Report TP2 Apprentissage Statistique Appliqué

HEDFI Houeida

KONDO Godson Leopold Junior

December 9, 2020

PART 1 : visualization

1. Re-ordering the barplot using standard month ordering

We have re-ordered the barplot using pd.Categorical() function of pandas.

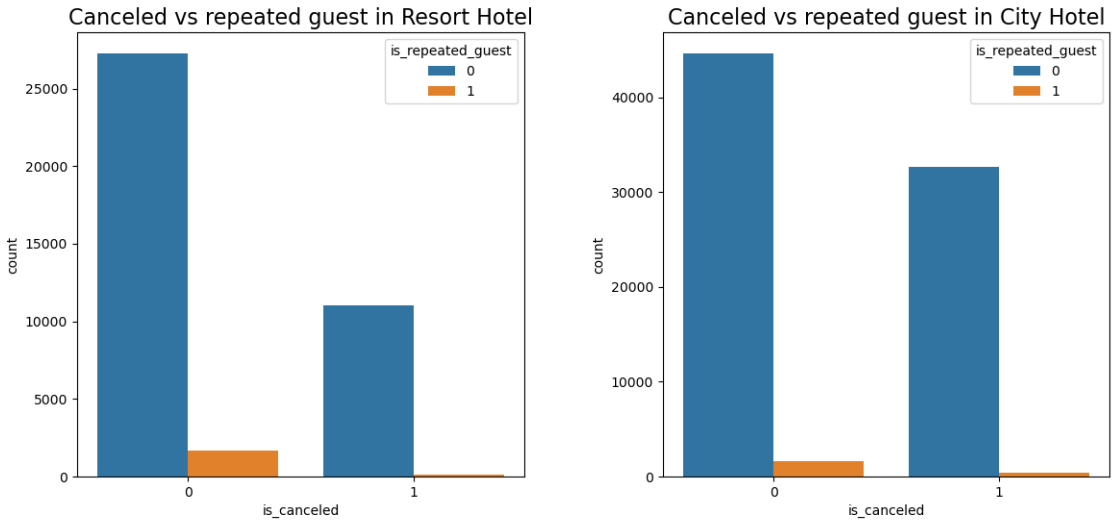
2. Interpretation of the plot

First of all we can notice that the cancellation rate of reservations over the year are higher in City Hotel than in Resort Hotel. The cancellations rate over the year vary between 15% and 35% in Resort Hotel and between 35% and 45% in City Hotel. The peak of cancellations rate in Resort Hotel is in August and in City Hotel, the peak is in April.

3. Most and the second most common country of origin for reservations

- For the Resort Hotel the most common country of reservation is Portugal, with 17630 bookings, and the second common is Great Britain, with 6814 bookings.
- For the Resort Hotel the most common country of reservation is Portugal, with 30960 bookings, and the second common is France, with 8804 bookings.

4. Plotting the number of cancellations for repeated and not repeated guests



5. Special request vs Cancellation in Resort Hotel

Contrarily to City Hotels, the rate of Cancellation reservations when no special requests are made

is 30%. However, when special requests are made the cancellation rate become slightly lower (approximately 20%).
Please find attached the figure 1 in the appendix and in the notebook.

PART 2 : ML

6. What is OneHotEncoder()? Why do we use it in our case?

OneHotEncoder() is a **sklearn** function which allows to encode categorical features as a numeric array. This function creates a binary column for each category and returns a sparse matrix or dense array.
In our case, OneHotEncoder() is used with the parameter *handle_unknown='ignore'* which converts unknown category to a resulting one-hot encoded columns filled with zeros.

7. Inserting a Normalization Step in the pipeline

We have inserted a normalization step in the pipeline. We have used **Standard Scaler** (StandardScaler()). The latter resolved indeed the warning.
Please find the normalization step code in the notebook.

8. Our favourite ML method

We have used random forest method. As a first step, we made our random forest model without a data split and we got as best hyper-parameter 100 and an accuracy 0.9671. In second time, we have divided our Data set into *train* and *test*. the train set represent 80% of the data. The results are as follows. For the accuracy on test set we have 0.9218 and the best hyper-parameter is equal to 1500.

PART 3 : The homework

We will construct a prediction algorithm for the target variable that we will define ("need_car_parking_spots") in order to advise the manager whether to send an SMS to a costumer or not. To do so, we will start by defining our target variable. Second, we will explore the data in order to exclude irrelevant features. Third, we will explore data using descriptive statistics methods in order to highlight the relevant features to use in our algorithm. We will end up by constructing our prediction algorithm and presenting our results. To see the usefulness of our choices in the precision of our model we considered producing two models, one with the selected variables and the other with all the variables

available and with no missing values. The choice of these variables goes through several procedures that we present below.

1. Defining the label

We will, first of all, define our label "need_car_parking_spots". We have used the variable "required_car_parking_spaces" and we have assigned the value of 0 to our label is the variable "required_car_parking_spaces" is equal to zero and the value 1 if the variable "required_car_parking_spaces" is strictly superior to zero.

2. Features exclusion

There are several features we need to exclude directly for the following reasons:

Features exclusion because of non availability:

Among the variables, there are some that we need to exclude because they are not available at the moment of reservation, such as IsCanceled, ReservationStatus, ReservationStatusDate and ADR.

Features exclusion because of non consistency:

The variable AssignedRoomType is not consistent because it does not depend on the custumer.

Feature exclusion because of missing values:

The variable "company" contains 94.3% of missing values, which consists of a total of 112593 missing values. Despite the fact that this variable seems useful for our prediction, we have chosen to not use it because of the important amount of missing values. For the same reasons, we have also decided to exclude the variable "agent" that contains 16340 missing values.

Feature exclusion conclusion:

We will operate a Data exploration of the features that remained, they consist of the following:

numeric features: "arrival_date_week_number", "adults","children", "babies" ,"booking_changes" ,"total_of_special_requests", "stays_in_weekend_nights", "stays_in_week_nights", "previous_bookings_not_canceled", "arrival_date_year", "arrival_date_day_of_month", "lead_time", "days_in_waiting_list" and "previous_cancellations".

categorical features: "hotel", "market_segment","reserved_room_type", "deposit_type","customer_type", "distribution_channel", "meal", "country", "arrival_date_month" and "is_repeated_guest".

3. Data exploration and feature consistency

Features consistency using the correlation ratio:

With the aim to identify the most valuable features for our model, we calculated the correlation ration between all numerical features and the target variable (Table 1). For this matter, we needed to exclude some reservations. As a matter of fact, the feature *children* contains 4 missing values.
From this table, we can notice that most of the correlation ratio are low but not null. The feature *lead_time* has the higher correlation ratio with our target variable (0,01%). The features *arrival_date_week_number* and *arrival_date_day_of_month* have the lowest correlation ratio and we chose to not use them in the model with selected features. Thus, we can used all these variables to make sure to capture all the information they can provide.

Table 1: Correlation ratio

Features	**0**
arrival_date_week_number	0,00000815
adults	0,00021370
children	0,00323819
babies	0,00142659
booking_changes	0,00427547
total_of_special_requests	0,00697489
stays_in_weekend_nights	0,00038703
stays_in_week_nights	0,00068509
previous_bookings_not_canceled	0,00229292
arrival_date_year	0,00022134
arrival_date_day_of_month	0,00007300
lead_time	0,01395211
days_in_waiting_list	0,00106037
previous_cancellations	0,00034678

Features consistency using Khi2 test:

We use the *khi2 test* to detect categorical variables related to the target variable. For this matter, we needed to exclude some reservations. As a matter of fact, the feature *country* contains 488 missing values.
With the results of the test below, we can see that the p-values are all below 5%, which allows us to say that all these variables are not independent to the target variable. We can therefore keep them for the prediction.

Features consistency using the correlation matrix:

```
SpearmanrResult(correlation=0.2191334235873404, pvalue=0.0)
SpearmanrResult(correlation=-0.0311011156912504, pvalue=7.617144317002537e-27)
SpearmanrResult(correlation=0.12154566894142614, pvalue=0.0)
SpearmanrResult(correlation=-0.09399919661772066, pvalue=1.7615471062237836e-231)
SpearmanrResult(correlation=-0.050708630584633256, pvalue=1.5232677954551167e-68)
SpearmanrResult(correlation=-0.13773117959885858, pvalue=0.0)
SpearmanrResult(correlation=-0.03524701902003104, pvalue=5.2280608196530004e-34)
SpearmanrResult(correlation=-0.005165978038646425, pvalue=0.07485782955879906)
SpearmanrResult(correlation=-0.01984225032263943, pvalue=7.774685511887175e-12)
SpearmanrResult(correlation=0.07598438392243852, pvalue=9.612825746387989e-152)
```

In order to avoid multi-colinearity problems, we had built the correlation matrix between all our numerical features (Figure 10 in Appendix). When we analyse the correlation matrix, we can notice that the variables are not very correlated with each other. So at this step we do not remove any feature.

Features consistency using data exploration:

We used the exploration of data to have another look on the dependence between our target variable and the other features. We can conclude from the figures that many features seem to be correlated to our target variable. As a matter of fact, in the figure 8, we observe that clients of resort hotel are more likely to need a car spot than the clients of city hotel. Besides, from the figure 4, we can observe that needing a parking spot depend on the market segment.

Please find attached the figures 2 to 9 in the appendix.

Features selection conclusion:

We have chosen to remove the following features: "arrival_date_week_number", "arrival_date_year" and "arrival_date_day_of_month".

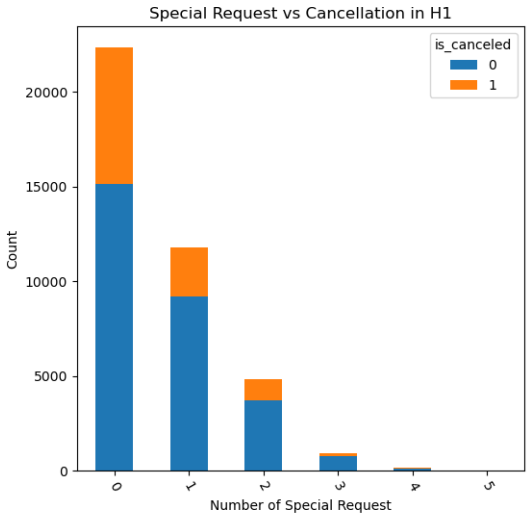
4. Interpretation of results

First of all, we have built a logistic regression with all the observable features. The classification accuracy on test is equal to 0.9383. Secondly, we had built the logistic regression with the selected variables and here we got an accuracy of 0.9381.

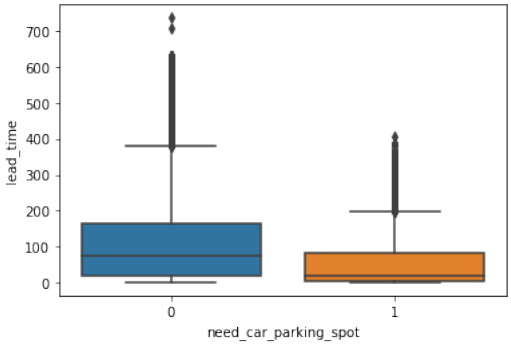
Third, We have built two random forest algorithms with and without features selection. For technical reasons we couldn't get the results of our favorite algorithm (the code has been running for two days). Thus, we decided to restrict ourselves to the results of the logistic regression algorithm to advise the manager about the customers to whom we will have to send an SMS message. For each new reservation, we will predict the value of the variable "need_car_parking_spots" using logistic regression with all the observable features because it has the highest prediction accuracy on test.

PART 4 : Appendix

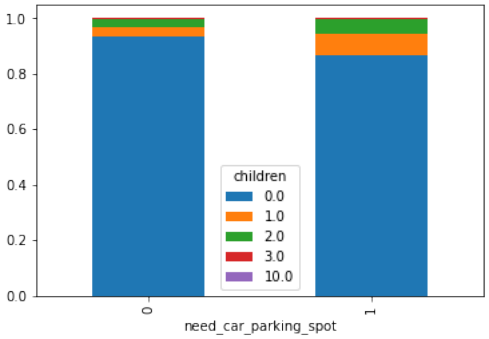
1. Figure 1: Special request vs Cancellation in H1



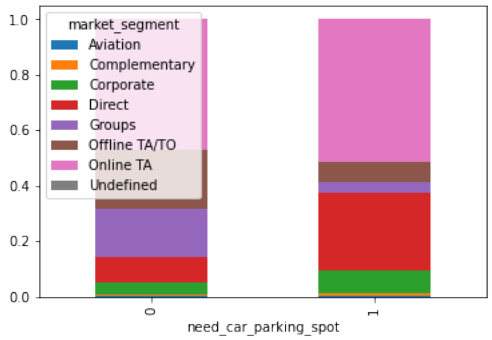
2. Figure 2: Lead time



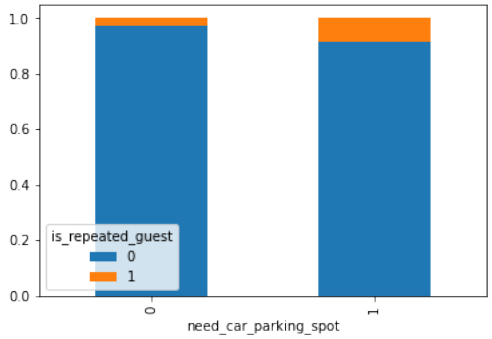
3. Figure 3: Children



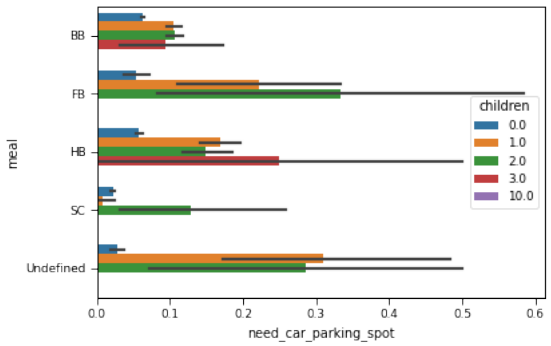
4. Figure 4: Market segment



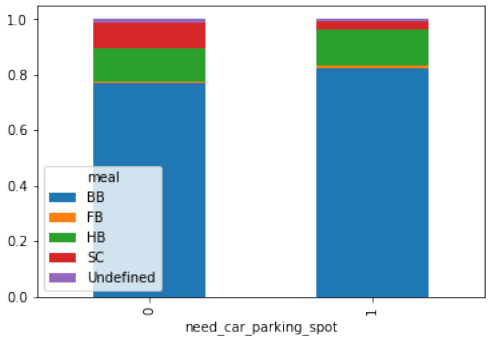
5. Figure 5: Repeated guest



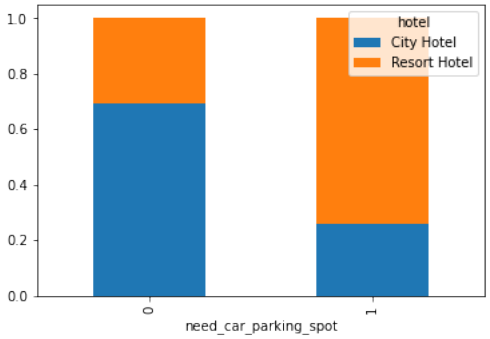
6. Figure 6: Meal children



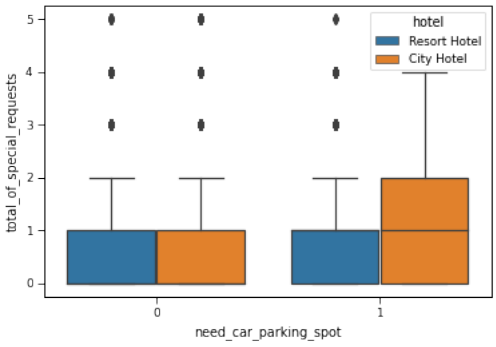
7. Figure 7: Meal



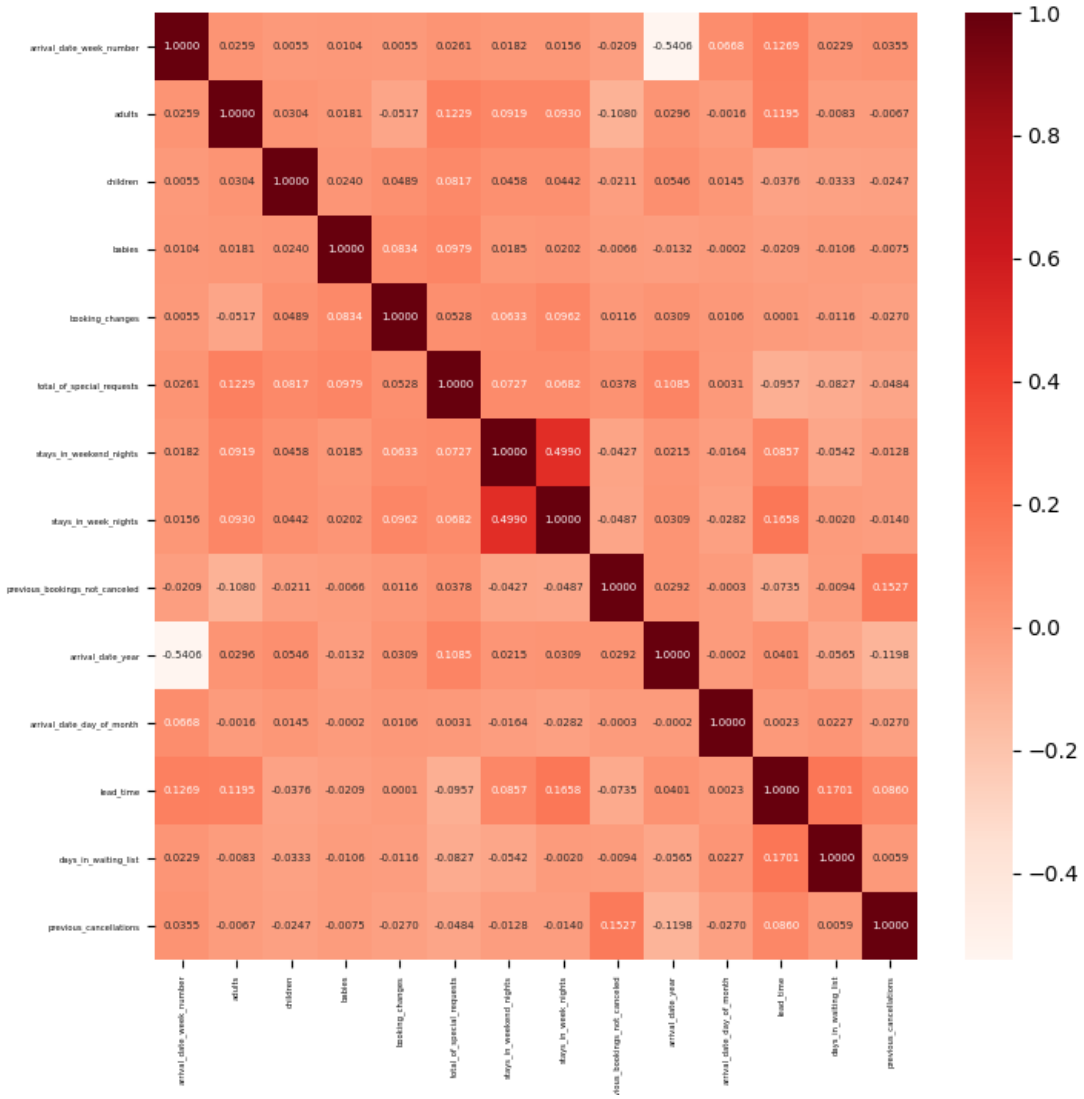
8. Figure 8: Hotel



9. Figure 9: Special request



10. Figure 10: Correlation matrix


arrival_date_week_number 1.0000 0.0259 0.0055 0.0104 0.0055 0.0261 0.0182 0.0156 -0.0209 -0.5406 0.0668 0.1269 0.0229 0.0355
adults 0.0259 1.0000 0.0304 0.0181 -0.0517 0.1229 0.0919 0.0930 -0.1080 0.0296 -0.0016 0.1195 -0.0083 -0.0067
children 0.0055 0.0304 1.0000 0.0240 0.0489 0.0817 0.0458 0.0442 -0.0211 0.0546 0.0145 -0.0376 -0.0333 -0.0247
babies 0.0104 0.0181 0.0240 1.0000 0.0834 0.0979 0.0185 0.0202 -0.0066 -0.0132 -0.0002 -0.0209 -0.0106 -0.0075
booking_changes 0.0055 -0.0517 0.0489 0.0834 1.0000 0.0528 0.0633 0.0962 0.0116 0.0309 0.0106 0.0001 -0.0116 -0.0270
total_of_special_requests 0.0261 0.1229 0.0817 0.0979 0.0528 1.0000 0.0727 0.0682 0.0378 0.1085 0.0031 -0.0957 -0.0827 -0.0484
stays_in_weekend_nights 0.0182 0.0919 0.0458 0.0185 0.0633 0.0727 1.0000 0.4990 -0.0427 0.0215 -0.0164 0.0857 -0.0542 -0.0128
stays_in_week_nights 0.0156 0.0930 0.0442 0.0202 0.0962 0.0682 0.4990 1.0000 -0.0487 0.0309 -0.0282 0.1658 -0.0020 -0.0140
previous_bookings_not_canceled -0.0209 -0.1080 -0.0211 -0.0066 0.0116 0.0378 -0.0427 -0.0487 1.0000 0.0292 -0.0003 -0.0735 -0.0094 0.1527
arrival_date_year -0.5406 0.0296 0.0546 -0.0132 0.0309 0.1085 0.0215 0.0309 0.0292 1.0000 -0.0002 0.0401 -0.0565 -0.1198
arrival_date_day_of_month 0.0668 -0.0016 0.0145 -0.0002 0.0106 0.0031 -0.0164 -0.0282 -0.0003 -0.0002 1.0000 0.0023 0.0227 -0.0270
lead_time 0.1269 0.1195 -0.0376 -0.0209 0.0001 -0.0957 0.0857 0.1658 -0.0735 0.0401 0.0023 1.0000 0.1701 0.0860
days_in_waiting_list 0.0229 -0.0083 -0.0333 -0.0106 -0.0116 -0.0827 -0.0542 -0.0020 -0.0094 -0.0565 0.0227 0.1701 1.0000 0.0059
previous_cancellations 0.0355 -0.0067 -0.0247 -0.0075 -0.0270 -0.0484 -0.0128 -0.0140 0.1527 -0.1198 -0.0270 0.0860 0.0059 1.0000
arrival_date_week_number
adults
children
babies
booking_changes
total_of_special_requests
stays_in_weekend_nights
stays_in_week_nights
previous_bookings_not_canceled
arrival_date_year
arrival_date_day_of_month
lead_time
days_in_waiting_list
previous_cancellations
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4