**Data Analysis 3. Second Assigment**

## Data exploration and cleaning process

I investigated the Airbnb dataset of Vienna to build a price prediction model for the project. The original dataset contains 11955 real estates. The task was to examine only apartments which accommodates 2-6 people after filtering accordingly 8801 observations remained. The dataset can be downloaded here. In the data exploration part, I decided to focus on the following property types, since these are the only ones which have a significant number of observations: Entire condo, entire loft, entire home, entire rental unit, entire serviced apartments.

| property_type | N | Percent |
|---|---|---|
| Casa particular | 1 | 0.01 |
| Entire bungalow | 5 | 0.06 |
| Entire cabin | 1 | 0.01 |
| Entire chalet | 1 | 0.01 |
| Entire condo | 1007 | 11.48 |
| Entire cottage | 4 | 0.05 |
| Entire guest suite | 12 | 0.14 |
| Entire guesthouse | 3 | 0.03 |
| Entire home | 77 | 0.88 |
| Entire loft | 155 | 1.77 |
| Entire place | 10 | 0.11 |
| Entire rental unit | 7043 | 80.29 |
| Entire serviced apartment | 377 | 4.30 |
| Entire townhouse | 11 | 0.13 |
| Entire vacation home | 38 | 0.43 |
| Entire villa | 6 | 0.07 |
| Room in aparthotel | 11 | 0.13 |
| Tiny home | 10 | 0.11 |

*Table 1*

I eliminated extreme values that would have significantly distorted my results. In order to see what the extreme values are, I used the data summary function. Based on the result filtered out the unrealistically low (less than $27) and high values (above $380.42) of price column.

| | P0 | P01 | P25 | P50 | P75 | P99 | P100 | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|
| price | 10.00 | 27.00 | 63.00 | 87.00 | 120.00 | 380.42 | 999.00 | 103.49 | 75.38 |

*Table 2*

The first figure shows us the price distribution of the apartments of Vienna. We can clearly see that majority of the apartments are under 200 dollars. The second figure helps us understand the trend a little bit more in detail, we can see for each property type how many observations we have, also what is the average price for each.

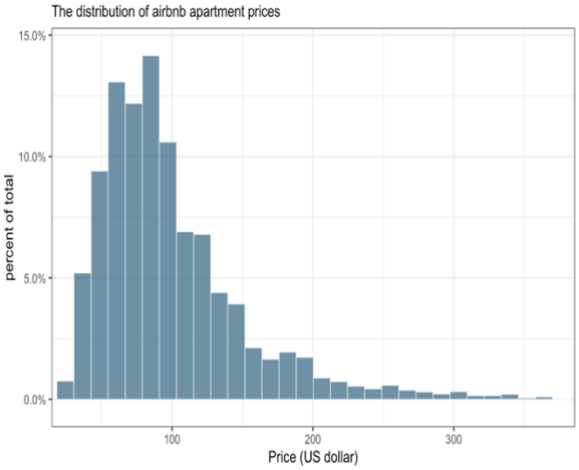


*Figure 1*


*Figure 2*

**Data creation and model definition**

The amenities column contained additional information about the properties, decided to further investigate it. Created different binary columns for each aminities that can be found in the dataset and included the most important ones for each Airbnb apartment. This allowed me to include relevant information in the models, such as whether an apartment has TV or not, whether it has air conditioning or not.

I split the dataset into two parts, one of them is the training (70% of the data) on which I built the models, the other is the holdout (30% of the data) on which I tested the performance of them on unseen data. Four models were built to predict the price for an apartment in Vienna. Two regression models (OLS & Lasso) and two machine learning models (random forest & gradient boosting machine).

**Model evaluation**

The model evaluation was performed based on *Table 3*. Overall, the machine learning models performed better than the regression models. Based on the cross validated RMSE the random forest model (38.46) performed slightly better than the gradient boosting method (38.53). However, the opposite happened based on the holdout RMSE. I would choose the random forest model, since cross-validated RMSE provides a more robust estimate of the model's performance by averaging the RMSE across multiple iterations of splitting the data into training and validation sets.

| models<br><chr> | CV RMSE<br><dbl> | Holdout RMSE<br><dbl> |
|---|---|---|
| GBM | 38.53131 | 38.78429 |
| LASSO (model w/ interactions) | 40.83963 | 41.38097 |
| OLS | 40.09806 | 41.37754 |
| Random forest model | 38.46225 | 39.23473 |

4 rows

*Table 3*

The following figures represents the partial dependence plot for the two machine learning models, based on the number of accommodates and property types. *Figure 3* is connected to the random forest model meanwhile *Figure 4* show the gradient boosting machine model. For the accommodates the models on average predict somewhat similar trend meanwhile for the property types on average they predict relatively different prices based on the holdout dataset.

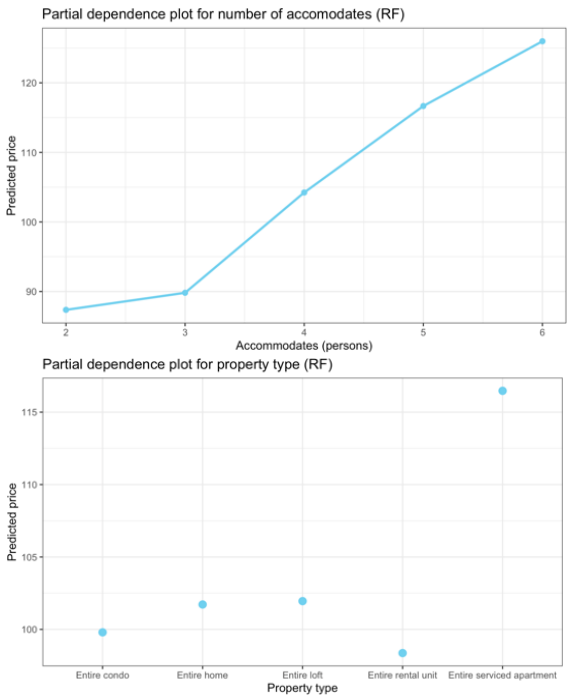


*Figure 3*

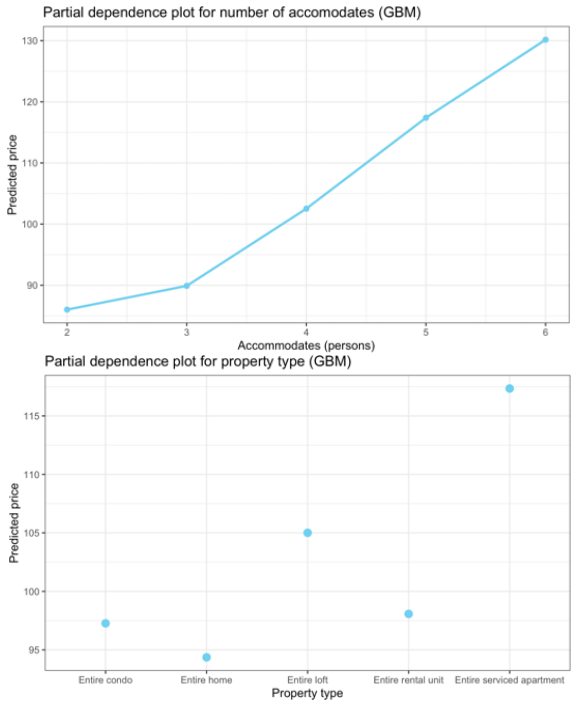


*Figure 4*

*Figure 5* shows the actual prices against the predicted prices using the two best performing models, GBM and RF. The points in the plot represent the individual observations, where each point shows the actual and predicted price for a specific observation. The dashed line represents the perfect prediction, the closer the points are to the dashed line, the better the prediction of the models. This plot helps us to compare the performance of the two models in terms of their ability to predict the prices accurately.

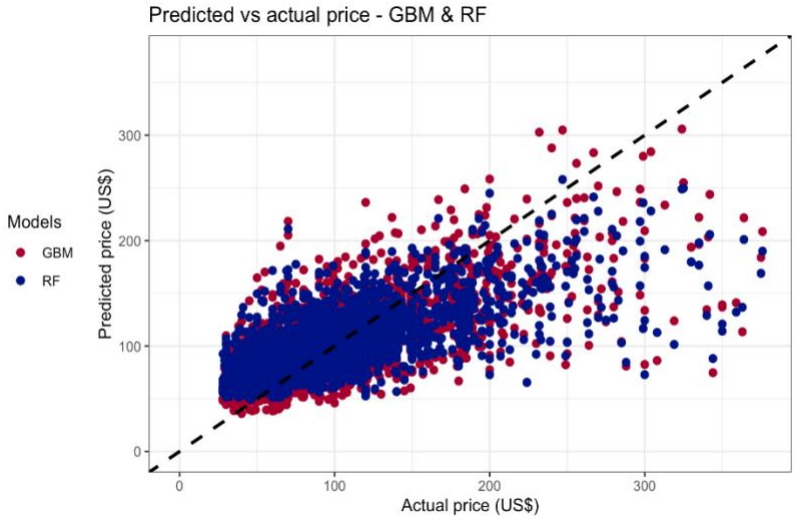


*Figure 5*

**Random Forest model**

Since the best performing model is the Random Forest model, I decided to investigate what are the most important variables in terms of their explanatory power. *Figure 6* shows the top 10 variables.

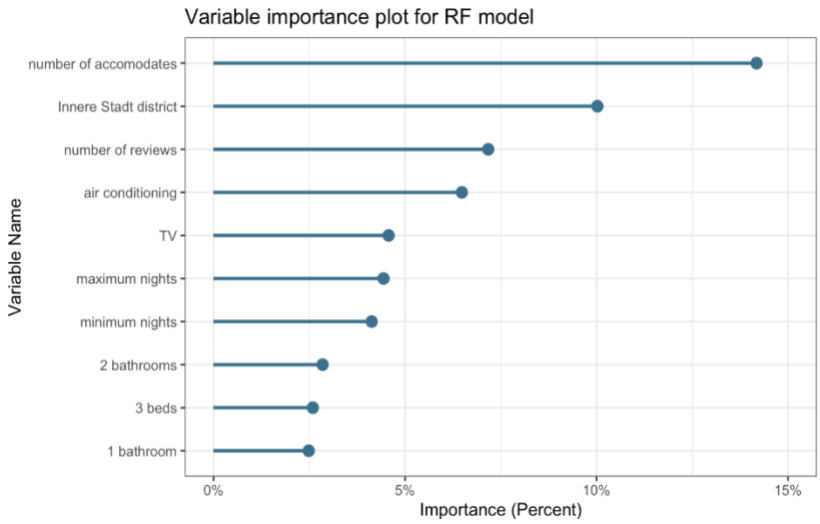


*Figure 6*

**Side note**

I realized that many people write the size of the apartment in the description column, I wanted to check if I could find it since it's an important variable. Unfortunately, this was not the case, more than half of the apartments didn't include this information or I couldn't gather the information based on a pattern.

If you are interested in the whole coding process please click here.