**Airbnb Price Prediction**

Team 1: Veronika Junková, Filip Rott, Lucie Pinterová, Daniel Borner

Course: Data X

# 1 Data Understanding

**Listings**

8366 lines and 82 rows

There are in total 7,82 % missing cells in Listings dataset.

Numeric variables: latitude, longitude, accommodates, bathrooms, bedrooms, beds, minimum\_nights, maximum\_nights, minimum\_minimum\_nights, maximum\_minimum\_nights, minimum\_maximum\_nights, maximum\_maximum\_nights, minimum\_nights\_avg\_ntm, maximum\_nights\_avg\_ntm, calendar\_updated, host\_id, host\_response\_time, host\_acceptance\_rate, host\_total\_listings\_count

Categorical variables: neighbourhood\_cleansed, property\_type, room\_type, host\_is\_superhost, host\_neighbourhood, host\_verifications, host\_has\_profile\_pic, host\_identity\_verified

Unstructured variables: name, description, picture\_url, host\_location, host\_about, amenities

Datetime variables: host\_since

# 2 Data Preparation

**Added Columns:**

We created some other columns, which we thought would be useful for the modelling.

Obsah obrázku Vykreslený graf, snímek obrazovky, diagram, text

Popis byl vytvořen automatickyamenities\_count – after parsing the amenities column we added the column amenities\_count, which represents how many amenities the Airbnb has

count\_verifications – instead of having one column (list of verifications), we decided to have a number of how many verifications host has in another column

distance\_from\_city\_centre – as we all know, when booking an apartment, we all consider how far is the Airbnb from the city centre, so we took the location of Old Town Square as a city centre and calculated thanks to longitude and longitude of the apartment distance in kilometres

Obsah obrázku text, snímek obrazovky, Vykreslený graf, diagram

Popis byl vytvořen automaticky

season, seasonal\_availability – thanks to availability\_30, availability\_60, availability\_90, availability\_365 we added a column which represents the seasonal availability

minimal\_rating, maximum\_rating

**Modified Columns:**

bathrooms\_text: text descriptions of bathrooms were extracted into numeric values for easier analysis and aggregation (2 baths into 2)

has\_availability, host\_has\_profile\_pic, host\_identity\_verified, host\_is\_superhost, instant\_bookable: boolean values were converted to 0 and 1, which allows simpler logical operations and integration into mathematical models

host\_acceptance\_rate, host\_response\_rate: Percentage values were extracted from the text format and converted to a numeric format

price: We removed the currency symbols and commas, converted it into a numeric format

Obsah obrázku text, snímek obrazovky, displej, software

Popis byl vytvořen automatickyneighbourhood\_cleansed: we unified the neighbourhoods

**Removed Columns:**

scrape\_id, calendar\_last\_scraped, last\_scraped: columns about data collection, not useful for our analytical models

latitude, longitude, neighbourhood, neighbourhood\_group\_cleansed: we decided to use the information from these columns into new columns neighbourhood\_cleansed, distance\_from\_city\_centre, which are more useful for modelling, so we decided to remove the additional columns

description, neighbourhood\_overview, amenities: these textual columns were not useful for other modelling, thus they were removed

availability\_30, availability\_60, availability\_90, availability\_365: instead of using these columns, column season and seasonal\_availability replaced them

Obsah obrázku text, snímek obrazovky, diagram, Paralelní

Popis byl vytvořen automatickynumber\_of\_reviews, number\_of\_reviews\_ltm, review\_scores\_rating – since we decided to just use minimal\_rating and maximum\_rating and don’t do a sentiment analysis these columns were dropped

minimum\_minimum\_nights, maximum\_minimum\_nights, minimum\_maximum\_nights, maximum\_maximum\_nights, minimum\_nights\_avg\_ntm, maximum\_nights\_avg\_ntm – we thought that enough information is included in minimum\_nights and maximum\_nights so the others were removed

# 3 Data Visualization

Here we see that most of the Airbnbs are Entire homes/apartments.

Obsah obrázku text, snímek obrazovky, diagram, Vykreslený graf

Popis byl vytvořen automaticky

Obsah obrázku text, snímek obrazovky, Vykreslený graf, diagram

Popis byl vytvořen automatickyObsah obrázku text, snímek obrazovky, displej, software

Popis byl vytvořen automatickyMost of the being in the city centre in Prague1, Prague 2 or Prague 3.

Most of the Airbnb is only available between 40 % and 80 % of the time. Most of the Airbnbs are medium-low available during the time.

Obsah obrázku text, diagram, Vykreslený graf, snímek obrazovky

Popis byl vytvořen automaticky

Obsah obrázku text, snímek obrazovky, diagram, Obdélník

Popis byl vytvořen automaticky

The number of amenities listed.

Obsah obrázku Vykreslený graf, snímek obrazovky, diagram, text

Popis byl vytvořen automaticky

Obsah obrázku text, snímek obrazovky, diagram, řada/pruh

Popis byl vytvořen automatickyThe median price is very low compared to the maximum, most prices are in the lower range. The outliers may affect the modelling.

# 4 Modelling

**Model Selection**

In our analysis, we evaluated various predictive models to identify the one best suited for our dataset. Linear Regression, despite being a fundamental approach, yielded very poor performance with an R2 of only 3%, indicating a weak explanatory power for our data's variability. The Random Forest model performed significantly better, achieving a 34% R2 score, making it the second-best model in our trials. However, the CatBoost model outperformed all others with an initial R2 of 56%. Through the application of Recursive Feature Elimination (RFE), we were able to enhance its performance significantly, boosting the R2 to 89%. This substantial improvement highlights CatBoost's capability in handling categorical features and its robustness against overfitting.

**Feature Selection**

The model performed better as more features were removed

Obsah obrázku řada/pruh, diagram, Vykreslený graf, snímek obrazovky

Popis byl vytvořen automaticky

**Model limitations and considerations**

While it provides superior accuracy, its training time is considerably longer, which might be a constraint.