Price prediction model for a specific Airbnb in Vienna (Austria)

Introduction

A company who operates small and mid-sized apartments (usually hosting 2-6 guests) asked us to help them setting their prices for new apartments in Vienna, Austria. As there are currently many hosts in the city, the data from Inside Airbnb[1] may contain false information, thus a thorough cleaning may be necessary to get a usable data frame to work with. Next, multiple models are used to determine which one may predict the actual worth of the company's apartments the best, based on a prediction fit factor (such as RMSE).

Data & Cleaning

The dataset itself could be cleaner, many modifications had to be done. The most important one, is that many listings **have missing values for the price** variable. These rows have been deleted from the dataset, as filling them with any number, the models will use false information for its dependent variable and filling them any other values than an integer/float, will break the code. The dataset further filtered to only include **2 to 6 accommodates**, based on the company's request and **Hotel rooms were excluded**. Many types of properties remained, thus more filtering was needed and now it includes **just apartment-like properties** (Rental Units / Condos). Flags are introduced for two review columns and the host columns are used mainly to get dummy variables for each listing. The reason behind the latter one, is that how the host "advertises" him/herself through Airbnb, how does the person care about his/her appeal in it, may influence the price. Last one, the bathroom variable has been changed from string to float. Except baths and flags, no new column creations felt necessary, the original names of them are understandable.

Usable Models

Using only one model has its danger, that the one we used may not be the best one among the possible models known to us. For this reason, three models are used to predict the price as best as it is possible. These models are the following:

a.) Linear OLS Regression
b.) Random Forest
c.) LASSO

The reasons of the usage of the said models are the following:

A.: Linear OLS Regression

The OLS Regression let's us to choose the variables from the dataset ourselves. While we have to make some important decisions in this model, by choosing which variables will be

[1] http://insideairbnb.com/get-the-data/#:~:text=Vienna%2C%20Vienna%2C%20Austria

used, the code chunk *LinearRegression()* takes care the rest of the calculations and will give us back a simple RMSE value.

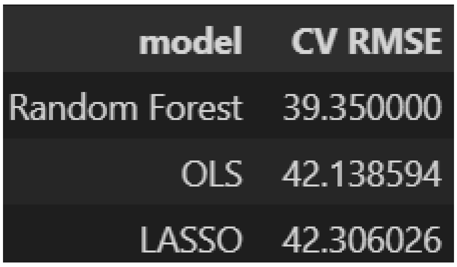
B.: Random Forest

Random Forest needs relatively less tuning than for the other prediction models, while more coding is needed in python, due its “Black Box Model” attribute. This model, unlike the OLS regression model, while giving all the variables that have chosen previously, will only use a randomly selected amount of it. It will combine the results of multiple trees and should give back a really good RMSE value. Due its Black Box attribute, feature importances will be also coded to get a better understanding of the model.

C.: LASSO

Because the dataset contains many columns/features (75 columns), the Least Absolute Shrinkage and Selection Operator model will select the variables in a way, to avoid overfitting, it tries to find the best measure.

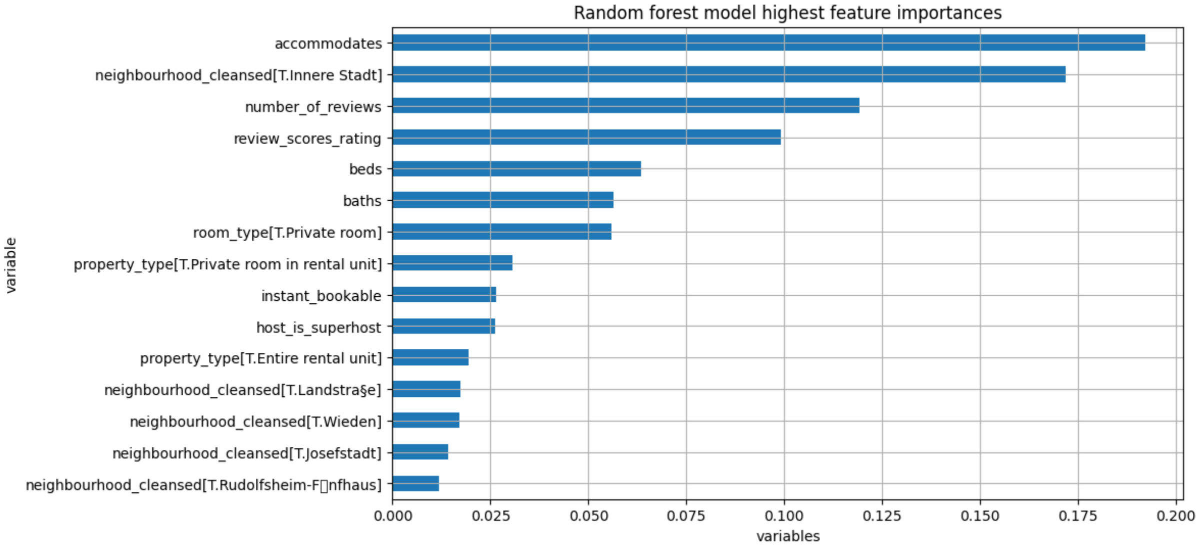
Models’ Results



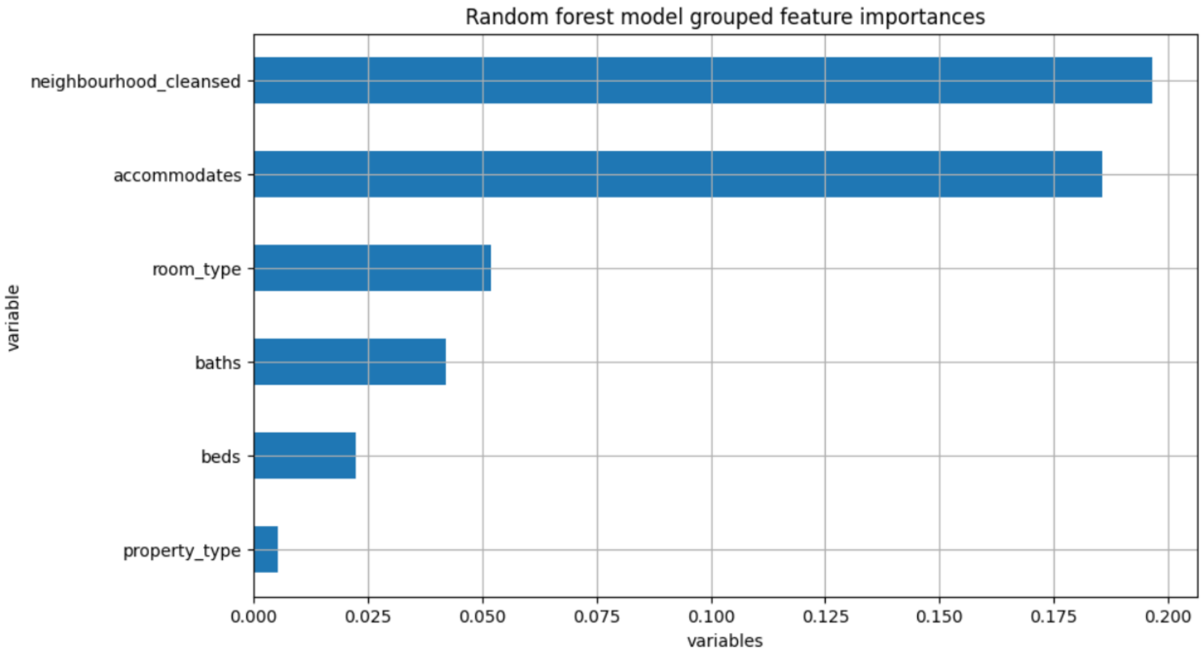
model	CV RMSE
Random Forest	39.350000
OLS	42.138594
LASSO	42.306026

As it can be seen from the CV RMSE comparing table, the Random Forest model came out with the lowest value, in other words, this model has the best fit amongst the used ones. While we should use the Random Forest, based on its RMSE, we don’t know yet which features influences the price the most, thus feature importances has been coded to check them.

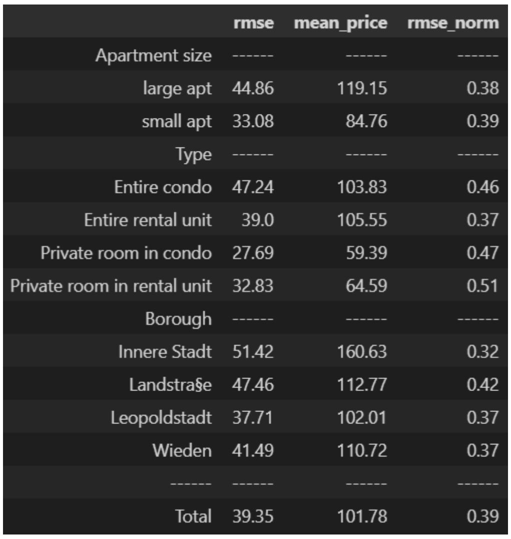
Features’ Importance



The graph above shows us, that (naturally) the number of people, the apartment can hold influences the price the most. As the place, it seems that Airbnb flats in Innere Stadt also gives a boost in price, while other places, like Landstraße or Wieden (while being a bit more significant than other variables) are not as influential. One of the most important values we can see, are the reviews. While they are important features, the company will just start to rent out their flats. **It is advised that the company should have somewhat lower prices, than the model's average says, for its initial deployment, until they get enough reviews and scores to be able to re-price.**



Looking at the grouped basic variables, the place of the apartment counts the most, even more than the number of accommodates, while the type of room does not mean that much, neither the type of property (Entire/Private room – Condo/Rental unit). From this, we can argue with the following table:



	rmse	mean_price	rmse_norm
Apartment size	------	------	------
large apt	44.86	119.15	0.38
small apt	33.08	84.76	0.39
Type	------	------	------
Entire condo	47.24	103.83	0.46
Entire rental unit	39.0	105.55	0.37
Private room in condo	27.69	59.39	0.47
Private room in rental unit	32.83	64.59	0.51
Borough	------	------	------
Innere Stadt	51.42	160.63	0.32
Landstra§e	47.46	112.77	0.42
Leopoldstadt	37.71	102.01	0.37
Wieden	41.49	110.72	0.37
------	------	------	------
Total	39.35	101.78	0.39

This comparison table shows us the mean prices of the apartments with their precise column values, for example: Entire condos tends to have a price of ~104, while Innere Stadt flats have around ~161. Entire flats fetch for a higher price, which is self-explanatory, as they have much more space, than private rooms, but the main argument is for the Boroughs. If the company owns a **large flat** (3 or more accommodates), which is an **Entire rental unit** in **Innere Stadt**, than they can list it with a high price, just be advised of the lack of reviews, that has been already written before!