Amsterdam Airbnb Price Prediction

Business Report

2022-02-10

INTRODUCTION

The purpose of this project was for Airbnb companies to predict prices for small to mid size apartments accommodating between 2-6 persons. The task included using a dataset on a city for a particular day. In my case it is Amsterdam for 7th Septermber 2021. The task was accomplished using 5 different machine learning regression methods including, OLS, CART, Random Forest, LASSO & GBM. Each of these gave different models that best predicted the price for a night for an Airbnb.

THE DATASET

The dataset can be accessed through the Inside Airbnb (www.insideairbnb.com) website. The raw data contained 16116 observations. However, this required a bit of cleaning to be able to process the data for my analysis.

The target variable is the price per night in USD. I used different characteristics and features of a rental which became the predictors in my analysis. These included the number of people accommodated, the number of beds, the number of washrooms, the reviews given and the amenities provided by the rental.

DATA WRANGLING

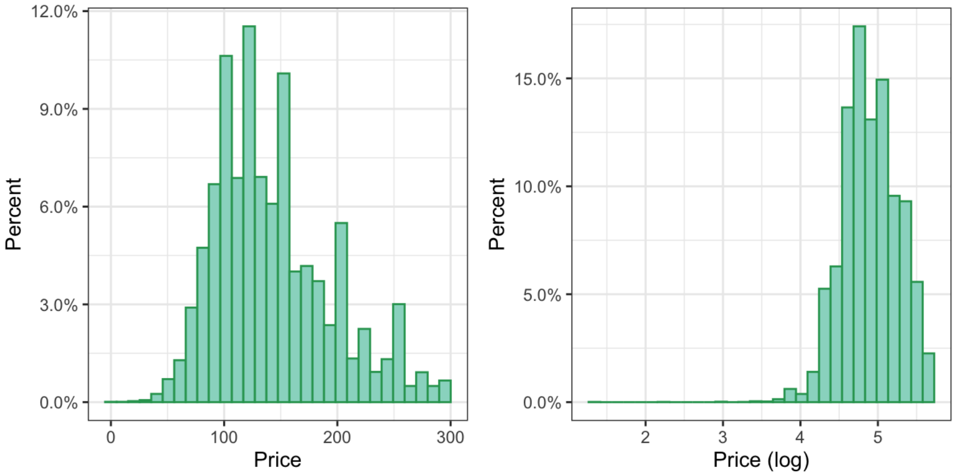
The initial requirement was to tweak the date, exclude and filter observations and variables. There were 74 variables initially. All of these were not used for the analysis so I filtered accordingly. Most importantly, the assignment asked to use apartments from all the given categories so I used apartments, lofts and rental units. I also dealt with missing values by either dropping the columns if they were not in used or imputing the values with median. The variables that were imputed were created as flagged variables. The property types were renamed and the amenities column was split into binary columns. After this I aggregated amenities by names using a function Using a function which searches for a particular key word (like “wifi” or “parking”). The categorical variables were converted to factors.

EXPLORATORY DATA ANALYSIS

After the data was cleaned, I conducted Exploratory Data Analysis to try to understand the characteristics of our data. This meant checking their descriptive statistics, their distributions and their relationship with price.

LABEL ENGINEERING

The target variable was price in USD. The missing values in the target variable were dropped. Looking at summary statistics below we can see the average price for an apartment was $162 while the maximum was $8000. This may be an extreme value but I decided to keep it as it was in our target variable and would be relevant for predictions. I used level prices as they gave a rather normal distribution compared to log price as shown the distrubution below. They also make interpretation easier.



FEATURE ENGINEERING

Next came deciding the functional form for the predictor variables and determine their relations with price. As amenities were binary columns they were used as it is. We can see that prices varied according to accommodation capacity. We can see an almost linear pattern using a non-parametric regression showing prices to increase as accommodation capacity increases. The box plot shows prices for apartments that are instantly available compared to those that are and there is not a significant difference in price for both. I have used quadratic accommodation for my models explained below.

The average price of serviced apartments tends to vary significantly as the capacity to accommodate people rises. This is different for entire rental units, as the accommodation capacity increases it results in a corresponding higher price. We looked at the distribution of beds, as beds is the same as accommodation capacity and decided to take logs to as the pattern of association with price was non linear. I pooled some of the observations in columns into groups. These included 'n_number_of reviews', 'n_number_of_bathrooms, 'n_minimum_nights'.For example if the accommodation had 0,1,2 or 5 bathrooms. Then I imputed the missing values depending on on the type of object it represents. For example, we filled '1' or '0' in missing values in 'n_number_of_nights' column. Using data driven modeling I tested many interactions between the dummy variables amenities and the factor variable price. I included those in my model that had a visible difference in price.

By the end I was left with 9477 observations and 31 variables including the dummy variables, factor variables, and some log transformations.

MACHINE LEARNING MODELING

I used three prediction models to predict Airbnb prices for Amsterdam. These models are in increasing complexity using different functional forms for variables and interactions. With these I estimated many models with different sets of variables or tuning parameters according to the model in use. Lastly, I calculated the RMSE on the test set and holdout set to determine which of the models is the performing best.

MODEL SELECTION

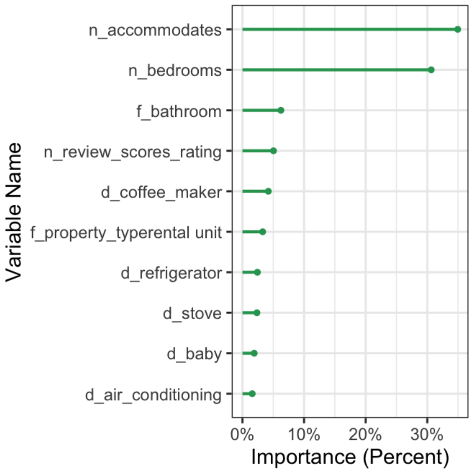
Each of the methods gave a different model as the best model for predicting prices. Using a five fold cross validation OLS regression I split the data intro training and a holdout set by 70% and 30% respectively. Model 1 was simply using the numeric variables, polynomials and factor variables. Model 2 added amenities to that and then model 3 added amenities interactions as well. With a test RMSE of 44.64 model 2 appears to predict prices better compared to the other 2 models. For LASSO, Its interesting to note that we received an RMSE of 44.61 with LASSO which is slightly lower than that of model 2 which we chose using OLS regression. Next, with random forest which is a similar technique to OLS, we constructed models using some tuning parameters. For this we used 3 models by adding more predictor variables with different functional forms each. Model 3 had the lowest average RMSE of 44.63 as compared to the other two model. So we can see that by adding more predictor variables for OLS didn't return a lower RMSE, however it did return a lower RMSE for random forest models.

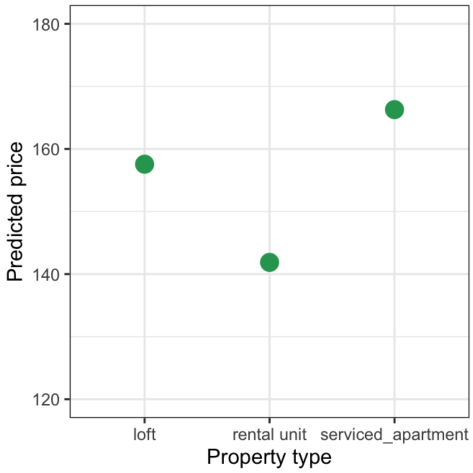
Comparing the CV-RMSE for all the models in the table below we can see that LASSO has the lowest CV-RMSE, followed by Random Forest and then OLS. These results are different compared to the case study predicting Airbnb prices for London.

Model performance comparison

	CV RMSE	Holdout RMSE
OLS Model 2	44.64	44.82
LASSO (model with interactions)	44.62	44.85
Random forest(with amenities)	44.63	45.22
GBM	44.90	45.66
CART	45.08	45.63

I also performed two diagnostics as part of Random Forest; variable importance and partial dependencies plot. Variable importance showed the 10 most important variables for predicting prices and the top of the list was the number of people an Airbnb can accommodate. The partial dependencies plot showed changes in price resulting from a change in predictor variable.For the property type, it concluded that prices for serviced apartments tend to be higher than lofts or apartments. This is due to the additional services offered to guest in serviced apartments, perhaps breakfast, laundry service or pick and drop facility.





CONCLUSION

I executed several models and chose the best one for each type of regression. The table below shows their cross-validated RMSE-s and the RMSE-s calculated using the holdout set. Compared with the results of the case study predicting Airbnb prices for London the ranking of models is exactly the same. Random Forest had the best performance followed by LASSO with a CV-RMSE OF 44.63. This means that in our live data, there is a possibility of making an error of 44.63$ on live data in the Airbnb of Amsterdam assuming that the external validity is high.

This gave me an understanding of the different approaches and choices one can make for predictions. This showed that not one single model is better but rather the decision depends on the situation at hand, requirement of the study and the understanding of the analyst.

Note: link to github repository: https://github.com/nawalhasan/DA3 (https://github.com/nawalhasan/DA3)