

A Research on Prediction of Airlines Passenger Satisfaction

Chandana Dokkadi, Kreepa Adhikari, Pragya Uprety and Someshwar Raman Suresh

School of Management, Clark University

BAN 4550: Analytics Programming

Project Proposal

Prof. Theyab Athwiti

12/06/2022

Abstract

a) Motivation of the study

The major purpose of doing this study is to examine factors that the aviation industry should consider so that they can improve the experience of their customer and increase customer satisfaction. The elements that we have observed are customer type, age, type of travel, class, flight distance, inflight Wi-Fi service, departure, convenience of online booking, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on board service, leg room service, baggage handling, check-in service, cleanliness, departure delay in minutes, and arrival delay in minutes and its relationship with customer satisfaction.

b) Method including data

We have obtained the Kaggle dataset for this study (Kaggle, 2018). The dataset contains 25 columns and 103905 rows. Logistic linear regression will be applied to develop information from the dataset. Data cleaning is done by removing the null values and the categorical data is converted into numeric data.

c) Analysis and results

The analysis is done using logistic regression and by fitting the model we get 83.64% accuracy for training data and 83.24% for the test data.

d) Conclusion

There are positive correlations that influence lower satisfaction and negative correlations that influence higher satisfaction in customers.

1. Introduction

a. Motivation of the paper

Aviation industry is one of the fastest growing industries with rapid technology advancement and high demand by customers all around the world. Airline industry connects people from all around the world. The competitive benefits of an airline in a customer-centric marketing strategy depend on how customers see its level of customer service(Gurau, 2003). After the whole world faced the COVID-19 crisis, the hospitality industry being one of the most affected, the airline industry encountered a major drop in its sales due to decreased flight schedules, flight cancellation and travel restrictions. As of now, many airlines have transitioned from crisis mode to recovery mode in 2021. According to Park et al. (2004), passenger satisfaction, service perception, service value, service expectations, and the airline's reputation are the main aspects commonly considered while modeling passengers' decision-making processes.

The aviation industry has started to pick up pace again. International flight has increased to 115.6% after August 2021. Therefore, all airline companies must focus on enhancing client satisfaction and experience to obtain a competitive edge. The purpose of this study is to determine how different airline-related characteristics connect to consumer satisfaction. Through statistical research, we will be able to identify trends and connections between airline firms' performance and customer happiness. The airline service factors: customer type, age, type of travel, class, flight distance, inflight Wi-Fi service, departure, convenience of online booking, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on board service, leg room service, baggage handling, check-in service, cleanliness, departure delay in minutes, and arrival delay in minutes will be taken into consideration as the independent variable. The dependent variable used is customer satisfaction. Therefore, to gain a competitive edge, all airline firms must concentrate on improving customer happiness and experience.

b. Specific problem under study

The paper focused on identifying factors which influence customer satisfaction in the aviation industry. Components in airlines services which includes inflight Wi-Fi service, departure,

convenience of online booking, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on board service, leg room service, baggage handling, check-in service, cleanliness, departure delay in minutes, and arrival delay are some of the major factors that can have significant impact on determining customers satisfaction in airline industry. To gain a competitive edge, all airline firms must concentrate on improving customer experience. Unhappy or disengaged customers naturally mean fewer passengers and less revenue. Hence, the major goal of the paper is to identify the gap between customer satisfaction and airline service factors.

c. Importance of the study

The significance of this paper is in its practical application to the airline sector in identifying the precise service elements that have a significant impact on customer satisfaction so that they may concentrate on making improvements. Making a wise choice on how to allocate their budget for increasing customer happiness could result in future financial success for the US airline business that is referenced in our dataset (Kaggle, 2018).

The results of this study will also enable us to answer our research questions on whether customers are satisfied, which factors they find most or least satisfying, and whether there is a difference in satisfaction between loyal and disloyal customers. The results will be helpful for future reference of the relationship between customer satisfaction and airline service factors.

Through this project, airline companies can develop a clear understanding on factors they should focus on for improving their service which eventually will result in customer satisfaction. Additionally, our project can also be beneficial to airline passengers who can get an idea of services that they should look at while choosing an airline for their use.

d. Research Questions

The research questions developed for the study are:

- To predict whether customers are satisfied with the service provided by the airline?

- To predict which factor positively influences satisfaction and which factor negatively influences satisfaction.

2. Method and Analysis

a. Data description and source

The dataset is taken from Airline passenger satisfaction data in Kaggle. It consists of both numerical and categorical data. It consists of 103905 rows and 25 columns. The dependent variable is 'Satisfaction' and the airline services like inflight wifi service, food and drink, seat comfort, on board service etc are the independent variables.

Source : <https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction>

b. Variable description

VARIABLE	VARIABLE TYPE	VARIABLE DESCRIPTION
#	Integer	Index number
Id	Integer	ID of the customers who has given their feedback
Gender	String	Gender of the passenger(Male or Female)
Customer Type	String	Whether the customer is loyal or disloyal
Age	Integer	Age of the customer
Type of travel	String	Whether the customer is traveling for a Business purpose or a personal one
Class	String	Whether the customer is traveling in a

		Business class, Economy class or Economy plus
Flight distance	Integer	The distance traveled by a customer in flight in kilometers
Inflight wifi service	Integer	Satisfaction level of a customer with Wi-Fi service on board ranging from 0-5(0: not rated).
Departure/Arrival time convenient	Integer	Departure or arrival time satisfaction level of a customer (0-5)
Ease of Online booking	Integer	online booking satisfaction rate (0-5)
Gate location	Integer	level of satisfaction with the gate location (0-5)
Food and drink	Integer	Level of satisfaction with food and drinks (0-5)
Online boarding	Integer	satisfaction level with online boarding (0-5)
Seat comfort	Integer	satisfaction level on seating (0-5)
Inflight entertainment	Integer	Level of satisfaction with inflight entertainment (0-5)
On-board service	Integer	Level of Satisfaction with on-board service (0-5)
Leg room service	Integer	Satisfaction level for leg room

		service(0-5)
Baggage handling	Integer	Level of satisfaction level with baggage handling (0-5)
Checkin service	Integer	Checkin service satisfaction level (0-5)
Inflight service	Integer	Satisfaction level with the inflight services (0-5)
Cleanliness	Integer	Satisfaction level with the cleanliness inside the airline (0-5)
Departure Delay in minutes	Integer	Minutes delayed during the time of departure
Arrival delay in minutes	Float	Minutes delayed during the time of arrival
Satisfaction	String	Airline Satisfaction level(satisfied, neutral or dissatisfied)

c. Data preparation

For this research, we have collected the dataset from Kaggle (Kaggle, 2018). It is based on the feedback given by the customers which we have classified into two types: Loyal customers and disloyal customers. It consists of 103905 and 25 columns.

The process of preparing raw data for subsequent processing and analysis is known as data preparation. The key processes are to gather, clean, and label raw data to make it appropriate for machine learning (ML) algorithms, and then to explore and visualize the data.

d. Data cleaning

After preparing the data which we collected from Kaggle, it needs to be cleaned so that it can be used for further analysis. The following columns consist of Categorical data: Satisfaction, Class, Type of travel, Customer type, Gender and are converted into numerical data (0's and 1's). The following are the steps for cleaning the data:

1. Importing the libraries : import pandas as pd , import numpy as np
2. Load the dataset using pd.read_csv
3. Find the number of unique values in columns using nunique()
4. Find the missing data: df.isnull().sum
5. Convert categorical data into numerical data

Out[239]:

	satisfaction	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	seat_comfort	Departure/Arrival time convenient	Food and drink	...	Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling
0	satisfied	Female	Loyal Customer	65	Personal Travel	Eco	265	0	0	0	...	2	3	3	0	
1	satisfied	Male	Loyal Customer	47	Personal Travel	Business	2464	0	0	0	...	2	3	4	4	
2	satisfied	Female	Loyal Customer	15	Personal Travel	Eco	2138	0	0	0	...	2	2	3	3	
3	satisfied	Female	Loyal Customer	60	Personal Travel	Eco	623	0	0	0	...	3	1	1	0	
4	satisfied	Female	Loyal Customer	70	Personal Travel	Eco	354	0	0	0	...	4	2	2	0	
...
129875	satisfied	Female	disloyal Customer	29	Personal Travel	Eco	1731	5	5	5	...	2	2	3	3	
129876	dissatisfied	Male	disloyal Customer	63	Personal Travel	Business	2087	2	3	2	...	1	3	2	3	
129877	dissatisfied	Male	disloyal Customer	69	Personal Travel	Eco	2320	3	0	3	...	2	4	4	3	
129878	dissatisfied	Male	disloyal Customer	66	Personal Travel	Eco	2450	3	2	3	...	2	3	3	2	
129879	dissatisfied	Female	disloyal Customer	38	Personal Travel	Eco	4307	3	4	3	...	3	4	5	5	

The above figure shows the dataset with categorical data.

	satisfaction	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	seat_comfort	Departure/Arrival time convenient	Food and drink	...	Online support	Ease of Online booking	On-board service	Leg room service	Baggage handling
0	1	0	0	65	1	1	265	0	0	0	...	2	3	3	0	3
1	1	1	0	47	1	0	2464	0	0	0	...	2	3	4	4	4
2	1	0	0	15	1	1	2138	0	0	0	...	2	2	3	3	4
3	1	0	0	60	1	1	623	0	0	0	...	3	1	1	0	1
4	1	0	0	70	1	1	354	0	0	0	...	4	2	2	0	2
...
129875	1	0	1	29	1	1	1731	5	5	5	...	2	2	3	3	4
129876	0	1	1	63	1	0	2087	2	3	2	...	1	3	2	3	3
129877	0	1	1	69	1	1	2320	3	0	3	...	2	4	4	3	4
129878	0	1	1	66	1	1	2450	3	2	3	...	2	3	3	2	3
129879	0	0	1	38	1	1	4307	3	4	3	...	3	4	5	5	5

This figure shows us the dataset in which the categorical data is converted into numerical data.

6. The missing data is removed using dropna()

e. Imputing missing values

Using isnull().sum() how many missing values are present can be found. Only 'Arrival Delay in Minutes' column has missing data and the null values in that column are removed using dropna().

```
satisfaction      0
Gender            0
Customer Type     0
Age              0
Type of Travel    0
Class            0
Flight Distance   0
seat_comfort      0
Departure/Arrival time convenient  0
Food and drink    0
Gate location     0
Inflight wifi service  0
Inflight entertainment  0
Online support    0
Ease of Online booking  0
On-board service  0
Leg room service  0
Baggage handling  0
Checkin service   0
Cleanliness       0
Online boarding   0
Departure Delay in Minutes  0
Arrival Delay in Minutes  393
dtype: int64
```

The new figure now shows the number of null values remaining in the dataset

```

satisfaction      0
Gender            0
Customer Type     0
Age              0
Type of Travel    0
Class            0
Flight Distance   0
seat_comfort      0
Departure/Arrival time convenient 0
Food and drink    0
Gate location     0
Inflight wifi service 0
Inflight entertainment 0
Online support    0
Ease of Online booking 0
On-board service  0
Leg room service  0
Baggage handling  0
Checkin service   0
Cleanliness       0
Online boarding   0
Departure Delay in Minutes 0
Arrival Delay in Minutes 0
dtype: int64

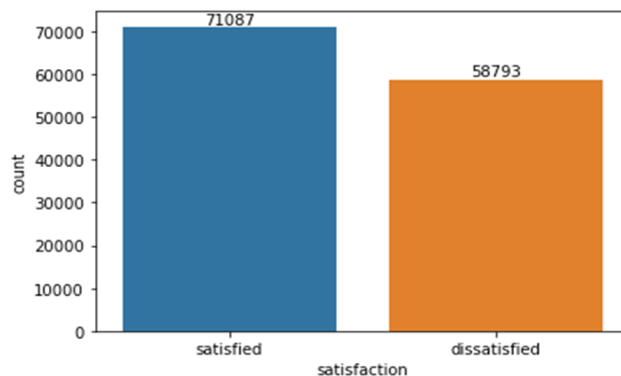
```

The figure shows us the variables in the dataset after dropping null values

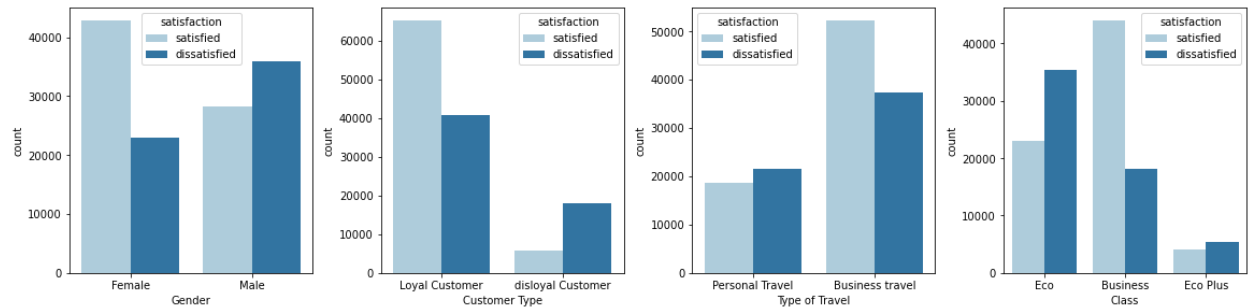
3. Results

a. Exploratory results: includes summary statistics, charts, relationships

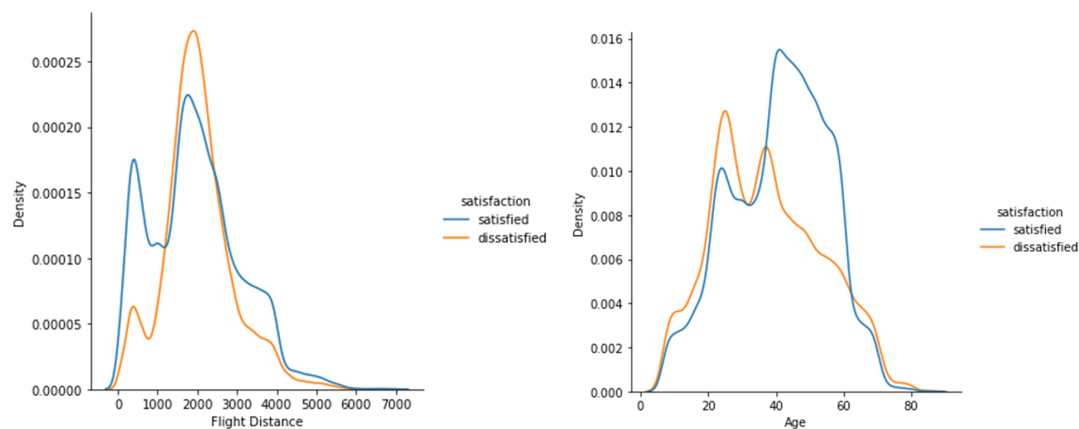
First, we begin by checking whether the customers are feeling satisfied with the service provided by the airlines or not. According to the data, provided out of the 129880 responses, 71,087 (54.73%) were satisfied with the overall services provided by the airline. Whereas 58,793(45.27%) were feeling dissatisfied with the services provided by the airlines. By this data, we can assume that more customers were feeling satisfied than dissatisfied.



Furthermore, through the graphical representations, we can find out more about the data:



From the graphical representation we can see that in relation to gender, females are more satisfied than males. Customer type graph shows a large number of loyal customers being satisfied in relation to disloyal customers. In the graph representing type of travel, we can see that there are higher number of customers travelling for business reasons who are also more satisfied. Finally, relating to the class, most customers travelled through business class and were also the population that were the most satisfied.



From the graphs above, we can see the relation between flight distance and age with satisfaction. From this we can see that people with flight distance of about 2000-3000 are more satisfied whereas the satisfaction decreases as the distance increases. On the other hand, people of age 20-40 are most satisfied whereas customers from age of 40 to 60 were most dissatisfied with the service.

b. Hypothesis testing to answer your research questions

Based on our research questions, we have the following hypothesis:

- To predict whether customers are satisfied with the service provided by the airline?

H0: The airline customers are satisfied with the airline services

H1: The airline customers are not satisfied with the airline services

Based on our data, we can conclude that the null-hypothesis is true as 54.73% of customers are satisfied with the airline services overall.

c. Modeling

We have used a logistic regression model for our modeling of our data. Since both our dependent and independent variables fall under categorical data, the logistic regression model gives us more flexibility for data profiling and prediction with the use of our binomial distribution. Much like logistic regression, linear regression tries to linearly approximate a specific function. With the use of results provided by the logistic regression we were able to determine the odds ratio via linear function of predictor i.e the probability of success and failure. In our data, satisfaction was labeled as 0 (hence making it the failure factor) and dissatisfaction was labeled as 1(making it the success factor)

We use the logistic regression method to answer our second research objective which is:

- To predict which factor positively influences satisfaction and which factor negatively influences satisfaction.

The table below shows the intercept and coefficients of the predicted satisfaction. Further analysis through logical regression shows that out of 21 factors, 11 factors had a positive relationship with the overall satisfaction, and the other 10 factors showed negative correlation with the overall satisfaction.

Intercept 0.3474139943586947		
Satisfaction Class [0 1]		
	Coefficient	Exp(coefficient)
Gender	-0.48	0.62

Customer Type	-0.80	0.45
Age	-0.12	0.89
Type of Travel	-0.42	0.65
Class	-0.31	0.73
Flight Distance	-0.10	0.90
Seat comfort	0.39	1.48
Departure/Arrival time convenient	-0.30	0.74
Food and drink	-0.31	0.73
Gate location	0.15	1.17
Inflight Wi-Fi service	-0.10	0.91
Inflight entertainment	0.94	2.56
Online support	0.12	1.13
Ease of Online booking	0.28	1.32
On-board service	0.41	1.50
Leg room service	0.29	1.34
Baggage handling	0.13	1.14
Check-in service	0.37	1.44
Cleanliness	0.09	1.09
Online boarding	0.22	1.25
Departure Delay in Minutes	0.09	1.09
Arrival Delay in Minutes	-0.29	0.75

We can see that for gender, the coefficient is -0.48. Since the sex for male is 1 and 0 is for female, it shows that being male reduces the odds of being satisfied to about 51.6% ($\exp(-0.48) = 0.62$) of the case where the sex female.

Similarly, in the case of Customer Type, loyal customers are denoted as 0 and disloyal as 1. From the intercept we can see the coefficient is -0.80 which shows that being a disloyal customer decreases the odds of being satisfied to about 45% ($\exp(-0.80) = 0.45$) of the case where the customer type is loyal.

Furthermore, looking at the other data we can interpret that:

- As the age of the customer increases, they are more likely to be satisfied. That is, with one unit increment in age, the satisfaction value will decrease by 0.12 which leads it closer to the value 0 which denotes satisfaction.
- Since the value for personal travel is 0 and business is 1, it shows that traveling for business reason decreases the odds of being satisfied to about 65% of the case where the customers are traveling for personal reason
- The value for Economy class, Business class and Economy Plus class is 0,1, and 2 respectively. This shows that being in the Economy Plus class will decrease the value of being satisfied to about 0.73% of the case where the class is Business.
- As the distance of the flight increases, the customers are more likely to be satisfied. That is, with one unit increment in flight distance, the satisfaction value will decrease by 0.10 which leads it closer to the value 0 which denotes satisfaction.
- As for the departure delay times, as the departure delay in minutes increases by one unit, the overall satisfaction also increases by 0.9 which takes it closer to 1. From this we can conclude that people are more likely to be overall dissatisfied if there is a higher delay in departure time.
- Furthermore, the increase in arrival delay in minutes by one unit will decrease the overall satisfaction by 0.29 which pushes it closer to 0. From this we can conclude that people are more likely to be satisfied even if they have more delay in arrival times.

For the data which show the satisfaction from 0-5 (0 being the least satisfied and 5 being the most satisfied)

- We can see a positive relation between factors like seat comfort, gate location, inflight entertainment, online support, ease of online booking, on-board service, leg room service, baggage handling, check-in service, cleanliness, and online boarding on overall satisfaction. This means that if the customer's satisfaction on any of these factors increases by one unit then the overall satisfaction also increases by their respective coefficient leading the overall satisfaction closer to 1 i.e. dissatisfaction.

- On the other hand, there is a negative relation between factors like convenience of departure and arrival time, food and drinks and inflight Wi-fi service with overall satisfaction. This means that if the customer's satisfaction with these services increase then the overall satisfaction will get closer to 0 making them more likely to be satisfied with the overall airline.

d. If you are predicting something, explain the model performance

Based on the results, we found that the logistic regression method was really useful in determining the relationship between the categorical dependent and independent variables. However, the factors that were analyzed were limited to only positive and negative relation and prediction of success and failure. On the other hand, more detailed analysis has to be done through visualization rather than the model itself.

4. Discussion and Conclusion

a. Interpret the results in detail and make recommendations

The results conclude that one factor alone does not determine the satisfaction or dissatisfaction of a customer. There are more factors that have negative correlation that influence the satisfaction level of the customer to a higher level and vice-versa for the ones that have positive correlation.

The airlines can focus on improving services for factors that have positive correlation in order to improve their overall satisfaction level. After research of individual factors, we see that there are more dissatisfied customers for the quality of Food & Drink, Gate location and Seat comfort. Considering these factors, the airlines can improve their services and customer satisfaction which in turn affects the Loyalty of the customers.

b. Try to tie back the results to the points discussed in the introduction section

According to the point discussed earlier, we have concluded that the percentage of satisfied is more in number than dissatisfied and the service factors influencing these have to be addressed in order to reach the ultimatum.

c. Limitations of the paper

- The dataset acquired does not provide the actual name of the airline as well as the methods in which the factors were considered.
- It does not specify if it pertains to domestic or international travel which can play a major role in accessing the region of origin.
- Majority of the service factors were rated according to satisfaction level which are categorical hence it was needed to convert in order to apply an analysis model.

d. Future work and conclusions

The airlines can look into the major factors that affect customer satisfaction in order to improve their overall customer experience. For example, customers in the age group between 40 - 60 are more dissatisfied hence coming up with serving such customers can provide better experience. It will also lead to customers returning to the same airlines which in turn leads to growth of the organization.

References

Airline Passenger Satisfaction (2018). Kaggle. Retrieved October 24, 2022, from [Passenger Satisfaction | Kaggle](#)

Gurău, C. (2003). Tailoring e-service quality through CRM. *Managing Service Quality: An International Journal*.

Park, J. W., Robertson, R., & Wu, C. L. (2004). The effect of airline service quality on passengers' behavioural intentions: a Korean case study. *Journal of Air Transport Management*, 10(6), 435-439.

Noviantoro, T., Huang J.P. (2022). Investigating airline passenger satisfaction: Data mining method. *Research in Transportation Business & Management*, 43, 100726. ISSN 2210-5395. <https://doi.org/10.1016/j.rtbm.2021.100726>.

Tsafarakis, S., Kokotas, T. & Pantouvakis, A. (2018). A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. *Journal of Air Transport Management*, 68, 61-75. ISSN 0969-6997. <https://doi.org/10.1016/j.jairtraman.2017.09.010>.

