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.

2. Brief description of the data

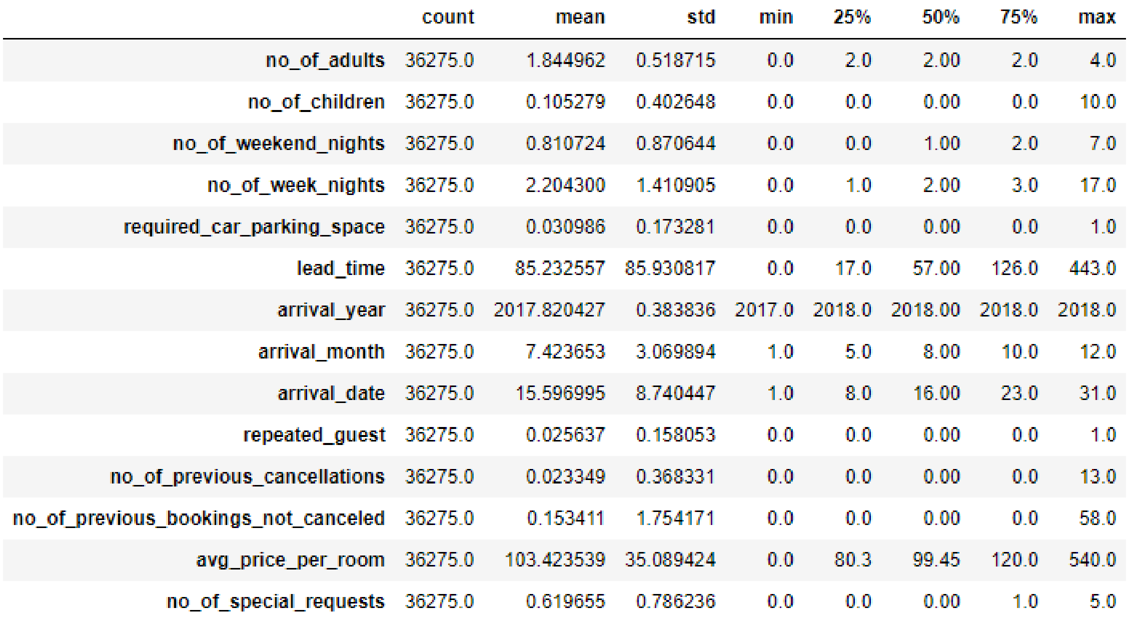
Dataset used in analysis presents 19 attributes of 36275 room reservations which are described below:

- 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)

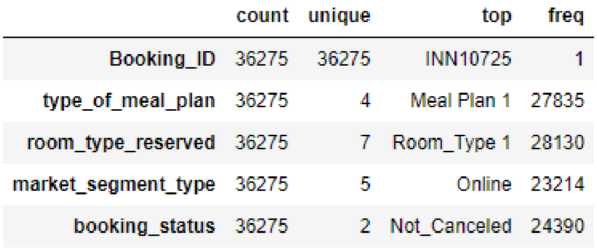
There are 15 numeric variables and 4 categorical variables. This dataset was part of the study developed by Nuno Antonio ,Ana de Almeida and Luis Nunes which was published in Hotel booking demand datasets, Data in Brief, Volume 22, 2019. With this analysis I try to predict whether a reservations meeting 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.



	count	mean	std	min	25%	50%	75%	max
no_of_adults	36275.0	1.844962	0.518715	0.0	2.0	2.00	2.0	4.0
no_of_children	36275.0	0.105279	0.402648	0.0	0.0	0.00	0.0	10.0
no_of_weekend_nights	36275.0	0.810724	0.870644	0.0	0.0	1.00	2.0	7.0
no_of_week_nights	36275.0	2.204300	1.410905	0.0	1.0	2.00	3.0	17.0
required_car_parking_space	36275.0	0.030986	0.173281	0.0	0.0	0.00	0.0	1.0
lead_time	36275.0	85.232557	85.930817	0.0	17.0	57.00	126.0	443.0
arrival_year	36275.0	2017.820427	0.383836	2017.0	2018.0	2018.00	2018.0	2018.0
arrival_month	36275.0	7.423653	3.069894	1.0	5.0	8.00	10.0	12.0
arrival_date	36275.0	15.596995	8.740447	1.0	8.0	16.00	23.0	31.0
repeated_guest	36275.0	0.025637	0.158053	0.0	0.0	0.00	0.0	1.0
no_of_previous_cancellations	36275.0	0.023349	0.368331	0.0	0.0	0.00	0.0	13.0
no_of_previous_bookings_not_canceled	36275.0	0.153411	1.754171	0.0	0.0	0.00	0.0	58.0
avg_price_per_room	36275.0	103.423539	35.089424	0.0	80.3	99.45	120.0	540.0
no_of_special_requests	36275.0	0.619655	0.786236	0.0	0.0	0.00	1.0	5.0

Figure 1. Main statistics for numerical variables



	count	unique	top	freq
Booking_ID	36275	36275	INN10725	1
type_of_meal_plan	36275	4	Meal Plan 1	27835
room_type_reserved	36275	7	Room_Type 1	28130
market_segment_type	36275	5	Online	23214
booking_status	36275	2	Not_Canceled	24390

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 criteria. 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.

3. 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

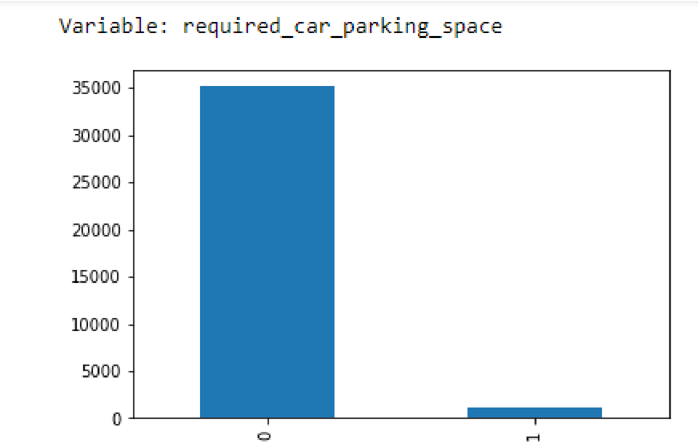


Figure 3. Distribution of Required car parking space variable

```
Variable: repeated_guest
```

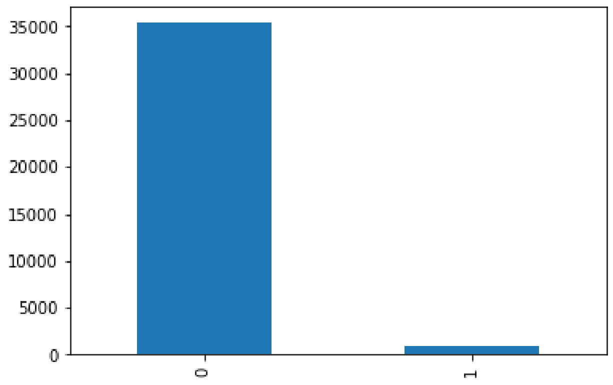


Figure 4. Distribution of Repeated guest variable

```
Variable: no_of_previous_cancellations
```

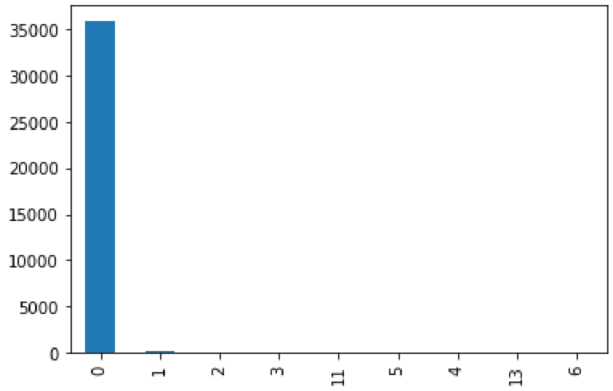


Figure 5. Distribution of Number of previous cancellations variable

```
Variable: no_of_previous_bookings_not_canceled
```

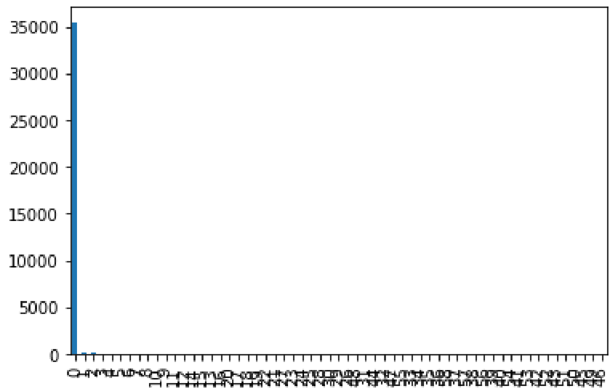


Figure 6. Distribution of Number of previous bookings not cancelled

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.

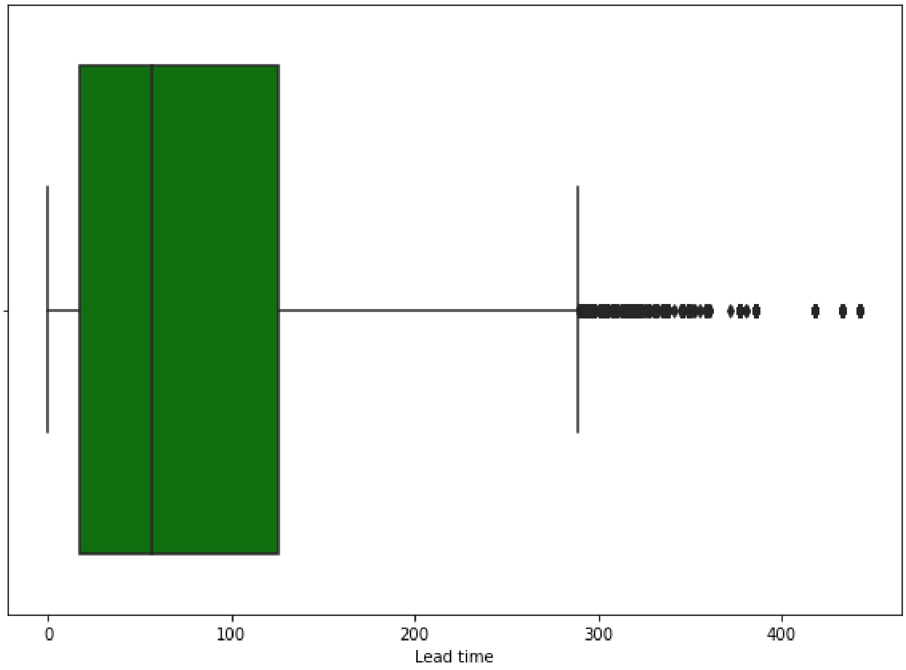


Figure 7. Boxplot for Lead time variable

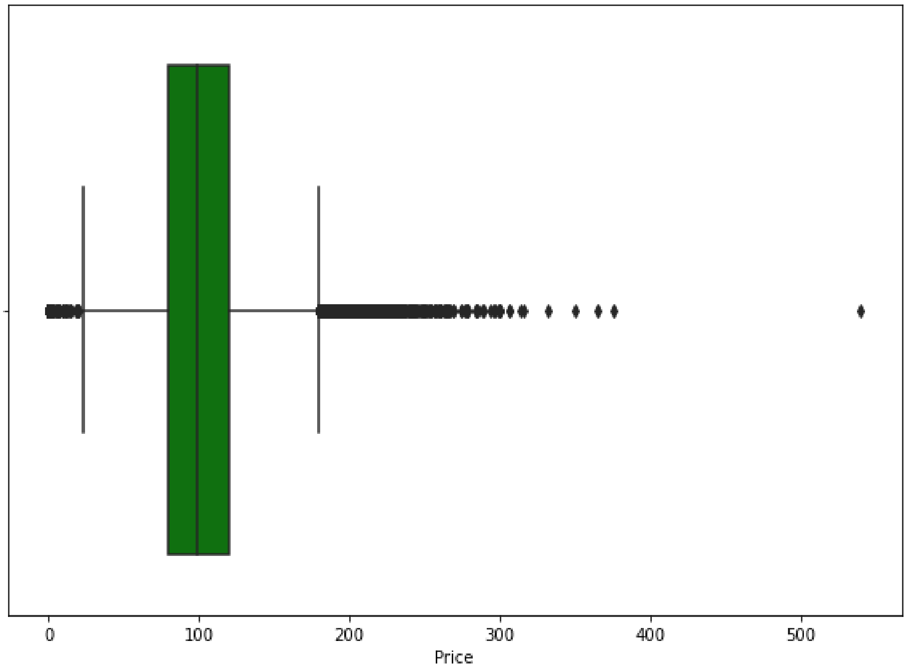


Figure 8. Boxplot for Price variable

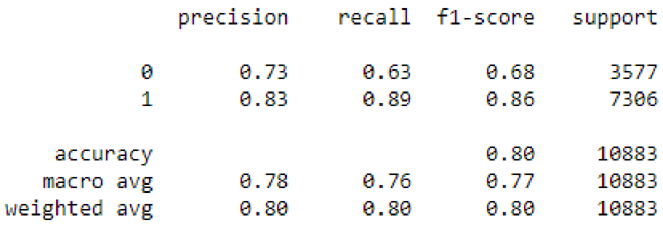
Both variables contain observations which might be perceived as outliers. Nonetheless I decided to keep all observations in the dataset as none of them seems to be an error and all of them on the other hand might bring some additional value into the analysis (assumptions being reservation with higher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

4. 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.



```
              precision    recall  f1-score   support

           0       0.73      0.63      0.68      3577
           1       0.83      0.89      0.86      7306

    accuracy                           0.80     10883
   macro avg       0.78      0.76      0.77     10883
weighted avg       0.80      0.80      0.80     10883
```

Figure 9. Evaluation metrics for Logistic Regression

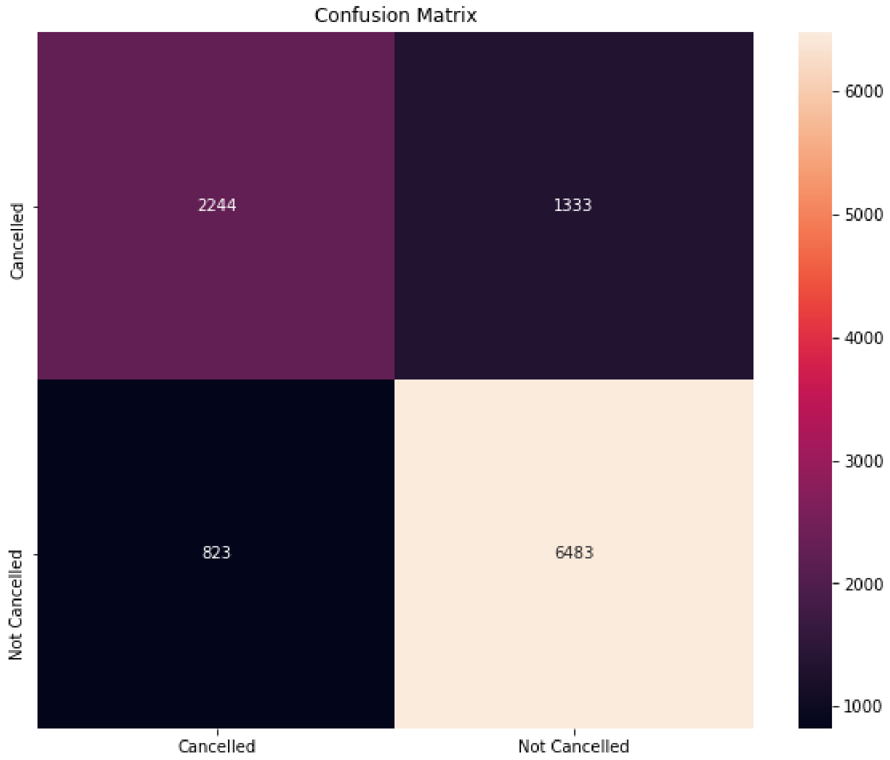
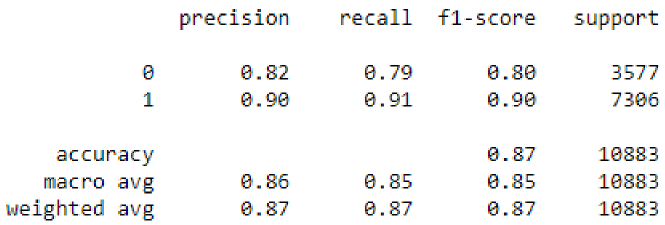


Figure 10. Confusion matrix for Logistic Regression

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 73% 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 63%. To summarize, this model doesn’t have sufficient accuracy and struggled especially with Cancelled reservations which are more important in our case.

Second algorithm that was used is K-Nearest Neighbours. As this algorithm operates on distance measures it is important to use scaled data there (this was actually already performed when working on Logistic Regression). I used the same datasets for testing and training as in previous algorithm. To find a most optimal KNN algorithm I used Grid Search CV tool which picks the best hyperparameters for this model. Below I present the key metrics’ results for this model:



```
              precision    recall  f1-score   support

           0       0.82      0.79      0.80      3577
           1       0.90      0.91      0.90      7306

    accuracy                           0.87     10883
   macro avg       0.86      0.85      0.85     10883
weighted avg       0.87      0.87      0.87     10883
```

Figure 11. Evaluation metrics for KNN

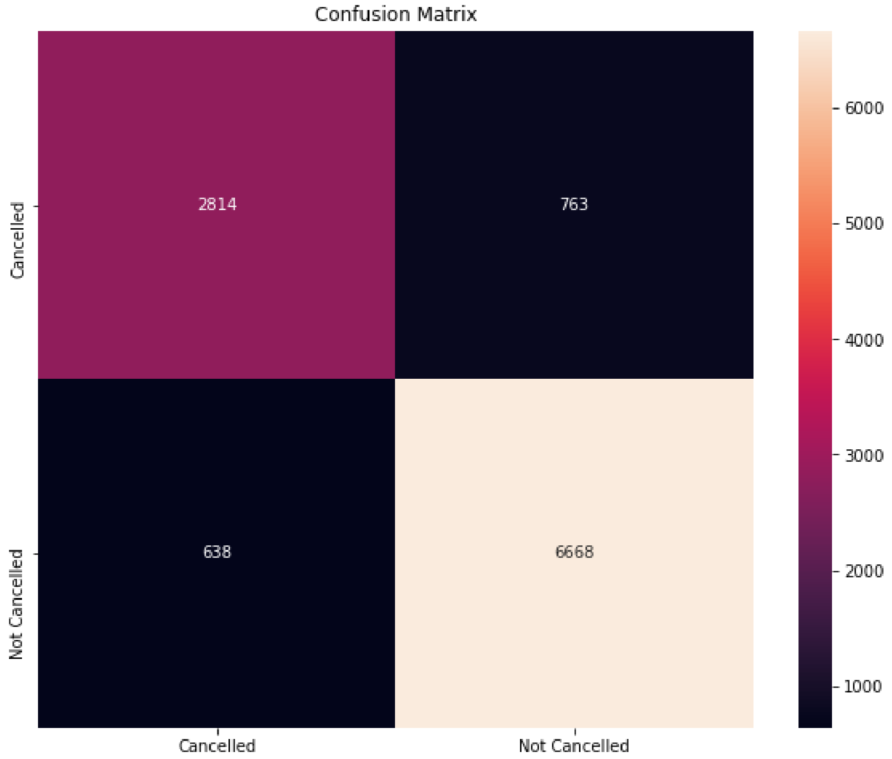
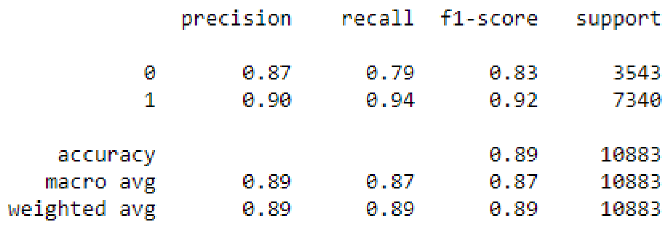


Figure 12. Confusion Matrix for KNN

We can see a big improvements comparing to Logistic Regression. All of the key metrics increased. Accuracy equals to 87% so it is 7 base points higher than in case of Logistic Regression. Precision which increased to 87% follows similar pattern. The biggest change though we can spot in recall which in this case equals 85%. KNN model then is significantly better in identifying the crucial category of cancelled reservations.

Last classification model that was built for hotel reservations data is Random Forest Classifier. In this case no data pre-processing is needed. Model was then prepared on non-scaled data with original categorical features coded as a numbers (which were transformed to binary variables in previous both examples). Again, 70% of records were taken into the training set and 30% into the testing set. Key evaluation metrics' results can be found on the following pictures.



```
              precision    recall  f1-score   support

           0       0.87      0.79      0.83      3543
           1       0.90      0.94      0.92      7340

    accuracy                           0.89     10883
   macro avg       0.89      0.87      0.87     10883
weighted avg       0.89      0.89      0.89     10883
```

Figure 13. Evaluation metrics for Random Forest

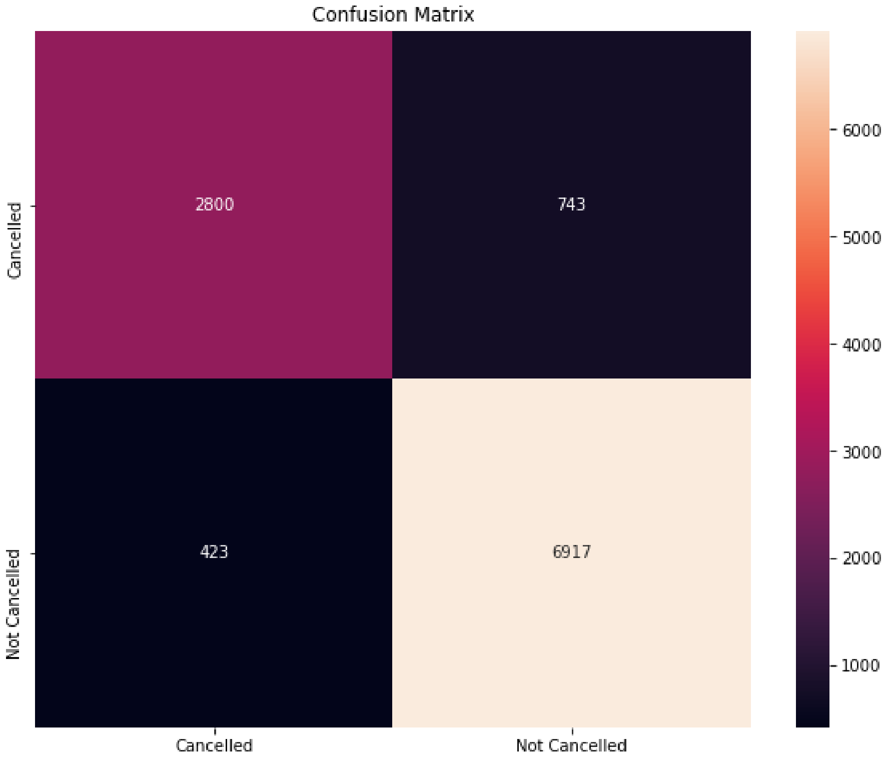


Figure 14. Confusion Matrix for Random Forest

Accuracy for Random Forest classifier equals 89% which makes it the highest from all of the analysed models. This is a slight improvement in comparison to what KNN algorithm has offered. Average precision result is 89% and average recall equals to 87%. Random Forest offers the best precision result for 'Cancelled' reservations which means it correctly recognizes cancelled cases. Precision for this class is the same as for KNN algorithm and equals 79%. This (i.e. optimizing recognition of all cancelled reservations) should be a main priority for the next model releases as all of the tested results struggle especially with this category.

5. Recommendation for a final model

To summarize in this analysis I prepared three classification models which task was to correctly identify cancelled hotel reservations. Machine Learning algorithms prepared in this paper were:

- Logistic Regression
- K-Nearest Neighbours
- Random Forest

All of the models were first trained on 70% of observations from the set and then tested on the rest of observations. After the fitting and predicting they were evaluated based on analysis of confusion matrix and metrics connected with it (accuracy, recall, precision). Model which achieved the greatest results is Random Forest. Accuracy for this classifier equalled 89%, average recall reached 87% and average precision was 89%. These results are slightly better than ones achieved by KNN model. Logistic Regression algorithm turned out to produce the worst results. Therefore I would recommend Random Forest as ‘go-to’ model for future releases of the analysis.

6. Summary Key Findings and Insights

Dataset used in the analysis turned out to be very clean and well prepared. There were neither duplicated nor missing values. For numerical variables some outliers existed but eventually all of the observations were taken into machine learning modelling phase.

One of the most interesting insights from the analysis of evaluation metrics calculated after predicting step in classification project was that all of the models struggled especially with differentiating and recognizing correctly cancelled reservations. Precision metric, saying what percent of all reservations predicted to be cancelled were indeed cancelled, achieved good but not perfect results. Recall metric for cancelled reservations category underachieved. This value says what % of all positive cases were predicted positive. This is not satisfying as cancellation is more important label in this project as we strive to find the ways to recognize potential cancellation cases in a hotel chain. Meanwhile all of the classification models had problems in recognizing all such labels (best result achieved by both KNN and Random Forest equals just 79%).

7. Suggestions for next steps in analysing this data

Further steps of the analysis should include two separate processes:

- Model optimization so that it will recognize the room cancellations with higher quality
- Finding the key factors driving people to cancel they stays. First part of analysis focused clearly on recognizing cases with high risk of cancellation. Further step would be to get some more insights which in longer run will allow hotels to adjust they offers and plans. Such an actions would potentially decrease the cancellation rate in the future and generate more income and eliminate uncertainty on the hotel side.

8. References

1. 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)