Hotel Cancellations

Group Project 7

Gabriel Deregnaucourt Paul Ritzinger

1 Introduction

A worldwide hotel brand wants to evaluate the cancellation rate or several reasons. Firstly, to better organize the schedule of their employees, then to adapt their policy by potentially discriminating between customers. If they can forecast which customer is probably going to cancel their trip, they can apply penalties to prevent profit losses. In this group project, we are presenting the dataframe containing every significant variables, such as the cancellation indicator, the guest's country of origin, which type of hotel is concerned, how many adults are involved and many more variables. The first step in the project is simply to describe the dataset with some graphs and tables to better visualize it. Then, we are presenting simple correlations and interaction between the variables, and looking at potential seasonality phenomena as well. The next step is to identify the biggest correlated variables with cancellation in order to prepare further analysis. Lastly, we build a simple Machine Learning decision tree model, to provide hands-on tools for the hotel brand. With this tool, they can easily identify the probability of a cancellation for each customer, once they have filled their reservation.

2 Proportions of cancellations in the dataset

The proportion of cancellations in the dataset is equal to 36.52%.

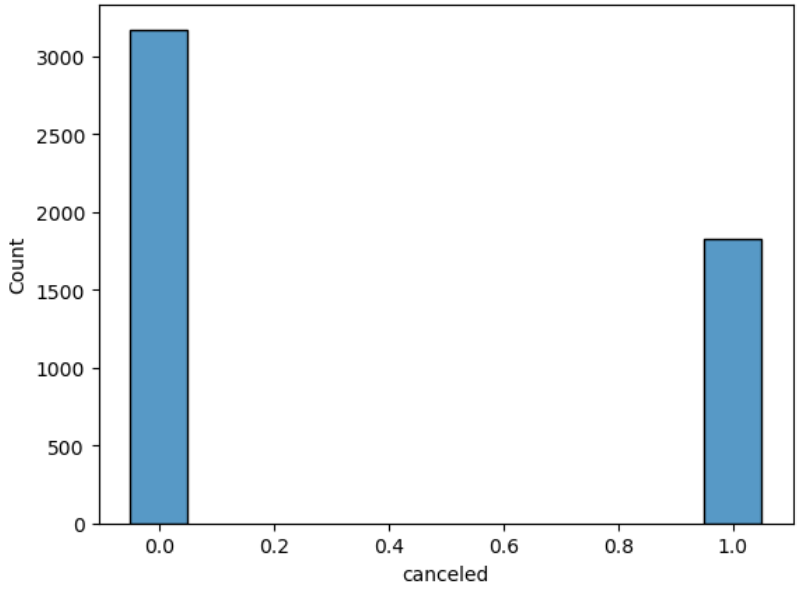

Figure 1: Rate of cancellation

3 How are the variables distributed in the dataset?

There are 5000 observations, without any missing value in the dataset. Here are some distribution graphs to better visualize the dataset: as shown in figure 2, most reservations are for with 3 adults with few different values represented.

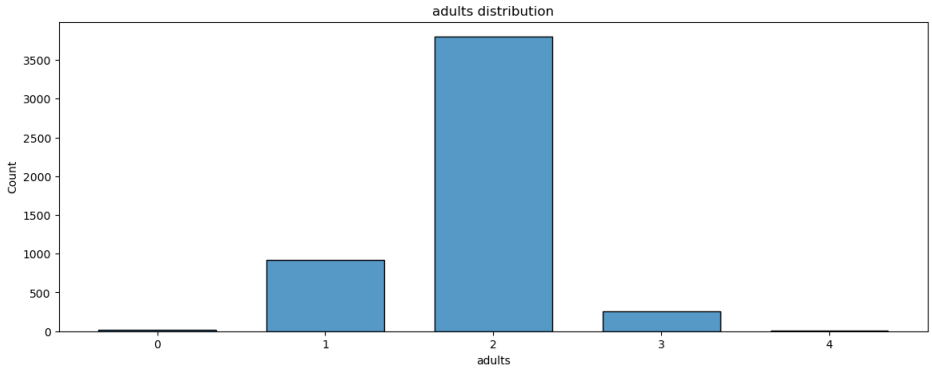

Figure 2: Number of adults per reservation

Figure 3 clearly shows a seasonality effect, with more arrivals in summer especially in august and fewer in winter. It could also be related to the location of the hotel.

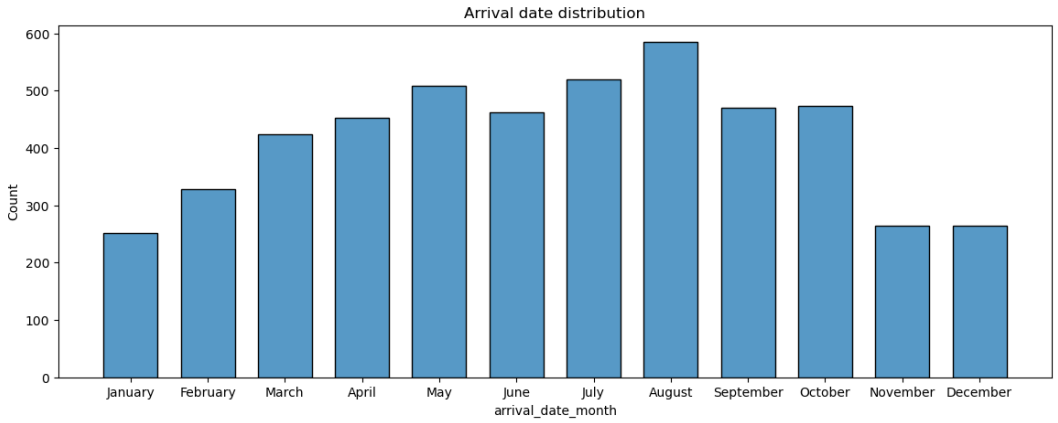

Figure 3: Number of arrival for each month

Figure 4 is interesting because it shows that most reservations are made by an online TA (travel agent). To maximize profit, the hotel brand could focus on this market segment.

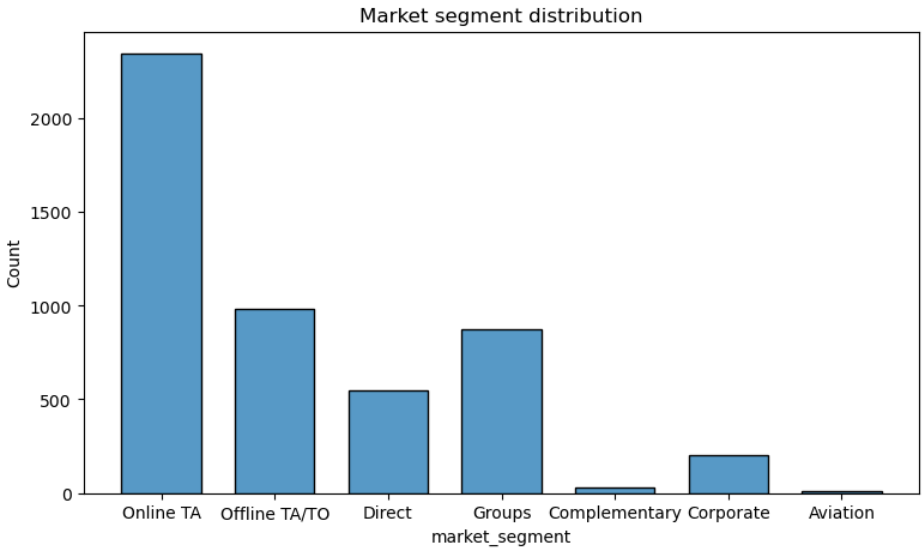

Figure 4: Market segment

4 How do variables interact with each other

4.1 Stays in nights with others variables

We are interested in identifying the variables associated with a longer duration of stay. We created the variable `stays_in_nights`, which is simply the sum of the number of weeknights and weekend nights.

First, we observe that if the booking is associated with a contract, the stay tends to be longer, averaging approximately 5 days, compared to the other groups (around 3 days each).

Then, if we do a geographical analysis, we collect the top ten countries (with at least ten observations) with the longest average stays

country	count	mean	std	min	25%	50%	75%	max
IRL	136.0	4.963235	2.6678	1.0	3.0	4.0	7.0	15.0
ROU	26.0	4.846154	2.361225	1.0	3.25	4.0	7.0	9.0
GBR	527.0	4.628083	3.004149	0.0	3.0	4.0	7.0	21.0
CN	59.0	4.559322	4.829047	1.0	2.5	3.0	4.0	25.0
POL	28.0	4.321429	1.88667	1.0	3.0	4.0	5.25	8.0
LUX	14.0	4.285714	2.267787	1.0	3.25	4.0	5.75	8.0
SWE	31.0	4.193548	3.709621	1.0	2.0	3.0	6.0	21.0
RUS	37.0	4.054054	2.655664	1.0	2.0	4.0	5.0	14.0
DNK	19.0	4.0	3.231787	1.0	2.0	3.0	4.5	12.0
FIN	17.0	4.0	2.0	1.0	3.0	4.0	6.0	7.0

Figure 5: Statistics of average stay for the top ten countries

The Irish are the population who stay the longest, with an average of almost 5 days, followed by the Romanians. The British complete the podium with an average stay of 4.6 days. We can observe that, with the exception of China, the top ten is composed mainly of European countries. Europeans seem to have longer stays on average.

If we look at the populations who stayed the least, we see the Spanish and the Portuguese, ranked 4th and 5th from the bottom, respectively. However, these countries represent a large portion of the customers (2,048 for Portugal and 392 for Spain).

If we analyze by continent, people from Africa stay for more nights on average. However, there are only a few cases (36), and the same goes for Oceania with just 24 observations, so we cannot draw any firm conclusions. Looking at Europe, Asia, and America (with at least 172 observations each), Europeans and Asians tend to stay slightly longer than Americans.

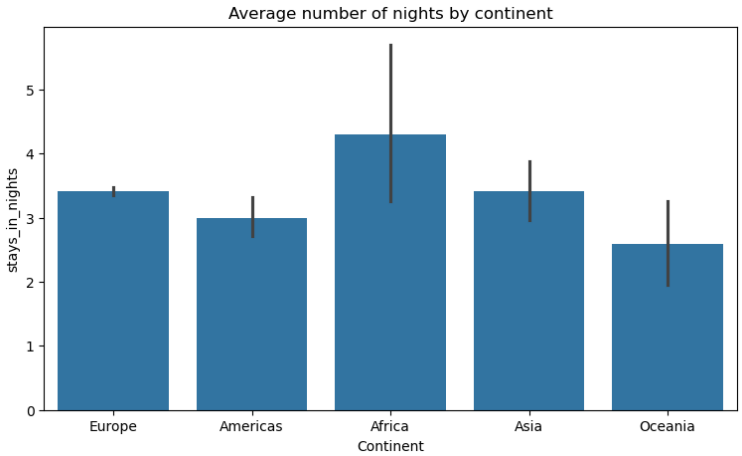

Figure 6: Average number of nights by continent

4.2 Average daily rate

We also analyze what variables are associated with the average daily rate.

What type of consumers spend the more?

The three countries with the highest daily rates are, in order, Luxembourg, Morocco, and Denmark. The three countries that spend the least per night are Portugal, Great Britain, and Finland.

Excluding Africa and Oceania, by continent, Americans spend on average $118.5 per night, followed by Asians at $110.9. Europeans spend less, with an average daily rate of $101.6.

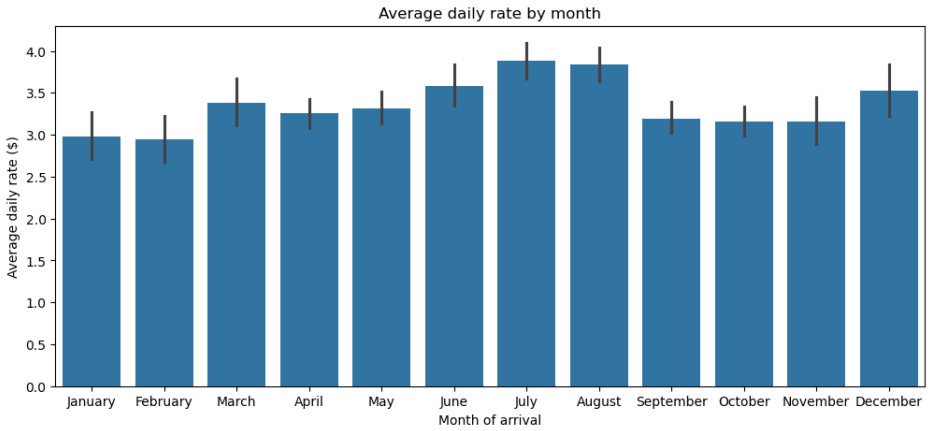

Figure 7: Average daily rate by month

The customers spend more money per night in the summer, especially in August, than during the other seasons. This could be explained by the fact that there is higher demand

during the summer holidays than during the rest of the year, particularly in the colder months.

5 What variables are associated the most with booking cancellations?

To conduct this analysis, we compute the correlation between booking cancellations and the other variables. This allows us to identify potential determinants of cancellations.

First, we see that the target variable has a strong positive correlation when the deposit type is Non Refund. It means that a deposit was made for the total cost of the stay. However, this does not make much sense, and furthermore, 598 out of 600 bookings with a non-refund payment were cancelled. We suspect there may be an error in the dataset.

Portuguese guests are highly correlated with the cancellation rate. Indeed, 56.45% of Portuguese guests cancel their trip. In comparison, the cancellation rate for the Dutch is only 9%, which is a substantial difference.

Also, the more demanding the customer is, the less likely they are to cancel. The cancellation rate when there is at least one special request is 22.84%, whereas the cancellation rate when no request has been made is 46.12%. We observe the same pattern for the number of car parking spaces required. Among the 323 guests who requested a parking space, none cancelled (which may also indicate a data issue).

Finally, increasing the price to discourage cancellations could be an option to consider.

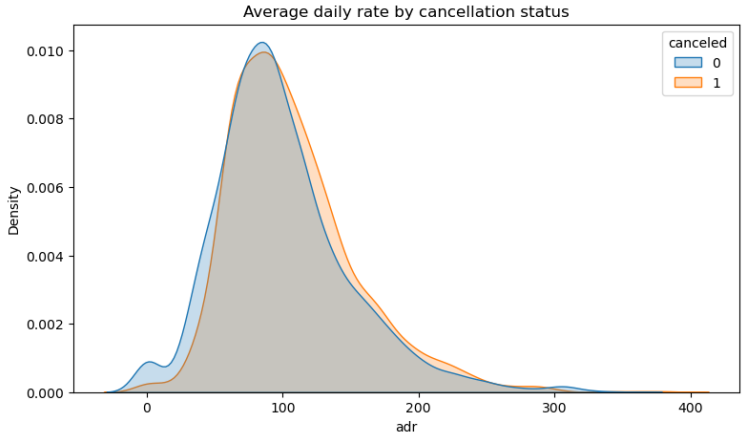

Figure 8: Average daily rate

However, if we analyze the distribution, the cancellation rate is higher when the average daily rate increases. And when a booking is canceled, the price is on average 7.45% higher than when it is not.

6 Machine learning classification model

We built a simple decision tree to help the hotel brand forecast which customer is likely to cancel.

Using the package Scikit-learn, we first splitted the dataset in two parts, the explained variable (y = canceled) and the explaining variables (X = database – canceled). After a few tests, we figured out that some variables were categorical data which was a problem for the following steps. So we one-hot encoded these variables to have dummies and to be able to run our classification problem. By using the standard 30% of the dataset for the test set, we obtained the following results:

- Depth of the tree: 35
- Number of leaves: 588

Which may be overfitted. By looking at the accuracy scores, we get the following results:

- Train accuracy: 0.9985714285714286
- Test accuracy: 0.772

Which shows us clearly an enormous overfitting. To prevent this issue, we are going to try a second tree with `min_samples_split = 8`, `max_depth = 10`. We get the following results:

- Train accuracy: 0.8508571428571429
- Test accuracy: 0.7913333333333333

Which is definitely better.

To find the better combination of `min_samples_split` and `max_depth`, we are using the function `GridSearchCV`, to run a cross-validation and evaluate each tree by its test accuracy. We obtained `max_depth = 10`, `min_samples_split = 2` for the best hyper parameters, and a test accuracy of 0.7939. To better visualize, we printed this tree:

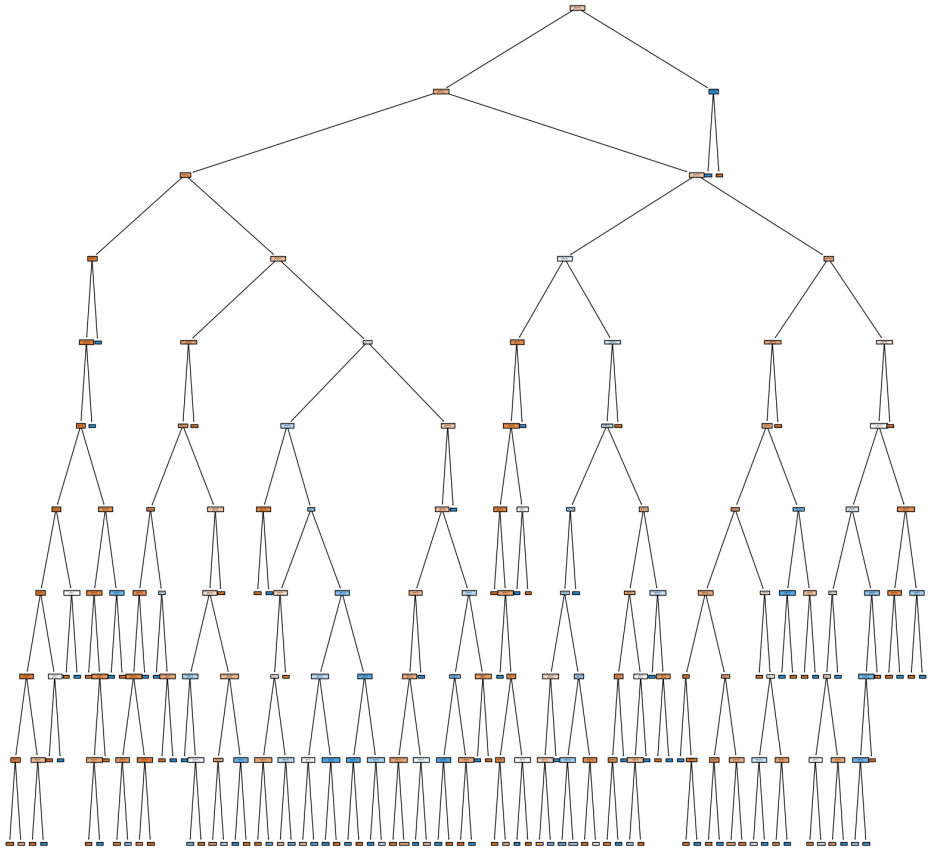

Figure 9: Decision tree with max_depth = 8 and min_samples_split = 6

The last step of our analysis was to run the same code after dropping `deposit_type` because of its weird correlation with cancellation. Not surprisingly, the accuracy score dropped by about 1%, due to the removal of the biggest explaining variable of the dataset. The new best hyper parameters are 8 for `max_depth` and 6 for `min_samples_split`, with a wider range for both of them:

```
param_grid = {
    'max_depth': list(range(5, 21)),          # 5 to 20
    'min_samples_split': list(range(2, 10))   # 2 to 10
}
```

We obtained the following confusion matrix, showing that the number of false negative is quite high.

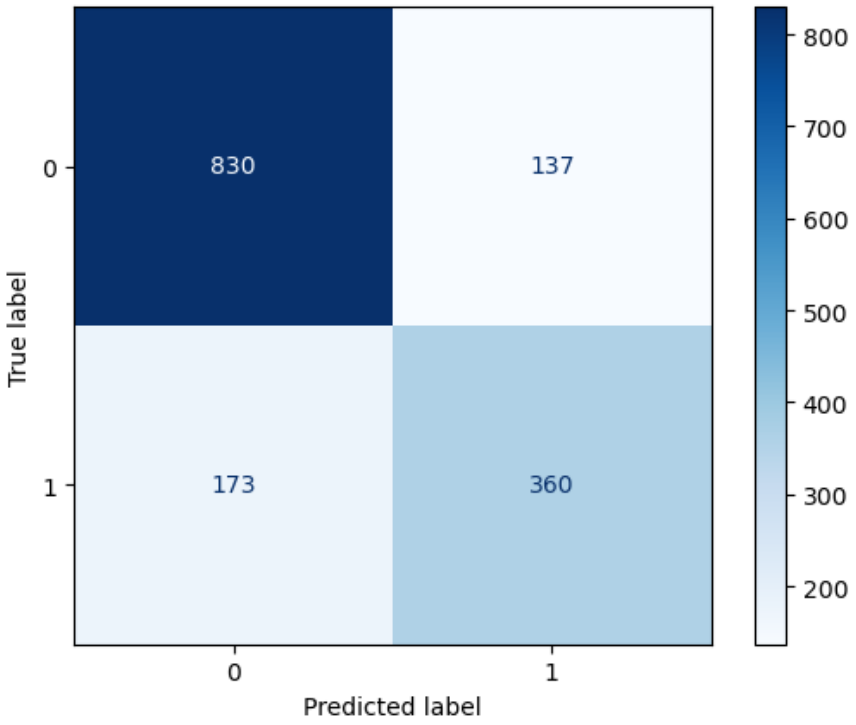

Figure 10: Confusion matrix for the best hyperparameters