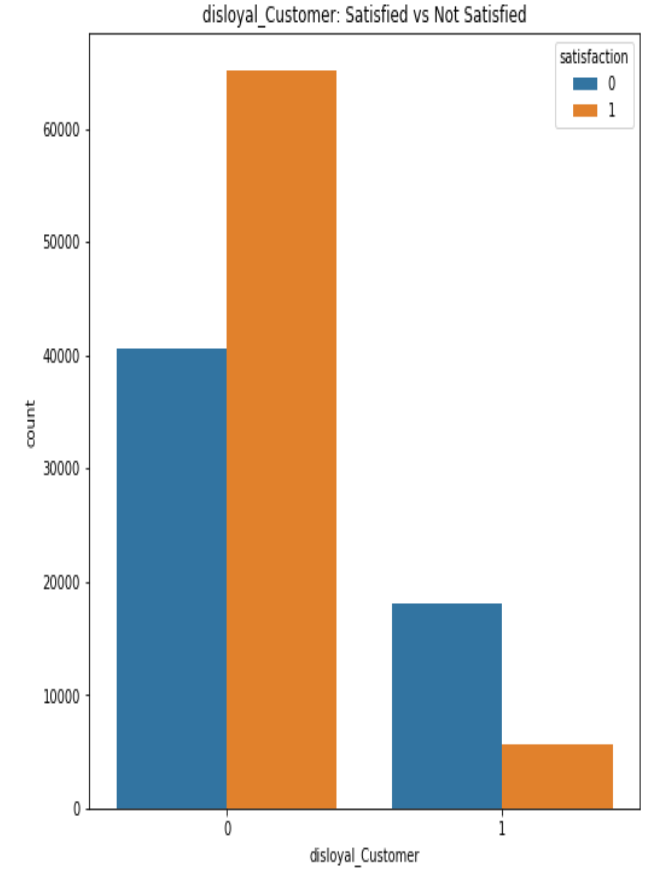
Females are more satisfied than males. Males are more dissatisfied with the airlines.

Disloyal customers of the airlines are less satisfied compared to loyal customer.



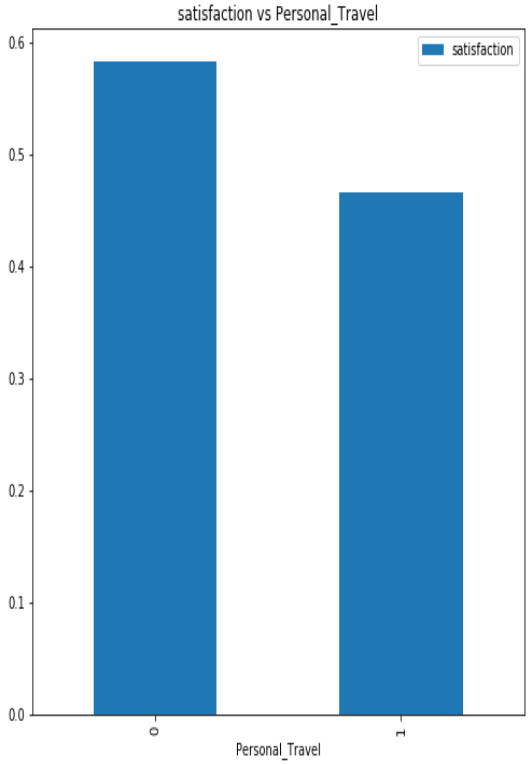


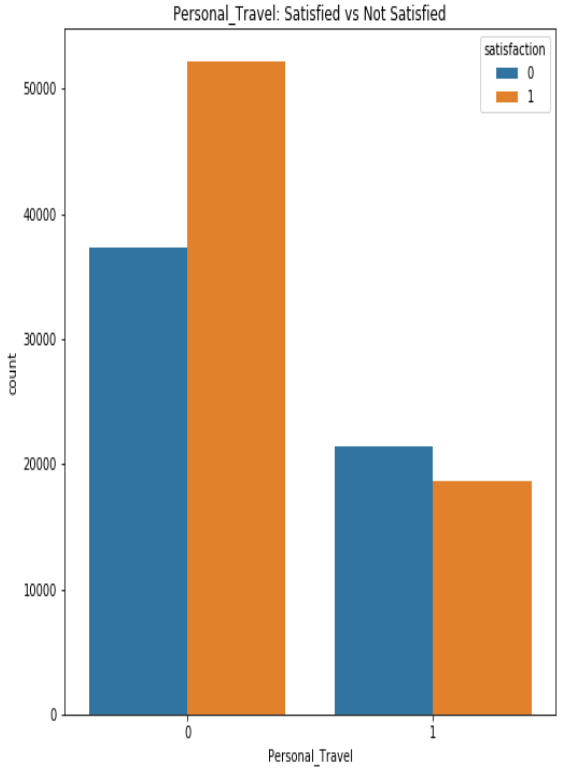




EDA (Exploratory Data Analysis)

Customers doing business travel are more satisfied compared to personal travel customers





A cleanliness level of 4 and 5 yields higher satisfaction among customers.

satisfaction	0	1	All
Cleanliness			
0	5	0	5
1	4624	3122	7746
2	7953	5408	13361
3	16307	7600	23907
4	20110	28555	48665
5	9606	26197	35803
All	58605	70882	129487

EDA (Exploratory Data Analysis)

A seat comfort level of 4 and 5 yields higher satisfaction among customers.

satisfaction	0	1	All
Seat_comfort			
0	10	4771	4781
1	11466	9416	20882
2	18396	10249	28645
3	18734	10362	29096
4	9858	18457	28315
5	141	17627	17768
All	58605	70882	129487

An Inflight_entertainment level of 4 and 5 yields higher satisfaction among customers.

satisfaction	0	1	All
Inflight_entertainment			
0	1007	1961	2968
1	9289	2479	11768
2	15861	3257	19118
3	19326	4807	24133
4	11696	30056	41752
5	1426	28322	29748
All	58605	70882	129487

EDA (Exploratory Data Analysis)

A better ease_of_online_booking service will yield more satisfaction among customers as suggested by the ratings of 4 and 5 for ease_of_Online_booking service.

satisfaction	0	1	All
Ease_of_Online_booking			
0	18	0	18
1	10815	2582	13397
2	14192	5695	19887
3	14366	7978	22344
4	11233	28574	39807
5	7981	26053	34034
All	58605	70882	129487

Inflight wifi service doesn't hold much importance for satisfaction because we can see that even if inflight wifi service has a got a rating of 2, the number of customers satisfied is more than no of customers dissatisfied.

satisfaction	0	1	All
Inflight_wifi_service			
0	72	58	130
1	10731	3939	14670
2	13415	13542	26957
3	13496	14022	27518
4	11382	20092	31474
5	9509	19229	28738
All	58605	70882	129487

EDA (Exploratory Data Analysis)

Better onboard service will yield better satisfaction among customers as suggested by the ratings of 4 and 5 for onboard service.

satisfaction	0	1	All
On_board_service			
0	5	0	5
1	9708	3515	13223
2	11283	5834	17117
3	15898	11061	26959
4	14301	26257	40558
5	7410	24215	31625
All	58605	70882	129487

Better leg room service will yield better satisfaction among customers as suggested by the ratings of 4 and 5 for leg room service.

satisfaction	0	1	All
Leg_room_service			
0	136	306	442
1	7953	3145	11098
2	13522	8161	21683
3	14071	8326	22397
4	12930	26653	39583
5	9993	24291	34284
All	58605	70882	129487

EDA (Exploratory Data Analysis)

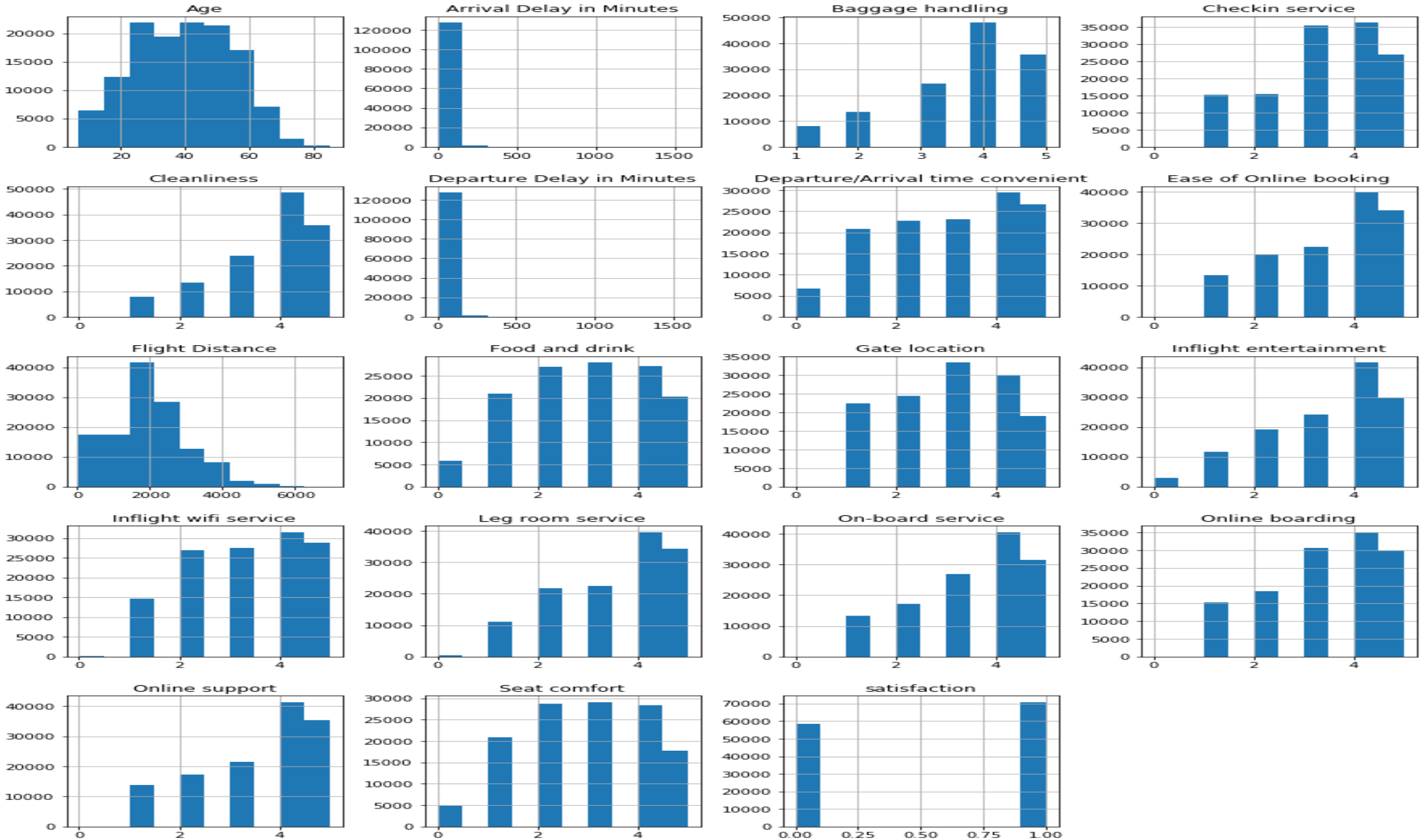
- Better food and drink will yield better satisfaction among customers as suggested by the ratings of 4 and 5 for Food and drink.

satisfaction	0	1	All
Food_and_drink			
0	1305	4617	5922
1	10326	10682	21008
2	15361	11717	27078
3	16036	12029	28065
4	11116	16013	27129
5	4461	15824	20285
All	58605	70882	129487

Data Pre-processing

- In the data preprocessing section, I have prepared the data where dummies have been created for the categorical data.
- I split the dataset into training and test dataset in the ratio of 80:20.
- Standardization have been done for only the numerical attributes.

❖*The next slide shows the distribution of the data before standardization*


Age
Arrival Delay in Minutes
Baggage handling
Checkin service
Cleanliness
Departure Delay in Minutes
Departure/Arrival time convenient
Ease of Online booking
Flight Distance
Food and drink
Gate location
Inflight entertainment
Inflight wifi service
Leg room service
On-board service
Online boarding
Online support
Seat comfort
satisfaction

Modelling

In the modelling section, to predict whether a future customer would be satisfied with the Airlines service, I fit a Logistic Regression model , a Gradient Boost Classifier model and a Decision Tree model to check which is a better model to perform the task.

To evaluate a better model, I look at the accuracy, AUC scores for both training and test datasets for all the algorithms.

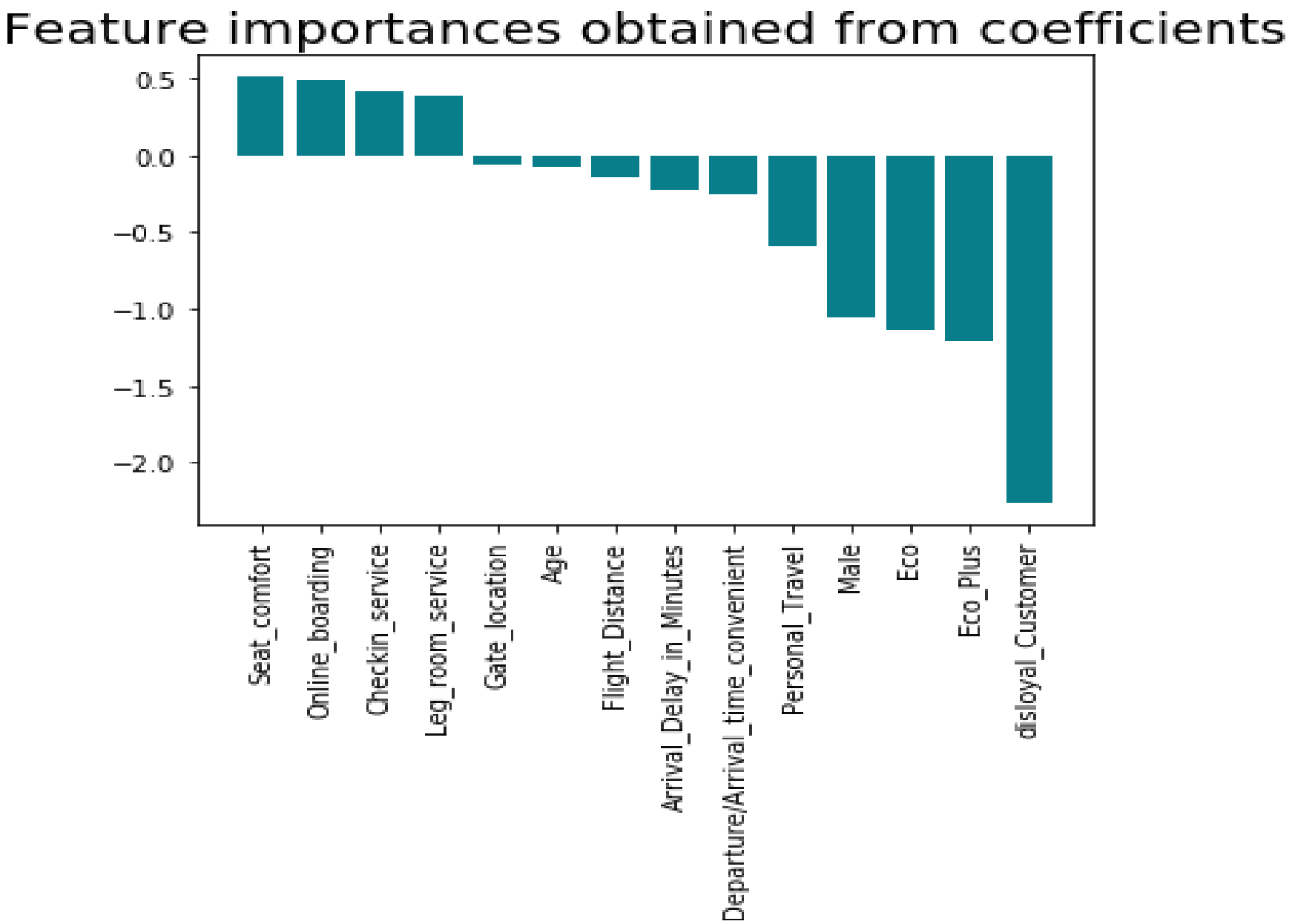
I also look at VIF (Variance Influence Factor) to look for any multicollinear variables. Any explanatory variable having a VIF>10 is a indication of multicollinearity present in the model. The variables having a VIF>10 were dropped and the model was re-estimated again. Hence, multicollinearity issues have been taken care of before finalzing the model.

Logistic Regression

Generalized Linear Model Regression Results

Dep. Variable:	satisfaction	No. Observations:	103589
Model:	GLM	Df Residuals:	103575
Model Family:	Binomial	Df Model:	13
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-47668.
Date:	Fri, 06 Jan 2023	Deviance:	95336.
Time:	16:29:22	Pearson chi2:	1.18e+05
No. Iterations:	5		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-0.0943	0.009	-10.874	0.000	-0.111	-0.077
Flight_Distance	-0.1423	0.009	-16.100	0.000	-0.160	-0.125
Arrival_Delay_in_Minutes	-0.2062	0.009	-23.527	0.000	-0.223	-0.189
Seat_comfort	0.4537	0.007	66.297	0.000	0.440	0.467
Departure/Arrival_time_convenient	-0.2794	0.007	-42.035	0.000	-0.292	-0.266
Gate_location	-0.2086	0.007	-29.366	0.000	-0.223	-0.195
Leg_room_service	0.1999	0.006	34.584	0.000	0.189	0.211
Checkin_service	0.2326	0.006	37.384	0.000	0.220	0.245
Online_boarding	0.3023	0.006	50.821	0.000	0.291	0.314
Eco	-1.4370	0.021	-68.314	0.000	-1.478	-1.396
Eco_Plus	-1.5837	0.032	-49.271	0.000	-1.647	-1.521
disloyal_Customer	-2.3923	0.025	-95.741	0.000	-2.441	-2.343
Personal_Travel	-0.4737	0.023	-20.583	0.000	-0.519	-0.429
Male	-1.2875	0.016	-78.765	0.000	-1.319	-1.255


Feature importances obtained from coefficients
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Seat_comfort
Online_boarding
Checkin_service
Leg_room_service
Gate_location
Age
Flight_Distance
Arrival_Delay_in_Minutes
Departure/Arrival_time_convenient
Personal_Travel
Male
Eco
Eco_Plus
disloyal_Customer

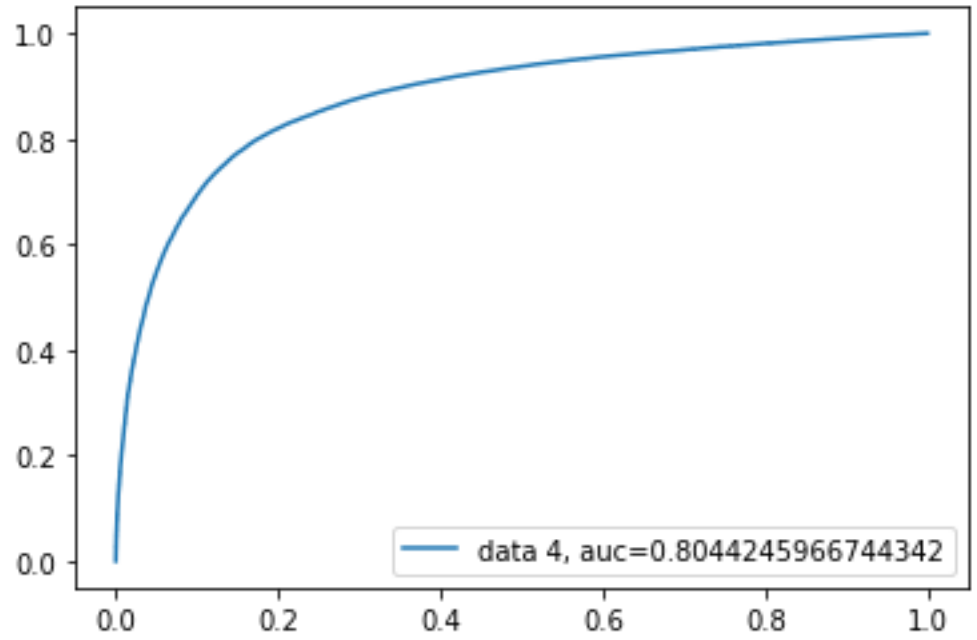
Logistic Regression Results

Some key findings that most likely explains the relationship between each of the independent and dependent variable.

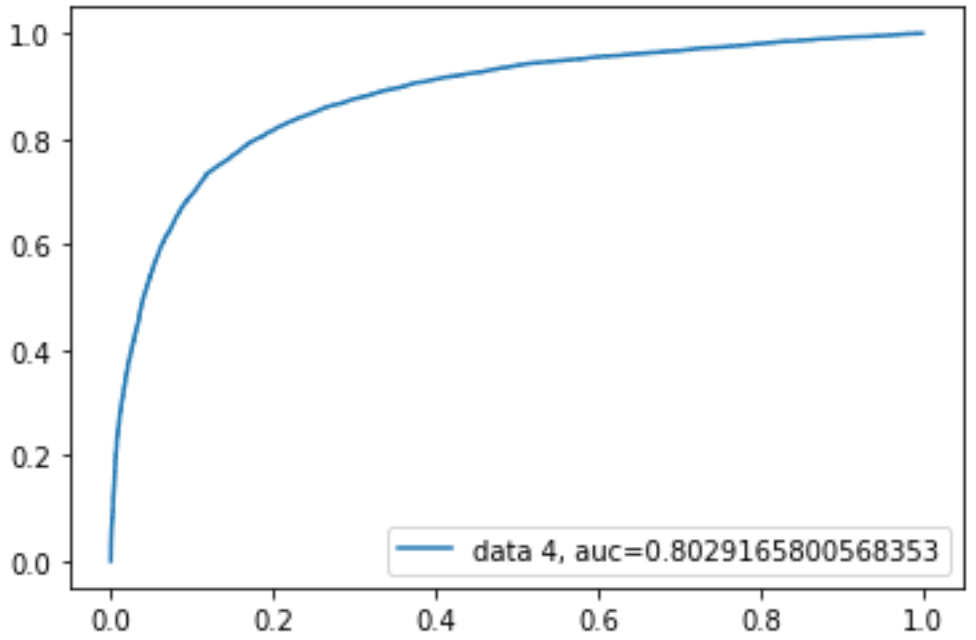
- In the Logistic Regression output, we find all our explanatory variables are highly significant.
- The negative coefficient signs of the features namely, more disloyal customer, Eco, Ecoplus, male customers, personal travel, customers travelling longer flight distance are more likely to generate dissatisfied customers for the Airlines.
- The positive coefficient signs suggest that better seat comfort, Inflight_entertainment, On_board_service, Leg_room_service, Checkin_service and Online_boarding are more likely to generate more satisfied customers for the Airlines.

Logistic Regression (AUC scores)

Training



Test



Performance Evaluation (Logistic Regression)

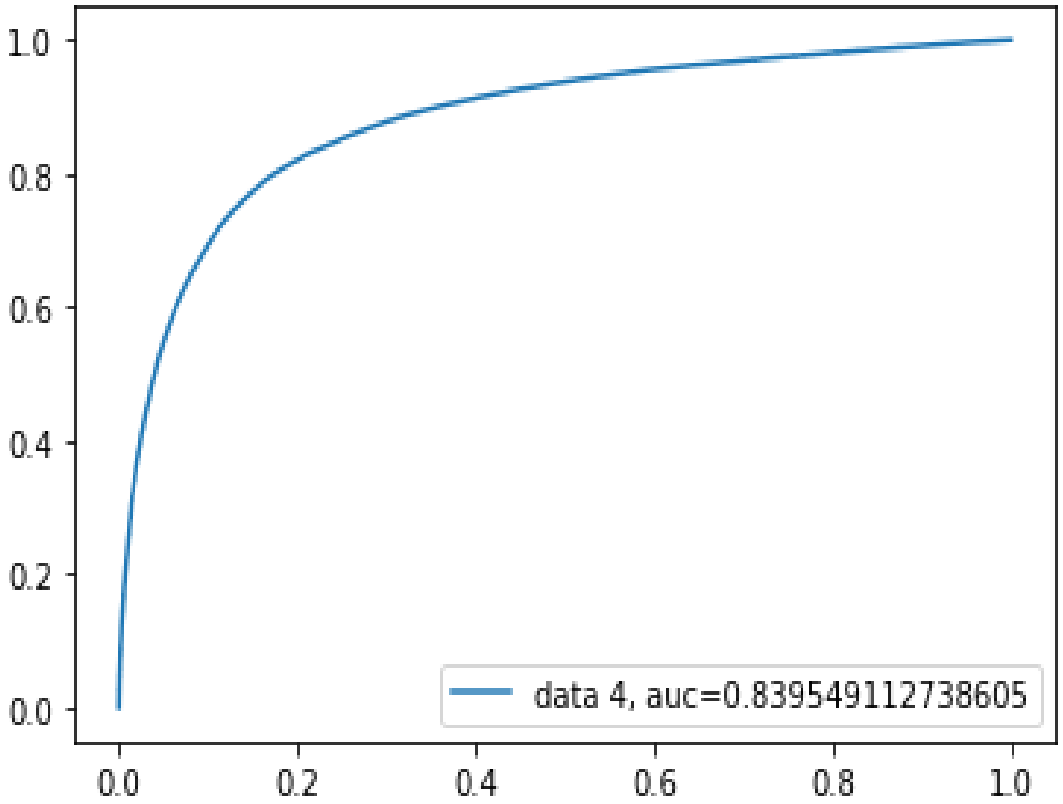
To evaluate the model, I look at the accuracy score and AUC scores for both training and test datasets.

The AUC score for Logistic Regression model is 0.803 approx. for both training and test data sets. The accuracy score is 0.81 (approx.)

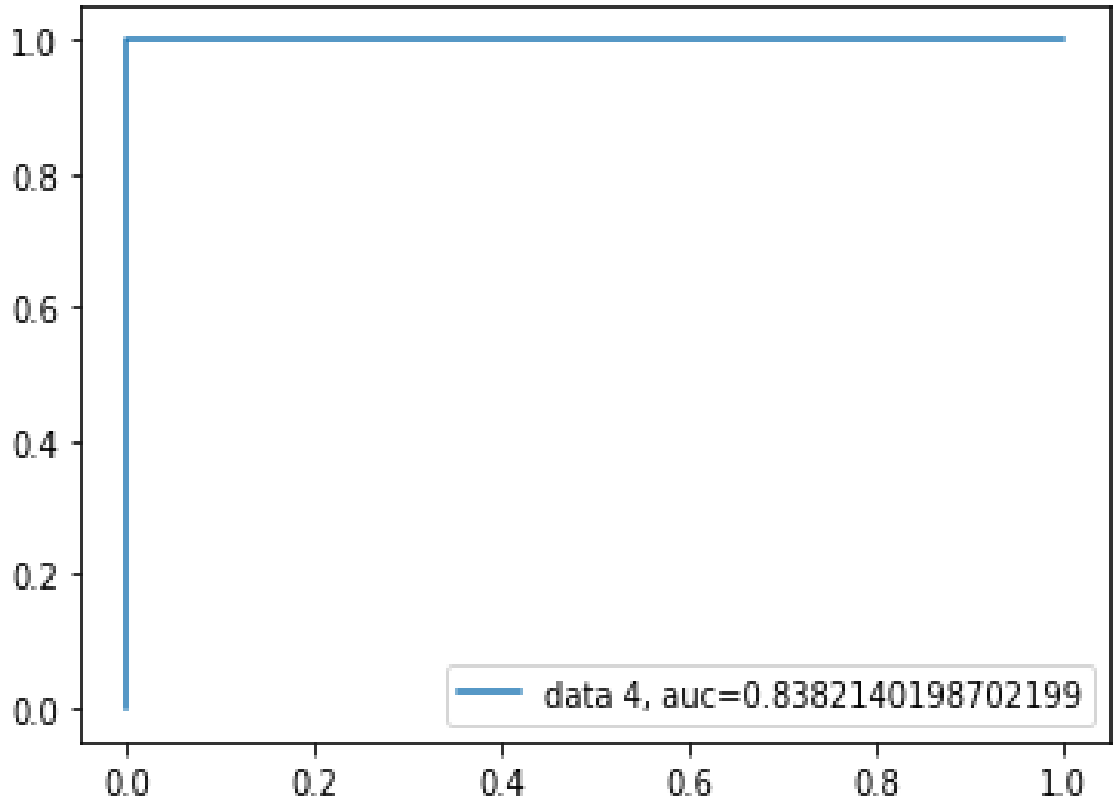
To achieve a better AUC score, I decided to refit the model using Decision Tree and Gradient Boost Classifier and check the model performance again.

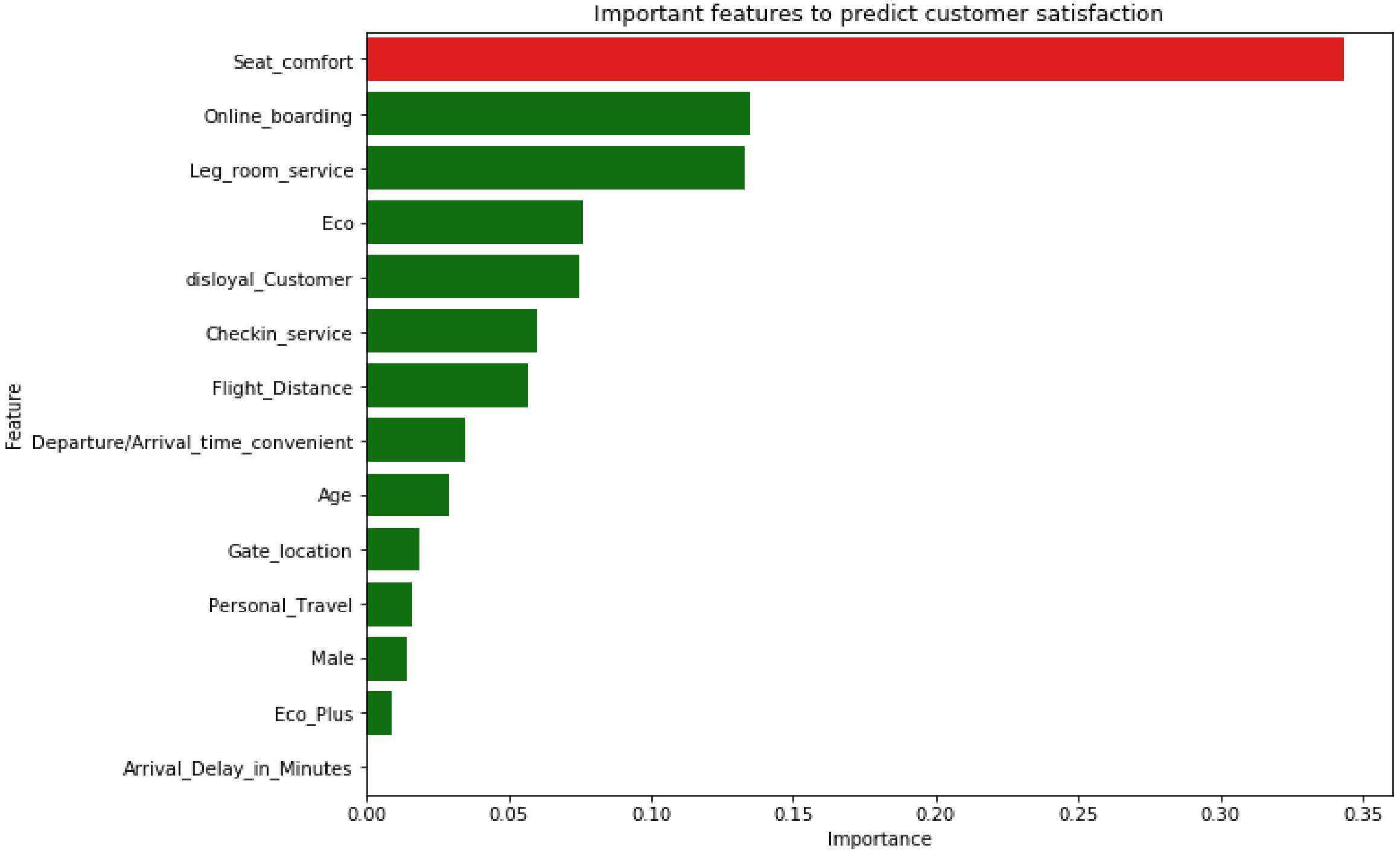
Decision Tree Model (AUC scores)

Training



Test




Important features to predict customer satisfaction
Seat_comfort
Online_boarding
Leg_room_service
Eco
disloyal_Customer
Checkin_service
Flight_Distance
Departure/Arrival_time_convenient
Age
Gate_location
Personal_Travel
Male
Eco_Plus
Arrival_Delay_in_Minutes
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance

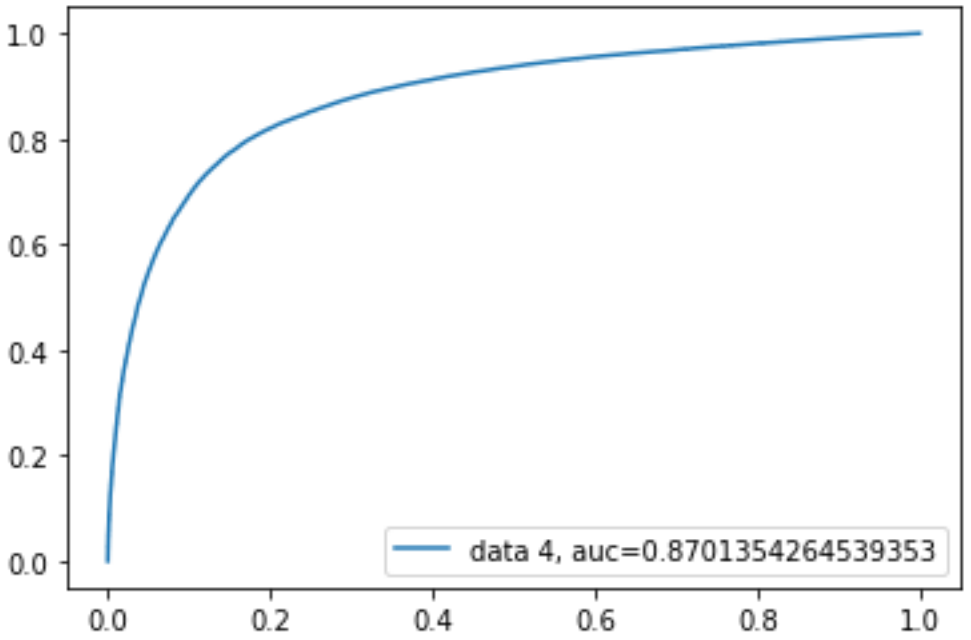
Performance Evaluation (Decision Tree model)

To evaluate the model, I look at the accuracy score and AUC scores for both training and test datasets.

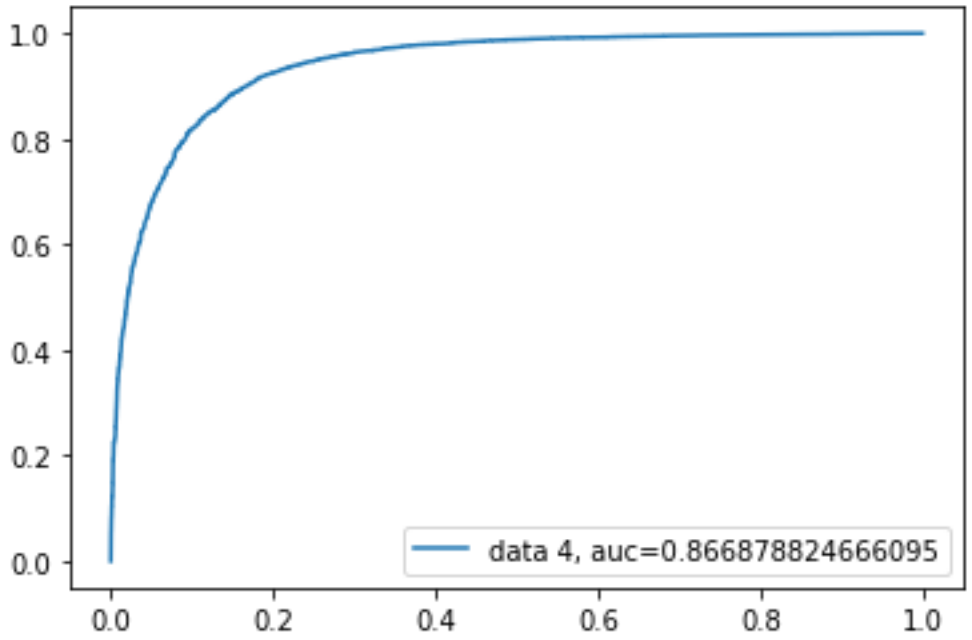
The AUC score for the model is 0.84 approx. for training data and test data set. The accuracy score is 0.84 (approx.)

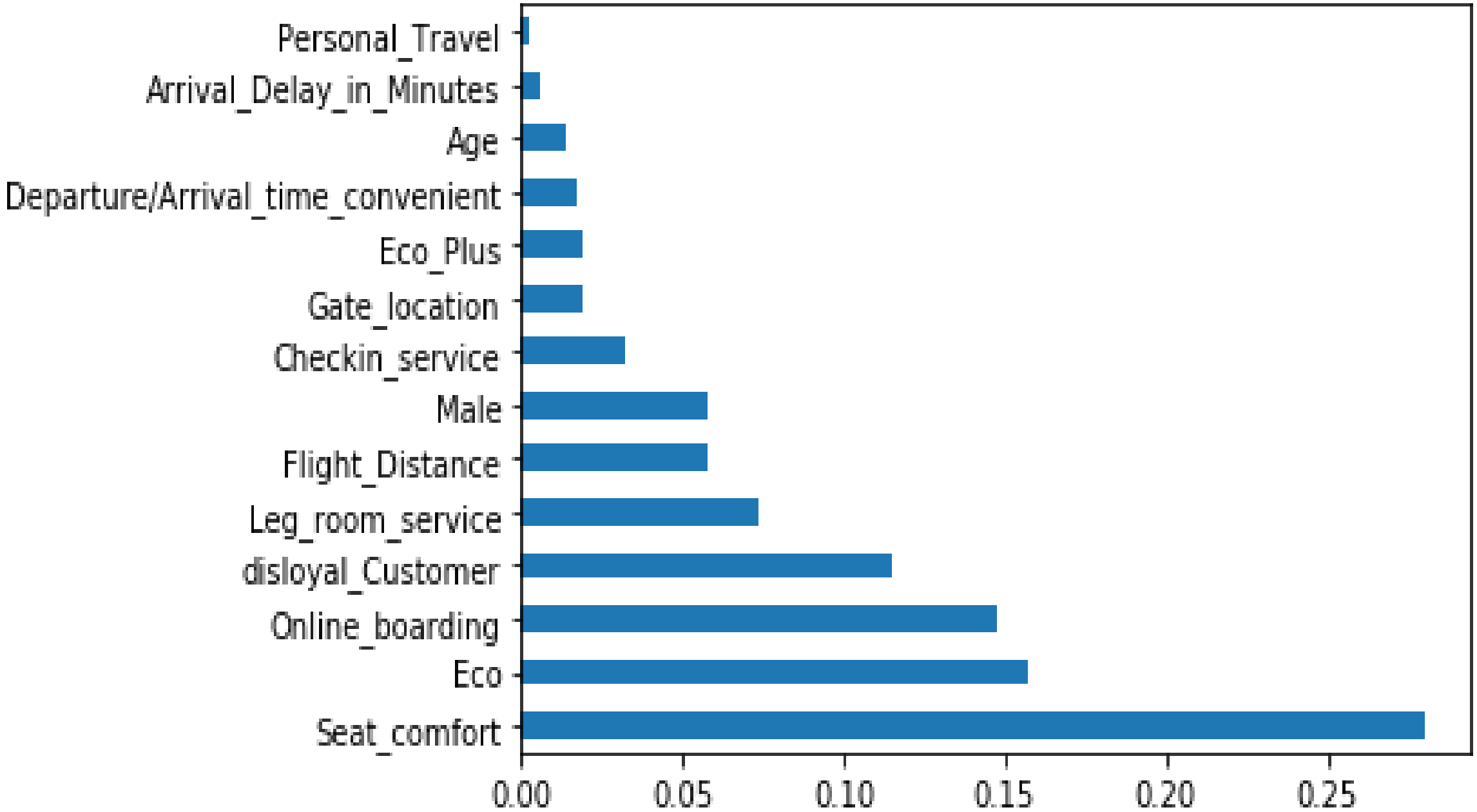
Gradient Boost Classifier (AUC scores)

Training



Test




Personal_Travel
Arrival_Delay_in_Minutes
Age
Departure/Arrival_time_convenient
Eco_Plus
Gate_location
Checkin_service
Male
Flight_Distance
Leg_room_service
disloyal_Customer
Online_boarding
Eco
Seat_comfort
0.00
0.05
0.10
0.15
0.20
0.25

Performance Evaluation (Gradient Boost Classifier)

To evaluate the Gradient Boost Classifier model, I look at the accuracy score and AUC scores.

The AUC score for the model is 0.87 approx. for both training and test data sets. The accuracy score is 0.87 (approx.)

Modelling

Looking at the accuracy score and AUC scores for all the models, we can conclude that Gradient Boost Classifier is a better model to analyze the given task.

Both AUC scores of training and test dataset are 0.87 (approx.) for Gradient Boost Classifier suggesting there is no form of overfitting in the model.

Conclusion

- The dataset contributed to the predictive power of the model.
- Out of the different ML models, Gradient Boost Classifier provided the best results.
- With 80%-20% splitting, for both the training and test data AUC is 0.87 (approx.) with the Gradient Boost Classifier algorithm.
- The five most important features that plays an important role in predicting customer satisfaction are: Seat comfort, Eco, Online Boarding, Disloyal customer and Leg room service.
- With more ideas, the model can be improved in the future.
- In the future I would like to spend more time gathering some additional data, if available, and also trying out different other classification algorithms as further robustness checks.

THANK YOU!