SMART SCAN

A MACHINE LEARNING BENCHMARKING OF AIRBNB LISTING

BT5153 Group Project - Group 9
Susan Koruthu | Felipe Chapa Chamorro | Widya Gani Salim | Georgius Gary Gunawan | Gino Martelli Tiu

BACKGROUND & MOTIVATION

BACKGROUND & MOTIVATION

COVID-19 has changed travel dynamics and Airbnb guest behaviors:

Remote work

Inflationary Pressures

28-30

Day Stays

Longer stays than before due to flexibility at work

+6.2%

Increase in Living Cost

High increase in living cost instills bargaining mentality

AIRBNB'S REVENUE MODEL

... is predominantly fee-based

- Shared
- Host Only

OPTIMAL PRICING

Pricing just right is important to maximize:

- Revenue
- Utilization

PROBLEM STATEMENT

The main focus is to empowering hosts to achieve a good balance between **maximizing property utilization** and **booking price**.

To achieve this, we will focus on providing **3 key insights**:

GET PRICE ESTIMATE

Given a listing's existing **meta-features**, hosts can get a price estimate.

IDENTIFY KEY FEATURES

Facilitates better understanding on which **features most influence prices**.

FIND COMPARABLE LISTING

Accommodates for **competitions** like hosts' own listing.

SOLUTION COMPONENTS

To acquire the key insights discussed, we implemented 3 models below.

NATURAL LANGUAGE PROCESSING (NLP)	REGRESSION MODEL	RECOMMENDER SYSTEM
Extract customer and common sentiment from listing reviews.	Output recommended pricing based on property features and customer sentiments.	Identify top accommodations that are closest to the host's property.
Emphasis: Capture and input sentiments in other models.	**Emphasis:** Uncover features which most influence prices.	**Emphasis:** Inform hosts of existing competitions.

Our proposed solution creates value for guests and hosts by enabling hosts **adjust prices** and **cater to guests' needs** better.

DATA SOURCING & EXPLORATION

DATA SOURCES

1. INSIDE AIRBNB LONDON DATASET

DETAILED LISTINGS FILE	USER REVIEWS FILE
66,641 Listings & **74** Features across	**1,043,004** reviews

Detailed Listings File:

- host
- amenities
- ratings
- location
- property

User Reviews File:

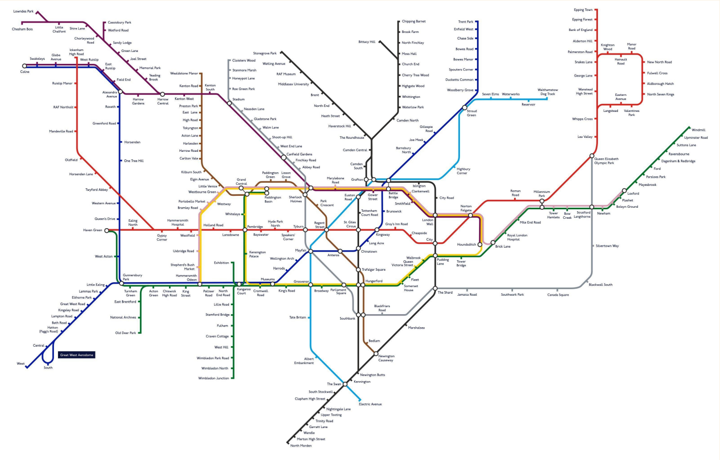
- date
- reviewer details

DATA SOURCES

2. GOOGLE API

Google Maps API was implemented to capture the location coordinates of all London tube stations. This information is then used to engineer location features.

Features extracted include **station name, latitude**, and **longitude** coordinates.



Station	Latitude	Longitude
Abbey Road	51.532	0.0037
Baker Street	51.523	-0.1569
…..	…..	…..

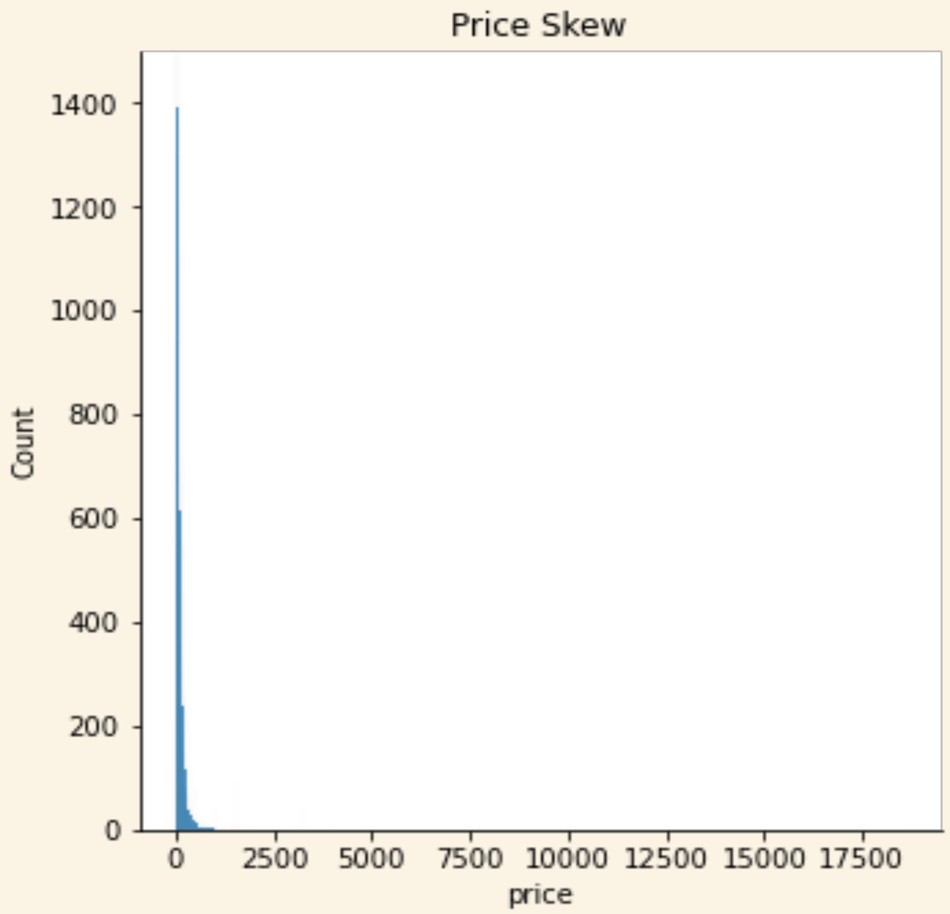
EXPLORATORY DATA ANALYSIS

EDA was used to provide direction on downstream data preprocessing and feature engineering

1. Are there any structural issues with the dataset?
2. Do certain attributes need to be transformed, normalized?
3. Are there features we need to delve into deeper?

EXPLORATORY DATA ANALYSIS

1. Price Skew

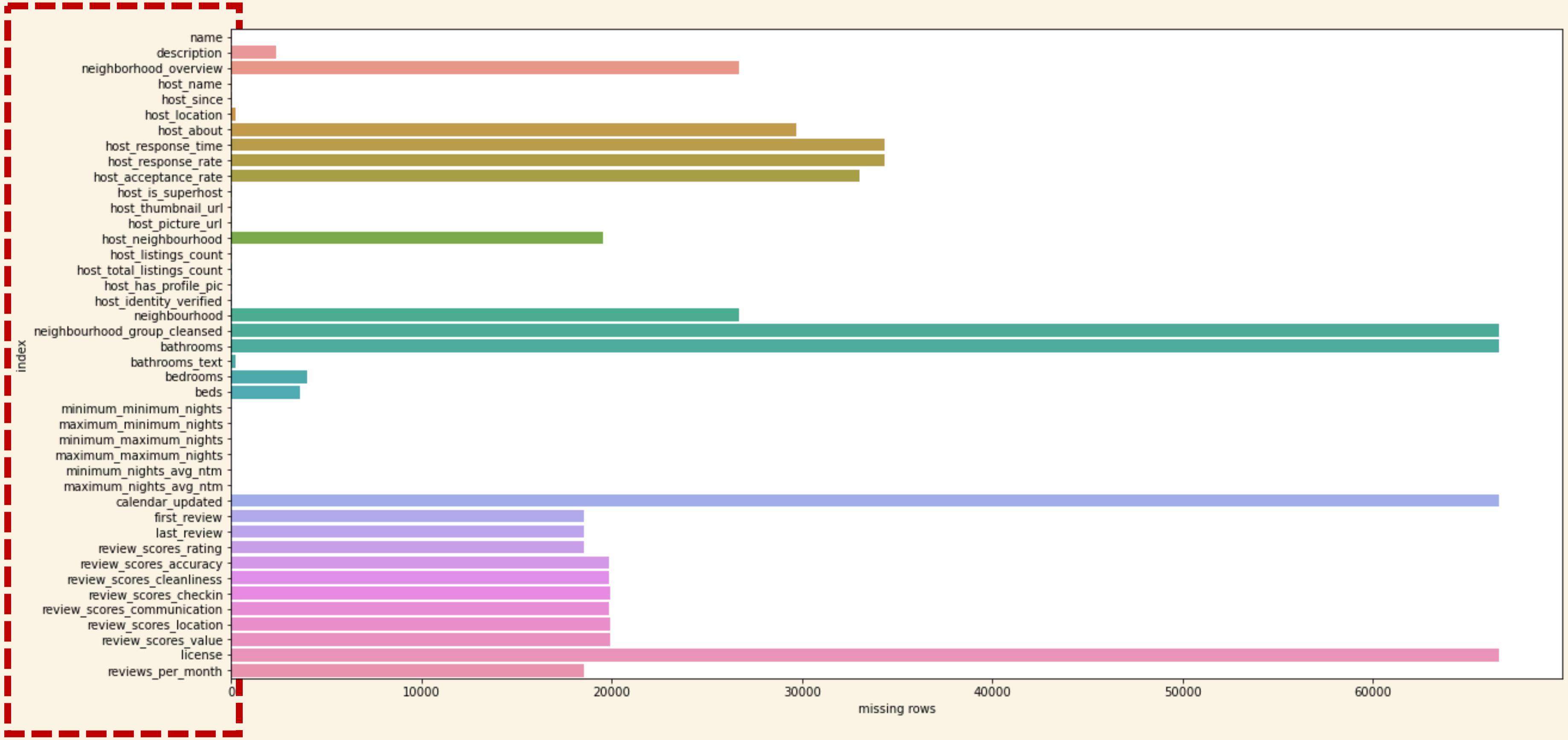


- **Right skewed** distribution for price
- **Performance impacts** of skewed data to non-tree-based models

- **Transformation methods** to address and remediate this gap (e.g., log transform or PowerTransform).

EXPLORATORY DATA ANALYSIS

2. Missing Values

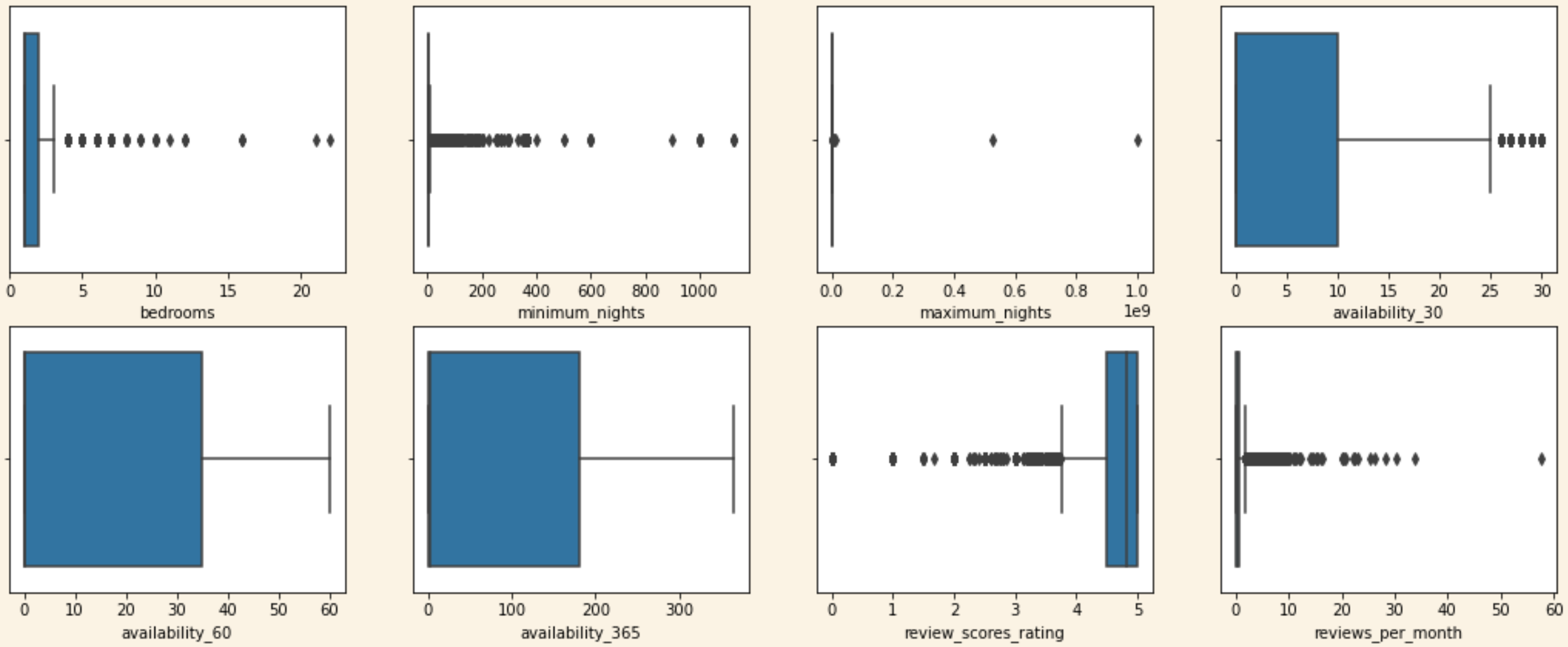


- **57%** of attributes have missing values
- **All rows** have at least 1 missing values

- **Dropping rows** not an option.
- **Impute** missing values instead.

EXPLORATORY DATA ANALYSIS

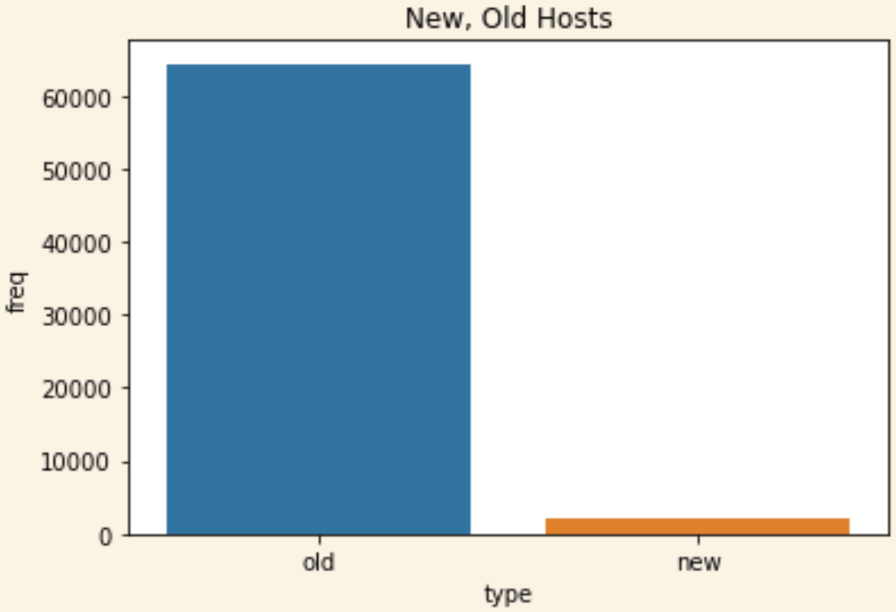
3. Attributes with Outliers and Skews

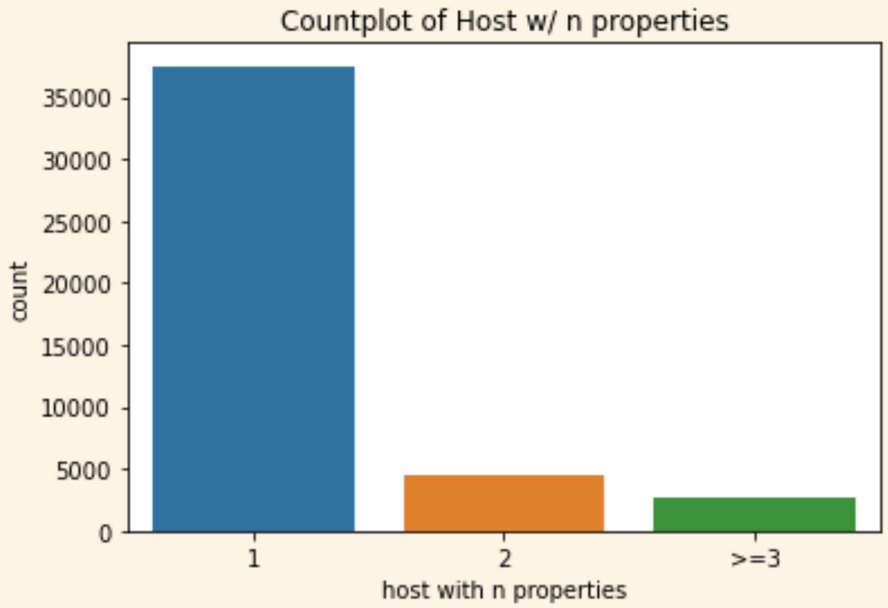


- **Box plots** were created for numerical values to easily spot outlier and skewed distributions
- **Winsorization & transformation methods** can be used.

EXPLORATORY DATA ANALYSIS

4. Host Tenure & Property Portfolios





- **Majority** of hosts have been in the platform for at least a year.
- **Most hosts** only have 1 listing, with only 6% having 3 or more listings

To what extent does this **impact price**?

DATA TRANSFORMATION

DATA PRE-PROCESSING

1. Data pre-processing was also done to the listings and reviews dataset
2. Feature engineering was done to enrich the host, property, location and sentiment

LISTINGS DATA	REVIEWS DATA
Dropping of non-value add columns	Dropping of non-value add reviews
Removal of inactive properties	Dropping of foreign language entries
Extract amenity features	Lemmatization
Imputation of missing values	
Winsorization of outliers	
Latitude and Longitude transformation	
One Key Hot encoding	
Log transformation of price	

FEATURE ENGINEERING

Additional relevant features related to host, property, location and sentiment were engineered to enhance our price prediction models and recommendation system.

Summary of the features engineered:

HOST/PROPERTY	LOCATION	SENTIMENT ANALYSIS
Host Duration	Nearest Station	Net Sentiment Score
Properties in London	Station Distance	
Professionally Managed Property	Walking Distance	
Occupancy Rate		

FINAL DATASET

After data pre-processing steps and feature engineering, our final dataset has:

Original Dataset (Reviews + Listings)

74
Original Features

55%
Numerical Features

45%
Categorical Features

Dropped & Engineered Features

25
Dropped Features

20
Engineered Features

Final Dataset (After Feature Selection)

65
Total Features

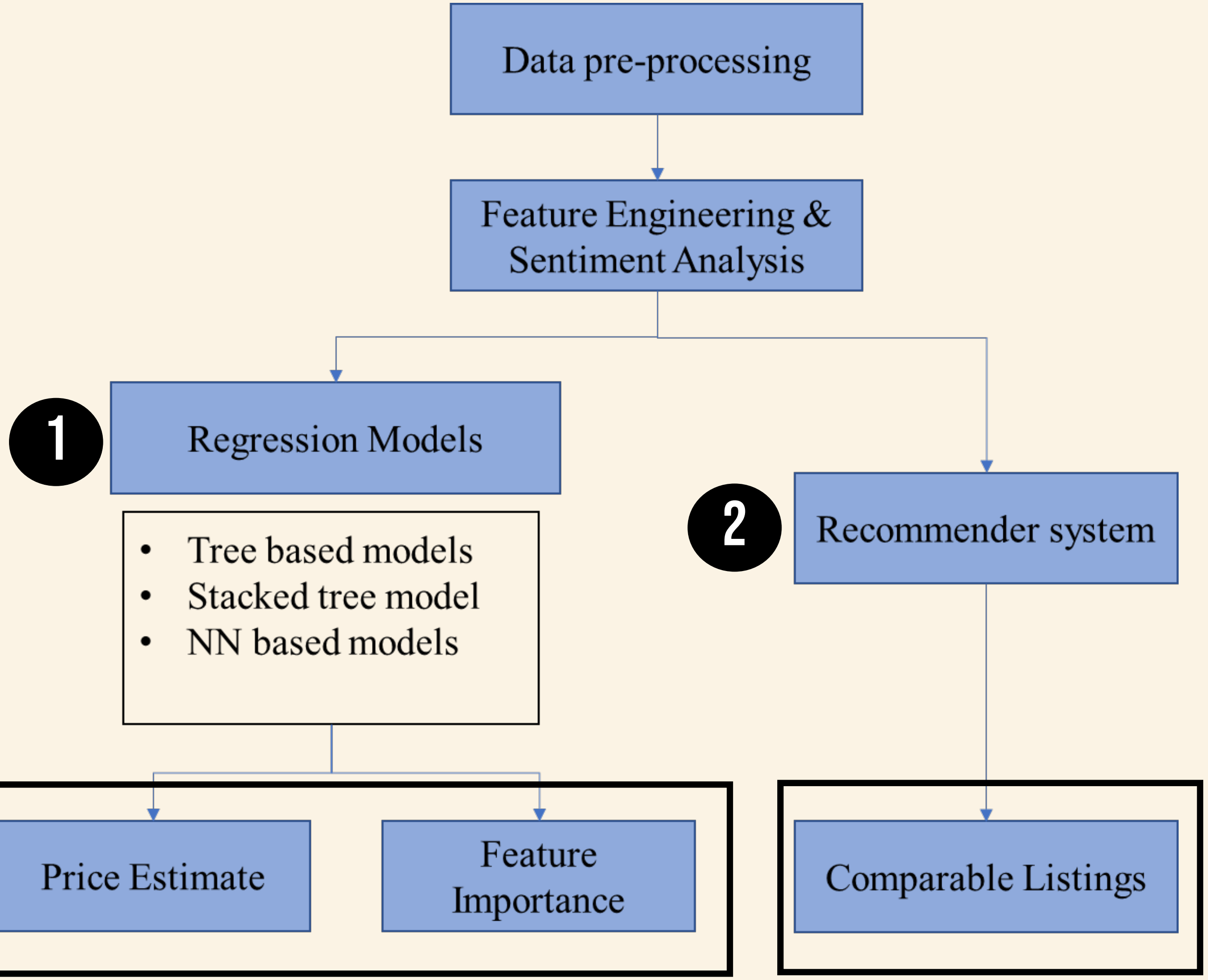
32
One Hot Encoding Features

33
Numerical Features

MACHINE LEARNING MODELS

MACHINE LEARNING MODELS

The implementation flow of the 2 models proposed follow the diagram below:



REGRESSION MODELS

REGRESSION MODELS: APPROACH

Key Objectives: Estimate listing price and identify important features

Data Preparation & Baseline Model	• Data was split into 80% training set and 20% test set. • Baseline model chosen was simple OLS Regression models without feature selection.
Model Selection & Training	• 2 families of models were selected: Tree (Ensemble and Boosting) and Neural Network based models. • Tree Based models: LightGBM, XGBoost, Random Forest, Stacked Regressor • Neural Network models: Baseline, Deep and Wide Neural Net
Hyperparameter Tuning	• 2-steps Approach: - RandomizedSearchCV → narrow search region for best parameters. - Then GridSearch CV → find the final best parameters to use in the model.
Results Evaluation & Selecting Best Model	• 5-fold CV was done on results from best performing models to avoid overfitting. • Best model was chosen based on R-Squared and MSE from validation set.

REGRESSION MODELS: APPROACH

Two families of non-linear regression models were chosen because...

TREE BASED MODELS

(Ensemble and Boosting Methods)

Chosen for performance and transparency top features influencing listing price.

Chosen Models

LightGBM, XGBoost, Random Forest and Stacked Regressor

NEUTRAL NETWORK MODELS

Chosen for flexibility and ability to model complex, non-linear relationships in large dataset.

Chosen Models

Baseline Neural Network, Deep Neural Network and Wide Neural Network

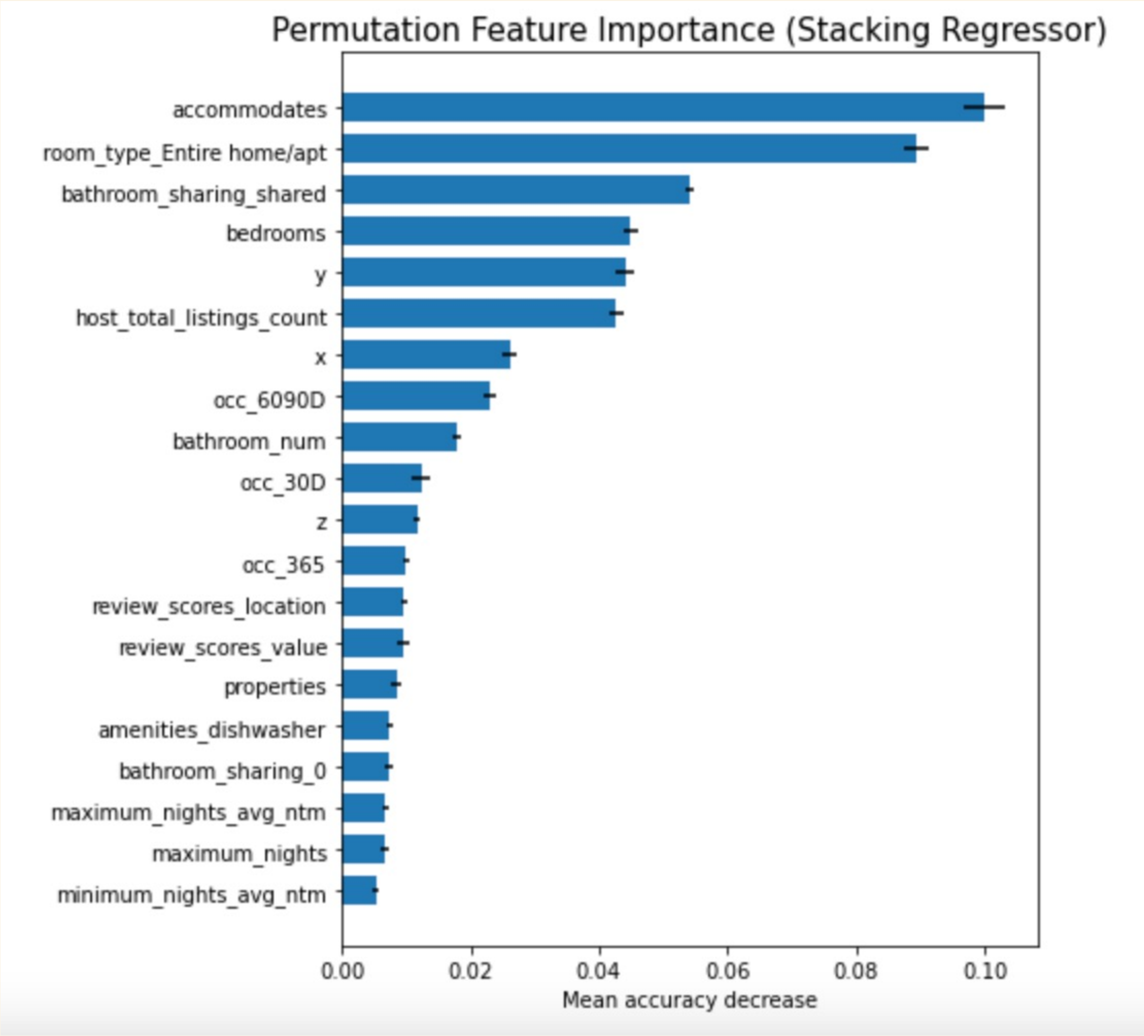
REGRESSION MODELS: EVALUATION & RESULTS

Each model is retrained using the best parameters and evaluated using 5-fold cross validation.

Model Name	Cross Validated R-Squared	Cross Validated MSE
Simple OLS	0.678	0.200
LightGBM Regressor	0.779	0.138
XGBoost Regressor	0.769	0.144
Random Forest Regressor	0.748	0.157
Stacked Regressor	**0.785**	**0.134**
Baseline Neural Network	0.691	0.183
Deep Neural Network	0.735	0.158
Wide Neural Network	0.696	0.180

REGRESSION MODELS: FEATURE IMPORTANCE

Permutation Feature Importance (PFI) was adopted to explain the results from non-linear models.

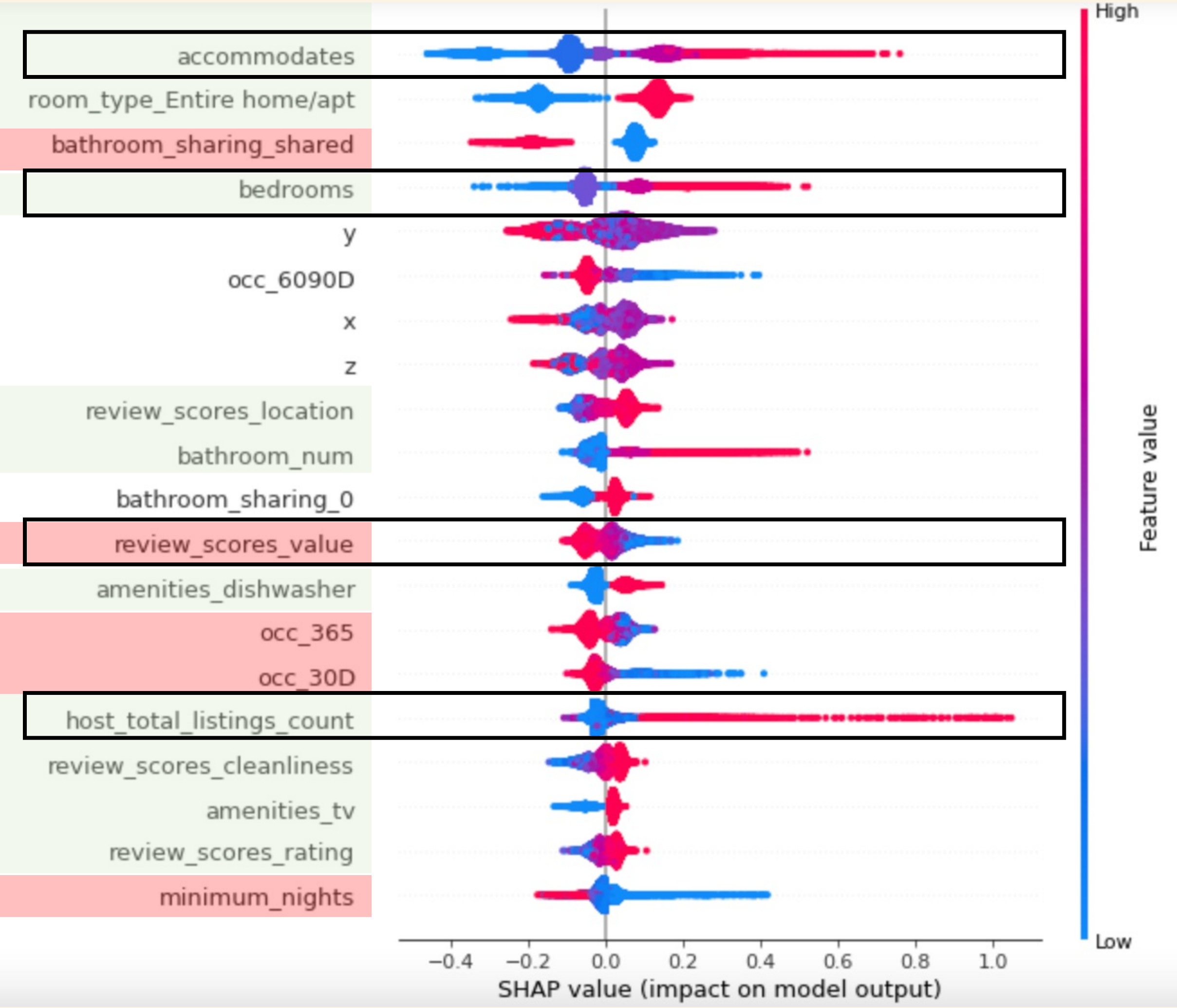


Variables which most explain prices are...

1. Those related to available space or number of rooms.
2. The ones which describe either the host's experience and status or the unit's listing position and availability.
3. Reviews scores, booking limitations and amenities availability.

REGRESSION MODELS: GLOBAL FEATURE IMPORTANCE

SHAP values further explain the effect of selected features on listing prices.



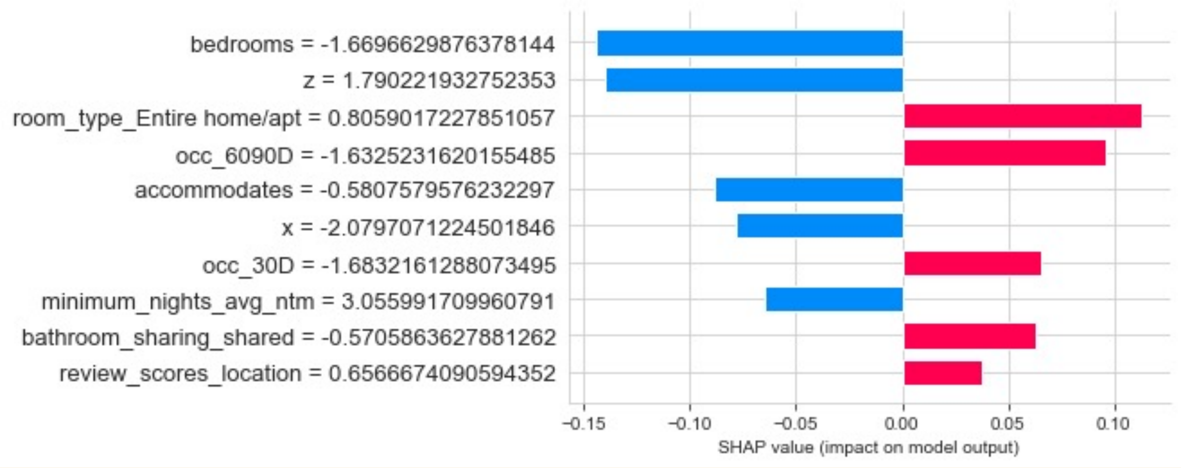
Listing prices will...

1. **Increase** when there are more rooms, space and amenities. Cleanliness, location and prior reviews also matter.

2. **Decrease** when bathrooms are shared, minimum nights to stay are higher, there is a high rating for value for money and high occupancy rates.

REGRESSION MODELS: LOCAL FEATURE IMPORTANCE

Local explainability graph also gives an interesting insight into a particular use case.



Listing price for this unit is...

1. **Increasing** as the unit is an entire apartment, bathroom is not shared, location review is high and the unit is available for longer term rental.
2. **Decreasing** because the unit size is below average and the number of minimum nights imposed is higher than average.

Further insights (on x and z variables) into location might give better understanding.

RECOMMENDER SYSTEM

RECOMMENDER SYSTEM: APPROACH

Key Objective: Benchmark a listed property against listings that are most similar to it.

1	Filter for features likely to appear in guest searches, including location features, and some relevant amenities.
2	Columns which are relevant from a result perspective but were used for feature comparison are dropped. This include IDs and prices.
3	Due to memory constraints, data is processed in 4 chunks and a similarity matrix is generated relative to the host property. The top 10 from each chunk are added to a shortlist.
4	From shortlist of 40 most similar properties, top 10 most similar are extracted and displayed to the host as a benchmark.

RECOMMENDER SYSTEM: LIMITATION AND EVALUATION

Due to the absence of user interaction data, it is not possible to come up with an evaluation metric. In lieu of this, we propose a future approach:

HOST SIDE

Track host interaction with outputted recommendation:

- Clicks on recommended properties.
- Rating on the quality of reviewed property recommendations.

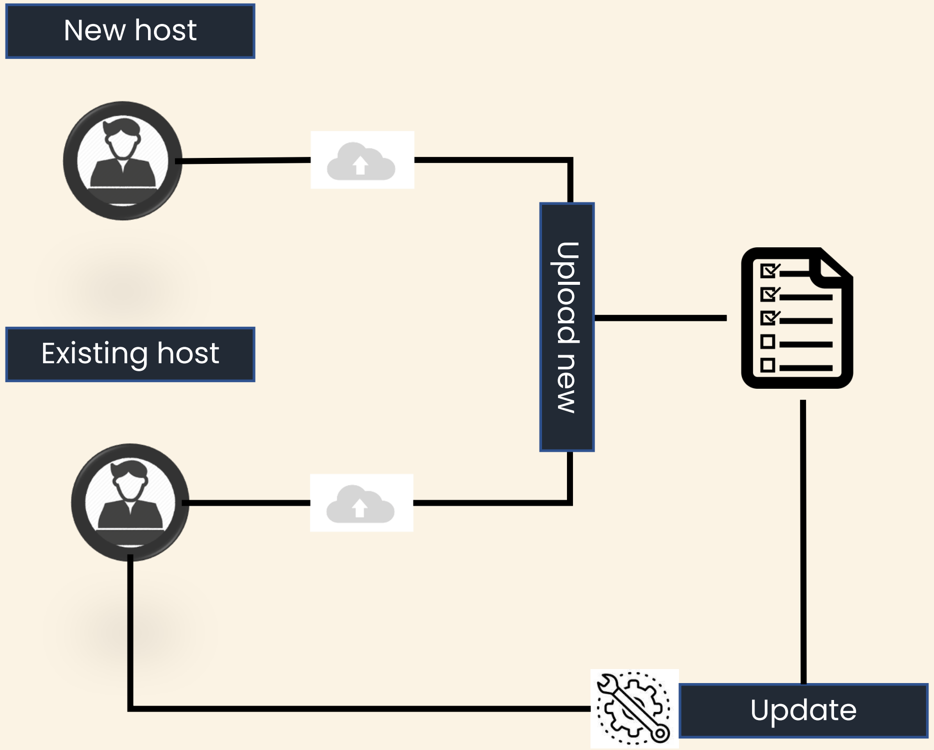
GUEST SIDE

Track how many of the top 10 lists recommended appear in guest searches.

BUSINESS APPLICATION

CONCLUSION: BUSINESS IMPLEMENTATION

Our models can be implemented seamlessly without disrupting existing operations for the hosts.



REGRESSION ENGINE OUTPUTS

New host

- Pricing recommendation
- Feature highlights

Existing host

- Top 10 most similar listings and their average features

RECOMMENDER ENGINE OUTPUTS

New host

- Pricing revision suggestions
- Feature revision suggestions

Existing host

- Top 10 most similar listings and their average features

CONCLUSION: BUSINESS IMPACT & CHALLENGES

Hosts could utilize our models in the following way to increase their earnings:

	ACTION	IMPACT	BLOCKERS/ IMPROVEMENTS
Price Recommendation	• Provide price adjustment suggestions	• Increased booking rate • Increased total earnings	• Hosts might be resistant towards changes/suggestions • Price seasonality to be taken into account • Influence of certain features might change overtime • Add customizability to competitor list by allowing hosts to choose features to be compared
Feature Highlights	• Highlights notable features noted as important by regression model	• Tagged features important to guests Increased booking rate • Increased total earnings	
Competitor List	• Compares property booking rate vs. that of top 10 most similar properties	• Callout to calibrate either price or features • Potential benchmark for occupancy rate	

THANK YOU!