Report on Airbnb price prediction in Paris

Péter ENDES-NAGY

Introduction

This is a report on *predicting Airbnb prices in Paris.* 4 models were built for predicting (log) USD prices for Parisian apartments, based on Airbnb data.

A LASSO, a CART and 2 random forest models were built - difference in the latter two is autotuning.
The models were prepared for a client looking forward to list mid-sized (2-6 accommodates) apartments on the Parisian market.

The data was downloaded for the month of June (2021) that is among the most touristy periods in Paris. The possibilities are open for fine-tuning the models, so they take seasonality into account.

Data

Data was downloaded from the Airbnb website (link here) , for the date of 07 June, 2021.
The raw dataset contained more than 60k observations and 74 variables.

Raw data was heavily transformed and cleaned before the filtering and meaningful transformations could be carried out. For details on the pre-cleaning, please consult the Technical Documentation available (here).

Retrieving and selecting amenities in the listings was the most challenging task. In total, there were 1010 differently worded amenities in the dataset. Only those were kept that appeared in at least 1000 observations: 75 amenity types in total, each stored in binary variables.

Data cleaning and feature engineering

Without getting into too many technical details (see more in the Technical Documentation), the potential *predictor variables* were inspected, many of them *simplified/pooled and transformed. Missing values* for predictors were mostly imputed with mean/median or a given meaningful value.

The target variable (price in USD) was transformed into log as it follows lognormal distribution. Extreme values above 2000 USD we discarded, as high-end luxury hotels in Paris are in the 1500-2000 USD range, so up to this value, entire (luxury) flats can be expected in future samples. Listings with 0 USD price were considered missing and discarded from the sample.

Our client is interested in renting out mid-sized apartments, fit for 2-6 accommodates, so listings out of this range were discarded.

For the LASSO model, quadratic and cubic forms were introduced for some numerical variables based on loess plots. Interaction terms were also created for the model based on field knowledge.

At the end, we arrived to a clean and filtered dataset of 53497 observations and 106 variables. The key descriptive statistics of the price is as follows:

Table 1: Prices of Parisian Airbnb listings in June 2021

	Mean	SD	Median	Min	Max	Range	P05	P95
Price (USD)	103.39	90.28	80.00	8.00	1800.00	1792.00	39.00	246.00
Price (log)	4.45	0.57	4.38	2.08	7.50	5.42	3.66	5.51

Model building and results

4 Models were built in total.

For the LASSO model, practically all variables were used, as well as quadratic/cubic forms and interactions. As I wasn't interested in building simple OLS models, a LASSO is a straightforward choice, it shrinks the number of variables. The model managed to shrink their number to 74 which is still relatively high.

For the CART and the 2 Random Forest Models (second one run with autotuning), all available variables were used, except for the quadratic/cubic forms and interactions.

	CV RMSE	Holdout RMSE
LASSO	0.3901	0.3979
CART	0.4462	0.4549
Random Forest	0.3799	0.3849
Random Forest (autotune)	0.3766	0.3807

Overall, our best performing model (both on the training, both on the hold out sets) is Random Forest with autotuning, although it didn't lower the RMSE considerably compared to the simple Random Forest model. LASSO was the 3rd performer and CART did the poorest job.

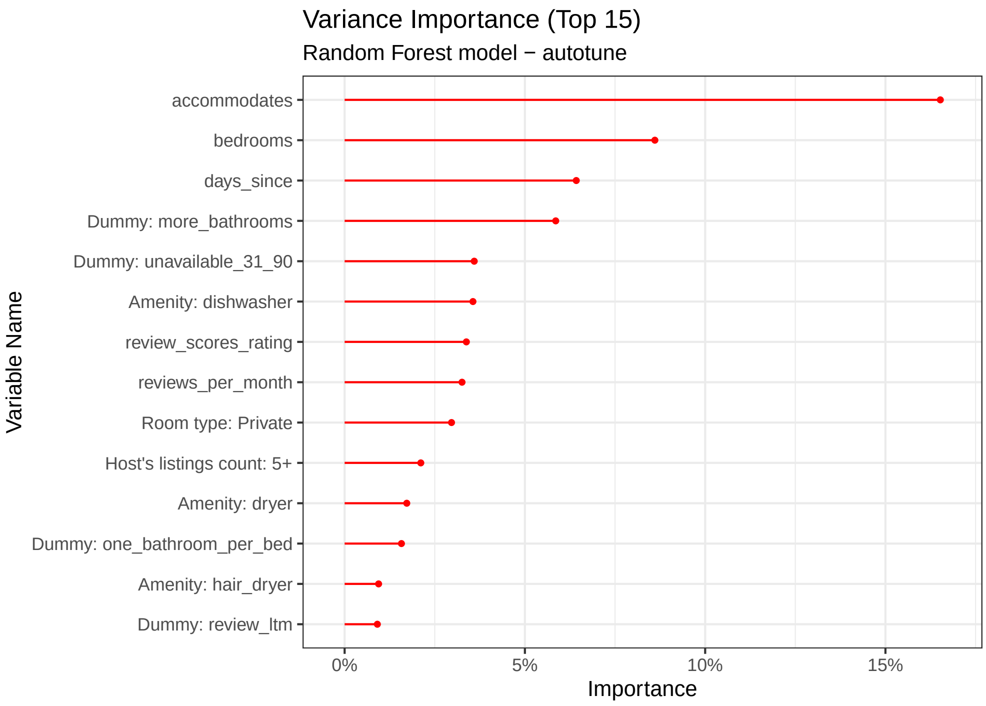
Model evaluation

Random Forest models (especially with autotuning) performed the best, so we are going to focus on it during the evaluation. Model performance plots for CART also included in the Annex.

The biggest disadvantage of Random Forest is that we can't really tell how the given variables are contributing to the predicted value of *price*. We have 2 main ways to identify which features our Clients shall focus on: Variance importance and Partial dependence.

Variance importance

Variance Importance plots show which variables capture the most for our prediction. The following charts display the TOP 15 most influential variables in the autotuned Random Forest model:



The most influential variable in the model is the number of accommodates and the number of bedrooms. Size matters. Number of days elapsed since the first review (implicitly measuring fo how long the apartment had been on the market) is also relatively important, out Client needs to build up their reputation.

The Client should also pay attention to the reviews, both their score both their numbers, both their timing (recently received) matters.The plot also implies that some rather rare/odd amenities are influential, like dryers and dishwashers.

Comfort also seem to matter, having more bathrooms than one and having at least one bathroom per bedrooms influences the price.

Partial dependence plots

The above mentioned variables influence the prediction, but we don't know how. Partial dependence plots helps us having an idea about them. Partial depence plots were created for almost each TOP 10 variables (see Annex). We should keep in mind, that these variables are highly likely to interact with others - e.g. "private rooms are predicted to have a lower price than shared rooms" sounds very unlikely, they have probably very different features that explains the surprising insight.
Both the number of accommodates both the number of bedrooms have a positive effect on the price, but the relationship isn't linear, more rooms/accommodates bring less extra revenue.

The review score is very interesting, until a score of 4 (out of 5), there isn't too much of a difference, but above 4.2, the predicted price increases sharply. Meaning that the Client should pay attention to keep the evaluations at least above 4.2 and keep as high as possible.

Regarding binary variables, having more than one bathroom increases the price just like the presence of a dishwasher. Interesting enough, if a given property isn't available the upcoming 2-3 months, it has a lower price - unfortunately we can't decide why unavailable. Fully booked or temporally retrieved from the market.

Entire homes are predicted to be considerably more expensive and apartments with hosts that have more than 5 listings (professionals, like out Clients), the predicted price is almost 25 USD higher.

Conslusion

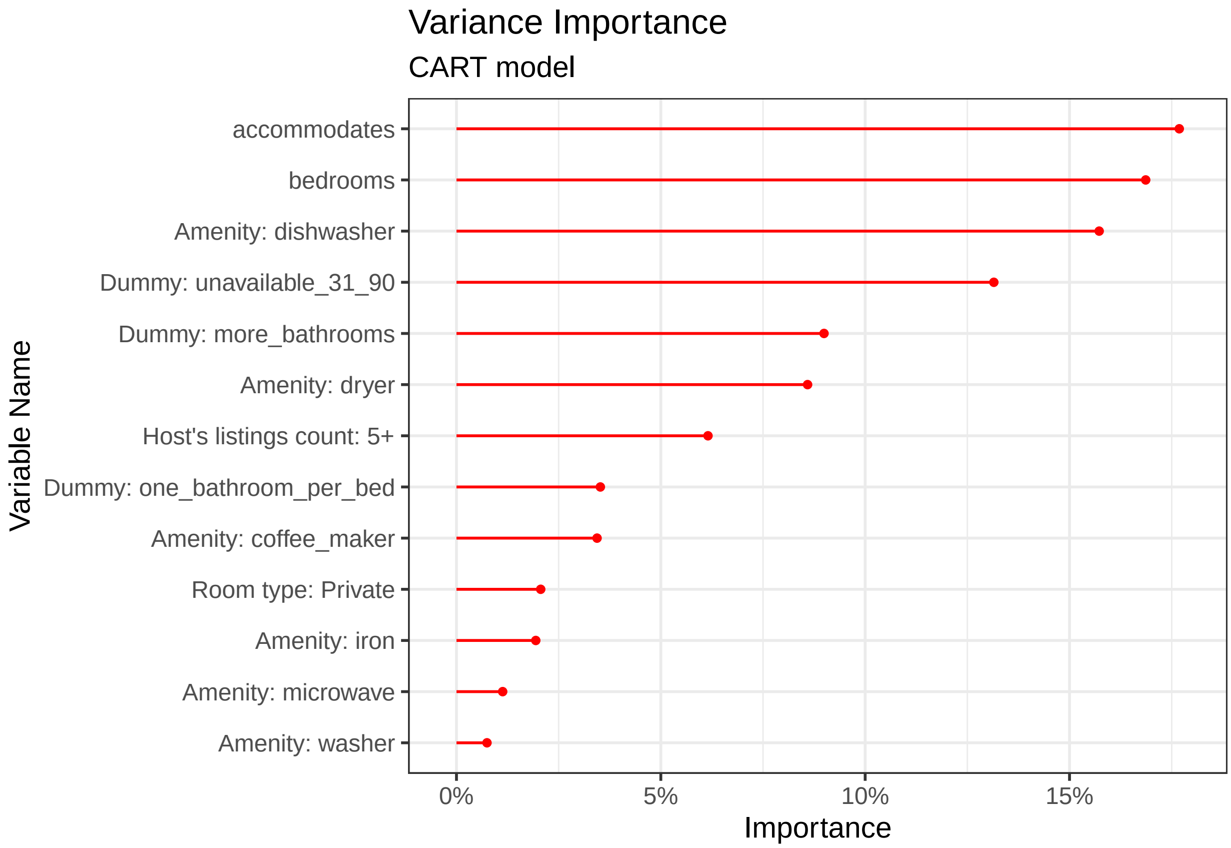
In the report, we investigated Airbnb prices in Paris and built 4 predictive Models for a Client looking forwards to rent out mid-size apartments.

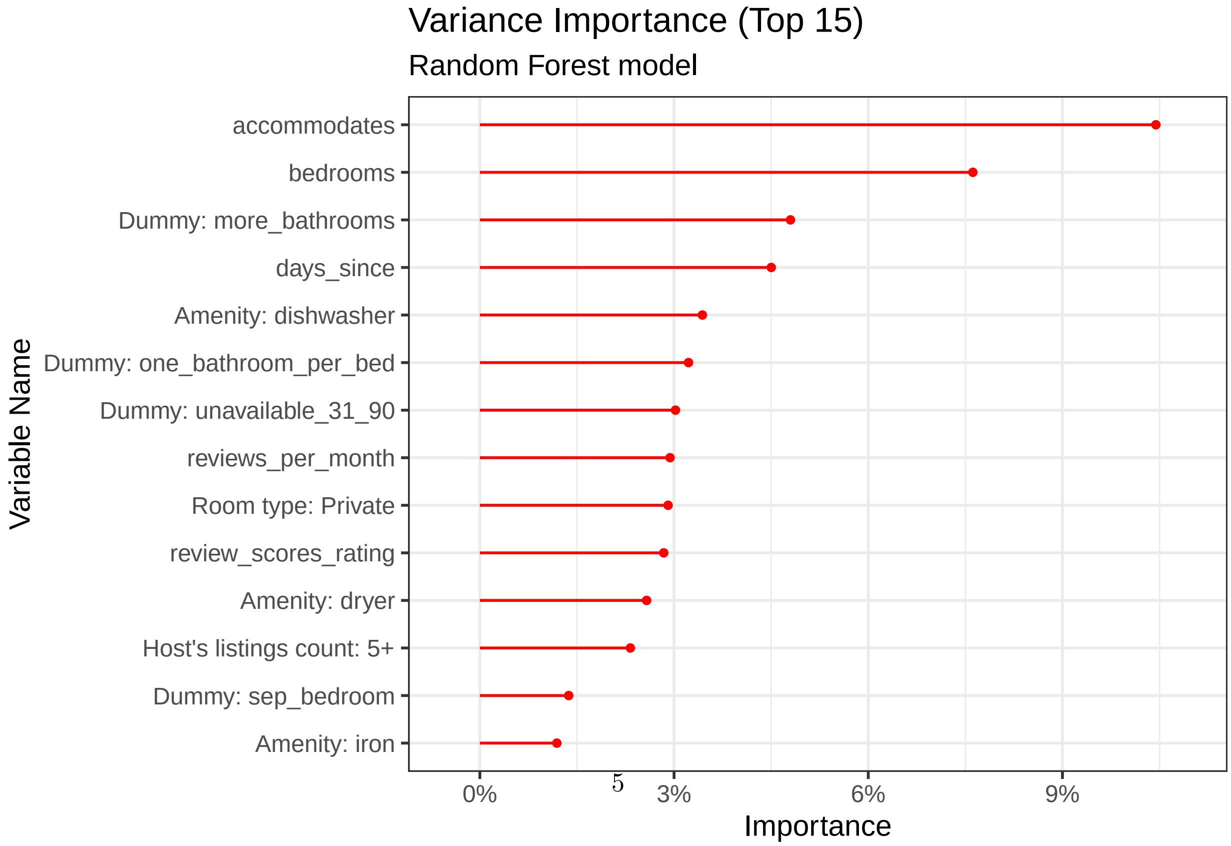
Among the 4 competing models, CART performed the worst, Random Forest (especially with autotuning) was the most accurate with predicting prices.

The variable importance and partial dependence plots suggest, that size matters (larger apartments with higher number of accommodates), but not in a linear way. We also see, that potential travelers care about comfort (see importance of having bathrooms for each room) and some special amenities, like dishwasher and dryers. Our Client also needs to build their reputation and pay attention to keep the review scores at least above 4.2, but the higher the better, so they can join the exclusive group of hosts with more that 5 listings, that are also more expensive in Paris.

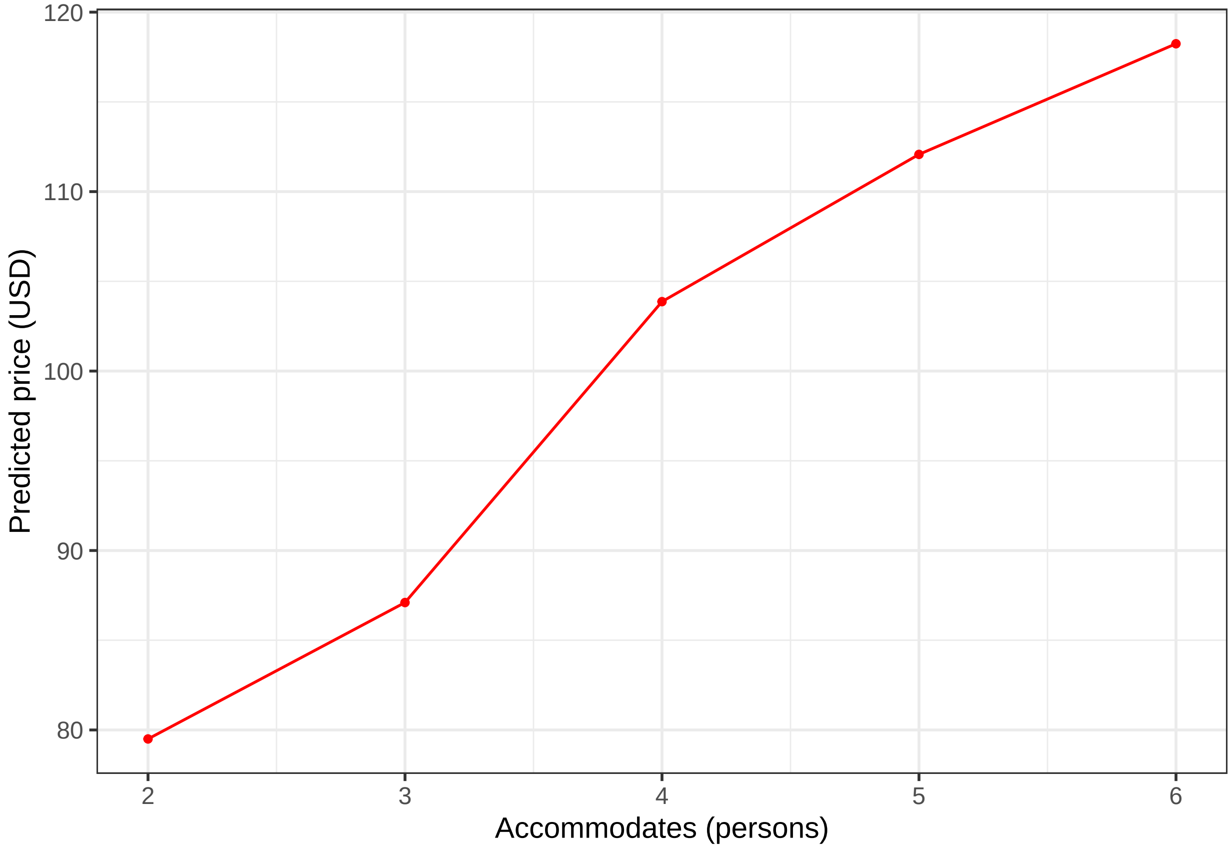
Annex

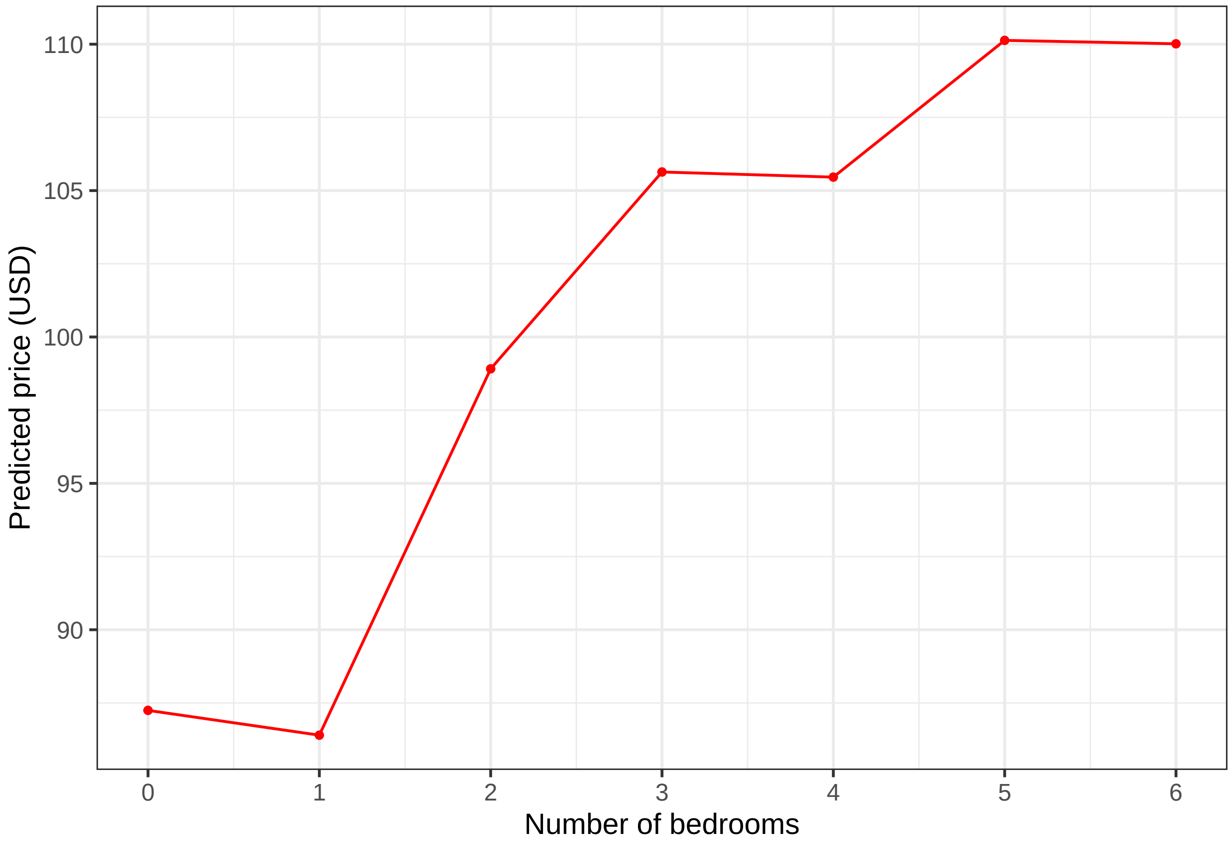
Variance importance plots for CART and Random Forest:

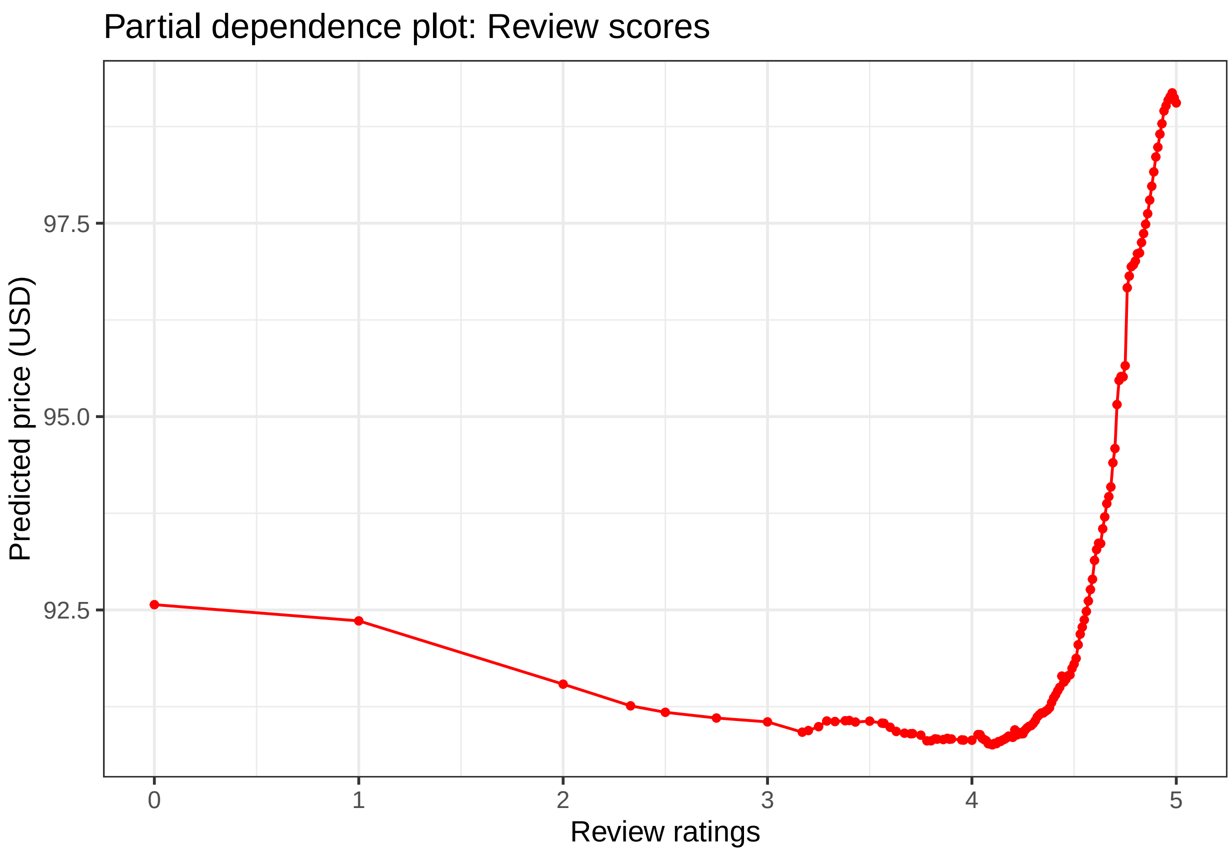


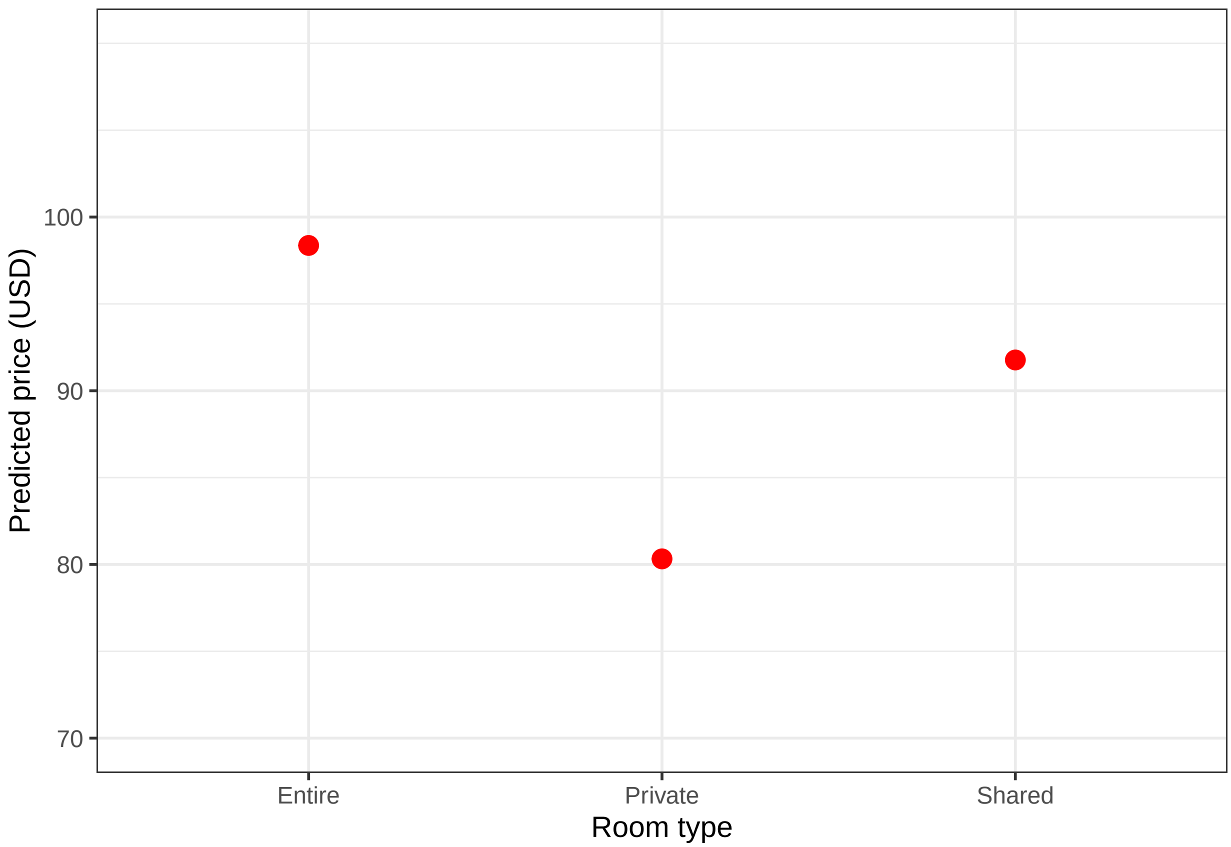


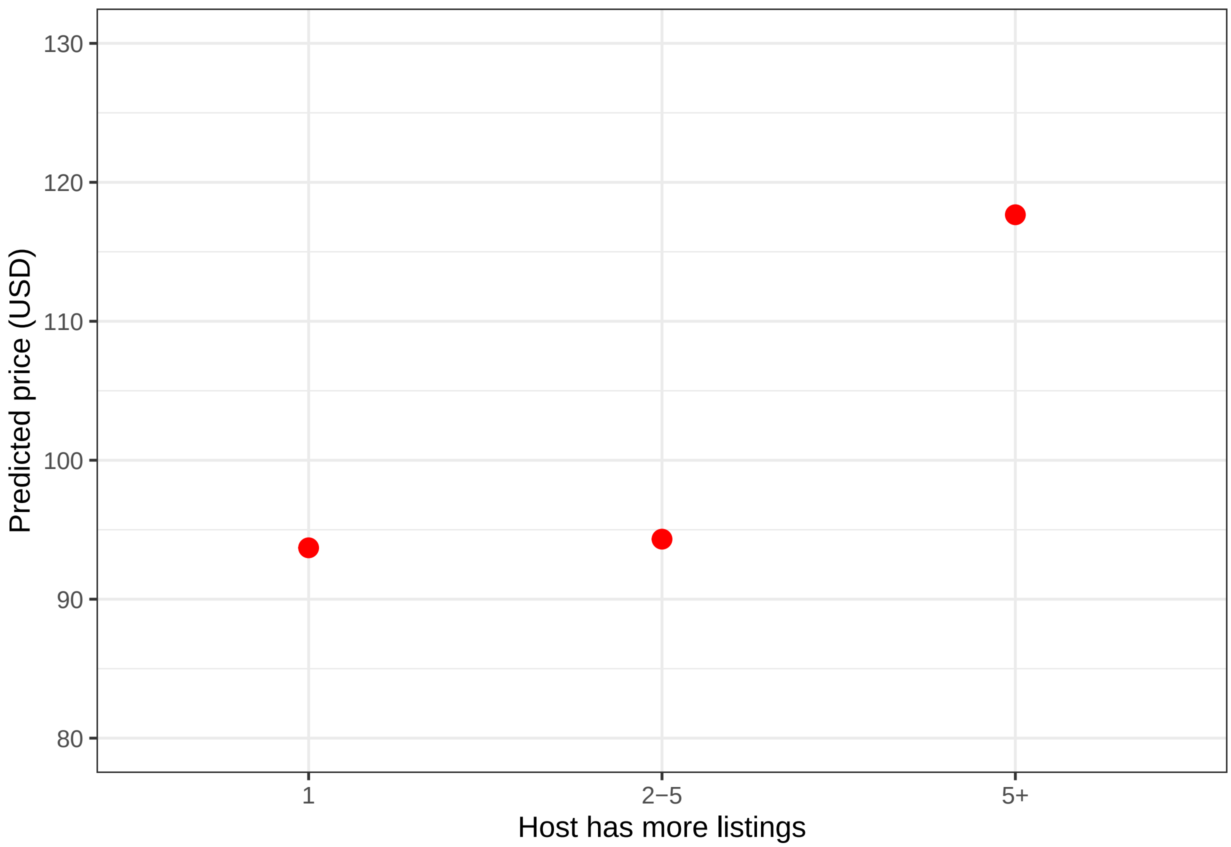
Partial dependence plots and tables






Partial dependence plot: Review scores
Predicted price (USD)
97.5
95.0
92.5
0
1
2
3
4
5
Review ratings


Predicted price (USD)
100
90
80
70
Entire
Private
Shared
Room type



Partial dependence table with predicted mean prices for the binary variables:

	More than one bathrooms	Unavailable for the upcoming months	Amenity: Dishwasher
No	92	108	93
Yes	103	91	107