ENSAE 3^{e} année
Apprentissage Statistique Appliqué
Année scolaire 2021/2022

HENG Anne-Marie
MARTIN Elise
OULID AZOUZ Noureddine

TP2 : Pandas, data analysis library

Professeur : Mohamed HEBIRI

Chargée de TD : Clara CARLIER

1 Predicting cancellation : Part I – visualization

Question 1.1

Propose a solution that will re-order the barplot above using standard month ordering

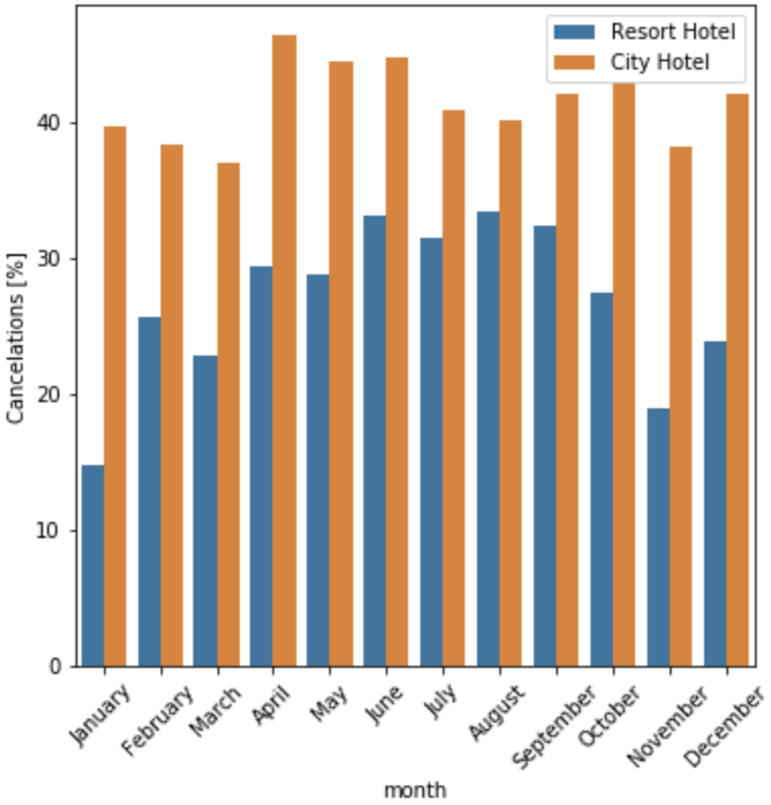

Figure 1 – Cancelations per month

Question 1.2

Provide interpretation of the above plot

Regardless of the month of the year, we can notice that there are more cancellations in the city hotel. The percentage of cancellations increases slightly between April and August. The cancellation rate is usually close to 40% for the City Hotel, while it is around 25% for The Resort Hotel. This rate decreases to 15% in January and 18% in November for the Resort Hotel. The highest cancellation rate is 45% for the City Hotel in April. When looking to the behavior of cancellations with respect to the month of the year, no clear relationships appears for both the hotels.

Main conclusions are :

- there is clearly always more cancellations on City Hotel than in Resort Hotel;
- there is no big seasonal trends in cancellations for City Hotel.

Question 1.3

What is the most and the second most common country of origin for reservations of each hotel ?

The most common country of origin for reservation in Resort Hotel is Portugal with 17630 cancellations, and the second most common country is Great-Britain with 6814.

For City Hotel, the most common country of origin for reservation is Portugal, and the second most common is France with respecteively 30960 and 8804 cancellations.

Question 1.4

Plot the number of cancelations for repeated and not repeated guests for both hotels

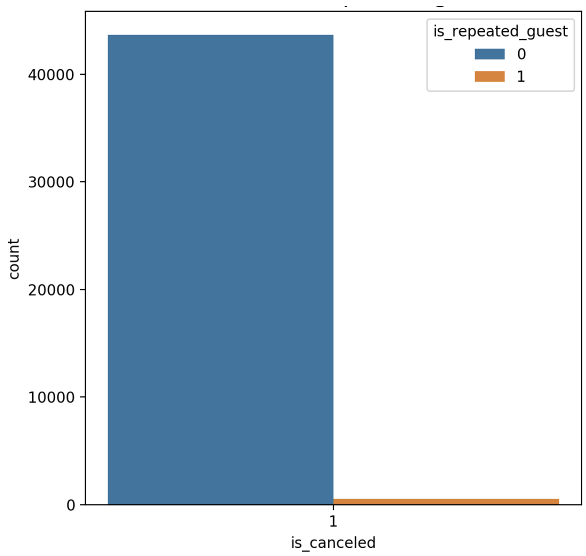

FIGURE 2 – Number of cancelations for repeated and not repeated guests

Question 1.5

Make the same plot for Resort Hotel. Make your conclusions.

The more clients ask for Special Request, the less they cancel the reservation. Indeed, we can see that when there is no special request, almost the third of the reservations are cancelled, whereas when clients do at least 3 special requests, there are barely no cancelations.

We can also add that there is a smaller proportion of cancellation in Resort hotel than in city hotel, no matter the number of special requests.

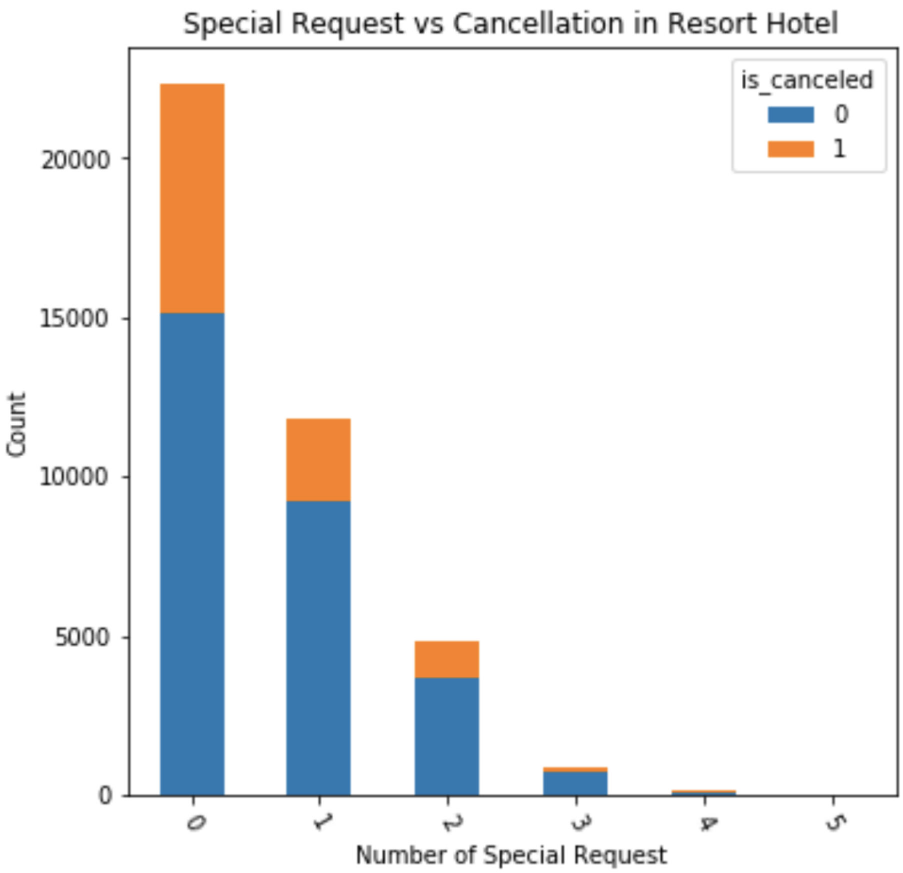

FIGURE 3 – Special Request vs Cancellation in Resort Hotel

2 Predicting cancellations : Part II – ML

Question 2.1

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

The "One-Hot-Encoding" is an approach that allows us to adequately treat qualitative variables with a number of categories greater than or equal to 3. This allows to encode categorical into numerical data to feed our ML algorithms with.

Let's suppose we have a qualitative variable, X with $k \geq 3$ categories. For one feature, "One-Hot-Encoding" creates $n_{categories}$ number of columns (as many columns as classes or values for the given feature). Each column now represents 1 class and the value will be 1 (resp. 0) if the entry row belongs to the class (resp. doesn't belong to the class). Indeed, if the realization of the qualitative variable is the modality $l \leq k$ then, the observation is represented as follows :

$$(0, 0, ..., 1_{l^{th}position}, 0, ..., 0)$$

This is a very good way to treat categorical features with no ordinal relationship.

Question 2.2

As you have seen, proper normalization can resolve the issue. Insert a normalization step in the pipeline. Note that we do not want to normalize the categorical data, it simply does not make sense. Be careful to normalize only the numerical data. Did it resolve the warning ?

Just as in the previous assignement, scaling our data did remove the convergence warning.

Question 2.3

As we can see, previous code uses only logistic regression. Modify the above code inserting your favorite ML method

For this question, please refer to the Notebook. We decided to use a KNN-classifier and it does not perform better than the Logistic Regression.

3 Homework

Question 3.1

Explore your data to help the manager and construct a prediction algorithm, using the above template as an inspiration.

Drop columns not available at the moment of the reservation

First of all, it is important to notice that some variables are not available at the time of reservation. We need to drop them for our modeling. The variables are the following :

is_canceled
days_in_waiting_list
reservation_status
reservation_status_date
booking_changes
assigned_room_type

Exploring the target column

Required car parking spaces	Number of values
0	111974
1	7383
2	28
3	3
8	2

We can see in the previous table the number of required car parking spaces. As only few people ask for a lot of parking spaces, let's dive into the people who ask for at least one parking space.

For now, we use an encoded version of this feature created as follow : 1 if a parking lot was asked and 0 otherwise.

Parking asked	Number of values
0 (No)	111974
1 (Yes)	7416

This rises an important issue we'll try to deal with later : our dataset is really imbalanced and this might make it difficult to deploy an effective ML pipeline ! Indeed, a simple classification would be quiete biased because of the over represented people not asking for parking spaces.

In a first time, we will try to analyze the relationship beetween this new variable called parking and the others, in order to do manual feature selection.

Correlation Matrix

One of the simplest way to analyse relation between couple of numerical features is the correlation matrix. This helps us confirm theoretical effect and basic idea we might have.

An important thing to notice before analyzing the correlation matrix is the fact that the variables "agent" and "company" are coded in a certain way that could false the result of correlation. Indeed, the value for the variable "agent" corresponds to an agent ID which introduces a false ordinal relation ! This falses completely the result of correlation. In order to deal with that, we will encode this feature so we know if the reservation was made by an agency (and thus we would have a non null ID that we replace by 1) or no (replace a null by 0). The same encoding is done for the variable "company".

Thus, we obtain the following correlation matrix :

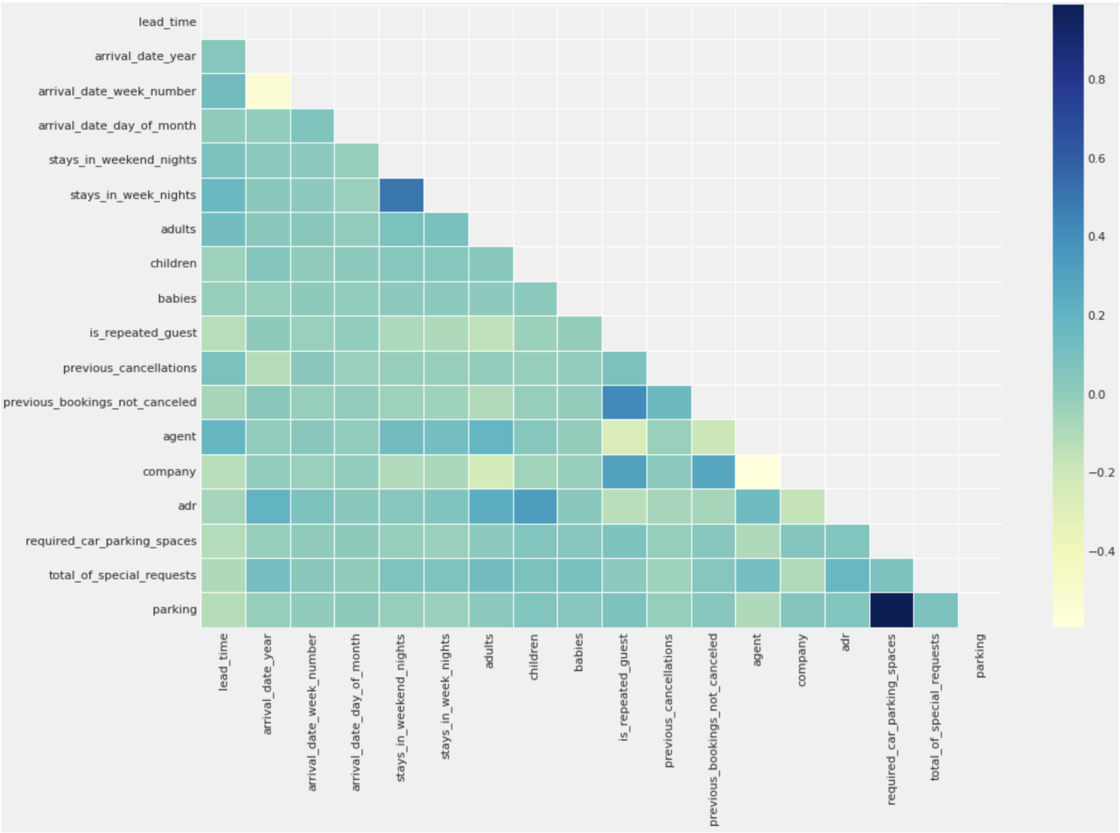

FIGURE 4 – Correlation Heatmap

The correlation heatmap helps us understanding better the linear relation between features and the target (i.e at least one parking space was asked).

First of all we can notice that there is little correlation between features (from -0.5 to 0.5). This means there is a small "intercorrelation" between variables : the dataset is not full of redundant features that express the same information, which means we need to study all the relations with our target variable.

Then we can say that the most interesting relations between our target column and numerical features are the following[1] :

- The number of special requests is likely to increase the will for a parking spot, furthermore, offering a parking spot to demanding clients might be not an issue as they won't overreact for an unecessary notification.
- A repeated guest is more likely to ask for a parking place.
- Clients with bigger lodging are more likely to ask for a parking spot.
- Booking through an agency might indicate a lower need for parking space.
- The longer between the booking date and the arrival, the lower the need for a parking space, we might explain this by a lower urgency.
- Obviously, features such as number of children and babies might explain the need for comfort, resulting into the ask of a parking spot too.

Now that one have in mind some theoretical effect about numerical feature, let us dig further into the most interesting ones and combine our analysis with a categorical features study

Further analysis

3.1 The type of hotel

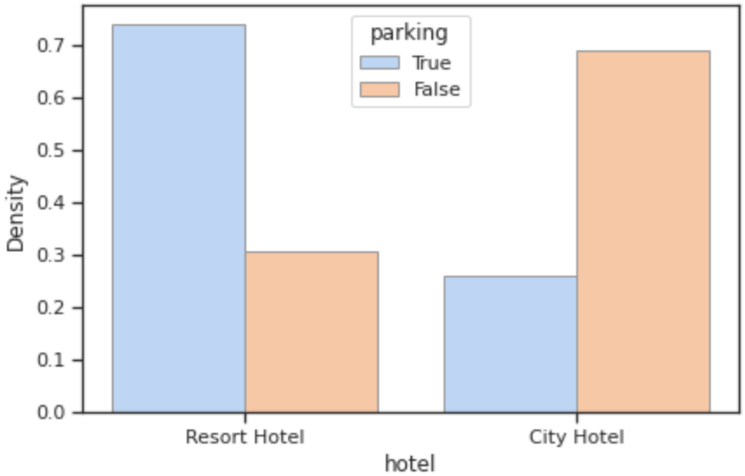

Figure 5 – Demand for a parking spot in the two types of hotel

We can see an over representation of people requiring a parking spot in the Resort hotel.The following table conforts us in this idea. Indeed, Resort Hotel clients are really more likely to ask for a parking spot (17.6% of the clients, vs only 4.7% for the city hotel).

Hotel	Percent Parking
City Hotel	4.734294
Resort Hotel	17.602953

[1] please refer to the notebook to see the most correlated features with the target

3.2 Children and babies

First of all, we have analyzed the two distinct variables "babies" and "children". We can see that having at least one child or baby increases the probability of requiring a parking spot.

We have also decided to look into the relationship between the fact of having a child, demanding a parking spot and the type of hotel. The following graph shows us that clients with 3 childs in the Resort hotels ask for a parking spot with a probability of 35%. We can also say that sending an sms to a client without a child in the city hotel is unnecessary considering that less than 2.2% of them ask for a parking spot.

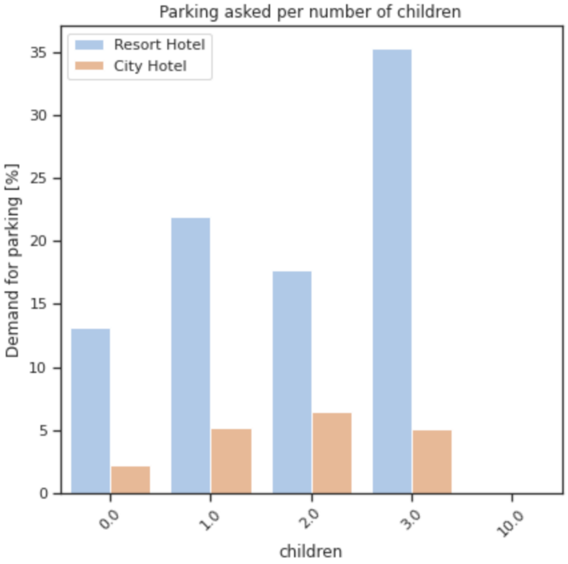

FIGURE 6 – Children in the type of hotel VS parking spot demand

We can conclude that there is a significant difference in parking ask wether you have a child or not. However, considering the small number of clients with more than one child or baby and requiring for a parking spot, and to avoid overfitting, we have decided to reencode theese two variable into just one. The new variable is called "have_child" and is equal to 1 if the variables "babies" or "children" are at least equal to 1, and is equal to 0 otherwise.

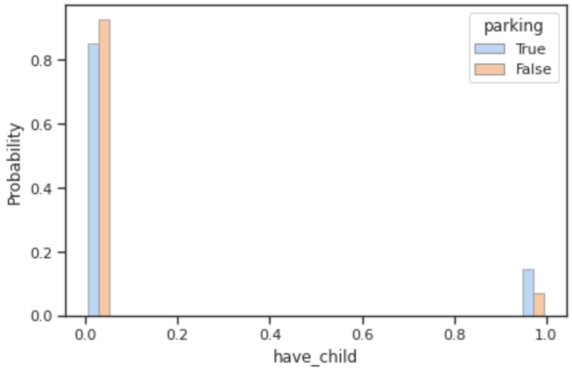

FIGURE 7 – Having a child VS Parking spot demand

For the rest of the study, we will keep this new variable and delete "babies" and "children". We also show in the notebook that this new encoded features offers a better correlation with the target.

3.3 Date columns

Date columns are :

lead_time
arrival_date_year
arrival_date_month
arrival_date_week_number
arrival_date_day_of_month
stays_in_weekend_nights
stays_in_week_nights

We can eliminate year column as we don't care if there was more demand in a certain year in the past.

Week number, day of the month are too precised : we also eliminate these variables.

When looking at the correlation matrix, one can also see that time variables such as 'arrival_date_year', 'arrival_date_week_number', 'arrival_date_day_of_month' have very little correlation with the parking demand : this confirms that there is no need to keep this features.

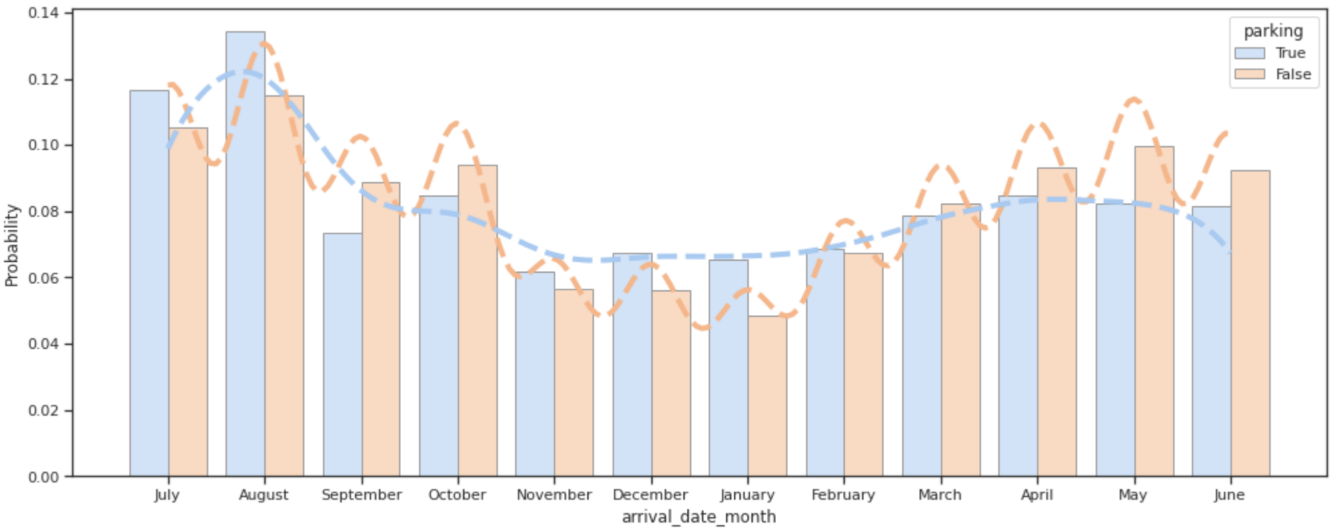

FIGURE 8 – Month of reservation VS parking spot demand

In the above plot, we can see the probabilty of asking (or not) for a parking space over the months. We see here that there is a higher demand in July and August. We might tend to believe that it is useful to send more SMS during this period.

3.4 Customer type

In this section we seek to understand whether there is a relationship between requesting a parking space and the type of consumer.

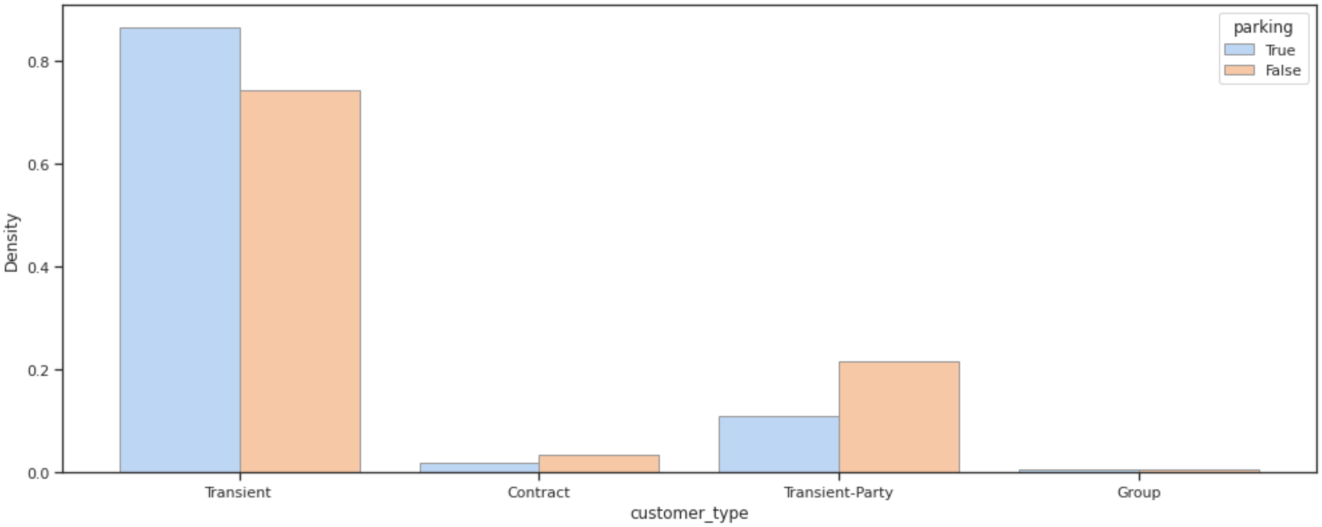

Figure 9 – Customer type VS parking spot demand

According to the above plot, we can say that not all type of customers have the same habit concerning parking spot demand. Indeed, more than 82% of the pepole asking for a parking spot are transient customers for example.

Secondly, we have studied the behavior of repeated guests.

is_repeated_guest	n_booking	n_parking	percent_parking
0	115580	6798	5.881640
1	3810	618	16.220472

According to the previous table, we can say that the percentage of people asking for parking varies wether or not the client is a repeated guest. As a hotel, you might offer more services to your repeated guest and send an sms for parking.

3.5 Country

In order to study the country features, we have deciced to built two lists : the list of countries that ask the most for parkings and the list of countries that ask the least for parking. The point here is to study the differences between the two lists. If the lists look quite the same, one might conclude that tere is no major difference regarding the country.

We observed that 9 countries were in common in the two lists. We concluded that this variable is not interesting for our modeling. We decided not to keep it.

3.6 Reserved room type

One variable that could be interesting for our modeling is the type of room reserved. Indeed, we could think for example that guests who book a superior room would have more money and therefore be more likely to book a parking space.

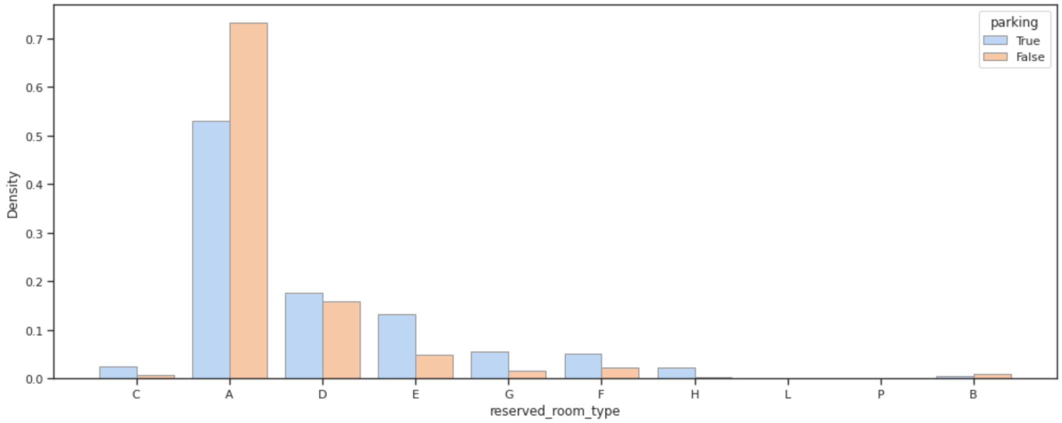

Figure 10 – Reserved room type VS parking spot demand

We can see that people who reserved room type C, E, D, F, H, G are more likely to ask for a parking spot.

3.7 Special requests

Finally, the numer of special requests might have an impact on wether or not you might offer a parking spot. We see that there is an increase in the demand for parking with the number of special requests, as suggested by the correlation matrix. Furthermore, you might already been in a dialogue with a custoemrs that have special requests which would make them less likely to be pissed by a parking notifcation.

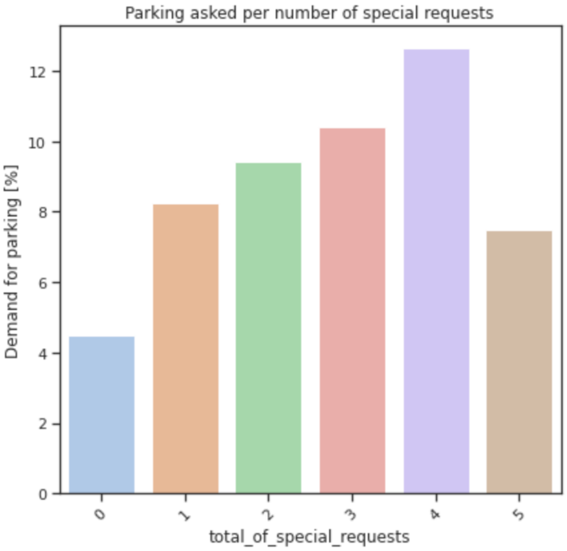

Figure 11 – Special request VS parking spot demand

Modelling

3.8 Evaluation metrics

In our problem, we would really avoid sending an sms to people we are not sure about whether they need a parking space or not. That is to say, we really want to avoid bothering someone who isn't interested in a parking spot.

Statistically speaking, we really want to avoid false positive (positive implying sms sending). The best metric to be sure about our positive predictions is **precision**, which we will use to select the best model. We'll watch the accuracy too, but prefer precision over it.

3.9 First model : logistic regression

We first tried a logistic regression and did the same preprocessing as the first part of this TP. The main advantage of Logistic Regression is to output probability which will clearly help us in the decision of sending an SMS. We obtain a very poor precision of **0.028** but a good accuracy of **0.94**.
This can be explained by the fact that our dataset is totally imbalanced. There are so few true positives that a lot of our predictions are false. In order to balance our dataset, we will use resampling method, more precisely we will use oversampling method. This method enables us to give more weight to the less observed class.

Before using oversampling method, we could try the already implemented method of **class weight = "balanced"** in the logistic regression from sklearn. Doing so, we already improve a lot the precision (**0.46**) but the accuracy (**0.75**) drop by almost 0.2.

3.10 Oversampling method

Here, we use the oversampling method from imblearn library [2] called SMOTE. The idea is the following : generate instances wanting a parking place that "look like" the one we have in our dataset based on their features.
With a logisitic regression combined to oversampling, we have a good trade-off between precision (**0.56**) and accuracy (**0.85**). Indeed, compared to the last method we increased our precision without degrading too much our accuracy.

3.11 Results

The following table summaries our results.

Models	Precision	Accuracy
Logistic Regression	0.028	0.94
Logistic Regression (balanced)	0.46	0.75
Logistic Regression (SMOTE)	0.56	0.85

[2] see see https://imbalanced-learn.org/stable/ for documentation

3.12 Feature importances

The idea of this part is to discuss our first suppositions on feature importance with the real feature importance from our model.

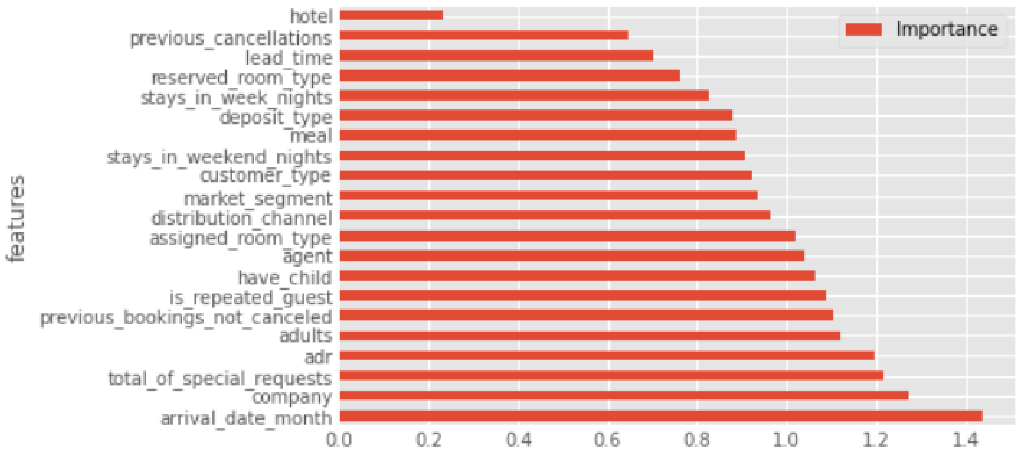

Figure 12 – Feature importance

The importance are obtained by taking the exponential of the coefficients obtained through Logistic Regression. Here the intuition is to compare each feature importance to another but this should not be interpreted as causal effect ! Even if many features we discussed earlier have a huge importance such as the month (seasonality), whether or not the reservation was made by a company, if the customer have child, some features are unfortunatly not as important as we would think like the hotel type. The next steps would be to try models using the best features we have (the most importants ones) and do some feature selection with this method.