Price prediction for newly introduced Airbnb listings in Vienna

Quick overview

The data for the prediction exercise is from the Inside the Airbnb site, from 2022. The data is restricted to entire homes and apartments with accommodates between 2 and 6 persons.

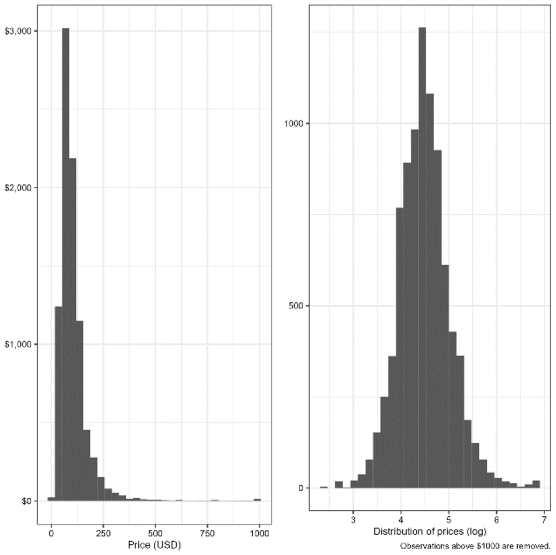

Figure 1

The target variable is nightly prices, whose distribution can be seen in Figure 1. Nominal prices are heavily skewed to the right.

Among the most important features, there appears the neighbourhood (Figure 2), number of accommodates, some host features (experience, being a superhost, etc.) and some amenities (what the listing includes, whether parking, pets, smoking is allowed, whether there is a view, etc.). From the neighbourhoods, the 1st district, Innere Stadt emerges as it has a much higher average price than the other districts as it lies within the heart of the city and closest to all the famous sights.

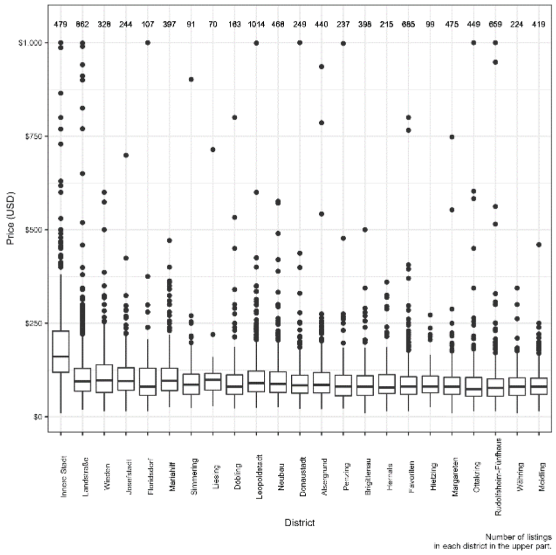

Figure 2

Important to note, that features regarding reviews are not used and not reported as the listings shall be new, so they cannot be rated so far. In the medium run though, they can be useful to maintain effective pricing.

Calibrated models

Several models were computed and tested, namely:

1. Multiple (5) OLS regressions with additional number of features
2. LASSO model based on the best-performing OLS model
3. Random Forest (RF) on all available features.

The performance of the several OLS models can be seen in Figure 3, where the fit is compared to the complexity of the models. It is obvious, that Model 4 within the OLS regressions performs the best, which includes basic features, host features and amenities. Model 5 (the most complex one) is also interesting, but the fit is worsened by some interaction terms.

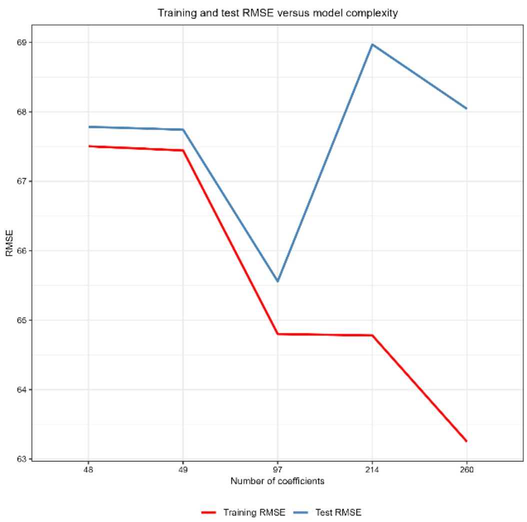

Figure 3

Model 4 among the OLS is the one which produces the lowest average error, so this model is used to estimate our next method, the LASSO. LASSO shrinks some of the many coefficients to zero that do not improve the fit much but increase the variance of the model. The performance of the LASSO and the other models is presented in Table 1.

Model	Test RMSE
OLS (Model 4)	65.559
LASSO	65.458
Random Forest	62.036

Table 1

The final model also presented in Table is Random Forest. As opposed to the models previously seen, it does not require a priori model or feature selection, it only needs some basic tuning. During the analysis, several tuning parameter specifications were tried, and the best RF model was chosen. It clearly outperforms both models. The shortcoming of this model can be that it is something like a “Black-Box”, more difficult to interpret and to follow the analytical steps performed. The variables that matter the most in reducing the variance can be seen in Figure 4.

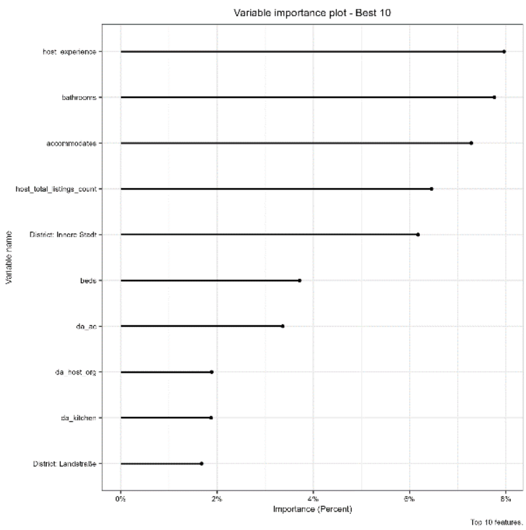

Figure 4

Unsurprisingly, number of accommodates, number of bathrooms, location in the 1st district are among the top features of the model, but host experience and host being an organization rather than a single person matters as much, as well as some amenities, such as air conditioning.

Not only it fits the training data better, but it also produces better predictions for the listings' prices.

Predictive performance

In Table 4, the order of the models is the same: Random Forest produces the lowest average error when predicting the prices of the listings based on their features. It can capture some patterns to better predict the prices of some more distinct observations, whereas the other models (because of the limitations of the OLS setting) cannot replicate this.

Model	Test RMSE
OLS (Model 4)	68.300
LASSO	68.273
Random Forest	63.236

Table 2

Conclusion

In this analysis, the task was to predict the prices of Airbnb listings in Vienna, 2022. The aim was to produce a benchmark on how to evaluate some future listings' prices yet to be on the market. The most important features are clearly basic identities (number of accommodates, bathrooms), host features (experience, organizational entity), location (inner districts) and amenities (air conditioning, kitchen). The best performing model in every aspect of the analysis was the Random Forest, so this model is highly advocated to use for the price prediction of the newcoming listings in Vienna.

Balint Parragi