COURSERA IBM MACHINE LEARNING –CLASSIFICATION – COURSE PROJECT

Author: Wojciech Gołębiowski

1. Main objective of the analysis

Online room reservations are constantly changing booking processes. Despite of the offer which is being constantly expanding the phenomena of ‘no-show’ at the hotel is still a big struggle for many chains acting in this branch of business. It is a result of very easy option of cancellation without being charged. As this seems to be a very attractive option for customers at the same time it is a big problem for hotel itself which needs to deal with a higher number of unexpected behaviours. The question is what usually features the situation in which client decides to cancel their reservation. Here comes a power of Machine Learning and especially the classification algorithms. Therefore a main objective of this analysis is to predict whether a given type of reservation is likely to be cancelled.

1. Brief description of the data

Dataset used in analysis presents 19 attributes of 36275 room reservations which are described below:

* Booking Id
* Number of Adults
* Number of children
* Number of weekend nights
* Number of week nights
* Type of meal plan
* Required Car parking space
* Room Type reserved
* Lead time
* Arrival year
* Arrival month
* Arrival date
* Market segment type
* Repeated guest
* Number of previous cancellations
* Number of previous bookings not cancelled
* Average price per room
* Number of special requests
* Booking status (target variable)

There are 15 numeric variables and 4 categorical variables. This dataset was part of the study developed by Nuno Antonio ,Ana de Almeida and Luis Nunes which was published in Hotel booking demand datasets, Data in Brief, Volume 22, 2019. With this analysis I try to predict whether a reservations meeting a specified criteria is likely to be cancelled. My classification algorithm will be capable of quickly detecting potential ‘risky’ reservations.

In the below tables I present a short summary of main statistics for each of the analysed variables.

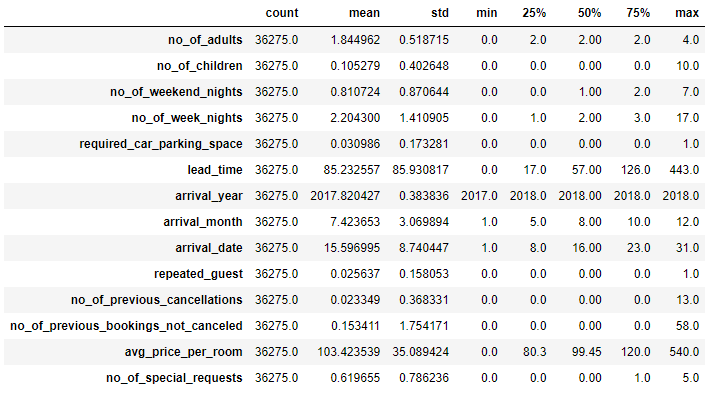


Figure 1. Main statistics for numerical variables

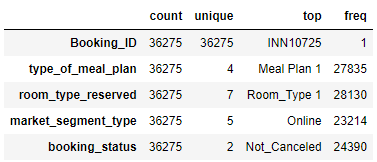


Figure 2. Main statistics for categorical variables

Data presented in figure 1 points to the fact that the data was collected at the ten of years 2017 and 2018. Most of the reservations didn’t include children (75th percentile equal 0), the same pattern follows the other categories like required car parking space of repeated guests. We can say that typical reservations didn’t meet these criterias. The average duration of stay was about 3 days (combining the average for both week and weekend days). The average lead time equals 85 days whereas the median for this category is 57 days. We can suspect that in the dataset there might be a few reservations which were booked far in advance.

Figure 2 shows some additional insights on categorical variables. The most interesting one is our target which presents quite positive pattern – most of the reservations in the dataset were not cancelled. For the other categories the most typical observations were meal plan1 for type of meal plan; room type 1 for room type reserved and online market segment type.

1. Brief summary of data exploration and actions taken for data cleaning and feature engineering

The plan of data exploration will contain the following steps:

* Searching for duplicates
* Examining missing data and potential data imputation process
* Feature engineering
* Examining outliers

First step was to look for duplicates. It turns out that dataset doesn’t contain any duplicated value.

Second action was to check how many missing values has each of the analysed variables. Nonetheless it turns out that there is no missing values in the whole dataset.

There is a big scatter of the brands. As it may not make sense to include such a variable in the model I will remove it from analysis – there is a lot of categories which are not significantly represented in the data set which may lead to some problems in later stage of analysis (after implementation of one hot encoding this may create a big number of variables which will not bring any additional information).

Third step included feature engineering process. First I excluded reservation ID from the analysis as this variable doesn’t bring any additional information. The next step was to analyse the distribution for the rest of the variables. If the vast majority of observations would pop into only one of the categories I decide to remove this variable from further analysis as the difference in single cases wouldn’t bring a lot of new information into the analysis. There are some variables which are dominated by one of the categories. The following figures present their distributions. These variables will be removed from the analysis:

* Required car parking space
* Repeated guest
* Number of previous cancellations
* Number of previous reservations not cancelled

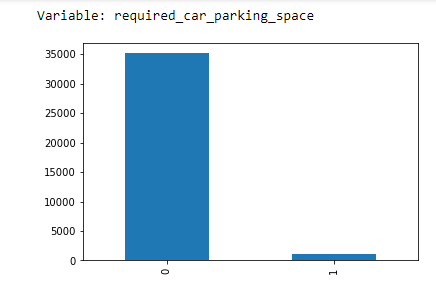


Figure 3. Distribution of Required car parking space variable

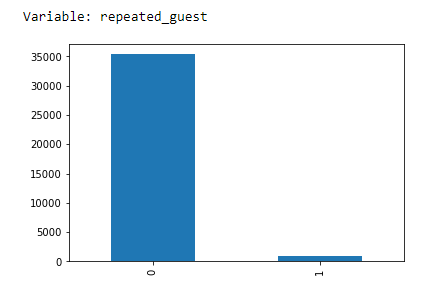


Figure 4. Distribution of Repeated guest variable

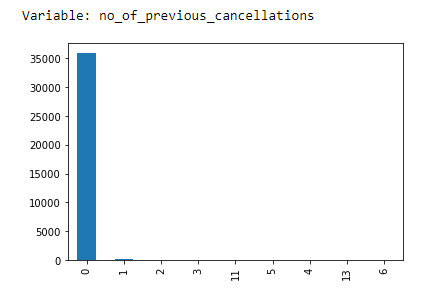


Figure 5. Distribution of Number of previous cancellations variable

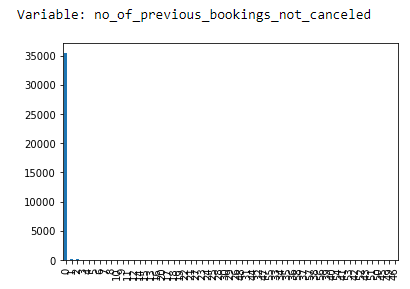


Figure 6. Distribution of Number of previous bookings not cancelled

Additionally all of the time variables (year, month, day) were removed. Nonetheless in the place of the month variable I put the season variable. It might be a good predictor of cancellations, as for example in the winter the unexpected changes in weather conditions might lead people to reservation cancellations. After the specified changes the dataset contains 12 columns.

Lastly, I examined whether there are any outliers for two variables: lead time and average price. For this purpose I used boxplots which are presented below.

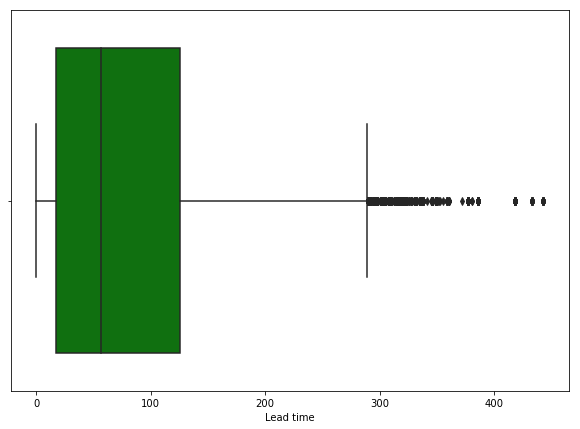


Figure 7. Boxplot for Lead time variable

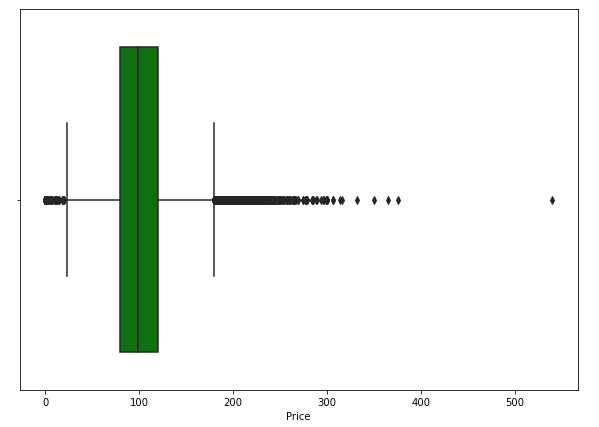


Figure 8. Boxplot for Price variable

Both variables contain observations which might be perceived as outliers. Nonetheless I decided to keep all observations in the dataset as none of them seems to be an error and all of them on the other hand might bring some additional value into the analysis (assumptions being reservation with higher lead time is more likely to be cancelled and reservation with higher price is more likely to be cancelled).

1. Summary of training different classifier models

First model that was trained was Logistic Regression. To do so, first I created some transformations:

* Convert target variable to binary
* Standardize all numerical variables
* Convert categorical variables into binary classes

30% (10883) of observations were taken into testing and 70% for a testing (around 25392 records). I used Grid Search feature to select the best hyperparameters for this model. Then the model evaluation was performed. Below pictures presents the key metrics’ results for logistic regression model.

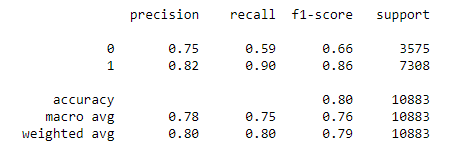


Figure 9. Evaluation metrics for Logistic Regression

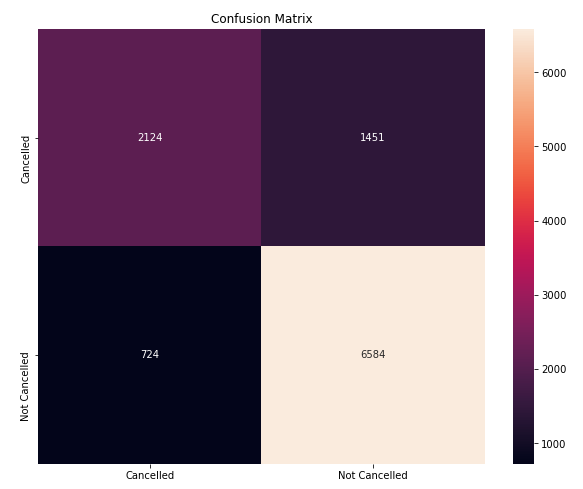


Figure 10. Confusion matrix for Logistic Regression

Logistic Regression doesn’t turn out to be a particularly good model – accuracy metric equal 0.8. Model achieves small precision and recall results for class 0 (Cancelled reservations). Only 75% of all reservations predicted to be cancelled by the model were indeed cancelled. Recall saying what % of all positive cases were predicted positive is even lower and equals 59%. To summarize, this model doesn’t have sufficient accuracy and struggled especially with Cancelled reservations which are more important in our case.

1. Recommendation for a final model
2. Summary Key Findings and Insights
3. Suggestions for next steps in analysing this data
4. References
5. Nuno Antonio, Ana de Almeida, Luis Nunes, *Hotel booking demand datasets*, Data in Brief, Volume 22, 2019, Pages 41-49, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2018.11.126.(https://www.sciencedirect.com/science/article/pii/S2352340918315191)