

Predicting Airbnb Prices: A Study of Key Features Using Machine Learning

Mintao Wang

University of Washington, Seattle, USA
mwang36@uw.edu

Abstract. Short-term rentals have become more popular in the housing market, driven by platforms like Airbnb, which have largely transformed how people engage with vacation stays and temporary housing. This shift reflects a growing demand to understand key drivers for rental rates. In order to examine various factors influencing Airbnb rental costs, this study analyzes 74,111 Airbnb public data sets. The number of bedrooms, guest capacity, bathrooms, and review score ratings were determined to be the main predictors of the rental price of the property through the use of linear regression and decision tree modeling. The linear regression model achieved an R^2 of 0.3567, the decision tree model improved the prediction with an R^2 of 0.4384, while the random forest model further strengthened the predictability with an R^2 of 0.4605. The findings suggest that rental prices may increase significantly by adding more rooms to accommodate more guests and maintaining a higher review score. These findings offer crucial information for policymakers to comprehend the economic effect of short-term rentals, as well as insightful advice for homeowners looking to maximize listings and travelers making well-informed decisions.

Keywords: Airbnb, price determinants, linear regression model, decision tree model, random forest model

1. Introduction

The sharing economy has fundamentally changed the hospitality industry, with property owners renting out their properties to travelers worldwide through platforms such as Airbnb [1]. Since its inception, Airbnb has quickly grown to become a leader in the short-term rental industry, offering a variety of accommodations, including studios, restaurants, single-family homes, and apartments. Research shows that as short-term accommodations like Airbnb become available for people to choose from, the market has been negatively impacted by the trend of increases in rent [2]. A price prediction model becomes especially important for many stakeholders. For example, property owners who want to set appropriate and competitive prices and travelers who want to make intelligent choices must understand what factors influence the rent [3]. The primary goal of this study is to examine the factors that influence property values and investigate the relationship between these factors and the prices of accommodations using data obtained from the public Airbnb repository.

Previous research [4] has examined numerous aspects of Airbnb pricing, with a focus on geography, advertisement, tourism, and the quality of individual units. Yet, researchers must investigate deeper into how different listing standards interact to establish pricing. The dataset in this research consists of practically 74,111 records offered by Airbnb over a fixed period of time, which serves as a good platform for further statistical research. Each of the characteristics within the dataset has the potential to affect the rental price of Airbnb listings significantly.

This study uses decision trees and linear regression models to identify and assess the impact of these attributes on Airbnb pricing. Compared to machine learning as used in many research [4,5], decision trees and linear regression models are easier to interpret. Decision trees capture non-linear relationships which is important for this research.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the dataset used and the data cleaning process. Section 3 outlines the methodology, including linear regression, decision tree, and random forest models. In Section 4, the results of the analysis are presented, followed by a discussion of the key findings and limitations in Section 5. Finally, Section 6 concludes the study by summarizing the insights and suggesting directions for future research.

2. Dataset

The dataset used in this study was sourced from a third-party data provider that curates publicly available information from Airbnb listings across various cities. The data contains 74,111 records from Airbnb listings and was retrieved from Kaggle, including eleven attributes such as the listing ID, room type (e.g., entire home/apartment), and the number of accommodations, bathrooms, and bedrooms. The dataset also includes cancellation policies, cleaning fees, instant booking availability, and review scores. For each listing, the price was represented as a log-transformed price variable, which serves as a key feature in the analysis.

In the data cleaning process, several steps were taken to make sure the data was consistent and complete for use. Roles with missing values are either removed or replaced with default values, such as the rating scores from guest reviews. Data types were standardized by converting options for immediate booking availability into a simple yes/no format. Cancellation policies were given numerical values according to strictness, and missing values in columns related to property features like the number of bathrooms, bedrooms, and guest capacity were filled based on logical assumptions.

The goal of this study is to analyze the factors that influence Airbnb listing prices and develop a predictive model to predict rents based on features of accommodations.

3. Method

3.1. Linear regression

Linear regression is a basic and commonly used model that can describe the linear relationship between a dependent variable and independent variables [6]. The biggest advantages are its simplicity and interpretability. The linear regression model works well in this study as it identifies the linear relationship between factors that influence the listing price.

The model aims to minimize the Residual Sum of Squares (RSS), which measures the difference between the observed and predicted values. The RSS is calculated as follows:

$$RSS(\beta) = \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$$

Where:

y_i is the actual value of the i -th sample

β_0 is the intercept of the regression model

β_j represents the coefficients for the j -th feature

x_{ij} is the value of the j -th feature for the i -th sample

n is the number of samples

p is the number of features

One key limitation of linear regression is that it assumes a linear relationship between the variables, which may not be the case in more complex datasets.

3.2. Decision tree

To address the limitation of linear regression in identifying non-linear relationships, the study uses decision trees to offer a more flexible alternative [7]. Decision trees divide the data into branches of the tree by selecting the feature and split point that minimizes the error at each node. The objective is to minimize the Mean Squared Error (MSE) at each split:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

The total MSE for a split is calculated as:

$$Total\ MSE = \frac{N_{left}}{N} \cdot MSE_{left} + \frac{N_{right}}{N} \cdot MSE_{right}$$

Where:

N is the total number of samples.

y_i is the actual value of the i -th sample.

$\hat{y}_i$ is the predicted value of the i -th sample.

N_{left} and N_{right} are the number of samples in the left and right nodes, respectively.

MSE_{left} and MSE_{right} represent the MSE for the left and right nodes, respectively.

Decision trees excel in capturing non-linear relationships, which linear regression struggles with. However, decision trees are prone to overfitting, especially when the tree becomes too deep.

3.3. Random forest

To address overfitting in decision trees, the Random Forest model combines predictions from multiple trees, reducing variance and improving generalization by using random data subsets [8]. For regression, the final prediction is the average of the tree predictions:

$$\hat{y}(x) = \frac{1}{T} \sum_{i=1}^T \hat{y}_i(x)$$

Where:

$\hat{y}(x)$ is the final prediction for sample x

T is the total number of trees in the forest

$\hat{y}_i(x)$ is the prediction from the i -th tree

Random forests address both non-linear relationships and overfitting by averaging the results of multiple decision trees, which reduces the risk of overfitting and improves the model's generalization to unseen data.

4. Results

4.1. Linear regression model results

The linear regression analysis across different room types ('Entire home/apt', 'Private room', and 'Shared room') focused on evaluating the coefficients of the features to understand their influence on the log-transformed price predictions. The model, trained by minimizing the residual sum of squares, yielded a MSE of 0.3296 and an R^2 of 0.3567, indicating it explained about 35.67% of the variance in the log-transformed prices.

4.1.1. Entire home/apt

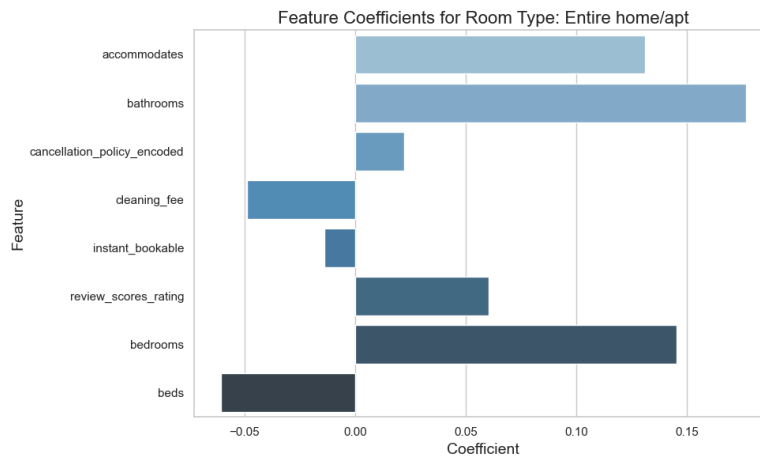


Figure 1. Feature coefficients for the room type 'Entire home/apt'

For 'Entire home/apt', the linear regression model showed the number of bathrooms and bedrooms as the most significant predictors of price, with positive coefficients of 0.1767 and 0.1454, respectively. Accommodates also had a positive impact (0.1311), while the number of beds exhibited a negative coefficient (-0.0609). This suggests that an increase in beds might correlate with a lower price, potentially due to layout or quality considerations. The R^2 for this room type was 0.3663, indicating that the model explained 36.63% of the variance, providing a moderately good fit for predicting prices in this room category.

4.1.2. Private room

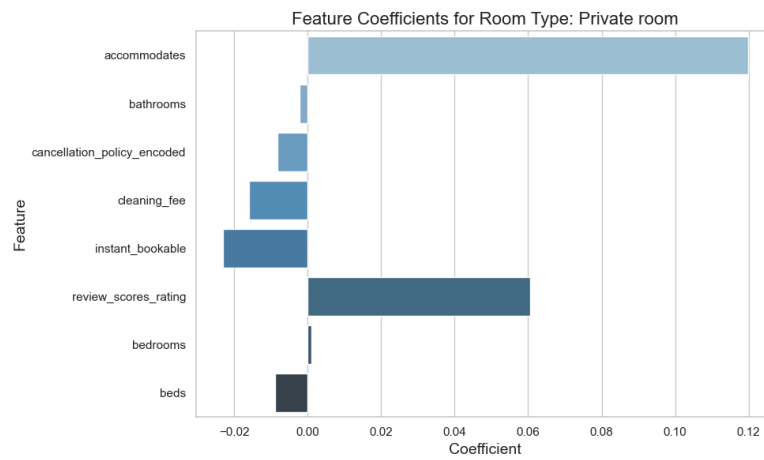


Figure 2. Feature coefficients for the room type 'Private room'

For 'Private room', accommodates had the largest positive coefficient (0.1198), with review scores rating also positively influencing predicted prices (0.0606). The number of beds showed a slight negative impact (-0.0088). Despite these reasonable feature coefficients, the model struggled to explain the variance in prices, with an R^2 of only 0.0653. This low R^2 value suggests that the model provided a poor fit for predicting prices in private rooms, indicating the need for further refinement or additional features.

4.1.3. Shared room

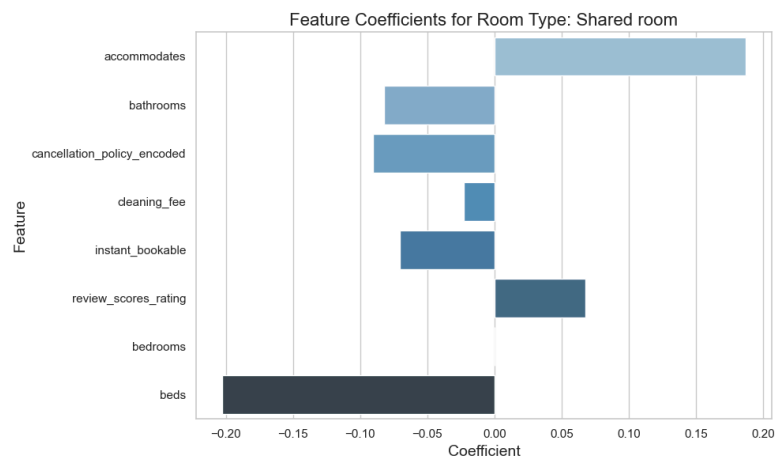


Figure 3. Feature coefficients for the room type 'Shared room'

In 'Shared room' listings, accommodates had the largest positive coefficient (0.1870), while the number of beds had a notably negative coefficient (-0.2026), indicating that more beds correspond to lower prices in shared rooms. The R^2 for this room type was also low at 0.0685, showing that the linear regression model had difficulty explaining the variance in shared room prices.

4.1.4. Cross-validation

To evaluate the performance of this mode across different room types, cross-validation was conducted. The results showed that the model performed best for 'Entire home/apt', with an average MSE of 0.2375 and an R^2 of 0.3607. However, for 'Private room' and 'Shared room', the R^2 values were significantly lower (0.0651 and 0.1422, respectively), indicating that the model had limited success in capturing the variance in prices for these room types. The lower R^2 values may be attributed to relatively smaller sample sizes for these two types of rooms.

4.2. Decision tree model results

The decision tree model improved upon the linear regression model by capturing the non-linear relationships. The study limits the maximum depth of the tree to 10 and sets the minimum number of samples required to split a node at 20. The decision tree achieved an MSE of 0.2878 and an R^2 of 0.4384, showing that it explained 43.84% of the variance in the data. This improvement over the linear regression model suggests that the decision tree was better able to capture complex relationships between features and price. However, the model still showed some sensitivity to overfitting, particularly with deeper trees.

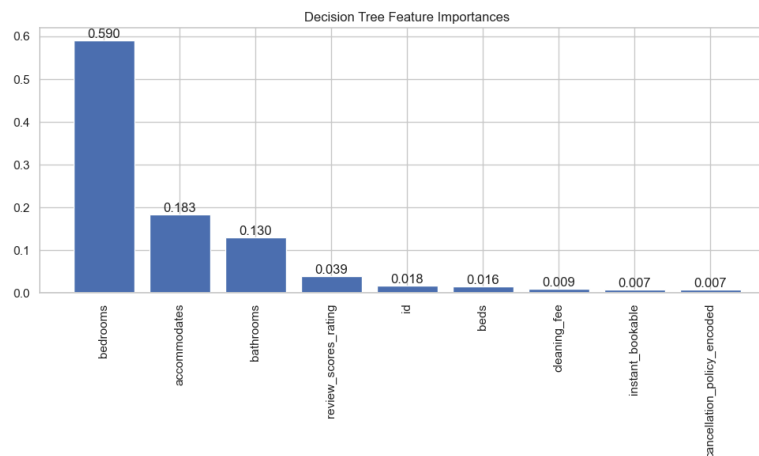


Figure 4. Feature importance plot for the Decision Tree model

The feature importance values, as illustrated in Figure 1, show which features had the most significant impact on the model's predictions. The most important feature was the number of bedrooms, with an importance score of 0.590, indicating that this feature contributed the most to predicting the price. The second most influential feature was accommodates, with an importance of 0.183, followed by bathrooms with an importance of 0.130. These three features were the key drivers of the decision tree's predictions.

Other features, such as review scores rating (0.039), number of beds (0.016), and cleaning fee (0.009), had less impact on the model. Additionally, features like instant bookable (0.007) and cancellation policy encoded (0.007) showed minimal importance, suggesting that these factors did not play a significant role in predicting the price for this dataset.

4.3. Random Forest model results

The Random Forest model was applied to predict the log-transformed price of listings, and the feature importance was extracted to understand which variables had the most significant impact on the model's predictions. The model resulted in an MSE of 0.2764 and an R^2 of 0.4605, indicating that it explained 46.05% of the variance in the log-transformed prices. It showed that the random forest outperformed both the linear regression and decision tree models, offering a more reliable prediction framework and reduced overfitting and improved generalization.

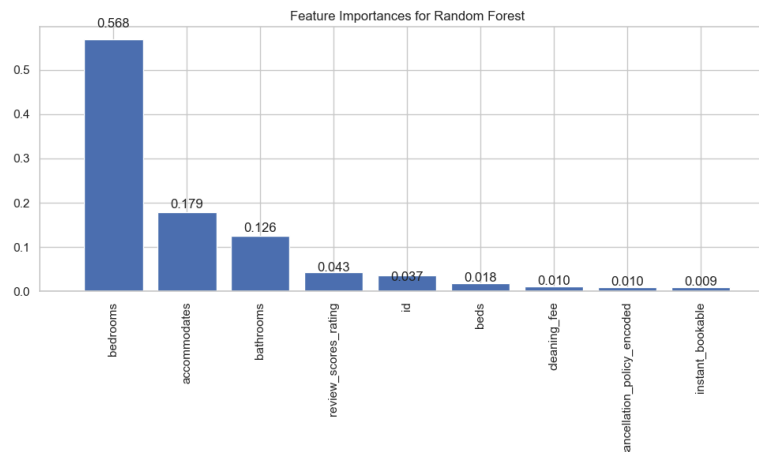


Figure 5. Feature importance plot for the Random Forest model

As shown in figure 5, the number of bedrooms was the most important feature, with a feature importance score of 0.568. The accommodates feature was the second most important feature and had a score of 0.179, followed by the number of bathrooms.

Other features like review score rating, beds, and cleaning fees had comparatively lower importance, indicating that they affect less to the overall price. Similar to the decision tree result, features such as cancellation policy and instant bookable had near zero scores, which means they had minimal influence on the outcome and played little role in predicting prices.

5. Discussion

The results of this study demonstrate many important factors influencing Airbnb listing prices across various room types. While the three models provided insights into the key factors of price, it is necessary to consider several limitations and implications of the results.

5.1. Key findings

The analysis found that three features influenced prices in all room types: the number of bedrooms, accommodations, and bathrooms, particularly for the 'Entire home/apt'. This result is consistent with expectations, as properties with more space and privacy typically charge higher prices. However, features such as cleaning fees, instant bookable, and cancellation policy had less impact on price.

The categorization of room types shows that the models performed best for 'Entire home/apt' and the coefficients suggested stronger relationships between the features and the price. For 'Private room' and 'Shared room', however, the models had much lower predictive power, which may be due to the limitations within the dataset.

5.2. Limitations

There are many limitations that affect the accuracy of the predictions in this study, as reflected particularly in the low R^2 values.

First, the dataset was relatively small, especially for 'Private room' and 'Shared room'. This may have limited the models' ability to identify the variance in these listings. The lack of data is likely the main reason that models had lower R^2 values for these room types. To improve the performance of the models, a larger and more diverse dataset is required, especially for room types with fewer listings.

Secondly, the dataset used in this study did not include geographic features, such as the location or neighborhood of the listing. Geographic features are important in predicting price. Airbnb homes located in highly sought-after areas like city centers or tourist attractions typically charge higher prices. The lack of these geographic factors likely led to the relatively low R^2 values across all models. Therefore, it is important to incorporate location-based data to help capture variations in pricing strategies across different locations.

There are many other factors that could largely influence pricing that were not included in the dataset of this study. Features like host experience, availability, and seasonal pricing strategies were not present in the dataset. These missing variables could explain some of the unexplained variance in the models and might help improve R^2 scores if included.

5.3. Implications for future research

The findings from this study help build a foundation for understanding the main factors influencing Airbnb listing prices. However, future research should consider including more listings and critical variables such as geographic location. It is also important to explore more complex models to better capture the complex relationships between features and price, particularly for room types like 'Private room' and 'Shared room'.

Lastly, as Airbnb markets continue to evolve, other external factors, such as travel trends and economic conditions, may play an increasingly important role in determining pricing. Therefore, future research should include dynamic and temporal variables to provide a more comprehensive view of the factors affecting Airbnb prices.

6. Conclusion

Ultimately, the study analyzed the data from 74,111 listings in order to investigate multiple variables that drive Airbnb postings. Three models have been used to discover the critical aspects influencing the sharing economy price. Results have shown that the number of bedrooms, guest capacity, bathrooms, and review score rating all have a major impact on pricing for rentals, with bedrooms being the most significant.

Homeowners can maximize their houses' earnings by focusing on the range and utility of their offers. For example, as shown in the study, larger accommodations tend to have higher listing prices, so homeowners can consider adding more bedrooms and bathrooms to enhance property value. At the same time, tenants may utilize this data to make more informed selections, balancing essential features versus costs. For example, the study shows that an increase in the number of beds tends to lower the rental price, and if the capacity matters the most for tenants, they can look for accommodations with extra beds instead of extra bedrooms. The approach laid out in the article may

provide a more reliable evaluation of the model's performance, enabling it to account for the majority of fluctuations in prices.

The examination nevertheless indicates that identifying the cause of regional and market differences remains challenging. Future research may combine sophisticated modeling techniques with additional contextual variables in order to enhance the accuracy of pricing forecasts in the fluctuating Airbnb market.

References

- [1] Hamari, J., Sjöklint, M. and Ukkonen, A. (2016), The sharing economy: Why people participate in collaborative consumption. *J Assn Inf Sci Tec*, 67: 2047-2059. <https://doi.org/10.1002/asi.23552>
- [2] Barron, K., Kung, E., & Proserpio, D. (2021). The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Science*, 40(1), 23–47. <https://pubsonline-informs.org.offcampus.lib.washington.edu/doi/10.1287/mksc.2020.1227>
- [3] Chen, Yong & Xie, Karen. (2017). Consumer Valuation of Airbnb Listings: A Hedonic Pricing Approach. *International Journal of Contemporary Hospitality Management*. 29. 10.1108/IJCHM-10-2016-0606.
- [4] Raul Perez-Sanchez, V., Serrano-Estrada, L., Marti, P., & Mora-Garcia, R.-T. (2018). The What, Where, and Why of Airbnb Price Determinants. *Sustainability*, 10(12), 4596-. <https://doi.org/10.3390/su10124596>
- [5] Lorenz, F., Willwersch, J., Cajias, M., & Fuerst, F. (2023). Interpretable machine learning for real estate market analysis. *Real Estate Economics*, 51, 1178–1208. <https://doi.org/10.1111/1540-6229.12397>
- [6] Olive, D. J. (2017). *Linear regression*. Springer. <https://doi.org/10.1007/978-3-319-55252-1>
- [7] Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group.
- [8] Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>