Apartment Pricing Prediction Model for Puglia Italy

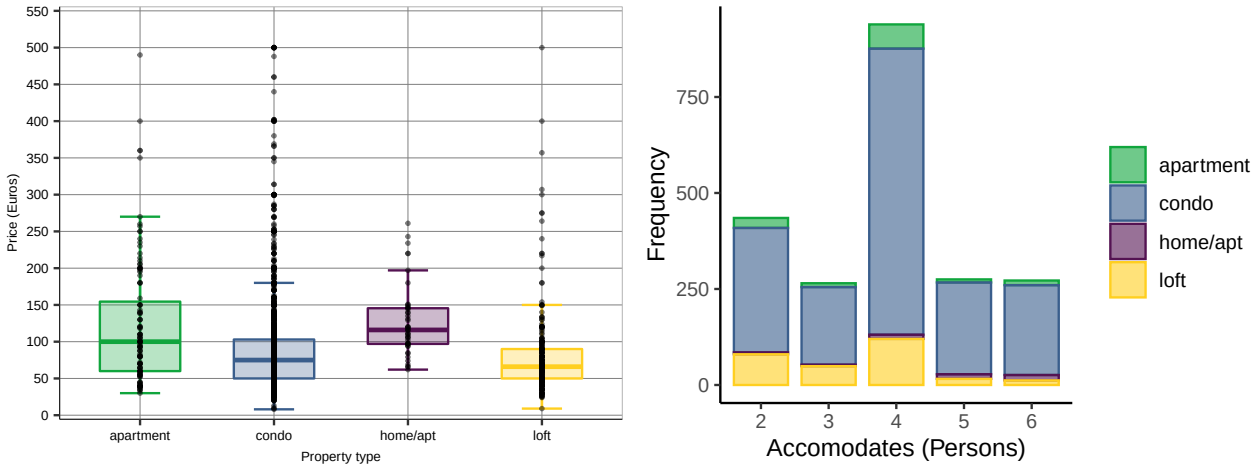
Shah Ali

Introduction:

The purpose of this project is to help Star Realtors set the on optimum rental price for their small and mid-size apartments hosting 2-6 guests, in the city of **Puglia** in Italy. To achieve this, several price prediction models were built based on data provided by Inside Airbnb which was scraped between December 30, 2021 and December 31, 2021. Several salient features of the accommodation were considered and four different kinds of models were analyzed to find the best prediction model, namely OLS, CART, Random Forest and GBM. Out of these the chosen the prediction results of Random Forest model to be the best one.

Data Engineering

The data for the chosen destination was a large dataset with more than 33000 observations and 74 variables. This data was fairly crude and required intense cleaning before we could begin our analysis. The data was filtered based on the specification provided in the brief; the property types were filtered to include, Apartments, Serviced Apartments, Loft and Condominium only each accommodating 2 to 6 guests. The main part of the cleaning process rested on cleaning the amenities. In our dataset all the amenities were written together in one large vector with total of 1442 unique amenities. Since it is essential to analyze the effect of each amenity on the target variable, meaningful dummy variables were created out of the names of amenities. Each amenity was separated from the list, grouped together if they were similar (Wifi & Internet, TV and cable) and only those which had at least 1% of measurable observations was chosen. In this way we are able to narrow down the amenities to 85.



While dealing with missing values, using domain knowledge, variables with no relevance on our price model and those with more than 50% of missing values were dropped from the dataset analysis for example picture, url, name, etc. For the main variables, imputation was employed along with creating a flag variable taking value of 1 for missing and 0 for NA. Imputation involved using logical assumption. For bathrooms, median

Table 1: Horse Race of Models CV RSME

	CV RMSE
OLS	60.6
CART	64.0
Random forest 1: Tuning provided	57.1
Random forest 2: Auto Tuning	55.8
GBM	58.7

value is imputed in place of missing values. For bedrooms, imputation included the value of reminder of the division by 2. For number of beds, we imputed the result of division by 1.5 (atleast 2 beds for 3 people). For target variable (price), all the observations where price was missing were dropped along with properties with price beyond 500 Euros. These constituted less than 1% of observations and we considered them as extreme values. Although the distribution price is right-skewed, we decided to proceed with without log transformation so that we don't incorporate while transforming back to normal prices.

Explanatory variable

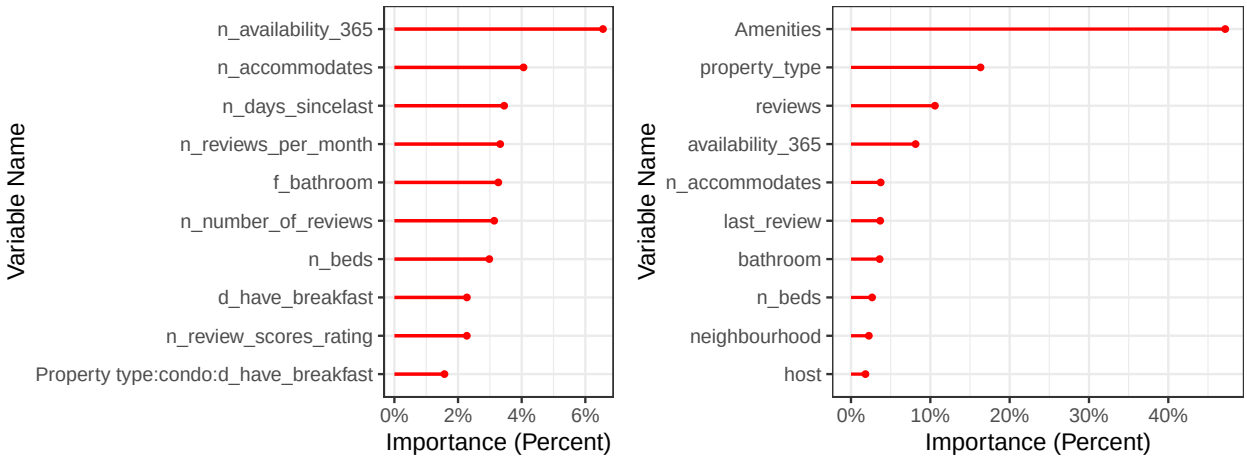
Numeric variables: These variables define size such number of beds, accommodates, bathrooms and bedrooms, the minimum number of nights required to rent the place and the availability of the apartment in 365 days of the year. *Factor variables:* These are categorical variables that are either in string or numeric form. For example, the neighborhood or the type of property. *Dummy variables:* These are binary variables describing the amenities offered in the property *Review variables:* These describe the reviews and ratings characteristics for an individual property. Number of reviews, the rating for property and the mean monthly reviews received. *Host variables:* These variables describe the characteristics of the host of the property. They are binary variables like host is a superhost or if the host is verified

Prediction Models

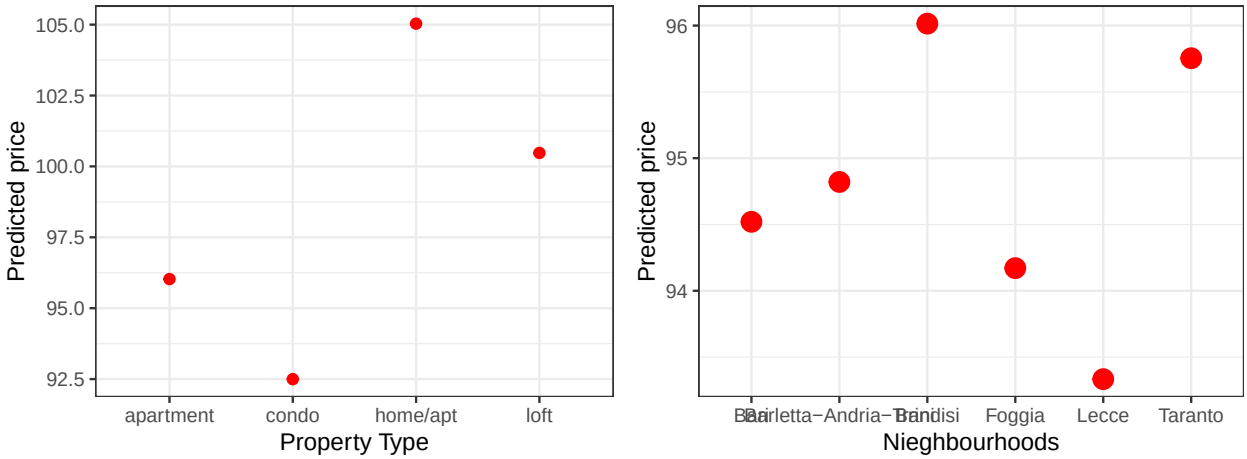
The methodology adopted for prediction analysis is to use LASSO as a variable selection method and use the non-zero predictors given out by the LASSO for all of the subsequent prediction models (OLS, Random Forest, CART and GBM). After running LASSO we are left with 102 predictors which we shall be using for the price prediction models. This is a data driven decision necessitated by a smaller number of observations in the data and limited business knowledge of Airbnb ecosystem. The four prediction models that we ran for this analysis include a simple Ordinary Least Squared (OLS) model starting with simple up to complex model, Classification and Regression Tree (CART) with pruning, two Random Forest (RF) models where one is provided with the tuning parameters and the other was run on automatic tuning and lastly a Gradient Boosting Machine (GBM) model. The results of the cross-validated Root Mean Squared Error (RMSE) on the training sample are shown below.

While the difference between the two RF models is negligible, and since the same models would be run again on a larger dataset, we advise to use the RF model with provided tuning parameters due to time efficiency and because it is more robust.

Moving one step further, we ran diagnostics on the results of Random Forest with auto tuning using Variable Importance (VI) plots, Partial Dependency Plots (PDP), and checking subsample performances. The figures below show the 10 most important variables. The second chart shows variables grouped based on their factor groups, it turns out that amenities account for more than 50% of importance.

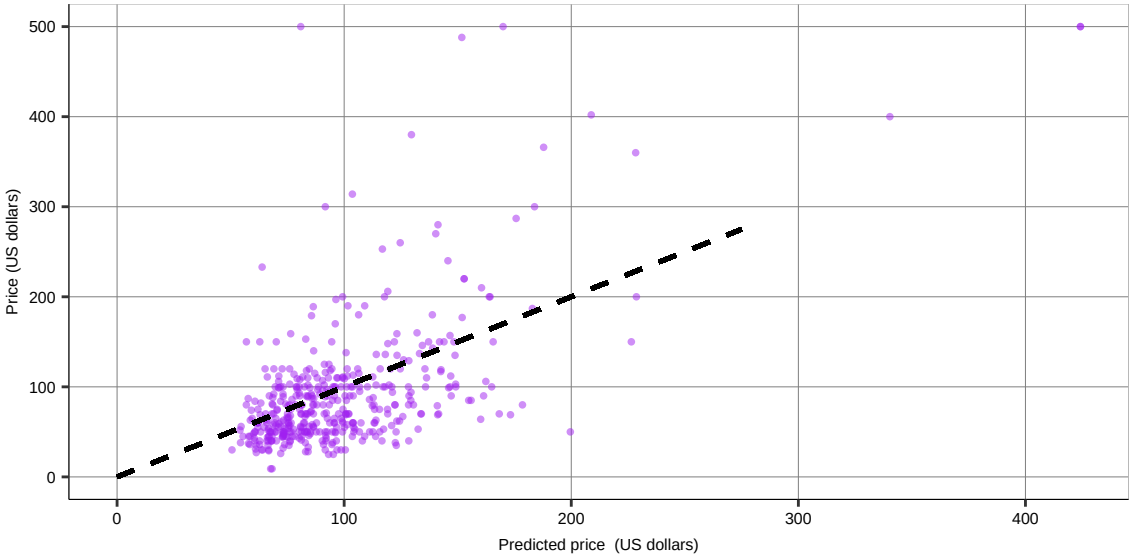


We deep dived into some of these important variables to create PD plots to see how changing the specific predictors value while keeping other predictors constant predicted the prices. These included, property type, number of guests accommodated and neighborhoods. The PD plot for property type suggests that our model predicts relatively higher prices for entire home/apartments than other property types, followed by loft and then condominium. Similarly, the plot for number of guests shows a positive linear relationship between number of guests and prices. Drilling down to the neighborhood suggests focusing more on properties in Brindisi neighborhood. We also looked at subsample performances for these variables to see how individually these would impact on the price prediction.



The outcome of sub sample supports the outcome of PD plots for property type. The prices of home/apartment are easier to predict as it has lowest prediction error and it also predicts the accommodation with 6 beds with lowest error. However, when we look at the sub sample of neighborhood we see that our model predicts the prices of Barletta-Andria-Trani neighborhood with lowest prediction error. This is a contradiction with PD plot which be accredited to the limitation in data and small number of observations for Barletta-Andria-Trani. In the case of size of apartment, the size, the model predicts the size of small apartments with at-most 3 guest capacity with lower prediction error.

Conclusion



To conclude our price prediction model with Random Forest that included auto tuned parameters gave the lowest RMSE values. After running subsequent diagnostic tools, this model predicts the prices of the home/apartment, accommodating at-most 3 guests, in Brindisi neighborhood with lowest prediction error. The management can apply this model for predicting prices of apartments with these specifications on a another dataset provided there is an high external validity of that data with this data.