Predictive Pricing Model for Airbnb Milan

Khawaja Hassan

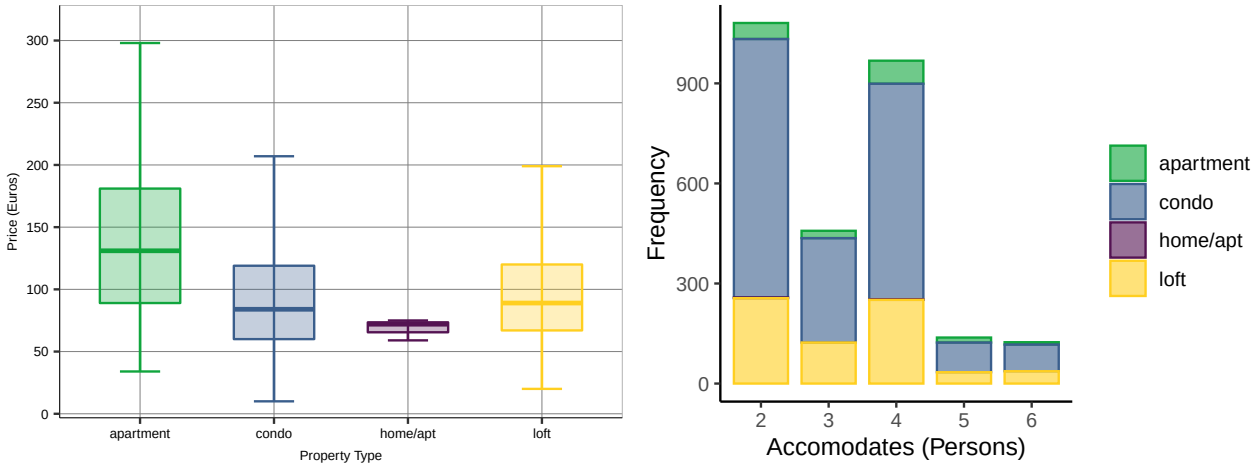
2/9/2022

Executive Summary

The aim of the study was to assist company to predict prices for their small and mid-sized apartment accommodating 2-6 people using different prediction model. To build these price prediction models we will be using data from Inside Airbnb. To find the best combination algorithms to assess the prediction model we will be building and comparing Airbnb predictive models for the city of **Milan**, Italy. We used 3 machine learning algorithms like OLS Linear Regression, Random Forest, Cart. The best model amongst these came out to be Random Forest (Auto-tuned) followed by OLS.

Data Engineering

Our choice provided a cross-sections of Milan listing with more than 16000 observations with last scrapping date of 9th January 2022. However, the data required immense cleaning in terms of extracting amenities into separate columns from their vector form and later clubbing them based on of some meaningful categories. We were able to narrow down amenities to 74 clubbed variables. After this we prepared our data based on the case we have been provided with, for which we filtered the number of accommodates between 2-6 and took property type which was full apartment (Entire Loft, Entire Serviced Apartment, Entire Apartment, Entire Condominium)



Moving forward, we had to check what variable can be imputed based on the number of missing values in each of them. Here domain knowledge about the case was essential to come up with some logical assumptions. For number of beds where the value was missing, we took the assumption that it should corresponds 1.5 number of accommodates so for every 3 accommodates there will be at least 2 beds. Whereas, for number of bathrooms we imputed the missing based on the median the value. In this case we did not create any

flag variable since the missing observation were below 5 % of the total observation. However, we had certain variable where missing observation was more than 30 % (example: Reviews per month, Review Score Rating) and for which we created flag variable first and then imputed their values with median.

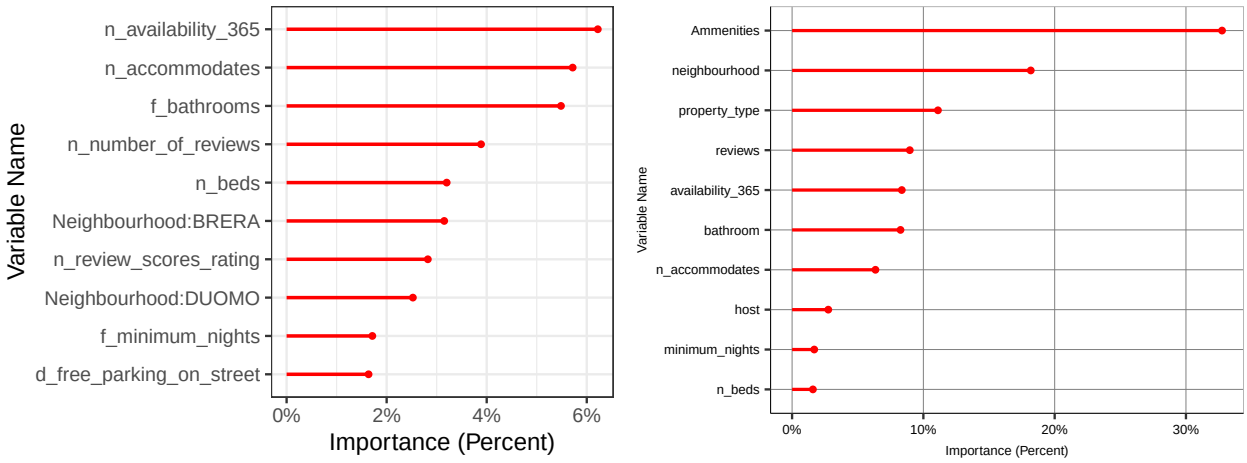
Explanatory variables The explanatory variables which were used in our models are as follows:

- *Factor variables*: For each Neighbourhood, type of property, including flag and factorized variable of size variables.
- *Reviews variables*: Review score rating and the number of reviews the apartment gets each month.
- *Host variables*: Created Dummies for host verification and they being a super host or not.
- *Dummies*: Binary variables consisting of all the amenities that are being offered by host.
- *Size variables*: This includes numeric variable like number of beds, baths, accommodates, and minimum nights.

Prediction

Our four-prediction methodology included Ordinary Least Squared (OLS), Two Random Forest with and without tuning parameter and Classification and Regression Tree (CART).The following table shows the cross-validated Root Mean Squared Error (RMSE) on the training sample.

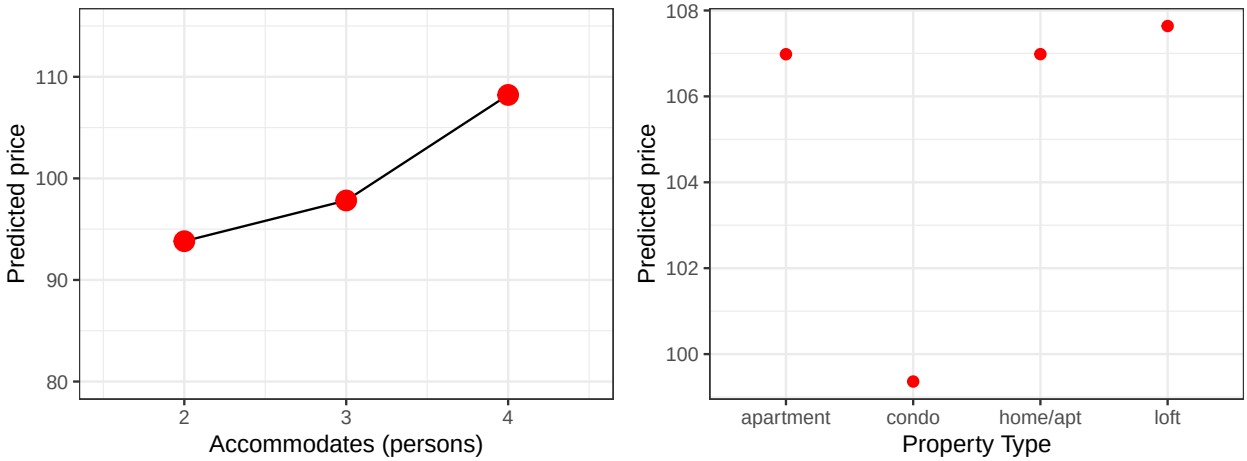
Based on the result we will further running diagnostic test on the RF auto model using **Variable importance**, **Partial dependencies plot** and **Sub-sample** to check for performance.The following top 10 important variable plot shows variables which had the largest MSE reduction. In second chart these variables are just grouped based on their factors along with other numeric variables. Based on the plots it suggests the company needs to focus more on the loft category since the predicted prices are higher, while taking the assumption of ceteris paribus. Along with that, the model also suggests higher predicted price for neighbourhood of Brera with highest yhat of $137 and property type accommodating 3 or more guest.



To further check our RF performance, we run sub sample on three x variables as shown in the figure below. The prediction error is similar across apartment size and number of beds but in type of property the relative RMSE had major deviation under the type of apartment (suggesting it is difficult to predict the price). However, even the RMSE of type apartment is low and their mean prices are way higher than others it might suggest that other factors impacting its price. It can be the case that there are few observations in our apartment category and those apartments might be located near the city centre leading to higher prices.

Table 1: Horse Race of Models CV RSME

	CV RMSE
OLS	42.4
CART	46.4
Random forest 1: Tuning provided	42.9
Random forest 2: Auto Tuning	41.9



Conclusion

To conclude, our final predictive pricing model is Random Forest with autotuned parameter as it was resulting in the lowest RMSE value. In accordance to our PDPs, it suggested to invest in loft category with accommodation greater than 3 and in the neighbourhoods of Brera and Duomo to earn higher yield. However, based on the marginal difference in the final RMSE of OLS it is also advised that using OLS model would also be beneficial on the live data. Lastly, to implement this prediction on live data we need to assess the external validity of the model with the socio-economic levels of Milan.

Predicted vs actual prices