Predicting Prices and Quantifying Popularity For Airbnbs in Barcelona City

Adam Amar, Oore Fasawe

September 2024

1 Abstract

In the rapidly growing Airbnb market, accurately pricing rental properties is crucial for maximizing revenue and ensuring long-term success. This study aims to develop a model that predicts the optimal rental price for Airbnb properties in Barcelona based on key features, including number of rooms, accommodation capacity, reviews, room type, neighborhood, and number of bedrooms. Additionally, the model generates a popularity index to estimate how attractive a given property configuration will be, independent of price. This research is valuable for property owners looking to establish competitive base prices for their listings while understanding potential market demand. Utilizing an extensive training dataset from airbnb.com, which includes hundreds of Airbnb listings in Barcelona, we applied linear regression and random forest regression to identify the most influential factors on pricing and property popularity. To enhance accuracy, we introduced derived features like proximity to landmarks, bedroom density, and amenities count. Through this research, we aim to provide a practical framework for optimal Airbnb property pricing in Barcelona, with potential scalability to other markets worldwide.

2 Introduction

The existing body of literature on Airbnb pricing has yielded promising results. For example, Pouya et al. [2] employed a combination of machine learning, deep learning, and natural language processing techniques to predict Airbnb house prices. Numerous other studies have explored various methods to address price prediction, but the highest R^2 value reported was achieved by Pouya et al., who obtained an impressive 0.69 using Support Vector Regression (SVR). This result highlights the potential of advanced predictive models in this domain. However, the generalization of such findings remains uncertain, as variations in datasets, output variables, and property markets can lead to inconsistent results.

Building on these prior studies, our research seeks to enhance and expand the methodologies explored in the literature. Specifically, we aim to uncover relationships using simple linear regressions with different features and a Random Forest regression while also developing a popularity index to estimate the demand for a listing. This dual focus not only addresses pricing but also offers insights into market dynamics. Current pricing strategies often rely heavily on average market rents as proxies, which can result in inaccurate and suboptimal pricing decisions. Our approach aims to provide a more nuanced and data-driven alternative, unlike conventional pricing strategies, which often rely on market averages.

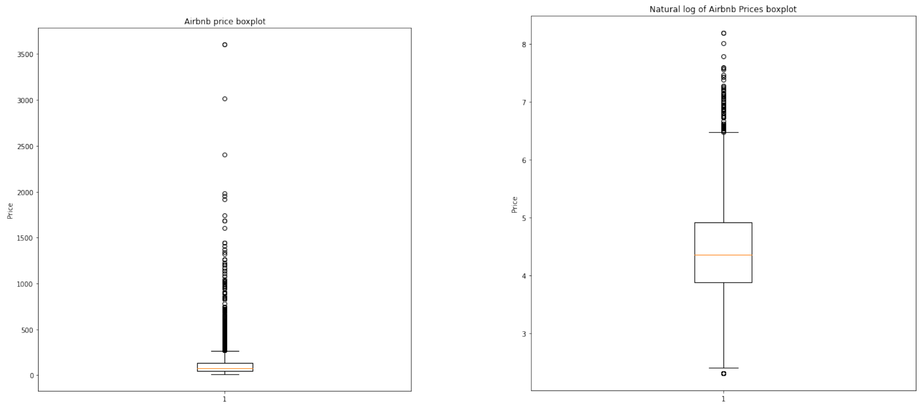
The relevance of this study is heightened in the context of Barcelona, a city facing significant challenges in its tourism sector. Rising tensions between local residents and an overwhelming influx of tourists [3] have led to new regulations aimed at curbing tourist numbers [5]. These measures are likely to impact the Airbnb market by reducing the supply of available rentals. For property owners, this presents a mix of risks and opportunities. On one hand, stricter regulations could increase the value of properties in prime locations by reducing competition. On the other hand, local dissatisfaction and unrest may diminish the appeal of certain areas, making it vital for hosts to strategically price their listings and identify features that attract guests.

In this evolving landscape, our project offers critical value by equipping property owners with actionable, data-driven insights, such as pricing optimization based on location and amenities. These insights can help them adapt to regulatory changes and optimize both the pricing and popularity of their properties in a market that is becoming increasingly complex and regulated [4].

3 Methods

The dataset we used includes information on 10,441 Airbnb listings across various neighborhoods in Barcelona [1]. Each data point consists of details such as room ID, room type, neighborhood, reviews, overall satisfaction, accommodation capacity, number of bedrooms, price, minimum stay, and geographic coordinates (latitude and longitude). Our first goal is to develop a model that predicts the best price for an Airbnb property given its unique combination of features. We used 60% of the dataset, comprising 6265 Airbnb homes, to train our model for predicting both the price and popularity of the properties. Another 20% was allocated for testing our predictions, while the final 20% was set aside for validation. This standard data split, commonly employed in scientific research, enabled us to evaluate the model's performance on unseen data effectively.

Cleaning the data: Our dataset initially consisted of 19 columns. Of these, "Country," "Borough," "Min stay," and "Bathrooms" were removed due to missing data throughout. Additionally, "room id," "name," "host id," "survey id," and "last modified" were excluded as they were irrelevant to the connections we aimed to explore. "city" and "location" were also dropped since the dataset exclusively focused on Barcelona, rendering "city" redundant, while "location" was better represented through latitude and longitude, which allowed for more meaningful geographic analysis. This left us with 8 columns to analyze: 'room type,' 'neighborhood,' 'reviews,' 'overall satisfaction,' 'accommodates,' 'bedrooms,' 'price,' 'latitude,' and 'longitude.' Each of these columns contributed valuable information to our analysis. 'room type' distinguished between shared and whole properties, enabling tailored analysis for different property types. 'Neighborhood' provided critical location-based insights into demand patterns and property suitability across areas. 'Overall satisfaction' captured guest experiences, offering guidance on high-performing areas. 'Price', which represents the rent price of 1 night, is the target variable, while 'bedrooms' and 'accommodates' provided information about property capacity. Finally, latitude and longitude were retained to calculate distances to key hotspots [2], which could significantly affect pricing. To mitigate this issue, we applied a logarithmic transformation to the price data, following a similar approach to Pouya et al. [2]. This transformation effectively reduced the influence of outliers, resulting in a more balanced dataset. The effectiveness of this adjustment was confirmed using box plots, which visually demonstrated the refinement of the data as in Figure 1, ensuring it was suitable for analysis.



(a) Before Transformation

(b) After Transformation

Figure 1: Comparison of box plots before and after transformation.

Derived feature generation: To capture more nuanced relationships and enhance the model's predictive power, additional derived metrics were calculated alongside the raw features. These metrics were carefully chosen to contextualize key variables, enabling more meaningful comparisons across listings. 'Bedroom density' was introduced to provide insight into how efficiently space is utilized within a property. This feature offers a standardized measure for evaluating affordability and demand across listings of varying sizes. Similarly, 'reviews per person' was calculated by dividing the total number of reviews by the maximum occupancy of each listing, allowing for a fair comparison of customer engagement regardless of property size. To further assess quality, 'satisfaction per person' was derived by adjusting satisfaction ratings relative to the number of people a listing can accommodate, offering a granular perspective on the customer experience. Drawing inspiration from previous research [2] that highlighted the importance of proximity to key hotspots, we incorporated haversine distances from prominent landmarks, including Plaça de Catalunya (a central hub), Sagrada Família (a major tourist attraction), Camp Nou (a renowned sports stadium), and Barceloneta Beach (a popular coastal area). These features capture the geographical attractiveness of a listing and its proximity to high-demand destinations, By including these derived metrics, the model achieves a richer representation of the data.

Linear Regression: Correlation matrices were constructed to identify the strength and direction of associations between features.

The correlation matrix in Figure 2 highlights that price has the strongest positive correlation with the number of bedrooms and accommodatees, while showing a negative correlation with satisfaction per accommodatee. This is intuitive, as properties with more bedrooms and accommodatees tend to be larger and more expensive, whereas lower satisfaction per accommodatee often signals poorer living conditions, diminishing a property's value.

With only 12 features—relatively few compared to those used in other studies—a linear regression model with all features was chosen as a baseline. Against, a random forest regression models was also explored and another linear regression was performed on the top 7 features alone. This approach offered a simple yet effective method to evaluate feature importance and establish a foundational

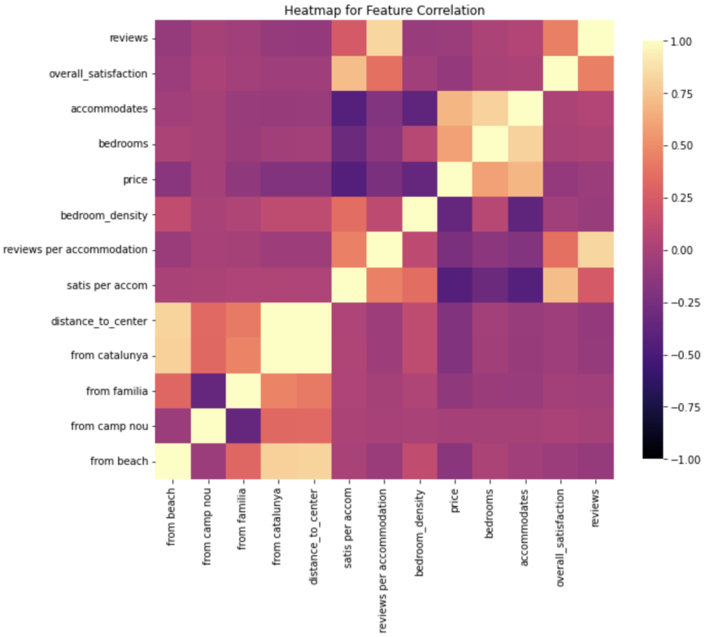


Figure 2: Correlation Heatmap

understanding of the data.

Popularity estimation: This will be approached using a weighted aggregation of key features. The popularity index is modeled as a combination of average rating, number of reviews, reviews per month, and annual listing availability. To ensure comparability across features, the data was normalized using Min-Max scaling, and weights were assigned based on their relative importance. Specifically, average rating was given a weight of 40% due to its strong influence on desirability, while number of reviews accounted for 30%, reflecting its role in building trust. Reviews per month were weighted at 20%, as they capture recent engagement trends, and annual availability contributed 10% to account for market consistency. The resulting weighted scores were aggregated to calculate the popularity index, providing a balanced representation of quality, activity, and accessibility.

4 Results

The models developed for this study demonstrated varying degrees of success in predicting rental prices and estimating property popularity, offering insights into the factors influencing the Airbnb market in Barcelona. The results highlight both the strengths of the predictive approaches and the challenges inherent in modeling such complex relationships.

Price Prediction: The baseline linear regression model achieved an R^2 of 0.503 on the training data and 0.509 on the test data, with an F-statistic of 645.2, indicating strong significance. The RMSE was 0.28 on the training set and 0.27 on the test set, suggesting consistent performance. These results indicate that the model explains about half of the variation in prices, identifying key trends such as higher prices for listings by superhosts, larger accommodations, and those offering more amenities. Negative impacts were observed for room-neighborhood interaction and bedroom density, while reviews per month and listing age showed smaller positive effects. However, the weak influence of some variables, such as bathrooms and number of reviews, points to areas for improvement.

The random forest regression model demonstrated high predictive accuracy, with an R^2 of 0.95 on training data but a reduced R^2 of 0.66 on testing data, reflecting the typical overfitting tendencies of ensemble models. RMSE dropped from 0.03 on training data to 0.19 on testing data. The "accommodates" feature emerged as the most significant predictor of price, while reviews and room-neighborhood interactions also played key roles. These results highlight the model's ability to capture non-linear relationships but suggest limited generalization due to dataset complexity.

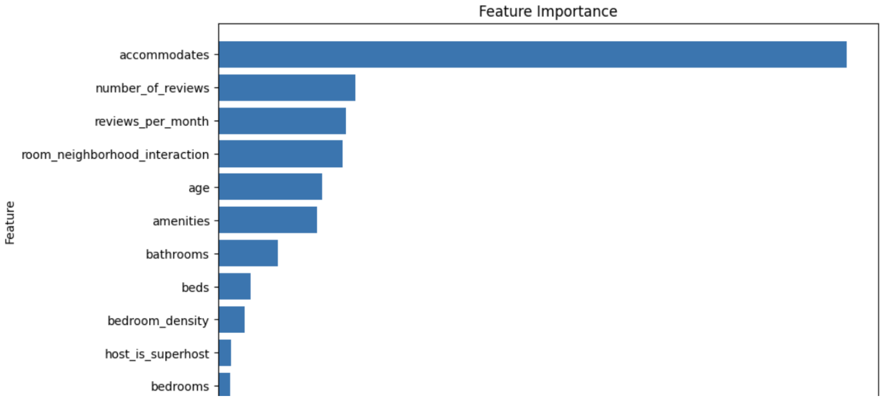


Figure 3: Graph of feature importance

Using linear regression with the top seven features yielded an R^2 of 0.48 and an RMSE of 0.28 across both training and testing datasets, indicating consistent performance and minimal overfitting. This simplified model provided comparable results to the baseline, demonstrating the effectiveness of feature selection in reducing complexity while maintaining predictive power. However, the reduced R^2 and high RMSE suggest that further refinement or non-linear approaches may better capture the nuanced relationships driving Airbnb pricing.

Popularity Estimation: The popularity index showed a mean of 0.427 with a relatively narrow distribution, as indicated by a standard deviation of 0.061. The median (0.434) was close to the mean, suggesting a balanced spread of scores without strong skewness. Most listings fell between 0.40 and 0.47, while a few exceptional listings reached as high as 0.70, marking them as clear standouts. Correlation analysis confirmed that ”number of reviews” had a modest positive correlation (0.086) with the popularity index, aligning with its weighted importance. ”Review scores rating” and ”reviews per month” were only weakly related to the index (0.011 and -0.004 respectively), while ”availability 365” showed a slight negative relationship (-0.021). These findings suggest that while all factors contribute, the number of reviews may have played a more consistent role in differentiating listings. Together, the descriptive statistics and correlation patterns imply that the popularity index effectively balances multiple inputs. Although review quality, frequency, and availability each contribute, the total number of reviews appears to influence the final score most strongly under the current weighting. The result is a nuanced index that captures key aspects of listing desirability and engagement.

5 Discussion

The results of this analysis offer valuable insights into the factors influencing Airbnb pricing and highlight opportunities for further refinement in predictive modeling. The findings suggest that while the linear regression model captures some meaningful relationships, its limited ability to generalize underscores the complexity of the pricing dynamics in the Airbnb market. This work contributes to understanding the relative importance of features such as property size, neighborhood, and guest satisfaction, which can inform property owners in setting competitive prices and improving listing appeal.

However, several limitations must be acknowledged. The model's reliance on a linear approach may oversimplify the intricate, non-linear relationships between features and price. Additionally, the dataset's limited number of features excludes other potential factors, such as dynamic seasonal trends, host experience, or marketing strategies, which could provide a more comprehensive view. Future work could address these limitations by incorporating more advanced modeling techniques, such as tree-based methods or neural networks, to capture complex interactions. Expanding the dataset with additional features, such as temporal data or sentiment analysis of reviews, could further improve model accuracy and applicability. Overall, this project lays the groundwork for more robust predictive models and offers a practical starting point for stakeholders seeking data-driven insights into the Airbnb market.

References

[1] AirBnb. "Airbnb Data Collection: Get the Data". In: (2022).

[2] "Airbnb Price Prediction Using Machine Learning and Sentiment Analysis". In: (2019).

[3] Monica Pitrelli — CNBC. *Protesters in Spain told tourists to 'go home.' Instead more arrived.* 2024. URL: `https://www.cnbc.com/2024/10/14/after-anti-tourism-protests-spain-receives-record-number-of-travelers.html`.

[4] Nina Godard — Le Monde. *Calculating the mixed impact of rent controls on French cities.* 2024. URL: `https://www.lemonde.fr/en/money-investments/article/2024/04/19/calculating-the-mixed-impact-of-rent-controls-on-french-cities_6668876_102.html?utm_source=chatgpt.com`.

[5] Sandrine Morel — Le Monde. *Barcelona property owners organize against tourist rental ban.* 2024. URL: `https://www.lemonde.fr/en/europe/article/2024/09/16/barcelona-property-owners-organize-against-tourist-rental-ban_6726220_143.html`.