The Warsaw Econometric Challenge

Second Edition

Prediction of an upgrade of a Class in LOT flights

Karol Augustyniak

Rafał Nojek

Filip Szempruch

Oleg Tkachenko

Cracow, the 20th of March 2022

INTRODUCTION

Class selection and passenger requirements vary widely and depend on several factors. In many cases people are not fully aware of all the conveniences that are offered in specific classes by the airline company - LOT. The wide range of ways in which we communicate with customers and the ability to collect extensive data sets allow us to analyze passenger behavior and preferences. This has always been an important topic for airline companies and in the world of economics, for instance: (Modelling choice of flight and booking class – a study using Stated Preference and Revealed Preference data) (Staffan Algers and Muriel Beser, 2001).

The hardest part of this content was to manage data in a proper way. The main issues were in the amount of observations and in connecting the databases. After the preprocessing of our dataset we achieved consecutive numbers - there were over 9 000 000 [9 277 193] passengers and only 27311 people decided to upgrade their flight’s choice. Which means that only 0.3% of the customers upgraded their class.

HYPOTHESIS

Hypothesis 1: Is there a dependence between choosing a long-haul flight and making an upgrade to a higher class? Customers can take more advantage of higher class amenities when traveling long-haul.

Hypothesis 2: Is the impact of summer significant and has a positive effect in the model? During summer-time people are claimed to take more leaves and travel choosing long-haul destinations, therefore they are believed to upgrade to a higher class.

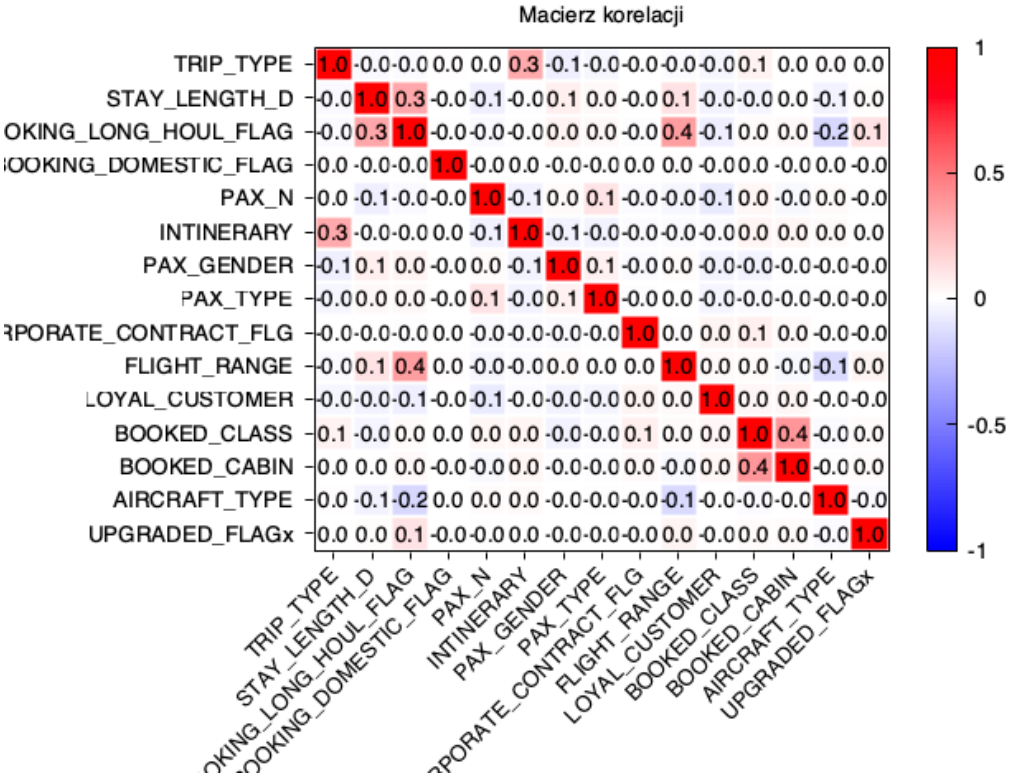
Hypothesis 3: Are business travellers more likely to upgrade to a higher class? Business people often rely on comfort and convenience to be well-rested for upcoming meetings.

Hypothesis 4: Are flights with high itinerary-rate less likely to upgrade their choice? If there are lots of stopover passengers will not decide to choose a higher class.

We hope that this article will be very helpful for LOT S.A. to improve their marketing proposition. Our analysis is focused on the passenger who had decided to upgrade to a class, thus we can find the most crucial variables that describe preferably if a passenger is going to make up his mind once again.

In the beginning, it is very important to decide what kind of model we should use to make the best estimation. This is the reason why we made a correlation matrix, to check if there exist linear relations.

Figure 1. Correlation between the variables used in the research.



Source: Own calculations.

As we may see in the picture above, there is no significant correlation in this dataset. This is the main reason we decided to not use any of the linear models such as linear regression etc.

EMPIRICAL ANALYSIS

HYPOTHESIS TESTING

Because of the huge disparity between those who expressed a desire to upgrade their ticket (0.3% of tickets in the entire dataset) and those who did not, we decided to focus on passengers who would tend to upgrade their tickets.

No 1.

In this hypothesis we have tested whether there is a dependence between choosing a long-haul flight and making an upgrade to a higher class. To verify it we used Chi-Square Test of Independence. Table of values that shows frequency counts of selected variables is shown below.

UPGRADED_FLAG.y	N	Y
BOOKING_LONG_HOUL_FLAG		
N	375299	24
Y	123811	866

stat=2490.75, p≈0.0

Probably dependent

The P-value estimated in the test is less than 0.05, so we reject the null hypothesis. So we may interpret it that these variables are dependent on each other.

No 2.

In this hypothesis, we test whether all summer months have an impact on long-term flights. Since the data are categorical, we use the chi-square correlation test

```
UPGRADED_FLAG.y
N  Y
BOOKING_DEPARTURE_TIME_UTC
2007-06-01 00:40:00
56  0
2007-06-01 01:55:00
3  0
2007-06-01 02:00:00
34  0
2007-06-01 02:05:00
56  0
2007-06-01 02:30:00
75  2
...
.. ..
2007-08-31 21:30:00
10  0
2007-08-31 22:25:00
22  0
```

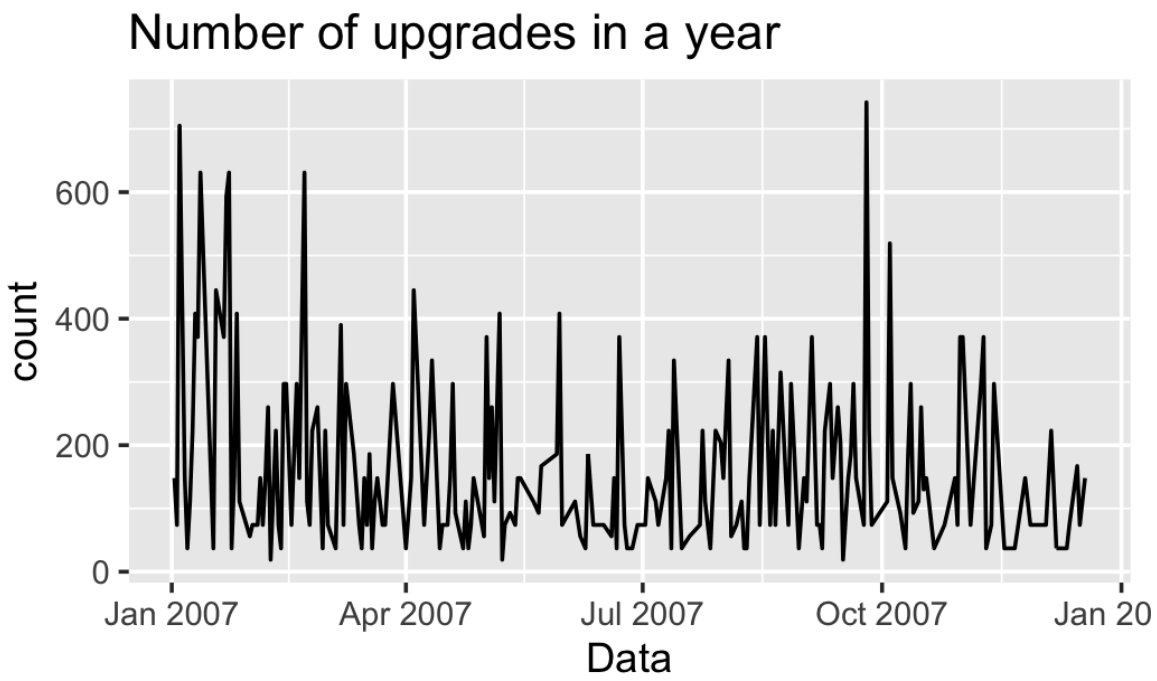
```
2007-08-31 23:05:00

31  0

2007-08-31 23:20:00

1  0

2007-08-31 23:25:00

35  1


[8891 rows x 2 columns]

stat=19396.29, p≈0.0


Probably dependent
```

The P-value estimated in the test is less than 0.05, so we reject the null hypothesis. So we may interpret it that these variables are dependent on each other.

No 3.

Time signature of the number of upgrades over one year

Figure 2. Plot of number of upgrades in a year.



Source: Own calculations.

The figure below shows us that in the summertime there number of people upgrading their choices is average. Indeed, we may also see the two picks in January and October. Thus there is no point in checking the hypothesis.

No 4.

By using the itinerary variable and chi-squared test we are going to check the influence of this variable on the final result. First of all the data was collected and organized in a table.

Table 1. Frequency of upgrades and not upgrades per number of itineraries.

Number of itinerary	Not upgraded	Upgraded
0	702935	430
1	2104507	4642
2	90598	550
3	947855	3442
4	69792	757
5	16841	114
6	1647	23
7	366	5

Source: Own calculations.

Secondly we conducted a chi2-test to check our hypothesis.

Pearson's Chi-squared test

data: table

X-squared = 8243.8, df = 7, p-value <

0.00000000000000022

As we may see on the result of the test. the p-value is less than 0.05, so we reject the null hypothesis. So we may interpret it that these variables are dependent on each other.

XGBoost Model

To analyze the research problem we applied the XGBoost model for classification (Tianqi Chen et al., 2016). The fact that our variables are nonlinear made it necessary to find a model that could handle nonlinear interactions in the data.

Our XGBoost modelling was relatively complex. We searched for the best hyperparameters in a grid search with the following setup:

```
'gamma': [0,0.1,0.2,0.4,0.8,1.6,3.2,6.4,12.8,25.6,51.2,102.4,
200],
'learning_rate': [0.01, 0.03, 0.06, 0.1, 0.15, 0.2, 0.25,
0.300000012, 0.4, 0.5, 0.6, 0.7],
'max_depth': [5,6,7,8,9,10,11,12,13,14],
```

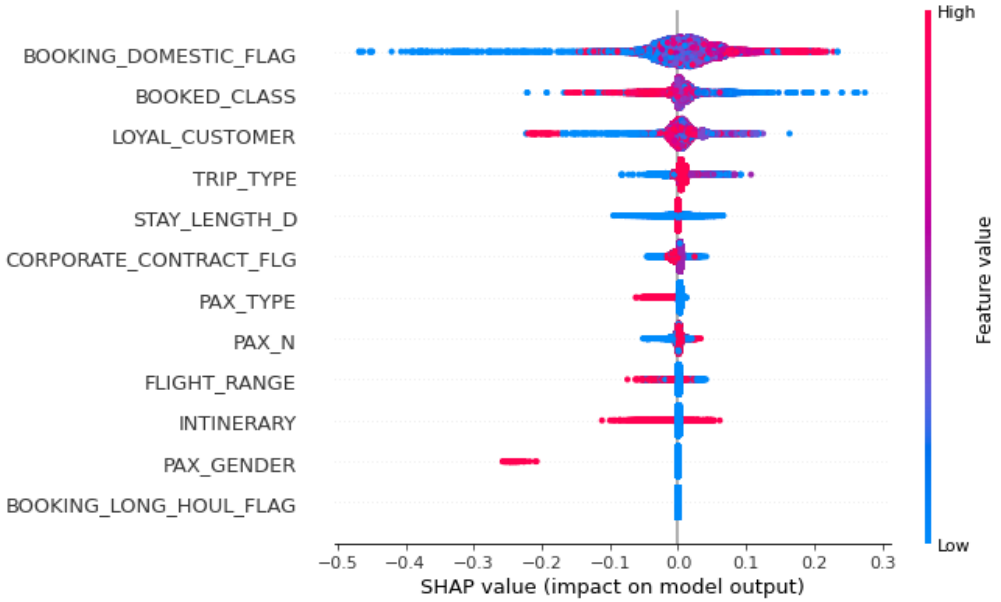
```
'n_estimators': [50,65,80,100,115,130,150],
'reg_alpha':
[0,0.1,0.2,0.4,0.8,1.6,3.2,6.4,12.8,25.6,51.2,102.4,200],
'reg_lambda':
[0,0.1,0.2,0.4,0.8,1.6,3.2,6.4,12.8,25.6,51.2,102.4,200]
```

Due to the equal selection of each category in the dependent variable, we chose to evaluate our metric in logistic regression for binary classification. Feature significance techniques allow us to analyze the significance of a given variable in the overall model and determine its quasi-share in the predictive power of the model. SHAP (SHapley Additive exPlanations) is a game theory-based approach for explaining the output of any machine learning model. Since we focused on summarizing the effects of all features, we used the SHAP summary table.

Results

The best model from our grid-search exploration has specified hyperparameters: 'gamma': 0.4, 'learning_rate': 0.5, 'max_depth': 10, 'n_estimators': 100, 'reg_alpha': 0, 'reg_lambda': 1.6. After applying the general approach to the detailed approach, the final fixed effects model obtained has 12 independent variables. The overall model results obtained using the SHAP Summary Plot are shown in Figure **(here place the number)**.

Figure 3. SHAP Summary Plot based on XGBoost model



Source: Own calculations

Let us first analyze the impact of this type of trip. We can easily conclude that passengers who flew domestic trips are more likely to upgrade their tickets. It can be explained that time-sensitive people want more convenience during their trip. In the case of booked class, as we expected, passengers with basic tickets are more likely to purchase an upgrade compared to people in business class. Additionally, the flights with high itinerary-rate are more likely to purchase an upgrade. The other variables show little effect on the final result.

To sum up, the model confirmed hypothesis 1 and 4 and rejected hypothesis 3.

Conclusion

The models confirmed most of the research questions posed at the beginning. Due to the small percentage of people redeeming upgrades, we mainly focused on them so as not to leave out any person. The first limitation of this study is a small percentage of people who upgraded their tickets. As a result we had to reduce the disproportion of the data, so here may be some miscalculations. Additionally, the features were linearly independent so we had to choose one of some black-box thinking models. We very much hope that our work will help LOT S.A. company to make strategic business decisions in the future. Moreover, we strongly believe that improvement of the dataset we were working on could enhance the results significantly and help our article achieve better conclusions.