Data Analysis 3 - Assignment 2

Máté Schieszler - 2024.02.11.

Introduction

The aim of the project is to create a predictive model, which is able to price mid-size apartments hosting 2-6 guests in Milan. The model will be chosen from 3 (4) different models, using different prediction methods. The four models are created with OLS, Lasso, Random Forest and GBM.

Data

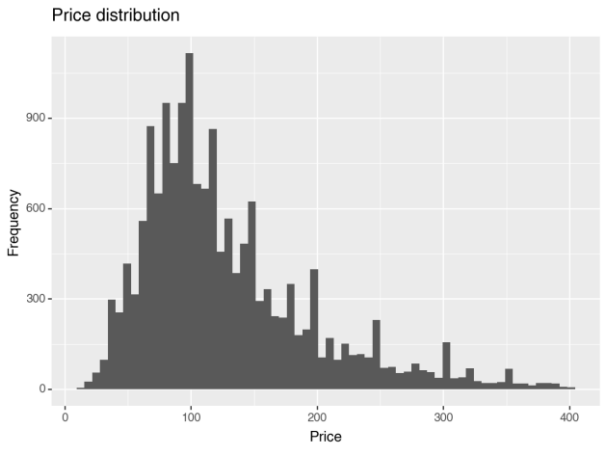
The data comes from the Inside Airbnb project. This project regularly scrapes and stores most the most important data from Airbnb's websites for the biggest cities. The data used during the project is from 2023.09.04., the most up to date data containing amenities for the accommodations.

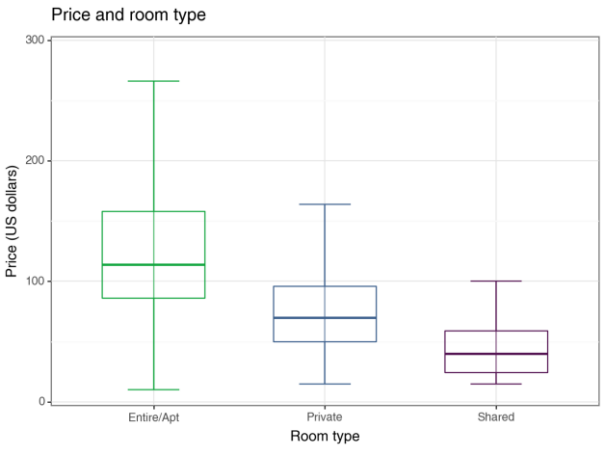
The initial shape of the data is (24177, 75) and contains information on the scraping, accommodations, reviews, availability, minimum and maximum nights, and host.

Data cleaning, Feature Engineering

The most important steps of the data cleaning process are choosing variables related to the accommodation meaning that columns relation to scraping and the host are dropped. The categorical columns are examined separately and filtered or redefine (grouped) to contain an adequate number of observations. Boolean variables are translated to dummies. Numerical variables are also filtered to drop errors (i.e. accommodation with 14 bathrooms) and to records that are in the intervals defined by the project aim. The 50 most frequent amenities are also converted to dummy variables. On price, percentage and bathroom columns transformations are made with regular expressions to kinyer the data. After the first round of cleaning the nan are examined. The columns containing a comparable number of NAN to all records are dropped for other columns first flags are created the missing values are imputed.

EDA

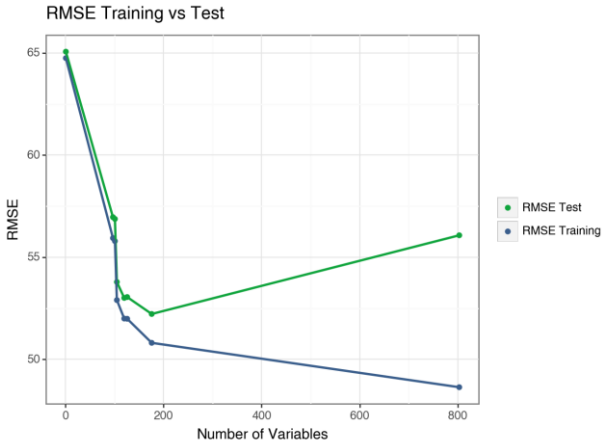




The most important step of the exploratory analysis is filtering the observations for price less than 400 USD at the 95% percentile. Then some amenity dummies are examined for possible good interactions with room type, already preparing for the OLS and Lasso models.

OLS

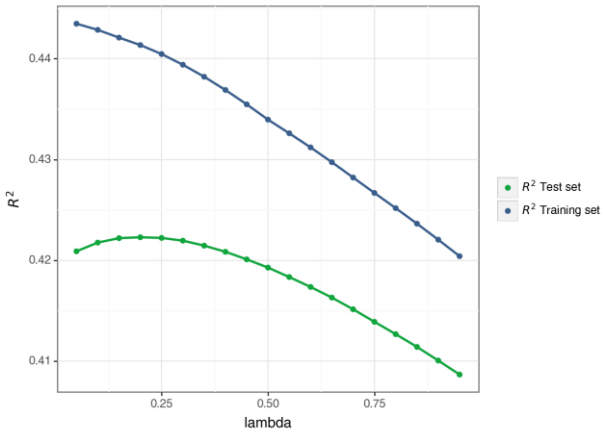
The best OLS model is chosen from 8 increasingly more complex models, containing basic, host, availability, and review related variables in for groups. Possible interactions discovered during the EDA process are also included in model 7 and in model 8 the perceived most important variables (chosen by domain knowledge) are interacted with all amenity dummies. All the models are cross validated on 5 folds. The average RMSE on the training data ranges from 64.75 to 48.75. When examining performance on the test data Model7 performs the best with 174 variables. OLS-Model7 has a 48.75 and a 52.34 RMSE respectively on the training and test dataset.

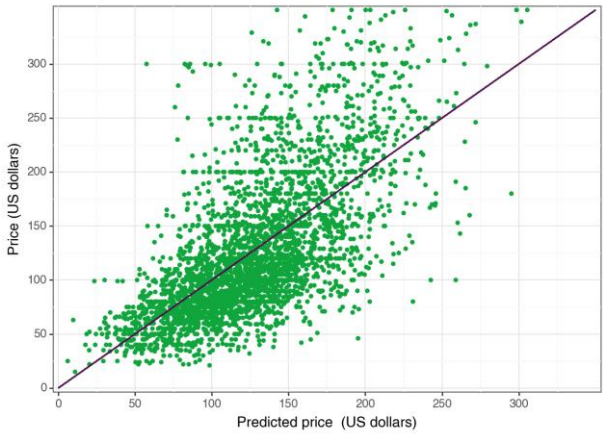


The graph comparing RMSEs on training and test data and number of variables shows that while Model7 is the best choice from this selection, its is most probably not the theoretically possible best model, but as it its nearly impossible to find the ideal combination of variables the project turns to alternative methods.

Lasso

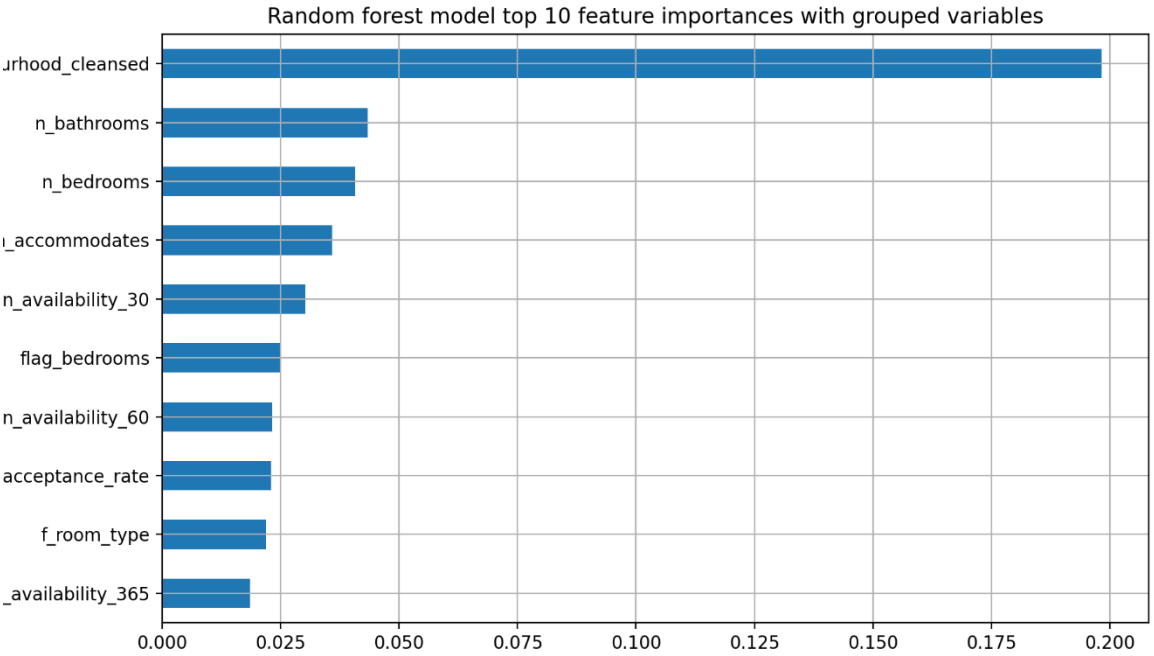
The lasso method is used on the best performing OLS model so that it can fine tune the variables. Lambda is examined from 0.05 to 0.95 by 0.05 increment. Graphing the prediction and the true price it's visible that the model overpredicts lower prices and underpredicts higher prices. Using the grid search method the Lasso Model is able to perform a 51.04 RMSE on the work dataset and a 50.94 RMSE on the test dataset.





Random Forest

The random forest is using the same variables as the previous models except for the predefined interactions and non-linear variables. The best preforming random forest has maximum 16 features and a minimum node size of 5. The method produces a model with 51.69 RMSE. The most important variables, when regrouped are neighbourhood, number of bathrooms, and number of bedrooms. Using GBM as a boosting method a significant improvement can be achieved as it performs a 47.94 RMSE (max_depth=10, max_features=10, min_samples_split=20, n_estimators=200), which is by far the best results of all the models.



Conclusion

Overall, out of the 3 (4) models built the GBM boosted Random Forest method has the best performance. Interestingly the Lasso model built using grid search is able to perform better than the unboosted Random Forest. The fact that amenities appear between the highest feature importances suggest that it is worth using not the most up to date, but the most complex data available, even if it means a few months old dataset. Regarding time, while only a simple grid search is used for the GBM, all the models are comparable and run in manageable timeframes, around a minute.

Model	RMSE
Gradient Boosting Model	47.9406
Lasso OLS	50.94186
Random Forest	51.59152
OLS CV	52.22308

github: https://github.com/mateschieszler/DA3-A2