Boston Airbnb Prediction of Listing Price

1201102251 Sabrina Amalyn Binti Aminur Rizal - 1201102478 Sarah Shahmina Binti Abdul Ra'uf - 1211307001 Chan Kah Kei - 1211307539 Amira Raina Binti Azlan Rahman

Introduction

Airbnb has transformed travel by offering a platform for hosts to list accommodations. However, determining the right price for a listing is challenging due to factors like location, property type, and amenities. Incorrect pricing can lead to lost revenue or reduced bookings. This project aims to predict Airbnb prices in Boston to help hosts optimize pricing strategies.

AI Problem Formulation

Problem Statement:

- Predicting Airbnb 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:

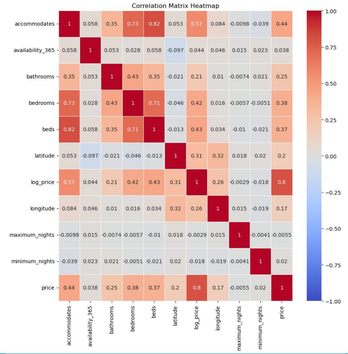
- Identify key features impacting pricing.
- Build a regression model to predict Airbnb prices.
- Achieve high accuracy to minimize pricing errors.

Data Preparation & Processing Pipeline

A. Data Cleaning

1. Relevant features were selected based on the problem statement and objective.
2. The price column was converted from string to float, and instant_bookable was adjusted.
3. One-hot encoding was applied to the amenities column.
4. Null values were handled by either filling them or removing records.
5. A new boolean column, has_public_transportation, was created from the transit column.
6. The space column was classified into room_size (small, medium, large) using Bart-Large-MNLI zero-shot classification.

B. Exploratory Data Analysis (Correlation)



	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	0.166753	3.099260e-24	Yes
3	Dog(s)	-0.015865	3.410882e-01	No
4	Indoor Fireplace	0.113085	1.210133e-11	Yes
5	Elevator in Building	0.296059	3.314107e-73	Yes
6	Pool	0.125781	4.492704e-14	Yes
7	Hair Dryer	0.139061	6.825506e-17	Yes
8	Washer	0.160145	5.887091e-22	Yes
9	Wheelchair Accessible	0.098457	3.690487e-09	Yes
10	Carbon Monoxide Detector	0.013163	4.315452e-01	No
11	Other pet(s)	-0.030754	6.604540e-02	No
12	Free Parking on Premises	-0.164996	3.169116e-23	Yes
13	Family/Kid Friendly	0.309472	3.692240e-80	Yes
14	Essentials	0.078165	2.899318e-06	Yes
15	Hot Tub	-0.092911	2.639400e-08	Yes
16	Washer / Dryer	0.029828	7.463135e-02	No
17	Safety Card	-0.013483	4.350303e-01	No
18	Smoking Allowed	-0.110297	2.822296e-11	Yes
19	TV	0.391242	5.055333e-131	Yes
20	Internet	0.195366	4.522771e-32	Yes
21	First Aid Kit	-0.097139	5.950332e-09	Yes
22	Breakfast	-0.085580	1.130764e-07	Yes
23	Heating	0.118748	1.072919e-12	Yes
24	Hangers	0.026020	1.199376e-01	No
25	Pets live on this property	-0.107247	1.301964e-10	Yes
26	24-Hour Check-in	0.172178	3.548622e-25	Yes
27	Laptop Friendly Workspace	0.039002	1.973075e-02	Yes
28	Free Parking on Street	-0.009226	5.814402e-01	No
29	Smoke Detector	-0.014968	3.710712e-01	No
30	Buzzer/Wireless Intercom	0.128199	1.447140e-14	Yes
31	Air Conditioning	0.352851	2.965584e-105	Yes
32	Fire Extinguisher	0.031792	4.340732e-02	Yes
33	Iron	0.156144	6.124685e-21	Yes
34	Cat(s)	-0.107874	1.014750e-10	Yes
35	Cable TV	0.351260	2.919893e-104	Yes
36	Lock on Bedroom Door	-0.233961	1.267824e-45	Yes
37	Gym	0.247617	4.622636e-51	Yes
38	Doorman	0.211725	1.711754e-37	Yes
39	Wireless Internet	0.035170	3.553808e-02	Yes
40	Paid Parking Off Premises	0.019372	2.469963e-01	No
41	Suitable for Events	0.027319	1.025318e-01	No
42	Kitchen	0.140044	4.114003e-17	Yes
43	Shampoo	0.100437	1.779963e-09	Yes
44	has_public_transportation	0.041831	1.239586e-02	Yes
45	price	0.803732	0.000000e+00	Yes

	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

Pearson's (numerical) results

a. Strong: accommodates, bathrooms, bedrooms, beds, latitude and longitude
b. Weak: availability_365, maximum_nights, minimum_nights

Point-Biserial (boolean) results

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

Kruskal-wallis (categorical) results

a. Significant across all categorical variables

AI Models Implementation

1. **Multiple Linear Regression**
   - Predicts using a linear equation with multiple inputs.
   - Simple, assumes linear relationships.
2. **Decision Tree Regression**
   - Splits data into branches for predictions.
   - Handles non-linear data, prone to overfitting.
3. **Random Forest Regression**
   - Ensemble of decision trees for better accuracy.
   - Reduces overfitting, robust, slower.
4. **Support Vector Regression**
   - Finds best-fit hyperplane within a margin.
   - Works for non-linear data, needs tuning.
5. **Gradient Boost Machine Regression**
   - Builds sequential trees (by correcting the mistake of previous tree)
   - Reduces bias in handling large dataset
6. **XGBoost Regressor**
   - Optimized Gradient Boosting for fast, accurate predictions.
   - Handles large data, builds sequential trees.
7. **Voting Regression Model**
   - Combines multiple model (using Gradient Boost and Random Forest)
   - Aggregates prediction from different regressors
8. **Neural Network**
   - Learns complex patterns using layered neurons.
   - Flexible, needs large data and tuning.

Experiments

1. **Model Training and Testing**
   - Trained the 8 models using 70% training set and 30% testing set.
2. **Hyperparameter Tuning**
   - GridSearchCV was used to systematically optimize the hyperparameters for the models. Each model will have different hyperparameters to tune.
3. **Building Model with Optimized Parameters**
   - Re-train each model using the best hyperparameters.
   - Test optimized models on the validation set to further refine performance.
   - Learning curves are generated to ensure models do not underfit or overfit.

Results

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	4

Discussion

- **Best Model:**
  Voting Regressor Model (lowest MAE, lowest MSE, highest R^2 - best overall accuracy and lower error)
- **Worst Model:**
  Support Vector Regression (highest MAE, highest MSE, lowest R^2 - this model struggles to capture relationships)