✻✻✻✻✻

AIRBNB CUSTOMER SATISFACTION

Group 11

Rachana Aithal
Vanesa Alcantara
Ambarish Narayan

Medha Nalamada
Kapish Krishna
Ayaan Khan

Table of Contents

The Context - Airbnb

Founded in 2008, Airbnb is an online platform that facilitates short-term lodging and rentals

We aim to use machine learning models to predict guest satisfaction for Airbnbs in Europe.

Understanding guest satisfaction can inform many avenues of business decisions.

Problem Statement

Predicting guest satisfaction targeted towards helping two primary stakeholders: Airbnb and Hosts

ADVERTISING CAMPAIGN FOR AIRBNB

New campaign to advertise offers with high guest satisfaction ratings

METRIC INFORMATION FOR POTENTIAL HOSTS

Important areas for hosts to consider before offering property to ensure success

IMPROVEMENT SUGGESTIONS FOR CURRENT HOSTS

Focusing on improving certain features to maintain satisfied customer base

Data

Location Related	Hosting related	Ratings
City	Price	Cleanliness Rating
City Centre	Day	Attraction Index
Metro Distance	Person Capacity	Restaurant Index
	Superhost	
	Bedrooms	
	Room Type	
	Multiple Rooms	

Exploratory Data Analysis

Cleaning the Data

Variables

1

- Initial variables include room type and one variable for each.
- Airbnbs with 0 bedrooms (Studios)

Target categories

2

- Some categories have low materiality
- Split the target in five categories from low(20) to high(100) satisfaction.

0	1	2	3	4
Ex low	Low	Middle	High	Ex high

Normalizing

3

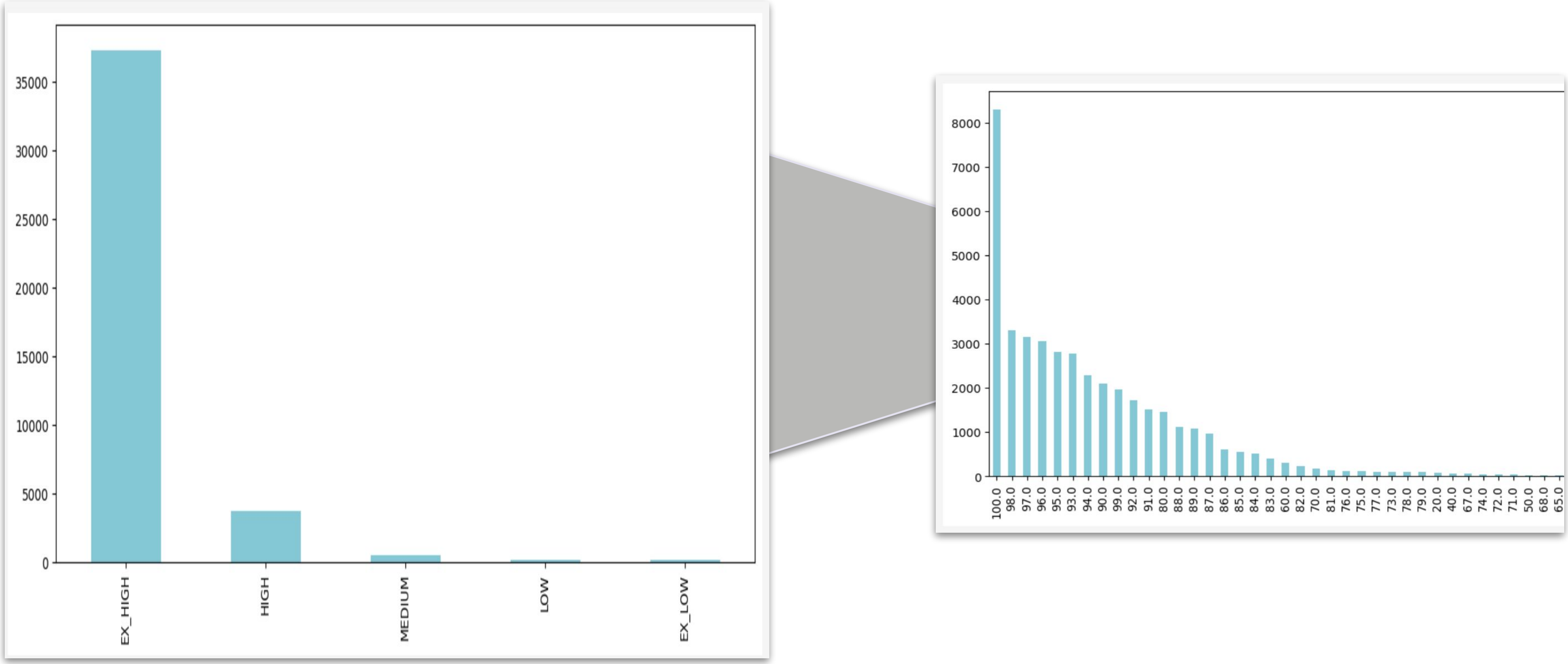
For Linear and Logistic regression and KNN

Categorical transformation

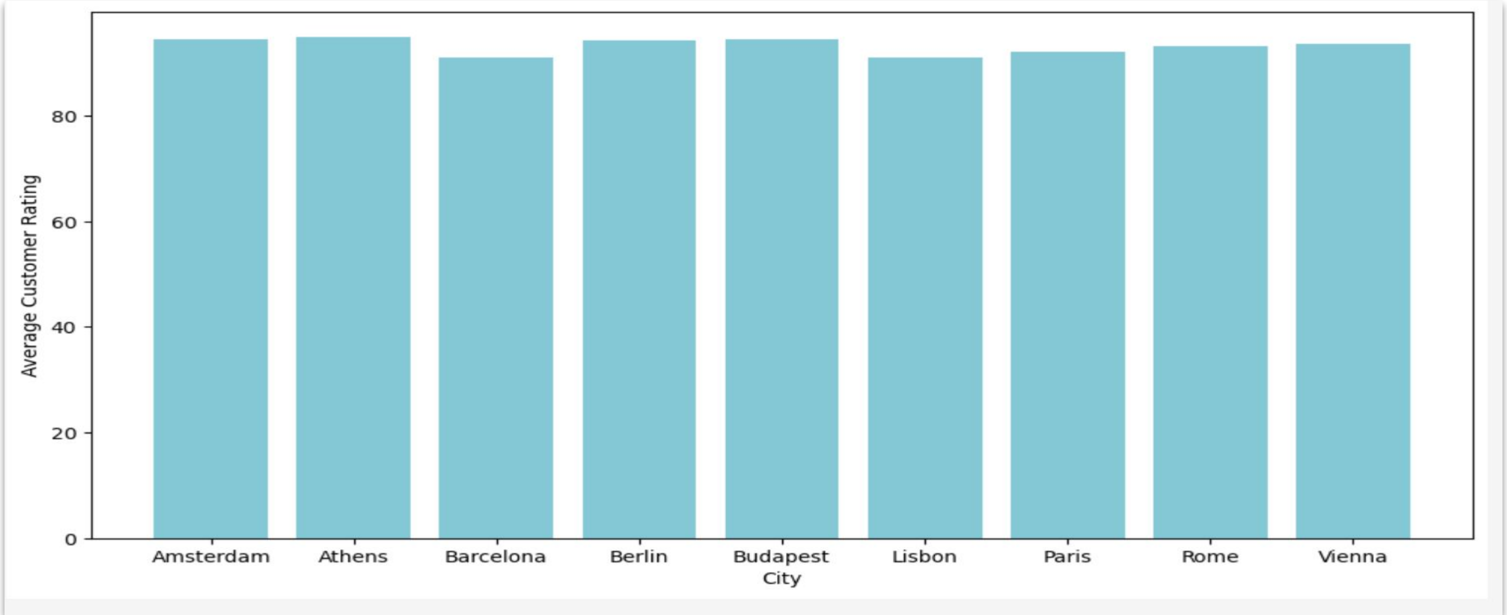
4

-Flags for binary variables
-Median or flags for multiclasses

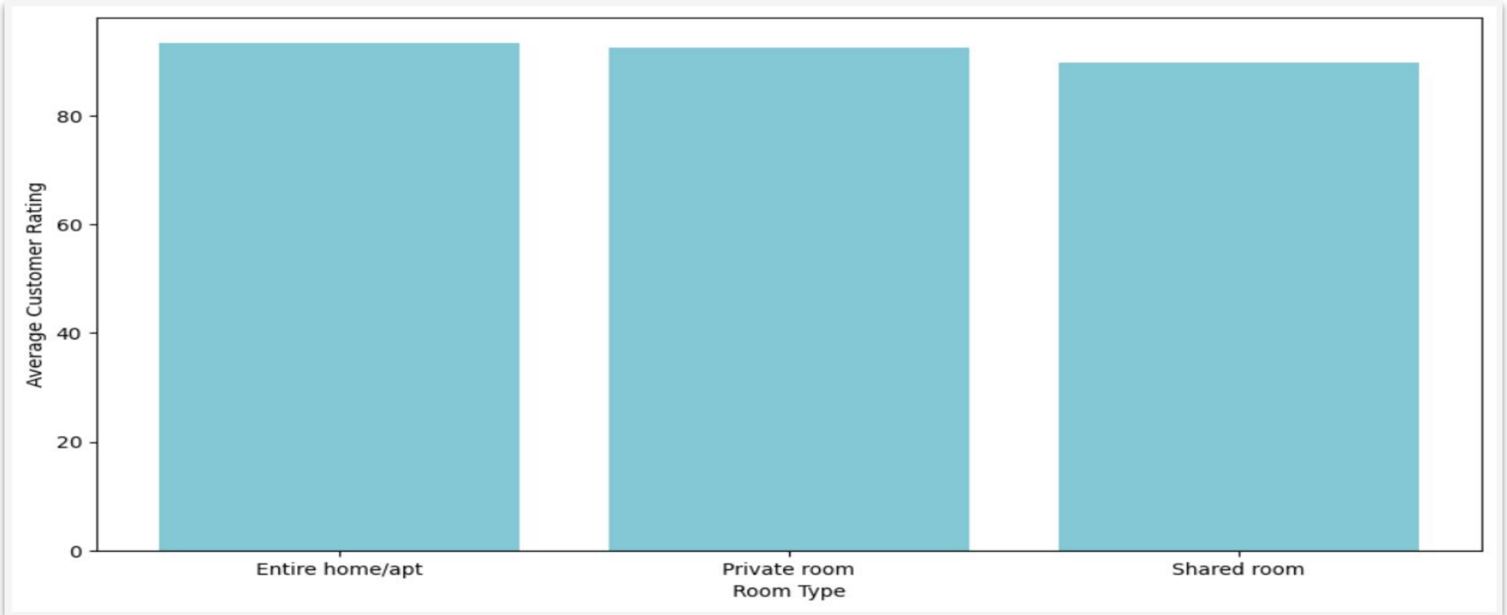
Most of the properties have a very high customer satisfaction rating



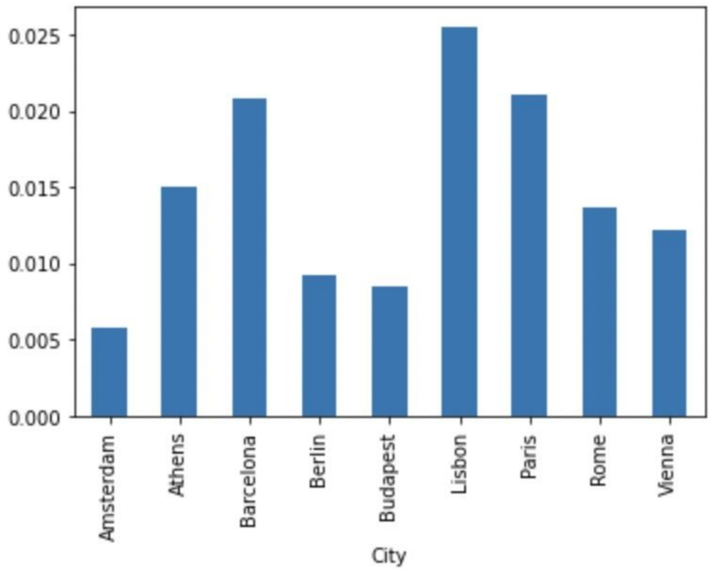
Among the cities Lisbon has the lowest guest satisfaction while Athens has the highest



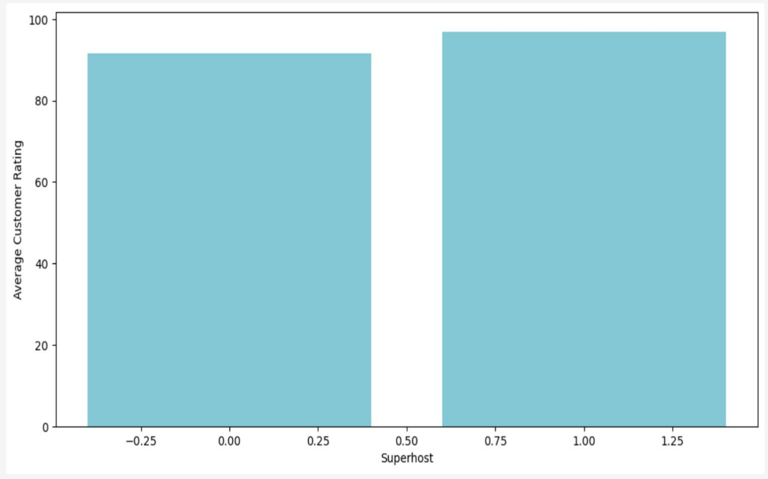
Shared rooms have lower average rating as compared to private

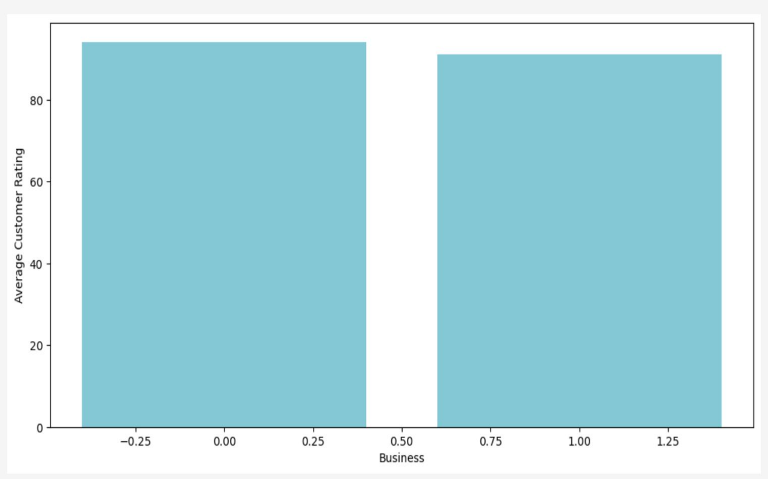


Lisboa, Paris and Barcelona have higher % of low guest satisfaction

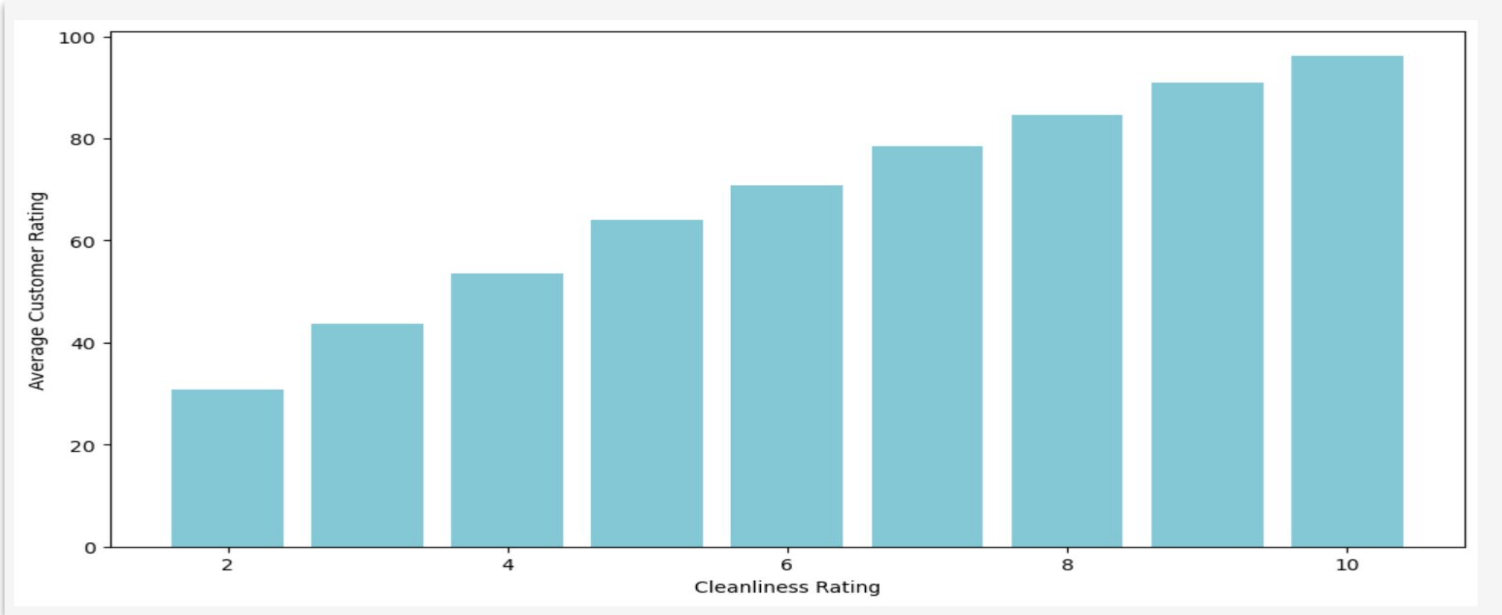


Properties of “superhost” owners have a higher guest satisfaction rating; while owners having multiple properties have a lower rating

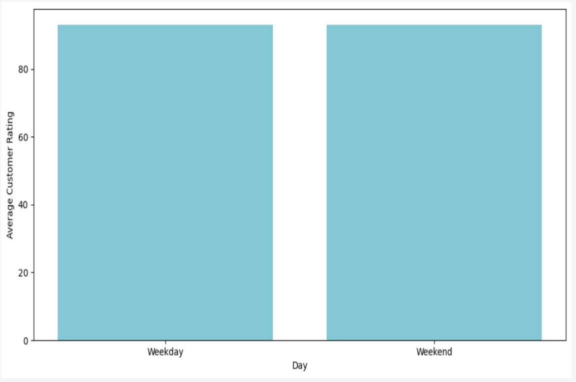


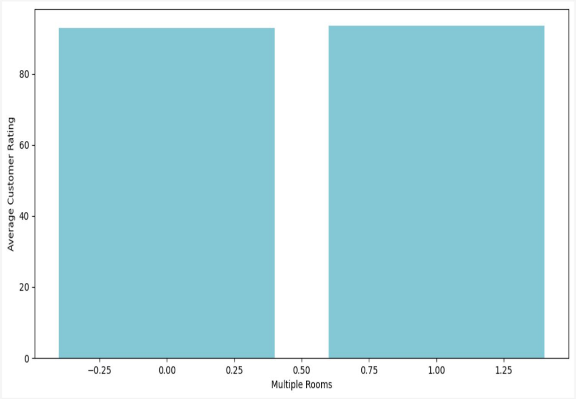


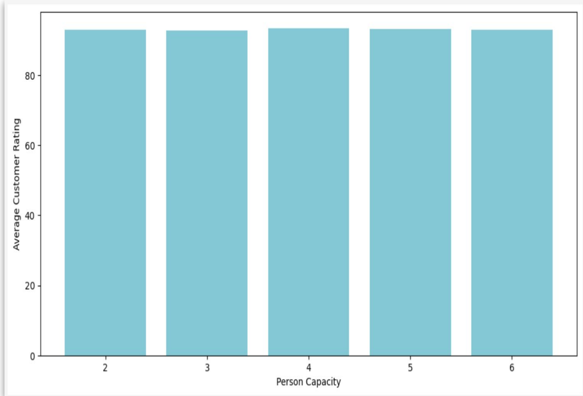
Higher cleanliness rating directly transforms to a higher customer rating



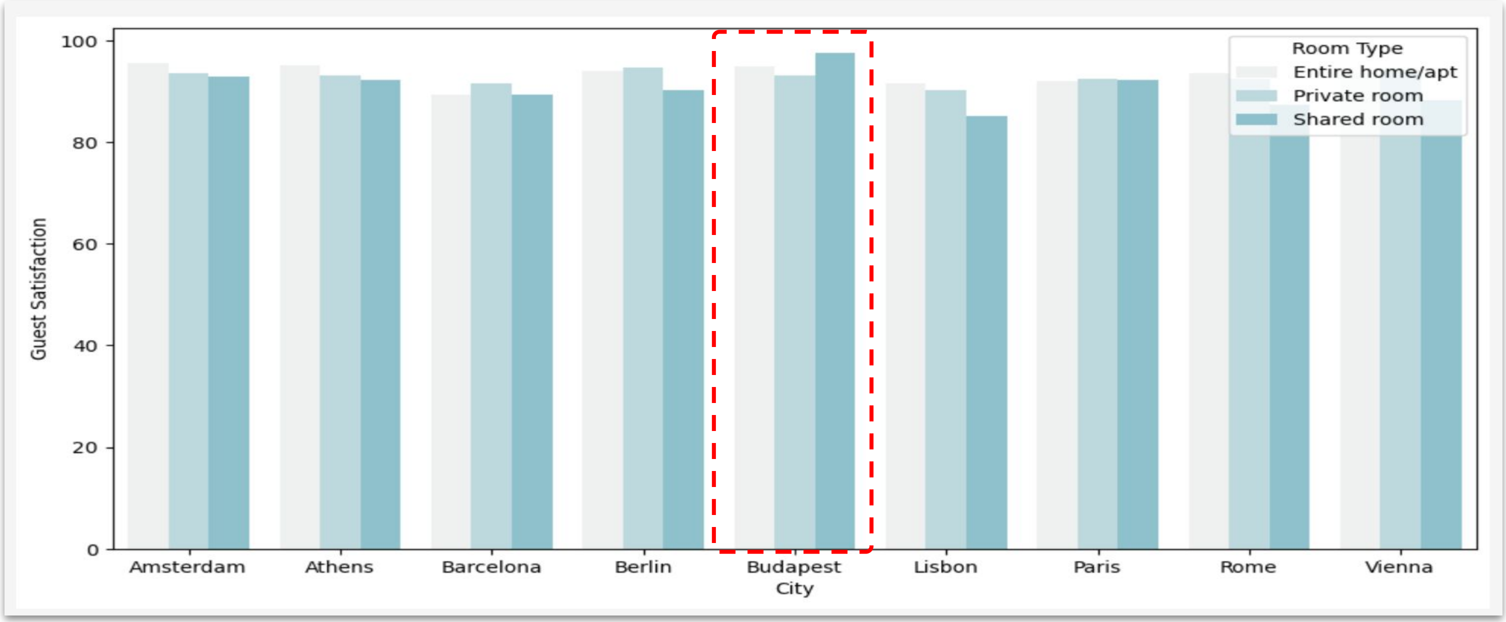
Customer Satisfaction is independent of day of the week during the stay, person capacity as well as presence of multiple room in the property



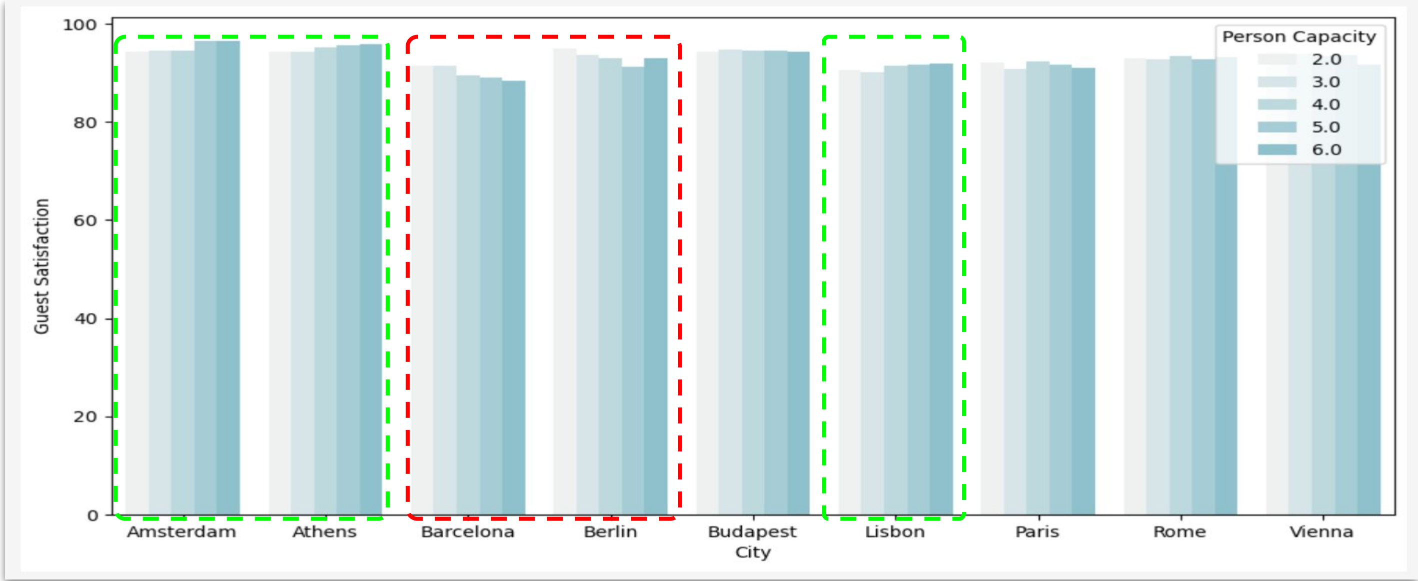




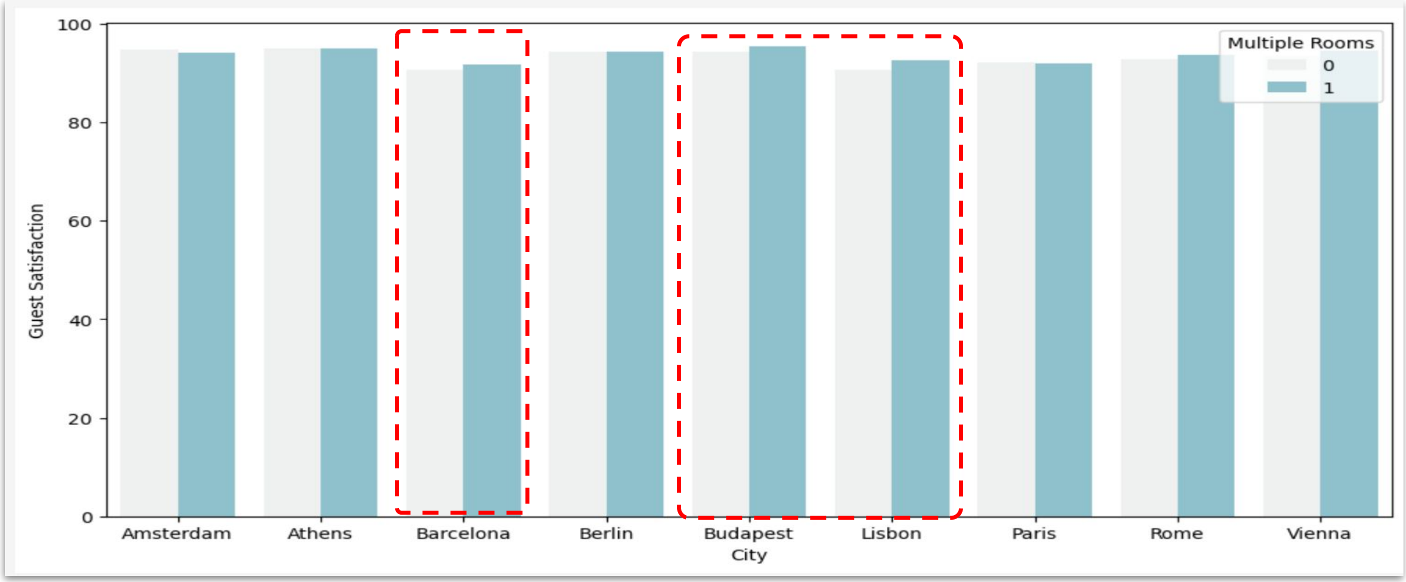
In contrast to other cities, shared properties in Budapest have higher guest satisfaction



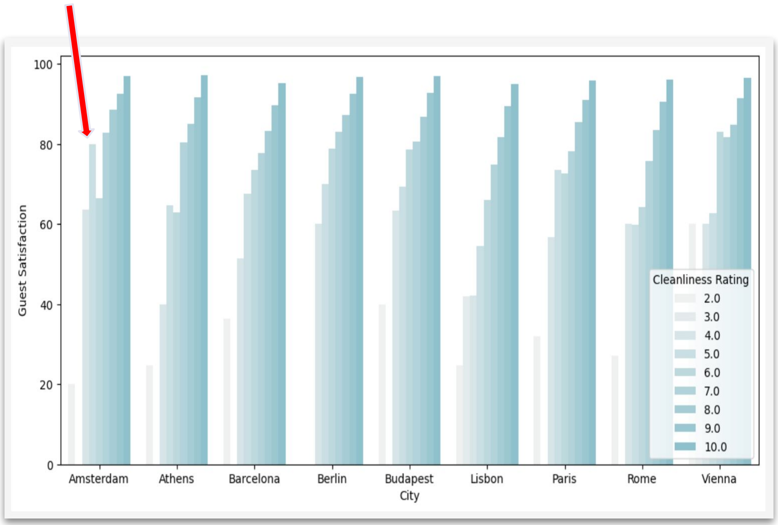
While people in Barcelona and Berlin prefer smaller capacity, those in Athens, Amsterdam and Lisbon have slight inclination towards higher capacity rooms

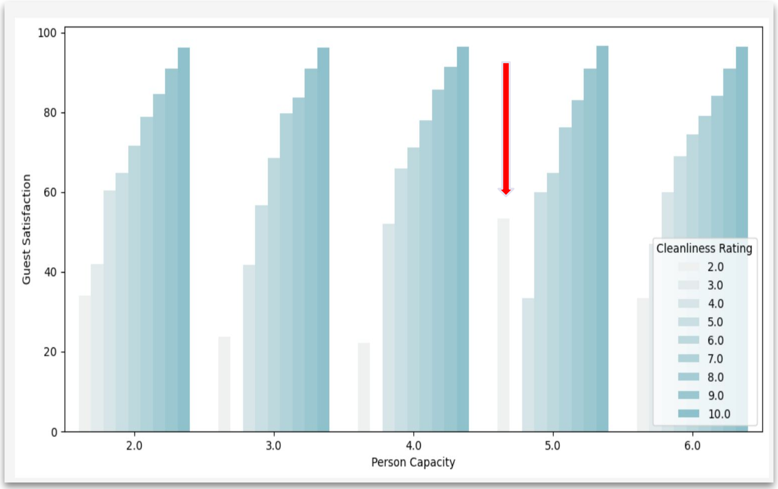


Customers in Budapest, Lisbon and Barcelona prefer multiple rooms

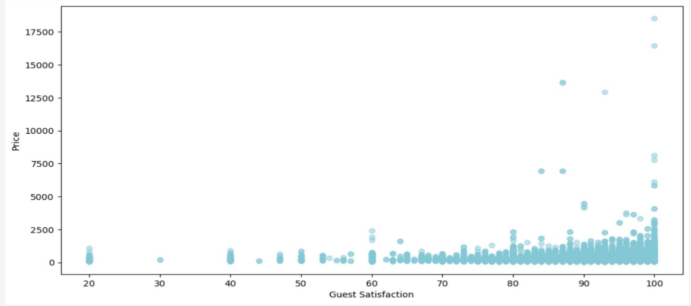


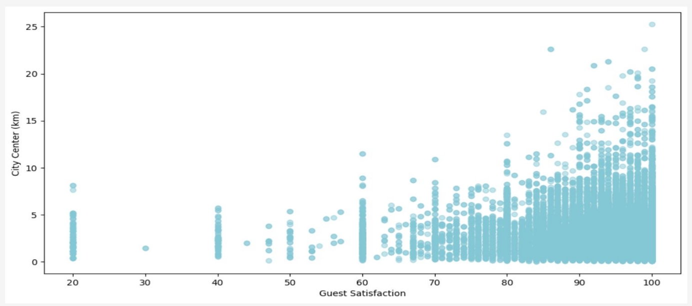
Although cleanliness directly affects customer rating, there are some exceptions

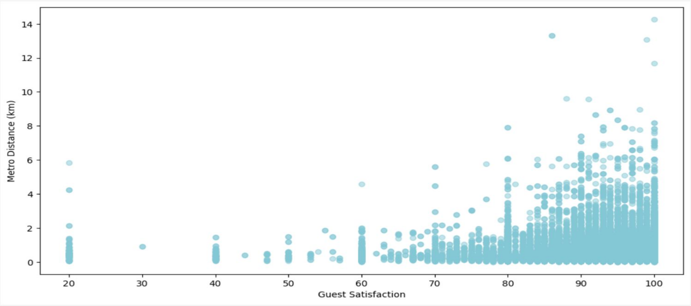


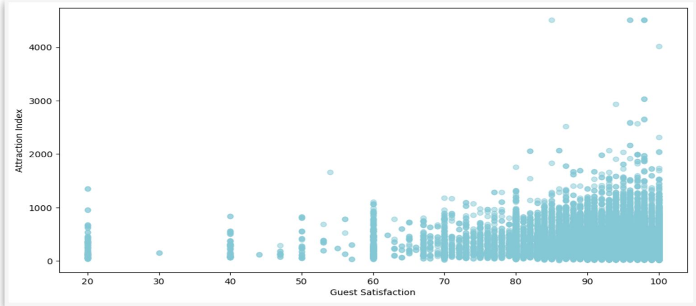


Parameters such as price, distance from nearest metro/city center do not have strong relationship with customer rating

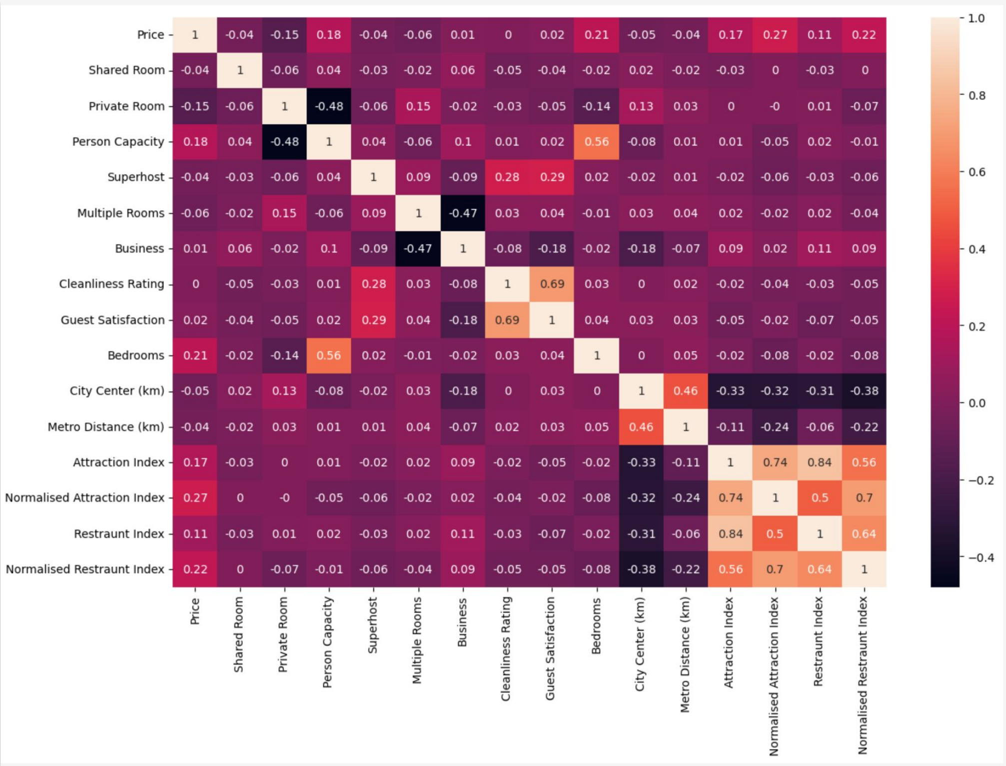








Correlation Matrix



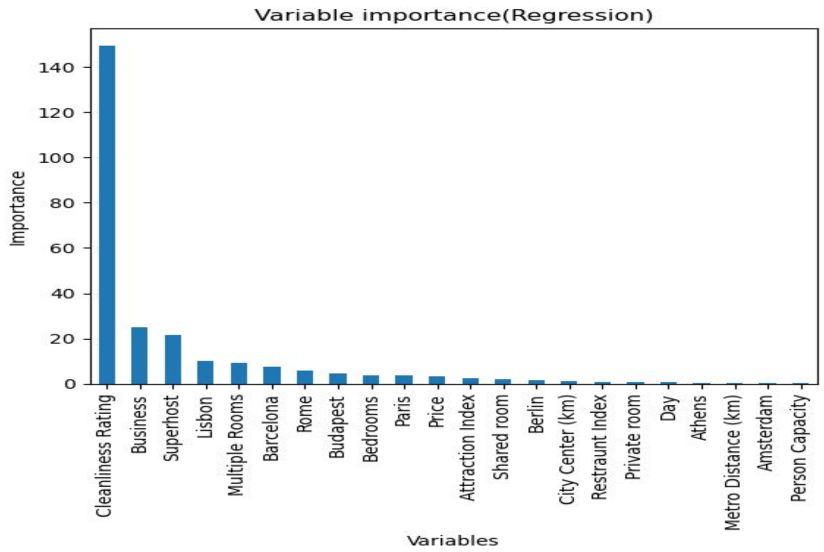
4. Models

0	Ex low
1	Low
2	Middle
3	High
4	Ex high

Regression

Logistic Regression, Ridge, Lasso and Linear regression are all very interpretable
Best Accuracy: 53%

Feature importance



Performance per class

	precision	recall	f1-score	support
0	0.00	0.00	0.00	18
1	0.18	0.17	0.17	18
2	0.37	0.31	0.34	114
3	0.15	0.86	0.25	770
4	0.98	0.48	0.65	7423
accuracy			0.51	8343
macro avg	0.33	0.36	0.28	8343
weighted avg	0.89	0.51	0.60	8343

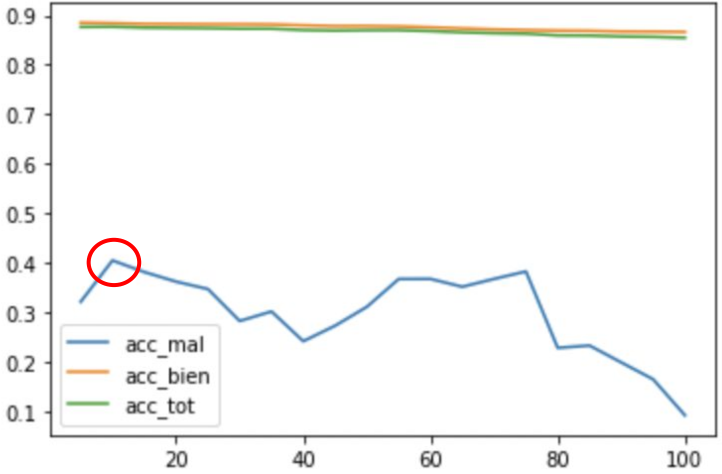
Confusion Matrix

	Extremely Low	Low	Medium	High	Extremely High
Extremely Low	0	9	7	1	1
Low	0	3	13	2	0
Medium	0	5	35	64	10
High	0	0	36	661	73
Extremely High	0	0	3	3830	3590

KNN

Non parametric tool . The best model have precision from 36% to 92% with 10 nearest neighbors

Accuracy



Performance per class

	0.EX_LOW	1.LOW	2.MEDIUM	3.HIGH	4.EX_HIGH
precision	47.62%	60.00%	36.00%	48.99%	92.45%
recall	50.00%	10.34%	8.41%	23.32%	98.48%
f1-score	48.78%	17.65%	13.64%	31.60%	95.37%

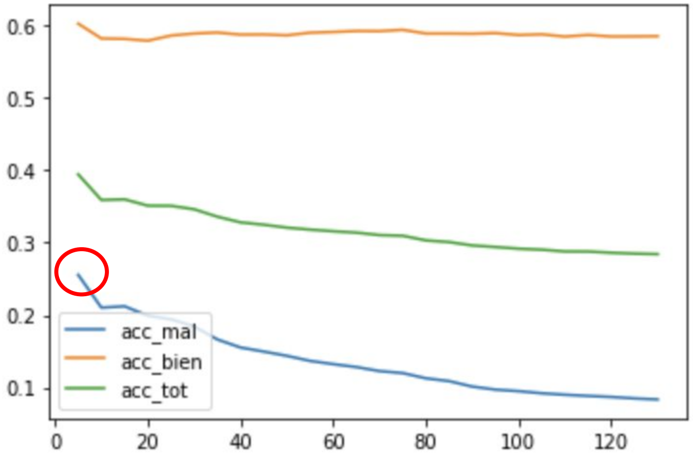
Confusion Matrix

	0.EX_LOW	1.LOW	2.MEDIUM	3.HIGH	4.EX_HIGH
0.EX_LOW	10	0	6	2	2
1.LOW	8	3	4	7	7
2.MEDIUM	3	1	9	55	39
3.HIGH	0	1	6	170	552
4.EX_HIGH	0	0	0	113	7345

KNN

Balancing the model by oversampling didn't improve the results in KNN

Accuracy



Performance per class

	0.EX_LOW	1.LOW	2.MEDIUM	3.HIGH	4.EX_HIGH
precision	40.91%	19.23%	16.54%	25.60%	94.89%
recall	45.00%	17.24%	19.63%	51.03%	85.44%
f1-score	42.86%	18.18%	17.95%	34.10%	89.92%

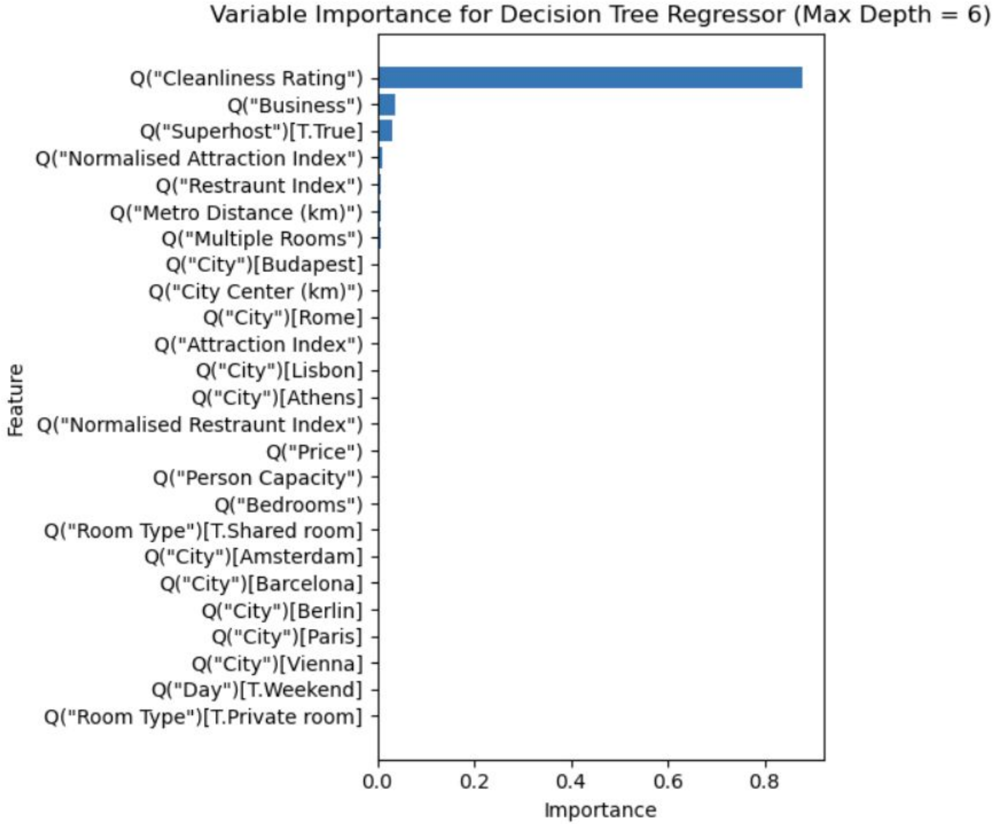
Confusion Matrix

	0.EX_LOW	1.LOW	2.MEDIUM	3.HIGH	4.EX_HIGH
0.EX_LOW	9	2	5	3	1
1.LOW	6	5	2	11	5
2.MEDIUM	5	5	21	57	19
3.HIGH	0	6	33	375	315
4.EX_HIGH	2	8	67	1013	6368

Decision Tree Regressor

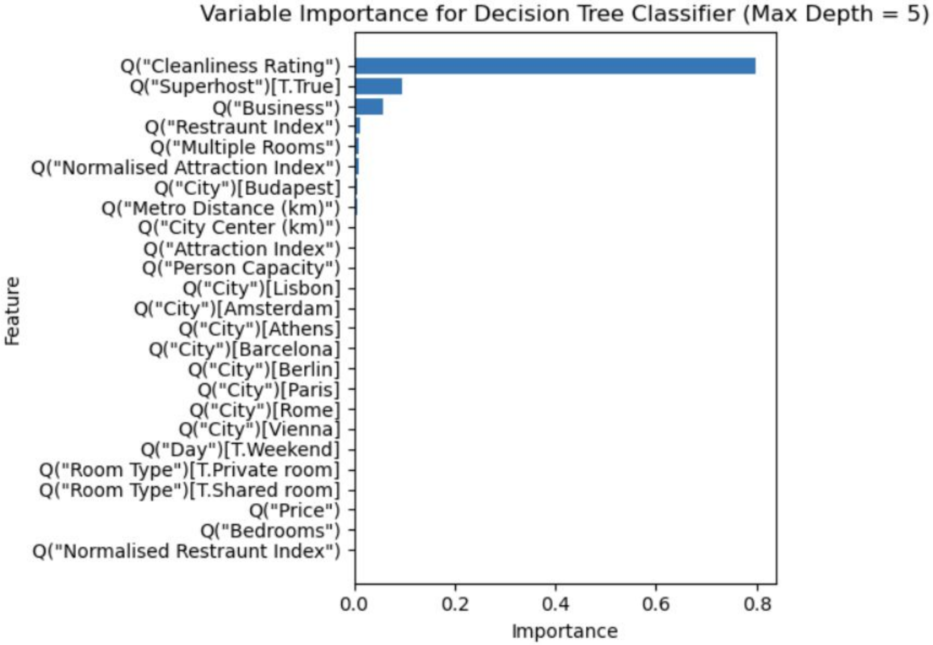
- Deep decision tree -
  max_depth = 46
  training mse = 0.0
  test mse = 28.92

- CV (k = 5)
  depth = 6
  training mse = 29.01
  test mse = 31.31



Decision Tree Classifier

- Deep decision tree - max_depth = 35 → 100% training accuracy. But, 93% test accuracy.
- CV depth = 5 with k = 5, training accuracy = 91.19% test accuracy = 91.21%



<u>Precision per class:</u>

	0.EX_LOW	1.LOW	2.MEDIUM	3.HIGH	4.EX_HIGH
precision	85.71%	100.00%	31.58%	56.69%	93.63%
recall	75.00%	13.64%	14.46%	36.13%	97.88%
f1-score	80.00%	24.00%	19.83%	44.14%	95.71%

<u>Confusion Matrix:</u>

Prediction

Actual

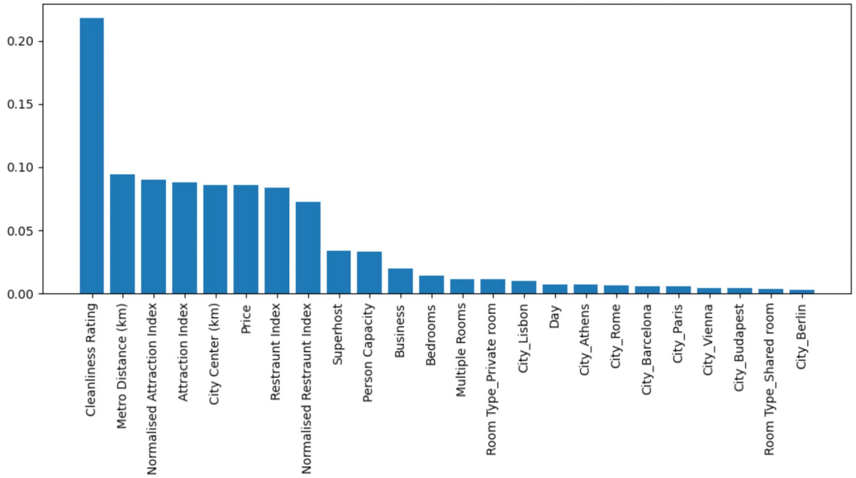
	0.EX_LOW	1.LOW	2.MEDIUM	3.HIGH	4.EX_HIGH
0.EX_LOW	18	0	4	1	1
1.LOW	2	3	7	8	2
2.MEDIUM	1	0	12	43	27
3.HIGH	0	0	12	271	467
4.EX_HIGH	0	0	3	155	7306

0	Ex low
1	Low
2	Middle
3	High
4	Ex high

Random Forest

The Random Forest model efficiently predicted Guest Satisfaction classes and identified significant predictors

Feature importance



Performance per class

```
Accuracy: 0.8929417961548591
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.90      0.81        21
           1       0.36      0.71      0.48        14
           2       0.46      0.41      0.44        78
           3       0.46      0.64      0.54       692
           4       0.96      0.92      0.94      6789

    accuracy                           0.89      7594
   macro avg       0.60      0.72      0.64      7594
weighted avg       0.91      0.89      0.90      7594
```

Hyperparameters

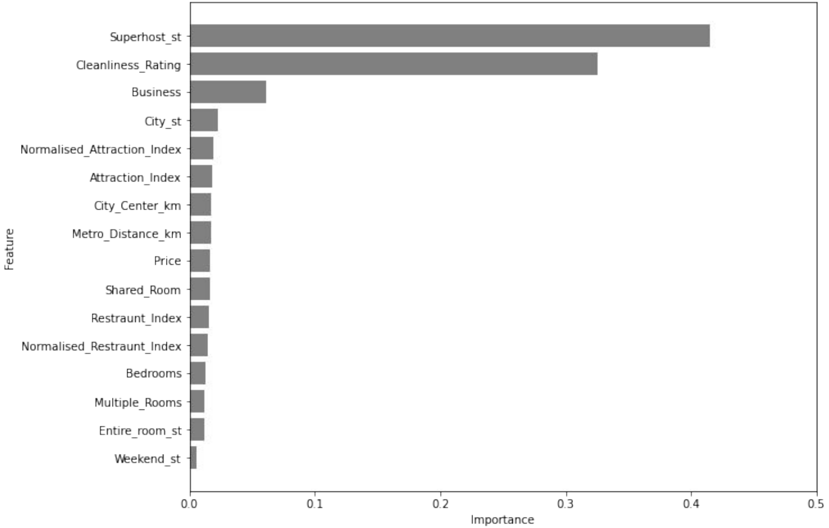
n trees	150
Max depth	None
Min obs leaf	10

Using ten fold cross validation

XGBoost

0	Ex low
1	Low
2	Middle
3	High
4	Ex high

Feature Importance



- Captures non-linear relationships
- High precision (**95%**) and recall (**95%**)
- Low interpretability of variables
- Increased number of parameters

Performance per Class

	precision	recall	f1-score	support
0	1.00	0.85	0.92	20
1	0.95	0.62	0.75	29
2	0.92	0.51	0.66	107
3	0.83	0.63	0.71	729
4	0.96	0.99	0.98	7458
accuracy			0.95	8343
macro avg	0.93	0.72	0.80	8343
weighted avg	0.95	0.95	0.95	8343

Hyper-Parameters

Regularization	**0.3**
Trees	**3000**
Subsample	**1.0**
Colsample	**0.3**
Max Depth	**12**

Model Summary

0	Ex low	Bad
1	Low	
2	Middle	
3	High	Good
4	Ex high	

	Accuracy	*Accuracy - good*	*Accuracy - bad*
Linear regression (bin)	51.4%	51.8%	24.2%
KNN	87.7%	88.57%	41.95%
Random Forest	89.2%	91.6%	48.1%
Trees	91.2%	91.5%	56.5%
XGBoost	94.9%	94.9%	93.3%

Proposals

The best model selected was XGBoost

Most Important Features
Superhost and Cleanliness Rating

Model Accuracy
94.9%

Airbnb:

Filter out offers on these two categories and create a new campaign to advertise them to increase customer satisfaction and revenue!

Current Hosts:

Focus more on keeping the house clean to ensure satisfaction!
Aim for more bookings and verification to become a Superhost.

Potential Hosts:

Gauge the probability of success early on and strategize/make improvements before listing!

Thank you!
Questions?