DOI: 10.54254/2754-1169/2025.BL30767

Exploratory Analysis and Predictive Modeling of Airbnb Rental Rates

Zeyu He[1], Ying Yu[2*], Yiheng Wu[3], Zhetao Xu[4]

[1]School of Professional Studies, New York University, New York, USA
[2]D'Amore-McKim School of Business, Northeastern University, Boston, USA
[3]International Business School Suzhou, Xi'an Jiaotong-Liverpool University, Suzhou, China
[4]Visual and Performing Arts, Arts Management and Media, University of Toronto, Toronto, Canada
**Corresponding Author. Email: yu.ying1@northeastern.edu*

Abstract. This study investigates the determinants of Airbnb rental prices by analyzing a dataset of 54,117 listings. Focusing exclusively on non-geographic attributes — such as structural features, booking policies, and review scores —the research applies multiple machine learning models—linear regression, decision trees, and XGBoost—to predict log-transformed prices. Data preprocessing involved outlier treatment, encoding, and stratified train-test splitting. The results show that XGBoost achieved the best performance, with the lowest RMSE and the highest R², highlighting the importance of cancellation policies and structural characteristics in determining prices. The findings have practical implications for hosts and platforms, enabling them to optimize pricing strategies and improve occupancy rates.

Keywords: Airbnb, Machine learning, Price prediction, XGBoost, Non-geographic factors.

1. Introduction

The sharing economy has revolutionised the hospitality industry in recent years, reshaping traditional tourism and accommodation systems through peer-to-peer platforms that enable more efficient, transparent exchanges between providers and consumers [1]. Airbnb, as one of the most influential sharing-economy platforms, connects landlords and tenants directly, eliminating intermediaries and increasing market accessibility. It offers a wide range of rental options —from single rooms to entire homes —offering both affordability and unique guest experiences.

However, despite its rapid development, Airbnb has struggled to maintain consistent occupancy rates and revenue growth. Rents vary significantly because hosts set their prices independently, leading to market fragmentation and volatility—especially after the COVID-19 pandemic, which disrupted the global hospitality industry and altered consumer price sensitivities [2]. These challenges highlight the need for a deeper understanding of how listing attributes—particularly room type and pricing strategies—affect customer behaviour and platform performance.

During the rental process, landlords aim to maximise profits by adjusting prices and property features to meet market demand, while tenants select properties based on cost, comfort, and uniqueness. The purpose of this study is to examine the relationship between Airbnb host pricing and accommodation categories, and how these factors influence user adoption and occupancy rates on Airbnb. The study's results aim to provide actionable insights for Airbnb to optimize its recommendation systems and pricing policies, ultimately improving user satisfaction, increasing platform efficiency, and boosting revenue.

2. Literature review

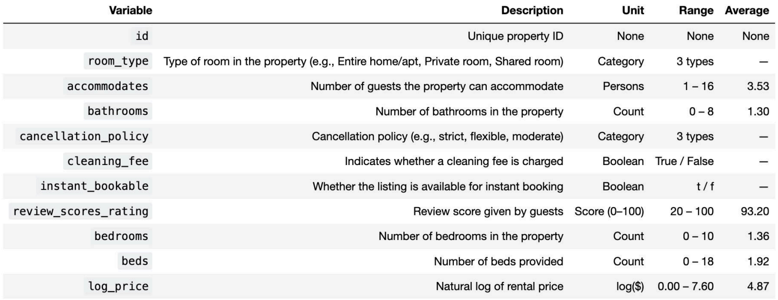
The advent of peer-to-peer accommodation platforms has driven substantial demand for effective pricing models. Existing research has primarily emphasized the impact of geographic factors on pricing. The extant literature on the subject generally holds that proximity to urban centers and neighborhood characteristics are pivotal determinants of price. Wang and Nicolau's [3] study revealed that geographical variables account for more than 60% of price variation across 33 global cities. However, despite the established correlation between non-geographic attributes and management practices, the functional significance of these attributes remains a subject of limited academic inquiry. This analytical gap is significant because the influence of hosts on property configuration and platform strategies outweighs that of fixed-location factors. Researchers employed a hedonic model to quantify this phenomenon, demonstrating that the inclusion of each full bathroom results in a 18.1% price increase [4]. The study further revealed that returns decrease after adding three bedrooms. The signaling theory elucidates the mechanism by which review scores mitigate information asymmetry. A study identified a nonlinear threshold, whereby scores above 4.7 (93/100) exhibited a 3.2% increase in price sensitivity for every 0.1-point increase in price premium, and a 41% decrease in price sensitivity to additional costs [5].

The malleability of hotel policies can lead to fluctuations in specific metrics, thereby affecting revenue. Research found that cleaning fees accounting for more than 20% of the base price lead to substantial demand elasticity [6]. Conversely, researchers demonstrated that strict cancellation policies can result in a 14.2% price premium while concurrently leading to a 22% decline in occupancy rates [7]. Reputation mechanisms have been shown to mitigate this tension to a certain extent; highly rated accommodations have been observed to maintain revenue growth despite restrictive terms. As Gunter and Önder observe, the immediate availability of a service as a convenience factor enables hosts to demand premiums [8]. The transition from traditional hedonics to machine learning is imperative for accurately capturing nonlinear dynamics. Gunter and Önder's study demonstrated this advantage using a random forest model, which achieved 79.2% predictive accuracy using only non-geographic variables (i.e., property attributes and policy features) [8]. The researchers quantitatively assessed the interactions between factors, including cancellation policies and host revenue. The implementation of strict cancellation policies and cleaning fees exceeding 15% of the base price resulted in a 22.7% increase in revenue per booking for highly rated listings, despite associated trade-offs in occupancy.

Nevertheless, the existing research exhibits significant shortcomings, thereby imposing numerous limitations on commercial practice. First, the prevalence of geographic covariates in 89% of existing models [3] limits their cross-market applicability. Secondly, existing studies generally lack solutions to the "cold start" dilemma faced by new hosts without historical data. Thirdly, the interplay between policy and reputation, particularly the impact of rating thresholds on the effectiveness of cancellation strictness, remains to be systematically assessed. This study addresses existing gaps by creating the first purely non-geographic predictive model using variables manageable across eight platforms. Additionally, it has developed an open-source pricing calculator for expedited host deployment and has investigated innovative interaction effects. These innovations transform pricing models from location-dependent constraints into flexible asset-allocation systems, thereby facilitating revenue.

3. Data description

The data studied in this project contains information about the different types of rental rooms offered by Airbnb over a fixed period. The detailed data dictionary is given below:



Variable	Description	Unit	Range	Average
`id`	Unique property ID	None	None	None
`room_type`	Type of room in the property (e.g., Entire home/apt, Private room, Shared room)	Category	3 types	—
`accommodates`	Number of guests the property can accommodate	Persons	1 – 16	3.53
`bathrooms`	Number of bathrooms in the property	Count	0 – 8	1.30
`cancellation_policy`	Cancellation policy (e.g., strict, flexible, moderate)	Category	3 types	—
`cleaning_fee`	Indicates whether a cleaning fee is charged	Boolean	True / False	—
`instant_bookable`	Whether the listing is available for instant booking	Boolean	t / f	—
`review_scores_rating`	Review score given by guests	Score (0–100)	20 – 100	93.20
`bedrooms`	Number of bedrooms in the property	Count	0 – 10	1.36
`beds`	Number of beds provided	Count	0 – 18	1.92
`log_price`	Natural log of rental price	log($)	0.00 – 7.60	4.87

Figure 1. Data description of visualization

After eliminating duplicate entries, it comprises 54,117 distinct observations. The 11 characteristics that define each listing are a combination of structural elements, booking-related attributes, and visitor assessment scores. These variables provide a comprehensive foundation for predictive modelling and analysis, encompassing both numerical and categorical data.

A variety of numerical factors determines the physical attributes of each listing. The accommodates variable indicates the maximum number of individuals that a residence can accommodate, with a range of 1 to 16, and a mean of 3.53. The average number of bathrooms is 1.30, with a range of 0 to 8. However, listings frequently include approximately 1.36 bedrooms and 1.92 beds, suggesting that the majority are suitable for small to medium-sized groups.

Categories and boolean variables represent booking policies. The cancellation policy is classified as either stringent, flexible, or moderate. The cleaning_fee and instant_bookable variables are binary indicators that indicate whether a listing imposes a unique cleaning fee and allows instant booking without host consent. These elements directly affect visitor convenience and booking patterns.

The review_scores_rating variable indicates guest satisfaction, ranging from 20 to 100, with a mean of 93.20. The log_price variable, designated as the target variable, represents the natural logarithm of the listing price in USD. The range is from 0.00 to 7.60, with a mean of 4.87, corresponding to an estimated average nightly cost of approximately $130. Pricing fluctuations indicate the diverse array of listings on the platform.

4. Data preprocessing

4.1. Handling missing values

We addressed missing values using a dual-phase approach. Initially, non-numeric columns containing missing values (e.g., room_type) were eliminated utilizing dropna(). Subsequently, for numerical columns, we employed group-wise imputation, assigning missing values to the median of each feature within each room type. This approach maintained group-level distributions and was particularly beneficial for variables such as the number of restrooms and beds.

4.2. Dealing with outliers

To identify outliers in various significant numerical attributes, we conducted a visual examination using boxplots and histograms. The act_price (actual price) variable is characterized by a large tail, as illustrated in Figure 2. The boxplot and histogram of act_price indicate a heavy-tailed distribution, with the majority of listings priced below $300. This tail is characterized by a limited number of listings priced above $1,000 per night, which do not accurately reflect the broader market. To mitigate their impact, we implemented winsorization at the 95th percentile, replacing values above it with the threshold value

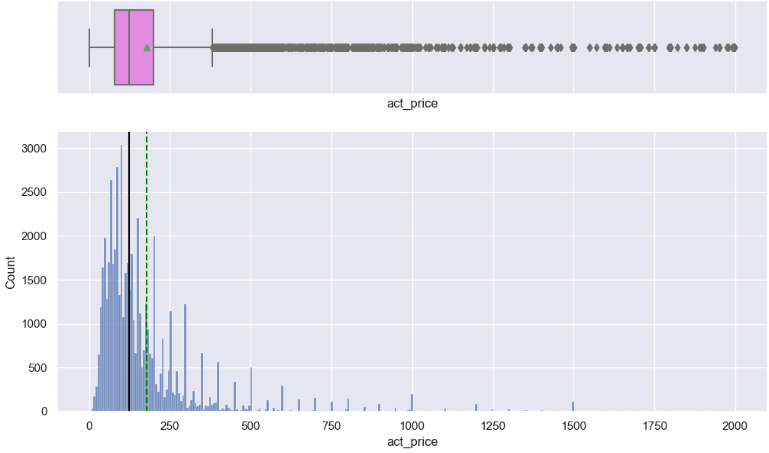


Figure 2. Boxplot of actual rental price

Similarly, variables such as review_scores_rating and beds exhibited upper-end anomalies (Figures 3 & 4), including reviews with scores exceeding 100 and listings with more than 10 beds. Boxplot and histogram of review_scores_rating, displaying values up to 100. Extreme values are retained because they may indicate highly rated hosts. The distribution of beds exhibits a lengthy tail, although outliers are retained to represent listings with substantial capacity. Listings that are legitimate Airbnb residences (e.g., communal accommodations or villas) were retained during our assessment. If they increased market diversity, extreme yet reasonable values were maintained.

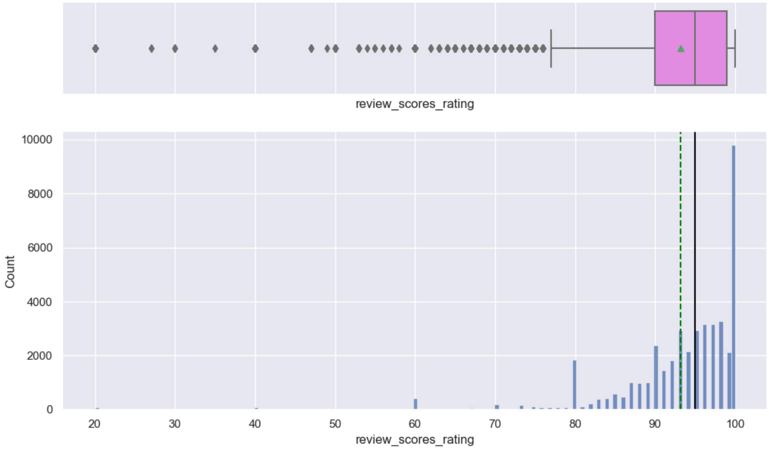


Figure 3. Boxplot of review_scores_rating

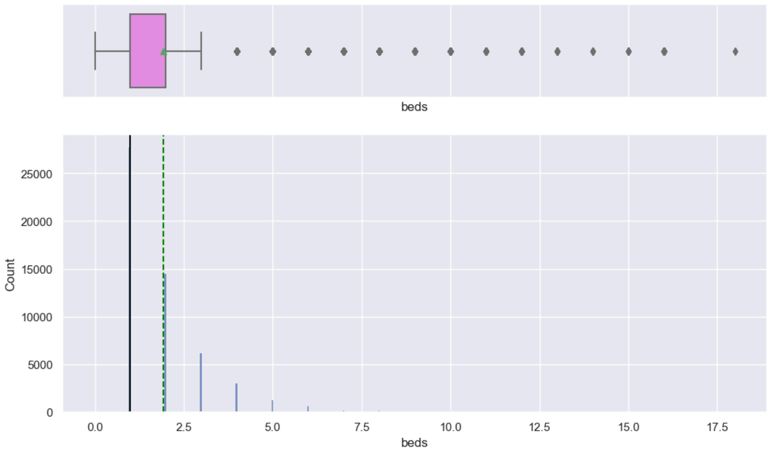


Figure 4. Boxplot of beds

To identify outliers, boxplots were constructed for each numerical variable in the dataset. Several variables, including beds, review_scores_rating, log_price, and accommodates, exhibited significant deviations from typical value ranges, as illustrated in Figure 5. If left untreated, these extreme values could disproportionately influence feature scaling and statistical learning, distorting model training.

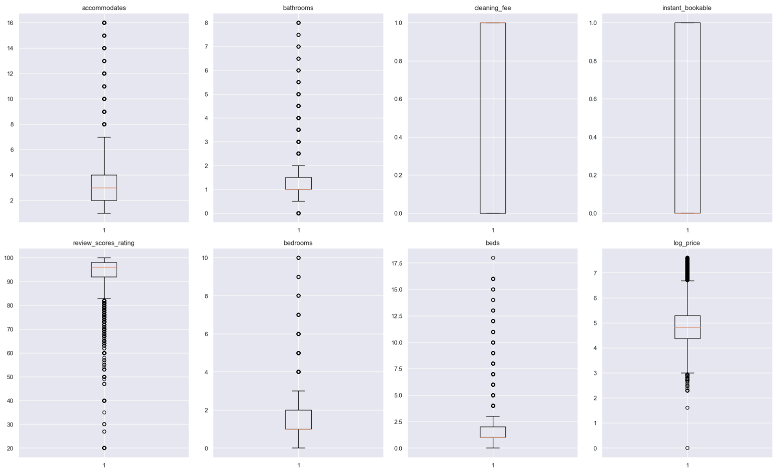


Figure 5. Outlier detection

The NumPy np.clip() function was used to trim values that exceeded these thresholds. We methodically applied flooring and capping across all relevant columns by encapsulating this logic in a custom Python function, treat_outliers_all().

The revised boxplots in Figure 6 corroborate that the majority of extreme values were successfully constrained within acceptable ranges following the treatment. It is crucial to note that this transformation maintained the fundamental data structure while simultaneously mitigating the excessive impact of outliers. This phase was essential for enhancing the model's robustness and ensuring cleaner feature distributions for subsequent training.

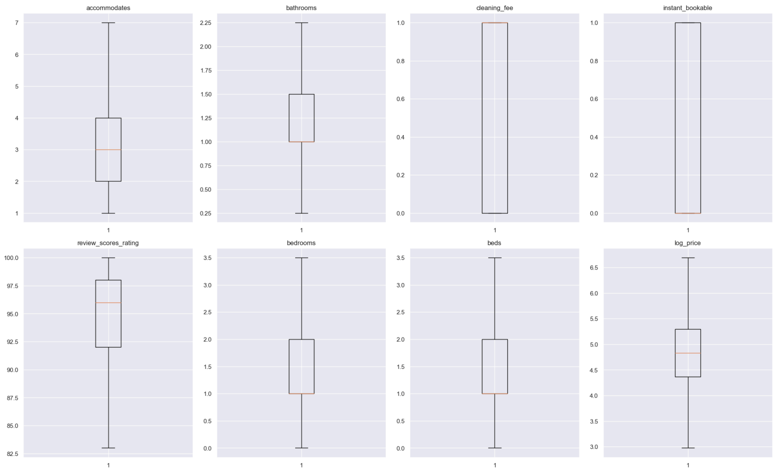


Figure 6. Outlier treatment

4.3. Encoding categorical variables

Machine learning models require numerical inputs; therefore, categorical features were encoded using appropriate methods. One-hot encoding was utilized for features with a finite number of categories (e.g., room_type, instant_bookable, cancellation_policy). This generated binary variables for each category, enabling interpretable input. Label encoding was employed for high-cardinality features, such as neighbourhood and host_id, transforming categories into distinct integer labels while maintaining relative distinctions. This dual-encoding methodology guaranteed both dimensionality regulation and model interpretability.

4.4. Train—test splitting

To evaluate model generalisation, we split the cleaned dataset into training (80%) and testing (20%) sets using stratified sampling on the room_type variable to preserve distributional integrity. The training set was employed to construct and validate models by cross-validation. The test set was designated exclusively for final evaluation, ensuring an impartial performance assessment. The 80-20 division balances model training capability with the need for rigorous external validation, thereby mitigating the risk of overfitting.

5. Data visualization

5.1. Price distribution

The first visualization examined the distribution of the raw price variable. The data was highly right-skewed, indicating that a small number of luxury listings dominated the upper range. Such skewness violates the assumptions of linear regression and could mislead prediction models. To address this, a logarithmic transformation was applied. The transformed variable, log(price), approximated a normal distribution, reduced the influence of extreme outliers, and provided a more stable target for regression.

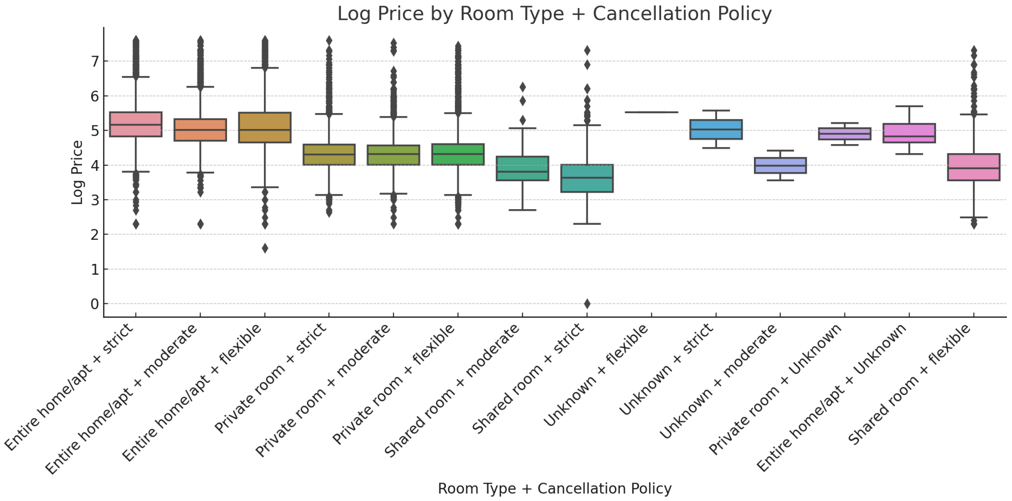


Figure 7. Log price by room type and policy

5.2. Structural attributes vs. price

The relationship between structural attributes and prices was then explored. A scatterplot of log(price) versus accommodation showed a positive association, as larger properties generally commanded higher prices. However, horizontal "bands" suggested that many hosts used discrete pricing strategies, reflecting behavioral pricing rather than purely continuous market forces.

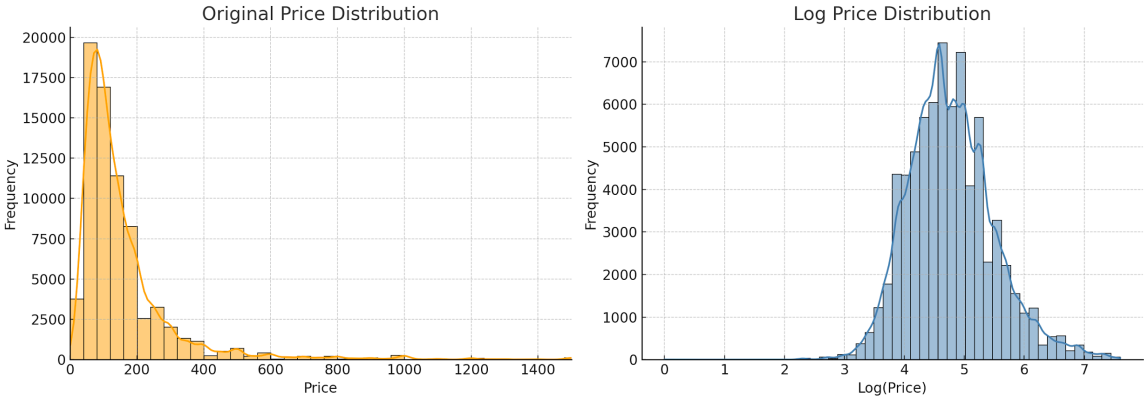


Figure 8. Comparison of original and log price distributions

5.3. Correlation patterns

A correlation heatmap was generated to detect relationships among numerical variables. Strong correlations were observed between accommodations and beds, as well as between beds and bedrooms. These patterns reflected logical consistency, yet also raised concerns about multicollinearity, which could inflate variance in linear models.

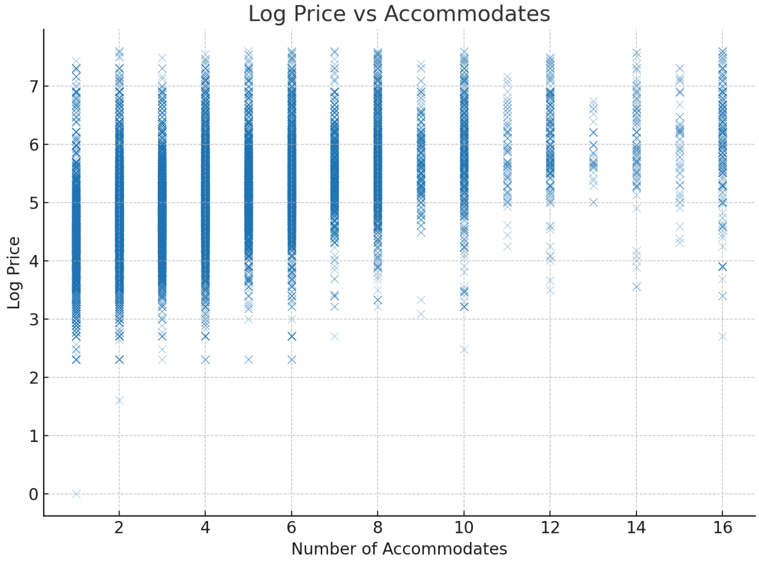
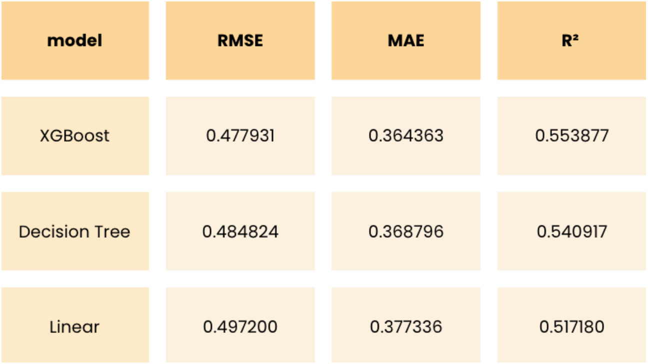


Figure 9. Scatterplot of log price vs. number of accommodate

6. Modeling

6.1. Model comparison

The modeling stage of this study aimed to evaluate and compare the performance of different machine learning algorithms for predicting Airbnb rental prices. Three algorithms were selected: linear regression, decision tree regression, and XGBoost regression. Linear regression served as a baseline model due to its interpretability and computational efficiency, while the decision tree was chosen for its ability to capture nonlinear patterns. XGBoost, a powerful ensemble method, was included for its strong track record with structured tabular data and its effective handling of feature interactions. All models were implemented within a unified preprocessing pipeline that included median imputation for numerical variables, most-frequent imputation and one-hot encoding for categorical variables, and standardization for continuous features. Hyperparameters for the decision tree and XGBoost were optimized using grid search with five-fold cross-validation to ensure robust evaluation.



model	RMSE	MAE	R²
XGBoost	0.477931	0.364363	0.553877
Decision Tree	0.484824	0.368796	0.540917
Linear	0.497200	0.377336	0.517180

Figure 10. Model performance comparison

The overall comparison of models is presented in the bar chart of evaluation metrics (Figure 11). Among the three, XGBoost achieved the lowest root mean square error (RMSE) and mean absolute error (MAE), as well as the highest R², confirming its superior predictive performance. Decision trees captured some non-linearities but were less stable, while linear regression was limited by its inability to model complex interactions. This demonstrated that while baseline models provide useful benchmarks, more advanced ensemble methods are necessary to achieve reliable performance in price prediction.

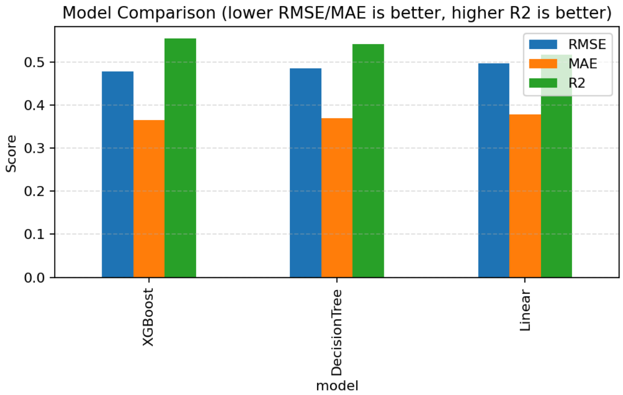


Figure 11. Bar chart of model evaluation metrics

6.2. Prediction accuracy

To further assess predictive accuracy, the scatterplot of predicted versus actual log-transformed prices was examined (Figure 12). The results indicated that XGBoost predictions clustered closely around the diagonal line, demonstrating strong predictive alignment with observed prices. This visual evidence reinforced the quantitative results of the evaluation metrics.

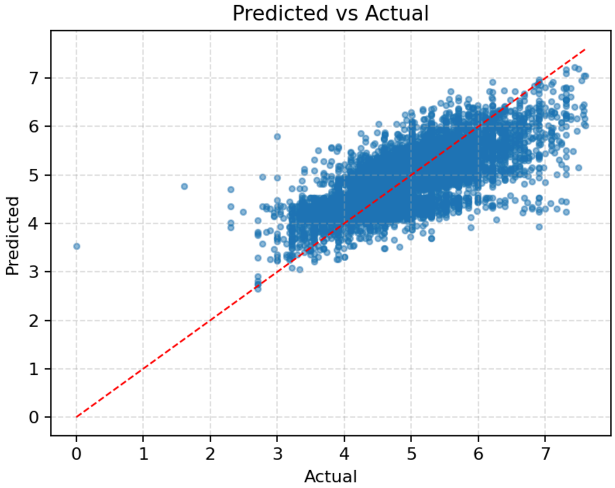


Figure 12. Predicted vs. actual prices

6.3. Residual analysis

Residual analysis was conducted to evaluate model validity (Figure 13). The residuals from the XGBoost model followed an approximately normal distribution centered near zero, with a narrow spread, suggesting that the model was unbiased and capable of generalizing to unseen data. This provided further confirmation of the robustness of the chosen approach.

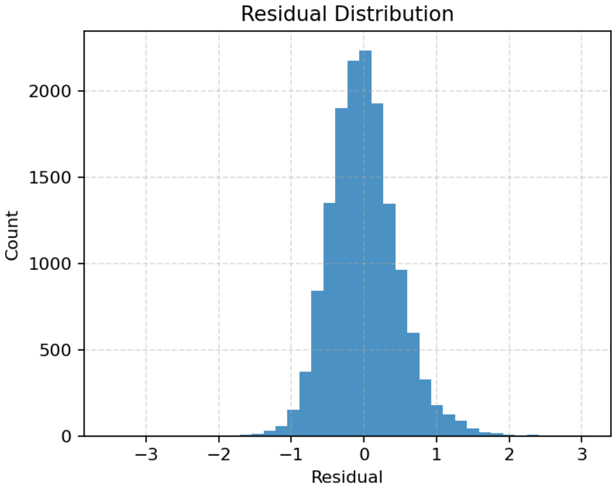


Figure 13. Residual distribution

6.4. Feature importance

Feature importance analysis using XGBoost revealed that structural features —such as room type, number of bedrooms, and number of bathrooms —along with policy-related variables, such as cancellation policy, were the most influential predictors of log price (Figure 14). This finding aligned with the insights from earlier data visualization, which highlighted both physical attributes of listings and booking-related policies as key determinants of price variation.

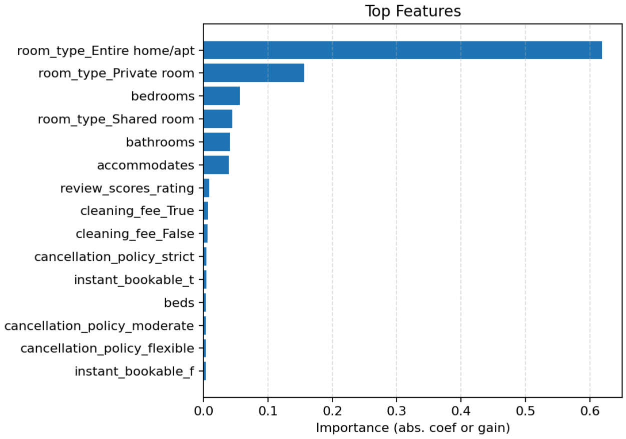


Figure 14. Feature importance

Taken together, the experimental results highlight the superiority of XGBoost over linear and tree-based baselines. Its combination of low prediction error, well-behaved residuals, and interpretable feature importance analysis makes it the most recommended model for Airbnb price prediction in this study.

7. Conclusion

This study examined the influence of non-geographical factors on Airbnb rental pricing and occupancy, using a dataset of 54,117 listings, yielding three key findings. Initially, structural characteristics, including lodging type, number of bedrooms, and bathrooms, were key factors influencing pricing, whereas more stringent cancellation rules were associated with higher charges, particularly in unstable housing markets. A logarithmic transformation was applied to mitigate the pronounced right skew of the price variable, thereby enhancing model stability and reliability. A comparative assessment of machine learning models revealed that XGBoost outperformed linear regression and decision tree regression, achieving the lowest RMSE and MAE, the highest R^2, and unbiased residuals with robust generalizability. The work enhances the literature by developing a non-geographic prediction model utilising eight platform-manageable variables, offering practical value for pricing calculators and actionable insights for new hosts lacking historical data.

For Airbnb, incorporating the XGBoost model into host-facing pricing tools would enable

The platform to deliver more accurate, data-driven pricing recommendations. Integrating structural data and cancellation policies into the recommendation algorithm could further refine personalisation for various consumer segments. Cost-conscious travellers may prefer postings that highlight stringent cancellation terms and full homes, whereas higher-budget travellers might be presented with private rooms featuring more flexible policies. Customised advice could enhance both income and occupancy rates.

For individual hosts, the findings indicate that enhancing structural features is a viable approach to augment performance. Listings featuring one to two bedrooms and one bathroom are well-suited for small groups or individuals looking for a compact space. Adding additional beds for larger parties can boost occupancy without significantly increasing costs. These modifications align with market preferences and can improve both affordability and appeal, thereby increasing booking rates.

The research possesses multiple limitations. The analysis initially omitted time-sensitive factors, including seasonality and holiday effects, hence limiting the model's capacity to account for short-term market swings. Secondly, geographic variables were excluded to guarantee cross-market applicability; nevertheless, this diminishes sensitivity to regional price variations, such as differences between urban centres and suburban regions. Third, the dataset failed to consider host quality or property amenities, both of which could substantially affect pricing.

Subsequent research should expand the modelling framework to incorporate host- and amenity-level variables, along with temporal and geographic dimensions. Would improve predictive accuracy and provide a more comprehensive understanding of market dynamics. The development of more robust and competitive pricing strategies would be facilitated by broadening the research to include occupancy rate projections and incorporating data from competing rental platforms. Enhancing revenue and mitigating the risks associated with occupancy fluctuations would benefit both hosts and platforms.

References

[1] Roblek V, Stok ZM, Mesko M. Complexity of a sharing economy for tourism and hospitality. In: Faculty of Tourism and Hospitality Management in Opatija, Biennial International Congress. Tourism & Hospitality Industry [Internet]. University of Rijeka, Faculty of Tourism & Hospitality Management; 2016. p. 374. Available from: https: //www.researchgate.net/publication/301612962_Complexity_of_a_sharing_economy_for_tourism_and_hospitality

[2] Janssen E. Changes in pre- and post-pandemic pricing decision factors: an overview of South Africa’s luxury accommodation sector. Res Hosp Manag [Internet]. 2021; 11(1): 37–44. Available from: https:

//www.tandfonline.com/doi/abs/10.1080/22243534.2020.1867385

[3] Wang D, Nicolau JL. Price determinants of sharing economy based accommodation rental: a study of listings from 33 cities on Airbnb.com. Int J Hosp Manag [Internet]. 2017; 62: 120–31. Available from: https: //doi.org/10.1016/j.ijhm.2016.12.007

[4] Gibbs C, Guttentag D, Gretzel U, Morton J, Goodwill A. Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. J Travel Tour Mark [Internet]. 2017; 35(1): 46–56. Available from: https: //doi.org/10.1080/10548408.2017.1308292

[5] Zervas G, Proserpio D, Byers JW. A first look at online reputation on Airbnb, where every stay is above average. Mark Lett [Internet]. 2020; 32(1): 1–16. Available from: https: //doi.org/10.1007/s11002-020-09546-4

[6] Benítez-Aurioles B. Why are flexible booking policies priced negatively? Tour Manag [Internet]. 2018; 67: 312–25. Available from: https: //doi.org/10.1016/j.tourman.2018.02.008

[7] Li J, Moreno A, Zhang D. Agent behavior in the sharing economy: evidence from Airbnb. SSRN Electron J [Internet]. 2015. Available from: https: //doi.org/10.2139/ssrn.2708279

[8] Gunter U, Önder I. Determinants of Airbnb demand in Vienna and their implications for the traditional accommodation industry. Tour Econ [Internet]. 2017; 24(3): 270–93. Available from: https: //doi.org/10.1177/1354816617731196