Group 1
Predictive Analytics
12 December 2024

Predicting Airbnb Prices Across Europe

__Introduction (219/250)__

The short-term rental market, led by platforms like Airbnb, has transformed travel and real estate investment globally. Accurate pricing predictions have significant value for stakeholders aiming to optimize revenue and identify profitable properties. Key factors such as location, room type, customer reviews, and neighborhood characteristics play crucial roles in determining prices. For hosts, understanding these dynamics can drive occupancy rates and revenue, while investors can leverage insights to acquire high-potential properties. This project focuses on major tourist cities in Europe, where varying economic and cultural factors interact to influence Airbnb pricing. By analyzing datasets with features such as room type, person capacity, and location, we aim to uncover critical metrics that drive significant Airbnb price changes. Our resulting research question is: Based on features like room type, location, and ratings, can we predict Airbnb prices across major European cities? Leveraging predictive analytics, we will explore complex interactions among these variables to develop robust predictive pricing models. Our findings will contribute to both research and industry practice by offering actionable insights and strategies for stakeholders. Hosts and property managers can use these insights to refine pricing strategies, while real estate investors can make informed business decisions. This project highlights the importance of data analytics in addressing challenges in the short-term rental market, enabling stakeholders to navigate a competitive industry effectively.

__Research & Background (556/600)__

Due to the rise of digital platforms, changing travel habits, and greater internet access, the short-term rental market has experienced rapid growth in recent years. This study focuses on major European tourist cities to examine the current trends, challenges, and opportunities in the short-term rental industry. Platforms like Airbnb and Booking.com have revolutionized the hospitality sector by creating new possibilities for stakeholders such as property owners and travelers. According to researchers at Rey Juan Carlos University, these platforms have encouraged peer-to-peer accommodation sharing, enabling individuals to rent out their properties on a short-term basis. Data from AirDNA shows that short-term rental listings grew by 9.4% in 2021, with an estimated 20.5% increase in 2022. Moreover, Key Data Dashboard reports that properties with strong reviews and popular amenities see occupancy rates increase by up to 15%. These statistics further highlight the importance of understanding the factors that affect pricing in this flourishing industry.

European cities like Paris, London, Madrid, and Amsterdam have seen substantial increases in short-term rental activity, which has had both positive and negative impacts. While these rentals boost local economies and tourism, they also raise concerns about housing affordability and urban regulation. The European Union has introduced laws to promote transparency in this industry, requiring platforms to verify property details and comply with local regulations. Cities have also implemented measures such as rental caps, zoning rules, and permit requirements to manage the industry's growth. For example, in Amsterdam, rental days are capped at 30 per year for hosts. These regulatory changes reflect the evolving environment of short-term rentals and their impact on local communities.

Understanding what drives rental prices is crucial for stakeholders such as property owners and investors. Research from Rey Juan Carlos University shows that factors like property size, location, local market conditions, and amenities play key roles in determining rental pricing. In addition to these static features, many platforms use dynamic pricing algorithms to adjust rates based on market demand and conditions in real-time. According to AirDNA, properties utilizing dynamic pricing strategies can see revenue increases of up to 40% compared to those with static pricing. These tools not only maximize revenue, but also highlight the importance of data analytics as a cornerstone of decision-making in the short-term rental industry.

Property managers are increasingly using analytics to track performance, optimize pricing, and forecast demand trends. According to Key Data Dashboard, these tools help owners make informed decisions about their properties and improve the overall guest experience. For example, performance tracking enables property owners to identify patterns in bookings and adjust their strategies accordingly. Demand forecasting helps managers anticipate high-demand periods and set competitive prices, while tailoring to guest preferences tends to lead to better reviews and repeat bookings. Together, these tools are reshaping how the short-term rental industry operates, making data-driven decision-making a necessity for success.

This project focuses on identifying the most important factors affecting short-term rental pricing, particularly in Airbnb listings across Europe. By analyzing data from multiple cities, we aim to uncover underlying patterns that explain price variability. For instance, understanding how amenities like proximity to tourist attractions or property size contribute to pricing can provide actionable insights for stakeholders. We will use predictive modeling techniques to explore these features and identify the combination of features that best predict Airbnb prices across major European cities.

Methods & Architectural Design (448/400)

We analyzed 20 datasets of Airbnb listings from 10 major European cities, including Amsterdam, Berlin, and Paris, with separate datasets for weekday and weekend listings. They were sourced from Kaggle, curated from the work of Gyódi, K. and Nawaro Ł, and included features such as price (calculated for a two-night stay), room type, and attraction index (derived from TripAdvisor data). Data preprocessing involved merging the datasets and introducing “City” and “Weekend” columns for weekend listings. To address price skewness, we applied a natural log transformation. We removed non-essential columns, such as dataset indices, and checked for missing values. Feature dimensionality was reduced using *Principal Component Analysis (PCA)*, retaining 13 components that explained 79% of the data variability. This streamlined the dataset for further modeling.

We began our analysis with *Multi-Linear Regression (MLR)* to establish a baseline model and explore the linear relationships between price and predictor variables. Data preparation included encoding categorical variables, then splitting the data into training and test sets for evaluation. This model provided us with initial insights into how various features influence pricing. However, the presence of outliers and non-linear relationships affected performance, suggesting the use of more advanced models like *Gradient Boosting* and *Deep Neural Networks*.

We then implemented *Gradient Boosting (GB)*, an ensemble learning algorithm, to address the non-linear relationships and interactions among features that *MLR* could not capture. Data preparation was consistent with *MLR*, focusing on encoding categorical variables and scaling. This model constructs an initial model and iteratively improves it by minimizing the residual errors. Each iteration computes the residuals, selects a weak learner–a predictive model like decision tree (with small depth and low complexity)–to generate predictions for some pseudo residuals. It then compares these pseudo residuals with actual residuals to minimize the existing error resulting from previous iterations based on patterns discovered, which helps to prevent overfitting. Then the model updates by using a weighted combination of these weak learners. *Gradient Boosting* was particularly effective at handling feature interactions and variability across cities, making it a valuable enhancement over *MLR*.

Finally, we utilized *Deep Neural Networks (DNN)* in an attempt to further enhance performance by capturing high-dimensional relationships within the data. Similar to *MLR* and *GB*, data preparation included encoding categorical variables and scaling. Our *DNN* model consisted of multiple dense layers–288, 320, 384, and 32 units–followed by one output layer for price prediction. The model used *ReLU* activation across hidden layers and the Adam optimizer with a learning rate of 0.0002. To prevent overfitting, early stopping was added, halting training when validation loss plateaued for 10 consecutive epochs. While the *DNN* model provided valuable insights into feature interactions, its performance didn't surpass the *Gradient Boosting* model.

<u>*Results (388/400)*</u>

The results of our analysis allow us to compare the performance of the three models: *Multi-Linear Regression (MLR)*, *Gradient Boosting*, and *Deep Neural Network (DNN)*. For the *Multi-Linear Regression* model, we achieved a Mean-Squared Error (MSE) of 0.12 and adjusted R^2 of 0.6619 *(Table 1)*. The *Gradient Boosting* model outperformed the *Multi-Linear Regression* model with an MSE of 0.086 and an adjusted R^2 of 0.7621 *(Table 1)*. And the *Deep Neural Networks* model, while better than the *Multi-Linear Regression* model with an MSE of 0.105 and adjusted R^2 of 0.7126 *(Table 1)*, did not surpass *Gradient Boosting* in predictive accuracy. Overall, the *Gradient Boosting* model emerged as the most effective model, achieving the lowest MSE and highest adjusted R^2, indicating its predicted prices were the closest to the actual price and it was able to explain a greater percentage of the data variation *(Figure 2)*.

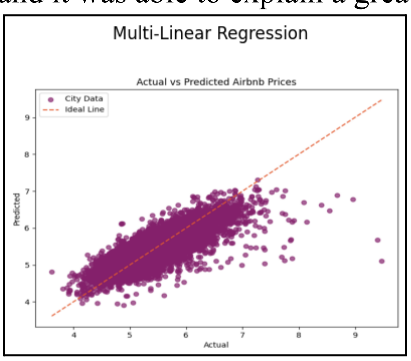


Figure 1

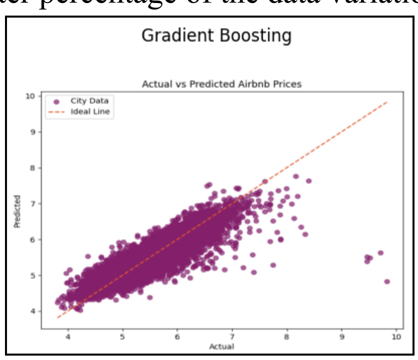


Figure 2

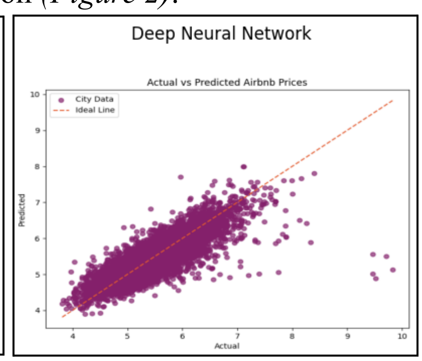


Figure 3

Performance of Models

Model	MSE	R^2	Adjusted R^2
Gradient Boosting	0.086	0.7628	0.7621
Deep Neural Networks	0.105	0.7136	0.7126
Multi-Linear Regression	0.12	0.6632	0.6619

Table 1

The *Gradient Boosting* model demonstrated strong performance at accurately predicting Airbnb prices, particularly in cities like Amsterdam, London, and Barcelona *(Figures 4-6)*. For these cities, this model achieved an MSE below 0.07 and an R^2 score of approximately 0.80. These results portray this model's optimal performance to account for city-specific variations and interactions between features.

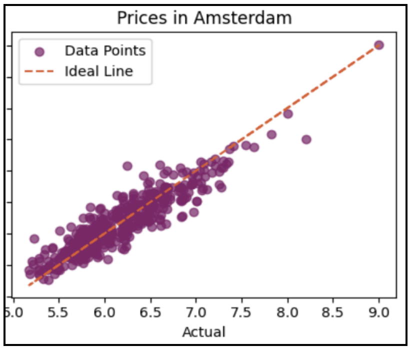


Figure 4

Figure 5

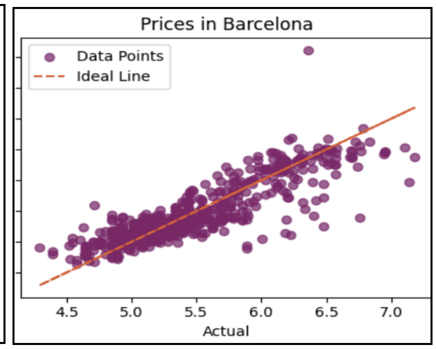


Figure 6

To further explore the *Gradient Boosting* model's success, we analyzed feature importance. Among the least influential features *(Table 2)* were room_type_Private_room, room_private, and weekend columns, each contributing less than 0.1% to the final predictions. And columns like multiple_rooms and host_is_superhost also had minimal impact, contributing approximately 0.05%.

Least Influential Features

Feature	Importance
room_type_Private room	0.001
room_private	0.0009
weekend	0.0009
multiple_rooms	0.0005
host_is_superhost	0.0005

Table 2

Most Influential Features

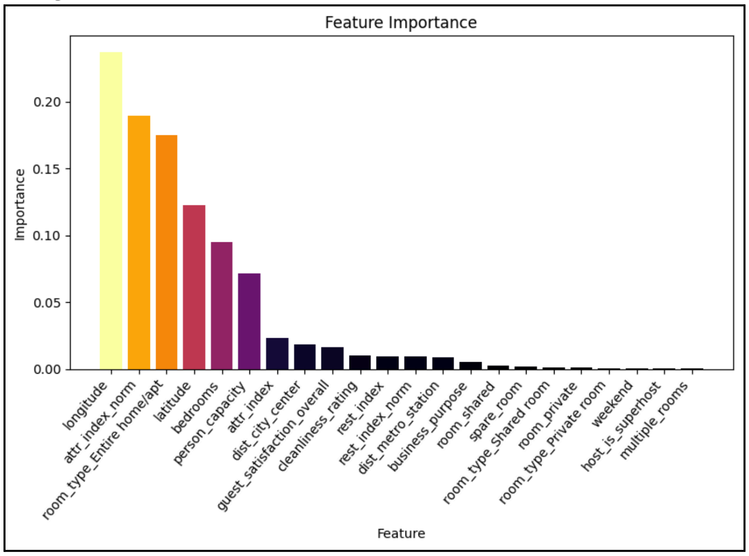
Feature	Importance
longitude	0.24
attr_index_norm	0.19
room_type_Entire home/apt	0.18
latitude	0.12
bedrooms	0.09

Table 3

On the other hand, the most influential feature *(Table 3)* was longitude (location) that contributed 24% to the final price predictions. Other significant features include attr_index_norm (a score representing how close the Airbnb listing is to tourist attractions), room_type_entire_home/apt, latitude, and bedrooms columns. These features contributed a little less to the final price prediction compared to the longitude column: 19%, 18%, 12%, and 9% respectively. These findings align with expectations, as proximity to attractions and property size are critical determinants of Airbnb pricing. By leveraging these insights, our *Gradient Boosting* model not only achieved high accuracy but also provides actionable interpretations of key factors that drive Airbnb prices across major European cities.

Interpretation & Conclusion (707/750)

Our analysis demonstrated that *Gradient Boosting* is the most effective model for predicting Airbnb prices across major European cities. *Gradient Boosting* outperformed both the *Multi-Linear Regression* and *Deep Neural Networks* models in accuracy and interpretability. *Gradient Boosting* successfully captured non-linear relationships and feature interactions that were overlooked by *Multi-Linear Regression*, and provided more actionable insights than *Deep Neural Networks*. As a result, these observations guided our decision to base the results and analysis primarily on the *Gradient Boosting* model, which demonstrated the most robust predictive performance. By leveraging this model, we identified the most influential features driving price variations, including location (longitude and latitude), proximity to attractions, whether the whole home or apartment was listed, and person capacity. These findings provide many business insights for stakeholders, enabling them to make data-driven decisions in a competitive market. On the other hand, features such as whether a host is a superhost or whether the room is private had minimal impact on pricing. This distinction between influential and non-influential features emphasizes the value of using advanced models like *Gradient Boosting* to refine business strategies.



These findings also provide us with many implications that extend beyond our academic interest: these findings provide practical applications for stakeholders such as hosts, property managers, investors, and travel agencies. For hosts and property managers, understanding key price determinants enables them to refine pricing strategies, market their properties more effectively, and maximize revenue. Property managers can apply these insights to optimize portfolios across various regions by focusing on high-performing properties and ensuring competitive pricing. Investors can benefit from these predictive analytics by identifying lucrative properties in desirable locations, which can guide acquisition strategies and portfolio diversification. Meanwhile, travel agencies can use these findings to understand market trends, tailor their marketing focus, and develop strategies that align with consumer preferences and the dynamic demand trends. These general insights set the stage for stakeholders to leverage specific recommendations and suggestions based on the most and least influential features.

The most influential features on price identified by our *Gradient Boosting* model–location, proximity to attractions, listing entire home or apartment, number of bedrooms, and person capacity–highlight actionable strategies for stakeholders. First, stakeholders should focus on emphasizing proximity to key areas in their listings to justify premium pricing. For example, the attraction index, reflecting the proximity to nearby attractions like landmarks, can be leveraged to promote properties in high-tourist areas. Properties listed as entire homes or apartments tend to attract higher prices, suggesting that stakeholders could prioritize marketing or investing in these property types. Finally, emphasizing the number of bedrooms and the person capacity can target larger groups like families that are willing to pay for additional space and amenities.

On the other hand, the least influential features on price identified by our *Gradient Boosting* model are multiple rooms, host is a superhost, weekend availability, and private room. These features can also provide valuable insights for stakeholders. For instance, "superhost" status showed minimal impact on pricing, so hosts should not make it a priority to maintain or achieve this status (especially since there are many strict requirements to qualify as a superhost such as maintaining an overall rating of 4.8 or higher). Similarly, weekend availability, whether the room is a private room and whether the listing has multiple rooms also have limited influence on price optimization. Therefore, stakeholders can allocate fewer resources towards these features while focusing on the top predictor variables for maximizing occupancy and revenue.

Predictive analytics provides a significant advantage in the ever-changing short-term rental market. By leveraging complex models like *Gradient Boosting*, stakeholders can uncover key relationships between features and pricing that drive both revenue and competitiveness. This capability supports more informed decision-making, from setting individual property prices to designing broader strategies for market optimization. The insights from our analysis highlight how advanced machine learning techniques can translate complex data into actionable outcomes that align with business goals. The actionable findings from this research project empower stakeholders to refine their operations and marketing strategies, ensuring they remain agile in responding to fluctuations in consumer demand and short-term rental market conditions. Ultimately, our project highlights the transformative potential of integrating advanced analytics into rental and hospitality markets, ensuring that stakeholders remain well-equipped to navigate an ever-changing market.

Role of Each Member

Millie and Wenli collaborated on researching datasets, conducting data preprocessing, and performing exploratory data analysis. Reese implemented the *Multi-Linear Regression (MLR)* model. Yan then developed the *Gradient Boosting (GB)* model, while Vishnu focused on building and optimizing the *Deep Neural Networks* model. Finally, Millie and Reese worked to develop the project's overall analysis and conclusion. Together, our group was able to effectively divide tasks to ensure a comprehensive analysis.

Incorporation of Feedback

To address feedback, we switched from using R^2 to an adjusted R^2 for a more accurate evaluation of model performance. Using an adjusted R^2 allowed us to account for the number of predictors in the model, providing a more accurate measure of model performance and preventing possible overfitting. We also refined our tables for better clarity and readability when incorporating them into our paper. Finally, we incorporated a clearer and more detailed description of our business problem and the dataset used, ensuring improved communication of our research question and methodology.

References

The Devastator. (2024). *Airbnb Prices in European Cities* [Dataset]. Kaggle. https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities/data

Gyódi, K., & Nawaro, Ł. (2021). Determinants of Airbnb prices in European cities: A spatial econometrics approach (Supplementary Material) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4446043

de Jaureguizar Cervera, D., Pérez-Bustamante Yábar, D. C., & de Esteban Curiel, J. (2022). *Factors affecting short-term rental first price: A revenue management model*. Frontiers. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.994910/full

Bergeron, P. (2021, December 3). *Short-term rental demand on the rise*. benefitspro.com. https://www.globest.com/2021/12/03/short-term-rental-demand-on-the-rise/

European Parliament. (n.d.). *Short-term rentals: New EU rules for more transparency: Topics: European parliament*. Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20231127STO15403/short-term-rentals-new-eu-rules-for-more-transparency

Rentals United. (2024, February 16). *A guide to dynamic pricing for short-term rentals*. Vacation Rental Channel Manager. https://rentalsunited.com/blog/a-guide-to-dynamic-pricing-for-short-term-rentals/

Key Data Dashboard. (n.d.). *How to utilize short-term rental analytics for Success: Key Data*. Key Data Dashboard. https://www.keydatadashboard.com/blog/short-term-rental-analytics