https://github.com/pilipentseva/predictions

Goal of the project: Predict prices for small and medium apartments in San Diego

Data set: for my project I used a data set from Airbnb Inside. Initial dataset has 9673 (almost 10 K) observations and is collected for the data 25th of September 2021.

Data cleaning and transformations: Firstly, I got rid of the columns which didn't have any meaning for my analysis. Secondly, I deleted the columns which repeated information listed in another column or stating the obvious information. For, example it was the case for the neighborhood columns. I could see that the neighborhood column has a lot of missing data (almost 1/3 of the observations). Looking more carefully into the data I could see that neighborhood information is much better depicted in neighborhood cleansed (and it has no missing values). Thus, I dropped variable neighbourhood. Host neighborhood sometimes also had missing values and looked very similar to neighbourhood cleansed, but with more missing values for reasons I couldn't identify. I decided to only keep the neighborhood cleansed. The column with the number of bathrooms as a number variable was missing, but I extracted information about the number of bathrooms from the column which had it in text format. Prices and percentage columns contained symbols which I didn't need for my analysis - I got rid of them. I encoded variables containing true and false (t, f) to dummy variables 1, 0.

After this I looked at data types and transformed those who belonged to a mistaken data type. I also evaluated the missing values. Depending on the logic and structure of the data, I took decisions on how to replace missing values. For some observations I replace them with 0, for others with mean or mode. I checked duplicate rows and found none. I created several important variables. Firstly, I calculated the distance from the city center to the location of the airbnb. I could do it using coordinates of the city center of San Diego and given coordinates of airbnb locations. Amenities and host verifications had many variables encoded in one cell. For each amenity and host verification method I created a dummy variable. I purposely didn't group amenities in groups, because I don't have a field expertise and tried to avoid making a mistake combining amenities in wrong groups. There are also amenities with strange symbols in the beginning. Even after cleaning amenities I left them as separate ones since I consider these symbols to be specifications. I worked with available time variables and created a variable for host which referred to the duration a person has been a host calculated on the 25th of September 2021 date. Moreover, I created a variable determining the overall number of amenities for each of the accommodations, since it also may be an important variable.

Last but not least, I encoded the categorical variables.

Subsampling: Since I am only interested in small and medium apartments, I only considered places that could accommodate less than 8 people. I also subsampled it to the type of room: Entire home/ apartment, which should reflect my targeted types of apartment.

Model building: I built 3 models. One with basic predictors (distance, beds, number of amenities and etc). Second one with basic predictors and variables for reviews. Third with all features included in the data set. The main difference about the 2nd and 3rd one is the inclusion of amenities. I got the following RMSE for all 3 models.

Model evaluation: RMSE Model A - 114.280
Model B - 109.301

Model C - 124.721

The lowest RMSE for Model B which corresponds to model 1. It makes sense because it has the lowest number of predictors and they are most likely the most important.