


IMPROVING SERVICE QUALITY OF THE AIRLINE USING CUSTOMER SATISFACTION LEVEL



IS4007 - Stat in Practice II

Individual Activity 2

U.A.B.M. Gunasekara

S14330

Abstract

The airline industry has been revived with the removal of the Covid-19 virus. Airline companies will have to improve their service quality or follow new techniques to bring their airline back to the way it was. This study was done to find the what are the most important variables that impact the customers' satisfaction level. And what are the improvements that can be done for services that provide by airlines? The dataset that has more than 129000 observations was used to do this study. Both descriptive analysis and as advanced analysis Logistic Regression have been done. IBM SPSS 25 software and sklearn library in python language were used to do the study. As result, In-flight service Leg room service for women, and Online Boarding service for men were recognized as improvement needed services. And Online Boarding and In-flight Wi-Fi services are the most important variables that could be affected customer satisfaction have found.

Table of Content

Abstract	1
Table of Content.....	2
Table of Figures	3
1 Introduction	4
2 Literature Review	5
3 Theory and Methodology	6
3.1 Theory	6
3.2 Methodology	7
4 Data	8
5 Exploratory Data Analysis	9
6 Advanced Analysis	13
7 General Discussion and Conclusion	16
8 References	18

Table of Figures

Figure 3.1.1: Interpretation of ROC Curve	7
Figure 5.1: Pie chart of dependent variable (Satisfaction)	9
Figure 5.2: Pie chart of gender variable.....	9
Figure 5.3: Stacked bar chart of Satisfaction with Gender	9
Figure 5.4: Stacked bar chart of In-Flight Service Satisfaction Levels with Gender	10
Figure 5.5: Stacked bar chart of Online Boarding Service Satisfaction Levels with Gender	10
Figure 5.6 : Stacked bar chart of Leg Room Service Satisfaction Levels with Gender.....	10
Figure 5.7: Stacked bar chart of Baggage Handling Service Satisfaction Levels with Gender	11
Figure 5.8: Stacked bar chart of In-flight Wi-Fi Service Satisfaction Levels with Gender.....	11
Figure 5.9: Stacked bar chart of Food and Drink Service Satisfaction Levels with Gender.....	11
Figure 5.10: Stacked bar chart of Cleanliness Satisfaction Levels with Gender.....	11
Figure 5.11: Stacked bar chart of In-flight Entertainment Service Satisfaction Levels with Gender	11
Figure 5.12: Frequency Table of satisfaction levels of Ease of Online Booking	12
Figure 5.13: Frequency Table of satisfaction levels of In-flight Wi-Fi Service	12
Figure 6.1: Heat-Map between independent variables	13
Figure 6.2: Confusion Matrix of Train Set.....	14
Figure 6.3: ROC curve of the model.....	15
Figure 6.5: ROC Curve of the Test Set	15
Figure 6.6: Confusion Matrix of Test Set	15

1 Introduction

There are over 5000 international airlines around the world. From there around 300 flies between continents (Continental Airlines). But among them only a few numbers of airline companies are popular. During the period of Covid 19 epidemic was severe some airline companies were bankrupt. Because some countries close their countries for foreign tourists. With the decrease of the covid 19 pandemic, people all around the world tend to travel around the world as usual. And most countries have removed their traveling restrictions. Because of those reasons, competition among airline companies is increased widely.

Therefore, it is important for airline companies to re-establish their names and win this competition, taking into account the level of satisfaction and customer expectations. This study will be important for airlines to move forward with their company. And those who look forward for make new changes for their airlines can get new ideas from this analysis.

Objectives of the study

The main objective of the study is to improve the service quality of the airline using customer satisfaction data. To fulfill this objective some sub-objectives were assigned.

- Study the satisfaction level of each service with respect to gender.
- Building a model for checking the satisfaction probability of customers.
- Finding the most important variable that can impact customer satisfaction level.

Any airline service can use the results of this study to improve their service quality.

2 Literature Review

There are many studies about customer satisfaction levels of airline service. Some of the studies have focused on different classes separately, in some studies, they have more focused on the costs. Most of the researchers based on their study only one airline service. The different researcher has used to collect their data in different ways. Most of the researchers have used questionnaires, some have used passenger reviews (Sezgen, 2019), some have used online ratings (Sudhakar, 2020), etc. According to the journal papers can be seen most analyses are based on similar types of factors.

When considering the airline industry their service can be divided into a large number of sub-services. Checking service, baggage handling service, food, and drink service, online booking, in-flight services, and cleaning services are some of them. In addition to that when studying the satisfaction of airlines have to consider some other facts also. Gender, Age as biological factors, and distance to the destination, delay of the flight like physical factors are also considered. So, analyzing this type of various types of variables does not limit only to traditional statistical techniques. Have to use new techniques like Machine Learning (Hulliyah, 2021).

When using techniques like Machine Learning have to be more careful. After fitting the model there is an important step like checking the accuracy, having to investigate whether the model is overfitted or not because there is a higher chance to overfit the ML models. Especially when there is higher accuracy it is good to check the model has over fitted or not.

3 Theory and Methodology

3.1 Theory

Logistic Regression

Logistic Regression (LR) is a special case of linear regression. When dependent variable categorical and binary variable logistic regression can be used. It can be known as a classification technique also. The predicted probability of mutually exclusive events include in the dependent variable according to the effect of independent variables is given by the logistic regression model. Though ordinary regression equations cannot be used for that purpose. LR model is used logit function to involve the probability of happening binary events. To convert the function values to a value between 0 and 1 sigmoid function is used by the model. (Nath, 2020)

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

$$P = S(y) = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\dots\dots+\beta_px_p)}} \dots\dots\dots (2)$$

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\dots\dots+\beta_px_p)}} \dots\dots\dots (3)$$

Logistic Regression Model

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$$

Parameters of this model will be estimated by Maximum Likelihood Estimation.

Predictions of this model are based on the probability given by this model. If the probability is greater than some cut-off value that observation belongs to class 1. If not, it belongs to remain class (class 2).

Generally, the cutoff value is 0.5 for the majority. But it can be changed according to the situation.

Performance of Logistic Regression Model:

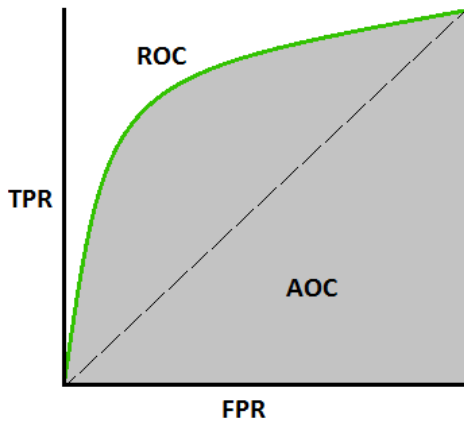
- One way of calculating model performance is checking the model deviance.
- Another way is by using Confusion Matrix.

	Predicted	
	True (TF) Positive	False (FN) Negative
A c t u a l	False (FP) Positive	True (TN) Negative

$$\text{Accuracy of the Model} = \frac{(TP) + (TN)}{(TP) + (TN) + (FP) + (FN)}$$

- Area Under the Curve (AUC)

ROC Curve plotted with True Positive Rate (TPR) Vs False Positive Rate (FPR)



AUC Range	Classification Accuracy
0.9 – 1.0	Excellent
0.8 – 0.9	Good
0.7 – 0.8	Fair
0.6 – 0.7	Bad
0.6 – 0.5	Very Bad

Figure 3.1.2: Interpretation of ROC Curve (Narkhede, 2018)

3.2 Methodology

1. Data Preprocessing

First, how many null values are in each variable was checked. Only 393 null values in the data set. And also, Arrival Delay was the variable to that null values belong. There was not an arrival delay that might be the reason for that null value. When comparing the number of records of the dataset number of null values is inconsiderably small. So, records that have null values were removed from the dataset. 43.45% and 56.55% were the percentages of (Satisfied and neutral or dissatisfied) two categories of the dependent variables. Those results were shows there is no need of balancing the data separately.

2. Exploratory Data Analysis

By choosing a suitable chart type univariate analysis was done for all variables as an initial step. Then according to the objectives, type of the variable, and the number of categories (in categorical variables) rest of the charts were plotted after applying or not relevant filters for the dataset. IBM SPSS 25 software was used for the whole descriptive analysis.

3. Advanced Analysis

Using heatmap, ideas about the correlation between variables were taken. Dataset was split into two parts 0.2 and 0.8. 80% part was taken as the train set and remain part was taken as the test set. After that dummy variables were created for nominal variables. Next, the model selection procedure was applied to drop lesser significant variables. The model was built using the Logistic Regression model in sklearn library in python. Finally, the model was applied to the train and test sets and checked the accuracy using the confusion matrix and AOC value.

4 Data

The data set is freely available on kaggle.com. Customer satisfaction scores from 120,000+ airline passengers and additional information about each passenger under 24 variables have been included. Five of them are quantitative and remain 19 are qualitative. From qualitative variables, 5 variables are nominal, and remain 14 are ordinal.

Field	Description
ID	Unique passenger identifier
Gender	Gender of the passenger (Female/Male)
Age	Age of the passenger
Customer Type	Type of airline customer (First-time/Returning)
Type of Travel	Purpose of the flight (Business/Personal)
Class	Travel class in the airplane for the passenger seat
Flight Distance	Flight distance in miles
Departure Delay	Flight departure delay in minutes
Arrival Delay	Flight arrival delay in minutes
Departure and Arrival Time Convenience	Satisfaction level with the convenience of the flight departure and arrival times from 1 (lowest) to 5 (highest) - 0 means "not applicable"
Ease of Online Booking	Satisfaction level with the online booking experience. ("From 1 (lowest) to 5 (highest) - 0 means "not applicable" ")
Check-in Service	Satisfaction level with the check-in service (" ")
Online Boarding	Satisfaction level with the online boarding experience (" ")
Gate Location	Satisfaction level with the gate location in the airport (" ")
On-board Service	Satisfaction level with the on-boarding service in the airport (" ")
Seat Comfort	Satisfaction level with the comfort of the airplane seat (" ")
Leg Room Service	Satisfaction level with the leg room of the airplane seat (" ")
Cleanliness	Satisfaction level with the cleanliness of the airplane (" ")
Food and Drink	Satisfaction level with the food and drinks on the airplane (" ")
In-flight Service	Satisfaction level with the in-flight service (" ")
In-flight Wi-Fi Service	Satisfaction level with the in-flight Wifi service (" ")
In-flight Entertainment	Satisfaction level with the in-flight entertainment (" ")
Baggage Handling	Satisfaction level with the baggage handling from the airline (" ")
Satisfaction	Overall satisfaction level with the airline (Satisfied/Neutral or unsatisfied)

Techniques that were used for data pre-processing was mentioned in the methodology.

5 Exploratory Data Analysis

Basic and important charts of univariate analysis are included as usual.

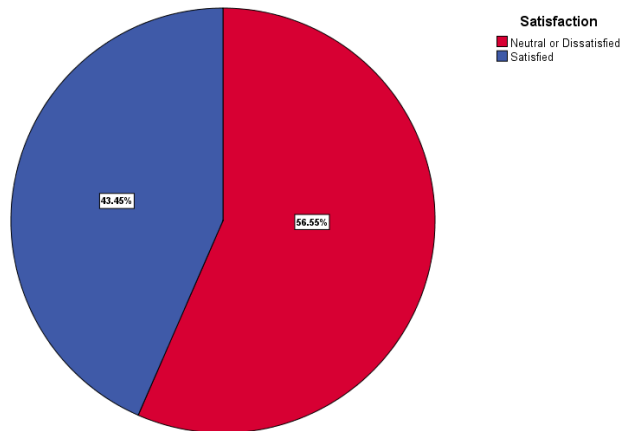


Figure 5.1: Pie chart of dependent variable (Satisfaction)

According to this pie chart, two responses in the dependent variable have been distributed nicely. No need for any extra effort to balance the dataset. The percentages of satisfied and dissatisfied/ Neutral are 43.45% and 56.55% respectively. Logistic Regression can be applied to this dataset without any changes.

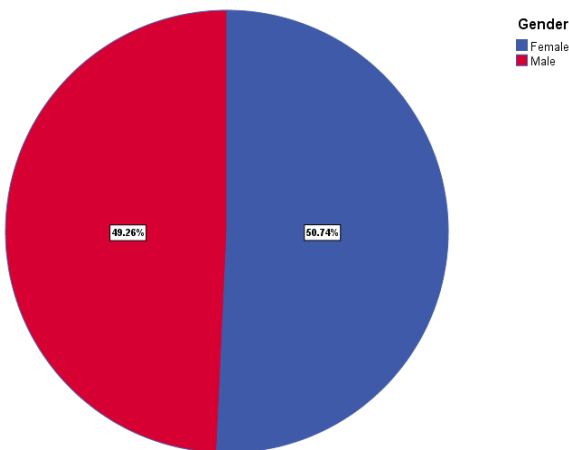


Figure 5.2: Pie chart of gender variable

There is an almost 50/50 gender distribution have this dataset 49.26% are male and 50.74% are female in this dataset. So, this dataset is not biased toward any gender. Gender can be a good variable to take as a base variable to analyze other variables because of this unbiased distribution.

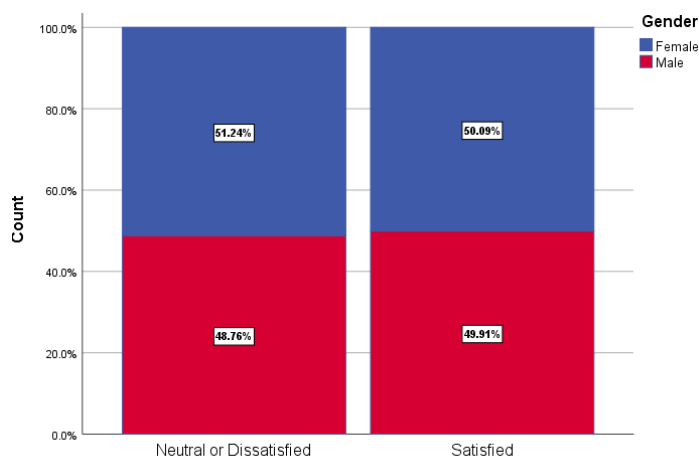
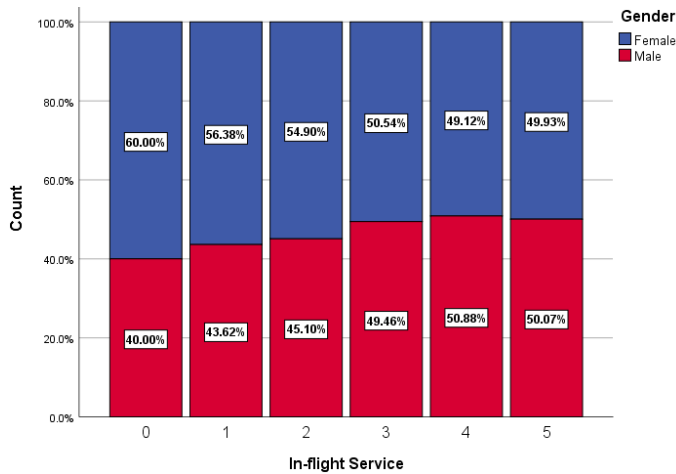


Figure 5.3: Stacked bar chart of Satisfaction with Gender

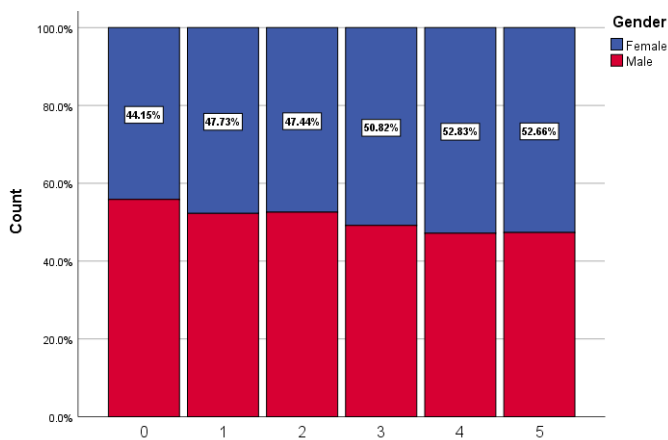
This chart shows the distribution between two categories of the gender variable inside each satisfied and dissatisfied/neutral. This proves the fact that was mentioned just above. Male and females have distributed 50/50 inside each category.

Satisfaction level of services with Gender variable



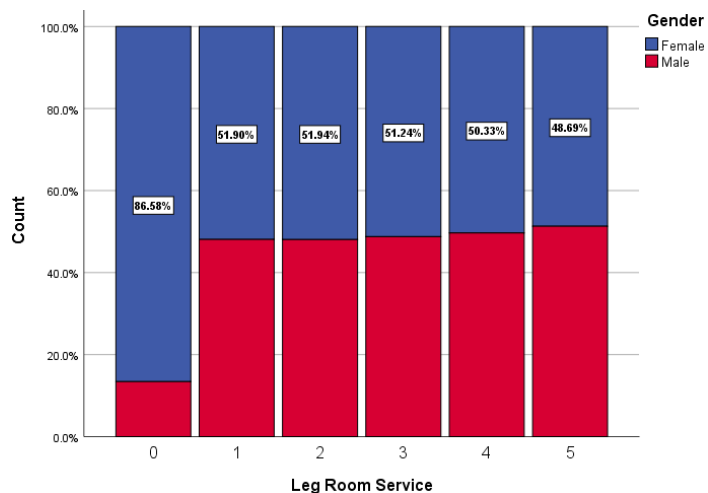
This chart shows a clear decrease in the female percentage (comparing the male) with the increase in satisfaction level of In-flight service. According to the graph, there is a clear decrease in women's percentage when going to higher satisfaction levels. But for men, the scenario is the other side of it.

Figure 5.4: Stacked bar chart of In-Flight Service Satisfaction Levels with Gender



This chart shows a similar de-trend as the above chart for men (not for women). The percentage of selecting the lower satisfaction level for online boarding is higher for men than women. But when goes to higher satisfaction levels it changes to other way women percentage higher than men.

Figure 5.5: Stacked bar chart of Online Boarding Service Satisfaction Levels with Gender



This chart shows the higher female percentage in the dissatisfaction (0th level) bar. It is a very huge percentage difference between genders. And this is un-regular behavior. As above cases here also have a small percentage decrease when goes to higher satisfaction level.

Figure 5.6: Stacked bar chart of Leg Room Service Satisfaction Levels with Gender

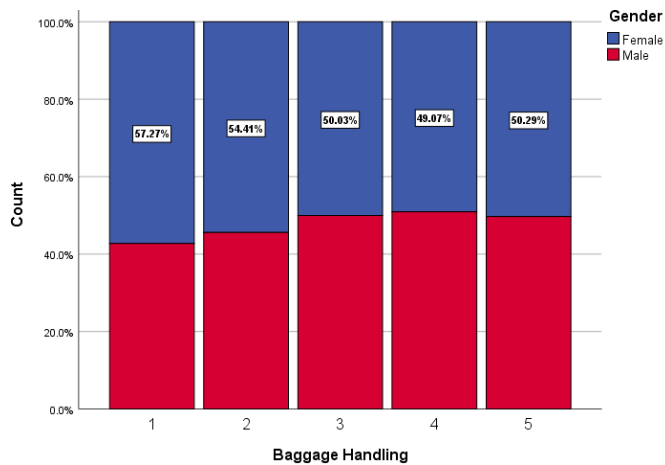


Figure 5.7: Stacked bar chart of Baggage Handling Service Satisfaction Levels with Gender

In here there is one special situation than above-mentioned services. Any of the customers have not selected dissatisfied (0th level). As previous charts in here also some female percentage decreasing trend when goes to lower from the highest satisfaction levels.

Except for the above-mentioned services all other services' satisfaction levels do not show any gender biases. Almost all satisfaction levels of that services have 50/50 male and female percentages. Some of those charts are included below.

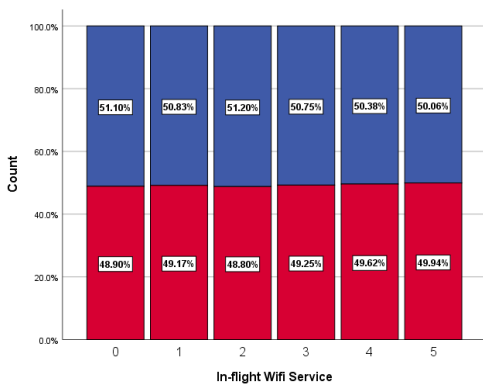


Figure 5.8: Stacked bar chart of In-flight Wi-Fi Service Satisfaction Levels with Gender

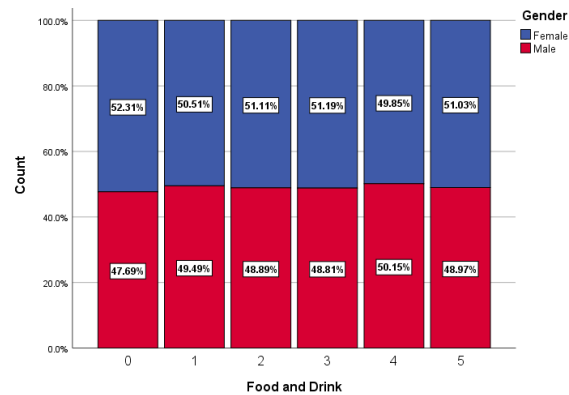


Figure 5.9: Stacked bar chart of Food and Drink Service Satisfaction Levels with Gender

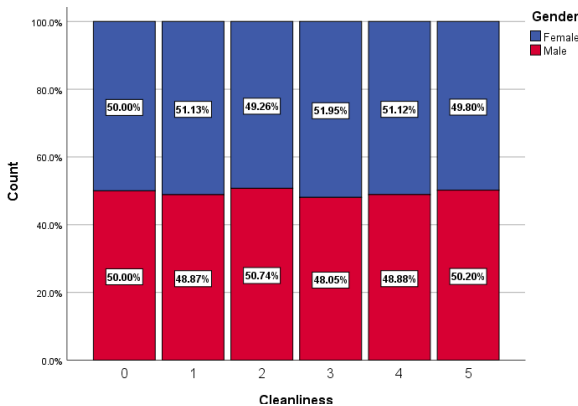


Figure 5.10: Stacked bar chart of Cleanliness Satisfaction Levels with Gender

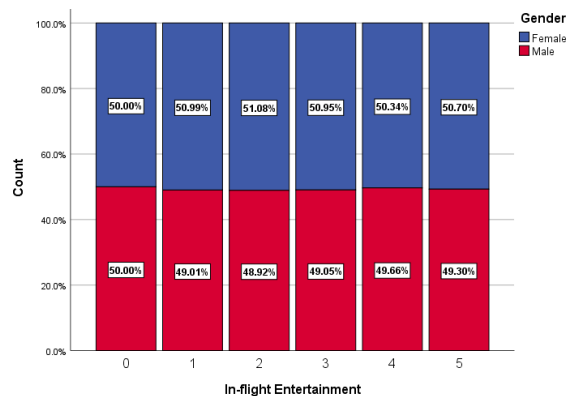


Figure 5.11: Stacked bar chart of In-flight Entertainment Service Satisfaction Levels with Gender

Ease of Online Booking					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	5666	4.4	4.4	4.4
	1	21808	16.8	16.8	21.2
	2	29983	23.2	23.2	44.4
	3	30297	23.4	23.4	67.8
	4	24362	18.8	18.8	86.6
	5	17371	13.4	13.4	100.0
	Total	129487	100.0	100.0	

Figure 5.12: Frequency Table of satisfaction levels of Ease of Online Booking

In-flight Wifi Service					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	3908	3.0	3.0	3.0
	1	22250	17.2	17.2	20.2
	2	32236	24.9	24.9	45.1
	3	32087	24.8	24.8	69.9
	4	24702	19.1	19.1	89.0
	5	14304	11.0	11.0	100.0
	Total	129487	100.0	100.0	

Figure 5.13: Frequency Table of satisfaction levels of In-flight Wi-Fi Service

When comparing the other Services these two services have the higher cumulative percent of the first 2 (lower) levels. In every other category, this percentage is around 12%. We can't give up saying small probabilities. When comparing 100 this percentage can be small. But comparing other services this is not a good trend. The airline team has to take necessary actions for these lower satisfactions. Both of these services are related to new technology. This is more important because the new generation expects more like these technological services. Improving these services may be a good investment for the future.

6 Advanced Analysis

sklearn library of python language is used to do this advanced analysis.

1. Train Test Split module in sklearn library used to split data set in to 80% (train) and 20% (test) parts.
2. A heatmap was plotted between independent variables to check is their multicollinearity among independent variables.

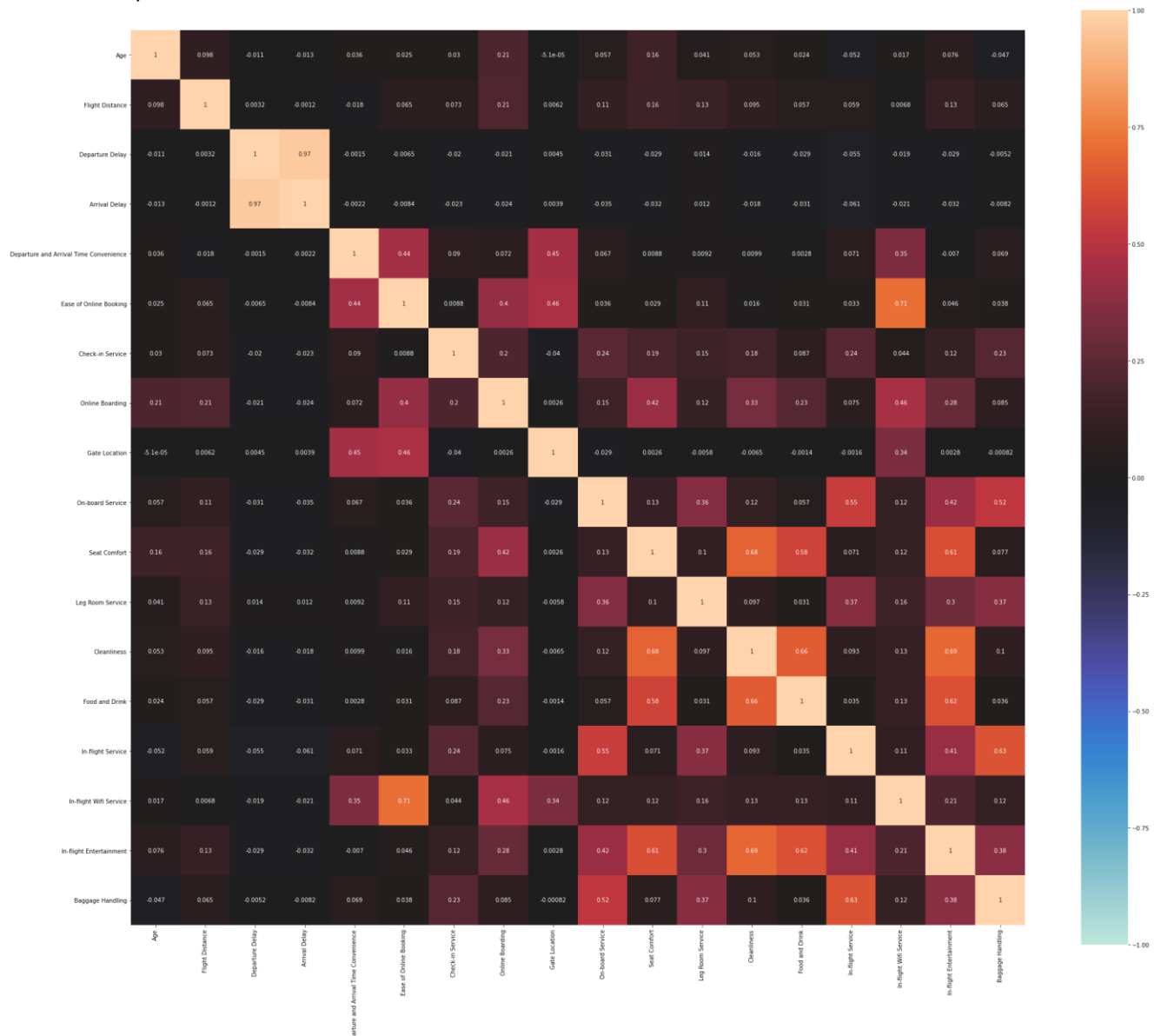


Figure 6.1: Heat-Map between independent variables.

According to this heatmap, we can see there are some moderately correlated variable pairs. Therefore, the possibility of using all independent variables was very low. Because a model selection procedure has been used.

- As a variable selection method Recursive Feature Elimination Cross Validation module was used. For all values, 5,4,3,2 with minimum features to select were given the same set of variables. So, to build the logistic regression model only the variables that were selected using the RFECV method were used.

Variables were used to build the logistic regression model:

Departure and Arrival Time Convenience(X1), Ease of Online Booking(X2), Check-in Service(X3), Online Boarding(X4), On-board Service(X5), Leg Room Service(X6), Cleanliness(X7), In-flight Service(X8), In-flight Wifi Service(X9), In-flight Entertainment(X10), Gender_Male(X11), Customer Type_Returning(X12), Type of Travel_Personal(X13), Class_Economy(X14), Class_Economy Plus(X15)

- Logistic Regression Model: -

$$\log\left(\frac{P}{1-P}\right) = -7.9259433 - 0.12311573(X1) - 0.13820771(X2) + 0.33240462(X3) \\ + 0.59572626(X4) + 0.31832882(X5) + 0.25006928(X6) \\ + 0.23345395(X7) + 0.1903167(X8) + 0.41027576(X9) \\ + 0.08358501(X10) + 0.07356658(X11) + 1.88603465(X12) \\ - 2.66679345(X13) - 0.7291348(X14) - 0.81774577(X15)$$

- After the model was fitted to check its accuracy of the model confusion matrix of the train set was obtained. And using the confusion matrix, the accuracy score of the model was calculated.

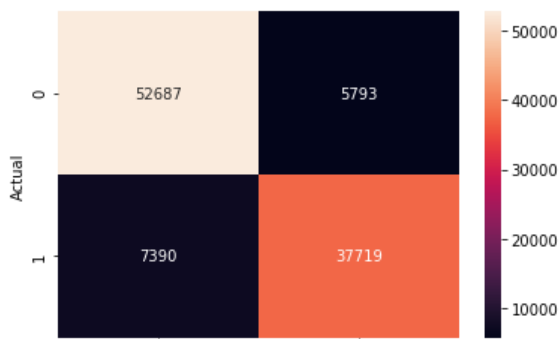
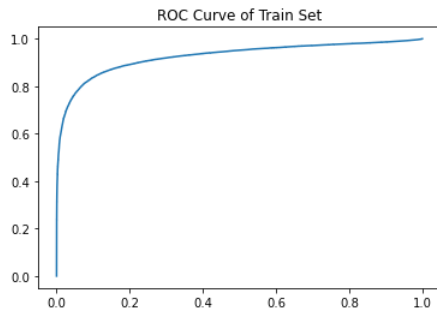


Figure 6.2: Confusion Matrix of Train Set

$$\text{Accuracy of the model} = \frac{(TP) + (TN)}{(TP) + (TN) + (FP) + (FN)} = \frac{52687 + 37719}{52687 + 37719 + 7390 + 5793} = 0.87273745$$

6. ROC curve and AUC value also calculated for model



AUC value of the model = 0.925449403

Figure 6.3: ROC curve of the model

7. Confusion Matrix, ROC Curve and AUC value of test dataset were calculated for checking the model was overfitted.

8.

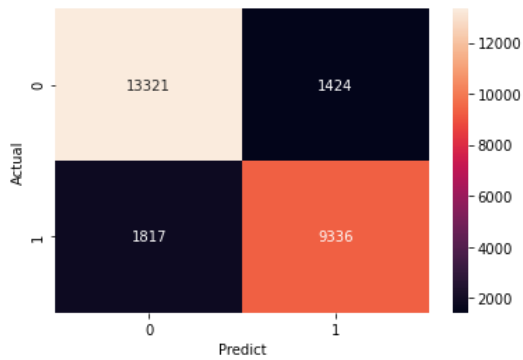


Figure 6.6: Confusion Matrix of Test Set

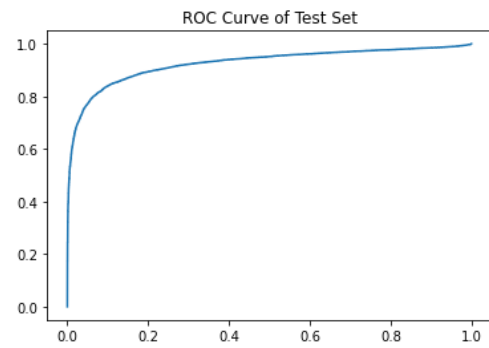


Figure 6.5: ROC Curve of the Test Set

AUC value of the test set model = 0.925995602

7 General Discussion and Conclusion

- The dataset used to analyze was balanced in many directions. According to the first pie chart, the variable satisfaction was balanced. And next one has shown that when seeing gender points of view are also balanced. And next stack bar chart shows inside of each category of the satisfaction variable was balanced under the gender variable. The fact that unbiased is a good aspect of a dataset.
- The satisfaction variable is unbiased for any category of gender variable so, using the gender variable for analyzing service satisfaction may be useful for getting more insights.
- Graphs that were included first in Satisfaction level of services with Gender variable section show us some important insides of related service.
- Satisfaction level of in-flight service increases female percentage tends to decreases. In the dissatisfied column female percentage is 60%. Finding the cause of this reason may good point for improving their airline rating. It may cause missing one important service only related to women. Doing a separate analysis for finding the reason for that will be better.
- The satisfaction level of online boarding increases the male percentage of each level decreases. For the dissatisfied column is 56% and for satisfaction level 5 (highly satisfied) 47%. There is a gradual decrease. But not a big difference as above. The cause for this may be some male goes sudden trips. They did not plan their trips for months. When they try to book seats there may be not enough seats to book as they wish. As a solution airline team can have some seats for bit higher prices in one area for a number of fixed days. (Until one day before flight departure day. It may change according to distance, destination, etc.)
- In Leg Room Service Satisfaction levels women were taken a huge percentage in level 0 (dissatisfaction). It may cause different reasons. A major reason for that may be the pregnant ladies. Pregnant women's body shape is different from that of a normal person so they may need more room space for their legs. And, than a normal person, pregnant ladies should be seated more relaxed way. And because of differently-abled people satisfaction level goes down. But gender will not be affected by that scenario. To reduce this percentage inside this plane airlines can introduce separate seats or areas for pregnant ladies. If that number of seats will not be booked, they can give those seats for a bit higher price than normal ones. And also, for differently-abled persons also introduce new areas. And for their special services a separate service provider can attach.
- In Baggage Handling service also have similar kind of trend as above mentioned with the satisfaction level go higher female percentage of each row decrease until level 4. Before taking action on this situation it is good if can be done a small survey. The aim of the survey should find if there are any special kinds of criteria mentioned by women than men when they give their baggage to airlines. According to the results of that survey, airline services can take necessary action. And also there is a positive point in baggage handling service because there is one who chooses 0th level (dissatisfied) for this service.
- For remaining services satisfaction levels, there is no special trend like the ones previously mentioned. As in the dataset most all the satisfaction levels percentage of both genders are closer to 50%.

- Logistic Regression Model was built using sklearn library in python has shown good accuracy value that is 0.87273745. AUC value (0.925449403) is also in the excellent accuracy region. Both values prove the high accuracy level of this classification model.
- With the high accuracy value of the model, doubt come to our mind as is this model overfitted. This analysis has included answers to that question. The model was built using a training dataset have given a high AUC value for the test dataset also. The difference between the AUC value of the train and test dataset is less than 0.001 so we can conclude this model was not overfitted.
- Although the sklearn library is not given the p-value of the independent variables (RFECV, Model selection method have used before fitting the model so there is no chance of selecting not significant variables to the model) there is a small chance of guessing the most important variable for the probability using the coefficients of the variables. Because almost all the variables in this model were the same type (Ordinal / Dummy variables). According to the (2) equation in the theory part when the whole value of the model increase value of the P increases.
- Accordingly online boarding and in-flight Wi-Fi service have the highest coefficients respectively. Although those variables have lower satisfaction levels they impact the Probability. So, Online Boarding Service and In-flight Wi-Fi service are the most important variables for airline passenger satisfaction.
- Departure and Arrival Time Convenience(X1), Ease of Online Booking(X2), Type of Travel_Personal(X13), Class_Economy(X14), Class_Economy Plus(X15) variables have negative coefficients. The value of that variables cause to reduce the probability of satisfaction.
- Departure and Arrival Time Convenience(X1), Ease of Online Booking(X2), Check-in Service(X3), Online Boarding(X4), On-board Service(X5), Leg Room Service(X6), Cleanliness(X7), In-flight Service(X8), In-flight Wifi Service(X9), In-flight Entertainment(X10), Gender_Male(X11), Customer Type_Returning(X12), Type of Travel_Personal(X13), Class_Economy(X14), Class_Economy Plus(X15) variables are cause to increase the probability satisfaction.

8 References

1. Hulliyah, K., 2021. Predicting Airline Passenger Satisfaction with Classification Algorithms. *IJIS: International Journal of Informatics and Information Systems*, 4(1), pp. 82-94.
2. Narkhede, S., 2018. *Towards Data Science*. [Online]
Available at: <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>
[Accessed 10 6 2022].
3. Nath, R., 2020. *medium*. [Online]
Available at: <https://medium.com/@rajwrita/logistic-regression-the-the-e8ed646e6a29>
[Accessed 10 6 2022].
4. Sezgen, E. M. K. a. M. R., 2019. Voice of airline passenger. A text mining approach to understand customer satisfaction.. *Air Transport Management*, Volume 77, pp. 65-77.
5. Sudhakar, S. a. G. S., 2020. Examining online ratings and customer satisfaction in airlines. *Anatolia*, Volume 31(2), pp. 260-273.