Analysis on hotel prices

Zsombor Hegedus

2021 february 6

Summary

This document is to build a random forest model on the `hotels_europe` dataset that is freely available on Gabors Data Analysis sit. I wish to build a predictive model for hotel prices, using this versatile dataset that contains more than 20k observations on hotels from all around the globe. After cleaning my data, I will experiment with different tuning approaches to arrive and see how they can improve the predictive power of the model. Once the best model is identified, I examine variable importance plots to showcase the most important predictors that the model used for price prediction

The analysis is submitted as data exercise from Chapter 16 exercise 5. Workfiles are available in my github repo.

Data

After joining the `hotels_europe_price` and `hotels_europe_features` csvs, I merged them and filtered for June, 2018. This seemed like a good choice for a date given that it is not very far into the past, yet data is collected in a non-stressed environment (free of the the COVID-19 pandemic e.g.). I further filtered out observations with not available accommodation type, and decided to keep only those, that I identified as either *hostels*, *hotels*, *apartments* or *guest_houses*. Furthermore I filtered out accommodations that had prices higher than $1,500 as those were extreme values that were suspiciously expensive. I filled NAs for the `rating` and the `rating_reviewcount` variables - I used the median and mean respectively - and left a flag in case imputation was used. The total number of observations was 14.000.

Overall I had the following predictors: whether the accomodation was advertised with a discount, if room was noted as scarce, the city of the accommodation, distance from city center, average rating, count of ratings, distance from alternative city center, accomodation type and the flags I created for imputed metrics. This is 10 variables in total, however since some of them are factor variables that were transformed to dummies when used by models - the final number of variables included was 57.

Modeling

I used machine learning for prediction by using a random forest model. I first stripped out a 30% random sample and trained models on the remaining 70%. I used 5-fold cross validation with RMSE being my loss function and grew 500 trees with every iteration. For tuning, I first looked at 4 scenarios regarding how many variables I allowed to stay in a terminal node - these are under column: *Minimum node size*. I also looked at 5 other scenarios in which limited the number of variables to choose from (further referred to as m parameter) for a split in any tree of the random forest - these are under the *Vars to select* column. The end result is the below table, where one can see RMSE in 20 different scenarios:

Table 1: RMSEs for Model comparison based on different tuning parameters

Vars to select	Minimum node size	RMSE
3	5	108.7132
3	7	108.7149
3	10	108.8051
3	15	108.8466
9	5	103.0562
9	7	102.9961
9	10	103.2075
9	15	103.2596
11	5	102.9766
11	7	102.8317
11	10	102.9175
11	15	102.8989
13	5	102.9869
13	7	102.8729
13	10	102.8945
13	15	102.8163
15	5	103.0407
15	7	102.9806
15	10	102.8404
15	15	102.8200

The best tuning scenario was the case when I allowed the tree to have 15 observation in the terminal nodes, while variables to be selected was limited to 13 variables only. When I was looking at different scenarios it was apparent that I can mildly reduce the prediction error if I increase the m parameter from 3 to 9, but it just slighly improved the RMSE afterwards - hence the golden rule of using the square root of the number of variables seems to be true here as well. In terms of minimum node size, it seemed that there is not much of a difference when allowing 5, 7, 10 or 15 observations, but generally, the higher this was, the lower the resulting RMSE.

The best model had a cross validated RMSE of 102.8163. Not only that, it also faired quite well on the holdout set and the whole dataset as well. Table 2 summarises the RMSE results in their raw form and when normalised by the mean price. We can see that RMSE is highest in the cross validated case, and slightly lower on the holdout set, and surprisingly low when measuring it on the whole dataset.

Table 2: Model performance on the working set, holdout set and the whole dataset

	RMSE	Mean Price	RMSE - norm
CV RMSE	102.816	156.528	0.657
Holdout RMSE	101.642	154.098	0.66
All data RMSE	86.14	155.8	0.553

I also created two charts to look at variable importance of the best rf model, which can be seen in Figure 1. The first chart visualises variable importance for the top 10 predictor, while the second shows every predictor, but the city dummies are grouped together. From these charts we can't draw conclusions on casuality, or associations but what we can say is that, when predicting the price of our inspected accomodations, the trees in the random forest were mostly split by these categories. So if they were important for the random forest, they are important for the predictions as well. If anyone wishes to price their accomodations, they should

make sure that they include variables on the rating of the listing, the distance from the city center, and also in which city, their accomodation is located at. The grouped importance chart shows it very clearly that the city variable is essential when predicting, and could probably mean that there are very different price levels for different cities in the dataset.

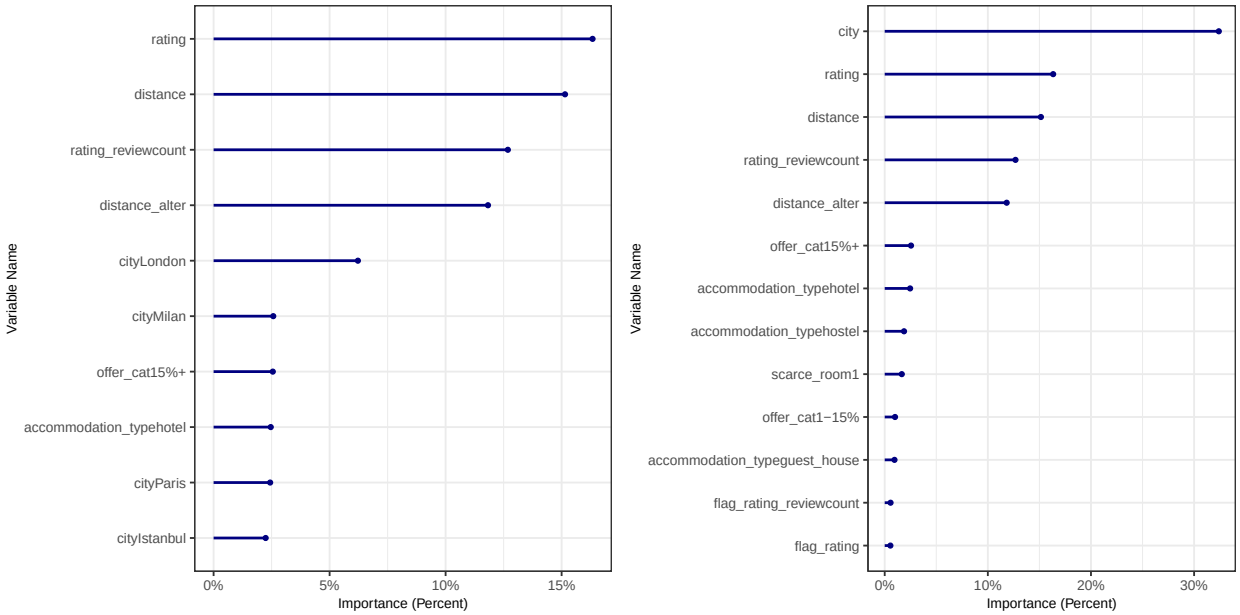

Figure 1: Top 10 and Grouped city variable importance plots

Conclusions

I set out to examine the `hotels_europe` dataset and to build a model to predict the prices of accommodations using machine learining techniques. After experimenting with different tuning parameters my best model had a minimum node size of 15 and limited the number of variables to select from to 13. One of the conclusions of the analysis is that tuning can improve the RMSE, until a point, but then it looks to be less effective afterwards. Another important note is that the `city` predictor was the most important variable for the model signalling that there is a big variation between accommodation prices between different cities.