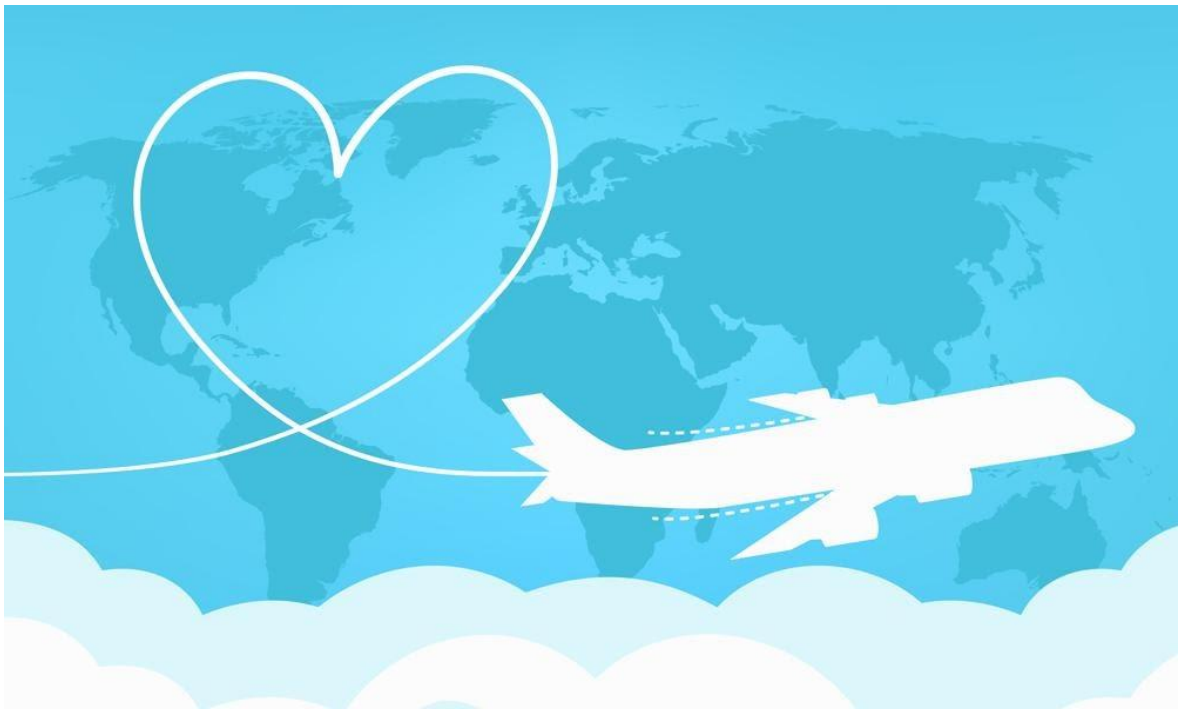
Duc Thanh Nguyen

REPORT

FLIGHT SATISFACTION

December 27, 2021



I. Introduction

This report shows the analysis results on the dataset of flight satisfaction. Dataset has dimensions (129880,23), in which there are 129880 rows and 23 columns. The target value of this dataset is the "Satisfaction" column with two classes: Satisfied and Neural or Dissatisfied. There are 21 variables with different data types include objects, integers, and floats. In particular, 15 variables are integers, but they are categorical variables with six classes on a scale of 0 to 5 since the data is collected from the survey. The report includes four main parts: Data Pre-processing, Data Exploration, Data Modelling, and Further Analysis.

0	satisfaction_v2	129880 non-null	object
1	Gender	129880 non-null	object
2	Customer Type	129880 non-null	object
3	Age	129880 non-null	int64
4	Type of Travel	129880 non-null	object
5	Class	129880 non-null	object
6	Flight Distance	129880 non-null	int64
7	Inflight wifi service	129880 non-null	int64
8	Departure/Arrival time convenient	129880 non-null	int64
9	Ease of Online booking	129880 non-null	int64
10	Gate location	129880 non-null	int64
11	Food and drink	129880 non-null	int64
12	Online boarding	129880 non-null	int64
13	Seat comfort	129880 non-null	int64
14	Inflight entertainment	129880 non-null	int64
15	On-board service	129880 non-null	int64
16	Leg room service	129880 non-null	int64
17	Baggage handling	129880 non-null	int64
18	Checkin service	129880 non-null	int64
19	Inflight service	129880 non-null	int64
20	Cleanliness	129880 non-null	int64
21	Departure Delay in Minutes	129880 non-null	int64
22	Arrival Delay in Minutes	129487 non-null	float64

Figure 1: Summary of Dataset

II. Data Preprocessing

1. Handling missing data

   All dataset columns are cleaned without missing data except the column "Arrival Delay in Minutes". The number of missing values is 393, accounting for only 0.3% of the dataset. The rows, which have missing data are dropped out of the dataset

#	Column	Non-Null Count	Dtype
0	satisfaction_v2	129880 non-null	object
1	Gender	129880 non-null	object
2	Customer Type	129880 non-null	object
3	Age	129880 non-null	int64
4	Type of Travel	129880 non-null	object
5	Class	129880 non-null	object
6	Flight Distance	129880 non-null	int64
7	Inflight wifi service	129880 non-null	int64
8	Departure/Arrival time convenient	129880 non-null	int64
9	Ease of Online booking	129880 non-null	int64
10	Gate location	129880 non-null	int64
11	Food and drink	129880 non-null	int64
12	Online boarding	129880 non-null	int64
13	Seat comfort	129880 non-null	int64
14	Inflight entertainment	129880 non-null	int64
15	On-board service	129880 non-null	int64
16	Leg room service	129880 non-null	int64
17	Baggage handling	129880 non-null	int64
18	Checkin service	129880 non-null	int64
19	Inflight service	129880 non-null	int64
20	Cleanliness	129880 non-null	int64
21	Departure Delay in Minutes	129880 non-null	int64
22	Arrival Delay in Minutes	129487 non-null	float64

Figure 2: Checking for missing data

2. Checking data balancing
   Since the target value has two classes, the primary type of modeling is classification, so it is essential to check for balancing between 2 classes. As a result, the dataset is balanced; Satisfied class accounts for 43.45% and Neural or Dissatisfied class take 56.55%. This dataset is already balanced, which is good for classification, but it is not good in the area of Customer Service since the number of participants who joined this survey that are Neural or Dissatisfied about the Airline was dominated.

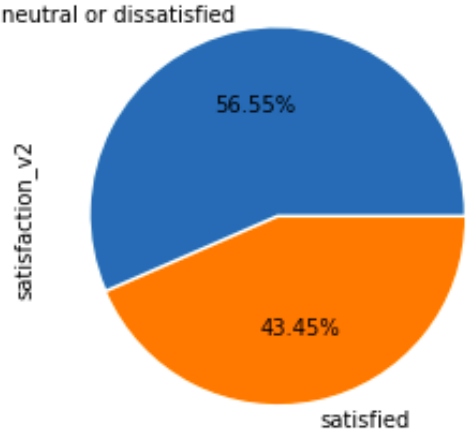


Figure 3: Checking for balancing

III. Exploration Data Analysis

1. Object type variables

There are 4 variables that are object data type: Gender, Customer Types, Class, Type of travels.

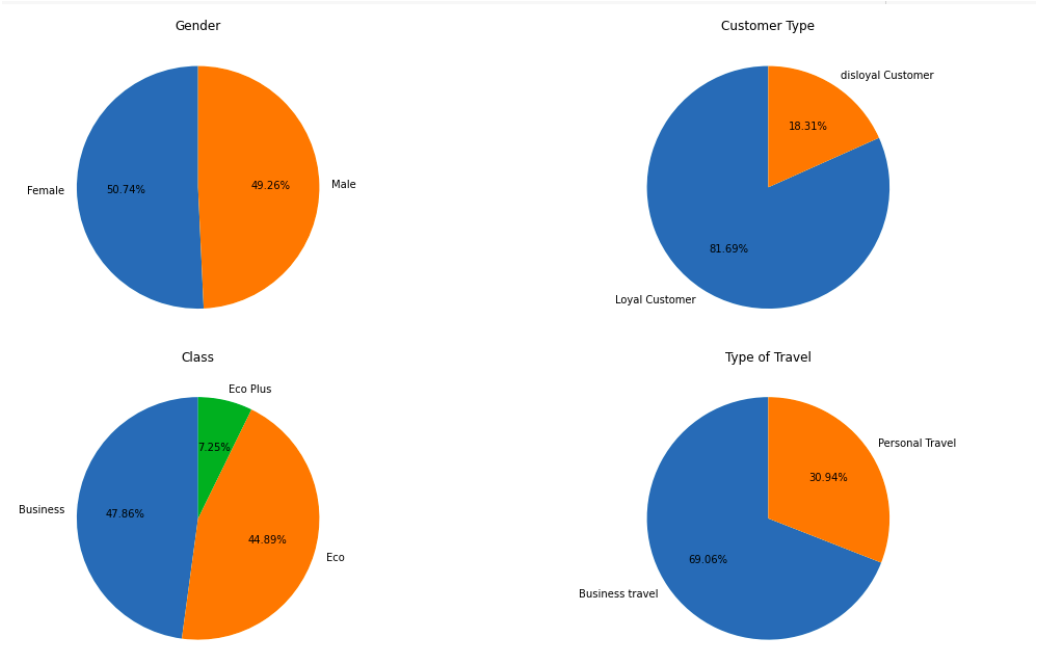


Figure 4: Object data variables

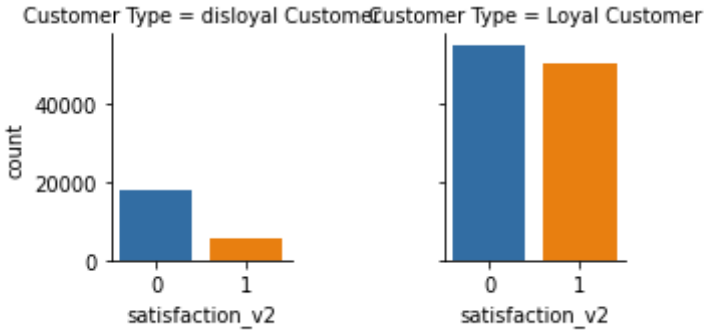
As can be seen, the chart of Gender shows that the data also balance between male and female, avoiding bias from a specific gender. The data is also balanced between Business class and Economy class in Class's data. Business travel and Loyal Customer account for a large amount of number of participants. All variables have 2 classes except variable "Class", which has Economy Plus and Business and Economic, but only took 7.25%.

2. Object type variables vs target value

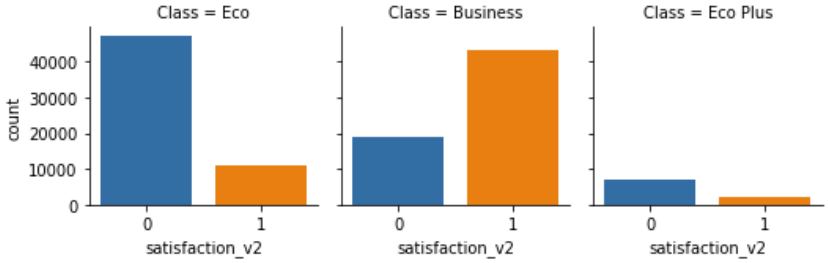
   a. **Gender**: It is observed that the gender-wise distribution of dissatisfied and satisfied customers is quite similar. For both male and female passengers, no. of unhappy customers are side compared to no. of satisfied customers.

   b. **Customer Type**: Loyal passengers are very high in number. Even among loyal passengers, satisfied and dissatisfied ones are almost 49:51.

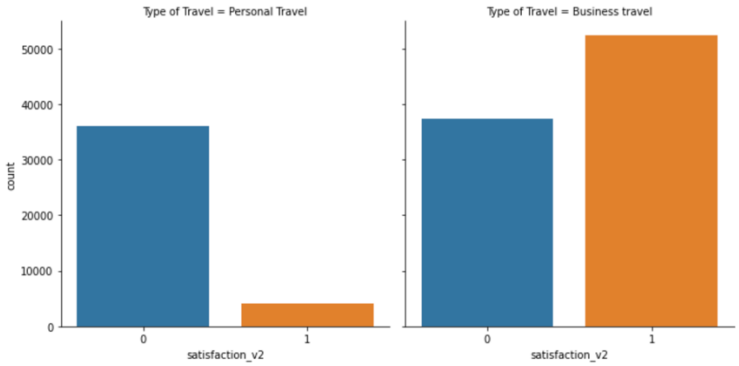
Although we have a lot of Loyal passengers, the disappointed amount seems to be high; there must be something that makes that customer come back.



c. **Class**: 3 classes of customer: Eco, Business, and Eco Plus. For Eco and Business class: a huge gap between the satisfaction and dissatisfaction in each class, the number of participants who are not happy about the service is much higher.



d. **Type of Travel**: The ratio of satisfied and dissatisfied of Personal travel is quite imbalanced, most of it are dissatisfied. On the other sides, Business travel give more positive feedback although they are not spending their own money or price is the driven decision factor



3. Corellation matrix

There are many variables that has low correlation with target value “Satisfaction”, in which the change of them has no significant impact on the outcome.

- No linear correlation variables: Departure/Arrival time convenient, Gate location, Departure delay in minutes, Arrival delay in Minutes
- Redundant variables: Arrival Delay in minutes→ dropped.

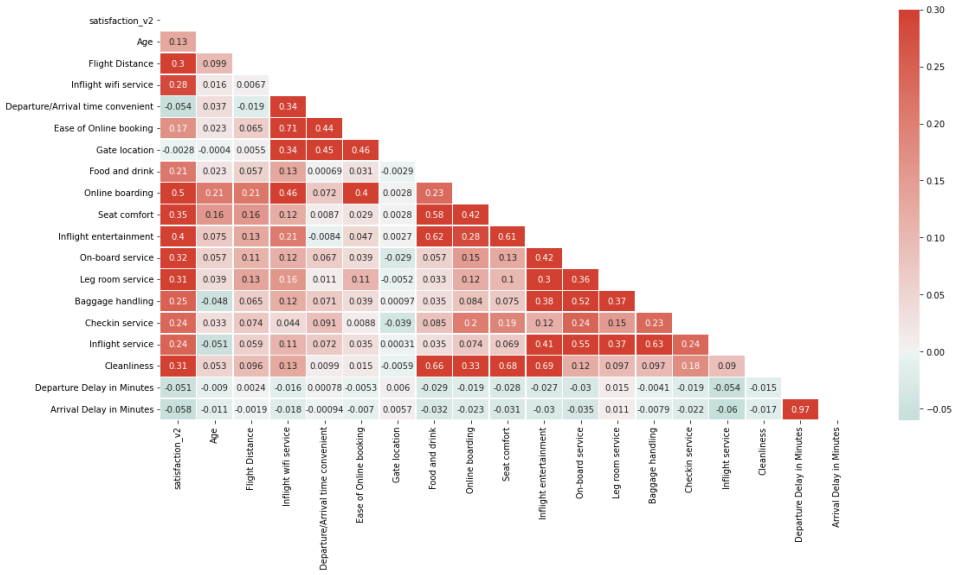


Figure 5: Correlation Matrix

4. Low correlation variables

   There is no interest pattern in the data of Departure/Arrival time convenient and Gate location, no participant gave zero score on Gate location. The distribution of Departure delay in Minutes is skewed to the left, showing that most people are being delayed for less than 100 minutes

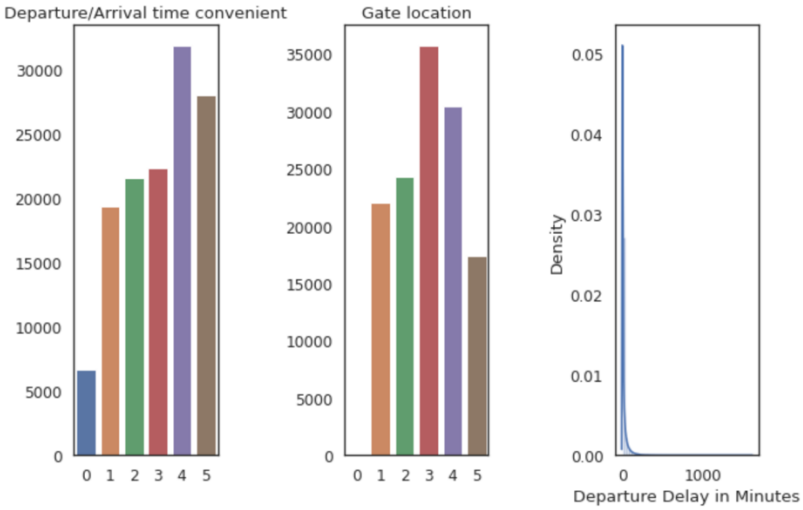
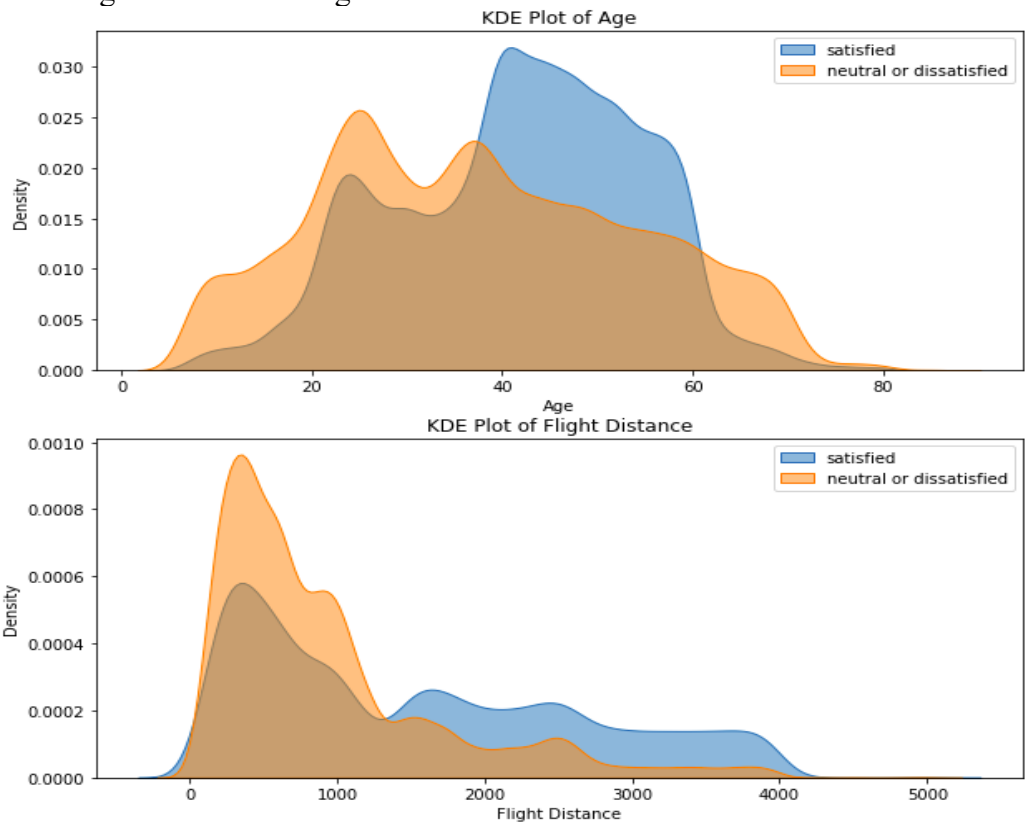


Figure 6: Low correlation variables

5. Other variables that are not in scale 0 to 5

There are to numeric variables in the dataset that is not in score type: Age, Flight Distance

- Older customer seems easier on the review than the youth.
- Interesting point is that long flight customer satisfied more than short flight. Most of the bad review comes from flight distance less than 1500km. We could divide flight distance into short and long-distance flight instead of using the number.



IV. Modelling

1. Data Preparation for Modelling

There are many steps has performed as listed below in order to prepare for classification. As a results, all data are in type of integer(int64).

- Drop low correlation variables: Arrival Delay in Minutes
- Encoding object type data: Customer Type, Gender, Type of Travel, Class
- Covert Target Value to Binary: Satisfied =1, Dissatisfied =0
  → As a result, all data types are int64

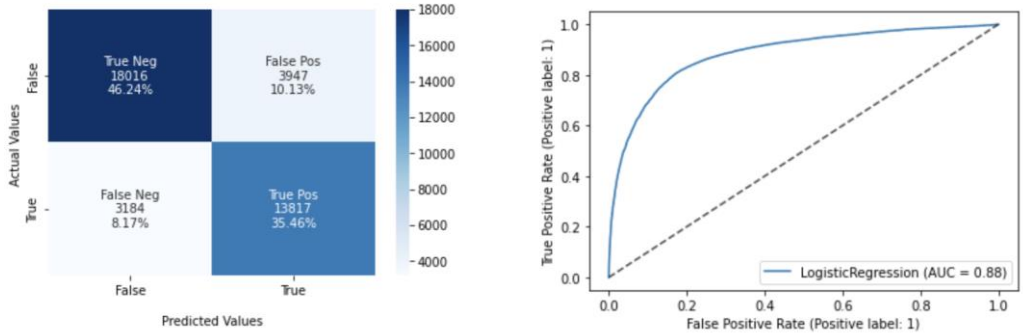
- Drop the "id" column → unuseful information
- Spit Data with the ratio: 70/30
- Train Set = (90916, 22)
- Test Set = (38964, 22)

```
0   index                                129880 non-null  int64
1   satisfaction_v2                      129880 non-null  int64
2   Gender                               129880 non-null  int64
3   Customer Type                        129880 non-null  int64
4   Age                                  129880 non-null  int64
5   Type of Travel                       129880 non-null  int64
6   Class                                129880 non-null  int64
7   Flight Distance                      129880 non-null  int64
8   Inflight wifi service                129880 non-null  int64
9   Departure/Arrival time convenient    129880 non-null  int64
10  Ease of Online booking               129880 non-null  int64
11  Gate location                        129880 non-null  int64
12  Food and drink                       129880 non-null  int64
13  Online boarding                      129880 non-null  int64
14  Seat comfort                         129880 non-null  int64
15  Inflight entertainment               129880 non-null  int64
16  On-board service                     129880 non-null  int64
17  Leg room service                     129880 non-null  int64
18  Baggage handling                     129880 non-null  int64
19  Checkin service                      129880 non-null  int64
20  Inflight service                     129880 non-null  int64
21  Cleanliness                          129880 non-null  int64
22  Departure Delay in Minutes           129880 non-null  int64
```

Figure 7: Clean data for Classification

2. Logistic Regression

   Because the target value "Satisfaction" has 2 classes, which can be converted into binary classes, logistic regression is the most popular model for binary classes. The advantage of logistic regression is that the model is regression model that we can obtain the coefficients of each variable. Thus, we will know how variables impact on the target values. The model's accuracy is 0. 8169, there is 10.13% of Positive class (Class 1) and 8.17% of Negative Class (Class 0) is misclassified. Based on the ROC curve on the right, the optimal point is 0.8 for True positive rate. Beyond optimal point, the false positive rate starts to increase.



3. Decision Tree

   The next model in this analysis is decision tree. The performance is this model is very accurate with 94,5% accuracy. 2,63% of Positive class (Class 1) and 2.36% of Negative Class (Class 0) is misclassified. However, there are 23 variables and variables have 6 classes from score 0 to 5, the model split into different groups of scores that has no specific name on that node. By the reasons, it is hard to interpret the results by observation.

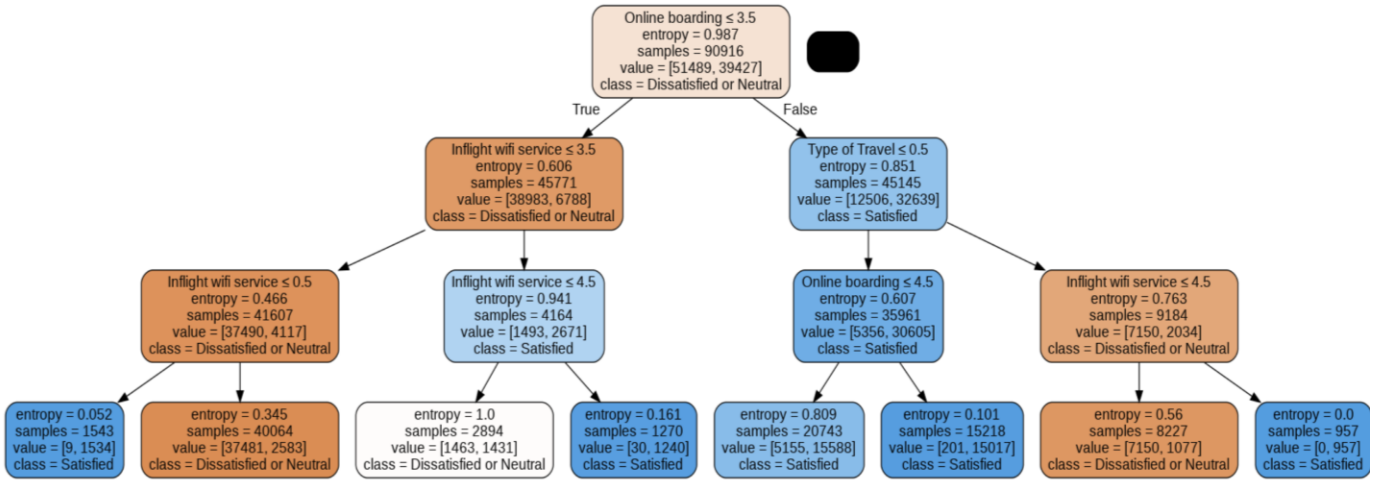


Figure 7: Decision Tree

Since the tree can not give tangible results, the important variables that significantly impact the airline service are chosen by their impurity in the decision tree by calculating the Gini index. Variable with higher Gini index is more important, since the prediction of target values in on the brands of the tree that has that variable has smaller number of misclassified data compare to actual values.

- Important variables for decision tree
  As a result, Online boarding is the most important variable with the highest impurity. In other words, if a participant gives 5 scores for online boarding, the actual outcome is "satisfied", a very percent that the prediction is also "satisfied"—the top 3 most important variables: Online boarding, Inflight wifi service, Type of Travel.

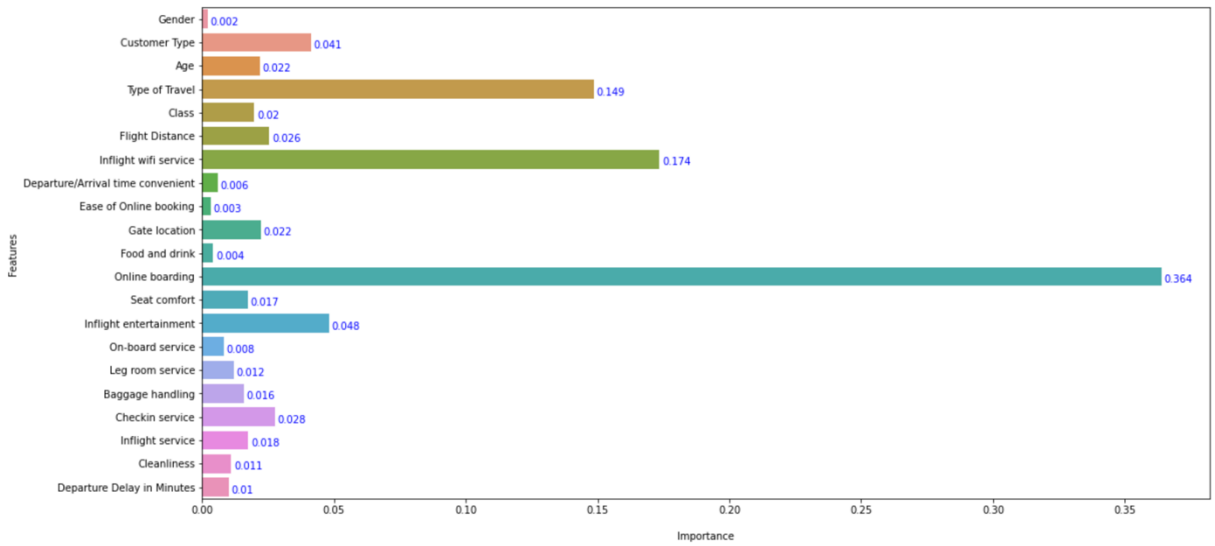


Figure 8: Importance of variables in Decision Tree

4. Random Forest
   Random forest will generate a combination of decision tree to limit overfitting and reduce bias. Random Forest is an ensemble method of decision tree, aims to optimize the performance of decision tree. However, the model's accuracy is 94.5% which is

the same with the Decision Tree. Since the final model is still a tree, we can obtain the important variable by Gini index. The most important variable is Online Boarding and the top 3: Online Boarding, Inflight Wifi Service and Type of Travel.

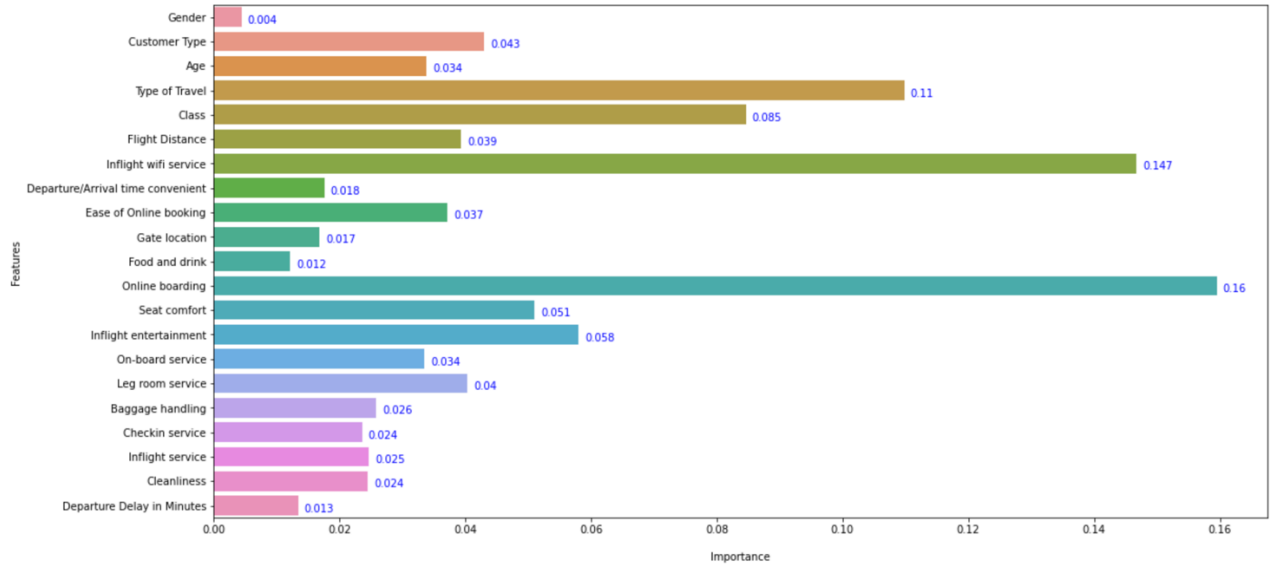


Figure 8: Importance of variables in Random Forest

5. Adaptive Boost

   Adaptive Boost (Ada boost) is also an ensemble method of decision tree, it is sequential ensemble method, weight of misclassified will be increase in the next tree. The model's accuracy is 92.5%, which is lower than Decision Tree. Because this is an ensemble method, final model is also a tree, thus we can obtain the importance of variables.

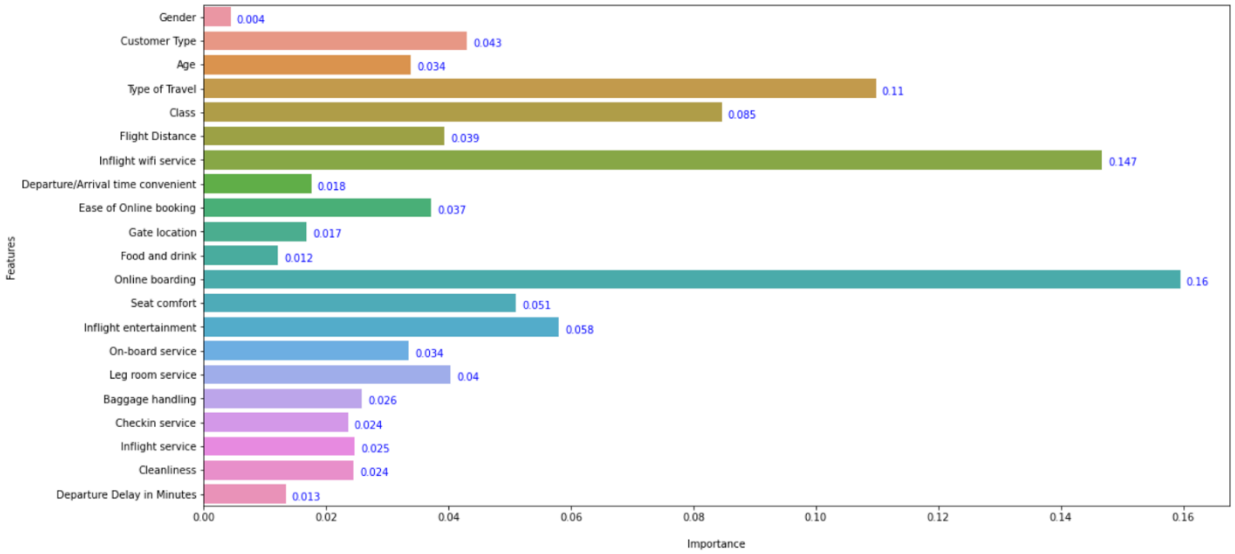
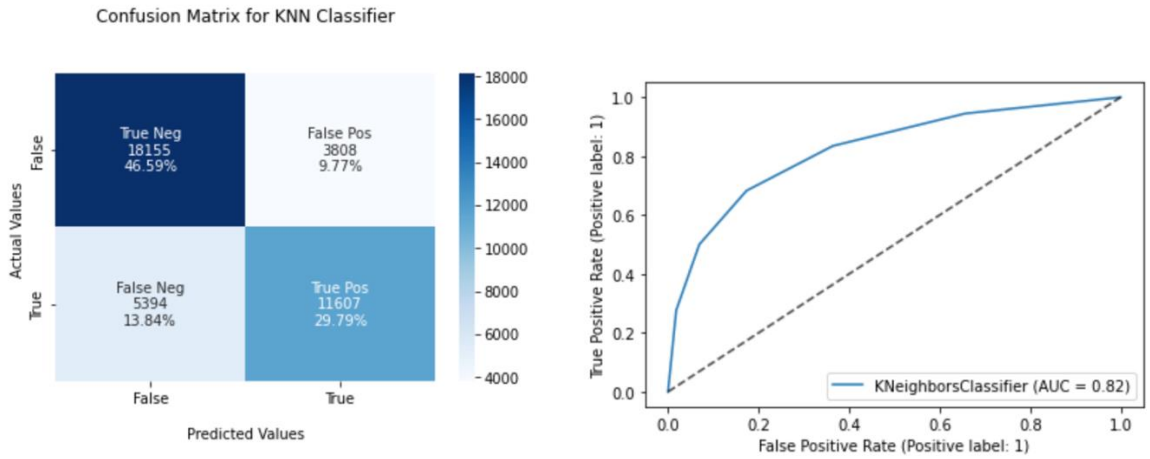


Figure 9: Importance of variables in Ada boost

6. KNN

The next model of classification for this analysis is KNN. The accuracy of the model is 76,38%. There are 9.77% of positive class and 13.84% of negative class is misclassified. However, beside knowing the performance of the model in order to predict the outcome of target value, there is no further useful information as a decision tree. The purpose of performing the model is to show that Decision Tree's performance is the best option for this dataset.



V. Further Analysis

After analyzing the dataset, many issues need further analysis that can't be obtained from the whole dataset. As mentioned in the Explore Data Analysis part, there is a high percentage of loyal customer who neural or dissatisfied about the airline services. Thus, some decision variables makes returning customer using the airline service, even they are not happy about it.

1. Subset of Loyal Customer

   The data of Loyal Customers is subtracted from the dataset. As can be seen in Figure 10, there is 52.2% of returning customer neural or dissatisfied about the service, while only 47,8% people happy about it

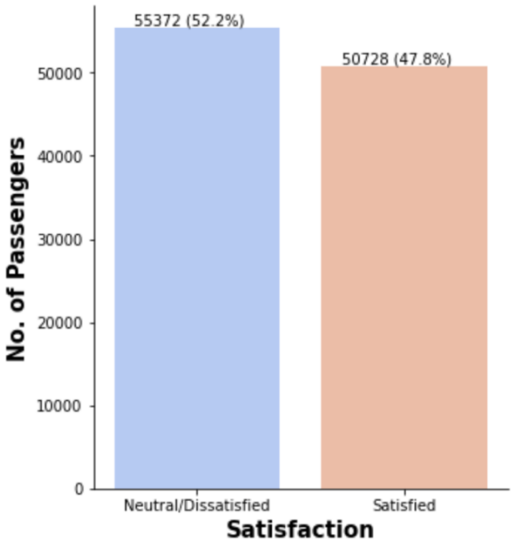


Figure 10: Loyal Customer

2. Correlation matrix of subsets

The correlation matrix is performed to find highly correlated factors with Loyal customers. There is not necessary to performed decision tree and obtained the most important variables because we are not looking for accuracy but relationships. The correlation matrix is performed within the data of Loyal Customer who are neural or dissatisfied about the airlines.

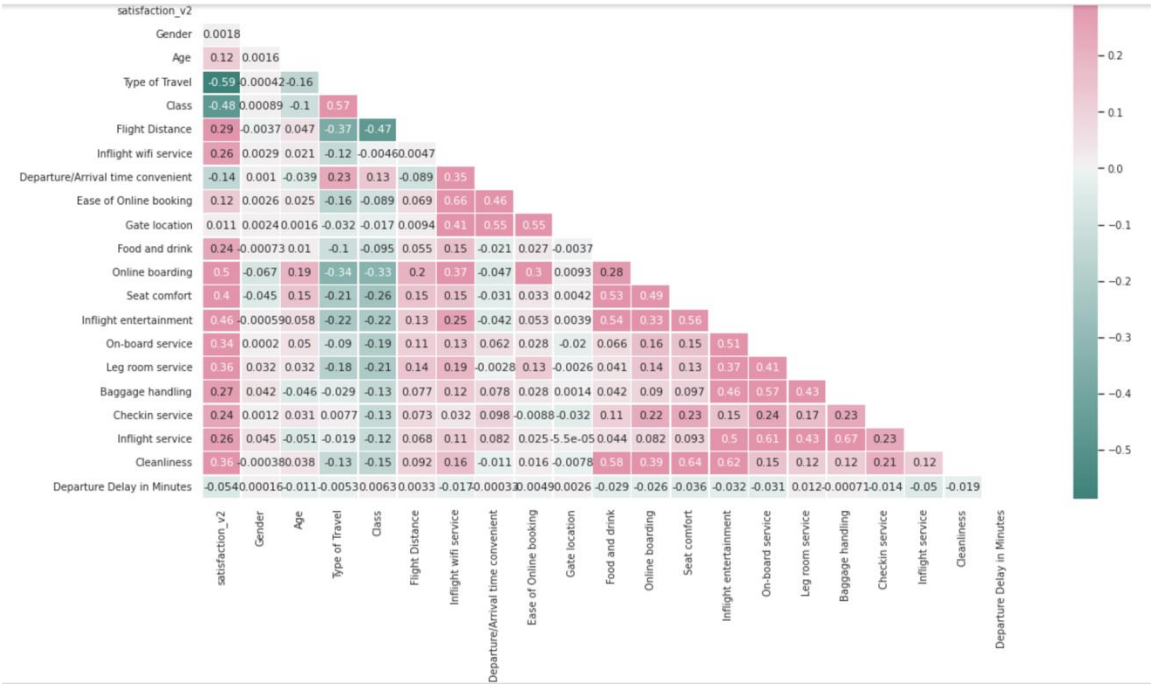


Figure 11: Correlation Matrix of Loyal Customer subset

Top 5 variable that has highest correlation with Neural or Dissatisfied Loyal Customer includes: Online boarding, Seat Comfort, Inflight Entertainment, Leg room service and Cleanliness. The mean score of class 1 – satisfied in the left table shows the satisfied class's average score. The Satisfied column of the table on the right shows the number of people who give a good score. The number of people who gave good scores in the top 4 variables is higher than those who gave low score. The mean of class 0 is also very high, approximately 3, showing that most people are neural about these variables. Thus, even the outcome is dissatisfied, there are many variables in the dataset having good score, in which can be the reason for unhappy returning customer keep using the services.

satisfaction_v2	0	1
Online boarding	2.735010	4.070533
Seat comfort	3.050441	4.072169
Inflight entertainment	2.843567	4.059711
Leg room service	2.932240	3.868889
Cleanliness	2.893719	3.820632

	Dissatisfied/Neutral	Satisfied
Online boarding	48768	57332
Seat comfort	42746	63354
Inflight entertainment	47608	58492
Leg room service	49940	56160
Cleanliness	53563	52537

VI. Conclusion

In general, the report shows data analysis with visualizations and 5 machine learning models is applied to find the data pattern. Decision tree model performed with highest accuracy rate and lower computation load than its ensemble models with 94.5%, ensure the predictions are close to actual values. Therefore, this algorithm can be used to evaluate the airline services. Furthermore, most importance variables can be obtained by their impurity on Decision Tree. This is very important for managers to evaluate the services by only looking into these important variables, instead of looking into hundreds of charts and plots of many variables. The further analysis shows the reasons why unhappy Loyal customers keep using the services, which can be obtained by only analyzing the subset of Loyal Customer. Further suggestion for this analysis is Type of Travel variable, people whose their company pays flight ticket are happier than those who paid by themselves. Therefore, data of price should be included for further analysis