Data Analytics: Assignment 1

Gertraud Malsiner-Walli

WS 2024

Linear Regression: Airbnb data

Goal

In the context of the Vienna Airbnb market in Vienna, the purpose of this study is to to analyse factors that predict pricing.

Data description

The dataset provides a comprehensive look at Airbnb prices in Vienna. Each listing contains various attributes such as room type, cleanliness and satisfaction rating, bedrooms, distance from the city centre, price is on weekdays or weekends. We would like to identify the determinants of Airbnb prices in Vienna, represented by the variable `realSum`. The data set contains the following variables:

Description of variables

- realSum: the full price of accommodation for two people and two nights in EUR
- room_type: the type of the accommodation
- room_shared: dummy variable for shared rooms
- room_private: dummy variable for private rooms
- person_capacity: the maximum number of guests
- host_is_superhost: dummy variable for superhost status
- multi: dummy variable if the listing belongs to hosts with 2-4 offers
- biz: dummy variable if the listing belongs to hosts with more than 4 offers
- cleanliness_rating: cleanliness rating
- guest_satisfaction_overall: overall rating of the listing
- bedrooms: number of bedrooms (0 for studios)
- dist: distance from city centre in km
- metro_dist: distance from nearest metro station in km
- attr_index: attraction index of the listing location

- attr_index_norm: normalised attraction index (0-100)
- rest_index: restaurant index of the listing location
- rest_index_norm: normalised restaurant index (0-100)
- lng: longitude of the listing location lat: latitude of the listing location
- weekend: dummy variable if the listing was on a weekend

Tasks

1. Read the data into R and check whether the date are represented correctly.

   Perform data pre-processing. Especially, investigate:

   - Are there missing data? Which types of variables are in the data set?
   - Eliminate variables which are not meaningful predictors.
   - Turn categorical variables into factors (e.g. dummy variables, 'room_type").
   - Remove doubled information (e.g. the attractivity index is given in two variables, measured on different scales, thus one of these variables can be excluded)

2. On the final data set, perform a descriptive statistics of the variables (summary, structure of the data, correlation among numeric variables) and visualize the data.

   Especially, investigate graphically the relationship between the outcome variable `realSum` and possible predictors. Which variables seem to be related to `realSum`?

3. Model 1: Fit a linear regression model in R using **all** variables. Which variables have an significant impact on the price?

4. Model 2: Perform stepwise variable selection using the function `step()`. You can use either forward or backward search. For the finally obtained model interpret the results:

   - Which variables have a positive/negative impact on the price? Comment on the sign of the coefficients: is the direction of the impact what you have expected intuitively?
   - Interpret in detail the coefficients of the predictors 'roomtype', 'dist' and 'weekend'.

5. Perform an in-sample comparison of model 1, model 2 and model 3.

6. Perform an out-of-sample exercise for model 1 and 3:

- Split the sample into 80% train vs 20% test sample.
- Fit the full model to the training data.
- Perform the stepwise procedure on the training data to obtain a smaller model with less predictors. Which variables are omitted?
- Use both the full and the stepwise regression model trained on the trainingsdata to make predictions on the test data. Calculate the training MSE and test MSE.
  - According to the theory, which of the two models should have a larger training MSE, which a larger test MSE?
  - Let's turn to the data: Which model performs actually best on the train MSE? Which model performs best on the test MSE? Which MSE (training or test) should be chosen for selecting the final model?

Submission:

- The assignment should be completed in an R Notebook.
- Submit the .nb.html file together with the .Rmd.
- Only one submission per team