INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 6)
Available online at: https://www.ijariit.com

And the airline is… (Airline Satisfaction Review)

Niveditha N.
nive03648@gmail.com
Coimbatore Institute of Technology, Coimbatore, Tamil Nadu

Nandhini S.
nandhinivadalur@gmail.com
Coimbatore Institute of Technology, Coimbatore, Tamil Nadu

ABSTRACT

Like how ratings are crucial in a hotel booking process, they are equally important in an airline industry. The most primal reason for such ratings is the expectation of service for the amount paid (since being the most expensive way of transport). A rating can boost a newbie up or pull a giant down. Again, it is all due to the service. Identifying the reasons and where to correct them is an example of a business model. The key objectives of this problem would be (a)To identify if changes are necessary to implement based on customer feedback. (b)To measure the modifications and minimize overall dissatisfaction. (c)Develop corresponding visualizations. And (d)To identify the satisfaction of a customer based on their feedback by fitting the appropriate model. In real time, this study can be used for reviewing the performance of the airline's front-end operations. With machine learning techniques, the user can find the best fit model. Furthermore, with the financial data, it can also be used to alter the existing model to bring better expenditure versus profit per seat. The final model should also be able to predict satisfaction based on ratings and analyse the ways to minimize dissatisfaction.

Keywords: *Visualization, Airline Industry, Rating, Business Model*

1. DATASET

The dataset describes the ratings offered over multiple customer services and the customer details. This dataset is semi-structured in nature since the factors like satisfaction, gender, class opted are in text rather than numerical. Hence, corresponding pre-processing of the data should be done.

2. INTRODUCTION

What flew earlier as a sole mode has now become a mass competition. In the earlier stages of commercial aviation, there were no competitors making very few airlines globally and only the rich flew. As soon as the jet age approached, new carriers emerged and competed with the existing airlines. Meanwhile the old carriers started ceasing due to financial loses and lack of trust (due to peak of terrorism). The jet age saw wider network through the transcontinental flights.

More airlines emerged at the digital age of aviation. It was at this period that the 'low cost' business models started emerging. Ever since, to attract customers became the primary goal of every brand.

3. FACTORS AFFECTING A COMMERCIAL AIRLINE'S OVERALL PERFORMANCE:

a) Passenger Choices
b) Destinations
c) Service plans
d) Cabin configuration
e) Delays
f) Pricing
g) Loyalty programs
h) Check-in Experience
i) Assistance provided
j) Baggage handling
k) Safety

a) Passenger Choices: This is the first factor for any airline. Their target is to pull customers through various offers and try to gain their trust.

b) Destinations: Instead of offering full term routes, an airline can switch to focus cities and seasonal routes. Further, seasonal routes can be rotated with holiday destinations and pilgrimage destinations while the focus cities being perennial.

c) Service plans: The service a passenger opts for the journey is the service plan. Airlines offer only economy (LCC) or up to four classes (Full time carriers)

d) Cabin configuration: An optimal cabin configuration will have a balance between the legroom and the number of seats. This optimality will also give satisfaction from the customer.

e) Delays: There always lies a strong correlation between the delays of a flight. Delay is a primary yet major factor that affects the opinion of an airline even if they compensated by an impeccable performance at other areas because none likes delay.

f) The correlation between the delays is up to 0.95 in this dataset.

g) Pricing: The pricing of a ticket is again, influenced by various factors like competition, parking fare, fuel surcharge and on. Full time carriers vs. low cost airlines is the crux here. Low cost airlines try to keep the charges as low as possible and use economy class throughout the configuration.

h) However, this airline has three classes (Eco, Eco Plus, Business; as seen in the dataset)

i) Loyalty programs: A loyalty program is a privilege to the customer and a badge of excellency to the airline. More customers in the loyalty program means more trust within the airline.

j) Check-in Experience: A hassle free check-in is something every flier would wish. Faster the check-in, more the hassle free it would remain. As for web check-in, smoother experience would leave a good impression.

k) Assistance provided: Assistance from the moment of stepping in to the moment of leaving the airport, some assistance will be needed. Be it the escorting or the

l) Baggage handling: A mishandling in baggage or lost baggage would make the airline to pay for the damages on one hand and dissatisfaction from the customer on the other hand.

m) Safety: The prime concern of any airline or any passenger is the safety. More than any delays, safety is the big factor. One accident can pull down the strings of that airline irrespective of its magnitude.

Pan Am was an example for how terrorism blew its lids eventually landing into bankruptcy.

3. TOOLS USED

Besides the conventional process, two extra tools were used.

1. Pandas Profiling
2. Apache Pyspark

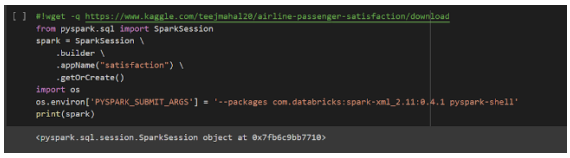
Pandas profiling is a simple tool to provide correlation plots, histograms and summary statistics.

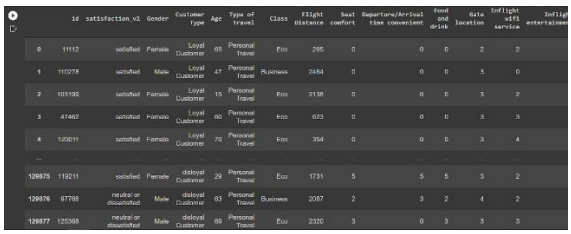
4. CONCEPTS DEPLOYED

1. Linear Regression
2. Logistic Regression
3. Decision Tree Classifier
4. Random Forest

5. CLEANING THE DATA

Considering the project and interpretations to be more generic instead of being specific, some details will be dropped keeping the ratings and satisfaction alone for the machine learning models. At first, a spark session is created and the data is imported into the work environment.

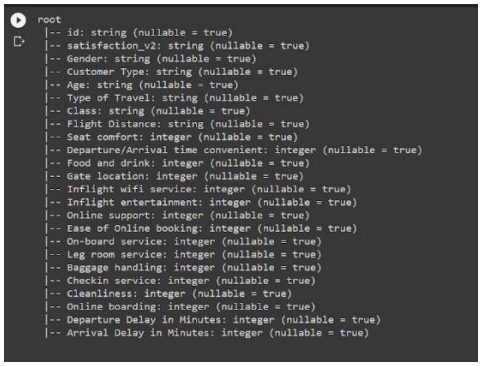




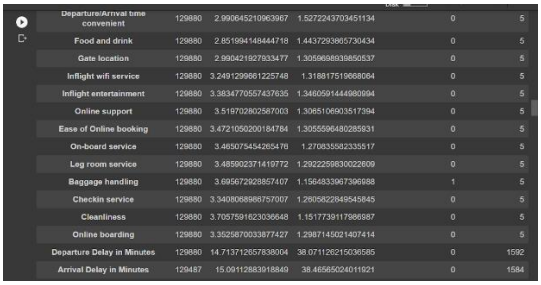
	id	satisfaction_v2	Gender	Customer type	Age	Type of Travel	Class	Flight Distance	Seat comfort	Departure/Arrival time convenient	Food and drink	Gate location	Inflight wifi service	Inflight entertainment
0	11112	satisfied	Female	Loyal Customer	65	Personal Travel	Eco	265	0	0	0	2	2	
1	110278	satisfied	Male	Loyal Customer	47	Personal Travel	Business	2464	0	0	0	3	0	
2	103199	satisfied	Female	Loyal Customer	15	Personal Travel	Eco	2138	0	0	0	3	2	
3	47462	satisfied	Female	Loyal Customer	60	Personal Travel	Eco	623	0	0	0	3	3	
4	120011	satisfied	Female	Loyal Customer	70	Personal Travel	Eco	354	0	0	0	3	4	
...														
129875	119211	satisfied	Female	disloyal Customer	29	Personal Travel	Eco	1731	5	5	5	3	2	
129876	97798	neutral or dissatisfied	Male	disloyal Customer	63	Personal Travel	Business	2087	2	3	2	4	2	
129877	125368	neutral or dissatisfied	Male	disloyal Customer	69	Personal Travel	Eco	2320	3	0	3	3	3	

On checking the schema, it is found that all columns take the string value.

On conversion of the ratings into int datatype, the schema obtained will look like this, with string and integer being two datatypes.

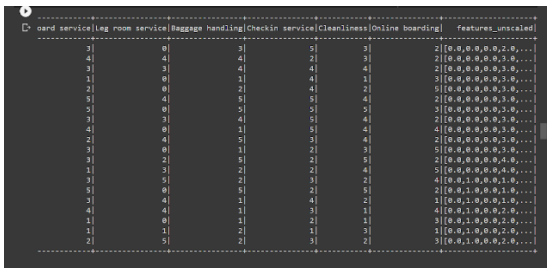


EDA (Exploratory Data Analysis) is carried out with the newly modified data and it is found that the average ratings lie between 2.5 – 3.7

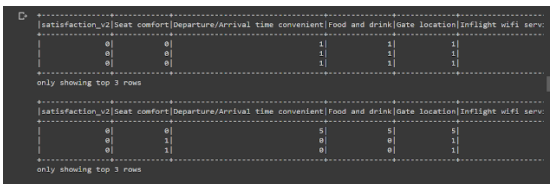


Departure/Arrival time convenient	129880	2.990645210963967	1.5272243703451134	0	5
Food and drink	129880	2.851994148444718	1.4437293805730434	0	5
Gate location	129880	2.990421927933477	1.3059698938850537	0	5
Inflight wifi service	129880	3.2491299661225748	1.318817519668064	0	5
Inflight entertainment	129880	3.3834770557437635	1.3460591444980994	0	5
Online support	129880	3.519702802587003	1.3065106903517394	0	5
Ease of Online booking	129880	3.472105020018476	1.3055596480285931	0	5
On-board service	129880	3.465075454265476	1.270835582335517	0	5
Leg room service	129880	3.485902371419772	1.2922259830022609	0	5
Baggage handling	129880	3.695672928857407	1.1564833967396988	1	5
Checkin service	129880	3.3408068986757007	1.2805822849545845	0	5
Cleanliness	129880	3.7057591623036648	1.1517739117986887	0	5
Online boarding	129880	3.3525870033877427	1.2987145021407414	0	5
Departure Delay in Minutes	129880	14.713712657838004	38.071126215036585	0	1592
Arrival Delay in Minutes	129487	15.091128839188849	38.46565024011921	0	1584

Making the data to be more generic, only the ratings and satisfaction are taken in and the rest are discarded. At the end, a vector is created by combining the ratings. The mapping would be Satisfaction → (Features_unscaled) and look like this.



After all pre-processing, the dataset is taken to the final step: splitting into train and test for machine learning algorithms. Here the data is separated in the ratio of 7:3 in random and the first the rows of both splits are displayed.



6. THE CASES

6.1 Machine Learning Models

i) Linear regression: The linear regression is applied to the new dataset and the accuracy is tested. The accuracy of the model turned to be 3.7% which is very poor in nature.

The following are the obtained values:

```
Intercept: 0.5004716050702425
RMSE: 0.488468
r2: 0.037239
Accuracy--> r2*100
```

With the coefficients almost zero and intercept being 0.5, which is contradiction to the mean.

Hence, the model is declared unfit and is ruled out.

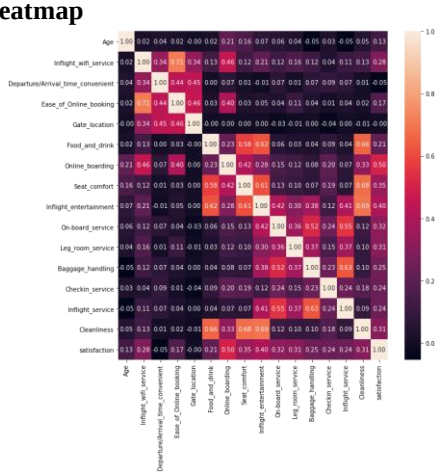
ii) **Logistic regression:** The accuracy obtained from this model is approximately 80%

iii) **Decision tree classifier:** The accuracy obtained from this model is 85% which makes this model much suitable for this dataset. Decision tree helps one to find the exact path of process.

iv) **Random Forest:** The accuracy obtained from this model is 82% which makes this model also suitable for this dataset. Random Forest has nearly same parameters as decision tree classifier and adds additional randomness while growing trees.

6.2 Visuals

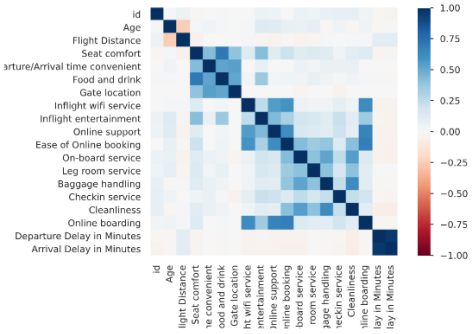
i) Heatmap



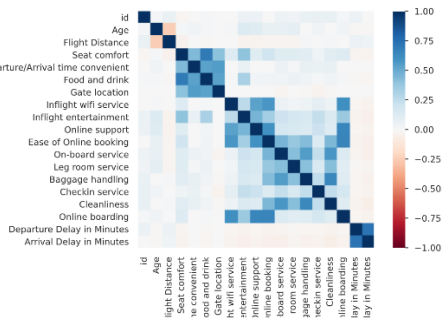
From the heatmap, we can conclude the following:

- Gate location is irrelevant to the onboard services but shares a little importance with online booking and arrival/ departure time convenience.
- Cleanliness of the cabin, food, seat comfort and inflight entertainment are correlated with one another.
- Inflight service and online booking are highly correlated.
- Age is irrelevant for satisfaction but is related with inflight entertainment, meaning that a variety of channels across different ages is highly appreciable.

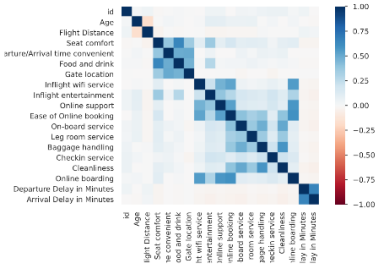
a) Heatmap from pandas-profiling:



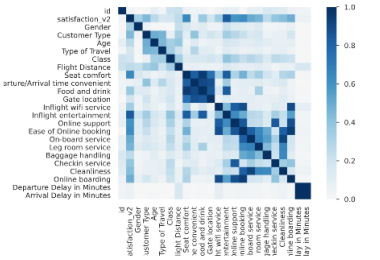
Pearson



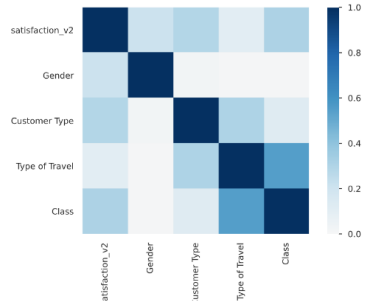
Spearman



Kendall

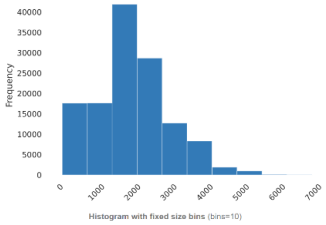


Phik



Cramer

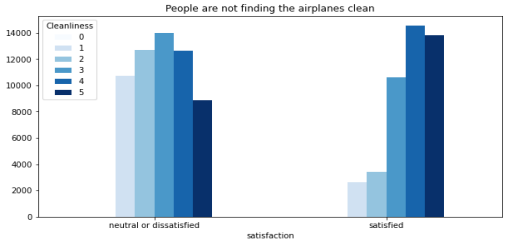
ii) Histogram:

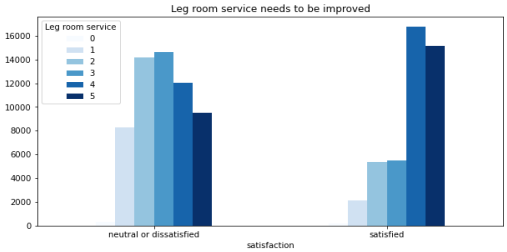


The distance histogram

The distance histogram can suggest the airline to focus on the route ranging between 1500 km and 2500 km since most passengers prefer it and on the other hand, the make good short haul routes.

iii) Rating Bar graphs:





7. CONCLUSIONS

1. It can be concluded that this airline X is a mediocre brand with average ratings.
2. Cabin service plays an important role than the rest.
3. The airline can focus on inflight entertainment and cleanliness of the cabin.
4. With business purpose scoring a majority, the airline can focus the metropolitan cities and cut the leisure destinations.
5. Improve the leg room and cabin cleanliness.

8. REFERENCES

[1] https://www.researchgate.net/publication/306050083_The_Effect_of_the_Cabin_Service_Quality_on_Customer_Loyalty_and_Airline_Image

