Boston Airbnb Prediction of Listing Price

Made by Sabrina Rizal, Sarah Abdul Ra'uf, Chan Kah Kei and Amira Azlan

PROBLEM STATEMENT

- Predicting prices is **difficult** due to various influencing factors.
- Incorrect pricing can result in **underpricing** (lost revenue) or **overpricing** (fewer bookings).
- Understanding key factors can **improve pricing accuracy** and **booking rates**.

OBJECTIVES

1. Identify **key features** impacting pricing.
2. Build a **regression model** to predict Airbnb prices.
3. Achieve **high accuracy** to minimize pricing errors.

METHODOLOGY

DATA PREPARATION & CLEANING

↓

EXPLORATORY DATA ANALYSIS (EDA)

1. Distribution Charts (Numerical/Class)
2. Correlation Analysis (Boxplots, Pearson, Point-Biserial, Kruskal-Wallis)

↓

MODEL DEVELOPMENT

1. Multiple Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. Support Vector Regression
5. Gradient Boosting Machine Regression
6. XGBoost Regression
7. Voting Regression Model
8. Neural Networks

↓

HYPERPARAMETER TUNING

1. Grid Search CV

↓

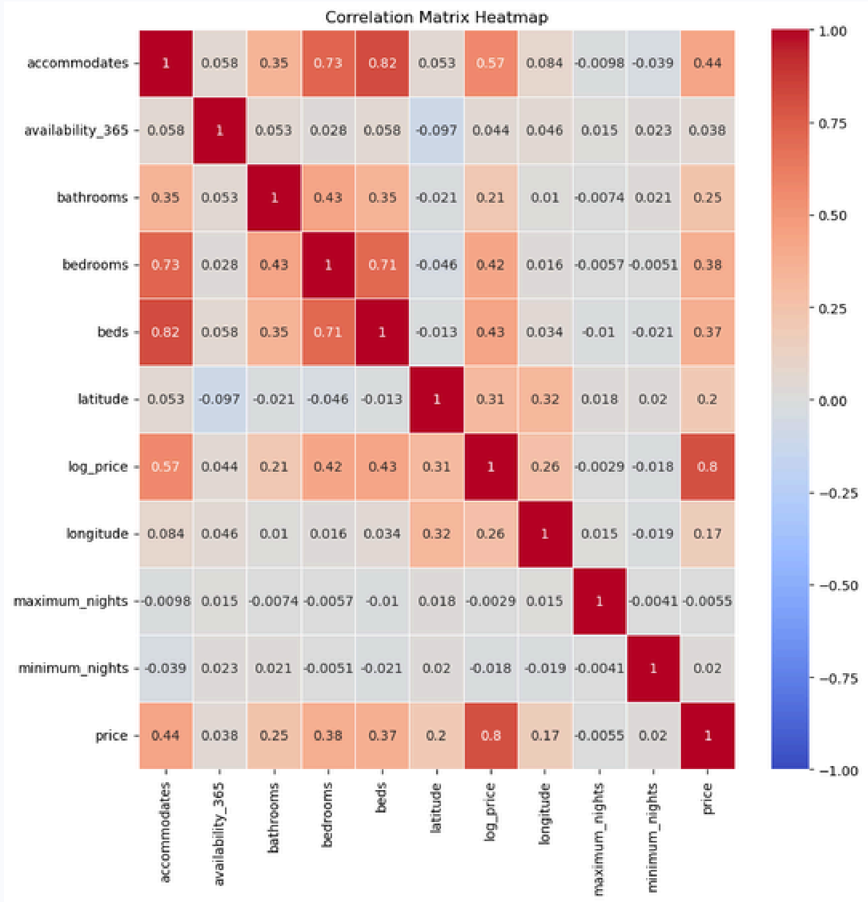
MODEL EVALUTATION

1. RMSE Learning curve
2. MAE
3. MSE
4. R^2

CORRELATION ANALYSIS

1. Pearson's (Numerical) Results

- **Strong:** accommodates, bathrooms, bedrooms, beds, latitude and longitude
- **Weak:** availability_365, maximum_nights, minimum_nights



2. Point-Biserial Correlation (Boolean Variables)

- **Strong Positive Correlations:** TV, air conditioning, cable TV, family/kid friendly, elevator, gym, doorman.
- **Strong Negative Correlations:** Lock on the bedroom door, free parking on premises, allow smoking, cats, pets on the property.

	Feature	Correlation	P-value	Significant
0	instant_bookable	-0.076629	4.527309e-06	Yes
1	Pets Allowed	0.086800	2.031561e-07	Yes
2	Dryer	0.168753	3.099260e-24	Yes
3	Dog(s)	-0.015865	3.430882e-01	No
4	Indoor Fireplace	0.113085	1.210133e-11	Yes
5	Elevator in Building	0.296059	3.314107e-73	Yes

3. Kruskal-Wallis Test (Categorical Variables)

- Significant differences observed across all categorical variables.

	Column	Statistic	P-Value	Significance
0	bed_type	119.070502	8.436792e-25	Significant
1	cancellation_policy	296.396510	5.993080e-64	Significant
2	neighbourhood_cleansed	1159.058571	1.301998e-229	Significant
3	property_type	243.510265	3.081255e-45	Significant
4	room_size	134.897939	5.096648e-30	Significant
5	room_type	1779.100883	0.000000e+00	Significant

MODEL PERFORMANCE

Model	MAE	MSE	R^2	Ranking
Multiple Linear Regression	0.273	0.138	0.681	4
Decision Tree Regression	0.301	0.164	0.620	7
Random Forest Regression	0.251	0.121	0.719	2
Support Vector Regression	0.357	0.216	0.497	8
Gradient Boosting Machine Regression	0.258	0.123	0.714	3
XGBoost Regression	0.286	0.149	0.655	6
Voting Regression Model	0.253	0.120	0.722	1
Neural Network Model	0.270	0.141	0.673	5

Best Model

- The **Voting Regressor Model** achieved the best overall performance with the lowest MAE (0.253), lowest MSE (0.120), and highest R^2 (0.722), indicating excellent accuracy and minimal error.

Worst Model

- The **Support Vector Regression Model** performed the worst, with the highest MAE (0.357), highest MSE (0.216), and lowest R^2 (0.497), showing difficulty in capturing relationships effectively.

REFERENCES

1. Dataset used for this project was taken from the Boston Airbnb Open Data dataset on Kaggle: [Boston Airbnb Open Data]
2. Full code for this project is available on GitHub: [Boston Airbnb Price Prediction].
3. This project was developed as part of our AI group assignment at Multimedia University (MMU).