Price Prediction Using Airbnb Data

Prediction Models with OLS, LASSO, CART and Random Forests

Ozan Kaya

February 5, 2021

This simple paper tries to create a fit on the available Airbnb accomodations on Barcelona, using prices along with other variables, to provide a prediction for new listings that had no related market information before. That is to say, we will be trying to use the existing housing prices to find out how a new entry should be priced. This exercise is limited by apartments only and doesn't offer insight for other property types.

As the first step into the exercise, I have tried to transform the raw data into a tabular format and also clean some undesirable notation or symbols. The problem with the data at that point was the custom nature of the way the hosts' enter the amenities information about their listings. I choose to create new variables using the mostly used factors and end up with 15 dummy variables that represents the available amenities. I have also created a bathroom type variable with factors shared and private from a set of strings and dropped numerous other variables that I find irrelevant to our question.

The data I have used can be downloaded from here. This a project by Murray Cox and more information about it is available here.

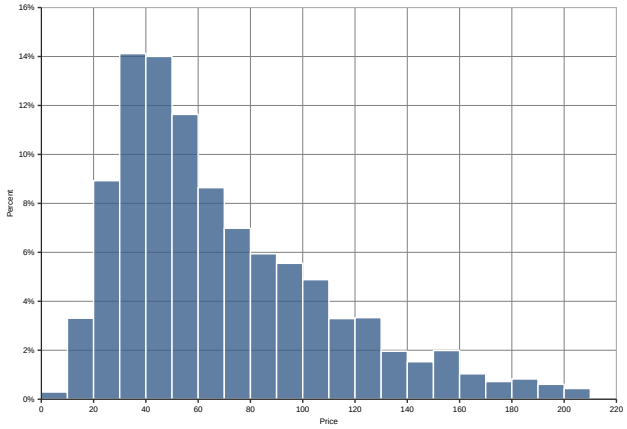
Data

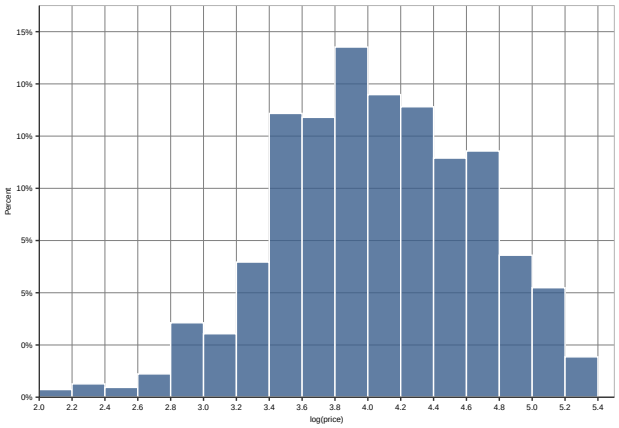
The raw data consists of many categorical variables, binaries and some numerical values. Since we want to predict the price, I will try to understand which variables have the strongest pattern of association with the accomodation price. At first glance, the accomodation itself, its location, reviews of past users and host of the apartment seems to be the major categories we can include variables from.

Statistics	Price	Accomodates	Bathroom Count	Review Count	Review Scores
Mean	85	3	1	34	92
Median	59	3	1	4	94
Std	196	1	0	67	9
IQR	59	2	0	37	5
Min	8	2	1	0	20
Max	9,999	6	4	773	100
numObs	13,924	13,924	13,924	13,924	13,924

Table 1: Selected summary statistics

I start with checking our dependent variable, the price. It seems that a large portion of the observations are under 200 Euros per night. However there are distinct outliers such as 10.000 Euros for a single night along with a wide dispersion among the observations. The distribution itself, that can be seen below on the left, is skewed and it has a relatively long right tail. A log transformation is also presented on the right and it provides a distribution more closely related with that of a normal distribution.

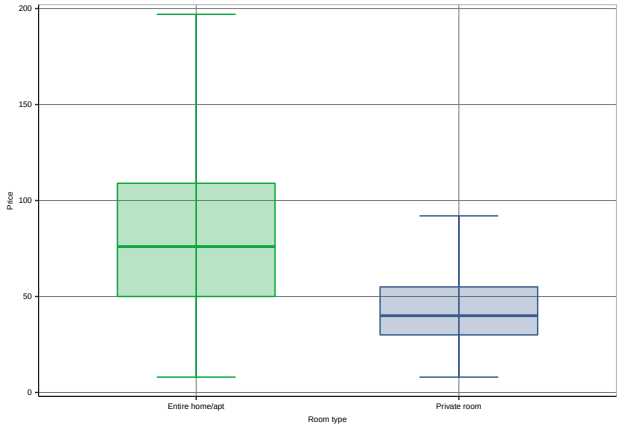


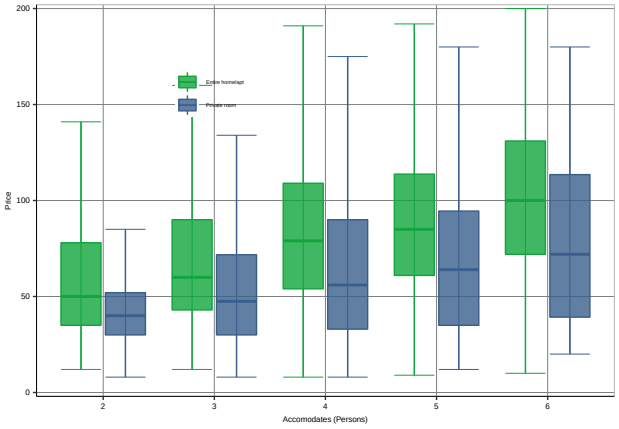


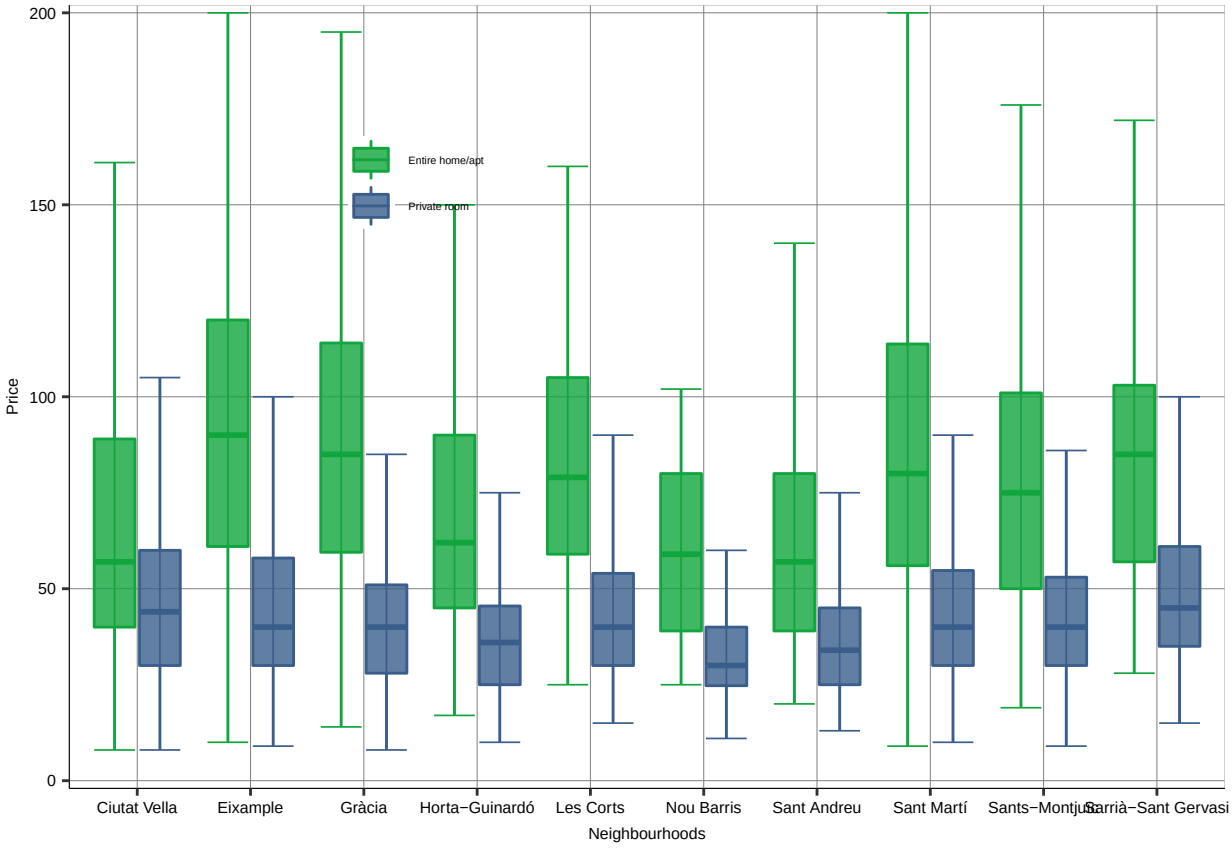
Another important determinant of price seems to be the type of the room. The majority of our observations belong to entire houses but a certain number of listings are for private rooms as well. Not surprisingly, we see a notable difference between their prices that can be onserved on the first graph below.

On the other hand, when we take a more granular look into the observations, we see that with higher number of accomodation opportunuties, the prices are naturally increasing. But interestingly, the dispersion of private room prices are increasing as well. That is mostly due to lower number of observations since low accomodation rates in small houses seem to suggests homeowners' renting out their property whereas large accomodations in private rooms seem to offer a more professional setup. This can be seen at the graph to the right.

By common sense, we can also guess that the neighbourhood of a house is an important determinant of its price. There is a stark distinction between some of the neighbourhoods but this is fairly visible with the entire apartment prices. For private rooms, the difference is visible but relatively less notable. This can be seen at the bottom graph below.







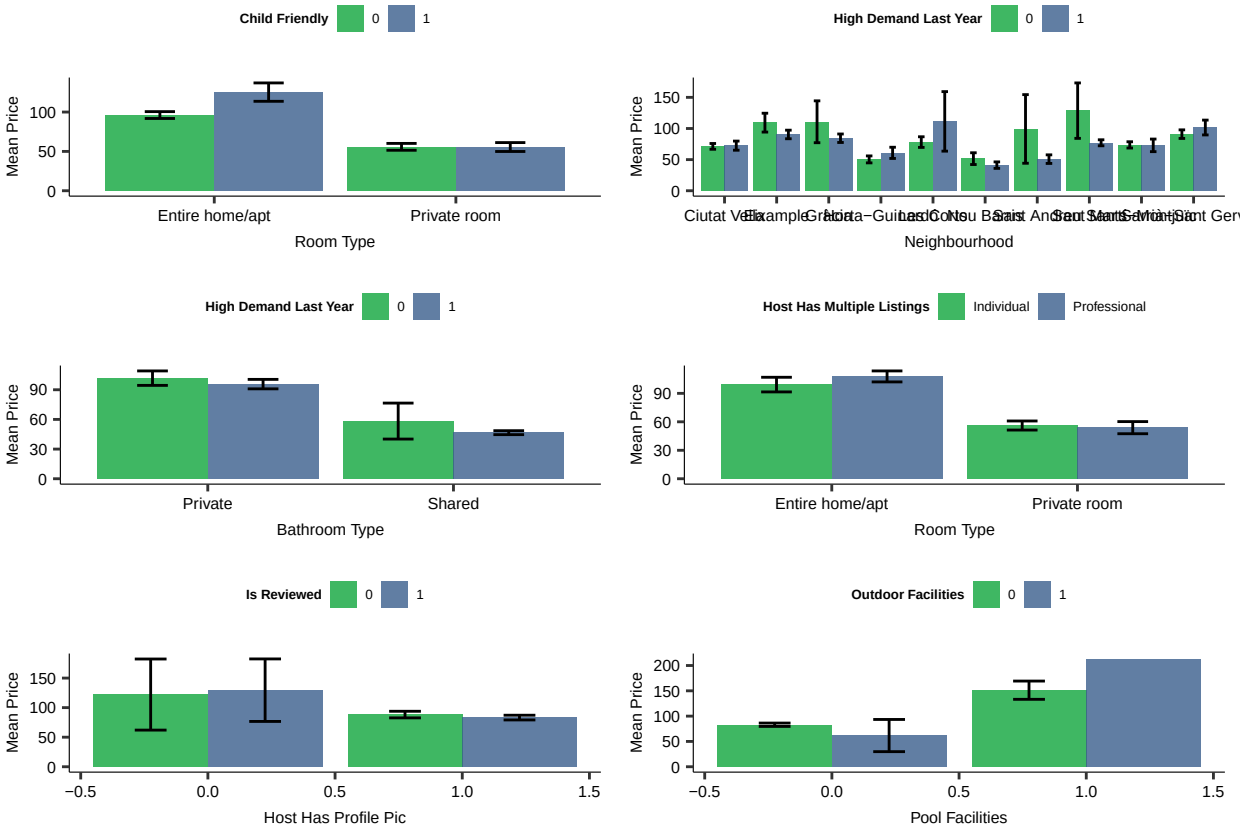
Model

I have started considerin the structure of my linear models by trying to take a look at the interactions between variables as well as trying to check the summary statistics of the numerical variables. Price of any listing seems to be heavily influenced by some simple variables that anybody would think of first when trying to price an apartment. In my initial model, I decided to include accomodation number, bedroom and bathroom counts, bathroom and room types and the neighbourhood.

As discussed before, I also wanted to include information about the host and the experiences of past users into my models as well. While checking relevant variables for this purpose, I have decided to include three more variables to my model. The first one is about the total listings a single host owns. I decided a large number of listings would propose a professional behind the daily operations of this apartments and the pricing mechanisms for these listings might be different than others. The second inclusion was about the review counts. It seems that a good portion of the listings have no reviews. This could suggest that the listing is relatively new and not many people have stayed there so far but it also means the pricing of the listing doesn't reflect the market and people are not demanding it. In either case, the price information might be different for them relative to others. The last variable is closely related to this idea, where I controlled for the availability variable which shows how many days the listing was available in the past 30, 90 and 365 days. I choose the 1 year alternative since there could be serious seasonality regarding airbnb listings and I wanted to avoid that. The new variable is a dummy variable that takes the value 0 if the apartment was available more than 300 days in the past year.

These being dealt with, I have included the dummy variables that shows if the host is a superhost, has a profile picture, if its identity is verified and finally if it has more than 3 listings in total as my variables regarding the host of the apartment. I have later on determined the number of reviews, the review score, is the apartment is reviewed and finally was the apartment available more than 300 days in the past year as my variables about the reviews or namely the past experiences. Another inclusion will be the amenities provided

to the customers such as a pool, outdoor facilities like a barbeque or a garden, a balcony and so on, each represented as a dummy variable. The final addition will be the possible interactions between the variables. I have tried to explore as many alternatives as I can and I tried to plot some of the more interesting ones. The plots can be seen below. At the end, I decided to control for the bathroom types and the neighbourhoods interactions with last years availability measure, as well as interactions with room type and child friendly apartments among others in my models.



OLS

At the end, I have decided to use the log transformation of price as my dependent variable and built 6 different linear models with relatively simple setups. I have started with accomodates as the only variable in a simple regression and slowly increased the number of regressors with each model. The initial model had an R^2 of 0.26 and an RMSE measure of 0.602. Inclusion of the basic variables I have listed above greatly reduced the mean squared errors to 0.578. The most inclusive model with 48 variables had the best test RMSE at the end with 0.570 and also with an R^2 of 0.34. Additionally, when we check our most inclusive model with our holdout set that we have put aside in the beginning of our analysis, the RMSE comes out to be 0.574. Compared to other alternatives, our last model is the best when used in an out of sample analysis as well.

Model	N predictors	R-squared	BIC	Training RMSE	Test RMSE
(1)	1	0.26	20348	0.602	0.602
(2)	14	0.32	19502	0.577	0.578
(3)	18	0.32	19506	0.576	0.577
(4)	22	0.32	19497	0.574	0.576
(5)	37	0.34	19436	0.569	0.572
(6)	48	0.34	19425	0.566	0.570

##LASSO

I later on tried the LASSO method using my most inclusive model and tried to check if LASSO provides a better fit. Since LASSO is particularly useful when there are many variables available relative to observation points, for our case I wasn't expecting a significant improvement in my OLS regressions. Due to this fact, the lambda parameter is choosen as 0.05 by our algorithm and it resulted with 7 non-zero coefficients in the end. However, the test RMSE value was worse then our simple regression.

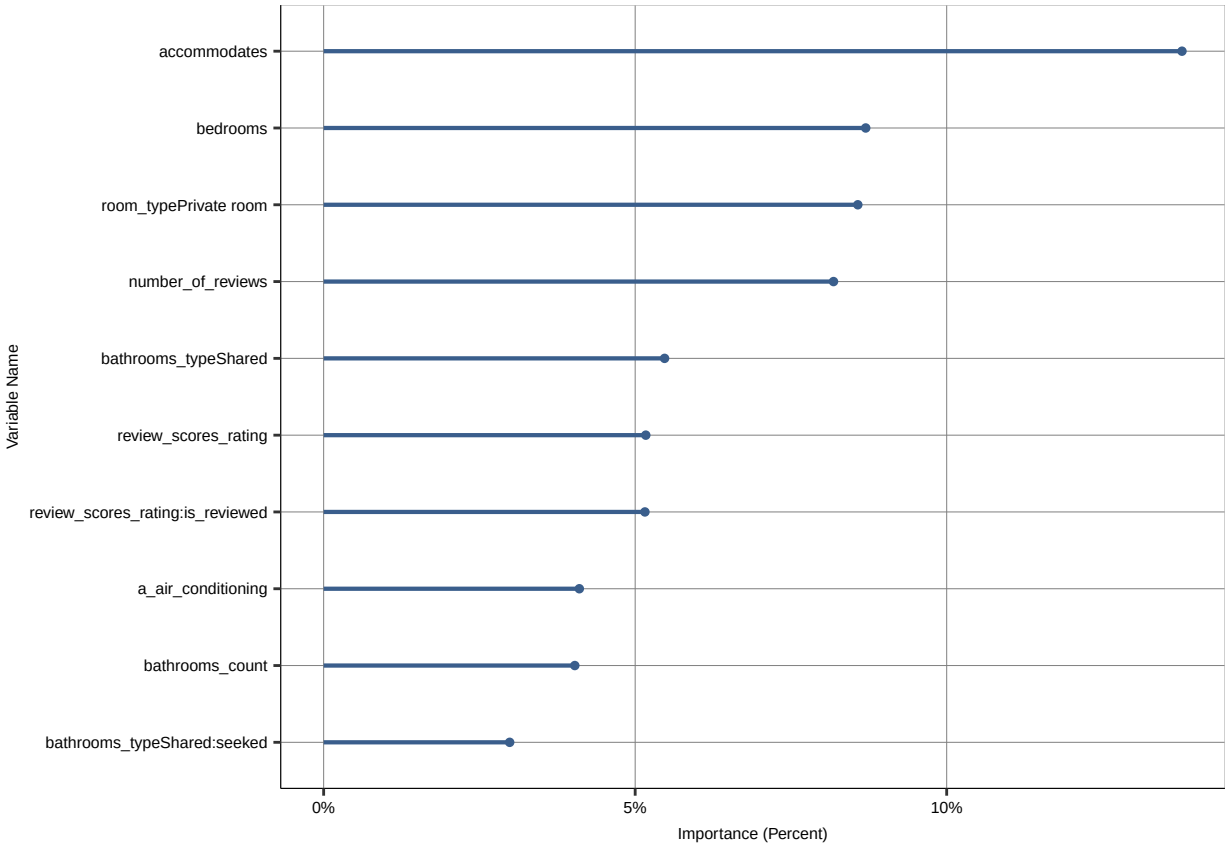
Model	N predictors	R-squared	BIC	Training RMSE	Test RMSE
(1)	1	0.26	20348	0.602	0.602
(2)	14	0.32	19502	0.577	0.578
(3)	18	0.32	19506	0.576	0.577
(4)	22	0.32	19497	0.574	0.576
(5)	37	0.34	19436	0.569	0.572
(6)	48	0.34	19425	0.566	0.570
7	7	0.26			0.606

CART and Random Forest

After exploring the more conventional methods, I also decided to implement regression trees into my analysis. I begin with a simple CART with my most inclusive variable set. The previous linear regressions coefficients have shown rather little variance so as expected, CART have performed not teribly but still lacks the random forest's 'wisdom of crowds'. Nevertheless, with a complexity parameter of 0.0032 and an RMSE of 0.581 it is a relatively robust predictive model.

Model	N predictors	R-squared	BIC	Training RMSE	Test RMSE
(1)	1	0.26	20348	0.602	0.602
(2)	14	0.32	19502	0.577	0.578
(3)	18	0.32	19506	0.576	0.577
(4)	22	0.32	19497	0.574	0.576
(5)	37	0.34	19436	0.569	0.572
(6)	48	0.34	19425	0.566	0.570
7	7	0.26			0.606
CART	-	0.31			0.581
Forest	-	0.41			0.538

However, I expect a significant improvement with a random forest where the aggregation of single trees into a tree would naturally smooth the variance that is natural within single regression trees. I wanted to see how the inclusion of other variables effect the performance of my forests and I was amazed by how reliable results that it can produce. With the most basic variable set we have, the forest's RMSE is 0.567 whereas as a result of the inclusion of new variables the model improves significantly and with the most inclusive variable set the RMSE falls to 0.538. This is a significant improvement among all of our models. The interpretability might be an issue with trees or forests of course by the below variable importance plot solves most ouf our problems and offers a good interpretation of the variables.



Summary

In this humble analysis, I have tried to find the association between house prices in the city of Barcelona using the publicly available airbnb listing prices from inside airbnb project. The aim was to provide an educated guess for new entries to the market, that is to say to provide price predictions for houses with given characteristics. I have built linear models and used methods like OLS and LASSO and I also built regression trees and also forests for this task. The end result is the uncontested victory of the random forest followed by conventional OLS estimates.