Finding Best Hotel Deals in Vienna

Attila Serfozo

2021-02-14

Abstract

This project aimed to build a hotel price prediction model to find best hotel deals in Vienna on November 2017 and compare the 5 best deals to the results of book Data Analysis for Business, Economics, and Policy by Gabor Bekes and Gabor Kezdi. During the exercise I cleaned the data, grouped variables and created 5 different models. Based on the results the Random forest provided the best prediction with an RMSE of 0.20 USD on log price target variable and an R-squared of 62.4%. But overall we can state that the model tends to underestimate the price of more expensive hotels with prices above 125 USD. In the end I selected the 5 best deals according to my model based on log price differences, and 4 deals out of 5 matched the highlighted hotels from the Gabor Bekes and Gabor Kezdi book.

Introduction

In this project my goal was to build a price prediction model for 3-4 star hotels in Vienna and compare the best deals to the results of Book Chapter 10 in the Data Analysis for Business, Economics, and Policy by Gabor Bekes and Gabor Kezdi Cambridge University Press 2021. The dataset represents the hotels available on a price comparison site on November 2017. In the exercise to create an accurate prediction I used key variables such as distance from center, number of stars, neighbourhood and review variables such as number of ratings and ratings, tripadvisor ratings. I looked at 5 different models including 3 OLS, a CART and a Random forest and in the end I wanted to finish with a final model with relatively good prediction power regarding it's RMSE and R-squared. Based on the final model I selected the 5 best deals from the dataset and compared them to the highlighted 5 from the Data Analysis for Business, Economics, and Policy book.

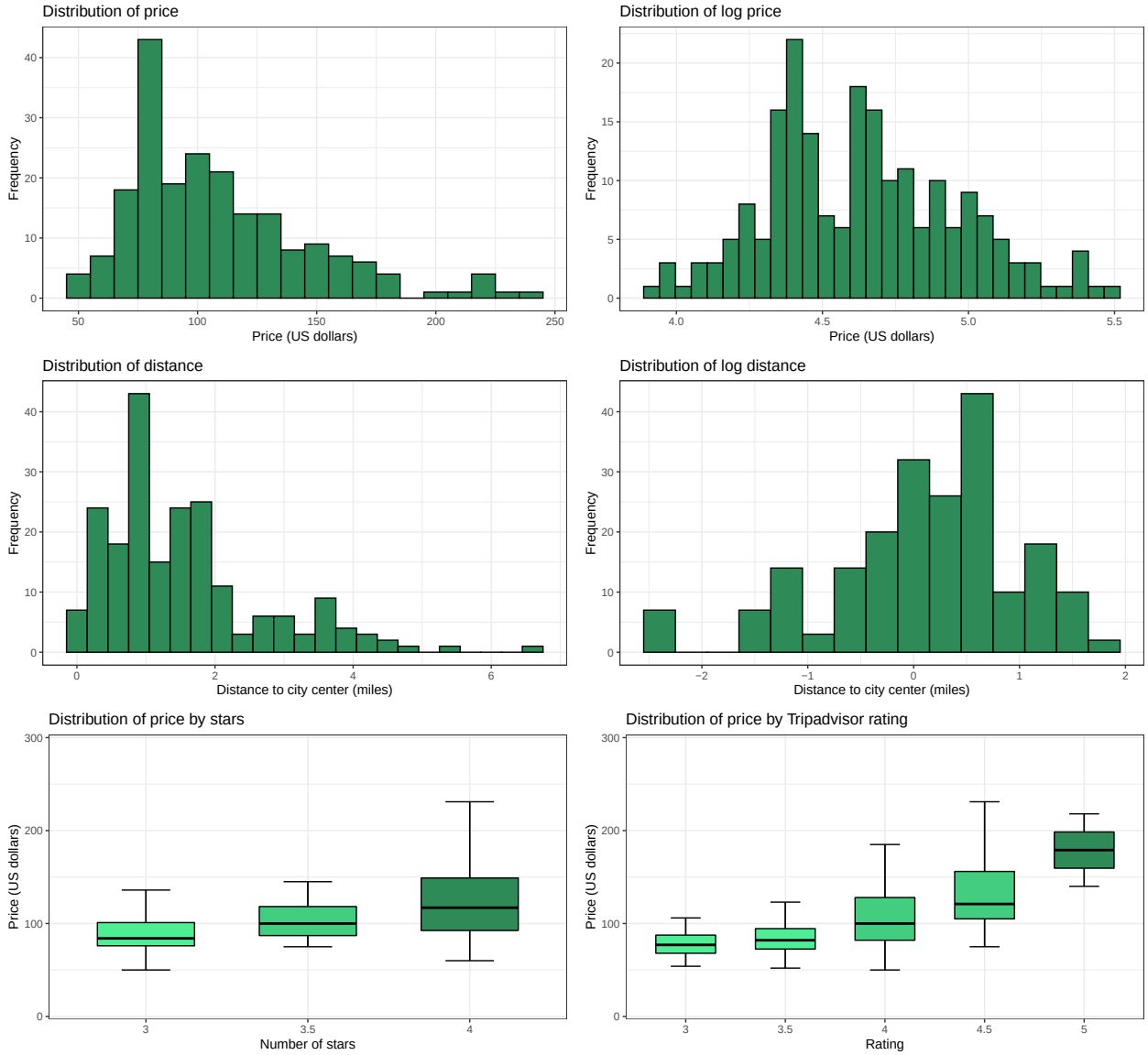
Data Preparation

In the beginning of the data preparation I started with 430 observations and 24 variables, all filtered on November 2017 which got reduced to 428 rows after removing duplicated observations. I also needed to treat missing values, as there were several hotels with missing data on ratings. Therefore I flagged missing values and decided to substitute all with the median of the variables. As a result I finished with 19 cleaned variables including numerical, factored category variables, binaries and flags.

To fit the dataset on the analysis purposes I filtered the clean data to hotels between 3 and 4 stars in Vienna actual. In addition when I had a look at extreme values, I excluded hotels with a one night price above 300 USD as there were one with a price of 383 USD and one with 1012 USD and they significantly lower the prediction accuracy of models. Thus, I finished with 206 observations on 3-4 stars hotels in Vienna.

Finally I looked at the distributions of key variables and their statistics. As it can be seen on the price distribution below, hotel prices are skewed with a long right tail. To transform it into a closer to normal form I decided to have a look at the logarithmic form and seeing the results I decided to use log price during the modeling exercise as a target variable. Also, based on the similar observation on distance from center

I applied a log-form on this variable as well. I also had a look at the distribution of price by stars and tripadvisor rating, the boxplot diagram results can be seen below.



As a last step of the preparations I separated my variables to 3 groups:

- **Main variables:** including distance in log, number of stars, ratings and neighbourhood.
- **Review variables:** including rating count, tripadvisor rating and tripadvisor rating count. I did not include flags as they did not improve the results significantly.
- **Other variables:** including alternative distance, offer available dummy, type of offer and flag if room was noted as scarce.

Modelling

In this time I did not use 5-fold cross validation for modeling purposes as we just would like to build a model to find the 5 best deals in this data and we are not planning to use this model for external prediction.

I created 5 different models:

- **OLS** using main variables
- **OLS** using main and review variables
- **OLS** using all the variables including main, review and other.
- **CART** using all the variables including main, review and other.
- **Random Forest** using all the variables including main, review and other.

The results of the models can be found below. Based on the model results the Random Forest model performed significantly the best, it had 0.20 USD RMSE on log price and an approximately 62.4% R-squared. Based on the results which can be seen below in the 2 tables my final model choice is the Random forest.

Table 1: Models R-squared

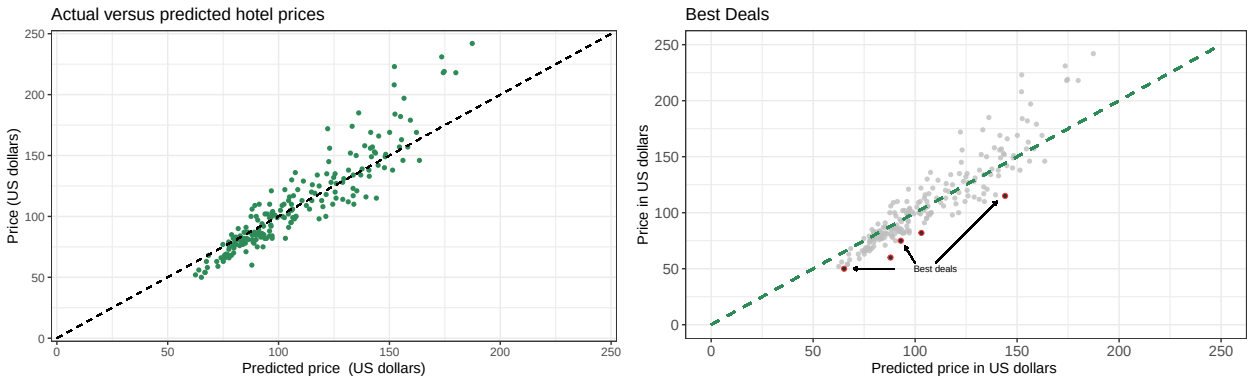
	CV Rsquared
OLS (main)	0.5452
OLS (main + reviews)	0.5733
OLS (main + reviews + others)	0.5765
CART	0.4907
Random forest	0.6237

Table 2: Models RMSE

	CV RMSE
OLS (main)	0.2210
OLS (main + reviews)	0.2173
OLS (main + reviews + others)	0.2170
CART	0.2433
Random forest	0.2086

Actual versus predicted price and Best Deals

As the aim of the project was to predict hotel prices and find the best deals, I had a look at the distribution of predicted prices versus actual prices of hotels, which can be seen below on the scatterplot chart. As we can see the final model tends to underestimate prices of more expensive hotels with prices above 125 USD. On the right scatterplot the 5 best deals can be seen highlighted in red according to the Random forest model.



Finally the 5 best deals according to the model and the 5 best deals from the book can be found below in the tables. In the selection of the 5 best deals the main decision point was the log price residuals as it can take

into account the percentage difference in prices instead of the absolute differences. Out of the 5 best deals highlighted in the book I had only one, hotel ID - 22080, which did not appear to be a best deal according to my final model.

Table 3: Best 5 hotels according to the model

ID	Price USD	Predicted price	Residual	Log Residual	Distance	Neighbourhood	Stars	Rating
21912	60	87.94	-27.94	-0.3745	1.1	Alsergrund	4	4.1
22344	50	65.19	-15.19	-0.2574	3.9	Vienna	3	3.9
22118	82	103.08	-21.08	-0.2209	1.7	Landstrasse	4	4.1
21975	115	144.11	-29.11	-0.2178	0.1	Innere Stadt	4	4.3
22184	75	92.99	-17.99	-0.2072	0.7	Leopoldstadt	3	4.1

Table 4: Best 5 hotels in the book

ID	Price USD	Predicted price	Residual	Log Residual	Distance	Neighbourhood	Stars	Rating
21912	60	87.94	-27.94	-0.3745	1.1	Alsergrund	4	4.1
22344	50	65.19	-15.19	-0.2574	3.9	Vienna	3	3.9
21975	115	144.11	-29.11	-0.2178	0.1	Innere Stadt	4	4.3
22184	75	92.99	-17.99	-0.2072	0.7	Leopoldstadt	3	4.1
22080	54	66.86	-12.86	-0.2057	1.1	Josefstadt	3	3.2

Summary

In this report I tried to find good deals among 3-4 star hotels in Vienna on November 2017. I created 5 models to predict prices from which I selected the Random forest as the final model seeing its predicting power with 0.20 USD RMSE on log price and an approximately 62.4% R-squared value. Overall we can see on the prediction that the model tends to underestimate prices of more expensive hotels with values above 125 USD. In the end I found the 5 best deals in Vienna according to our model based on percentage differences in prices. In the 5 best deals 4 out of 5 matched the ones in the Gabor Bekes and Gabor Kezdi book.