Assignment 2 in Data Analysis 3

MS in Business Analytics

Jo Kudo

Introduction

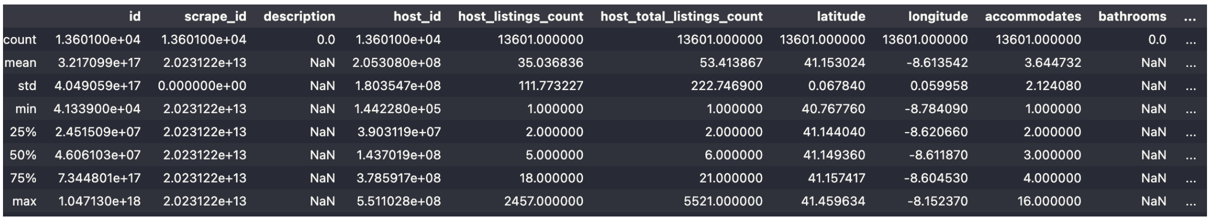
In Porto, our company manages a portfolio of small to medium-sized apartments that accommodate between 2 to 6 guests per unit. Currently, we are in the process of developing a pricing strategy for our upcoming apartment listings, which have yet to be introduced to the market. This report is dedicated to the construction of a predictive pricing model tailored for Airbnb accommodations located in Porto.

Our approach involves a thorough evaluation and comparison of various predictive models. The key criterion for this comparison is the RMSE, or Root Mean Squared Error, which is a standard measure used to quantify the accuracy of price predictions. By examining the RMSE values, we aim to determine the most suitable model for forecasting prices that have not been previously established.

The objective is to select a predictive model that not only provides precise pricing recommendations but also complements our strategic objectives for market entry. Identifying the optimal model is critical to ensuring that our new listings are competitively priced, thereby positioning our company as a strong contender in the local accommodation sector.

Feature Engineering

The information used for this report comes from an Airbnb website. It has 75 different details about the places people can rent. These details help us build a good model to suggest prices for these rentals. The original dataset looks like below:



	id	scrape_id	description	host_id	host_listings_count	host_total_listings_count	latitude	longitude	accommodates	bathrooms	...
count	1.360100e+04	1.360100e+04	0.0	1.360100e+04	13601.000000	13601.000000	13601.000000	13601.000000	13601.000000	0.0	...
mean	3.217099e+17	2.023122e+13	NaN	2.053080e+08	35.036836	53.413867	41.153024	-8.613542	3.644732	NaN	...
std	4.049059e+17	0.000000e+00	NaN	1.803547e+08	111.773227	222.746900	0.067840	0.059958	2.124080	NaN	...
min	4.133900e+04	2.023122e+13	NaN	1.442280e+05	1.000000	1.000000	40.767760	-8.784090	1.000000	NaN	...
25%	2.451509e+07	2.023122e+13	NaN	3.903119e+07	2.000000	2.000000	41.144040	-8.620660	2.000000	NaN	...
50%	4.606103e+07	2.023122e+13	NaN	1.437019e+08	5.000000	6.000000	41.149360	-8.611870	3.000000	NaN	...
75%	7.344801e+17	2.023122e+13	NaN	3.785917e+08	18.000000	21.000000	41.157417	-8.604530	4.000000	NaN	...
max	1.047130e+18	2.023122e+13	NaN	5.511028e+08	2457.000000	5521.000000	41.459634	-8.152370	16.000000	NaN	...

The report begins by selecting data for accommodations suitable for 2 to 6 guests, in line with the guidelines provided. Following this, variables initially presented as text are converted into numerical form for analytical purposes. Additionally, the dataset retains a selection of property types to investigate the hypothesis that certain categories of accommodations may significantly impact pricing.

Model building, prediction and model selection

Three models are built and compared with their RMSE. The following models are considered:

1. Random Forest

   The first model is from Random Forest, which is a machine learning algorithm that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It uses a technique of random sampling of training data points and features when building trees.

2. Linear Ordinary Least Squares (OLS) Regression

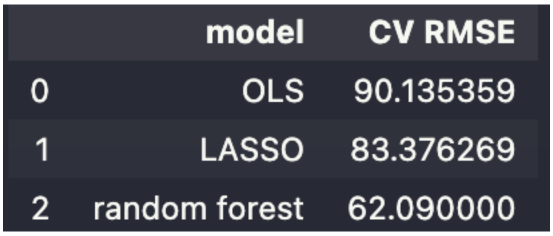
   The second model is from Linear OLS Regression. The aim is to draw a line that best fits the data by minimizing the sum of the squared differences between the actual data points and the line.

3. LASSO (Least Absolute Shrinkage and Selection Operator)

   The third model is from LASSO, which is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. The key feature of LASSO is its ability to shrink the coefficients of less important variables to exactly zero.

Comparing the three model with RMSE

Here is a table, which illustrates the comparison of the three models with RMSE. To calculate RMSE, cross validation method is used. According to this table, RMSE from Random Forest see the lowest score, which means Random Forest is the best model for predicting the price of the future accommodation.



	model	CV RMSE
0	OLS	90.135359
1	LASSO	83.376269
2	random forest	62.090000

Conclusion

In the end, when looking at how well three different ways of guessing prices did, each one had a different score for how close the guesses were to the real prices. The score used is called RMSE, which stands for how much the guesses are off, on average. Out of all of them, the Random Forest was the best at predicting what prices will be like later on. This is because it combines a lot of simple guessers into one big guesser that makes better predictions and doesn't mess up when things get tricky. So, for figuring out future prices well, Random Forest is the best choice.