1. **Overview**

The airline industry is one of the largest transportation service industries in this current fast-paced world. The first commercial airline in 1912 was using a form of the dirigible to transport more than 34,000 passengers before World War I. Early air travel began with balloons, gliders, and ultimately airplanes in 1890. Now in the 21st-century air travel is the most essential way of travel from one end to another.[1] As it is the most essential way of traveling, a great concern comes from the quality of travel that is associated with it. To get the answer, we found airline passenger satisfaction data containing information about the satisfaction survey by passengers after completing the trips. The data includes basic information about passengers, and what they think of the flights. We will be researching airline passenger satisfaction factors. The factors that satisfy the passenger so that airlines can develop a specific mechanism of passenger satisfaction measurement. Studying such a data set can help airline companies to improve their quality of service. To cover the drawbacks, the company can enhance passenger’s experience in order to improve passenger satisfaction. We are also intending to find out what features make loyal customers to an airline. To answer these two business problems, classification models are created from the flight satisfaction survey data from Kaggle to identify the critical factors that lead to customer satisfaction and loyal customers.

1. **Prior Researches**

We found some prior research on this dataset, mostly focusing on the model building using different classifiers that can predict if a customer is satisfied by the airline's service. Some of the research included oversampling the dataset to get more accuracy in finding loyal customers.

1. **Dataset Description**

The original dataset contains about 130,000 survey entries and passenger/flight details from a US airline. In total, there are 21 feature columns and 2 binary target columns. Out of all the features, 14 are survey entries where passengers rate the flight experience on a scale of 0 to 5. After removing some NaN values, the resulting data set for the model building has about 129,000 entries. Here is a brief summary of the columns of the dataset.

- **satisfaction**: Airline satisfaction level (satisfied, neutral or dissatisfied).

- **customer\_type**: The customer type (loyal customer, disloyal customer).

- **age**: The actual age of the passengers.

- **gender**: Gender of the passengers (female, male).

- **type\_of\_travel:** Purpose of the flight of the passengers (Personal Travel, Business Travel).

- **customer\_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 (0-5).

- **ease\_of\_online\_booking:** Satisfaction level of online booking

- **inflight\_service:** Satisfaction level of overall inflight service.

- **online\_boarding:** Satisfaction level of online boarding service.

- **inflight\_entertainment**: Satisfaction level of inflight entertainment.

- **food\_and\_drink**: Satisfaction level of inflight food and drink.

- **seat\_comfort**: Satisfaction level of seat comfort.

- **onboard\_service**: Satisfaction level of on-board service.

- **leg\_room\_service**: Satisfaction level of leg room service.

- **departure\_arrival\_time\_convenient**: Satisfaction level of departure/arrival time convenient.

- **baggage\_handling**: Satisfaction level of baggage handling.

- **gate\_location**: Satisfaction level of gate location.

- **cleanliness**: Satisfaction level of cleanliness.

- **checkin\_service**: Satisfaction level of check-in service.

- **departure\_delay\_in\_Minutes**: Minutes delayed when departure.

- **arrival\_delay\_in\_minutes**: Minutes delayed when arrival.

1. **EDA and SMART Questions**

After finding an accessible source of data from the airline's service survey, there are several relevant questions we can ask about the passenger information and features of the airline.

* **SMART Questions:**

**1.1.** Do arrival/departure delays have any effect on customer satisfaction?

As obvious that the delays are mostly not occurring explicitly by the airlines but the weather or airports, we are intending to find out if these unavoidable circumstances are affecting the passenger satisfaction.

**1.2.** Which age group is traveling more frequently and how satisfied they are?

Each age group has different demands and experiences from air travel and the airlines. We want to find out the most frequent age group passengers and also intend to know which age group is having the most satisfaction/dissatisfaction.

**1.3.** Which ticket class passengers have more satisfaction?

Airlines are very particular about the travel class and there is a huge difference in the services between each of the classes. We are interested in finding out which class is giving the most satisfaction. Also, we would like to know, as business class passengers are paying way more than others, how satisfied they are.

**1.4.** Who are more satisfied - Males or Females?

When it comes to satisfaction from any service, there is a significant trend of difference between males and females. We would like to find out from this dataset if we can see any difference or not.

**2.** What are the factors that satisfy the passenger?

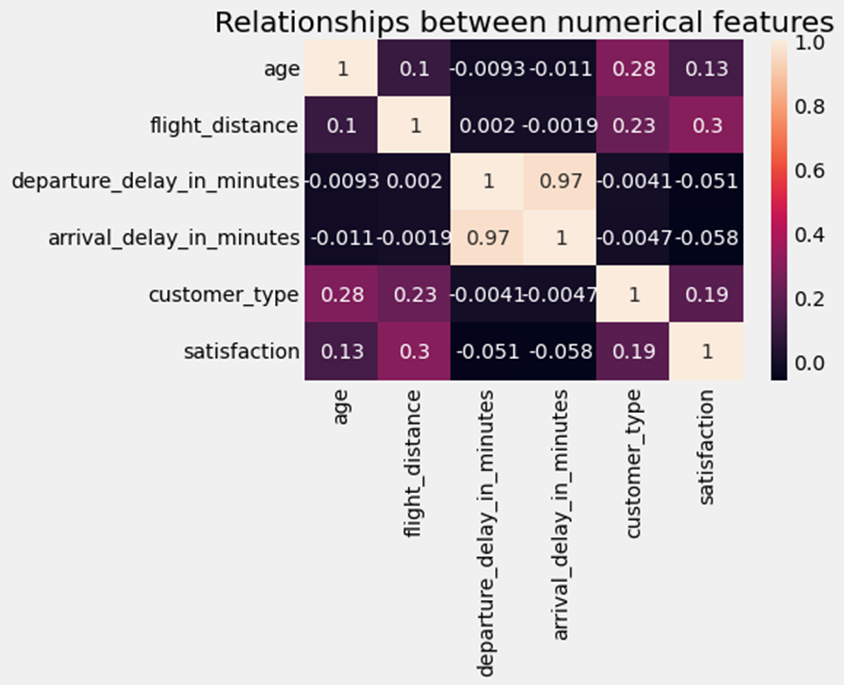
We are interested in finding out the most important features that lead to customer satisfaction by doing classification modeling so that the airlines can give more focus on maintaining the service for them.

**3.** What factors mostly lead to loyal customers?

We will be picking up the top few features that have the most important and build a model that can predict whether a customer is going to be a loyal customer type so that airlines can specifically focus on the potential customers and maintain the service of those important features.

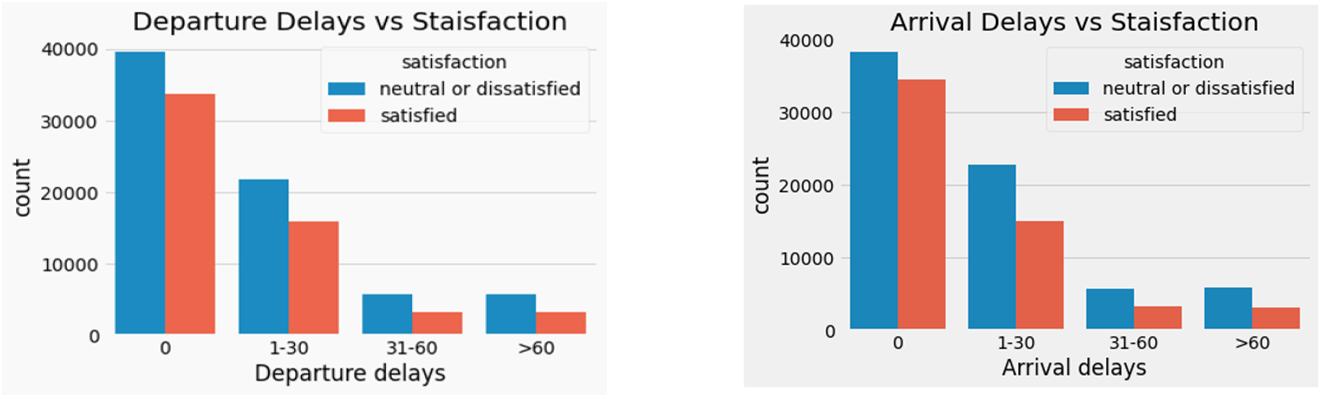
1. **Analysis:**

**1.1.** Arrival Departure Delays and Customer Satisfaction



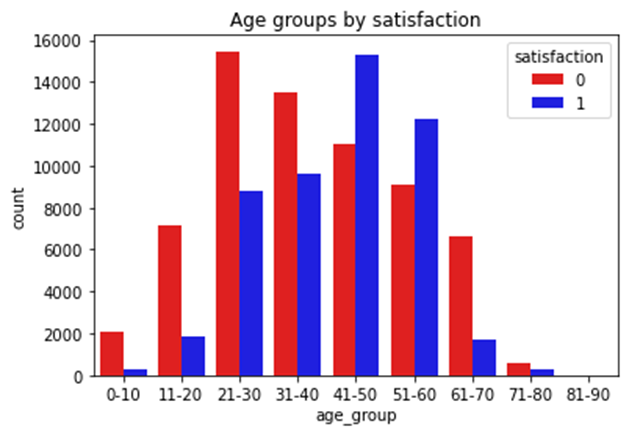
From the correlation matrix, it doesn't seem there is that much effect on the delays and passenger satisfaction. It could be because the delays are mostly not controlled by the airlines but by airport or weather conditions.

Now we'd like to find out if there can differentiate the delays and satisfaction by categorizing delay ranges. For this step, firstly we divided departure delay time into four groups which are not delayed, shorter than half an hour, between 30 to 60 minutes, and longer than 1 hour.

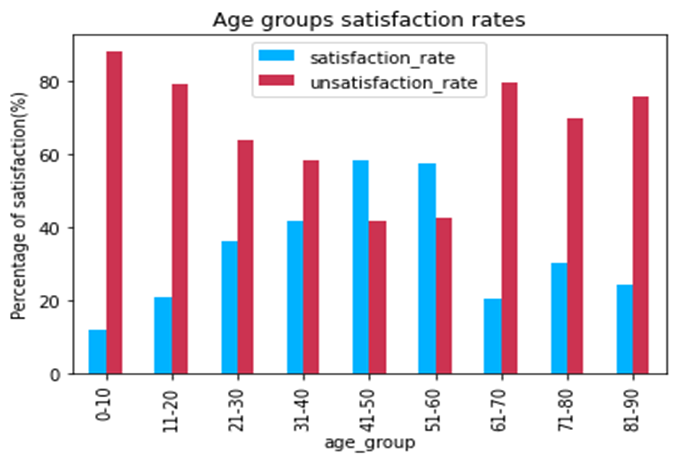


After that, we did a bar plot to show the proportion of passenger's satisfaction in each branch. We found that in each branch the proportion is similar, which means departure is an influential factor. The reason is mentioned in the correlation matrix that the delays may be caused by extreme weather or by other unavoidable circumstances by the airport. Extreme weather will cause many flight delays in the airport, not only this flight. So, the majority of people know that it is not the airlines' fault to delay the flight.

**1.2. Age Groups and Passenger Satisfaction**

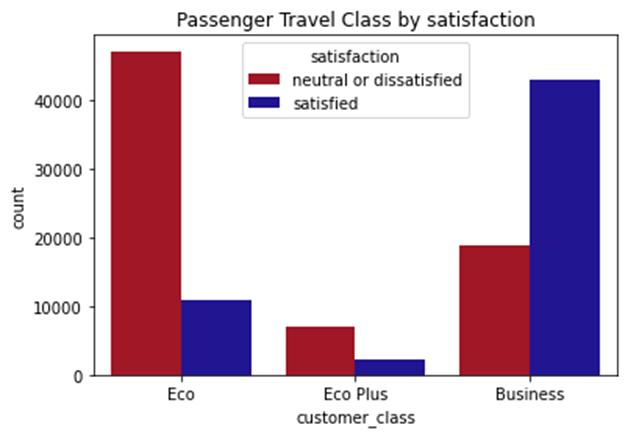
To find out the satisfaction level among different age groups, at first, we made the age group by dividing ages into several intervals of 10. 

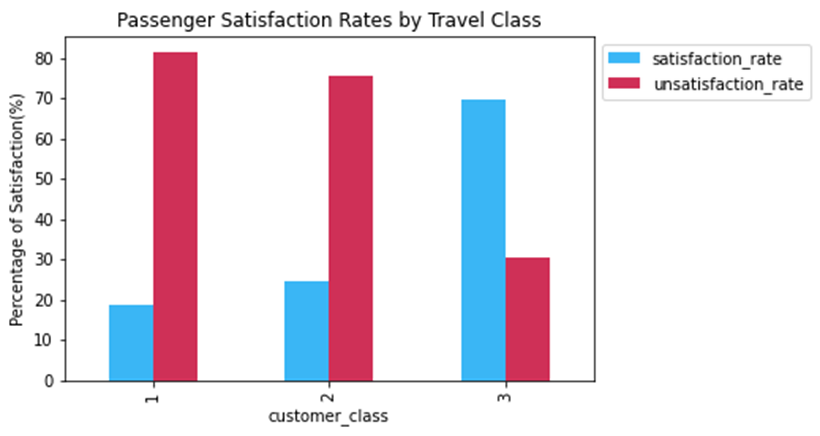
It looks like most travelers are between the ages of 21 to 60. The highest number of travelers are of age 21-30 but most of them are not satisfied. The second highest travelers are 41-50 but most of them are satisfied with the airline service. We also tested our hypothesis by running a Chi-Squared test of independence to determine whether there is an association between age groups and satisfaction. We found the p-value was 0 and very significant, hence we reject the null hypothesis that satisfaction is not associated with age groups and conclude that there is some relationship between them.



Looking at the rate of satisfaction and neutral/dissatisfaction of each age group, the highest rate of neutral or dissatisfaction is from children’s groups. Even though the airplane isn't a child-friendly place, airlines should find some way to improve their satisfaction. On the other hand, 70+ aged passengers are also having a great percentage of dissatisfaction. Even though the number of passengers in these groups is low, airlines should take a look at them as well. As we have observed before, the highest traveler age group is 21-30. Looking at their rates of satisfaction, airlines need to find out what is making them have less satisfaction as less than 40% of them are satisfied. Overall, only the age group 41-50 and 51-60 are having more satisfaction rate than the neutral/dissatisfaction rate.

**1.3.** Satisfaction Among Different Travel Classes

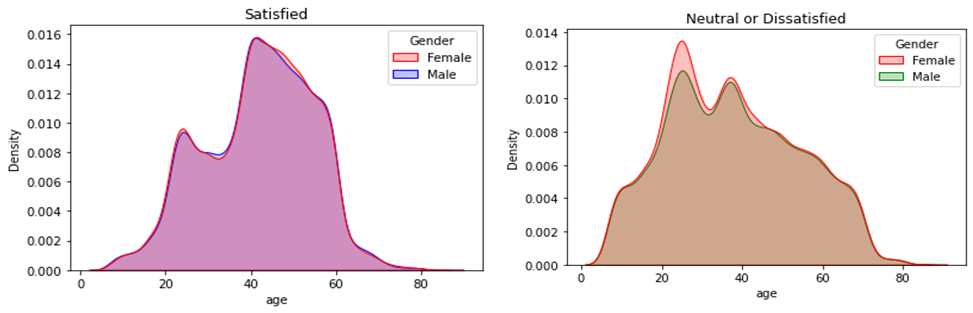


In this airline, most of the passengers are from economy class and business class has the second-highest passengers. 

Here it looks like economy class passengers are not that satisfied with the airline's service. We can also see, around 80% of economy class passengers are not so satisfied and more than 70% of business class passengers are mostly satisfied. Also, as we observed before, economy plus passengers is least in numbers but now we can see only less than 30% of them are satisfied. Airlines need to make the economy class better and also improve themself more to satisfy business passengers as well because they are paying more than others.

**1.4.** Satisfaction Among Male and Female passengers

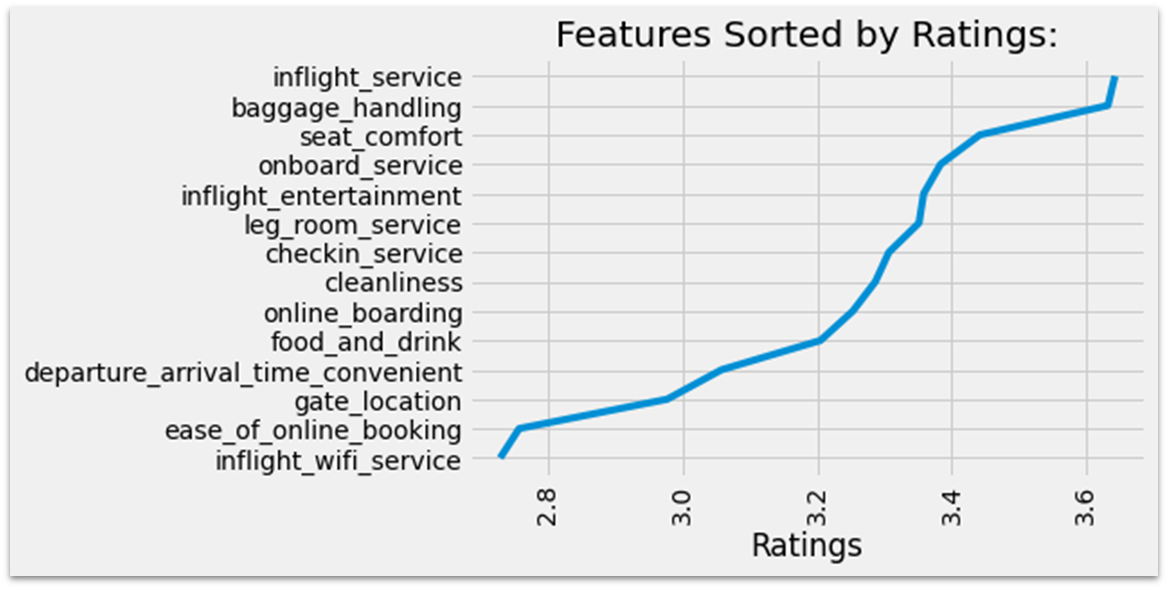
As obvious, when it comes to satisfaction from any service, there is a significant trend of difference between males and females’ satisfaction and comfort. Here we found out females are overall less satisfied than males. Then we explored it by their ages to find out if all ages are having equal satisfaction among both males and females. We compared the satisfied passengers and neutral or dissatisfied passengers separately.



In terms of only satisfaction data, from this kernel density plot, both males and females of the same ages are equally satisfied. But when we look into the neutral/dissatisfied passengers’ distribution, we could find a small imbalance there. Here it looks like more females of age between 20-30 are a bit more unsatisfied than males. Probably, airlines can try to find out what is making them dissatisfied as this group is a very frequent traveler.

1. **Classification & Modeling:**

Our goal is to find out the features that are most important for customer satisfaction and finding loyal customers. For that, we are building models through decision tree classifiers. But before that, we took a look into the features and their overall ratings.

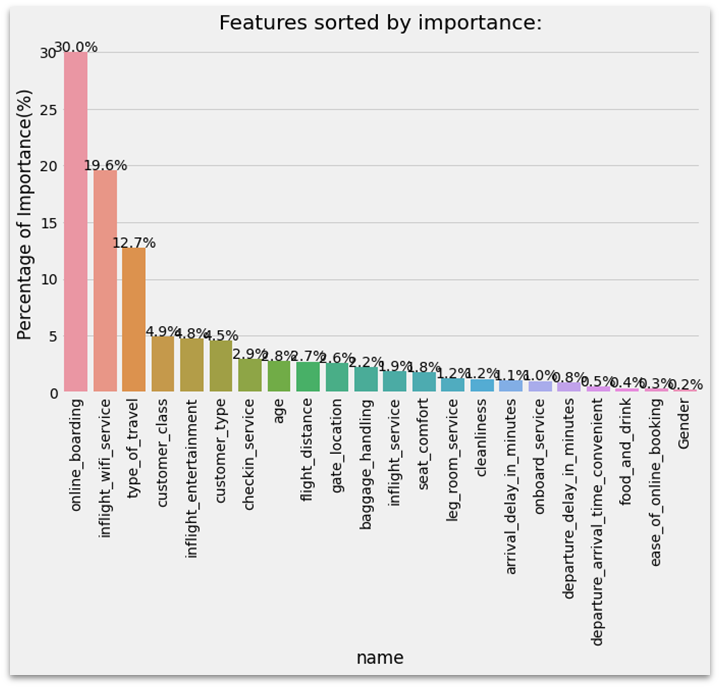


From here we found out the current overall ratings aren’t that satisfactory. The highest rating is just above 3.6 which is neutral in nature. Inflight service and baggage handling are having the highest rating overall whereas inflight Wi-Fi service is having the lowest rating. Observing the ratings, we can say that the airline definitely needs a lot of improvements to do. To make it easier we would like to find out the features that are having the most importance so that they can start improving from there. The average ratings and distribution of each feature are included in the appendix below.

We have divided the dataset into 80% train and 20% test datasets for all the classifications.

**2.** What are the factors that satisfy the passenger?

For this answer, we build many decision trees by changing the feature numbers and depths. We at first built a classification tree with satisfaction as the dependent variable and all other variables as independent, as a result, we got a train set with 100% accuracy which indicates that the classification is heavily overfitted. The test accuracy was 94% and we decided to move on by removing less important features.

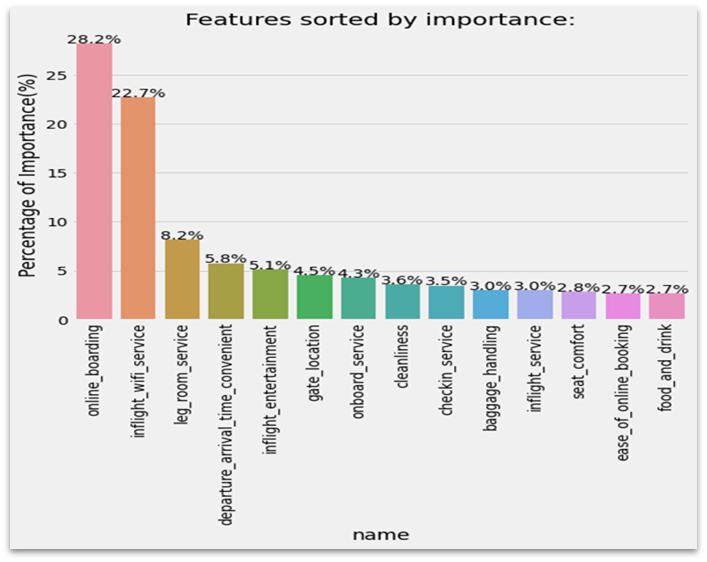


By looking into the feature importance’s from this decision tree model, we can see that online boarding is having the highest percentage of importance. There are some features that aren’t that relevant to customer satisfaction and not going to help much. For instance, customer class, type of customers arrival/departure delays, and gender of the passenger. We can also notice that gender is having the least importance here. This makes sense as looking at the gender of the passenger we cannot predict whether he or she would be satisfied or not. We can also notice that the customer class is having a good percentage of importance, but it doesn’t really add much value. The higher class should be having better service and their satisfaction might be higher than others.

For the next step, we took the top 5 features from here and built another tree classifier. The classifier had 93.86% train and test accuracy. But this tree wasn’t that valuable in terms of the features.

**Decision Tree with All The 14 Review Features**

We further made a tree with all the 14 review types of features so that we can specifically take a count on the service quality of the airlines and find better relevant results.



From the importance of features, we can see that food and drink are weirdly having the least importance here. Seat comfort and ease of online bookings are also having the least importance. As we have seen before, overall, none of the features are having good overall ratings, airlines cannot be assured that these features aren’t that important. Probably they can keep maintaining the current standard of these features or think about improving them later. Overall, this tree has 99% train accuracy and 92% test accuracy with a test precision and recall rate of 89% and 91%. Considering the difference of accuracy between train and test data, it's not a good model. We are moving forward with only the top 5 important features to make another decision tree.

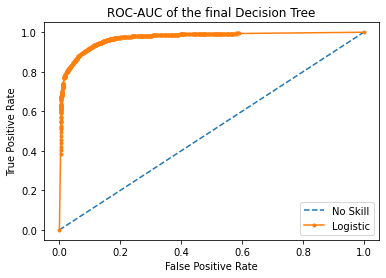
**Decision Tree 5 Most Important Features**

This classification tree gave us 91% train accuracy and 90.9% accuracy. In terms of accuracy, it is a good model.

|  |  |
| --- | --- |
|  |  |

The test data sensitivity is 90.9% and specificity is 91%.

The ROC curve shows the trade-off between sensitivity (or True Positive Rate) and specificity (1 – False Positive Rate), showing the diagnostic ability of these binary classifiers. The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/ (TP + FN)). Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/ (TN + FP)). ROC-AUC score of 80-90 is determined as a good score for the model.

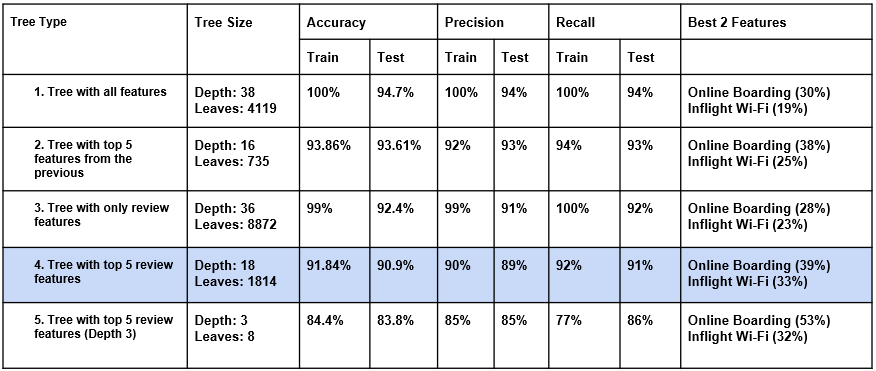


This decision tree’s ROC-AUC score is 97% which represents a very good model to predict the satisfaction of passengers.

The top 5 features and their importance in this model are: -

1. online\_boarding - 39.3%
2. Inflight\_wifi\_service - 32.64%
3. Leg\_room\_service - 10.49%
4. Departure\_arrival\_time\_convenient - 9.36%
5. inflight\_entertainment - 8.21%

Here is the chart of all the decision trees that we modeled and their corresponding scores:

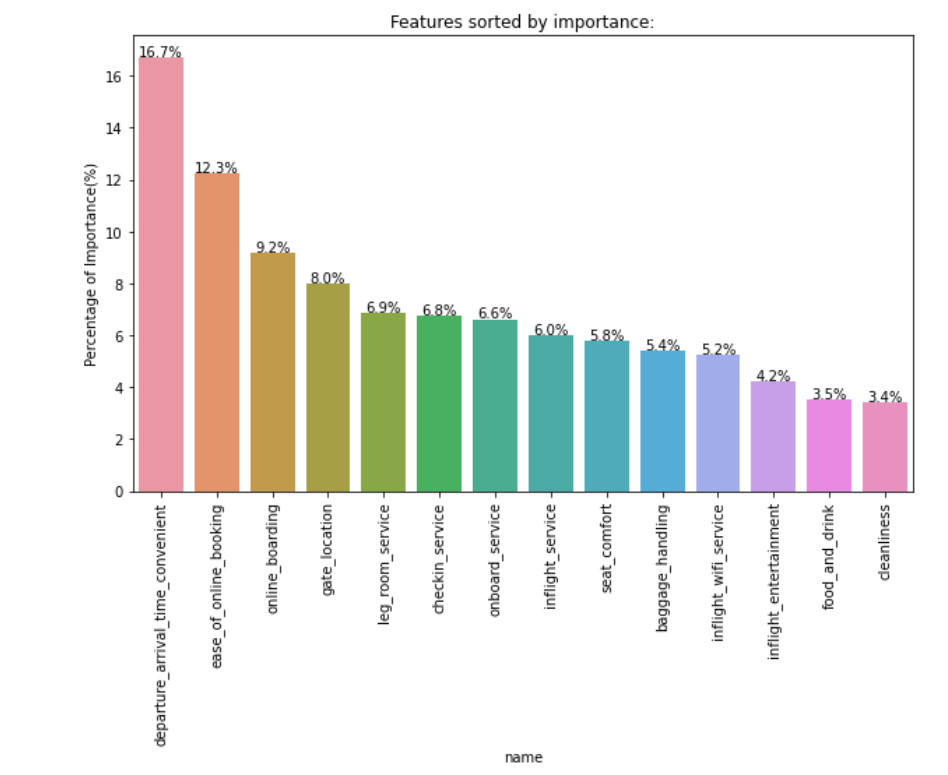


While precision refers to the percentage of our results which are relevant, recall refers to the percentage of total relevant results correctly classified by our algorithm. Under this paradigm, having high precision will be more important for our business problem. To correctly identify the crucial factors leading to customer satisfaction, the model prediction of the positive class, ‘Satisfied’ needs to be very reliable. For that we think the best decision tree would be the one with 5 most important features. It has 89% test precision and 91% recall rate. The most important features to focus on is online boarding and Inflight Wi-Fi service. Moreover, as we found out before these 2 features aren't having good average ratings as well. Online boarding rating is 3.25 and Inflight Wi-Fi is having only 2.73 ratings. So, Airlines should focus on these 2 features at first to get better customer satisfaction.

**3.**  What factors mostly lead to loyal customers?

Finally, we want to find loyal customers. What features impress them that they prefer this airline. To find out we are using the decision tree classifier again with the same 80% train and 20% test data.

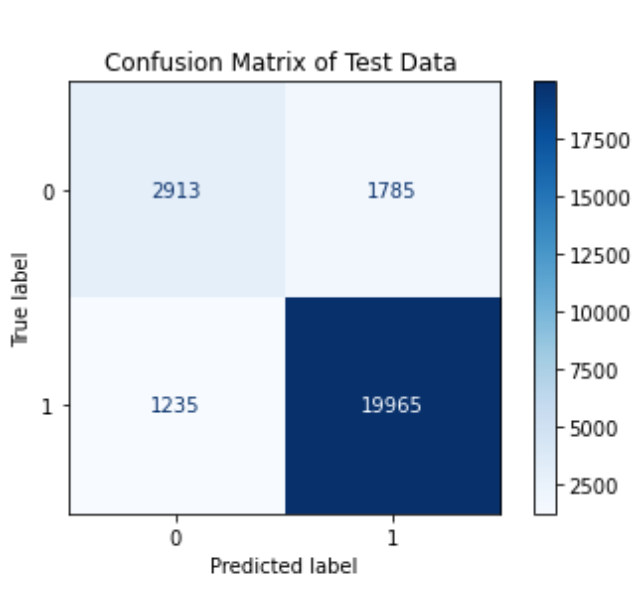
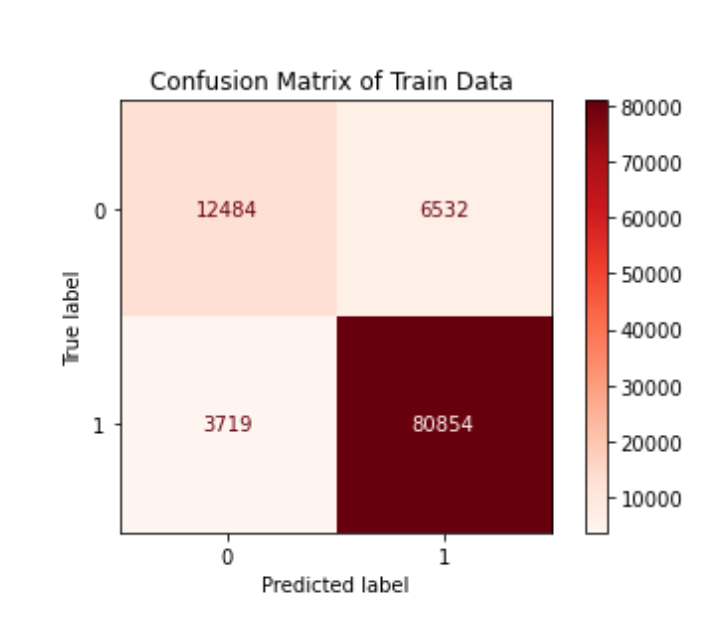
In this part, we think we should only consider the variables that the airline company can decide or improve. So, we built the tree model with the only 14 review features to specifically take a count on the service quality of the airlines and find better relevant results.



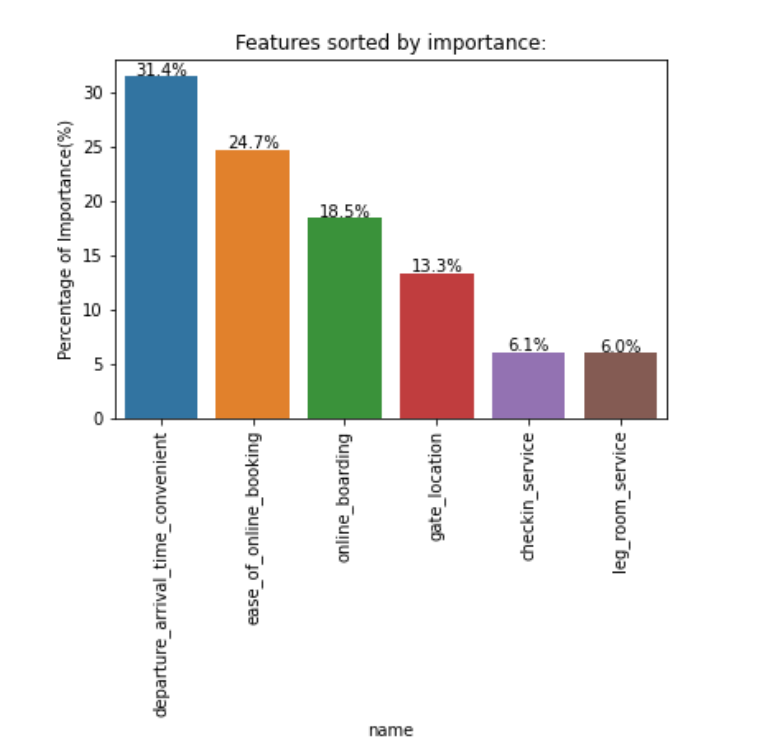
Firstly, we have created a list of importance of all 14 features to do the features selection. As we can see from the picture, food and drink, cleanliness and inflight entertainment have the least importance. And departure\_arrival\_time\_convenient, ease of online booking and online boarding have the most top 3 importance. Overall, this tree with all the features has 99% train accuracy and 92% test accuracy with the test precision and recall rate of 86% and 91%. However, obviously, there are too many features in this tree model, and it is not a good model. Therefore, we can choose the top 6 features to build a new tree model and see its performance.

**Decision Tree 6 Most Important Features**

In this classification tree model, it gave us 90% train accuracy and 88% test accuracy. It is much better than the model with all features.



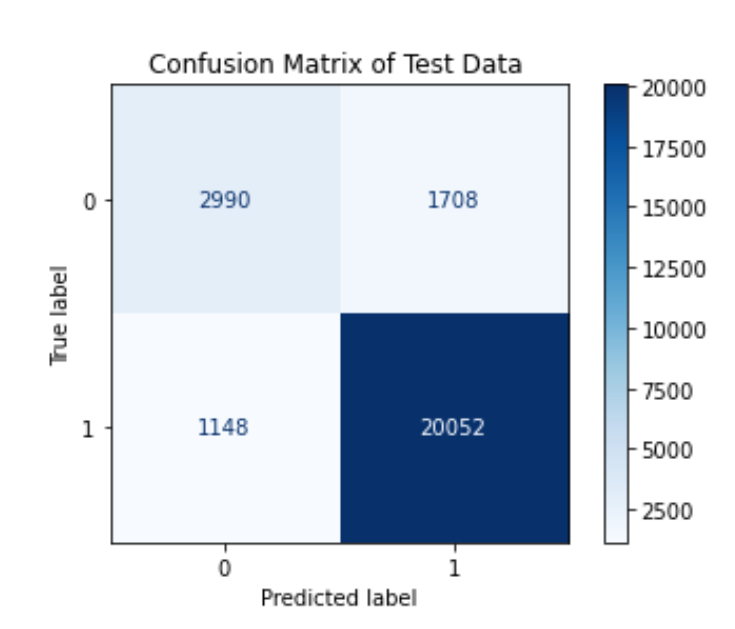
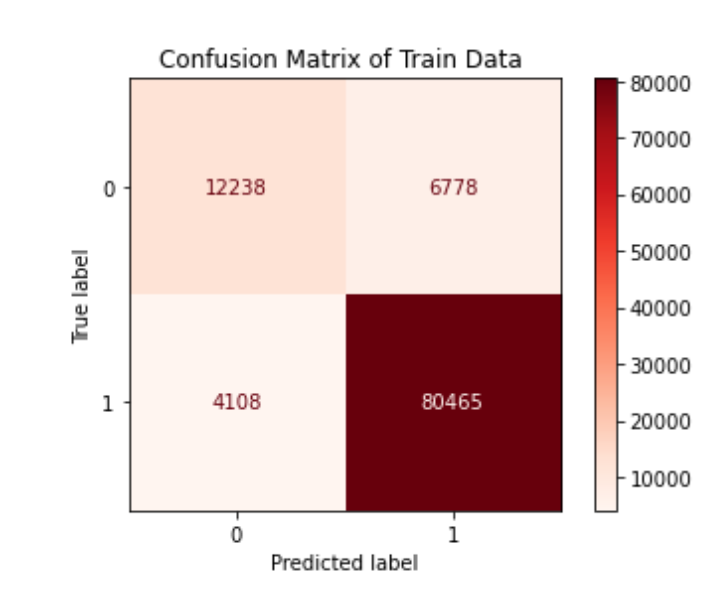
Observing the confusion matrix, this model is very good at predicting the loyal customers but not that efficient in predicting disloyal customers.



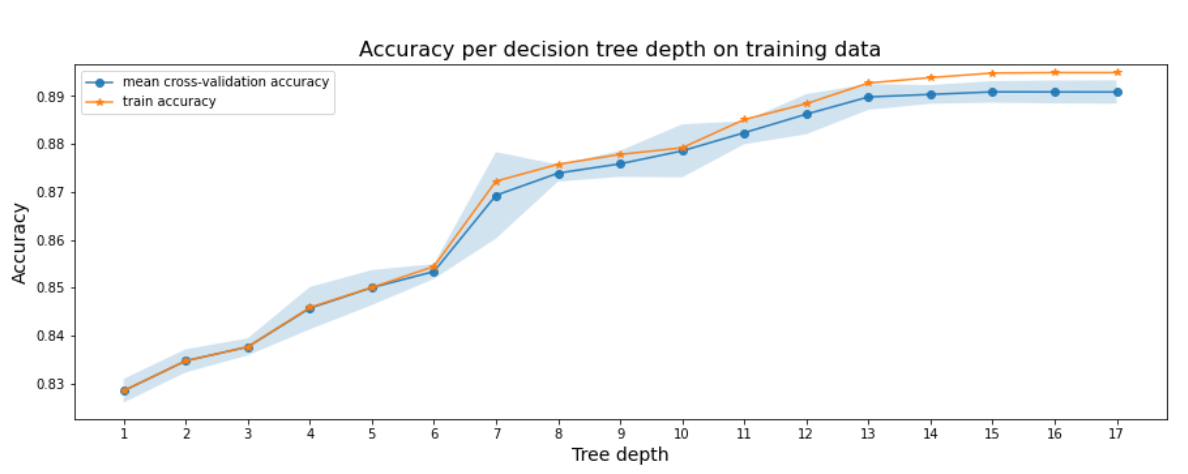
This is the list of importance with the features in this classification tree model. And we can try to choose the top 5 features to build another model to compare the performance with the tree model with the top 6 features.

**Decision Tree 5 Most Important Features**

This tree model with the top 5 features gave us 89.4% train accuracy and 88.9% test accuracy. Apart from this, we have tried other models with other different features. By comparing their performance with the model with the top 5 features, we found the model with the top 5 features is more appropriate. Therefore, we decide to choose the top 5 features to build the final model.



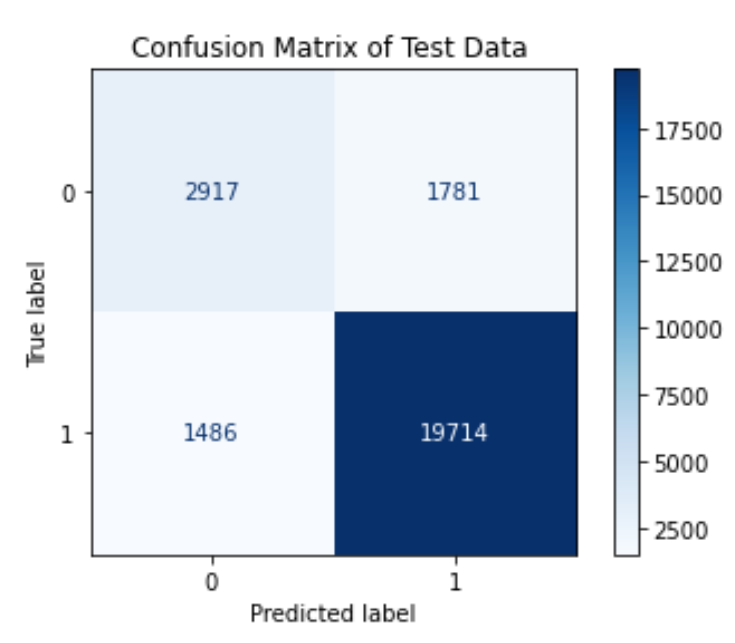
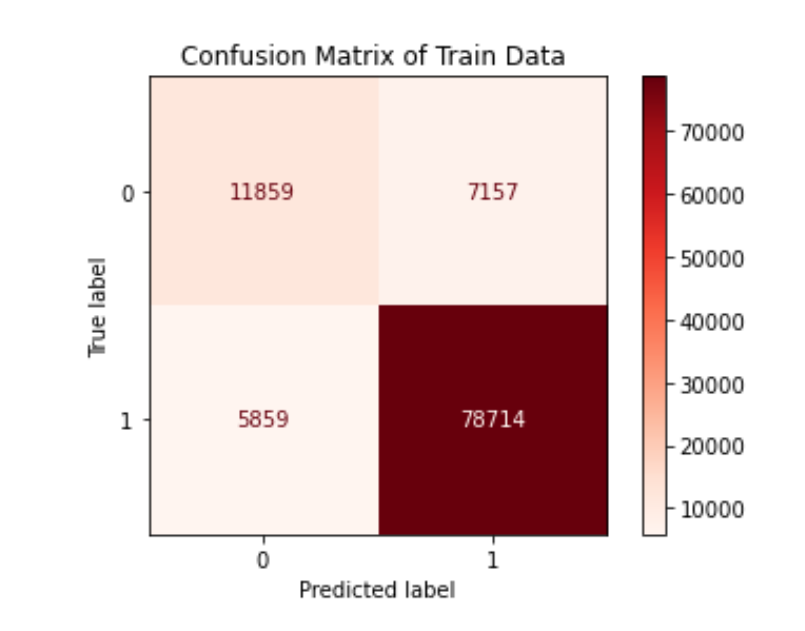
Then, we used the k-fold Cross-Validation to find the proper depths in this classification tree model.



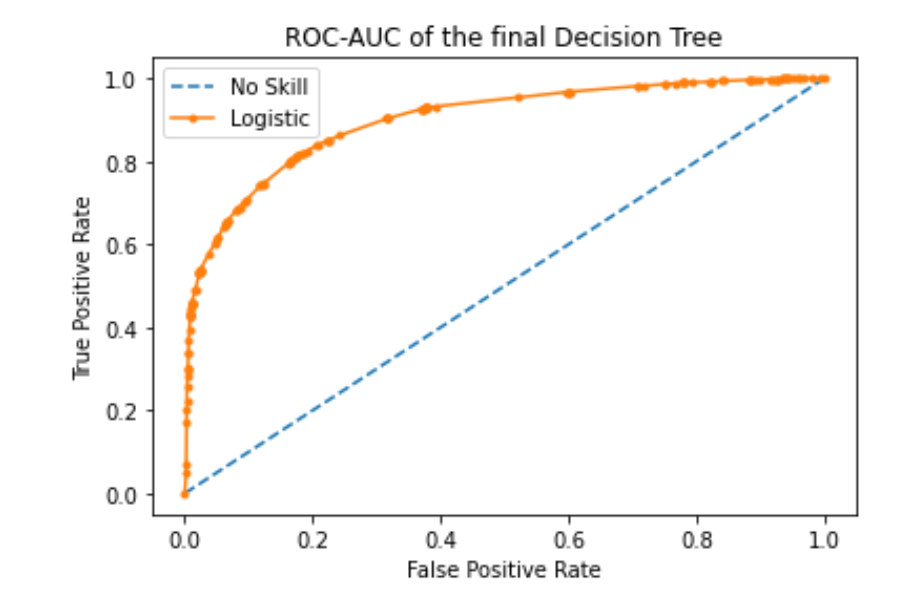
As we can see from the picture, when depth is 8, it achieves a high average accuracy score by using cross-validation folds with size 5. And from depth 13, the score is almost the same and there are not many improvements in average accuracy score from depth 8 to depth 13. And we can also see in depth 8, the lower bound of the confidence interval of the accuracy is high enough to make the value significant. Therefore, we have decided to choose depth 8 to build the final tree model. It can help us avoid overfitting and give a better chance to reproduce the accuracy and generalize the model on test data.

**Decision Tree 5 Most Important Features with Depth 8**

In this classification tree model, it gave us 87.43% train accuracy and 87.38% test accuracy.



This model is also very good at predicting loyal customers and in terms of accuracy, it is a good model. Apart from this, we can also check the ROC and AUC of this classification tree model.



The AUC (Area Under the Curve) is 0.9, which is much bigger than 0.8. Therefore, we can consider this tree model as a good model without any doubt.

The top 5 features and their importance in this model are:

1. Departure\_arrival\_time\_convenient 38.34%
2. Ease of online booking 26.6%
3. Online\_boarding 20.58%
4. Gate location 12.97%
5. Check service 1.51%

From what we have discussed above, departure\_arrival\_time\_convenient, ease of online booking and online\_boarding are the most important features to a loyal customer and in this model, they together cover more than 85%. So, airlines should focus on these 3 features at first to turn a customer to a loyal customer.

**7. Conclusion**

With finding out the important factor of passenger’s satisfaction, we can look forward to making air journey more comfortable and hassle-free. From this study, at first, we found out, among different age groups, the most satisfied age group is between 41 to 50. Also, for children and passengers aged greater than 70, the airline does not receive much satisfaction from them. It is highly suggested that airlines should try to make flights more comfortable for children and aged people, like providing helpful services. While exploring the satisfaction among different gender, we found out females aged between 20 to 30 are less satisfied than males. Young female teenagers may have higher requirements for the flight experience so the airline could aim at this point to find out the issues and make some changes according to that. We further explored the delays and passenger reactions and we concluded that arrival and departure delays have a small very effect on passenger satisfaction. The reason might be that passengers understand it is not the airline’s fault to delay the flights. The majority of delays are due to extreme weather which disrupts the whole airport and aviation system, not just one specific airline. Moving on to exploring the travel classes of the airline, it is suggested that the airlines are required to improve the service quality no matter of economy class or business class. Through our classification modeling, we can suggest that there are two most important features that are covering the major importance of customer satisfaction. They are online boarding and inflight WI-FI service. In recent years, more airlines provide online services and more technical services to make the flight experience more convenient, less time-consuming, and enjoyable. However, the quality of these services requires improvement as we found out this airline’s none of the features are having overall great ratings. The airline is suggested to build a more comprehensive online system. Also, airlines should highly focus on inflight WIFI experience as it has the lowest rating but high influence on satisfaction. Finally, to get a hold of the loyal customers, departure, arrival time convenience, and ease of online booking are major important factors to keep improving so that customers can rely on this airline without any second thoughts. To find out the potential loyal customers, the airline can use the decision tree-based model that we have created with 87% of accuracy from the reviews of features taken from them after the journey. Overall, in conclusion, the airline company still needs to make lots of improvements to increase satisfaction, and the problems are obvious. We believe the company will be better after solving these problems.

**8. Appendix:**

**Distribution of the columns:**

**Numeric Columns:**

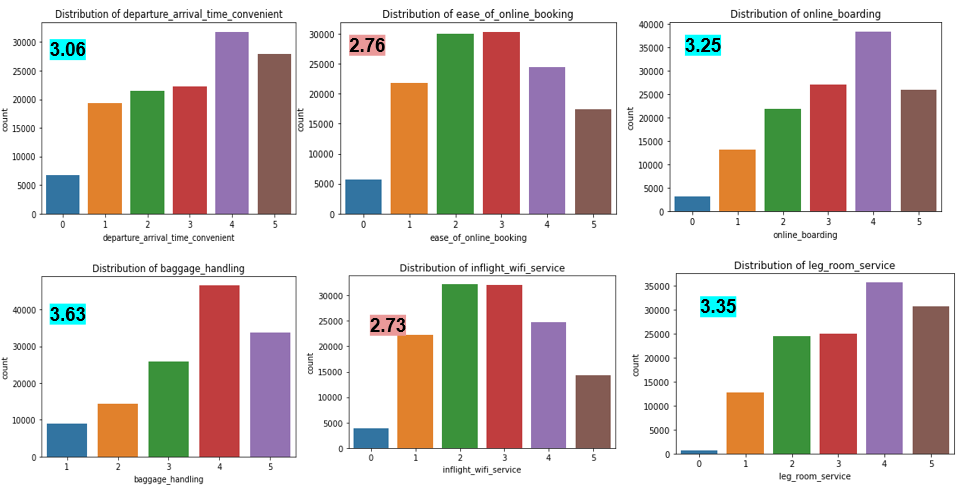
|  |  |
| --- | --- |
|  |  |
|  |  |

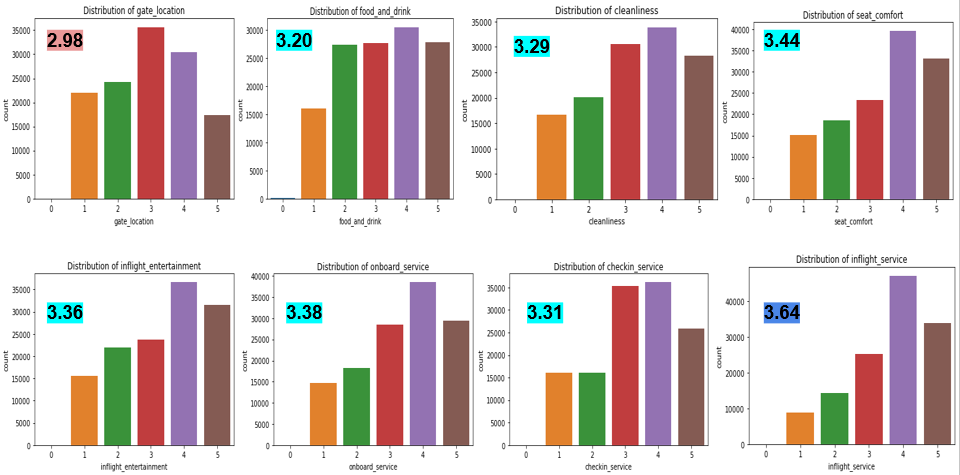
**Categorical Columns:**

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |
| --- | --- |
|  |  |

**Review Features:**

****

****

**Reference:**

**Link of the Dataset:** <https://www.kaggle.com/binaryjoker/airline-passenger-satisfaction>

[1] https://encyclopedia2.thefreedictionary.com/airline+industry