Balint Thaler - Data Analysis 3
Github repo: https://github.com/balint-thaler/DA3-Assignment-2

Summary report

Introduction

The goal of this report is to summarise the findings of the prediction models about the AirBnB dataset. For more technical details and specifics, please refer to the technical summary.

The data

The data is downloaded from the website of Inside AirBnB which periodically scrapes the AirBnB listings from specific cities. I have chosen Rome, Italy as the basis of the analysis. Rome is one of the most visited cities in Europe, which means that there is likely a high number of various AirBnBs available in the city. The clean dataset contains 14,422 observations and 46 features including “price”, the target variable.

Exploratory data analysis (EDA)

To understand the data a bit more, some of the variables are explored deeper. We observe that price follows a non-symmetric right-tailed distribution with a mean of 125.5 EUR. Given that the prediction is aiming to get absolute prices as an outcome, price will not be transformed despite the non-symmetric distribution. In terms of neighbourhood, most of the listings (more than 8000) are in the Centro Storico, while the second most frequent neighbourhood (Parioli/Nomentano) contains almost a magnitude fewer observations. The property type also shows similar patterns with “Entire rental unit” dominating the data with more than 9000 observations. Number of accommodates and beds show a somewhat more balanced distribution with 4 and 2 being their modes respectively, however it is noteworthy that there are less than 150 observations with at least 6 beds.

Modelling

For modelling a total of 3+1 models will be built: first a very simplistic OLS model is to be built. The intuition would dictate that the price of the apartments is mostly influenced by the qualities of the accommodation, secondly by the reviews and only thirdly by the hosts attributes. Therefore, the first OLS model will only contain the accommodation-related variables. This model is kept very simple intentionally, as it only serves as a point of comparison for the later models.

Secondly, a Random Forest model will be built and thirdly, one with Gradient Boosting. Lastly, using the outputs of the machine learning models the most important variables will be identified and will be plugged into another OLS model. This way the last model will likely utilise the benefits of the ML algorithms with the relatively easy-to-understand nature of the OLS models.

Evaluating the models

The total dataset is split into train and holdout data in an 80-20% ratio. All model building is conducted on the train dataset, while evaluation is done on the holdout data. To evaluate the models, their cross-validated root mean squared errors (RMSE) will be compared. The first very simple OLS model performs with an RMSE of 55.38 which is around 44% of the mean of the target variable. Despite its very simple nature, this is already a fairly good performance, but the expectation is that the machine learning algorithms will perform even better.

After doing a grid search on some of the tuning parameters of the random forest model, the best one performs with an RMSE 53.66 which is lower that the OLS, but considering the increased complexity of the model, the gain in RMSE is not substantial.

Lastly, a similar grid search was conducted with the Gradient Boosting (GBM) model, where the best performing model has an RMSE of 48.46 which is considerably better than the previous two models.

Understanding the ML models

Given that both the random forest and the GBM model work more or less like a black box, it is rather difficult to understand which variables are the most important in defining the price and what are the patterns of association. To understand these, we can look at the variable importances and the partial dependence plots.

Figure 1 shows the most important variables of the random forest model, while Figure 2 shows the same for GBM. We can observe that the most important variables overlap in the two models, and they mostly include accommodation and review-related variables – confirming the earlier intuitive thought. Interestingly, the top 7 variables in the random forest model contribute to around 50% of total importances, while to reach the same cumulative importance, the GBM needs 11 variables.

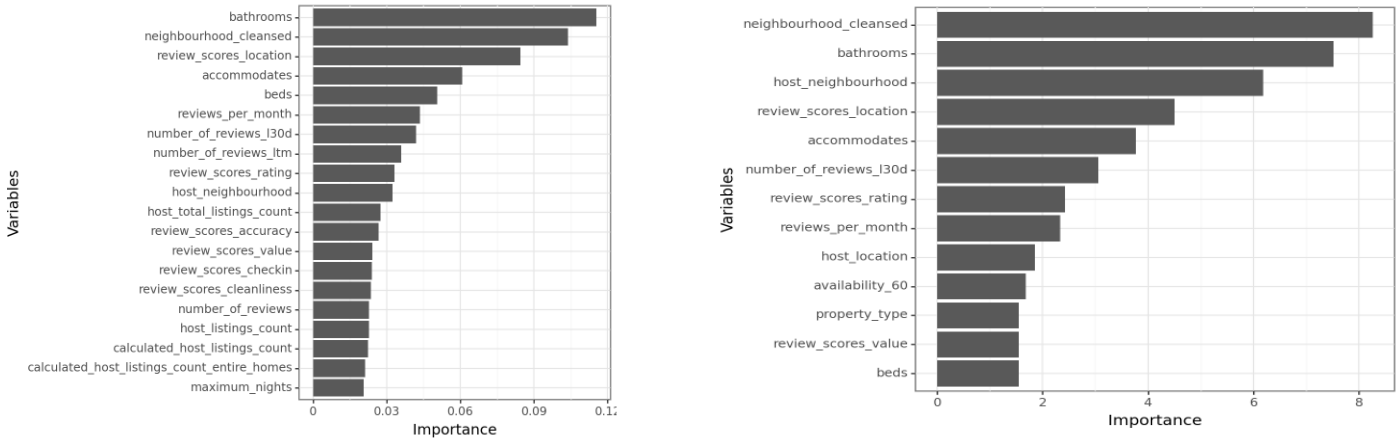

Figure 1 and Figure 2 variable importances of the random forest (left) and the GBM model (right)

Given that the most important variables are mostly consistent between the models, the partial dependence plots are only shown for the random forest. With these we find a S-shaped association between number of bathrooms and price, an exponential positive association between review scores – location and price, an almost linear positive association of number of accommodates and price, a non-linear negative association between number of monthly reviews, number of reviews in the last 12 months and price, and finally a strong positive association followed by a flattening of number of beds and price. The detailed charts can be found in the technical report.

Refined OLS

Using the outputs of the ML models it is now possible to focus on the most important variables in our data and build an OLS model with them. The non-linear functional forms and some possible interactions are also considered and thus the best performing OLS model has an RMSE of 53.16 which is lower than that of the random forest (53.66) and the original OLS (55.38) but higher than the GBM's (48.46). This means that we were more successful than the original OLS and outperformed the random forest as well, GBM was still superior to all other models which makes it the preferred choice from the presented models.

Evaluating the GBM model on the holdout set

To understand how the GBM model performs outside of the training set, its RMSE in the holdout set is calculated on the total set, and certain subsets of it. Its overall RMSE is 46.4 (somewhat below than on the train set), which is 36.9% of the mean of the prices in the holdout data. Interestingly, the model performs the weakest among cheap hotels (relative RMSE is 55%), well among expensive ones (relative RMSE is 34%) and best among mid-priced ones (relative RMSE is 29%). It also has a varied performance among different locations with Centro Storico (the most frequent location) being close to the average (relative RMSE is 35%), but in some locations this goes above 50%. The model performs rather consistently among different accommodate, bed, bathroom, and location rating numbers.

Understanding the top features using Shapley values

The GBM models Shapley values were also calculated to understand the top features behaviour more. We see that higher number of bathrooms consistently have a large positive impact on prices while smaller values tend to have a much smaller but negative impact. Number of accommodates shows a balanced impact on prices however the higher feature values have a higher positive impact on the model than the lower ones have negative. Among number of beds there are also some average values that have a positive impact on the prices.

The review scores variables (on location and rating) show a more interesting picture as even some of the higher scores can impact the model output in a negative way (see some red dots left of the 0 line). Lastly on number of reviews per month we again see that the lower values have a much stronger positive effect than the higher values have negative.

Overall, we see that the reviews have a more complicated impact on the prices compared to the other accommodation-related variables. A deeper analysis of these associations is needed to uncover the details of these patterns.

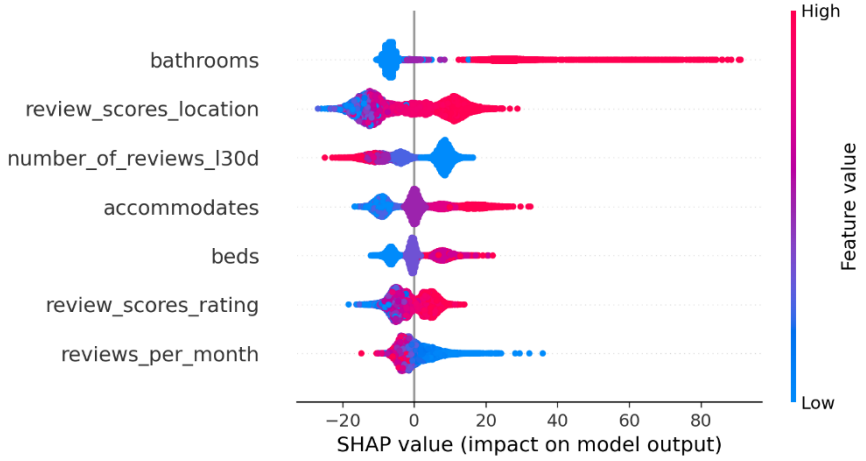

Figure 3: Shapley values of the most important features of the GBM model

External validity

The models' external validity is likely to be limited especially in terms of other locations. The patterns discovered in Rome may not hold in other places. Given that the data is from a specific date (15 December 2023), there may also be different patterns in different times of the year even within Rome. Although the most important features are likely to be more or less stable, the period before the holiday season is likely to be somewhat unique despite Rome not being a typical Christmas destination. Furthermore, the dataset was limited to full apartments that accommodate 2-6 people and cost less than 450 EUR. The results may not hold true for accommodations outside of these ranges.