esade

Best Practices for Hosts @ airbnb

AI II Group Project

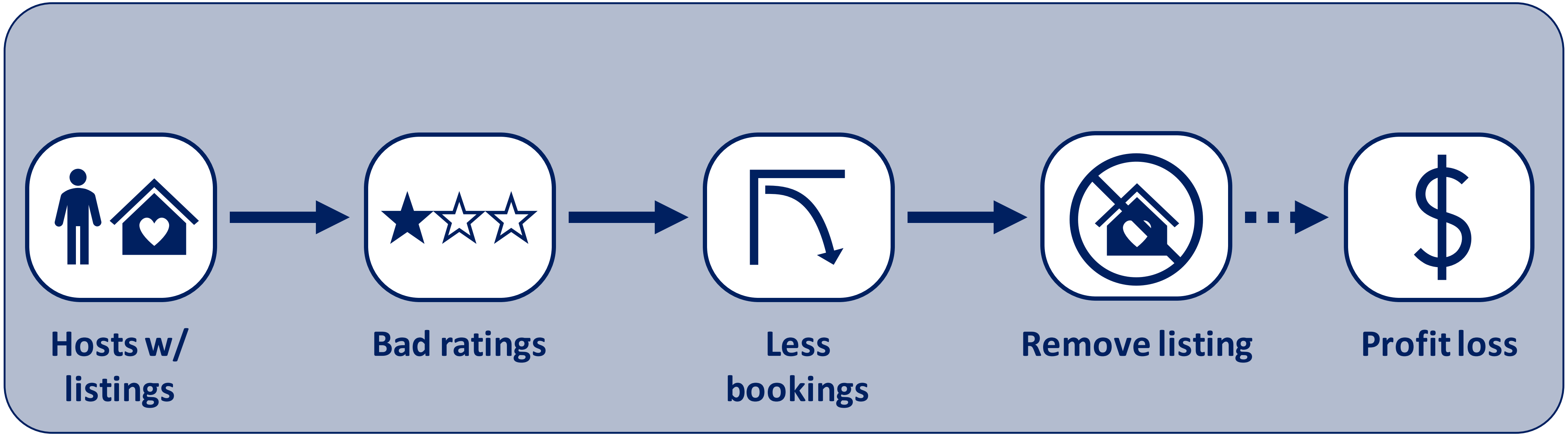
Team B2: Thibault, Tsaichi, Carles, Shatha, Miron

Problem Description

Goal definition | Data preparation | Model creation | Model interpretation | Model implementation

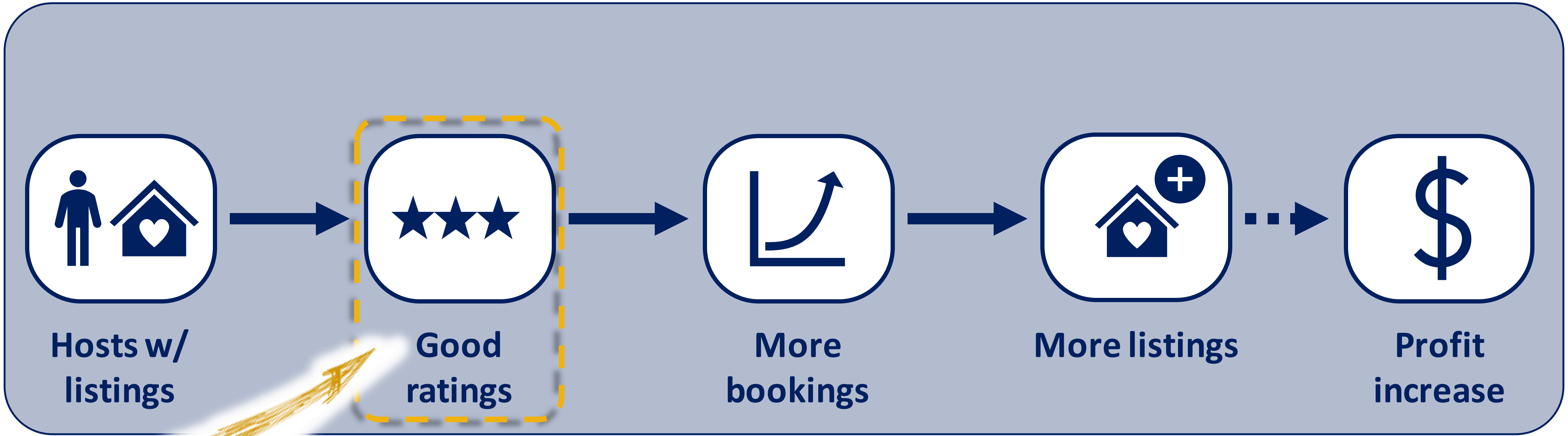
Bad ratings frustrate hosts and leads to a reduction of listings which leads to a profit loss for Airbnb

Problem



Sharing best practices with host will improve ratings and lead to profit increase for Airbnb

Motivation



- identify best practices on how to receive good ratings
- share best practices with hosts to improve their platform experience

The target review_score_rating will be predicted with a classification algorithm

Approach

review_score_rating > 4.85[1)]

1) Threshold was chosen to approximately split data set 50-50 in false-positive class of the target variable

esade

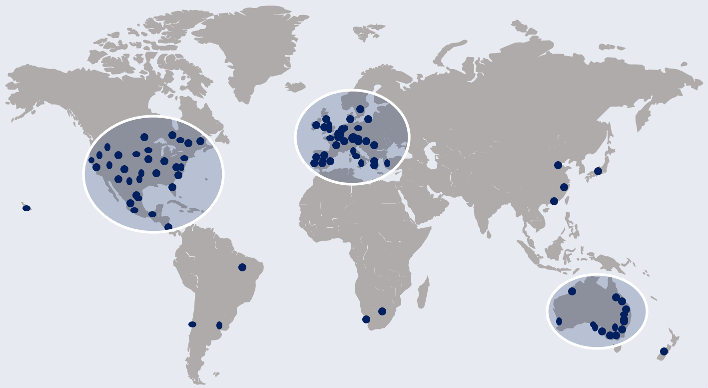
Data preparation

Goal definition → **Data preparation** → Model creation → Model interpretation → Model implementation

The acquired raw data set contains information on 82 features, for 1.2 m listings, across 108 cities

Raw data

108 Cities



1,171,739 instances

Out of which ~870k instances used in final analysis

82 features

Out of which 18 features used in final analysis

2 conditions for 21 selected features: 1) impact on rating is possible, 2) host's influence on feature is feasible

Selected Features

Property features (8)

Tangibles

- room_type
- acommodates
- bathrooms_text
- bedrooms
- beds
- amenities

Others

- price
- review_scores_location

Host features (7)

Profile

- host_about
- host_acceptance_rate
- host_verifications
- host_is_superhost

Service

- host_response_time
- host_response_rate
- review_scores_cleanliness

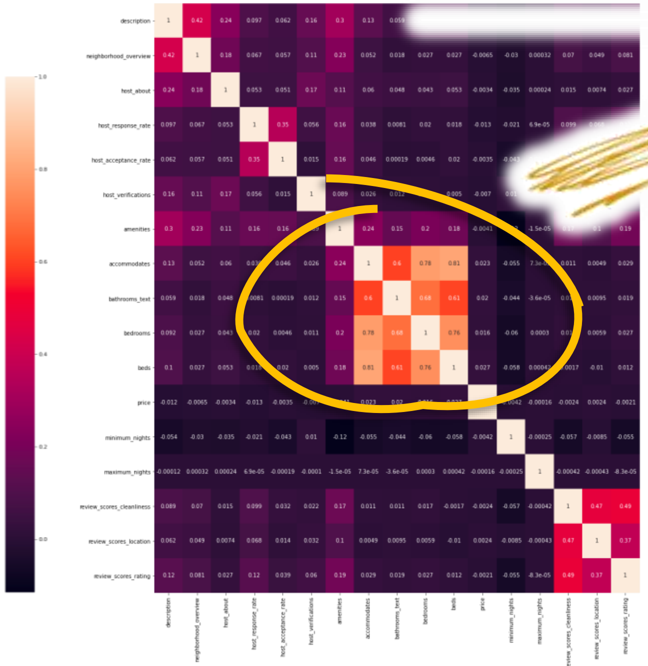
Presentation features (6)

- description
- neighborhood_overview
- minimum_nights
- maximum_nights
- instant_bookable
- host_has_profile_pic

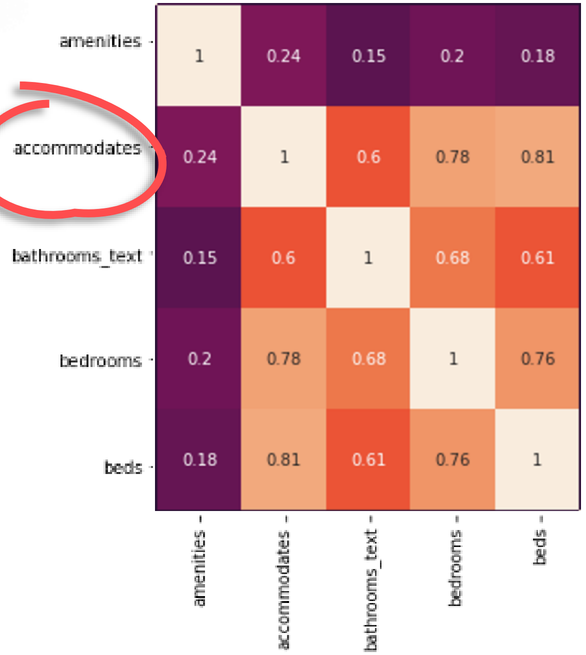
`Accomodates` was removed due to its high correlation with other features

Feature selection based on correlation

Correlation Matrix



High correlation

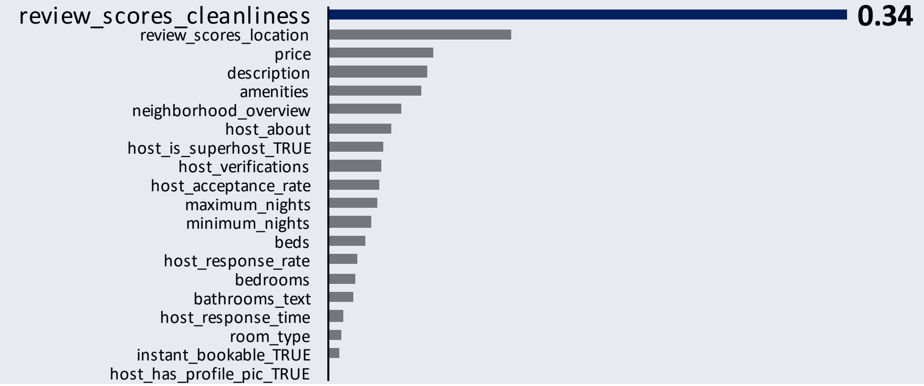


esade

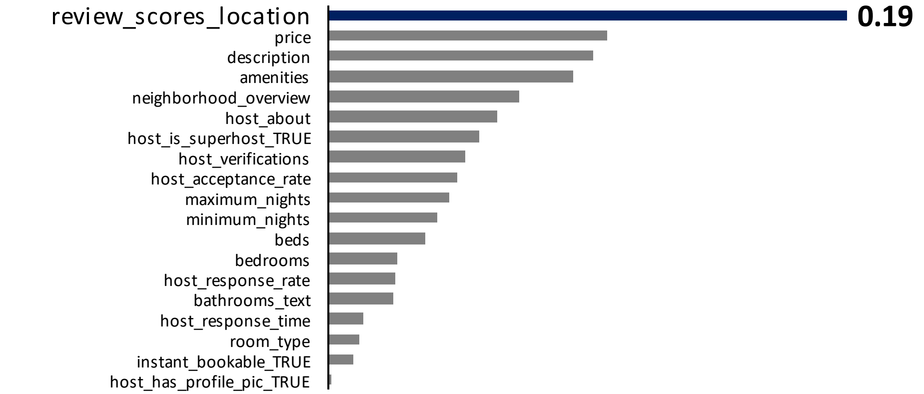
Cleanliness review & location review were removed due to their dominant impact on the prediciton

Feature selection based on feature importance

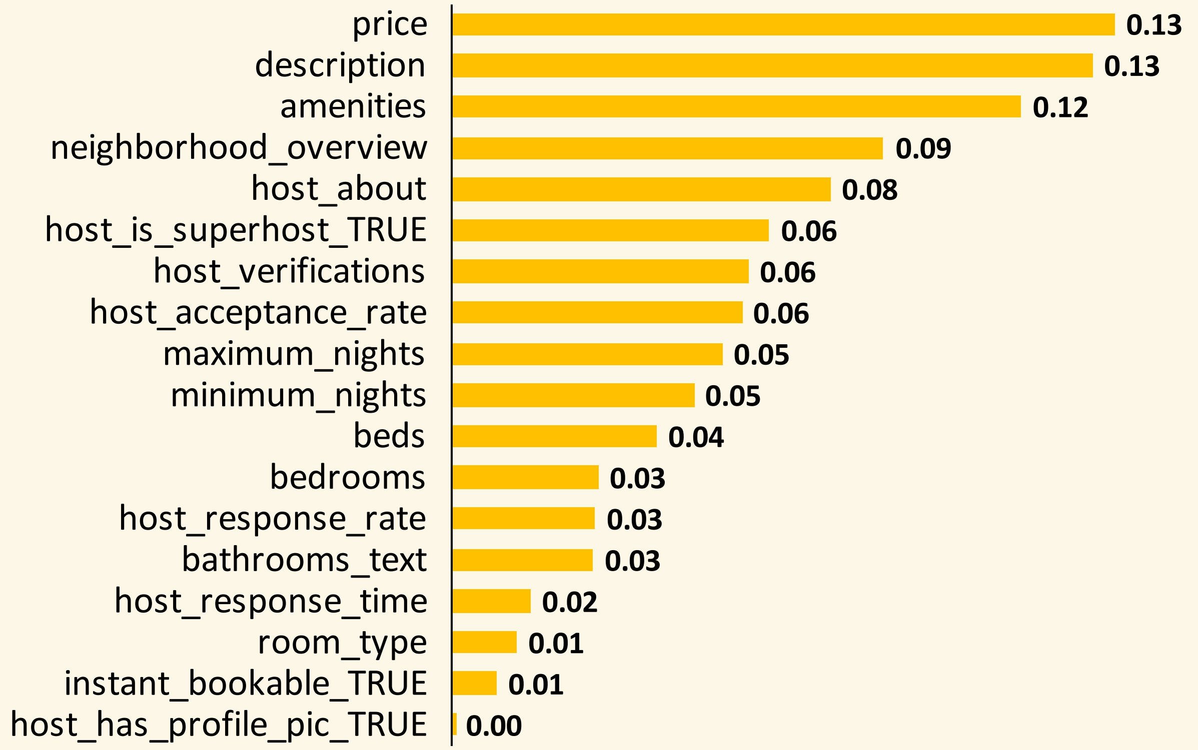
1. All selected features



2. Selected features excl. cleanliness review



3. Selected features excl. cleanliness review & location review



The final feature selection consists of 18 features, equally distributed across categories

Final feature overview

Property features (6)

Tangibles

- room_type
- bathrooms_text
- bedrooms
- beds
- amenities

Others

- price

Host features (6)

Profile

- host_about
- host_acceptance_rate
- host_verifications
- host_is_superhost

Service

- host_response_time
- host_response_rate

Presentation features (6)

- description
- neighborhood_overview
- minimum_nights
- maximum_nights
- instant_bookable
- host_has_profile_pic

The new data set makes up 74% of the raw data – N/As were eliminated or imputed by appropriate estimates

Treatment of NA values

Eliminated instances (281,564)

Target (23.8% N/As)
- review_scores_rating

Only 0.34% of raw data
- host_has_profile_pic
- bathrooms_text

Imputed by x

x = 0
- bedrooms

x = mean
- host_response_rate
- host_acceptance_rate
- beds

x = mode (most frequent)
- host_response_time
- host_is_superhost

Text features were converted to numeric values to become processable by the algorithm

Feature engineering

Text descriptions & lists

- description
- neighborhood_overview
- host_about
- amenities
- host_verifications

→ Converted to length in words (numeric)

"Enjoy your stay at our 4 person apartment in…" ⟫ 124

Text numerical

- bathrooms_text
- price

→ Converted to numeric

"2.5 shared baths" ⟫ 2.5

Percentages

- host_response_rate
- host_acceptance_rate

→ Converted to float

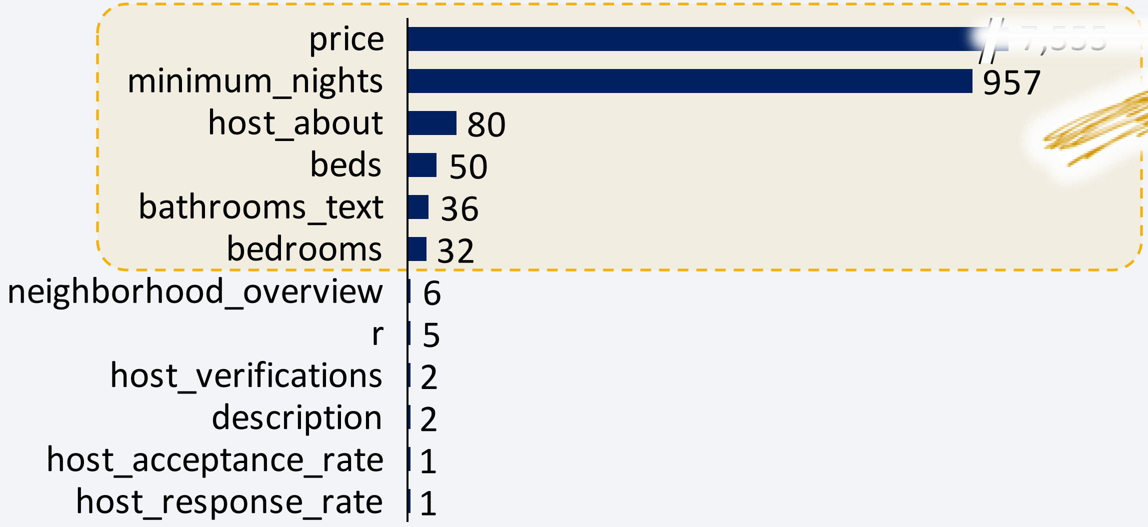
"75%" ⟫ 0.75

19 k instances where eliminated in the outlier removal process – Final data set has ~870k instances

Outliers

Features with suspicious instances

Maximum value as multiple of mean



Elimination strategy

The **highest 0.1 percentile** was removed

(18,938 instances)

→ **871,237 instances in the final data set**

As a final step, categorical features were encoded and the data set was standardized

Encodation & standardization

Encodation

OHE

- Host_is_superhost
- Host_has_profile_pic
- Instant_bookable

Ordinal

- Room_type
- Host_response_time

Standardization

StandardScaler()

Model creation & evaluation

Goal definition → Data preparation → **Model creation** → Model interpretation → Model implementation

The random forest outperforms all other models – However it comes at the cost of interpretability

Model Overview[1)]

Rank	Model	Interpretability	Performance (Accuracy)
1.	**Random forest**	◔	65.3%
2.	**Decision tree**	●	58.6%
3.	**Logistic regression**	◑	63.5%

Cross-validation applied

With Grid Search and Cross Validation, the best performing model achieves an accuracy of 65%

Random Forest

Grid Search Parameters

Criterion	• gini • entropy
Max_features	• sqrt • Log2 • None
Max_depth	• 10 • 15 • 20
Min_sample_splits	• 10 • 50 • 20 • 100

Best performing parameter

Evaluation

65.26% Accuracy

65.44% Precision

62.81% Recall

F1 **64.10%** F1 Score

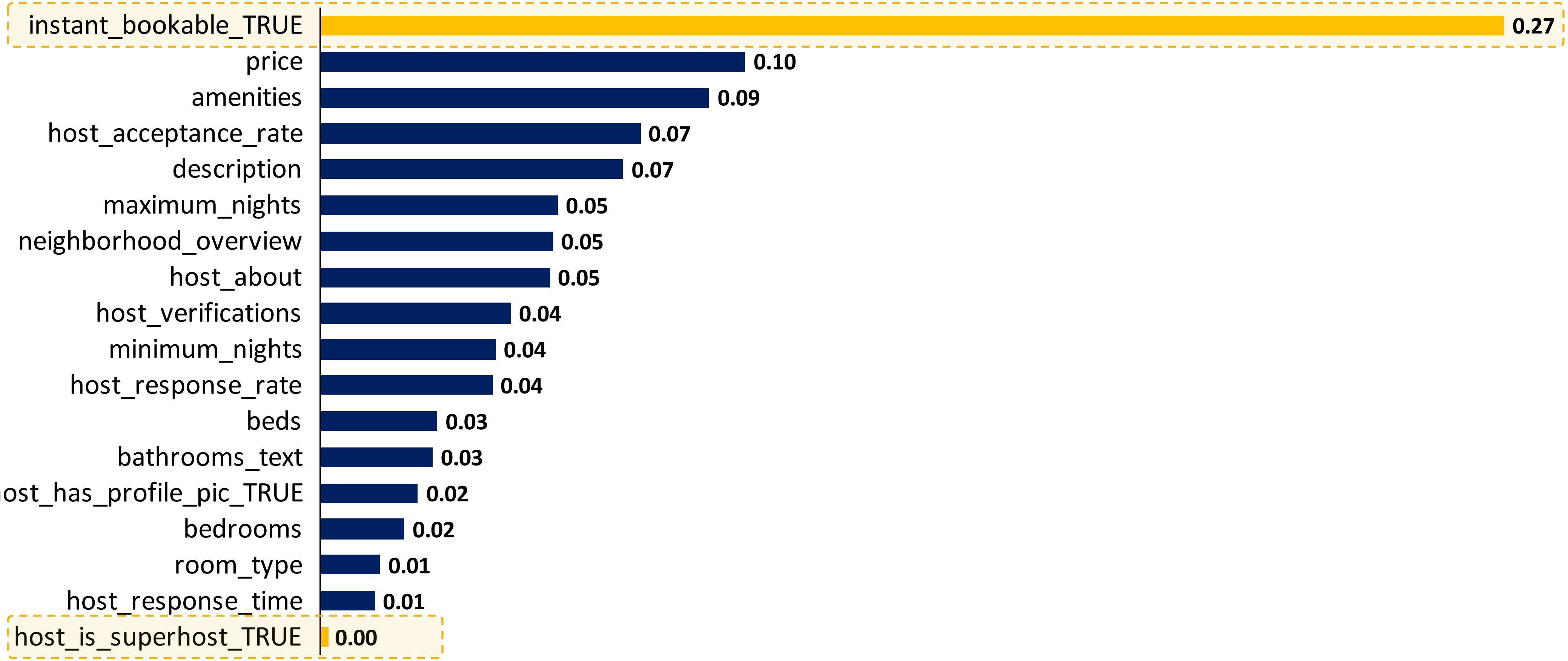
Cross-validation applied

Model interpretation

Goal definition → Data preparation → Model creation → **Model interpretation** → Model implementation

Surprising results of the feature importance analysis: instant_bookable leads, is_superhost has low importance

Feature Importance



Each feature can be translated into an implication for Airbnb hosts

Implications of top 10 features

Feature	Weight	Implication
instant_bookable_TRUE	0.27	Instant bookable feature provides more convenience and ease to customers and thus lead to higher satisfaction rate and higher overall ratings
price	0.10	Customers are extremely sensitive to prices, in which their overall score ratings hinges on whether they perceive a listing to be a good value for money or not
amenities	0.09	The number of amenities has a positive correlation with the overall score rating of a listing, the more amenities the listing has the better the overall rating gets.
host_acceptance_rate	0.07	Guests tend to like properties which’s hosts have a high acceptance rate. Most likely this is a corelation without major implication for hosts.
description	0.07	The description of a listing influences how customers perceive the overall quality of a listing; hence the more elaborate the description is the higher the overall score rating gets
maximum_nights	0.05	It is inconvenient for customers to relocate from a listing when the max nights available don’t match their total stay duration and thus higher max nights lead to more satisfied customers.
neighborhood_overview	0.05	Neighbourhood has a huge influence on customers overall experience whilst renting a property, and thus impacts the overall score ratings.
host_about	0.05	About host feature gives a better credibility to hosts and make their profiles more legitimate, the older the description of the profile is the better the overall score rating
host_verifications	0.04	Customer would feel safer and more relieved if an Airbnb host has a verified identity and thus it would enhance their overall experience and overall score rating
minimum_nights	0.04	Flexible stay duration in terms of minimum nights provide more convenience to short-stay customers and thus has a negative correlation with the overall score rating

Model implementation

Goal definition → Data preparation → Model creation → Model interpretation → **Model implementation**

esade

With the model we could identify action steps for each of the 3 parts of a host's user journey at Airbnb

Best practices handbook

1. Property selection

Know the area! **Choose good locations**.

A **fair price** is one of the most important aspects of a good listing. Make sure to buy properties for which you can offer fair prices.

Amenities enhance guests satisfaction! Make sure to equip the prop. with sufficient ammenities.

2. Platform presentation

Liked properties often have the **option for instant booking activated**. Make sure to follow their lead!

Properties with **high ratings tend to have better descriptions of property and neighbourhood**. Put in the effort to make it shine!

Guests value host flexibility. Try to make your listing as accessible as possible by **elongating the maximum nights a guest can stay.**

3. Guest service

Make sure guests arrive at a clean home!

Guests value reliability! Try to attend **to every guest or potential guest as quickly** as possible

In further steps, provide more details on identified levers and test causality

Steps for further analysis

More details

- **Descriptions**: Whats the optimal word count?
- **Location**: What defines a good location (e.g. center proximity, access to transport etc.)
- **Price**: How to approximate a fair price?
- **Amenities**: What are amenities that bring value?

Causality

ML only shows corelation!

Make sure best practices actually cause better ratings though experiments

(e.g. A-B Testing)

Best Practices for Hosts @