DA3 EX2

Peter Kaiser

2022 02 09

Airbnb price prediction for Manchester

Introduction

The aim of this project to help a company that operates in Manchester and rents small and mid-size apartments to prize their units. The data is from http://insideairbnb.com/get-the-data.html and is from the 21st of December 2021. I built 5 different models and horse raced them to see which model performed the best.

Data

For Data Cleaning and munging, I filtered on the airbnbs, which were condos, residential units, rental units, and rooms. Filtered for size for 6 or less people. Removed the records, where the price was missing, or was more than 100 usd a night as I do not think that is in the range of the company. For other missing values, I imputed them or set them to 1 and flagged them. The impressive part was the amenities where I created a dummy variables for every amenity which was listed. I created every listed amenity in every listing, and added this list to the data and set the dummy to 1 if the record had that amenity. In the end I only kept the amenities which had at least 300 occurrences. I also created groups of the variables d_ for the dummies, f_ for factors and n_ for numerics. Some squares, cubes and logarithms were also added.

Models

First I created a hold out set with 30% of the data. These are not used in creating the models just used during the final test even after that as the selection of the model should be done before that of the models they are to imitate new data not used before.

And I created 3 set of predictor variables:

- The basic: accommodates, beds, property type, room type, bathrooms, neighborhood
- The advanced: The basic + amenities and reviews

the extra set for lasso:

- The extra: The advanced + interactions

Models: I used the same models we used in class.

- Random forest with the basic set.
- Random forest with the advanced set.
- OLS with the advanced set
- CART with the advanced set
- LASSO with the extra set
- GBM with the advanced set

Results

	CV RMSE
OLS	58.31478
LASSO (model w/ interactions)	58.78068
CART	60.80508
Random forest 1: smaller model	57.89759
Random forest 2: extended model	57.65785
GBM	57.99805

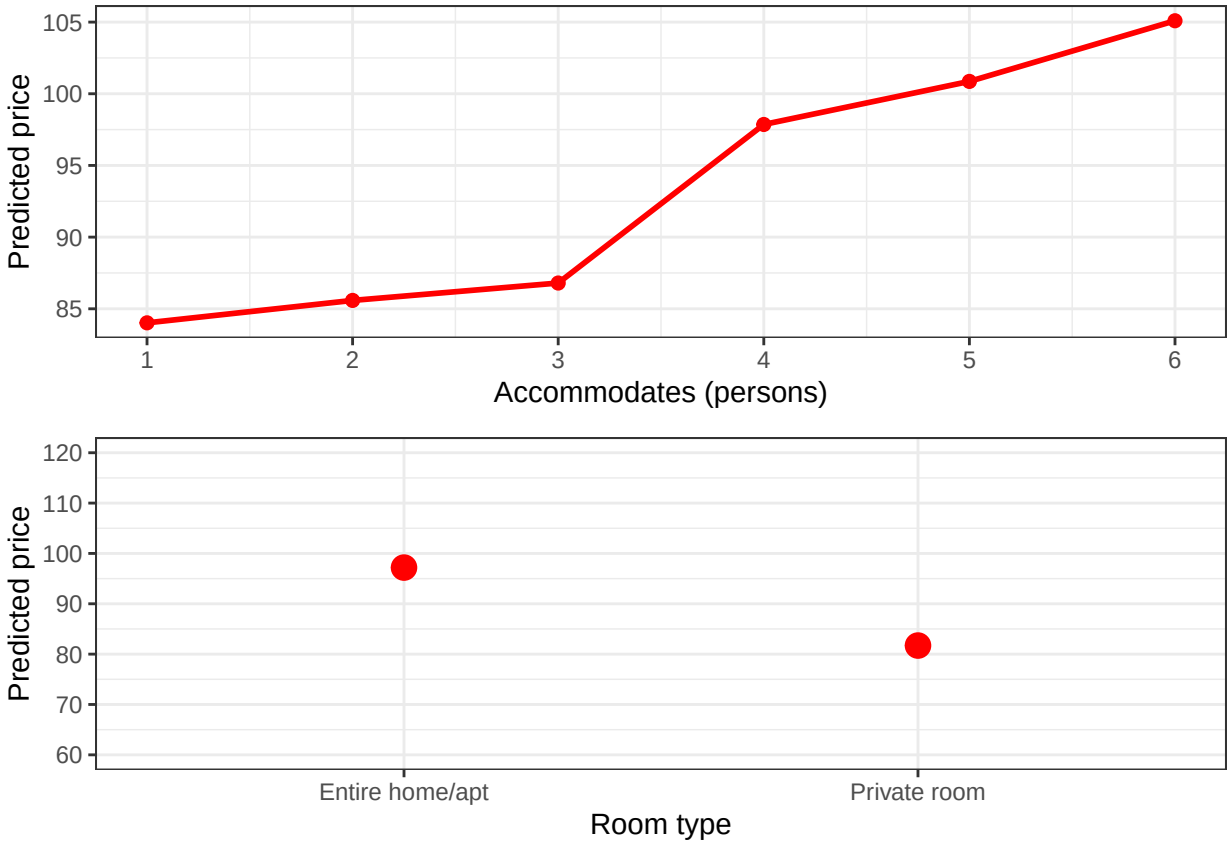
For the results we can see that the random forest and GBM did the best. Out of these the random forest with the advanced set did the best.

	Holdout RMSE
OLS	63.18380
LASSO (model w/ interactions)	64.68235
CART	64.65796
Random forest 1: smaller model	63.04198
Random forest 2: extended model	61.49551
GBM	61.12473

On the holdout set we can see that the random forest with the advanced set still did better than the others.

Random Forest

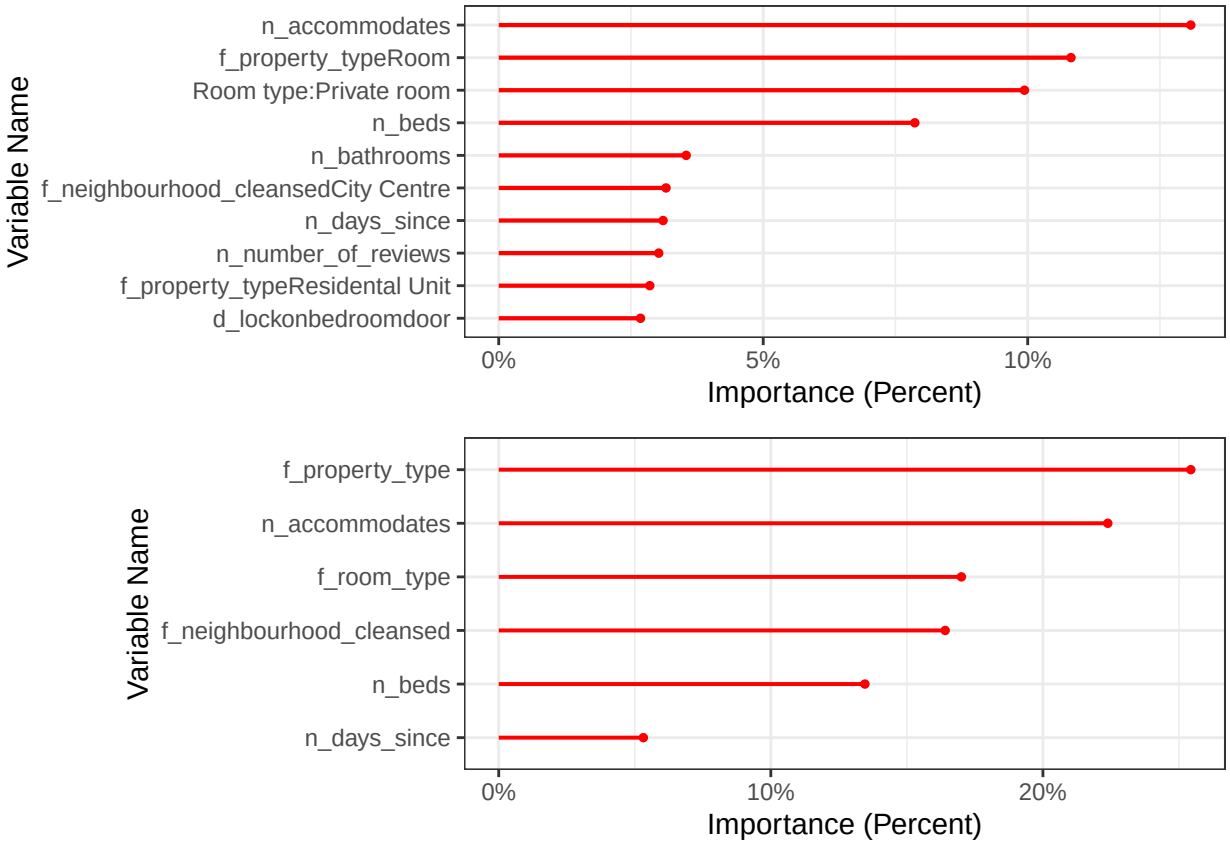
Both of the best model results came from black box models, but the advanced random forest did a little better. I checked partial dependence of accommodates and Room Type. We can see that the price is in a linear relation with the number of persons, also that the a entire unit pays more than a room as maybe we expected it. In the appendix I included the partial importance of the variables and the grouped importance of variables. The most important variable groups are, property type, number of accommodates, room type, and neighborhood.



Conclusion

In conclusion the best model we could use is the random forest model with the advanced variables, but all models are close in results. It is a black box model where we do not know the exact coefficients. About the external validity, this data was collected on the 21st of December, which maybe the holiday season so it may not be externally valid, also the molds have the Manchester city regions coded into them, so they are not usable for other cities. But i may be true that, more people cost more and that entire units cost more. In terms of the London models, the models were worse but that is maybe because of the lesser number of observations.

Appendix


n_accommodates
f_property_typeRoom
Room type:Private room
n_beds
n_bathrooms
f_neighbourhood_cleansedCity Centre
n_days_since
n_number_of_reviews
f_property_typeResidental Unit
d_lockonbedroomdoor
Variable Name
0%
5%
10%
Importance (Percent)
f_property_type
n_accommodates
f_room_type
f_neighbourhood_cleansed
n_beds
n_days_since
Variable Name
0%
10%
20%
Importance (Percent)