Price recommender for new listings in Airbnb Barcelona

Final Project Visual Analytics – UPF 2020/21

Javier Rando Ramírez

Table of Contents

1 INTRODUCTION 2

2 DESCRIPTION OF THE PROJECT 2

2.1 Data Acquisition 2

2.2 Visual Analysis of Tourism in Barcelona 2

2.3 Model design 2

2.3.1 Airbnb data exploration and transformation 3

2.3.2 Training a model 3

3 RESULTING TOOL AND USAGE 4

4 CONCLUSIONS 5

5 ANNEX 6

5.1 Visual Analysis of Tourism in Barcelona 6

5.2 Tourism and visitors overview 6

5.2.1 Airbnb evolution 7

5.2.2 Geographical analysis by neighborhoods 7

5.2.3 Conclusions 8

5.3 Data analysis and transformation 8

5.4 Data transformation 9

5.5 Data analysis 10

5.6 Models discussion 12

5.6.1 Predicting minimum price 12

5.6.2 Predicting total price 15

6 References 17

1 INTRODUCTION

The objective of this project is creating a **recommender system** to help hosts choose the right price for their new accommodation in Airbnb. The scope of the tool is limited to the city of **Barcelona** in Spain. Furthermore, each recommendation will be accompanied by an **explanation of why this price was chosen** by the model, using SHAP values [1]. This second step will help users understand the black-box recommender and decide whether to trust it or not.

The main motivation for this tool is **helping users that are new in the platform to establish a baseline for their accommodation price**. Airbnb current guidelines for this purpose are very limited and not helpful at all [2]. I consider this service a major step to make Airbnb more accessible for everyone and improve substantially the user experience. Also, the explainability module will play a key role to improve user trust in the system and transparency. To ensure reproducibility, the **whole code and report has been published in GitHub** [7].

2 DESCRIPTION OF THE PROJECT

2.1 Data Acquisition

To build this project, I used data coming from two different sources:

- Travelers in Barcelona. Source: Ajuntament de Barcelona [5].
- Airbnb monthly information since 2015. Source: Inside Airbnb [6].

The first dataset is very simple, containing aggregated number of travellers for each month during the last years. The second one will be discussed in detail in the section devoted to the model. It was scrapped and aggregated using the *Scrapper.ipynb* file.

2.2 Visual Analysis of Tourism in Barcelona

The first step was to understand the context in which this recommender will operate and whether it has potential in the long run. For this purpose, I conducted a visual analysis of tourism in the city of Barcelona and how Airbnb has evolved in the past 5 years. Data sources were presented in the Data acquisition section.

See Annex 5.1 for the discussion about tourism evolution in Barcelona using a Tableau Dashboard.

From the visualization we get a clear understanding of the increasing importance of tourism in Barcelona and the sustained growth of Airbnb in the city. This reinforces the idea of improving the quality of the service.

2.3 Model design

In this section, we will explore the models I have created and the process to load, edit and prepare the data.

The goal of the models is creating a recommendation tool for new accommodations. There are two very important amounts that must be predicted: **minimum price and total price**. Most accommodations offer a price for a minimum number and guests and then charge an extra fee for every additional person up to the maximum number of guests allowed. Our system will give an approximate value for both of them.

2.3.1 Airbnb data exploration and transformation

The very first is understanding the data we obtained from Inside Airbnb. Many variables won't be considered for the problem because they are related to reviews and the calendar. The features considered for the problem and their transformations are presented in Annex 5.2. This step is essential to build a good model. In this use case, it is especially relevant given the origin of the data. It is coming directly from Airbnb where users can input whatever they want. In fact, many outliers and errors were found in the dataset.

2.3.2 Training a model

The following step is training regression models to predict minimum and total price with our data. I used Dataiku to fine-tune the parameters of the different prediction models.

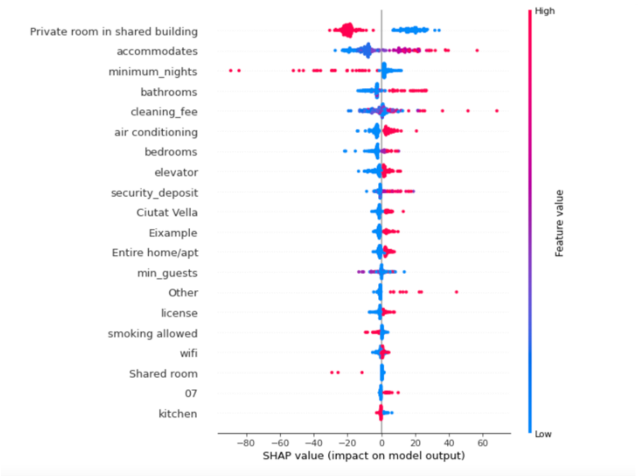
Our data will be splitted into 80% for training and 20% for testing.

I will consider a baseline model that is a simple Linear Regression. Then, I will try different options that improve the performance such as random forests or XGBoost. Other models like neural networks, lasso regression, etc. were discarded using Dataiku. The performance measures to compare the models will be R2 score and MSE. I will also display a scatter plot for prediction versus true values.

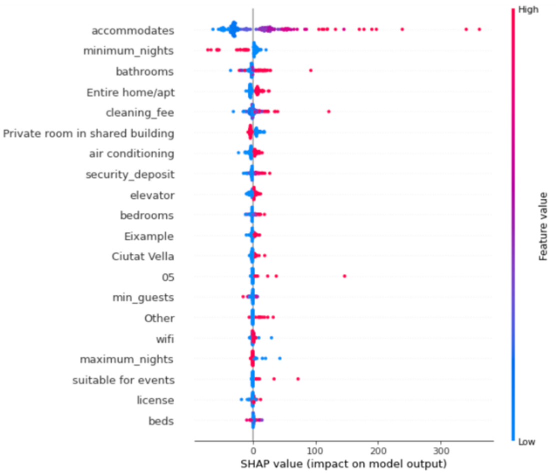
After finetuning the parameters as presented in Annex 5.3, the final models for the tool were **regression random forests** for both variables. These were the obtained metrics.

Total price regression	Minimum price regression
R2 score for test set: 0.677 MSE score for test set: 2580.033	R2 score for test set: 0.604 MSE score for test set: 2122.991

Annex 5.3 also contains the discussion for the model explanation using the following graphs were the value of each feature (red for higher values and blue for lower ones) is presented against the effect they have in the total price. Dots on the right-side show that the feature is contributing to increase the price while the left side represents a decline of the price.



Minimum Price explanation



Total Price explanation

3 RESULTING TOOL AND USAGE

After training the models, I built a compact tool that can be used for prediction. Ideally, it should be integrated with some user-friendly interface. This system will take as input the details for your apartment and generate a prediction. It is implemented in *Execute_recommender.ipynb*. Also, further instructions about execution will be found there.

If you input a listing with the same number of minimum and maximum guests, it will generate only one prediction. Otherwise, it will return two predictions for minimum (hosting just minimum number of guests) and total price (for maximum occupancy) that can be used as a baseline to decide a final price.

The system handles all data transformations required and **input can be typed in a user-friendly way**.

Moreover, the most powerful part is the explanation for these decisions that can be used by the user to improve trust in the system and to understand why he/she should choose this price. The class will return two HTML scripts that are later used to plot explanations.

This would be a **sample execution**.

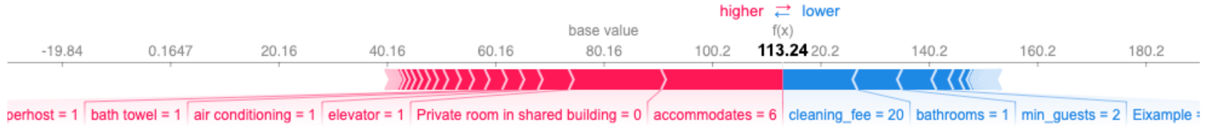
Defining the property

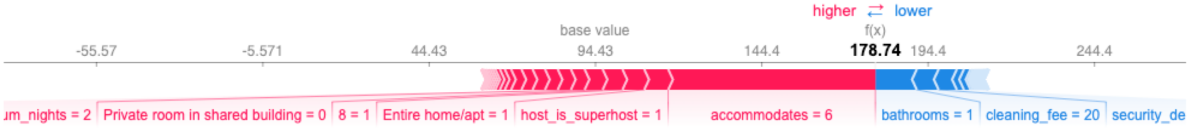
```
accommodates=6
bathrooms=1
bedrooms=2
beds=3
min_guests=2
security_deposit=0.
cleaning_fee=20.
host_is_superhost=True
minimum_nights=2,
maximum_nights=80
license=True
amenities=['air conditioning', 'balcony', 'bath towel', 'tv', 'elevator', 'heating', 'kitchen',
           'wifi', 'family/kid friendly', 'internet', 'essentials', 'washer']
neighborhood='Gràcia'
month=8
type_room='Entire home/apt'
type_accommodation='Apartment'
```

Obtaining the predictions

```
The estimated minimum price for your accommodation is: $113.24
The estimated total price for your accommodation is: $178.74
```

Explaining the predictions





These graphs are equivalents to the ones we saw in Annex 5.3 but are focusing on just one sample. Each model has a base value and then we will see how features push the price away from it. This base value represents the output for random noise or, what is the same, the average for all possible predictions.

Features in red are those increasing the value and variables in blue are decreasing the price of the accommodation. The size of the bins represents importance. The value for each of them is also printed next to the name of the feature.

In this case, we see that only having 1 bathroom for 6 people is a problem and thus, decreases the price. On the other hand, having an entire apartment will increase the price.

4 CONCLUSIONS

In this project, I tried to include visual power to recommendation systems through explanation. Being able to explain to customers why our models are behaving in a certain way will be really important in the feature. This will increase trust and transparency. In this use case, users of the system will easily understand why they should choose a price like asking some kind of oracle that can take into account all variables.

Moreover, I think it is highly valuable because current recommendations and guides for pricing are not focusing on a location and are quite general. Also, exploring similar accommodations can be tedious since there are many variables that must be taken into account. Once the system gives you a baseline price, then you can compare with existing properties if it is reasonable and ease the process.

The main limitation here is the origin of the data since I am getting it from someone else that is scrapping Airbnb. This process might produce information loss and subtle misunderstanding of the data. Ideally, this platform should scrap Airbnb regularly and retrain the models according to existing information and give more importance to recent samples. Also, the data is input by people and not reviewed, therefore this always yields outliers and errors that cannot be easily spotted.

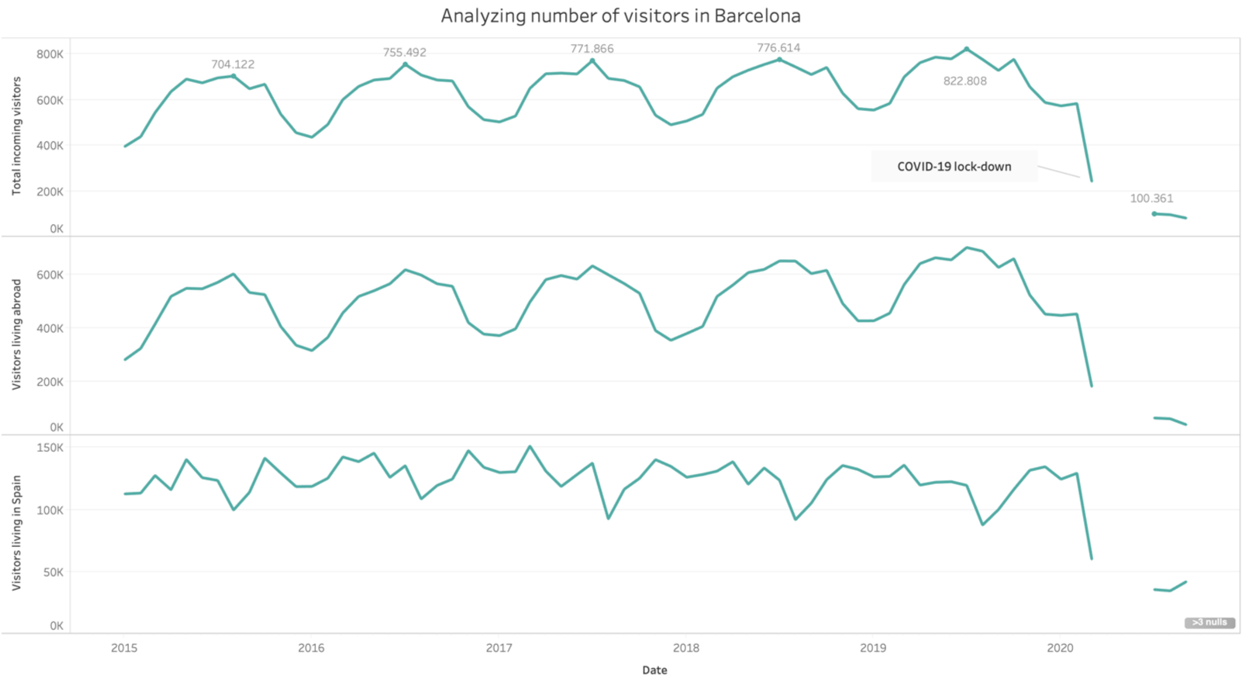
5 ANNEX

5.1 Visual Analysis of Tourism in Barcelona

All the visualizations that will be used along this section are available in this Tableau Dashboard [4].

5.2 Tourism and visitors overview

Tourism is claimed to be one of the most important activities in Barcelona by the city hall generating around 12% of the GDP and 9% of employment [3]. In the following graph we will analyze how many people visit Barcelona, their origin and seasonality of the trends.

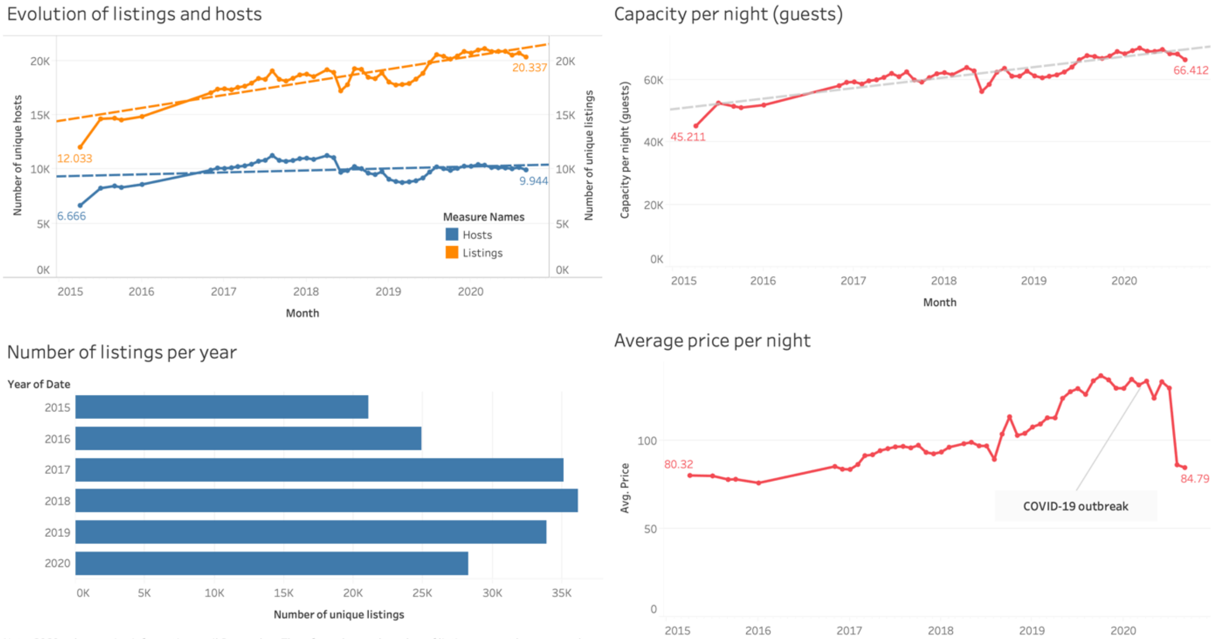


From the first row of this graph, we can easily spot a strong seasonality in tourism having maximum number of visitors every summer. Also, we can realize how in general this number of visitors is increasing every year. The historical maximum number of visitors was achieved in July 2019 with 822.808 people. It is also very relevant to notice how strong has been COVID-19 impact on the city that got only 100.361 visitors in July 2020.

The two remaining charts display how these visitors are divided between national and international travelers. People living in Spain represent a more stable distribution along the year but those coming from abroad are the ones that shape the resulting curve since they represent almost 85% of the visitors.

5.2.1 Airbnb evolution

It is clear that tourism in Barcelona is really important and increasing every year. The next question I had was whether Airbnb is also following a positive trend in order to create a new tool for it. This is what we will analyze next.


Evolution of listings and hosts

Capacity per night (guests)

Number of listings per year

Average price per night

Note: 2020 only contains information until September. Therefore, the total number of listings cannot be compared directly with the previous years.

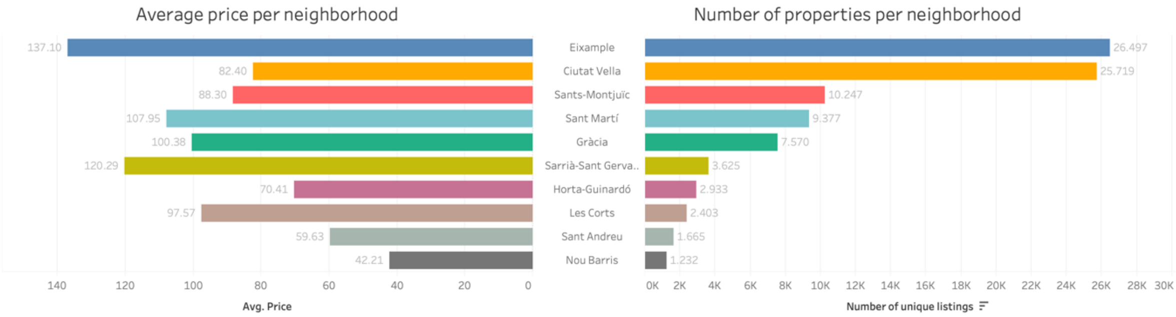
These charts show us some interesting insights about Airbnb in Barcelona. Starting from the top left plot, we see how there is a clear increasing trend both in the number of hosts and properties offered in the city. However, listings are growing much faster than the number of landlords. This is one of the main issues that Airbnb is facing around the world, speculation by big tenants.

In the second graph, we see how the number of guests that can be hosted by Airbnb each night has also grown proportionally to the number of listings along these years. The third one, represents the total number of unique properties that were offered each year on the platform. The maximum was reached in 2018 with 36.176 accommodations. Note that we only have data until September 2020 and therefore 2020 data cannot be compared directly with previous years.

Finally, in the bottom left corner, we find a line representing the average price per night in Airbnb Barcelona. It is also interesting to notice how COVID-19 had a great impact in the average price per night. However, we didn't see such an important decrease in the number of offers.

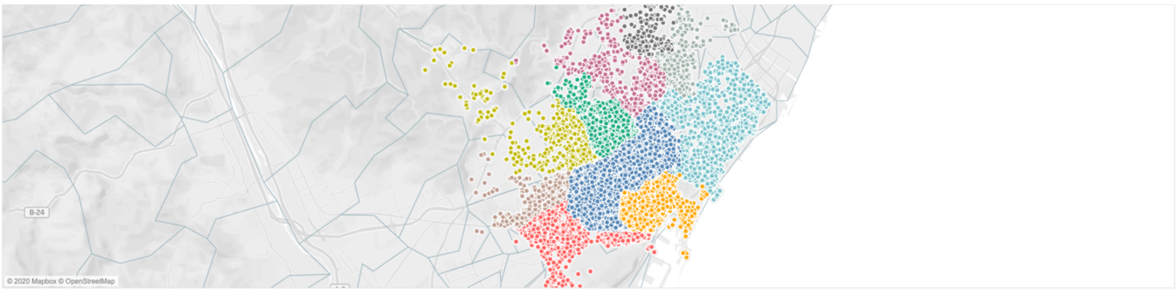
5.2.2 Geographical analysis by neighborhoods

Moreover, I was interested in seeing how properties were distributed in the city. Therefore, I built this last dashboard that shows how many listings are there per neighborhood, which is the average price and the location in the map for each of the neighborhoods and their accommodations.


Average price per neighborhood

Number of properties per neighborhood

Geographical distribution of properties



Eixample is the neighborhood with the largest number of available offers and, at the same time, the most expensive one. This is most likely due to the fact that it is a quite well-located area but still away from the center of the city. Also, buildings here are of higher quality.

Ciutat Vella is the second most popular area but here the price drops significantly. It is interesting because is the most centric neighborhood of the city. Most likely this is because buildings are older, usually they have no elevator, and this area is becoming really busy and noisy.

5.2.3 Conclusions

These graphs were useful to understand the problem and context we are facing. Also, they give the impression that every day more people post offers in Airbnb. Therefore, our tool can be useful in the future for many people.

5.3 Data analysis and transformation

- id: unique identifier of the accommodation.
- host_is_superhost: boolean value that defines whether the host is verified as “super host”.
- neighbourhood_group_cleansed: neighborhood for the property.
- property_type: defines which is the type of the accommodation. for instance, entire apartment, hose, loft, hostel, etc.
- room_type: determines how is the room. possible values: entire home, private room, shared room or hotel room.
- accommodates: maximum number of guests.
- bathrooms: number of bathrooms. 0 means shared bathroom.
- bedrooms: number of bedrooms.
- beds: number of available beds.
- amenities: list of services available in the house.

- square feet: size of the property.
- price: minimum price per night.
- security_deposit: optional amount charged as security deposit
- cleaning_fee: optional amount for cleaning services
- guests_included: number of persons included in the minimum price.
- extra_people: price for additional people.
- minimum_nights: minimum number of nights you must stay.
- maximum_nights: how many nights you can stay at most.
- license: includes the value of the license for the apartment. null if no license.
- host_since: date since the host is registered.
- date: date in which the data was crawled.

5.4 Data transformation

We cannot feed this raw data into our classifier. For this reason, several transformations were considered.

- License. We are not interested in knowing the value for the license but just whether the property has a license to be rented. It will become a Boolean value.

- Apartment type. There are many different values for the apartment type variable but many of them are really rare. For this reason, we selected the most common and aggregated the resting as 'Other'. We will consider the following values: apartment, bed and breakfast, house, loft, serviced apartment, private room in apartment, condominium and boat. Then, they are one-hot encoded.

- Maximum and minimum number of nights. If we analyze, we can basically distinguish two types of renting short-term and long-term. Therefore, the following transformations are performed:

    - Maximum nights becomes a boolean value. 1 determines that you can stay long-term (more than a month), else will be set to 0.
    - Minimum nights also becomes a boolean value. 1 determines if you need to stay at least a week, 0 if you can stay shorter.

- Amenities. These lists must be structured. Again, many values are present and we don't need to keep all of them. I select the most common and those that I consider the most relevant to distinguish properties. For instance, pool is not very common but really descriptive. On the contrary, very common ones as smoke detector don't add value to a hosting. They will be later one-hot encoded. The final values are: 'air conditioning', 'balcony', 'bath towel', 'beach view', 'breakfast', 'elevator', 'essentials', 'family/kid friendly', 'garden', 'heating', 'internet', 'kitchen', 'pool', 'smoking allowed', 'suitable for events', 'tv', 'washer', 'wifi'.

- Neighborhoods and room types are also one-hot encoded.

- Total price was calculated as: price + extra_people*(accommodates-min_guests)

5.5 Data analysis

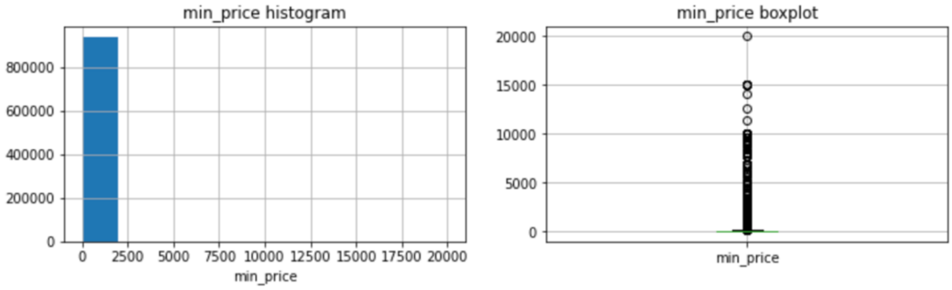
After performing this very first operations to structure our data and clean the different features, we can explore our variables to further improve our dataset. We have 944531 samples. Note that they are not unique since the same listing may appear in different months.

Analyzing null values

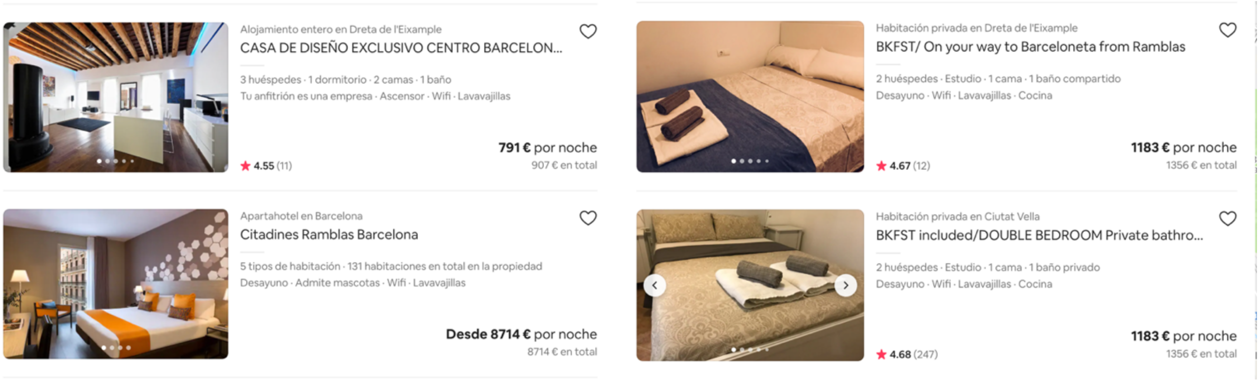
When we see how many null values there are for each variable, we can see that square feet is a very sparse feature. Although it could be a very useful variable, we remove it because we only have 32077 non-null samples. In a way, this information will be encoded by accommodates and number of beds and bathrooms.

Numerical distributions

Now, I will take a close look to our numerical variables. First of all, I explored our target value 'min_price'.



We can see that our distribution for the price has many outliers. We must recall that this defines the price for one night. Definitely, 20000 seems unreasonable. Nonetheless, I validated on AirBnB if there were this kind of accommodations available. If you visit the platform and filter for these prices, you will see very ordinary listings that are definitely not worth that money. This could be a strategy for hosts to hide their place without removing it from the platform. These are some examples.



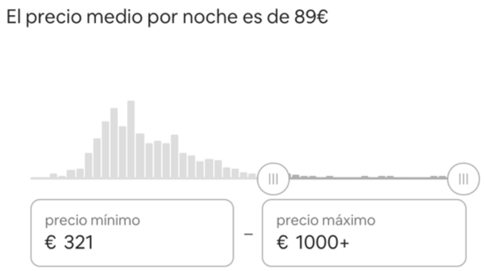
Alojamiento entero en Dreta de l'Eixample
CASA DE DISEÑO EXCLUSIVO CENTRO BARCELON...
3 huéspedes · 1 dormitorio · 2 camas · 1 baño
Tu anfitrión es una empresa · Ascensor · Wifi · Lavavajillas
★ 4.55 (11)
791 € por noche
907 € en total

Apartahotel en Barcelona
Citadines Ramblas Barcelona
5 tipos de habitación · 131 habitaciones en total en la propiedad
Desayuno · Admite mascotas · Wifi · Lavavajillas
Desde 8714 € por noche
8714 € en total

Habitación privada en Dreta de l'Eixample
BKFST/ On your way to Barceloneta from Ramblas
2 huéspedes · Estudio · 1 cama · 1 baño compartido
Desayuno · Wifi · Lavavajillas · Cocina
★ 4.67 (12)
1183 € por noche
1356 € en total

Habitación privada en Ciutat Vella
BKFST included/DOUBLE BEDROOM Private bathro...
2 huéspedes · Estudio · 1 cama · 1 baño privado
Desayuno · Wifi · Lavavajillas · Cocina
★ 4.68 (247)
1183 € por noche
1356 € en total

If we take a look to the distribution in the Airbnb filter, we find that almost the whole distribution is smaller than 300€.

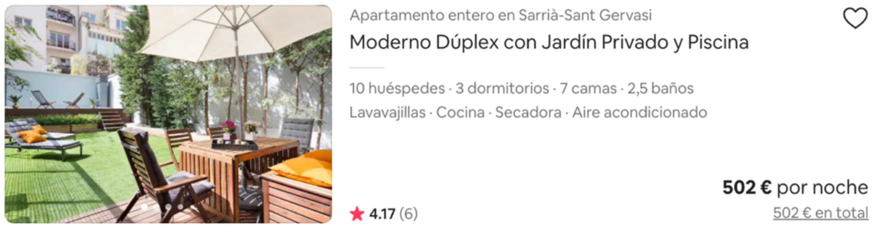


El precio medio por noche es de 89€

precio mínimo
€ 321

precio máximo
€ 1000+

However, we can find really luxurious accommodations such as boats that might be worth that money. In this tradeoff, I decided to generalize for the largest part of the distribution and discard outliers. The threshold was fixed at $600. This way, we can include some expensive listings such as these houses.



Apartamento entero en Sarrià-Sant Gervasi

Moderno Dúplex con Jardín Privado y Piscina

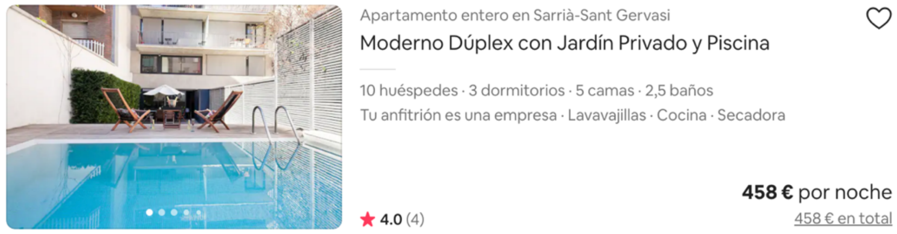
10 huéspedes · 3 dormitorios · 7 camas · 2,5 baños

Lavavajillas · Cocina · Secadora · Aire acondicionado

★ 4.17 (6)

502 € por noche

502 € en total



Apartamento entero en Sarrià-Sant Gervasi

Moderno Dúplex con Jardín Privado y Piscina

10 huéspedes · 3 dormitorios · 5 camas · 2,5 baños

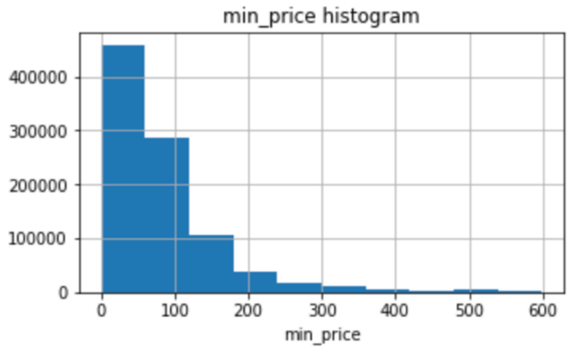
Tu anfitrión es una empresa · Lavavajillas · Cocina · Secadora

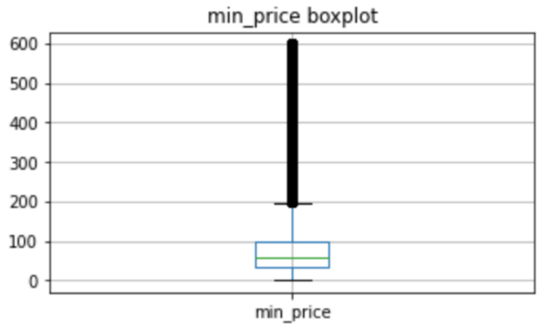
★ 4.0 (4)

458 € por noche

458 € en total

After deleting the outliers, the distributions became more realistic.





At this point, our data has decreased from 944531 to 931124 samples.

Now, I explored other numerical variables (details can be seen in the notebook) and realized several things that should be fixed. For instance there were samples in which beds, bathrooms or bedrooms were 0. This values made no sense at first but I decided to explore when this happened.

- Bathrooms equal to 0. After filtering the data, I could notice that this happened whenever the bathroom was shared. For example, in hostels or shared apartments. Then these rows were preserved. However, those entire apartments with 0 bathrooms were dropped.
- On the other hand, we won't consider properties with 0 beds or bedrooms.

Now, our dataframe has decreased to 889062 samples.
Finally, we must consider that the same property may appear several times for different months. If we train on this dataset, we will overfit because we would already have seen those properties during training for different months. Then, we need to drop these duplicates. This definitive dataset contains 83776 unique listings.

5.6 Models discussion

5.6.1 Predicting minimum price

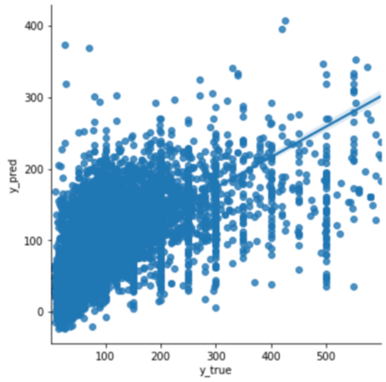
5.6.1.1 First approaches

First, I will build a model to recommend a minimum price for a accommodation. As I said, my baseline model will be a Linear Regression.

Linear Regression

R2 score for test set: 0.42448447526638877
MSE score for test set: 3145.103835862317

This model seems to fit somehow our data but it is still far from a good performance.

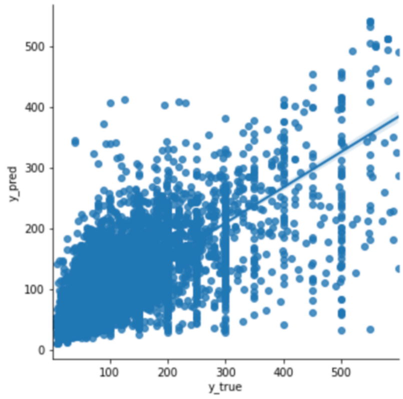


Random forest

Maximum depth: 20
Minimum samples per leaf: 5
Number of estimators: 200

R2 score for test set: 0.5951198353489635
MSE score for test set: 2212.6078345114233

The performance has increased considerably from our baseline. However, we may still improve our data for better performance.



5.6.1.2 Fine-tuning our features

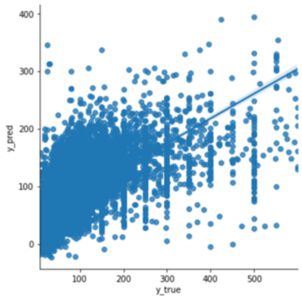
To improve our predictions, I considered change our features slightly. The following transformations will be done:

- One hot encode for months.
- Year will be dropped to ensure generalization.
- Numerical variables will be standardized.

After these changes, we can revisit the models.

5.6.1.3 Prediction with updated features

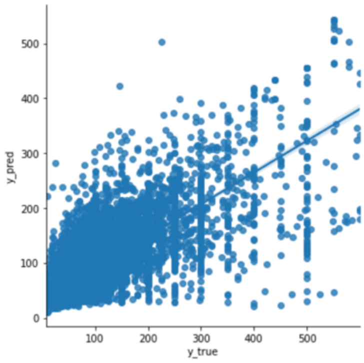
I will revisit the previous models to see if there was some improvement and try some further fine-tuning at the end.



<u>Linear Regression</u>

R2 score for test set: 0.4286256649385076
MSE score for test set: 3066.507788450275

We have slightly improved our model by reducing the MSE.



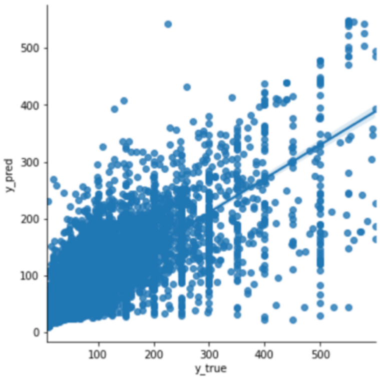
<u>Random Forest</u>

Maximum depth: 20
Minimum samples per leaf: 5
Number of estimators: 200

R2 score for test set: 0.5933454313515583
MSE score for test set: 2182.4735999651234

<u>Random Forest fine-tuned</u>

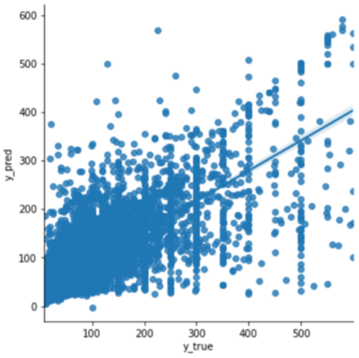
Using Dataiku, I found better hyperparameters for this new data.



Maximum depth: 20
Minimum samples per leaf: 2
Number of estimators: 120
Minimum samples for split: 6

R2 score for test set: 0.6044286602321987
MSE score for test set: 2122.9910408123687

Now, we are above 0.6 in our R2 score.



XGBoost

R2 score for test set: 0.5882693517307629
MSE score for test set: 2209.716401639593

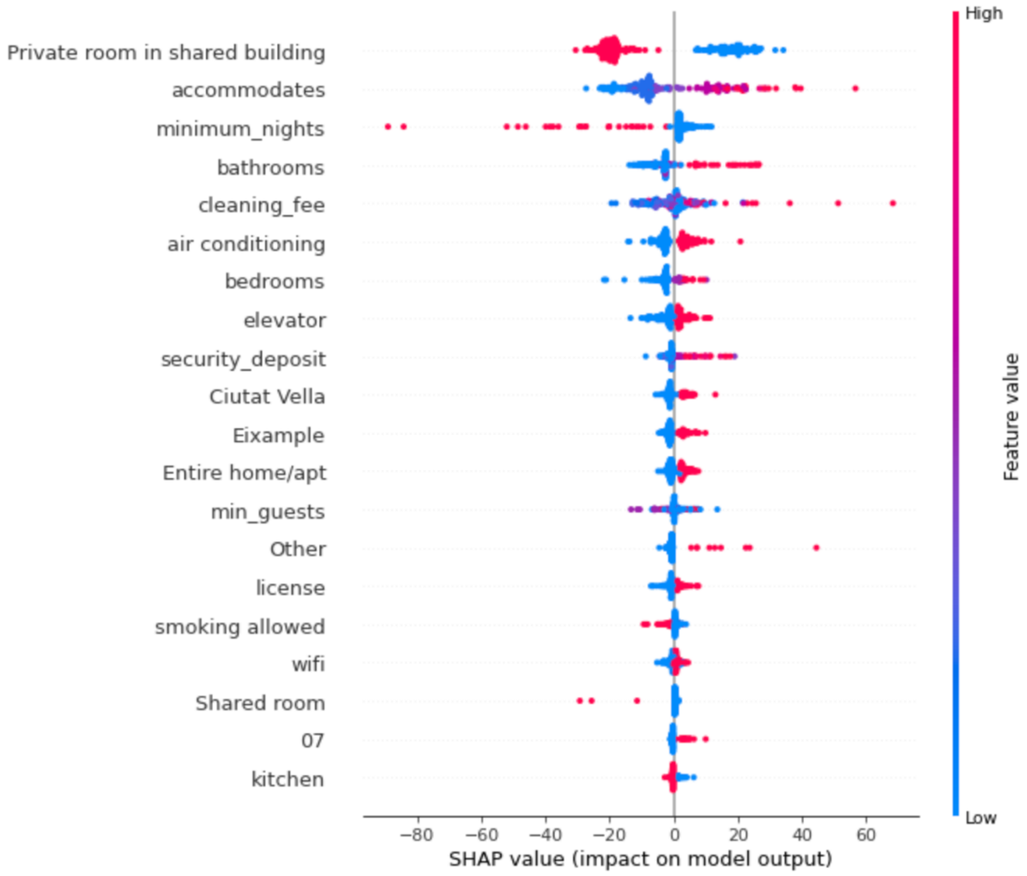
Even after tuning these parameters, the result was no better than the one obtained by the second random forest.

5.6.1.4 *Conclusions*

The second random forest will be stored as my final model. We have values for high true values that are penalizing a lot. Some of them can be accommodations with higher value than expected as we saw before.

5.6.1.5 *Explaining the model*

Finally, I will briefly try to explain the model using SHAP values.



What this plot is representing how a high/low value for a value affects the final output. Red dots represent high values for the variable (for Boolean they will mean 1) while blue are low values

(0 in Booleans). When these dots are placed on the right side of the 0 line, it means they are contributing to higher price. Those on the left contribute to lower prices.

Therefore, it is interesting to notice how important “Private room in shared building” is. Low values, meaning 0 in this Boolean and therefore whole apartments, have higher prices. Also, the number of people they can host is really important. Bigger properties will have higher prices. Recall that Eixample was the most expensive neighborhood and here we can notice how it appears as an important variable to increase the price.

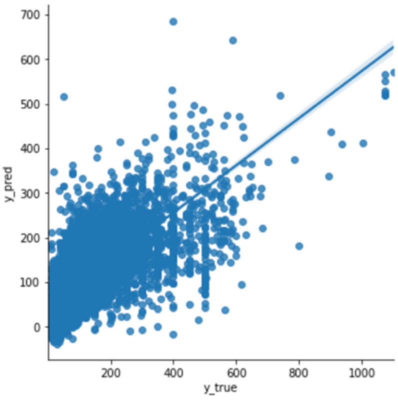
5.6.2 Predicting total price

Now, we will start from our transformed features to predict the total price of the accommodations. Again, Linear Regression will be our baseline.

Linear Regression

R2 score for test set: 0.5268601440544892
MSE score for test set: 3780.0841685037376

These values are much better than those for minimum price. Also, in the scatter plot we can notice that less outliers in the data are contributing to this improvement.

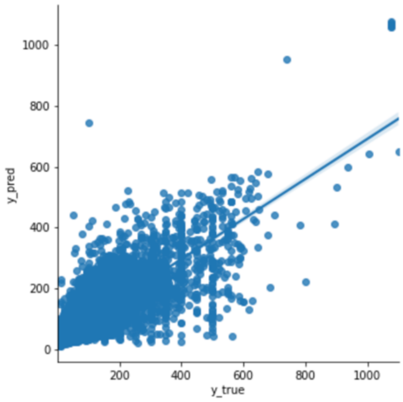


Random Forest

Maximum depth: 20
Minimum samples per leaf: 5
Number of estimators: 200

R2 score for test set: 0.6673353396578764
MSE score for test set: 2657.7773995956804

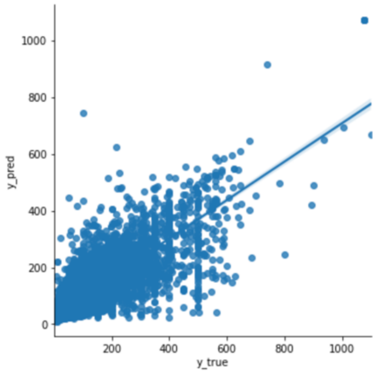
Great improvement can be noticed here.



Random Forest fine-tuned

Maximum depth: 20
Minimum samples per leaf: 2
Number of estimators: 120
Minimum samples for split: 6

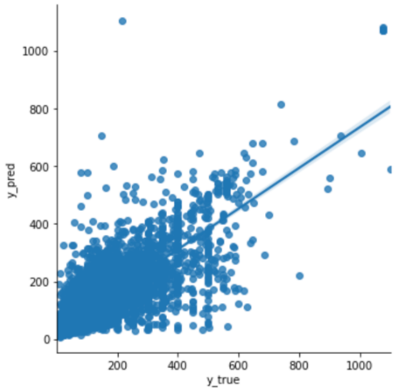
R2 score for test set: 0.6770662120341577
MSE score for test set: 2580.033966754234



XGBoost

R2 score for test set: 0.6708743645665064
MSE score for test set: 2629.5028590746356

Again, now improvements were obtained with XGBoost.

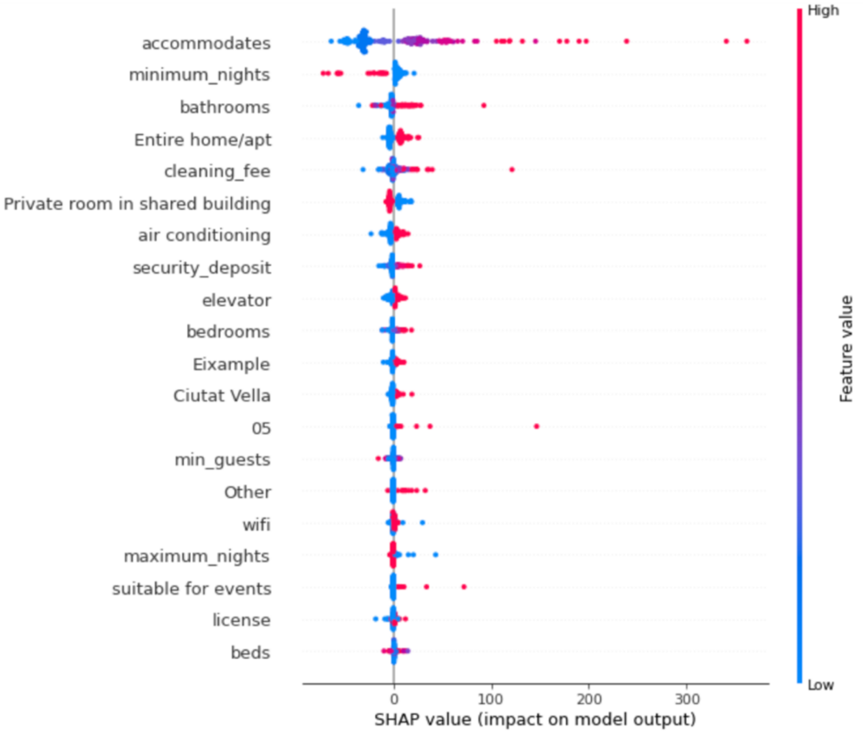


5.6.2.1 *Conclusions*

The second random forest will be stored again as the final model since it has the best accuracy on the test set and also it is quite efficient.

5.6.2.2 *Explaining the model*

As I did before, I will present an explanation for this model based in SHAP values.



Again, we see the importance of the number of people. Here it is even more relevant because the larger it is, the larger the total price will be. Also, entire apartments appear as important variables. Moreover, notice that those apartments with a minimum number of nights have lower prices. This is interesting because they are most likely long-term renting which are cheaper per night.

6 References

[1] SHAP values library for Python: https://github.com/slundberg/shap
[2] Airbnb recommendations for choosing a price in Spanish:
https://www.airbnb.es/help/article/52/c%C3%B3mo-elijo-cu%C3%A1nto-cobrar-por-mi-espacio?_set_bev_on_new_domain=1606414820_MGY3M2E1NjMwMGQ2
[3] Information about turism by Barcelona Town Hall:
https://ajuntament.barcelona.cat/economiatreball/es/turismo
[4] Tableau Public Dashboard: https://public.tableau.com/views/FinalProject-BarcelonaTourismAnalysis/Story1?:language=es&:display_count=y&:origin=viz_share_link
[5] Official travel statistics in Barcelona:
https://www.bcn.cat/estadistica/angles/dades/economia/teoh/evo/th06.htm
[6] Inside Airbnb data: http://insideairbnb.com/get-the-data.html
[7] Github repository for the Project: https://github.com/javirandor/airbnb-barcelona-price-predict