Ukrainian Catholic University

Faculty of Applied Sciences

Business Analytics & Computer Science Programmes

Econometrics
2nd interim report

Beyond the clouds

Team:
Daryna Kuzyshyn
Olha Havryliuk
Anastasiia Dynia

8th of April 2024

1. Introduction

In today's world, what customers say online can really make or break a brand, and this is especially true for airlines. The airline business is all about providing great service and making sure passengers are happy. But as the world changes, especially with everyone using the internet more, airlines have to keep up not just with flying planes but also with what passengers expect and say about them online. Passengers" reviews and stories can now reach people all over the world, influencing others to choose or avoid an airline. Understanding why passengers recommend an airline is now key for airlines to get better and keep their customers satisfied.

In this project, the authors are diving deep into what passengers say in their reviews to figure out what really matters to them. It is essential to carefully look at the reviews to see what parts of the flight service stand out, either in a good or bad way. The authors are interested in everything from how comfortable the seats are to how friendly the flight attendants are.

Also, the authors will compare different airlines to see how they stack up according to passenger reviews. This will help to see what some airlines are doing right and where others might need to improve.

An important part of the project is creating a model that can predict if a passenger would recommend an airline based on certain things we learn from the reviews. This could help airlines understand what makes a passenger's experience good enough to recommend their flight.

2. The aim of the project

The main goal is to help airlines understand better what makes passengers more likely to recommend their flights. By looking closely at passenger reviews, the authors hope to find out what makes a flight experience great. This is important because it can help airlines to improve based on what their customers say. The findings of this project could make flying a better experience for everyone involved.

3. Data description

The project utilizes a dataset that is compiled from an online airline review platform - https://www.airlinequality.com, it provides a diverse range of passenger feedback on various airlines and flight experiences. This dataset has detailed reviews that are spanned from January 2019 to December 2023, offering an extensive view of passenger satisfaction and service quality in the airline industry during this period. After data preprocessing, there are left more than 5700 review entries out of 128000 that where at the beginning, these data provide a complete and full foundation for the future analysis.

Key variables from the dataset:

- Aircraft: the name of the aircraft being reviewed.
- AirlineName: the name of the airline being reviewed. This is particularly useful for questions related to comparing airlines based on insights from passenger reviews.

- CabinType: the class of service used by the reviewer (Economy, Business, First Class). This variable helps in understanding how service experience varies across different classes.

- DateFlown: the month and year of the flight. The flight date can provide insights into seasonal variations in passenger satisfaction.

- EntertainmentRating, FoodRating, GroundServiceRating, SeatComfortRating, ServiceRating, ValueRating, WifiRating: passenger ratings for each of these aspects on a scale from 1 to 5. These specific ratings allow for a detailed analysis of what factors are most important to passengers when giving a review.

- TripVerified: a boolean variable indicating whether the review is verified.

- OriginCountry: the country from which the flight originated.

- OverallScore: an overall score for the flight experience on a scale from 1 to 10. As a comprehensive measure of passenger satisfaction, this variable is central to almost all analyses that will be made.

- Recommended: a boolean variable indicating whether a passenger recommends the airline ("yes" or "no"). This binary variable is key to our prediction model, aiming to determine what factors most influence a passenger's likelihood to recommend an airline.

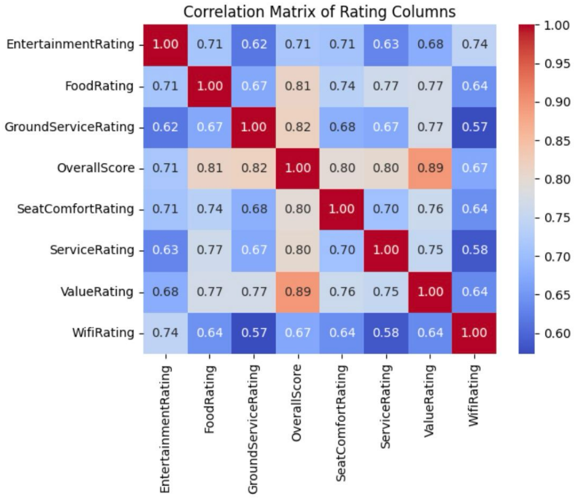
- Review: textual review of the flight. This is especially relevant for generating a word cloud to see what words are most important for people when giving reviews.

- Route: flight route.

- TravelType: the type of travel (e.g., Solo Leisure, Business). Different travel types can have distinct expectations and experiences. This variable helps in examining how service perceptions vary among different passenger segments.

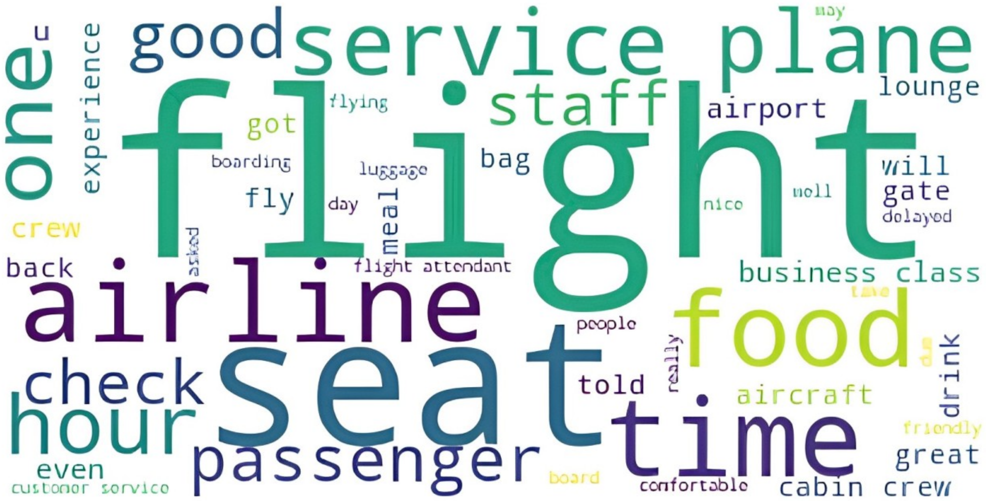
- TripVerified: whether the trip was verified.

Let's move to plotting **a correlation matrix**. It plays an important role by revealing potential connections between different variables. It guides this investigation, showing where strong links exist and helping to identify which features might not be relevant to our analysis.



The OverallScore has high correlations with all the individual service ratings, especially with FoodRating and SeatComfortRating. This indicates that passengers' overall satisfaction is heavily influenced by their experience with the airline's food and seat comfort. However it is interesting that WifiRating have the weakest correlation with OverallScore. This might indicate that while WiFi service contributes to overall satisfaction, it might not be as crucial as other factors like food and seat comfort.

During the initial analysis of the data, the authors became interested in which words people use the most when giving feedback on a flight. In authors' opinion, the word could help to understand what people pay most attention to.



From this word cloud, it can be stated that people pay attention to the following things:

- **"Flight" & "Airline":** these central terms are expected as they are primary focus of the reviews.

- The word **"service"** being so central indicates that the quality of service is a common thread in passenger feedback. **"Friendly"** and **"staff"** suggest that the the airline staff is an important factor in the reviews, likely contributing to both the "ServiceRating" and "OverallScore".

- Words like **"seat"** and **"comfortable"** suggest that the physical comfort during the flight is a significant concern for passengers, that directly impacts their satisfaction. Additionally, the presence of **"business class"** signals that the additional comforts and services provided in higher travel classes are important to reviewers.

- The visibility of **"food"** in the cloud shows the importance of catering services as a component of the in-flight experience, correlating with the "FoodRating". The term **"entertainment"** could also appear, but it is not as significant as "food", this might suggest that food is a higher priority than entertainment for the passengers providing reviews.

- The appearance of **"time"** and **"hour"** likely relates to the punctuality of flights or the duration of services, such as check-in and boarding processes. This may affect "GroundServiceRating" and could influence the "OverallScore" and whether passengers recommend the airline.

4. Data analysis & Methods

After having the data visually represented, the next step is to analyze the main questions and to derive conclusions. Let's start with determining what impact have rating variables on the overall score of customers' flights. For this the authors ran the multiple regression model with OverallScore as a dependent variable. Based on the results of the correlation matrix, the authors concluded that all the variables are significant enough to include them into the model. The model is as follow:

$$\text{OverallScore} = -1{,}8186{\cdot}\text{const}+0{,}2438{\cdot}\text{SeatComfortRating}+0{,}2745{\cdot}\text{FoodRating}+0{,}3270{\cdot}\text{ServiceRating}$$

$$+\ 0.8635\ {\cdot}\text{ValueRating} + 0{,}0852 \cdot \text{WifiRating} + 0{,}0566 \cdot \text{EntertainmentRating} +$$

$$0.4778\ {\cdot}\text{GroundServiceRating}$$

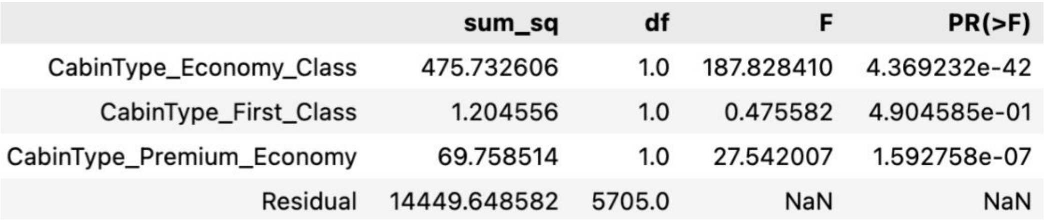
$$R^2 = 0{,}882$$

As it can be seen the R-squared value is really high, which indicates a strong fit of the model. All the independent variables have positive coefficients that are statistically significant (p-values is less than 0.05), it indicates that improvements in these ratings are associated with higher OverallScores.

During the work on this project, the authors came up with the hypothesis that the CabinType may affect the score of ServiceRating, which also has an impact on the Overall Score. To explore this impact, such dummy variables were created: CabinType_Business Class, CabinType_Economy Class, CabinType_First Class, CabinType_Premium Economy for CabinType with a CabinType_Business Class as a base group. For this, the $F-test$ was performed, testing such hypotheses:

H_0: CabinType affects the score of the ServiceRating
H_1: there is no observed impact of the CabinType on the ServiceRating

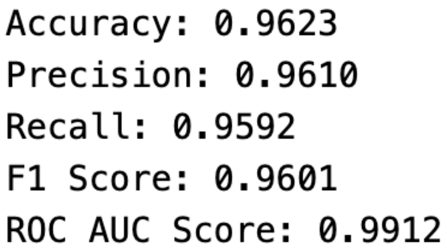
The results are presented on the following image:



	sum_sq	df	F	PR(>F)
CabinType_Economy_Class	475.732606	1.0	187.828410	4.369232e-42
CabinType_First_Class	1.204556	1.0	0.475582	4.904585e-01
CabinType_Premium_Economy	69.758514	1.0	27.542007	1.592758e-07
Residual	14449.648582	5705.0	NaN	NaN

The sum of squares for the Economy Class is relatively high, which indicates that there is a considerable amount of variability in service ratings that can be explained by whether a passenger is in Economy Class. This implies that the service rating is significantly different for passengers in Economy Class compared to the baseline category. The sum of squares for Premium Economy is lower than for Economy Class but still notable (69.76). The sum of squares for First Class is small (1.20), suggesting that it does not account for much variability in service ratings. **To sum up**, the service rating is significantly affected by whether a passenger is flying in Economy Class or Premium Economy, with Economy Class having a particularly strong effect. There's no significant effect of flying in First Class on service rating compared to the baseline cabin type. These results could inform airlines that service expectations and perceptions differ significantly across different cabin types, particularly between Economy and higher-class services.

Now, the last step is to build a model that will allow to predict whether a customer recommends an airline or not. For this we will use Logit model. We selected the model using **the AIC technique**, with dropped variables *"CabinType_Economy_Class, EntertainmentRating"*. All Service Ratings have positive coefficients that are statistically significant, and indicates that higher ratings in these areas are associated with a higher likelihood of recommending the airline. Also it is important to mention that Premium Economy type has the negative coefficient (-0.7227) that is statistically significant (p = 0.030). It suggests that passengers in Premium Economy are less likely to recommend the airline compared to those that travel in Business Class. We used obtained model to predict passengers likelihood to recommend their flight and we got quite a good results.



```
Accuracy: 0.9623
Precision: 0.9610
Recall: 0.9592
F1 Score: 0.9601
ROC AUC Score: 0.9912
```

These metrics together suggest that the logit model is performing very well in terms of predicting recommendations, with high reliability.

5. Conclusions

In conclusion, our project has effectively applied regression analysis to uncover key drivers of airline passenger satisfaction. The results strongly suggest that factors like seat comfort and food quality are crucial in influencing passengers' likelihood to recommend an airline. While cabin class was not a significant predictor, the value offered by an airline was found to be a decisive factor. The logistic regression model stood out for its accuracy, making it a valuable tool for airlines to predict and enhance customer satisfaction.