Airbnb Pricing in Shanghai / China

Data Analysis 3 - Assignment 2

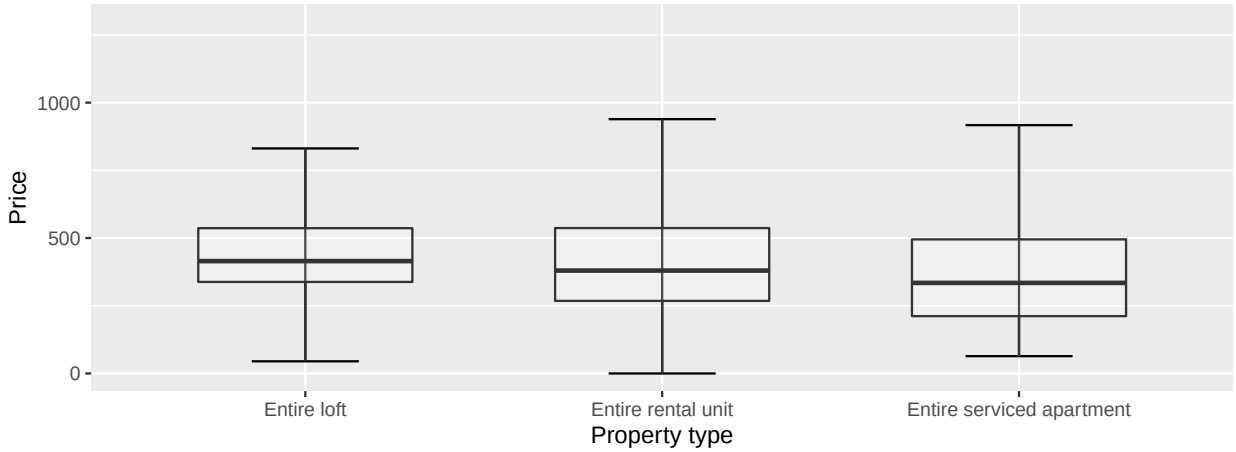
Ersan Kucukoglu

Introduction

This study focuses on a single location - Shanghai, China - and seeks to estimate airbnb rental costs per night. The aim of the study was to assist company to predict prices for their small and mid-sized apartment accommodating 2-6 people using different prediction model. I'll be using data from Inside Airbnb to create these pricing prediction models. I'll be creating and comparing Airbnb predicting models for the city of **Shanghai**, China, to determine the optimal combination algorithms for assessing the prediction model. OLS Linear Regression, Lasso, and Random Forest were three machine learning techniques that I utilized. The used dataset was downloaded from insideairbnb.com which is a site collecting data on Airbnb listings in numerous cities. As mentioned above the dataset contains data on listings in Shanghai and was last updated on 24th December 2021.

Feature engineering

Having decided about the functional form of the target variable I inspected the explanatory variables and their relationship with the target variable. Besides deciding on grouping of factor variables functional forms also had to be decided because two of the used prediction models were OLS and OLS with LASSO for which this step is necessary.



First, I inspected the two factor variables: property type and neighbourhood. The prices conditional on property type can be seen in the chart above. Since there are differences between the three categories in conditional means as well as standard deviation I decided to keep all three.

I examined the dummies after the categorical variables. I calculated the conditional mean price for each dummy and came to the conclusion that my assumptions were correct in the sense that an apartment with more extras priced more in general. I also checked the number of missing values for each of the variables. As for the explanatory variables : I imputed missing values using the variable's median according to some

methods. Lastly, I examined possible interactions between the property type factor variable and all the dummy variables by using plots. I created two lists of interactions: one for OLS and one for LASSO.

OLS and LASSO

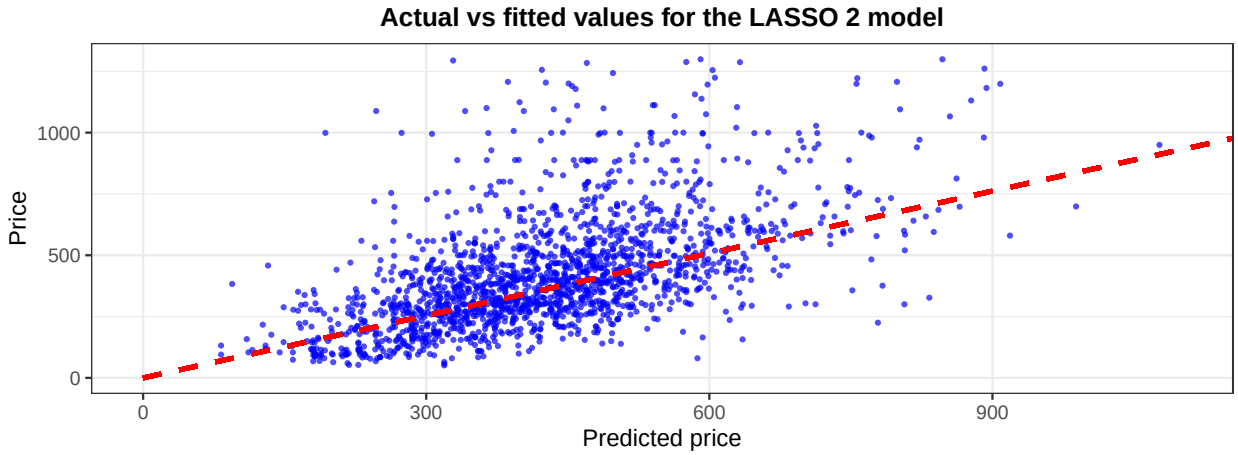
In total I estimated 3 OLS models with different sets of variables. The first one contained only the levels of variables, the second one contained the levels and the transformed versions, while the last one contained interactions as well. For LASSO, I estimated two models: one with levels and transformed versions of variables plus a few interactions and one with levels and transformed versions of variables plus all the interactions I previously determined.

The table below gives information about the models' performance as well as the number of coefficients they had. We can observe that the best performing OLS 3 and LASSO 2 models roughly have the same number of coefficients based on both cross-validated and holdout RMSE, but they do not totally overlap. The difference in their performance is tiny.

Table 1: OLS and LASSO performance

	Number of Coefficients	CV RMSE	Holdout RMSE
OLS 1	47	192.93	189.57
OLS 2	55	192.20	189.32
OLS 3	73	191.68	188.96
LASSO 1 (few interactions)	62	191.55	188.68
LASSO 2 (all interactions)	64	191.39	188.99

Because the LASSO 2 model performed the best, I created the plot below, which displays the actual prices as well as the prices predicted by this model. We can observe that it gives better predictions for cheaper prices and somewhat underestimates prices.



Random forest

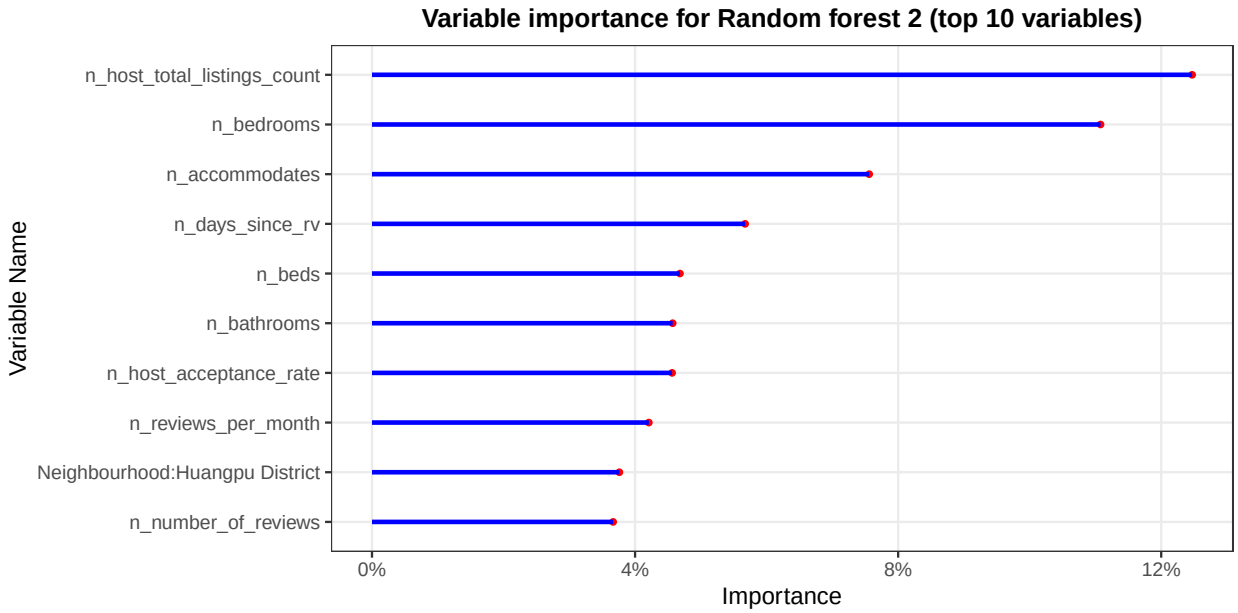
I estimated two random forest models in total using the levels of all variables. The models differ in the parameter sets that determine the number of randomly chosen variables at each split and the minimum number of observations in the terminal nodes for each tree. Following the rule of thumb first I set the number of randomly chosen variables at each split to 6 which is around the square route of all variables. However, as the table below shows both final models produced better results when setting the parameter to a higher number.

To be able to decide which model is better I calculated the cross-validated RMSE-s on the test sets. We can see in the table below that Random forest 2 performs better than Random forest 1. But again just like in the case of the OLS and LASSO models the difference between the two models is relatively small.

Table 2: Performance of random forest models

	RMSE
Random forest 1	163.05
Random forest 2	162.80

It is useful to see which factors contributed the most to the model's RMSE decrease.I created a variable significance plot for that model, which displayed the top ten variables that contributed the most to the reduction of RMSE in percentages. The very first variable indicates how many listings a host has in total.I was expecting the neighbourhood Huangbhu district in the first 3 variables, because it is heart of the city and popular.



Conclusion

After running different types, I choose the best results from each method. I described their cross-validated RMSEs as well as their RMSEs determined on the holdout set in the table below. According to these data, the random forest model outperformed the LASSO and OLS models. However, the pricing plan I would recommend for the firm would be determined by their choices. If they want to see the relationship between specific explanatory factors and the target variable and have coefficients, I would recommend using either the LASSO model or the OLS because their performance is fairly close. Otherwise, I would suggest using the random forest model since that one has the best prediction performance.

Table 3: Model performance comparison

	CV RMSE	Holdout RMSE
OLS 3	191.68	188.96
LASSO (all interactions)	191.39	188.99
Random forest 2	162.80	163.55