Data Science Summary of Price Prediction Modeling for Airbnb Apartments in Berlin, Germany

Ian Brandenburg (2304791)

GitHub Repository Link

1. Introduction

The objective of this project was to develop price prediction models for small to mid-size apartments hosting 2-6 guests in Berlin. These are a relatively popular option provided by Airbnb, so there is no shortage of data in this targeted range of accommodations. For this research report, three models were compared to determine which would be best used as a price prediction model. The three models that were compared were the OLS, Random Forest, and CART models.

2. Data Overview

The data was sourced from Inside Airbnb, a site that has collected publicly available listings information from the Airbnb website. This data was collected for the purpose of identifying the impact Airbnb has on a city. The data has been analysed, cleansed, and aggregated to create a space for researchers to discuss potential findings.

In the data cleansing process, the data was filtered to only reflect apartments that accommodate two-to-six guests. Hotels were filtered out from the analysis to shift the focus on apartments entirely. Additionally, the upper and lower quartiles of the prices were filtered out to avoid extreme values skewing the results.

3. Predictor Variables

Three sets of predictor variables were developed to incrementally test the addition of new variables, with each predictor becoming more complex. The data was also split into training and testing sets, at 70% in the training size.

3.1 Basic Variables

- **n_accommodates**: The number of guests a listing can accommodate.
- **n_beds**: The number of beds available.
- **f_property_type**: The type of property (e.g., apartment, house).
- **f_room_type**: The type of room offered (e.g., entire home, private room).
- **d_host_is_superhost**: Whether the host is classified as a "Superhost."
- **n_availability_365**: How many days a year the listing is available.
- **n_maximum_nights** and **n_minimum_nights**: Restrictions on bookings.

3.2 Review Variables

- **n_number_of_reviews**: The total number of reviews.
- **n_review_scores_rating**: The average review score.

3.3 Amenities Variables

Dynamic list creation captures the presence or absence of various amenities (e.g., WiFi, kitchen). Only the top 15 most frequent amenities were analyzed.

3.4 Room Booking Types

Variables beginning with d_ likely indicate different booking policies or room types not covered by f_room_type, providing additional granularity on the listing's offering.

3.5 Interaction Terms

- **X1**: Interactions between property/room types and accommodations.
- **X2**: Interactions involving host status and amenities with other features.
- **X3**: More complex interactions that aim to see deeper insights into how combinations of features, like property type with review scores or specific combinations of amenities, might be associated with pricing.

3.6 Predictor Sets

- **predictors_1**: A baseline set focusing on essential listing characteristics.
- **predictors_2**: An expanded set that adds reviews and amenities to the basic variable.
- **predictors_E**: The most complex set, including all basic, review, and amenity variables, plus complex interaction terms.

4. Modelling Approach

The three models used for predictive modelling were Ordinary Least Squares (OLS), Random Forest, and CART (Classification and Regression Trees). Each of these had a specific reason for being conducted and will be detailed here. Each of the three sets of predictors were ran through each model to determine which set of predictors would be best.

4.1 OLS Model

The OLS model is an important model in developing a stronger understand of the association between variables. Along with its interpretability and simplicity, this model was selected for comparison. As this is a simple model in nature, it acts as an excellent baseline for comparing other models.

4.2 Random Forest Model

The Random Forest model was used to handle non-linearity and interaction variables. This model allows for a more complex analysis of explanatory variables on a dependent variable. Additionally, the Random Forest model can provide insights into which variables are most significantly associated with the dependent variable, which is price of the apartments here.

4.3 CART Model

The CART model is a simpler model that allows for tree visualizations, which allow for easier interpretations. Additionally, the CART model can capture non-linear relationships in a flexible manner. This is helpful for this project, as there is a large array of that are very complex.

5. Results and Comparison

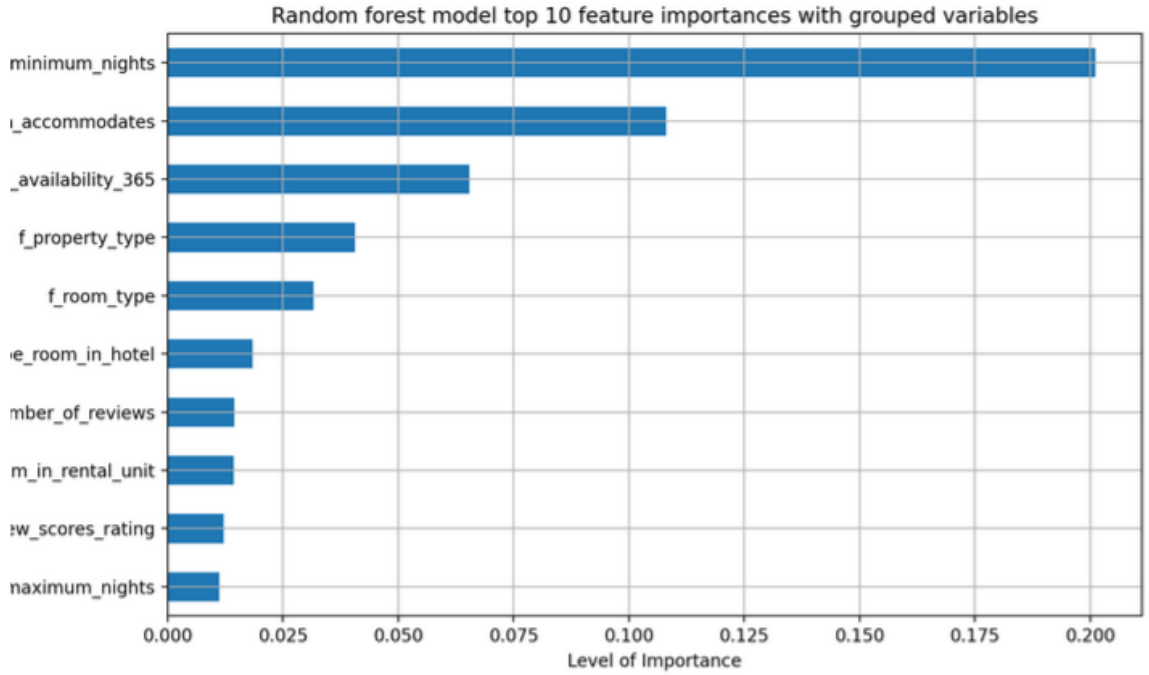
The three models were tested using the three sets of predictors, and their RSMEs can be visualized in the following table.

	Predictor Set	CART RMSE	OLS RMSE	RF RMSE
0	Pred 1	0.429047	0.415297	0.392397
1	Pred 2	0.503589	0.402598	0.381808
2	Pred E	0.476523	0.400506	0.383661

Predictor E performs the best, mostly because of the large number of variables. Nevertheless, the Random Forest has the highest prediction accuracy. Since the difference between predictor E and predictor 2 is miniscule, the ideal option would be to select the simpler model, which is predictor 2.

6. Conclusion and Recommendations

With an RSME, the best model for predicting apartment prices in Berlin is the Random Forest model. The variables from predictor set two is the preferrable model, as the RSME between the most complex Random Forest model, predictor E, is nearly the same. The simpler model, predictor 2, is preferred. Of predictor 2, the most important variables in predicting apartment prices, in order of most important first, are: minimum nights stay, availability, accommodation capacity, room type, and property type.



These variables should be analysed when considering prices for apartments that accommodate 2-6 people in Berlin. Further research should be conducted on external variables influencing prices, such as proximity to popular locations and seasonality.