Airbnb Rome Price Prediction

Attila Serfozo

2021.02.02

Abstract

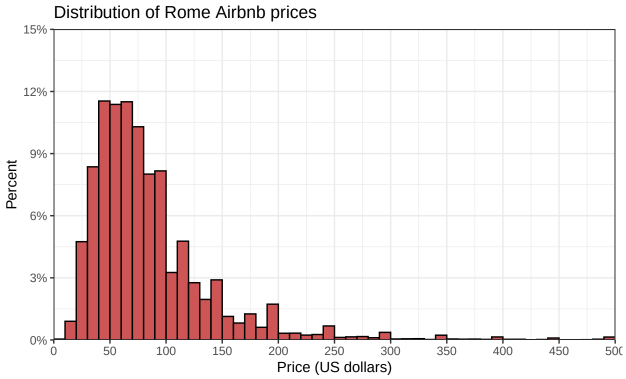
In this project my goal is to build a price prediction model for apartments able to host 2-6 persons in Rome based on local Airbnb data. During the exercise I cleaned the downloaded data, grouped variables and created six different models to predict prices. Based on the results the extended Random Forest model provided the best prediction with 44.9 USD RMSE and 40% R-squared values. The key predictor variables were number of accommodates, number of bathrooms, days since first review, review scores rating, room type, number of beds, air conditioning and free street paring availability, number of reviews and cancellation policy. These 10 predictors altogether account for approximately 48% of the model prediction power. From the subsample performance it is important to highlight that the model predicted prices of the III, IV and VI neighbourhood of Rome pretty badly, which probably caused some part of the RMSE effect of the model. Overall the model predicted the prices relatively good, but we can state that it tends to overprice the Airbnbs of Rome.

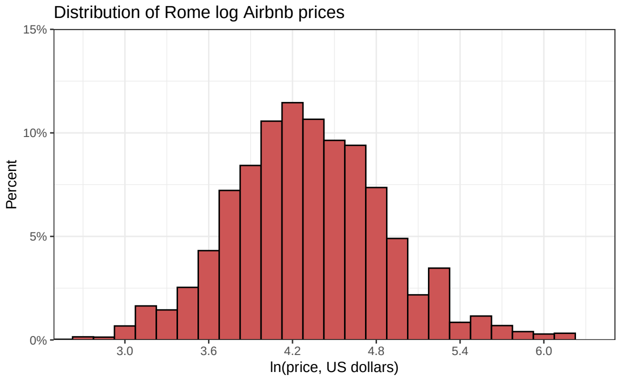
Introduction

The aim of this project is to build a price prediction model for apartments in Rome based on available Airbnb data downloaded from Inside Airbnb (http://insideairbnb.com/get-the-data.html). The dataset represents the available Airbnbs to rent on 2019.10.15. In the exercise to provide an accurate prediction I will select our key predictors from all variables of the dataset including both the basic property attributes, reviews and amenities of properties. During the project I will estimate 6 different type of models including two OLS, a LASSO, a CART and two Random Forest models and in the end I would like to finish with a final model with relatively good prediction power regarding it's RMSE and R-squared.

Data Preparation

In the begining of the data preparation I started with 30 814 observations and 106 variables describing all available Airbnb accommodations in Rome on 2019.10.15. The scope of the project aims to predict prices of apartments and condominiums able to host 2 to 6 people. After filtering the data to these criteria 21 366 observations remained in the analysis. The main variable of interest is price in the analysis as the goal of the project is price prediction. Thus observations with missing prices needed to be excluded, in addition I excluded observations with prices above 500 euros marked as extreme values. In the end I finished with 21 299 observations. The distribution of prices can be found below on the histograms.


Distribution of Rome Airbnb prices


Distribution of Rome log Airbnb prices

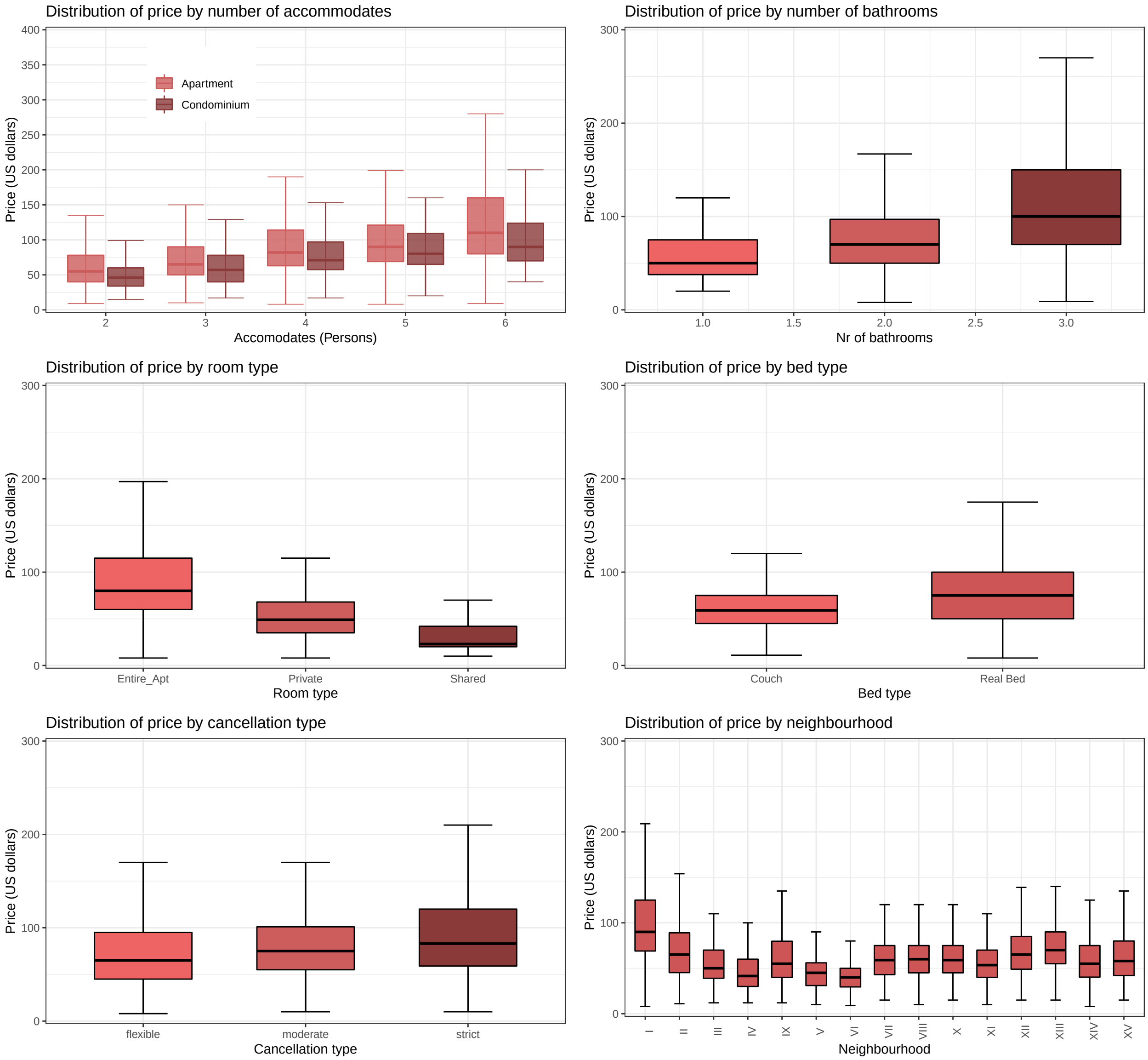
The larger task of the project was feature engineering. First, I needed to decide which variables to include in the analysis. Thus I separated my variables to 3 groups:

- **Basic variables** which consisted of the main variables of the prediction, including numerical variables such as number of accommodates (2-6), number of beds, days since first review and qualitative variables formatted to factors like property type (apartment or condominium), room type (Entire apartment, private or shared), bathrooms (1, 2 or 3), cancellation policy (flexible, moderate or strict), bed type (couch or real bed) and neighbourhood with the districts of Rome I to XV.
- **Review variables** which consisted of the main variables of guest reviews, including number of reviews, review scores between 0-100, and review score flags, which indicates missing reviews.
- **Variables of amenities** was the hardest group to create as on Airbnb the property attributions are not consistent, because they can be added manually by typing them in. Due to the fact that they are added by hand there are several amenities pointing at the same attribute but using different words, thus I finished with 198 different amenities in my dataset. To merge similar attributes to one I used a for loop to match similar variables, the loop can be inspected in the preparation code R file. As a results I kept or merged important attributes to one like air conditioning, washer, oven, pool, balcony, tub, free parking, elevator etc. In addition I dropped variables that based on my opinion would have insignificant effect on model results due to lacking enough number of true observations (less than 3% of the observations had it) or based on individual decision. In the end I finished with 50 amenities.

After creating the 3 groups the last task was handling missing values. There were 15 variables with missing values. From these 15, missing observations in bathrooms were replaced with 1 assuming there is at least 1 bathroom in an apartment, missing number of beds were replaced with the number of accommodates, missing minimum nights assumed 1 night stays and missing number of reviews were assumed as 1. Missing review scores rating were replaced with zeros indicating no reviews at the accommodation, an additional column was created with flags indicating that there is a missing value in that row. In the end USD cleaning fee and host response rate variables were dropped as there were too many missing values in them.

Exploratory Data Analysis

After the data cleaning as a next step lI had a look at the main variables of the models. The boxplots of variables can be found below. Overall we can see that the larger apartment the more it costs. On the first two boxplots we can see that apartments able to host more persons or have more bathrooms are more expensive which connects to their size. Also based on data apartments are valued higher than condominiums. The next two charts below, shows that entire apartments costs an extra compared to private rooms or shared rooms and real beds worth more for guests then pull-out couches. The last two boxplots shows surprisingly that strict cancellations are more expensive than flexible ones, for me it was interesting as I prefer flexibility. Also we can see that in Rome the district I, II and XIII districts are the most expensive in average compared to the IV, VI and VII which are the cheapest across Rome.



Modelling

Before starting the modeling I separated my observations to a training set (random 70% of our data) and a holdout set (random 30% of our data). During the modeling I used 5-fold cross validation as control and I based my modeling decisions on the average of the 5 RMSEs calculated from the five folds.

During the modeling exercise I created 6 models:

- **OLS** using basic variables
- **OLS** using basic and review variables
- **LASSO** using the extended model with basic variables, review, amenities and interactions
- **CART** using the extended model with basic and review variables and amenities
- **Random Forest** using basic and review variables
- **Random Forest** using the extended model with basic and review variables and amenities

The results of the models can be found below. Based on the model results the extended Random Forest model performed the best. Using 5-fold cross-validation it has an 44.9 USD RMSE meaning approximately

50% of the average apartment price (average price is 88.3 USD) and an R-squared of 40%. Based on these information I would choose this model as my final choice, but we should highlight that as the second OLS model, with basic and review variables, performed also relatively good, if the key decision point is interpretation then it can be also a good choice.

Table 1: Models R-squared

	CV Rsquared
OLS	0.306
OLS (basic + reviews)	0.341
LASSO (model with interactions)	0.375
CART	0.252
Random forest basic + reviews)	0.350
Random forest (with amenities)	0.400

Table 2: Models RMSE

	CV RMSE
OLS	47.9
OLS (basic + reviews)	46.7
LASSO (model with interactions)	45.5
CART	49.8
Random forest basic + reviews)	46.5
Random forest (with amenities)	44.9

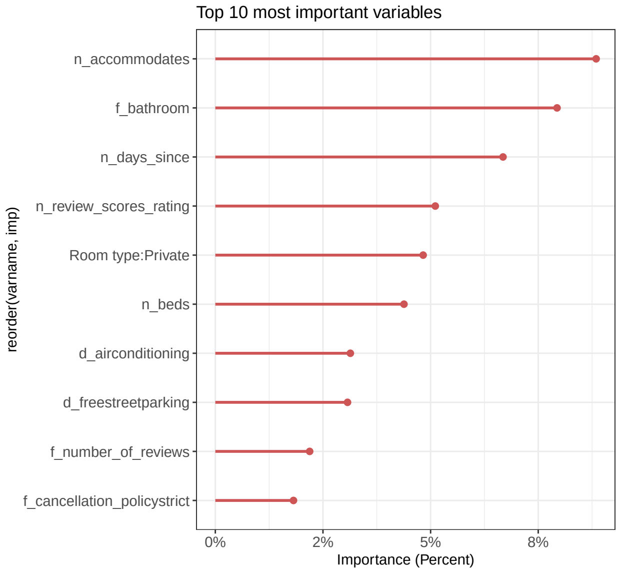
After running our models on the training sample let's have a look at their performance on the holdout set. Based on the results below we can see that the second random forest 46.1 USD RMSE performs a similar RMSE compared to the training set .

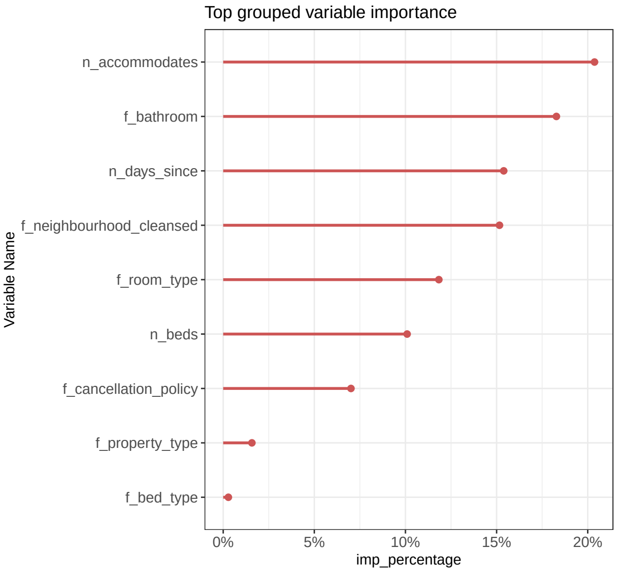
Table 3: Holdout models RMSE

	Holdout RMSE
OLS	50.1
OLS (basic + reviews)	48.5
LASSO (model with interactions)	47.1
CART	51.8
Random forest basic + reviews)	48.1
Random forest (with amenities)	46.1

Variable importance

I decided to select the extended random forest model. Below on the variable importance charts the key predictors of the selected random forest model can be seen. Interesting fact is that the top 10 predictor variable accounts approximately 48% of the model prediction power.


Top 10 most important variables


Top grouped variable importance

Partial Dependence

Based on the summary table below we can strenghten most of our inspections from the exploratory data analysis that larger apartments (4-6 persons) are way more expensive than small apartments (2-3 persons) and overall condominiums are cheaper. But there are no significant difference between the prediction error of these categories. The neighbourhoods supports our previous statements as well showing that the I, II and XIII are the most expensive. It is interesting that regarding the RMSE price we can see that in the III, IV and VI neighbourhoods our model performed much worse than in the others occuring more than 72-86% of uncertainty in price prediction. Interestingly there does not seem to be any pattern in the neihbourhoods RMSE regarding mean prices, so we can not state that the model tends to predict better lower prices or not.

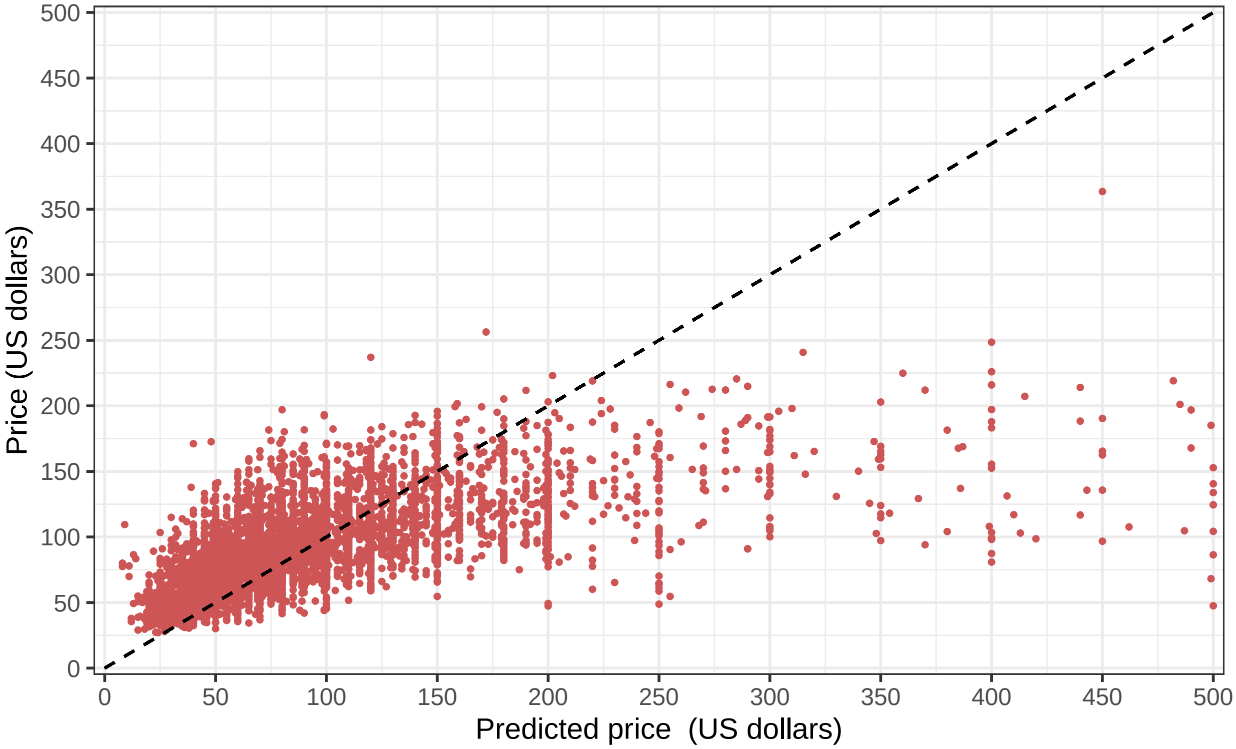
Table 4: Subsample performance

Var.1	RMSE	Mean.price	RMSE.price
Apartment size	NA	NA	NA
large apt	51.8	106.5	0.486
small apt	38.0	66.6	0.571
Type	NA	NA	NA
Apartment	47.0	90.2	0.521
Condominium	35.7	71.8	0.497
Borough	NA	NA	NA
I Centro Storico	54.2	108.9	0.498
II Parioli/Nomentano	41.9	76.0	0.551
III Monte Sacro	54.1	63.0	0.858
IV Tiburtina	39.4	54.7	0.720
IX Eur	44.3	69.2	0.640
V Prenestino/Centocelle	17.7	47.6	0.372
VI Roma delle Torri	36.2	50.0	0.724
VII San Giovanni/Cinecitta	31.1	64.3	0.484
VIII Appia Antica	30.8	62.7	0.490
X Ostia/Acilia	40.9	71.1	0.575
XI Arvalia/Portuense	32.0	60.2	0.530
XII Monte Verde	31.0	71.8	0.432
XIII Aurelia	37.6	80.9	0.465

Var.1	RMSE	Mean.price	RMSE.price
XIV Monte Mario	34.3	60.6	0.565
XV Cassia/Flaminia	37.9	67.7	0.560
All	46.1	88.6	0.520

Actual versus predicted price

Finally let's have a look at how the model predicted the actual prices. We can see that prices spread wide around the line, so we can not say that the prediction covers variety in prices. Also based on the chart below we can suspect that the model overpriced the cost of several apartments which probably caused the relatively large RMSE.



Summary

In this report I tried to find a good model to predict Airbnb prices in Rome. I created six models from which the extended Random Forest performed the best results with 44.9 USD RMSE and 40% R-squared values. Probably one of the key reason of the 40% performance is that the model overpredicts the prices of several Airbnbs. The created model highlights interesting attributes of Airbnbs which could be important for owners to consider as main price drivers, attributes like these includes number of accomodates, number of bathrooms, review scores raing, private room, airconditioning, free parking on street, cancellation policy or number of beds. Finally if someone would like to book a hotel based on the prediction of this model I would suggest to use it only if the person looks for an accomodation in the V Prenestino/Centocelle or XII Monte Verde neighbourhood of Rome as the model provides better prediction there compared to other parts of the city.