**Pricing model for Airbnb, based on New York City data**

V. Komarova

**Introduction**

This project focuses on price modeling for short term Airbnb rentals. The business goal is to come up with a service that would recommend new Airbnb users how to price their listings given certain parameters. Such service can be utilized within Airbnb itself or by an independent company. I will use Airbnb open data for New York. The data set is publicly available [here](https://www.kaggle.com/datasets/arianazmoudeh/airbnbopendata?resource=download).

From a research standpoint, real estate pricing models are interesting because of their complexity: they are regression, not classification problems; pricing projects deal with a variety of variables (continuous, binary, factor), there is a lot of data available on individual sales and listings, but it is not quite clear to what extent the price is determined by the features of an individual property, and to which by the macroeconomic factors, such as state of the job market, mortgage rates, etc.

**Data cleaning**

While preparing data for analysis, I had made several assumptions; in this section I will explain my reasoning.

**Selecting features**

The initial dataset contained 102,599 observations of 26 variables.

Data set contained several features that would not have determined the price or were purely technical. For the final dataset the following variables were dropped:

* Id, name, host id, host name, country, country code, house rules, last review, license.
* latitude and longitude - this information is in the neighborhood variable.
* reviews per month - had a lot of NAs and dropping these would lead to dropping a lot of columns with information.
* service fee, as it is a percentage of the price, so it will confuse the model.

The final model contains 12 variables: host\_identity\_verified, neighbourhood.group, neighbourhood, instant\_bookable, cancellation\_policy, room.type, Construction.year, price, minimum.nights, number.of.reviews, review.rate.number, calculated.host.listings.count

**Removing outliers**

From the data summary, it is clear there are outliers or mistakes done when scraping the data. The most concerning ones for this project are in the price variable, but negative values in minimum nights should also be addressed. Since the dataset has a lot of observations, I made a decision to filter out the outliers instead of filling them out with synthetic data such as mean values.

When removing values, I am making the following assumptions:

* private and shared rooms with price over $500 are most likely a mistake
* entire home for less than $100 is also most likely a mistake
* negative values in minimum nights are a mistake

Since the project focuses on price modeling for short term rentals, I am also dropping all listings that require more than 29 nights. Assumption here is that pricing strategy for short term rentals is very different from long term, so incorporating both into the model can lead to bias.

After removing outliers and missing values the final dataset contains 58,766 observations.

**Exploratory data analysis**

For a graphical representation of the summary of the variables as well as graphical distribution, please refer to the Appendix. Nothing stands out as unusual or suspicious in the exploratory data analysis.

**Data balance**

A picture containing table

Description automatically generated

Number of observations by neighborhood groups

Graphical user interface, application

Description automatically generated with medium confidence

10 best represented neighborhoods

For a table of 30 worst represented neighborhoods, please see Appendix

When looking at neighborhood groups, the data set has significantly less entries for Staten Island. When looking at the data grouped by neighborhood group and neighborhood, we see that the data set has very few entries for some neighborhoods in Staten Island, Queens and Bronx. The most well-represented neighborhood groups are Manhattan and Brooklyn which is very reasonable, given that these neighborhoods are the main tourist and business destinations.

A picture containing chart

Description automatically generated

Observations by property type

When looking at the balance between different room types, ‘entire house/apt’ is the best represented category (logical, as this would probably be the most popular), but we have enough observations here for both private and shared rooms, so balance for this category is not a concern.

Correlation table was not very informative with a mix of untransformed variables, so all variables were selected into the model.

**Models**

Pricing model is a regression problem, so linear regression and neural networks would be the two most obvious choices to try to apply to this problem.

For linear model I get the following results: Mean squared error: 77242.09 and R squared: 0.27

For neural networks I get the following results:

|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R squared** |
| 1 layer, 10 neurons | 77345.39 | 0.27 |
| 2 layers, 10 neurons | 77234.989 | 0.27 |
| 3 layers, 10 neurons |  | 0.27 |

I am very concerned with the high MSA and relatively low R squared. But the fact that both models produced very similar results hints that the problem might be in the data (or rather what’s missing from the data).

When re-testing the model on a dataset with only Manhattan and Brooklyn properties – on the assumption that these neighborhood groups were best represented – I got very similar results to the full set.

Linear model: MSE 78172.44, R squared – 0.26

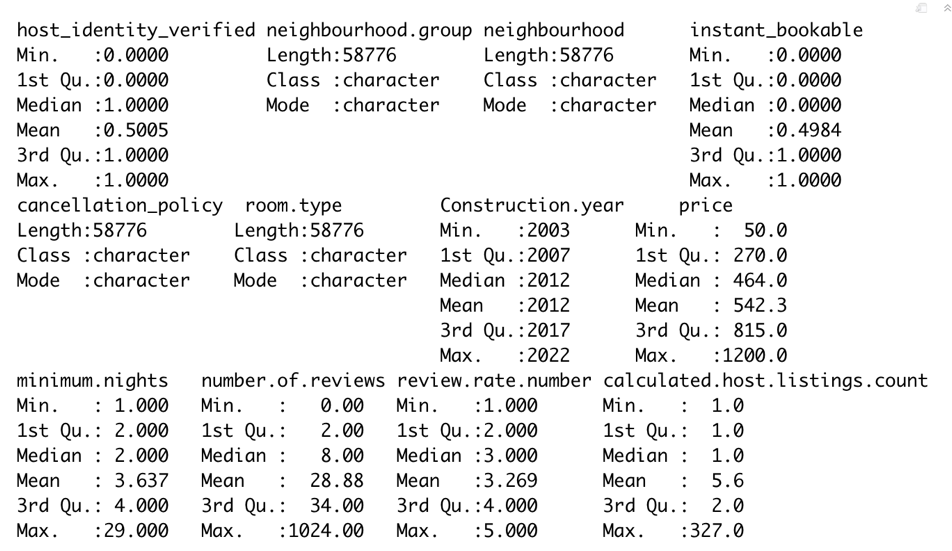
|  |  |  |
| --- | --- | --- |
| **Model** | **MSE** | **R squared** |
| 1 layer, 10 neurons | 78,183 | 0.26 |
| 2 layers, 10 neurons | 78,196 | 0.26 |
| 3 layers, 10 neurons | 78,174 | 0.26 |

Given the results, I can draw two main conclusions, that also point towards areas of future research:

* Some key features that determine the price could be missing from this dataset. Re-testing the models on a dataset that contains more features could potentially improve the model. I would be particularly interested in integrating two additional categories of features:
  + square footage, number of bedrooms and number of people a property can accommodate;
  + neighborhood walkability score and proximity to public transport, tourist attractions, as well as cafes and restaurants.
* Macroeconomic factors, such as overall health of the economy (as approximated by the GDP) could play a significant role in determining the price. Potentially we could look at panel data and see how the prices have changed over time and determine if they go up in the years when people and companies have more disposable income for personal and business travel.

**Appendix**

Summary of the final dataset



Timeline

Description automatically generated

Distribution of the numeric variables

Table

Description automatically generated

30 worst represented neighborhoods