Assignment 2 - Predicting Rio Airbnb Prices Summary

A. BOGNAR

Introduction

For our expansion in the Brazilian rental market initial predictive models were built to assess their efficiency for pricing. In this paper I present the main findings.

Data

Airbnb data for rental units is publicly available for several cities. In this exercise we focused solely on apartments similar in size to our new acquisitions in Rio de Janeiro.

While the data is from the Airbnb database, the information of the units is always provided by the sellers. This leads to human error. Only some fields require minimal validation. Systematic errors were corrected but the data issues may still affect the results. Specifically, amenities and services provided with the property are entered in free-text fields, thus not all relevant predictors may be captured.

The predictive variables used in the analysis can be grouped to four categories:

- **General characteristics** of the apartment: number of accommodates, bedrooms, location etc.
- Number of **reviews**.
- Characteristics of the **host**. These may be instructive for our marketing strategies to understand important qualities for rentees.
- **Amenities**.

For full explanation on feature/label engineering see the technical report.

Models

Several models were built with three different methods: linear regressions, LASSO and Random Forest. Below I present the results of one selected model from each type. The OLS model is not the best performing one of its type. However, it is instructive to understand underlying patterns. The main lessons:

- Proxies used for the size of the apartment (accommodates, bedrooms) have the strongest association with the price.
- Santa Teresa is the cheapest district, Leblon is the most expensive.
- The central district also offers very cheap accommodations, even cheaper than the aggregate of districts that have less than 100 units in the dataset.
- *Superhosts*, who are in Airbnb partnerships tend to offer cheaper accommodations.

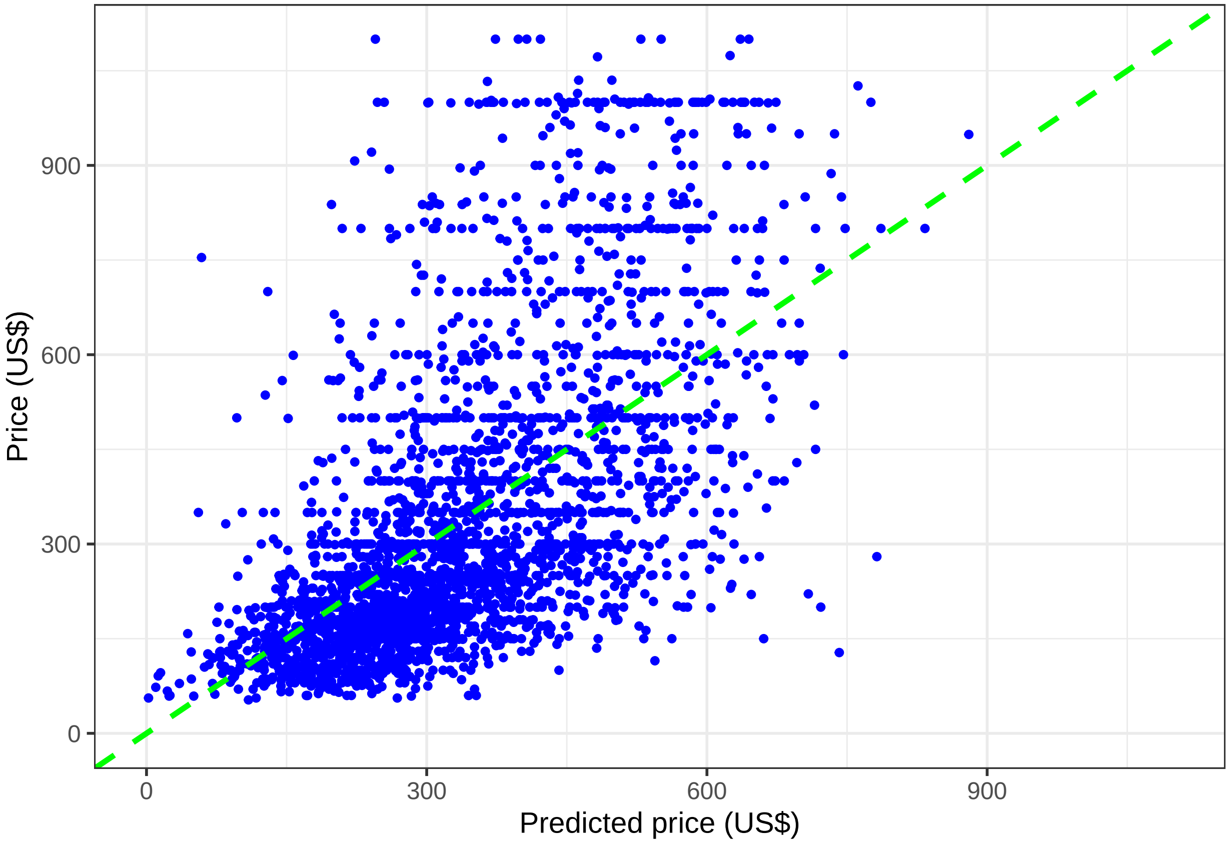
Even the best performing theory based models are outperformed by the data-driven alternatives. However, even these models produce really high prediction errors. Below you can see the expected average errors which are really high considering the avg. apartment price in the dataset is 340.

```
## character(0)
```

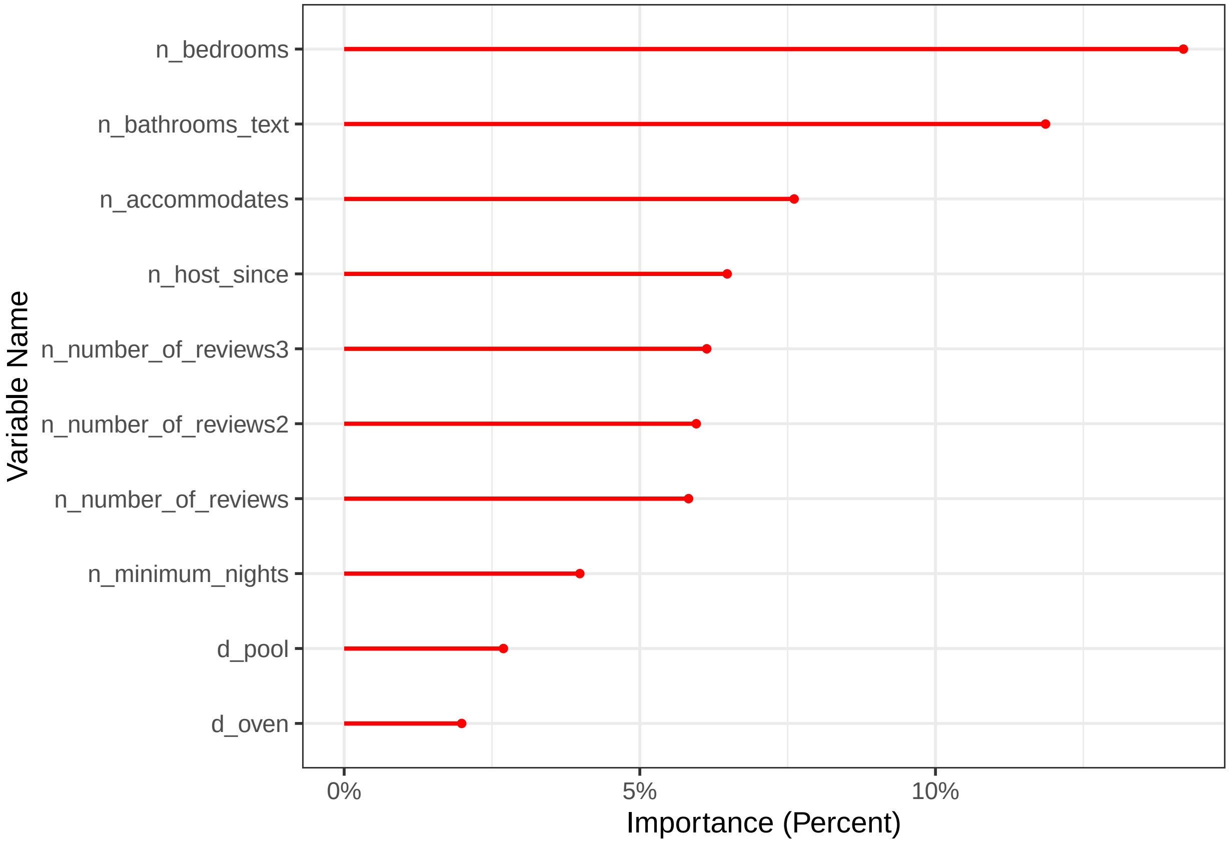
```
## character(0)
```

	RMSE on hold-out sample
OLS	193.4
LASSO	186.5
Random Forest	179.8

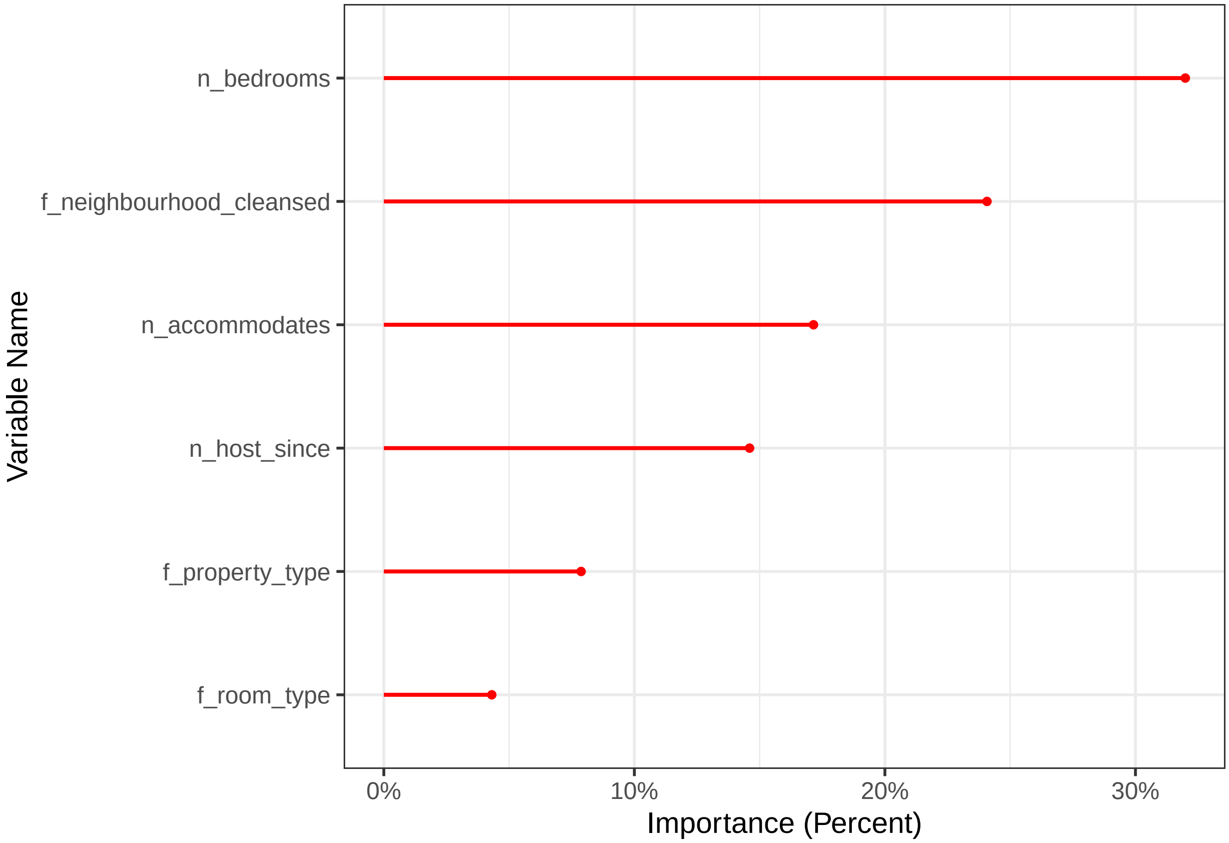
To illustrate, see the predictions of the LASSO model compared to actual prices. Especially above 500$ the model is heavily under-predicting in most cases but also overestimates often.



These data-driven models also offer less insight into the driving factors of their outcome than the OLS models. Below are two graphs trying to give some idea about the variables considered to be important by the Random Forest model. First, you can see the 10 most important variables in generating the overall predictions. The main predictors are similar to OLS, but reviews and certain items seem to be more important than the location.



If we group together categories and numeric values with ranges, we see that while individual districts may not be as important as a whole they are significant.



Conclusion, Further Analysis

In conclusion, model errors are too significant at this stage to provide a reliable solution. The main sources of the errors are:

- Limited data: consider adding similar units to our models, for example lofts or other units that are similar to apartments.
- Limited variables: The Rio dataset didn't include all relevant information such as cancellation policy of the host. Some potentially relevant features had no values or many missing observations. Nevertheless feature choices have to be further investigated.

Additional techniques may be explored. However, it is also worth noting that at the time of the recording of the data (2021/Nov), Brazil was between two major waves of the COVID pandemic. Especially since 2022 January the Omicron variant caused a major upswing in case numbers, thus any findings from this period has to be cautiously applied for predicting lockdown periods.