Prediction with Machine Learning for Economists - Assignment 2

The data source for this assignment is the Airbnb dataset (http://insideairbnb.com/get-the-data.html). And I chose Shanghai as the subject of my analysis. The task is to help a company that operates small to medium-sized apartments hosting 2-6 guests, which will price its new unlisted apartments.

1. Models Selection and Comparison

After cleaning the data, based on all the variables and the corresponding domain knowledge, I chose the following predictors for the price prediction:

- Basic variables: the number of accommodations, the number of beds, the number of bathrooms, days since the first review, property types, and neighborhood categories. These are the variables that people consider first when choosing an apartment based on their needs.
- Review variables: the number of reviews, the rating scores, etc. These variables are important criteria for people to make apartment selection comparisons.
- Dummy variables for amenities: a series of dummy variables, including the availability of Wi-Fi, air conditioning, kitchen, etc.
- Interaction term: a series of interaction terms are used to investigate the problem of heterogeneity, and mainly for the subsequent LASSO model.

Meanwhile, I chose five models and compared their performances, namely OLS, LASSO, CART, Random Forest, and GBM.

i. OLS

I chose basic variables, review variables, and dummy variables for amenities as predictors when running OLS. And the RMSE equals 2316.54.

ii. LASSO

In the LASSO model, I additionally included a series of interaction terms and let the LASSO algorithm automatically decide which variables become the final predictors and the optimal tuning parameter with 5-fold cross-validation. In the end, the optimal tuning parameter is 0.49 and RMSE equals 2304.68. By comparing these two models, we can see that the LASSO algorithm reduces the prediction error and its automaticity is also a great advantage.

iii. CART Model

The CART model contains the same predictors as the OLS model. It also applied a random search to select a "best" complexity parameter and ended with RMSE 3103.33. This shows that the CART model is not a good prediction model.

iv. Random Forest

A random forest needs a little tuning, i.e. three parameters: the number of bootstrap samples, the number of variables considered for each split, and the minimum number of observations in the terminal nodes of each tree as a stopping rule. For the number of bootstrap draws, I went with the default option of 500. For the number of variables, one rule of thumb is to pick the square root of the total number of variables, which would be around 10, so we tried 8, 10, and 12. For the minimum number of observations in the terminal nodes, we choose 5, 10, and 15. The results are shown in the following table:

min node size	5	10	15
max features			
6	2278.92	2283.33	2280.62
8	2273.78	2277.18	2277.23
10	2275.00	2276.82	2277.39
12	2274.75	2275.25	2278.81

From the table, we can find that the RMSE is the smallest with max_features of 8 and min_samples_leaf of 5, which is 2273.78. It turns out the Random Forest model that combines predictions from many imperfect models can produce very good predictions.

v. GBM

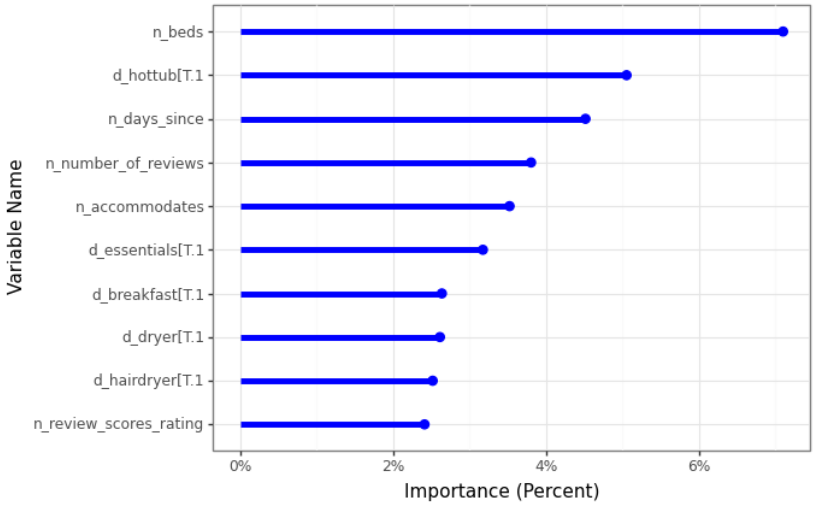
GBM is one of the boosting methods which are based on combining predictions of many trees that are built sequentially, each tree improving on the previous one. GBM requires more tuning parameters than random forest. They determine the complexity of trees, the number of trees, how we combine the trees to form the new prediction at each step, and how large each tree should be. We can build a search algorithm to find the parameter combination that provides the best fit (using cross-validation), but we need to specify a range for each parameter in which the algorithm should search. The final optimal model is constructed with max_depth of 1, max_features of 10, and min_samples_split of 20. And it gives RMSE which is 2285.67.

	model	CV RMSE
0	OLS	2316.540192
1	LASSO	2304.678268
2	CART	3103.326239
3	Random Forest	2273.784075
4	GBM	2285.667593

From this table, we can see that both the Random Forest and GBM models show better predictive power, and the LASSO model's ability to automate the shrinkage of predictors also gives it better predictive power than OLS. Although the CART model does not have good prediction, it can provide the interpretation of the model from a nonlinear perspective.

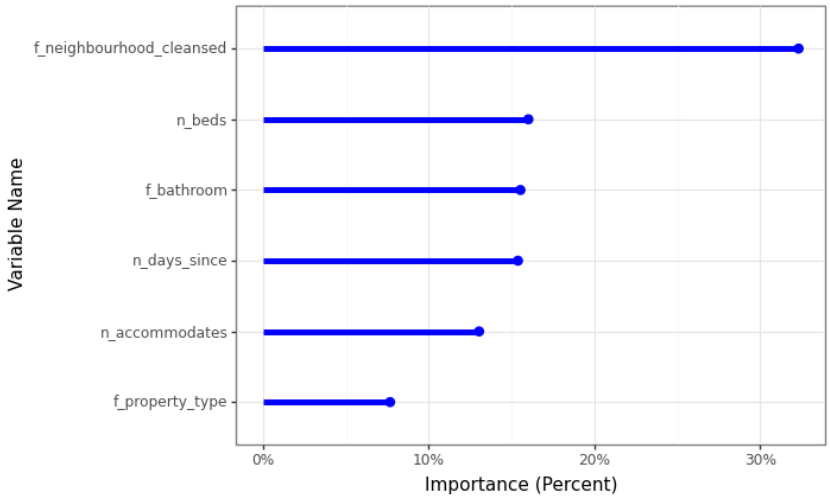
2. Model Diagnostics for Random Forest

i. Variable Importance



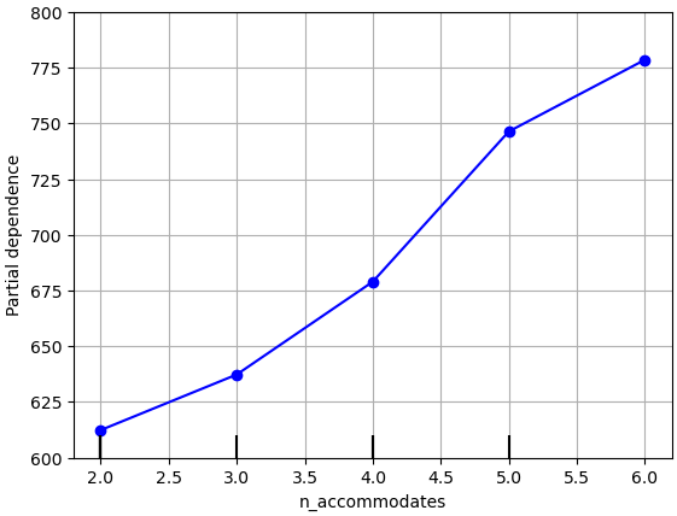
This is a graph of the importance of all the variables in the top 10, from which we are able to see that the most important variables for prediction are the number of beds, the availability of the hot hub, days since the first review, the number of reviews, the number of people it can accommodate.

Another way is to group qualitative (factor) variables that are entered as several binary variables. Variable importance then is a sum of gains (in terms of MSE reduction) by splits involving the factor variable. For instance, rather than considering splits by the borough binary variable, we now consider all the splits by any values of the neighborhood variable.



From this figure, we can find that neighborhood becomes the most important variable for prediction, followed by the number of beds and bathrooms categories variable. This is reasonable because the price difference between different areas of Shanghai is quite large, and the central areas located in Huangpu, Jing'an, Changning and Xuhui districts tend to be more expensive. However, when we look at this area variable separately, it also becomes less important due to the higher similarity of the adjacent areas.

ii. Partial Dependence Plots



The partial dependence plot of the number of accommodations shows that as the number of accommodations gets higher, the price increases accordingly. We can also perform corresponding analyses for other variables, which are not shown more in this report due to space limitations.

iii. Subsample performance

	rmse	mean_price	rmse_norm
Apartment size			
large apt	3793.78	934.97	4.06
small apt	1183.06	485.88	2.43
Type			
Entire condominium (condo)	733.99	479.53	1.53
Entire loft	2136.82	618.01	3.46
Entire rental unit	2827.72	678.52	4.17
Entire serviced apartment	1779.85	686.03	2.59
All	2445.64	644.24	3.8

Therefore, this diagnostic work concludes that price prediction is better for small condominiums than for large condominiums, and among the different property types, the Entire condominium has the best price prediction.