Lisbon School of Economics & Management
Universidade de Lisboa

MASTER IN MANAGEMENT (MIM)

MASTER FINAL WORK

DISSERTATION

PREDICTING AIRBNB PRICES USING MACHINE LEARNING ALGORITHMS

JOÃO RICARDO GALHARDO ALMEIDA COSTA

JURY:

PRESIDENT: PROF. MARIA EDUARDA SOARES

MEMBER: PROF. CARLOS P. BERNARDINO

SUPERVISOR: PROF. CARLOS J. COSTA

FEBRUARY-2025

GLOSSARY

CRISP-DM – Cross-Industry Standard Process for Data Mining.

GBM – Gradient Boosting Machine.

GWR – Geographically Weighted Regression.

KNN – K-Nearest Neighbors.

LR – Linear Regression.

MAE – Mean Absolute Error.

MAPE – Mean Absolute Percentage Error.

MGWR – Multiscale Geographically Weighted Regression.

NLP – Natural Language Processing.

NN – Neural Networks.

OLS – Ordinary Least Squares.

OSM – OpenStreetMap.

P2P – Peer-to-Peer.

RF – Random Forest.

RMSE – Root Mean Squared Error.

SDG – Sustainable Development Goal.

STR – Short-Term Rental.

SVM – Support Vector Machine.

URL – Uniform Resource Locator.

VIF – Variance Inflation Factor.

XGBoost – Extreme Gradient Boosting.

ABSTRACT

The hospitality industry has changed as a result of the growth of short-term rental websites like Airbnb. Since hosts and policymakers have the need to possess a deep understanding of the variables that affect price, in order to balance housing accessibility and economic opportunity, price prediction has become an essential tool for them. Overall, Airbnb hosts may improve their pricing tactics and policymakers can create more balanced policies to maintain urban housing stability and tourism-driven growth, by utilizing data-driven techniques.

This study evaluates several models to investigate numerous machine learning techniques for predicting Airbnb prices in Lisbon's parish. To comprehend the nature of the data and find crucial characteristics for price prediction, the data is first cleaned and pre-processed before the undergoing descriptive, prescriptive and exploratory analysis.

The predictive performance of four distinct models was assessed, namely Ordinary Least Squares, Geographically Weighted Regression, Random Forest and XGBoost. Among these four, the Random Forests model outperformed the others in terms of rental price estimation, demonstrating how well it captures intricates market trends.

Keywords: Airbnb, Lisbon, Machine Learning, Price Prediction, Short-Term Rentals, Sharing Economy.

RESUMO

O setor da hotelaria tem sofrido alterações significativas devido ao crescimento de plataformas de aluguer de curta duração, como o Airbnb. Dado que tanto os anfitriões como as entidades reguladoras necessitam de um conhecimento aprofundado sobre as variáveis que influenciam o preço — com o objetivo de equilibrar a acessibilidade da habitação e as oportunidades económicas —, a previsão de preços tornou-se uma ferramenta essencial. De um modo geral, os anfitriões do Airbnb podem otimizar as suas estratégias de definição de preços, enquanto as entidades reguladoras podem desenvolver políticas mais equilibradas para manter a estabilidade da habitação urbana e, simultaneamente, fomentar o crescimento impulsionado pelo turismo, recorrendo a técnicas baseadas em dados.

Este estudo avalia diversos modelos com o intuito de explorar diferentes técnicas de aprendizagem automática aplicadas à previsão de preços no Airbnb, especificamente nas freguesias de Lisboa. Para compreender a natureza dos dados e identificar as variáveis mais relevantes para a previsão, foi realizado um processo de limpeza e pré-processamento, seguido por análises descritivas, exploratórias e prescritivas, e consequente modelização.

Com base na análise do desempenho preditivo de quatro modelos distintos — Mínimos Quadrados Ordinários, Regressão Geograficamente Ponderada, Floresta Aleatória e XGBoost —, a Floresta Aleatória destacou-se como o modelo com melhor desempenho na estimativa dos preços de arrendamento, demonstrando uma elevada capacidade para captar as tendências intrínsecas do mercado.

Palavras-chave: Airbnb, Lisboa, Aprendizagem automática, Previsão de preços, Aluger de curta duração, Economia de partilha.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I want to sincerely thank my family for their everlasting belief in me and all their support along this journey. My greatest strength has come from your love and wisdom, and I will always be appreciative of all that you have done for me, especially during moments when I doubted whether I would be able to balance work and my thesis.

To my friends, for their solid support and for reminding me to take breaks when I needed them the most. Your companionship and support made this process much more manageable and pleasurable.

I also want to express my profound gratitude to my professor Carlos J. Costa, for all the advice, and priceless insights. His tolerance when I was feeling particularly nervous, numerous meetings that he kindly made time for, his direction, subject-matter knowledge and his constant support were crucial to finish this project.

I could not carry out without sincerely thanking to my incredible girlfriend, whose constant encouragement and willingness to offer a hand in any manner have been invaluable. Throughout the most challenging moments, your patience and comprehension have been absolutely indispensable. I am so grateful to have you by my side, and your belief in me has meant the world to me, even when I doubted myself.

Thank you all for the bottom of my heart. This accomplishment would not have been possible without each and every one of you.

1. INTRODUCTION

Platforms like Airbnb have become major participants in the worldwide lodging sector as a result of the rise of the explosive sharing economy. Airbnb's rise in Lisbon, a city known for its thriving tourism sector and cultural diversity, has brought both new possibilities and problems for all parties involved. It is now more crucial than ever to understand and anticipate pricing dynamics in the context of growing demand. Both hosts and visitors benefit from accurate price prediction, which maximizes value and host profitability, which helps shape laws and regulations and provides insight to the dynamics of the market.

With an emphasis on one of Europe's most vibrant tourist destinations, this thesis adds to the expanding corpus of research on price prediction in the sharing economy by combining these insights with machine learning techniques.

Furthermore, predicting accurate housing prices may improve sustainability by encouraging the use of existing infrastructures, which reduces the demand for new hotel projects that contribute to urban growth and, consequently, environmental damage. Machine learning algorithms can help with the distribution of tourists lodging in a more equitable way around the cities by allowing hosts to modify prices dynamically depending on real-time demand and availability situations and, therefore, limiting the negative consequences of over-tourism in specific districts.

The purpose of this thesis is to investigate how machine learning models can be used to predict Lisbon Airbnb prices. By providing flexible lodging alternatives, the short-term rental (STR) market, mostly driven by websites like Airbnb, has revolutionized the hospitality industry. However, this quick growth has also caused worries about gentrification, housing affordability, and regulatory issues. Using machine learning to accurately estimate rental costs can result in a more sustainable and balanced rental ecology, offering insightful information to hosts, visitors, and legislators.

The growing complexity of the STR sector and its effects on the economy, society, and regulations are the driving factors behind this study. Numerous studies demonstrate how a wide range of criteria, such as location, guest reviews, host characteristics, and property attributes, have the biggest impact on Airbnb pricing prediction. The dynamic nature of demand and supply, external economic conditions, and seasonality make it tricky to estimate these prices. Additionally, according to recent research, more

sophisticated prediction models are required to enhance pricing tactics since machine learning approaches can better forecast prices by capturing complex interactions and non-linear connections between features, compared to traditional econometric models used in earlier studies.

Beyond academic interest, this study has practical implications in sustainability and urban planning. In order to prevent local people from being disproportionately impacted by growing rental costs, policymakers may create regulations or laws, providing insights into pricing patterns to limit negative externalities, like housing displacement and speculative investment. This aligns with the Sustainable Development Goal (SDG) 11, which seeks to ensure fair and transparent pricing mechanisms for accommodations.

This study aims to answer several key research questions. The first question examines the most important elements that affect Airbnb listing prices in Lisbon, examining how property-related characteristics, location features, host traits and reputation attributes affect costs. Additionally, by contrasting several methods, this research assesses which machine learning models offer the best accurate forecasts for Airbnb rates in Lisbon.

By addressing these questions and closing the gap between pricing theory and practical implementations, this thesis seeks to provide a data-driven method of price forecasting in the STR market, using machine learning to help create a more sustainable, effective, and balanced rental market that benefits not only hosts and guests but also the environment and larger community.

The CRISP-DM (Cross-Industry Standard Process for Data Mining) approach will serve as the foundation for the methodology used in this study. It will guide the process from data understanding and preparation to modeling and evaluation. This method provides a more thorough structure and framework for guiding the data mining process, as well as clear and precise guidelines on handling data analysis initiatives successfully and efficiently. From problem identification to solution execution, the framework comprehends six phases (Costa & Aparicio, 2020, 2021). The business understanding stage establishes the main objectives and requirements of the project from a business perspective, followed by the data understanding stage, which seeks to comprehend, locate, assess, and explore the existing data. In the third stage, known as data preparation, data is refined, transformed, and integrated to construct a viable dataset for the following

phase, modeling, where modeling techniques are applied. The evaluation and deployment phases, represented by the fifth and sixth phases, respectively, involve the evaluation of the developed techniques from the previous stage, using performance and accuracy metrics on unseen data to identify which technique had the best results. Decisions are then made based on the insights obtained. This iterative process allows the possibility of revisiting earlier phases to redefine models or explore different strategies as the project progresses.

This document is divided in six chapters. Starting with the introduction, which provided an overview of the research, the importance of Airbnb price prediction is explained, outlining the study's goals and describing the methodological approaches. In the literature review chapter, key theories, concepts, and previous studies on Airbnb price prediction are discussed, focusing on the hedonic pricing model and machine learning. The methodology chapter covers data collection, exploratory data analysis, and the modeling techniques used to predict Airbnb prices. In results and discussion, the model outcomes are presented, and their performance is compared, and the findings are interpreted in relation to existing literature. The last chapter, the conclusion, summarizes the key findings, limitations acknowledged, and future suggested research directions.

2. LITERATURE REVIEW

Examining the literature to comprehend the methods, factors, and findings of earlier research is essential to this study. In order to investigate variables like location, facilities, reviews, and host traits, previous research has mostly used hedonic pricing model approaches. This analysis of the literature will serve as a basis for the creation of a customized prediction framework for the Airbnb market in Lisbon parish by summarizing the models and attributes employed in highlighting the research conclusions and shortcomings.

2.1. Sharing Economy

For ages, communities have combined resources and assets, a custom firmly anchored in human collaboration and support for mutual benefit. However, this ancient idea has changed with the rise of the internet and its growth, leading to the creation of what today is referred to as the sharing economy (Minami et al., 2021). To understand this concept, it is important to define it. The sharing economy is a new form of sharing resources that has developed with the broad adoption of digital technology (Sanasi et al., 2020). It represents a major change in the ways that people interact, work together, and share resources in the modern world. This process has been completely transformed by digital platforms, which now make it easier for resource seekers and asset owners to locate and do business with each other.

Also known as collaborative consumption, it is characterized by peer-to-peer (P2P) exchanges coordinated through community-based online platforms, where individuals share or rent goods and services rather than preserving them entirely (Hamari et al., 2016). Through user evaluations, safe payment methods, and intuitive user interfaces, these platforms build markets that improve convenience and trust while encouraging wider participation, enabling people to operate as micro-entrepreneurs and make money. Moreover, as noted by Huurne et al. (2017), this P2P connection enables individuals to share underutilized assets for mutual benefits. This concept represents a paradigm shift in the way value is created and shared that challenges the supremacy of traditional business models, in which big firms control distribution and resources. Through the combination of their own survey data with insights from previous research, Hamari et al. (2016) were able to validate two main incentives as the fundamental drivers

of sharing economy participation. On one hand, the economic benefits, since this collaborative use of goods and services allows property owners to monetize inactive resources and permits seekers to save money by accessing services at lower costs. On the other hand, the sustainability since this P2P transaction reduces the need for new production, contributing to the preservation of the environment with a more efficient way of utilizing resources.

The sharing economy has emerged as a disruptive force across multiple industries, especially disrupting traditional sectors, such as hospitality, transportation, and retail. By altering this business model and consumer behavior, it has shifted the focus from ownership to access, allowing people to rent or share unused assets in return for reciprocal advantages. Also, according to current definitions, its commonly defined as a system in which customers have a larger role as both resource producers and users (Lang et al., 2022) that not only promotes more effective resource usage but also democratizes market involvement. In Lang et al. (2022) article, it is gives an insight into the motivations behind individual's decision to become a providers in the sharing economy. Although financial incentives are still significant, motives like independence, ability, and connection have a much larger impact. That said, platforms that support these types of incentives have a better chance of succeeding and keeping satisfied providers.

One of the most prominent examples of this model, the known leader of the transformation of the short-term accommodation market, is Airbnb (Lee et al., 2021), which through its digital platform, connects hosts and guests, providing access to a variety of housing alternatives to visitors.

2.2. Airbnb: Lisbon case study

Over the past decade, Lisbon's real estate market has undergone an intense transformation into a global tourism center, a shift that has significantly impacted its housing market. A key driver of this change has been the rapid expansion of STRs, particularly through platforms like Airbnb. This trend has made Lisbon an important case study for examining how Airbnb pricing reflects and shapes broader economic and social dynamics, especially as the city struggles with the numerous challenges of these changes.

Lisbon's growing tourist industry, which has made the city one of Europe's most popular destinations, is directly linked to the rising number of STRs in this capital.

Property owners can transform residential spaces into tourist accommodations, thanks to platforms like Airbnb that have completely revolutionized how lodging is marketed and used. In addition to satisfying the growing need to accommodate tourists, this modification has had a significant impact on Lisbon's urban landscape.

Studies like Costa et al. (2019) highlight the complex interplay between the growth of tourism and the proliferation of STRs. On one side, these platforms have empowered property owners the ability to profit from the rise in tourists. On the other, they have played a role in rising property values, the commodification of housing, and in the gentrification of neighborhoods.

Moreover, as Lorga et al. (2022) stated, tourism, the foreign population, public regulations, and STRs all have a significant role in the increase of property prices, exceeding typical wage growth and reducing housing affordability.

That said, Lisbon's experience reflects how the growth of tourism and the rise of platforms like Airbnb may significantly alter urban housing markets, presenting both substantial social obstacles and chances for financial gain. The city's quick rise to recognition as a travel destination has been driven by both its appeal to tourists and the remarkably high profitability of STRs.

STRs are an appealing choice to traditional long-term leases because they allow property owners and investors to produce significantly higher profits. Cocola-Gant and Gago (2021) point out that this profitability has brought in professional investors who purchase properties with the purpose of turning them into STRs, frequently relocating long-term occupants in the process. Particularly in historically prominent neighborhoods, like Alfama, this approach has fuelled extensive gentrification. In these urban and cultural areas, changes have occurred, especially in property prices and increases in rent.

The effects of Airbnb's expansion in Portugal's capital go beyond just financial considerations. There have also been social and cultural changes triggered by the increase of tourism, which has weakened neighborhood cohesion and disrupted established communal institutions. Lestegás et al. (2019) explain how communities' identities are modified, and their sense of community is undermined when long-term inhabitants are replaced with temporary tourists. In parishes like Santa Maria Maior, the conflicts over space and culture are becoming more intense due to the outnumbers of the local population.

An engaging case study for comprehending the wider effects of STRs on urban property markets is Lisbon's experience with Airbnb. The city's unique environment for research is provided by its distinctive blend of increasing tourism demand, focusing on STR expansion and continuous initiatives.

2.3. Hedonic Price Model

The hedonic pricing model is an economics approach that helps determine how different product or service aspects impact its price. It acknowledges that the economic value of an asset is defined by the sum of its different traits or features rather than itself only. It provides a method for determining how much buyers are willing to pay for particular attributes, such as size, location, or other enhanced conveniences, which together determine the overall price (Zhan et al., 2024, Samadani & Costa 2021).

Originating from the work of Rosen (1974), this approach lies in the idea that products and services are not uniform. This means that they include a variety of characteristics that bring value to the customer. For example, in housing or accommodation sectors, considerations such as the size of the property (number of bedrooms and bathrooms), the amenities offered (Wi-Fi, pool, air conditioning), the proximity to city centers or tourist attractions, the reviews and ratings of the host and many others have a significant impact on pricing (Gibbs et al., 2018).

More specifically, in the Airbnb context, the hedonic price model can help measure the ways in which various factors affect the cost of STRs. This model offers insights into visitor decision-making and pricing strategies for hosts by breaking down the price into the contributions of different property features.

Studies, such as A. Gutiérrez and Domènech (2020) and Gyódi and Nawaro (2021), that used spatial hedonic price models, have shown that one of the most critical factors of Airbnb pricing is the access to crucial tourist attractions. They came to the conclusion that, rather than the distance to the city centre, this is the primary pricing factor that best explains the higher prices.

While not exclusively focused on the hedonic price model, the study of Eugenio-Martin et al. (2019) takes into account elements that influence location and pricing, as well as concepts from spatial analysis. The authors highlight that geography has a crucial role in determining whether Airbnb listings are present or not and how much they cost.

They also consider how safer, more accessible, and better-equipped neighborhoods attract both hosts and guests. The intersection of gentrification and community identity with the findings of Eugenio-Martin et al. (2019) is an intriguing angle to consider. With this, representatives might explore how hedonic price models and geographical data can guide tactics to strike a balance between the well-being of the community and financial benefits.

Although there's a high number of studies offering a starting point for comprehending pricing processes through features that can be measured (size of the property, for example), Gibbs et al. (2018) also leave room to explore the more subtle and intangible elements of what contributes the idea of value in the sharing economy world. This interaction of material and immaterial elements indicate wider trends in contemporary consumer behavior.

There have been recent empirical studies that increased the use of hedonic price modeling in the Airbnb atmosphere. The findings of Dogru and Pekin's (2017) research demonstrate that intangible factors, like cultural relevance or host responsiveness, are just as important, if not more, than traditional determinants like the location of the property. Travelers are increasingly looking for more than just one place to stay, and as the digital age progresses, customer preferences and demands in this industry are shifting, adding greater weight to intangible factors. They seek more meaningful experiences and guarantee that what they choose aligns with their ideals. By considering these non-physical attributes, it is presented a more comprehensive view of what impacts pricing, which underscores that in today's market, ethical and emotional considerations can be as important as tangible ones.

All these study results highlight how pricing in the Airbnb ecosystem is diverse, as consumer expectations reshape the dynamics of the market with more experiential factors. As the sharing economy world continues to evolve, hosts, policymakers, and researchers need to understand these intricate interactions through the hedonic pricing model to analyse the complex needs of this customer-focused competitive industry.

2.4. Predictive Algorithms and Price Determinants

There are an extensive number of studies using both statistical and machine learning techniques on this topic. Many of these models relied heavily on traditional hedonic pricing models derived from consumer theory. For example, Beijing's research

by Zhao et al. (2023) utilized Multiscale Geographically Weighted Regression (MGWR) to account for spatial heterogeneity in pricing, using and comparing hedonic attributes, such as property features (number of bedrooms and size of property), host characteristics (superhost and professional-host), and location (distance to hotspots and accessibility). By capturing the multiscale geographical effects of listing hedonic attributes, MGWR outperformed Ordinary Least Squares (OLS) regression and conventional Geographically Weighted Regression (GWR). It has been proven that in major cities such as Beijing, the ability to represent complex differences across multiple geographical dimensions is very useful in understanding price patterns. Similarly, a study by Iliopoulou et al. (2024) in Athens employed OLS and GWR to forecast STR rates, where GWR outperformed OLS by integrating spatial dependency, enhancing prediction accuracy. They came to that conclusion using several variables to predict the prices of STR listings, like property features, host qualities, and spatial features.

Machine learning approaches have been widely applied in Airbnb pricing research, owing to their capacity to handle non-linear correlations and high-dimensional datasets. Research developed by Wang and Huang (2023) in Hong Kong used Linear Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) to predict Airbnb prices. The authors came to the conclusion that the RF model exhibited better predictive performance, highlighting the effectiveness of ensemble approaches in price forecasting. These machine learning algorithms have gained more popularity in this type of study, especially in markets like New York and Los Angeles, where robust analytical tools are necessary due to intricate relationships between factors (Wang & Huang, 2023; Iliopoulou et al., 2024).

The COVID-19 pandemic brought extensive challenges to both Airbnb hosts and owners that had to reconsider their pricing methods in order to survive this outbreak. In a study by Gyódi (2021), the way how hotels and Airbnb hosts adjusted to this shifting legislation and demand was examined to analyze pricing trends before and during the pandemic. He used LR to quantify how various factors affect price changes, time-series analysis to track price patterns, and panel data analysis to assess price elasticity across various property types and different cities. The analysis showed a number of important variables affecting Airbnb rates, such as the declining demand that forced Airbnb hosts to lower their prices compared to hotels, the type of property, as listing entire homes

continued to have higher price points than those for single rooms, that saw more dramatic drops. The geographical location also played a role, as urban listings saw greater price declines than rural ones. Lastly, government regulations, such as travel restrictions and lockdown orders, had a big impact on pricing tactics that affected both immediate and long-term changes.

An article written by Zhao, H. (2023) about a study made to analyze the spatial-temporal heterogeneity and how built environment elements affect the cost of an Airbnb listing focused on how these impacts vary throughout time and space, using the OLS and the GWR model. While the OLS model provided a general understanding of how different variables influence listing prices across Copenhagen, the GWR revealed that the impact of built environment features has distinct effects in numerous locations. The study divided the factors into three main categories. First, accessibility indicators, such as the distance to the city center, to the nearest metro station, to green spaces, and to tourist attractions. The amenity indicators category was based on the proximity to needed locations, like restaurants and cafés, supermarkets and shopping plazas, and the presence of cultural landmarks. Lastly, the importance of neighborhood characteristics, mainly the population and housing density, crime rates, and socioeconomic status of the area. The author also highlighted the difference between the seasonal shifts, that while on peak tourist season (summer), the influence of tourist attractions and green spaces is strongest, on low tourist season (winter), the importance of built environment variables declines, as demand shows.

Moreover, by employing a combination of regression techniques and Natural Language Processing (NLP), Di Persio and Lalmi (2024) published a study that aimed to optimize both profitability and occupancy rates by analyzing key listing variables and guest preferences. The models chosen for this study include Support Vector Machine (SVM), XGBoost, and Neural Networks (NNs) due to their ability to represent non-linear correlations between listing prices and diverse influencing factors. NLP techniques, such as topic modeling and sentimental analysis, were used to examine guest evaluations and gather information about consumer preferences, perceived value, and satisfaction levels. This research considered multiple variables that were divided into three categories: location-based factors, listing characteristics, and host attributes. Proximity to commercial districts, tourist attractions, and public transit are examples of location-based

factors. The effect of host qualities on price is examined, including response rate, the quantity of listings under the host's management, and the experience level. The number of rooms, facilities, overall rating, and availability of unique features like balconies or swimming pools are all listed traits. In order to evaluate the relationship between textual feedback and price and occupancy rates, NLP was also used to process guest reviews. The author concluded that, among the regression models, NNs emerge as the best performers, simply because of their capacity to manage complicated interactions between variables, exhibiting great predictive capabilities. Pricing models are further improved by incorporating sentiment analysis from guest evaluations, which offers insights into consumer satisfaction beyond what can be obtained from numerical ratings alone.

The study by Lektorov et al. (2023) investigated how machine learning models can predict Airbnb rental prices in New York city. Their primary objective was to evaluate how different machine learning models perform in predicting rental prices, identifying the variables that influenced these prices best, and which model, linear and non-linear, was the most accurate. The Decision Tree was used as a baseline to understand how individual variables influence rental prices, splitting the data into branches based on the features. With the K-Nearest Neighbors (KNN) non-parametric, instance-based learning model, they predicted price based on average prices of the 'K', to determine the ideal number of neighbors in order to get the best accuracy. Multiple decision trees with additional randomization in feature selection were built using the Extra Trees models in order to take advantage of ensemble learning. The SVM model was applied to determine the best hyperplane, dividing several price points in a high-dimensional space. Lektorov et al. (2023) also explored the RF model, an ensemble approach combining multiple decision trees, averages several forecasts to improve accuracy and reduce overfitting. Lastly, with better precision, despite the cost of more computing complexity, the XGBoost was used as the primary model to achieve high predictability accuracy, adding trees that rectified the errors of previous iterations. In order to train these models, the study takes into account a variety of property-specific factors, such as the number of accommodations, the listing type, the square footage, the facilities provided, and more. Location-based elements and market factors are also examined, as well as host-related features. The results show that Airbnb rental prices have a complicated, non-linear

connection with all these variables, and in particular, RF, XGBoost, and Extra Trees demonstrated that complex ensemble models outperform simpler, traditional models.

According to a study by Milunovich and Nasrabadi (2025), to investigate predictive modeling for pricing on Airbnb in Sydney advanced predicting modeling, especially ensemble techniques like stacking regressions, greatly increases the accuracy of forecasting Airbnb rental prices in Sydney, exceeding simpler linear models. While proximity to highways tends to lower rental costs, other factors, including property capacity (number of bedrooms and bathrooms), upscale amenities, and accessibility to well-known areas, all have a favorable impact. The report revealed that forecasting accuracy increases as a result of feature engineering-enhanced datasets, especially in training ones, and that higher costs are predicted for properties with specific characteristics and guest capacity. Moreover, the results also show that professional hosts oversee a large position of Airbnb housing, pointing to the requirement for distinct regulatory frameworks that can help handle Sydney's housing affordability and availability issues more effectively. All things considered, their results highlight how crucial machine learning methods are in providing useful insights for hosts and legislators in the STR market.

Chen and Xie (2017) examined the factors that influence Airbnb listing prices using a standard hedonic pricing model with the OLS regression to evaluate how various listing criteria affect pricing. Property features, such as size and kind of lodging; geographical factors, which include accessibility to landmarks and city centers; host reputation, which is frequently measured by response rates and a number of reviews; amenities, like pools and air-conditioning; and booking flexibility, mostly cancellation policies, are the categories in which they divide Airbnb listing attributes. This traditional approach of classification captures the essential pricing factors well but does not account for spatial differences in attribute effects. Finally, his results show that while host reputation and amenities are important factors, property attributes and location have the most effects on cost.

In another study, by using an MGWR technique, Hong and Yoo (2020) expand on the hedonic price model and account for regional variation in price determinants. Similar to Chen and Xie (2017), they categorize qualities, but they place more emphasis on the geographical scales at which certain features affect price. They differentiate

between impacts that are local, regional, and global, highlighting qualities that include size, facilities, and the architecture of local traits. Neighborhood-level elements, such as safety, availability of food options, and entertainment choices, are all considered regional qualities. Lastly, global characteristics reflect more general market conditions (demand for tourists and citywide economic trends). This categorization method offers a more dynamic comprehension of how hedonic qualities affect Airbnb prices across different areas by including multiscale spatial effects.

By incorporating GWR in order to examine the role of streetscape elements in Airbnb pricing, Wand and Rasouli (2022) present an alternate classification technique. Their strategy places a strong emphasis on how urban environmental factors affect the platform rates. They divide qualities into three categories: location and accessibility features, streetscape elements, and conventional property-related criteria. Identical to Chen and Xie (2017), traditional property-related criteria include structural and service features. Geographical considerations, such as proximity to important places and access to transportation, are covered by location and accessibility features. A distinctive addition is streetscape attributes, which include road width, vegetation, the visual appearance of the nearby structures, and pedestrian hospitality. This categorization approach provides a more understanding of the factors that influence listing prices by extending the hedonic pricing model to account for the impact of urban environments.

The categorization frameworks presented in these three studies provide different viewpoints on the factors that influence Airbnb prices (Zhao et al., 2023). Chen and Xie (2017) aggregate qualities according to classic real estate price models, utilizing a traditional hedonic pricing technique. Although this approach lacks geographical distinction, it offers a strong basis for comprehending pricing factors. On the other hand, by classifying characteristics according to their magnitude of effect, Hong and Yoo (2020) suggested a spatial point of view that improves the model's capacity to capture changes across several geographic locations. Finally, in order to acknowledge the influence of urban environmental elements on Airbnb pricing, Wang and Rasouli (2022) adopt a unique technique by integrating streetscape features.

The OLS method makes it simple to apply in numerous situations by offering a methodical and transparent approach to pricing determinants analysis. Also, greater flexibility is introduced by the GWR method, which takes into consideration regional

differences in pricing drivers, and through the use of streetscape features, hedonic price research may capture a wider variety of factors that affect Airbnb pricing.

Pricing models continue to be interpretable thanks to the use of explainable machine learning techniques that offer hosts and market analysts insightful information. With the possibility of comparing several classification techniques, it is possible to increase the accuracy of Airbnb pricing forecasts, creating more thorough models.

3. METHODOLOGY

As previously noted, the CRISP-DM technique was utilized to find the optimal model for price prediction and identify the variables that best explain the price changes (Chapman et al., 2000; Costa & Aparicio, 2020). Because of its adaptability, it is an extremely useful framework for data science projects, ensuring an organized and flexible approach to the research (Costa & Aparicio, 2021).. This study's cleaning, pre-processing, and modeling source code is accessible on GitHub at https://github.com/Galhardo19/tese.git.

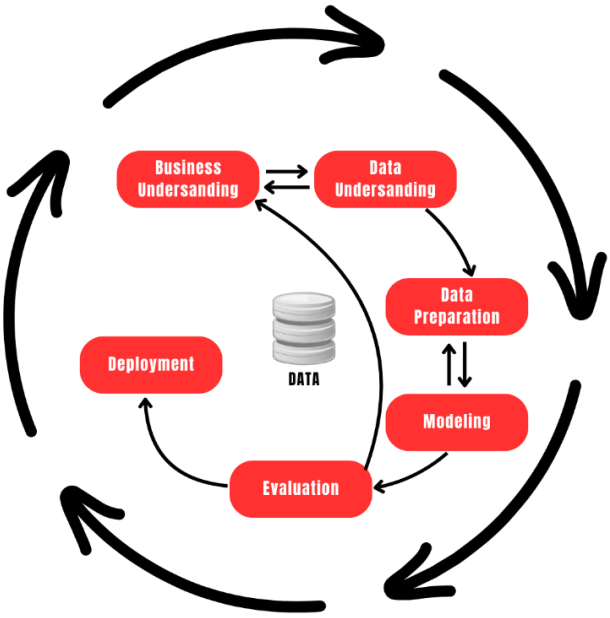

Figure 1 – CRISP-DM Methodology

Source: Chapman et al., 2000

3.1. Data Collection and Processing

Through the Inside Airbnb website, a third party with the main goal of compiling public data from Airbnb.com, accessed on October 4 of, 2024, was extracted the sample listing data in order to provide insights into the short-term market. A file containing 75 columns, which represented all potential variables, and 24204 data points, each of which represented a distinct listing.

As can be seen by looking at every variable (Annex A), many of them are initially not taken into account when predicting pricing, such as the URL variables. Additionally, since NLP will not be implemented, free text columns (*name* and *description*), and other columns that are not helpful for predicting price can be removed from the equation, such as *scrape_id*, *last_scraped*, *source*, *host_has_profile_pic*, *host_identity_verified*,

host_verifications, *calendar_updated*, *licence*, *instant_bookable*, *host_id*, *host_name*, *host_location*, *host_about*, *host_neighborhood*, *neighborhood_overview*, *bathrooms_text* and *calendar_last_scraped*.

Table I – Variables considered and discarded

Type of variables	Variables considered	Variables discarded
Review-related	numer_of_reviews	number_of_reviews_ltm, number_of_reviews_l30d, reviews_per_month, first_review, last_review
Score-related	review_scores_rating	review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, and review_scores_value
Stay-related	minimum_nights, maximum_nights	minimum_nights_avg_ntm, minimum_minimum_nights, maximum_minimum_nights, maximum_nights_avg_ntm, minimum_maximum_nights, maximum_maximum_nights
Availability-related	availability_30	has_availability, availability_60, availability_90 and availability_365
Neighbourhood-related	neighbourhood_cleansed, neighbourhood_group_cleansed	neighbourhood
Property-related	accommodates	beds, bedrooms, property_type
Host-related	host_is_superhost, host_total_listings_count	host_since, host_response_time, host_response_rate, host_acceptance_rate, host_listings_count, calculated_host_listings_count, calculated_host_listings_count_private_rooms, calculated_host_listings_count_entire_homes, calculated_host_listings_count_shared_rooms

Observing Table I, it was made a further selection of variables to include in the dataset. Regarding the review-related variables, due to their high correlation with the *number_of_reviews* variable, the rest of the attributes were discarded since it could introduce redundancy and, respectively, lower the model's efficiency. Along with improving the selection of this type, since the *first_review* and *last_review* variables do not provide any useful additional information and because of their lack of pertinence to the pricing prediction model in this study, they were excluded.

Similarly, for review score-related variables, the Inside Airbnb website provided multiple rating metrics, where, because it is a broad indicator that serves as a general

measure of guest satisfaction and listing quality, it was only kept the *review_scores_rating*. Even though the other review score variables offer more detailed information about several parts of the stay, their inclusion is considered less important, as they have a strong association with the total ratings.

In terms of attributes concerning the stay-related type, the Airbnb platform provided multiple variables and, once more, to avoid redundancy and maintain clarity in this model, only *minimum_nights* and *maximum_nights* were retained. These two represent the key limitations on a listing's stay requirements, while the others, that offer averages or adjusted values over time, were not taken into account because they could add complexity without significantly enhancing the ability to predict.

Following the same method, other attributes were dropped, due to their poor relevance or duplication, in order to further restrict the selection and increase the model efficiency. Among the availability-related variables, only *availability_30* was maintained. Short-term availability is succinctly and meaningfully represented by this variable, whereas the others had overlapping data that did not much improve prediction accuracy.

For neighbourhood-related attributes, *neighbourhood_cleansed* and *neighbourhood_group_cleansed* were considered over *neighbourhood*, due to data quality considerations. Unlike the last one, that frequently has inconsistent naming and formatting rules, the chosen variables provide a more organized and trustworthy geographic information, which makes them better suited for analysis.

In property-related variables, the *accommodates, beds* and *bedrooms* attributes showed a substantial association in a number of research, indicating that they deliver overlapping information. That said, since *accommodates* is the most thorough indicator of a listing's capacity, it was kept, eliminating repetition and simplifying the model. Furthermore, although this variable can provide information on a listing's nature, this analysis did not include *property_type*, since it was judged unnecessary for the purposes of this research.

In terms of host-related characteristics, a host's eligibility as a superhost is determined by several factors (*host_since, host_response_time, host_response_rate* and *host_acceptance_rate*). Because it combines all of the data into a single, more effective indication, *host_is_superhost* was added rather than each of the other variables separately.

Likewise, instead of including numerous attributes, only the *host_total_listings_count* variable was selected, as it essentially counts all of the attributes connected to a host, omitting all the others.

Iliopoulou et al. (2024) concluded that the information relative to the *amenities* was not entirely accurate due to the way each facility was differently represented. He stated that it was challenging to standardize the data, as the same amenity was frequently phrased variously. In addition, it was not often clear whether certain facilities were available. Parking, for instance, was usually referred to as either free street parking or paid parking instead of an on-site amenity. Including these amenities in the model only increased the r-squared by 1.5% to 2%, and because of that low percentage, this variable was not considered.

Table II – Null entries and data type

Attributes	Description	Null values	Data type
id	Identifier of each listing	0	int64
host_is_superhost	Whether the host is a "Superhost" (Yes/No)	870	object
host_total_listings_count	Total number of listings the host manages	0	int64
neighbourhood_cleansed	The cleaned name of the neighborhood	0	object
neighbourhood_group_cleansed	Neighborhoods parish	0	object
latitude	Latitude of the listing's geographical location	0	float64
longitude	Longitude of the listing's geographical location	0	float64
room_type	The type of room	0	object
accommodates	Number of people the listing can accommodate	0	int64
bathrooms	Number of bathrooms in the listing	2978	float64
price	Price per night for the listing	2971	object
minimum_nights	Minimum number of nights required for booking	0	int64
maximum_nights	Maximum number of nights allowed for booking	0	int64
availability_30	The number of available nights in the next 30 days	0	int64
number_of_reviews	Total number of reviews the listing has received	0	int64
review_scores_rating	The overall rating of the listing	3052	float64

The original data required to go through a number of procedures, including filtering, handling missing values and removing duplicates, in order to provide a more efficiently model. Python was used to build the whole cleaning process, exploratory analysis and modelling. Following the removal of all superfluous variables from the

dataset, the number of null entries for each of the 16 attributes was examined, as shown in Table II .

The variables with the most null entries, as can be seen by looking at the table, were *price*, *bathrooms* and *review_scores_rating*. In addition to identifying missing data, an analysis of duplicate entries was conducted, revealing that there were no duplicates present in the dataset. After this verification, 18084 data points were left in total. As the next step in the data preprocessing process, the id column was designated as the index, to facilitate further analysis.

Following the cleaning process, the data transformation came subsequently, that consisted in converting data type and renaming some columns to maintain uniformity throughout the data and enable additional analysis (Annex B). That said, in order to simplify, *host_is_superhost* was renamed to *superhost*, *number_of_reviews* to *reviews* and *review_scores_rating* to *ratings*.

It becomes evident from looking at Table II , that the variables *price* and *bathrooms* needed to be converted to proper data types. The attribute *price* needed to be changed to a float, and the attribute *bathrooms* needed to be modified to an integer. A closer look at the pricing values revealed that they were represented as dollar amounts, formatted with dollar signs (e.g., "$60.00"). To prepare these numbers for conversion into a float, the dollar sign and commas had to be removed. Additionally, since the analysis was conducted for properties in Lisbon, the *price* values had to be changed from U.S. dollars to euros. To do that, on february 4, 2025, it was used the Yahoo Finance platform.

In parallel, for consistency and additional analysis, the *superhost* variable, which had the categorical values "f" (false - indicating not a superhost) and "t" (true - indicating a superhost), required transformation into binary variables. This was achieved using one-hot encoding, where "f" and "t" was replaced by "0" and "1", respectively. A new binary variable, *professional_host* was also added during this data transformation process to differentiate between hosts who run one listing and those who manage two or more. Using the standards established by Deboosere et al. (2019) and Gibbs et al. (2018), which categorize hosts with two or more listings as professional hosts, this new variable was established from the *host_total_listings_count* column.

Once the columns were renamed and the data types standardized, a number of filters were used to further refine the dataset in line with the particular goals of this study.

Since the Lisbon parish is a well-known location with a large number of tourists, it was decided early on that the research area would be limited to this region. Therefore, only items with the label "Lisbon" were included in the variable *neighbourhood_group_cleansed*. Next, to further narrow the dataset, the *room_type* variable was filtered to solely include listings that were identified as "Entire home/apt". This helped to get more accurate results, considering that distinct Airbnb room types target a diverse group of visitors with varied value preferences, leading to distinct price mechanisms (Zhao et al., 2023).

More filters were used to guarantee the data's quality and applicability. In order to exclude entries with zero, the *price, reviews,* and *availability_30* variables were all filtered. Properties with no reviews, a price of zero euros or that were unavailable within 30 days were considered irrelevant for this analysis since it could be a sign of incomplete information or inactive listings that do not offer any useful information about the rental market that is currently operating. Consequently, only properties with positive values in these variables were kept for the final dataset.

The subsequence phase in this research was finding and resolving any outliers in the dataset, as it is crucial to detect data points that substantially depart from the predicted range. These anomalies have the potential to influence the study as a whole and provide false findings. The z-score method was employed to identify such outliers. This statistical method measures the standard deviations, or distances, between a given data point and the mean, helping to pinpoint values that are unusually high or low compared to the rest of the dataset. Upon applying the z-score technique, 46 data points were identified as extreme outliers with exceptionally high prices. These numbers were regarded as anomalies, most likely the consequence of inaccurate data entry or unusual properties that did not correspond to the overall trends in price. To improve the quality of the dataset in this study, all 46 outliers were eliminated. After implementing all necessary adjustments and cleaning steps, there was a reduction in the dataset to a total of 8298 listings.

Numerous studies highlight the significant impact that location has on property prices. Recognizing this, four key variables were created to represent the location-related attributes that could best impact pricing: *dist_hotspots, dist_metro, num_sports,* and *num_accomodation*. The *num-sports* measures the number of sports/leisure services within a 1 km range of the property. The *num_accomodation* represents the number of

accommodation services within 1 km of the property (Annex C). The *dist_metro* represents the distance of the listing to the nearest subway entry based on Euclidean distance. For all these three variables, it was necessary to make use of the Geopandas library and the Overpass Turbo, a web-based tool that enables querying and extracting geospatial data from OpenStreetMap (OSM), a collaborative mapping project that provides detailed location-based data (Figure 2). The *dist_hotspots* characterize the sum of the distance from a listing to the 20 most popular attractions in Lisbon using Euclidian distances. For this variable, 20 popular attractions in Lisbon were compiled along with the number of reviews as a measure of popularity based on TripAdvisor's rankings. Each attraction was assigned a weight according to its review count, and a weighted sum of Euclidian distances was calculated for each property. The latitude and longitude information were achieved by OSM.

Figure 2 – Visual Output num_accommodation, num_sports & dist_metro

3.2. Exploratory Data Analysis

Prior to using any machine learning models, it is crucial to thoroughly comprehend the depth in the dataset, once it has been cleaned. Matplotlib and Seaborn were the primary libraries used. Table III displays the descriptive statistics for each variable.

Table III – Descriptive Statistics

Variables	count	mean	std	min	25%	50%	75%	max
price	8298	148.28	82.73	17.43	94.89	124.91	174.29	716.54
accommodates	8298	4.08	2.02	1.00	2.00	4.00	5.00	16.00
bathrooms	8298	1.25	0.54	0.00	1.00	1.00	1.00	8.00
reviews	8298	109.26	117.90	1.00	19.00	67.00	163.00	861.00

Variables	count	mean	std	min	25%	50%	75%	max
ratings	8298	4.65	0.35	1.00	4.53	4.73	4.86	5.00
minimum_nights	8298	3.25	10.53	1.00	1.00	2.00	3.00	365.00
maximum_nights	8298	537.45	467.28	1.00	90.00	365.00	1125.00	3000.00
superhost	8298	0.36	0.48	0.00	0.00	0.00	1.00	1.00
professional_host	8298	0.84	0.36	0.00	1.00	1.00	1.00	1.00
num_sports	8298	77.93	22.94	10.00	64.00	77.00	89.00	241.00
num_accomodation	8298	86.28	63.12	0.00	30.00	79.00	136.00	206.00
dist_hotspots	8298	5.44	1.16	4.66	4.78	5.02	5.48	12.90
dist_metro	8298	0.74	1.02	0.00	0.27	0.46	0.74	7.86

By analysing the table, it is possible to extract important information. According to the statistics, for the variable *reviews*, each listing has an average of 109.26 reviews. Significant difference in the number of reviews among properties is shown by standard deviation. Some listings may have as little as one review, while others have up to 861, according to the minimum and maximum values. In terms of distribution, 50% of the listings have 67 reviews or fewer, whilst 25% have 19 or fewer. Furthermore, there is at least 163 reviews that are included on 25% of properties. A deeper look and analysis of this statistics can help better understand how each variable behaves.

With a mean of 148.28€, the target variable, *price*, ranges from 17.43€ to 716.54€. The median price of 50% of listings is 124.91€ or less, while 25% of properties are priced at 94.89€ or less. Additionally, 25% of listings also cost at least 174.29€, which means that 75% are priced at that value or less. According to this distribution, a lesser percentage of properties have a higher price, with the majority falling within moderate price range.

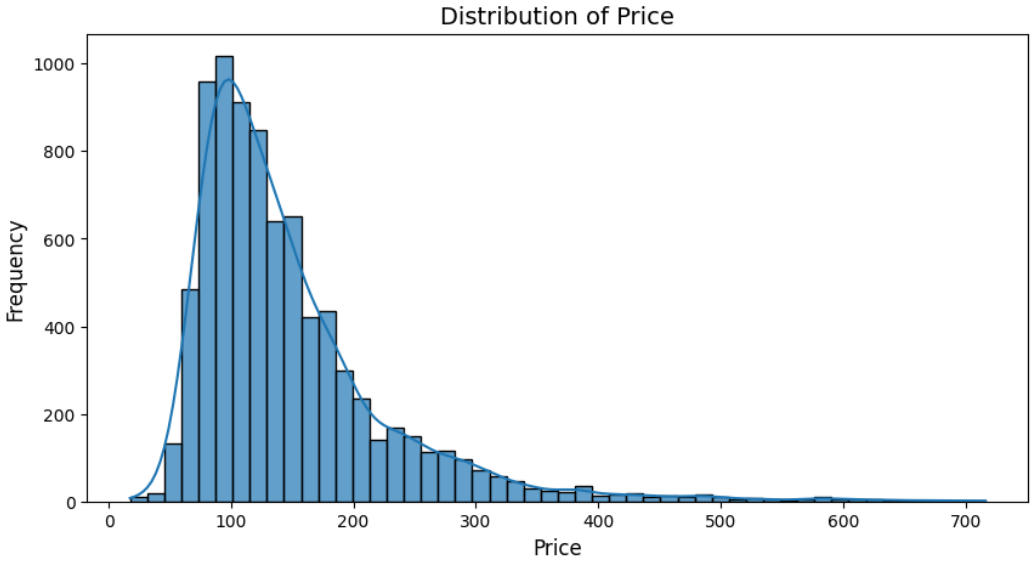

Figure 3 – Distribution of Price

The price distribution is favourably prejudicated, as shown in Figure 3, having fewer, more costly properties extending the right tail of the distributions, with the majority of the listings costing around 100€. In order to give a more accurate representation of central tendency, the median price was estimated across listings that could accommodate varying numbers of guests, due to the uneven distribution, as seen in Figure 4. It should come as no surprise that properties with more rooms often charge more per night. After about 11 guests, is possible to see the price rising slowly down, indicating a decreasing marginal return for bigger lodgings.

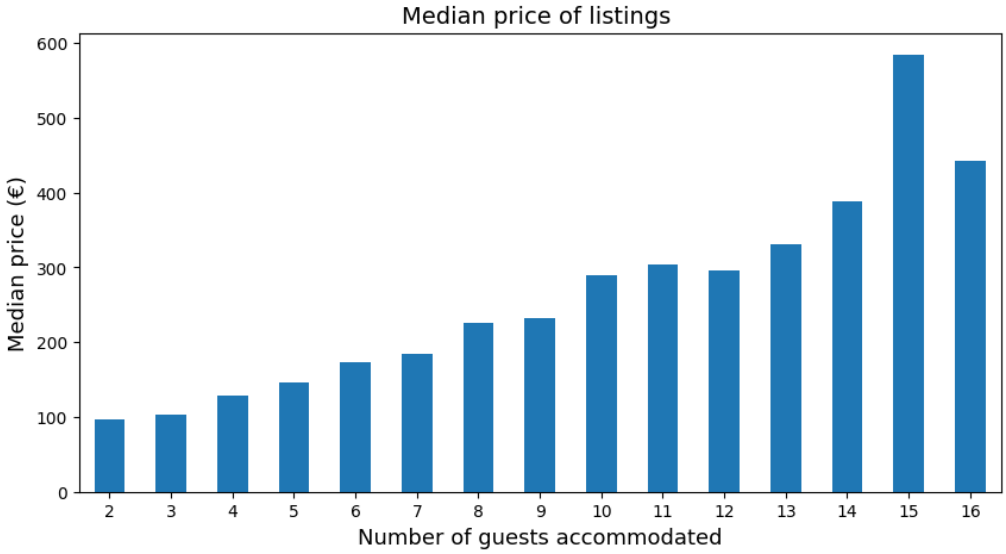

Figure 4 – Median Price of Listings

To get a more visual picture, the geographic distribution of price was plotted. The priciest listings are located close to the sea (Figure 5). Given that the average prices in Santo António and Misericórdia are 163.60€ and 162.311€, respectively, these are the most expensive neighborhoods (Annex D). On the contrary, with respective prices of 94.89€ and 103.49€, Carnide and Beato are the most affluent neighborhoods.

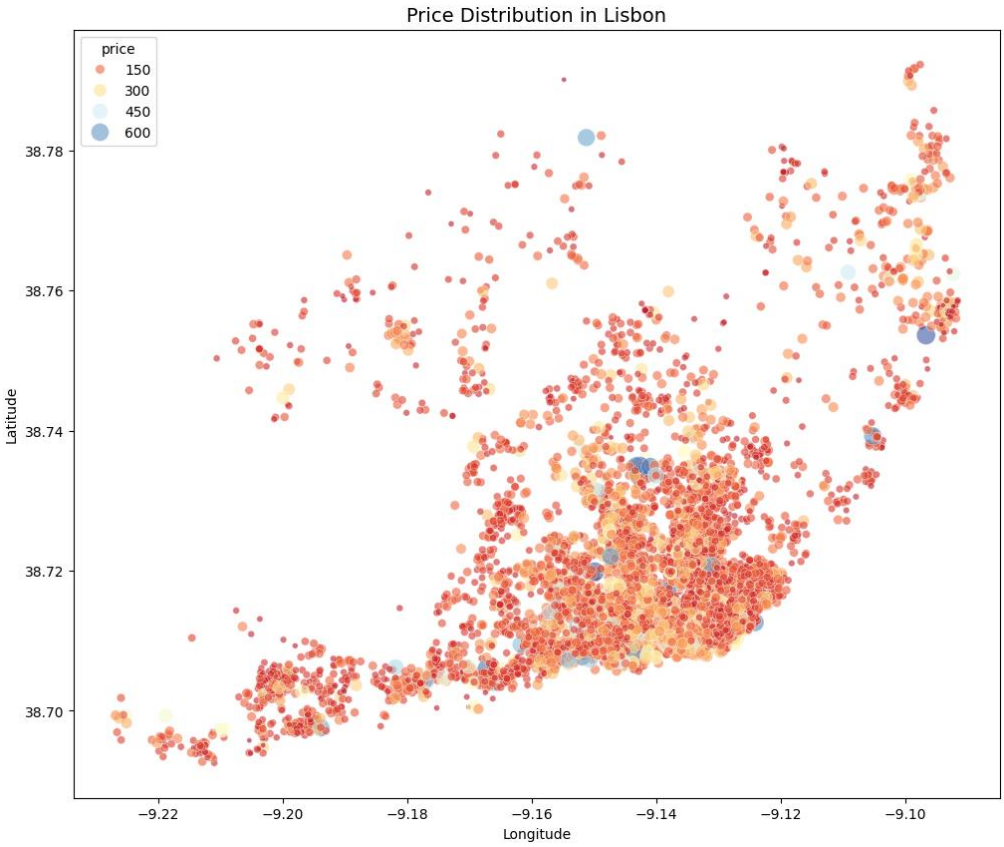

Figure 5 – Map of the Price Distribution in Lisbon

3.3. Modelling

As stated in the literature, Airbnb listings are heterogeneous products with a range of different attributes, and as a result, their utility and customer satisfaction vary. That said, to properly evaluate the factors that Airbnb listing considers, it is necessary to use a hedonic approach. In this section, for the modeling part, 12 variables were selected and grouped into 4 categories: property attributes, host attributes, reputation attributes, and location attributes (Table IV).

Table IV – Modelling categories

Group	Variable	Definition	Unit	Source
Property	accommodates	Number of people the listing can accommodate	numeric	Inside Airbnb
	bathrooms	Number of bathrooms in the listing	numeric	Inside Airbnb
	minimum_nights	Minimum number of nights required	numeric	Inside Airbnb
	maximum_nights	Maximum number of nights allowed	numeric	Inside Airbnb
Host	superhost	Whether the host is a "Superhost": 1 - yes and 0 - no	boolean	Inside Airbnb

Group	Variable	Definition	Unit	Source
Host	professional_host	Whether the host owns or not more than two properties: 1- yes and 0 - no	boolean	Inside Airbnb
Reputation	reviews	Total number of reviews of the listing	numeric	Inside Airbnb
	ratings	The overall rating of the listing	numeric	Inside Airbnb
Location	num_accomodation	Total number of accommodation services within 1 km range	numeric	Overpass Turbo
	num_sports	Total number of sports/leisure services within 1 km range	numeric	Overpass Turbo
	dist_hotspots	Sum of the distance from a listing to the 20 most popular attractions	numeric	OSM, TripAdvisor
	dist_metro	Distance of the listing to the nearest subway entry	numeric	Overpass Turbo

Before beginning the modelling process, standardization was employed due to the disparate scales of the variables. For the purposes of testing for multicollinearity and preventing model distortion, the variance inflation factor (VIF) of each variable was computed from potential correlation between variables (Table V). Since all variables' VIFs were less than 5, multicollinearity was not an issue.

Table V – Variance Inflation Factor

Variables	VIF
price	-
accommodates	1.563368
bathrooms	1.559270
reviews	1.172584
ratings	1.212077
minimum_nights	1.006080
maximum_nights	1.022454
superhost	1.250946
professional_host	1.013845
num_sports	1.190402
num_accomodation	1.812785
dist_hotspots	1.711257
dist_metro	1.486390

This study addresses the issue of setting listing prices with supervised learning algorithms, employing 4 machine learning techniques: OLS, GWR, XGboost, and RF.

Among these four, the OLS is the simplest one, described as a fundamental linear regression technique that minimizes the sum of squared errors to assess the connection

between dependent and independent variables. By using the reduction of the first-order requirements and maximizing the fit of sample points to guarantee low squared residuals, OLS provides a straightforward approach for estimating unknown parameters in a linear model (Wang & Huang, 2023).

Through the explicit inclusion of geographic context and the calculation of parameter estimates for each place of interest, the GWR algorithm loosens the assumption of spatial stationarity (Brunsdon et al., 1996). Presuming that the processes being simulated take place at a certain local scale, GWR determines a generic bandwidth to offer a measure of geographical size. It estimates local regression coefficients at various locations, accounting for spatial heterogeneity.

For increased accuracy, the RF machine learning method builds several decision trees and aggregates their output (Ho, 1995). Using a random selection of the sample of the data, each tree is trained, where the final prediction is determined by either average for regression or majority voting for classification. By improving efficiency and decreasing overfitting, RF becomes an effective tool for a range of prediction applications.

Lastly, the XGBoost is a powerful machine-learning technique for classification and regression applications (Chen, & Guestrin, 2016). To increase overall performance, it generates weak models (typically decision trees) in an orderly manner, with each new tree learning from the mistakes of the ones before it. This method improves prediction accuracy while preserving efficiency by improving the model in the downward gradient direction, using a loss function. It is considered extremely scalable and ideal for big datasets since it allows distributed learning and multi-threading.

3.4. Key Performance Metrics

A predictive model's accuracy and dependability may be measured using a number of important assessment indicators. The following described metrics, which are often employed, especially in regression tasks, were used in this study utilizing the Scikit-Learn library (Pedregosa et al., 2011).

As a measure of how well the model described the variation in the target variable, the R-squared metric demonstrated an enormous fit, which suggests whether the model

was successful in identifying patterns in the data. Conversely, it was taken into account that a high R-squared by itself does not always imply precise predictions.

For regression issues, three error metrics were used. First, with greater weight, given to larger mistakes resulting from squaring, the Root Mean Squared Error (RMSE) calculated the average magnitude of errors between anticipated and actual values. A lower RMSE indicated that the model's predictions were closer to the actual values, making it an important indicator for evaluating overall accuracy. When comparing performance across datasets with different ranges, Mean Absolute Percentage Error (MAPE) was really helpful in expressing prediction errors as a percentage of actual values since it allowed better interpretability. However, while assessing the data, sensitivity to minor real values was taken into consideration. Finally, by computing the average absolute differences between anticipated and actual values, Mean Absolute Error (MAE) offered a clear indicator of the model's prediction performance. This metric is a balanced statistic for assessing model performance since it handles all mistakes equally, contrary to the RMSE metric.

4. RESULTS

This section presents all the results from the analysis of the techniques utilized in section 3. With the interpretation of the coefficients of the linear regression model (Annex E), as well as the feature significance, each category (Table IV) is emphasized with the impact of their variables and their importance that either favorably or unfavorably influence the final pricing calculation.

Property attributes play a significant role in pricing, with the *accommodates* and *bathrooms* having an exceptionally positive effect. Larger facilities often provide guests with more value, so an increase in the *accommodates* results in a significant price increase. Pricing is positively impacted by the number of *bathrooms,* which reflects the increased comfort and convenience. On the other hand, *minimum_nights* have a slight negative coefficient, indicating that listings with more restrictive booking conditions might not be as appealing to prospective guests. However, the little to no impact of *maximum_nights* on pricing suggests that neither perceived value nor demand is greatly impacted by this feature. Also, it is not statistically significant, given that its p-value (0.749) is greater than 0.05, rejecting the null hypothesis.

Price formation is also favourably impacted by the host category, which comprises the *professional_host* and *superhost* variables. A significant benefit of being a superhost is that it highlights the higher prices that guests are ready to pay for reputable, well-regarded hosts. Similarly, listings managed by professional hosts exhibit a minor price rise, perhaps as a result of increased property care, better management techniques, or better guest' experiences.

Reviews and, *ratings*, reputation attributes have conflicting impacts on prices. The correlation for the number of reviews is negative, indicating that listings with more reviews often have lower pricing. This might indicate that properties that are booked regularly serve tourists on a tight budget or even that competition keeps costs down due to increased availability. On the contrary, ratings have a beneficial impact on pricing, demonstrating that high-rated establishments fetch higher costs since guests typically associate high ratings with dependability and quality.

Lastly, location attributes have both favorable and unfavorable effects on costs. *Price* and the *num_accommodation* in a given location are positively correlated, indicating that developed areas with more rental possibilities may provide value or draw

in higher-paying guests. However, the *num_sports* variable in the neighborhood negatively affects pricing, possibly due to the abundance of leisure/sports services that can indicate higher competition among hosts or differing guest preferences, meaning that if the primary customer base of STR in the area does not highly value these amenities, their presence might not justify the higher prices. Additionally, proximity to hotspots (*dist_hotspots*) has a negative effect, since prices tend to drop as one gets farther away from well-known attractions. A similar pattern is observed with the *dist_metro* variable, though its effect it is nearly insignificant, indicating that accessibility via public transportation has minimal bearing on Airbnb pricing in Lisbon's parish. Additionally, it is not statistically significant given that its p-value (0.764) is greater than 0.05, rejecting the null hypothesis.

Overall, the results indicate that all categories influence pricing dynamics, with the size of the property being the variable most impactful on price.

In Table VI is depicted the results of each model using a 5-fold cross-validation:

Table VI – Model's Results

Models	R-squared	MAE	MAPE	RMSE
GWR	49.08%	39.46	29.37%	58.14
OLS	46.73%	41.22	53.59%	60.32
RF	56.73%	36.18	53.49%	54.37
XGBoost	56.57%	36.24	54.32%	54.47

OLS is clearly the model that presents the lowest r-squared since only 46.73% of the price variance is explained by this model (Annex F). With an average prediction error of 41.22 monetary units, it also has the highest MAE, at 41.22. This is further supported by the RMSE of 60.32, which indicates that forecasts are normally off by 60.32 monetary units. A large mean absolute percentage error is also revealed by the MAPE, which is equal to 53.59%. To put it plainly, the OLS model performs poorly in price prediction.

In contrast, the GWR model performs better than OLS. Its forecasts are more accurate, as seen by its RMSE of 58.14, MAE of 38.46, and lowest MAPE of all models, at 29.37%. With an R-squared value of 49.08%, the price volatility can be netter explained. All things considered, these measurements show that GWR is a more reliable and robust model than OLS.

When compared to both OLS and GWR, the XGBoost model exhibits even more performance gains. With a substantially lower RSME of 54.47 and MAE of 36.24, XGBoost produces predictions that are more accurate. Although the MAPE of 54.32% is slightly higher than GWR's, it represents an improvement over OLS. The explanation of price variation has significantly improved, as indicated by the R-squared value of 56.57%.

Lastly, the RF model exhibits a very similar performance to XGBoost. Predictions are far more accurate, as evidenced by its lowest RMSE of 54.37, lowest MAE of 36.18, and MAPE of 53.49%. Furthermore, RF achieves the highest R-squared value of all models at 56.73%, suggesting that 56.73% of the price variance can be explained by the independent variables in the model.

In summary, the RF model emerges as the most effective, given that it has the highest R-squared value and the lowest RMSE and MAE among the four evaluated models. These results show that, by utilizing the characteristics included in the final dataset, RF offers the most accurate price predictions. The model's reasonable R-squared indicates that it can account for a sizable amount of the price variation, which makes it the most solid and trustworthy option for estimating Airbnb pricing in this specific research.

5. DISCUSSION

To place these findings in a larger academic perspective, it is crucial to compare the collected results with the study's conclusions and other research outcomes. The results demonstrate the complexity of pricing dynamics in this particular market and allude to important influencing key factors that the corpus of current knowledge has already confirmed. According to the examined studies, there is a considerable positive relationship between price and property size, especially the number of bathrooms and the amount of people the housing can accommodate. This aligns with the general notion that larger properties are often more expensive. Additionally, factors like the minimum and maximum number of days required for a reservation are not frequently included in price prediction models, which explains their lack of significant impact on cost.

When determining pricing, elements related to reputation are also crucial. It is commonly acknowledged in the literature that while a higher number of positive reviews causes a proportionate price increase, an elevated number of reviews often leads to lower prices. Due to their favorable link with pricing, hosts who have many properties, or the superhost badge, are especially important in price forecasting. There are still some cities where these factors can negatively impact pricing, even though this conclusion is consistent with the majority of studies.

Regarding location-based variables, all researchers agree on their importance, although there are some differences in which factors are considered and how they affect price. For example, variables like the number of accommodation services close to a property and its distance from the closest metro station provide inconsistent results. This study aligns with Zhao, H. (2023), who discovered that the number of local lodgings positively impacted pricing while the distance to the metro was not a particularly significant element. However, research by Zhao et al. (2023) in a different city concluded that metro proximity was not only relevant but also had a promising influence on price, whereas a higher number of nearby accommodations had a negative effect on pricing. This does not necessarily imply that one research study is incorrect; rather, it suggests that the same outcomes may be reversed in different cities under specific circumstances.

In line with previous research, variables related to the distance from tourist attractions are typically linked to positive price variations. Contrary to earlier findings, the number of leisure activities near a lodging was found to negatively impact price in

this study, possibly due to context-specific factors, as a result of increased competitiveness among hosts brought on by the abundance of leisure and sports options.

Table VII – Results of previous studies

Articles	Models	R-squared	MAE	RMSE
Zhao et al. (2023)	OLS	44.30%	-	-
	GWR	61.80%	-	-
	MGWR	73.30%	-	-
Iliopoulou et al. (2024)	OLS	38.50%	-	-
	GWR	53.60%	-	-
Wang and Huang (2023)	OLS	5.20%	-	2224.60
	RF	12.40%	-	2650.31
	XGBoost	6.00%	-	2423.51
Lektorov et al. (2023)[1]	RF	63.00%	0.30	0.42
	KNN	56.00%	0.33	0.45
	SVM	62.00%	0.30	0.43
	Decision tree	47.00%	0.36	0.50
	Extra Tree	61.00%	0.30	0.43
	XGBoost	63.00%	0.30	0.42

On the modeling component, a comparison of this work with earlier research demonstrates significant differences in model accuracy and the types of models used. When assessing the effectiveness of predictive models, the coefficient of determination, or R-squared, is crucial. Larger values indicate more accuracy and a more thorough explanation of price variation by the variables utilized.

Examining the table above (Table VII), several models on Zhao et al. (2023) and Lektorov et al. (2023) research outperform the one employed in this study. However, like this research, several additional studies support the usefulness of ensemble approaches in price forecasting. This is evident in the excellent performance of the RF model and XGBoost across different investigations, reinforcing their ability to capture complex, non-linear relationships. Nonetheless, it is crucial to note that comparing accuracy metrics across research might be challenging. Results may be influenced by differences in market situations, methods used to process price variables (such as logarithmic transformation), and datasets. Therefore, while this model demonstrates strong performance in its particular setting, its results should be used as a starting point for further study rather than as a final standard.

[1] Used a logarithmic transformation on price

6. Conclusions, Limitations, and Future Research

6.1. Conclusions

Using a variety of machine learning models, such as RF, OLS, GWR, and XGBoost, this study aimed to predict Airbnb prices in Lisbon's parish. After a thorough examination and assessment of the model's performance, the RF model emerged as the most dependable, demonstrating the best capacity to explain price variation and the highest predicted accuracy in the dataset. However, it is crucial to remember that the conclusions obtained in the results are exclusive to this dataset and might not apply generally to other markets, locations, or situations.

In order to ensure a methodical approach to data collecting, processing, modeling, and assessment, this research was based on the CRIPS-DM methodology. A wide range of variables that affect Airbnb prices were also considered and divided into four major categories: location-based criteria, reputation indicators, property traits, and host-related qualities. This categorization provided a thorough understanding of the different attributes that contribute to the price swings in the market for STRs.

Additionally, as discussed in the literature review chapter, this study was carried out in the large framework of previous studies on Airbnb pricing. This study adds to the current discussion on pricing factors in the STR market by combining knowledge from earlier research with actual results from Lisbon's parish dataset defined. When examining pricing trends and market dynamics, machine learning algorithms, like the ones employed in this study, can be useful resources. Their use goes beyond specific price schemes and may be advantageous to policymakers attempting to strike a balance between housing accessibility, social equity, and economic opportunities in the sharing economy landscape.

That said, it is reasonable to assume that Lisbon may continue to strive toward preserving its solid standing as a popular travel destination while tackling important issues with affordable housing and neighborhood integrity. The parish can continue to take advantage of the opportunities provided by the Airbnb industry without sacrificing the welfare of its residents, finding a balance between social responsibility and economic gains.

6.2. Limitations

As with any research, this study on predicting Airbnb prices in Lisbon's parish has several limitations that must be noted. One of the main limitations is that the research was conducted using data from a certain time period, and because of that, seasonal price variations and alternations in the relative importance of particular property attributes over time were not specifically taken into consideration. The Airbnb market is extremely dynamic, and factors like local events, seasonal demand, and economic conditions affect prices. Nevertheless, the absence of these temporal elements in the study may have limited how far the results may be applied. Also, external variables such as global crises, economic downturns, or government regulations related to STRs also have a big impact on Airbnb pricing, and their lack of use in the dataset may prejudice the predictive models.

Methodological limitations also influence the study's results. Within every machine learning model examined, the RF model was the most dependable one. However, experimenting with deep learning techniques, ensemble learning or hyperparameter tuning may reveal other ways to enhance model performance. Additionally, the MGWR model, which may have offered more profound geographical insights, was not used due to technological issues. That said, by overcoming computational obstacles or using different spatial modeling approaches, future research might close this gap.

The dataset itself is the source of another restriction. A large percentage of missing values led to the removal of certain potentially significant data points. Although this choice was required to preserve the integrity of the data, it could have left out important data that could have improved the accuracy of the model. Furthermore, the dataset lacked important characteristics, including socioeconomic aspects, that might help solve the endogeneity issue, which is commonly mentioned in comparable research.

Finally, potential discrepancies are introduced by the study's reliance on publicly accessible data. Since some hosts use different pricing schemes or operate on numerous platforms, Airbnb listings might not be an accurate representation of all STR lodgings in the city studied.

Despite these limitations, which are essential to recognize in order to evaluate the findings and direct future research, the study's results offer insightful information about the factors influencing Airbnb prices in Lisbon's parish.

6.3. Future research

By addressing the limitations specified previously, adding more variables, experimenting with other modeling techniques, and investigating policy actions that maximize the cohabitation of tourists and local housing requirements, future research should be built upon this basis.

One key area for improvement would be adding temporal dynamics to the analysis. Future research might employ time-series models or dynamic pricing algorithms to capture these variances and enhance price predictions accurately.

Another promising avenue is the inclusion of other factors that were not accessible in this study. A more thorough knowledge of pricing drivers may be possible by taking into account socioeconomic elements, such as income and education. Including textual elements in the models through the use of NLP approaches would also be interesting. The most intriguing way to put this into practice would be to extract the most frequently used terms from listing descriptions to find important characteristics that drive demand or to do a sentiment analysis on guest reviews to evaluate customer satisfaction and its impacts on price.

Future research might also examine deep learning approaches, including NNs to compare their performance with traditional machine learning models. Additionally, the comprehension of localized pricing drivers may be improved by effectively applying the MGWR model, which would further boost the geographical component of the analysis.

Lastly, expanding the study's geographic focus to include comparisons between Lisbon's parish and other locations may provide a more broadly global view of the STR market and the factors that influence their price. Cross-city comparative analysis could highlight differences and similarities in pricing factors across diverse urban areas.

As a result, this study serves as an important contribution to the growing corpus of knowledge on Airbnb price prediction, but it should be seen as a part of a much larger and more comprehensive research.

REFERENCES

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000)"CRISP-DM 1.0: Step-by-step data mining guide", SPSS inc, vol. 9, pp. 13, 2000.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM. https://doi.org/10.1145/2939672.2939785

Chen, Y., & Xie, K. (2017). Consumer valuation of Airbnb listings: A hedonic pricing approach. *International Journal of Contemporary Hospitality Management*, *29*(9), 2405–2424. https://doi.org/10.1108/ijchm-10-2016-0606

Cocola-Gant, A., & Gago, A. (2021). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in Lisbon. *Environment and Planning a Economy and Space*, 53(7), 1671–1688. https://doi.org/10.1177/0308518x19869012

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost Data Science. 2022 *17th Iberian Conference on Information Systems and Technologies* (CISTI), 1–6. https://doi.org/10.23919/cisti49556.2020.9140932

Costa, C.J., Aparicio, J.T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In: Arabnia, H.R., et al. Advances in Parallel & Distributed Processing, and Applications. Transactions on Computational Science and Computational Intelligence. Springer, Cham. https://doi.org/10.1007/978-3-030-69984-0_7

Costa, C., Stević, I., Veríssimo, M., & Da Silva, M. F. (2019). Short-term accommodation and overtourism in Portuguese urban centres. In *CAB International eBooks* (pp. 167–189). https://doi.org/10.1079/9781786399823.0167

Deboosere, R., Kerrigan, D. J., Wachsmuth, D., & El-Geneidy, A. (2019). Location, location and professionalization: A multilevel hedonic analysis of Airbnb listing prices and revenue. *Regional Studies Regional Science*, 6(1), 143–156. https://doi.org/10.1080/21681376.2019.1592699

Di Persio, L., & Lalmi, E. (2024). Maximizing profitability and occupancy: An optimal pricing strategy for Airbnb hosts using regression techniques and natural language processing. *Journal of Risk and Financial Management*, *17*(9), 414. https://doi.org/10.3390/jrfm17090414

Dogru, T., & Pekin, O. (2017). What do guests value most in Airbnb accommodations? An application of the hedonic pricing approach. *Boston Hospitality Review*, *5*(2). https://vtechworks.lib.vt.edu/handle/10919/79602

Eugenio-Martin, J. L., Cazorla-Artiles, J. M., & González-Martel, C. (2019). On the determinants of Airbnb location and its spatial distribution. *Tourism Economics*, 25(8), 1224–1244. https://doi.org/10.1177/1354816618825415

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2017). Pricing in the sharing economy: A hedonic pricing model applied to Airbnb listings. *Journal of Travel & Tourism Marketing*, 35(1), 46–56. https://doi.org/10.1080/10548408.2017.1308292

Gutiérrez, A., & Domènech, A. (2020). Understanding the spatiality of short-term rentals in Spain: Airbnb and the intensification of the commodification of housing. *Geografisk Tidsskrift-Danish Journal of Geography*, *120*(2), 98–113. https://doi.org/10.1080/00167223.2020.1769492

Gyódi, K. (2021). Airbnb and hotels during COVID-19: Different strategies to survive. *International Journal of Culture Tourism and Hospitality Research*, *16*(1), 168–192. https://doi.org/10.1108/ijcthr-09-2020-0221

Gyódi, K., & Nawaro, Ł. (2021). Determinants of Airbnb prices in European cities: A spatial econometrics approach. *Tourism Management*, *86*, 104319. https://doi.org/10.1016/j.tourman.2021.104319

Hamari, J., Sjöklint, M., & Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, *67*(9), 2047–2059. https://doi.org/10.1002/asi.23552

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Hong, I., & Yoo, C. (2020). Analyzing spatial variance of Airbnb pricing determinants using multiscale GWR approach. *Sustainability*, *12*(11), 4710. https://doi.org/10.3390/su12114710

Iliopoulou, P., Krassanakis, V., Misthos, L., & Theodoridi, C. (2024). A spatial regression model for predicting prices of Short-Term rentals in Athens, Greece. ISPRS

International *Journal of Geo-Information*, 13(3), 63. https://doi.org/10.3390/ijgi13030063

Lang, B., Kemper, J., Dolan, R., & Northey, G. (2022). Why do consumers become providers? Self-determination in the sharing economy. *Journal of Service Theory and Practice*, 32(2), 132–155. https://doi.org/10.1108/jstp-09-2020-0220

Lee, K., Hakstian, A., & Williams, J. D. (2021). Creating a world where anyone can belong anywhere: Consumer equality in the sharing economy. *Journal of Business Research, 130*, 221–231. https://doi.org/10.1016/j.jbusres.2021.03.036

Lektorov, A., Abdelfattah, E., & Joshi, S. (2023). Airbnb rental price prediction using machine learning models. 2022 *IEEE 12th Annual Computing and Communication Workshop and Conference* (CCWC), 0339–0344. https://doi.org/10.1109/ccwc57344.2023.10099266

Lestegás, I., Seixas, J., & Lois-González, R. (2019). Commodifying Lisbon: A Study on the Spatial Concentration of Short-Term Rentals. *Social Sciences*, 8(2), 33. https://doi.org/10.3390/socsci8020033

Lorga, M., Januário, J. F., & Cruz, C. O. (2022). Housing Affordability, Public Policy and Economic Dynamics: An analysis of the City of Lisbon. *Journal of Risk and Financial Management*, 15(12), 560. https://doi.org/10.3390/jrfm15120560

Milunovich, G., & Nasrabadi, D. (2025). Airbnb pricing in Sydney: Predictive modelling and explainable machine learning. *Applied Economics*, 1–18. https://doi.org/10.1080/00036846.2024.2446593

Minami, A. L., Ramos, C., & Bortoluzzo, A. B. (2021). Sharing economy versus collaborative consumption: What drives consumers in the new forms of exchange? *Journal of Business Research, 128*, 124–137. https://doi.org/10.1016/j.jbusres.2021.01.035

Pedregosa et al (2011), Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55. https://doi.org/10.1086/260169

Samadani, S., & Costa, C. J. (2021). Forecasting real estate prices in Portugal: A data science approach. In 2021 16th Iberian Conference on Information Systems and Technologies pp 1-6. IEEE. https://doi.org/10.23919/CISTI52073.2021.9476447

Sanasi, S., Ghezzi, A., Cavallo, A., & Rangone, A. (2020). Making sense of the sharing economy: A business model innovation perspective. *Technology Analysis and Strategic Management*, 32(8), 895–909. https://doi.org/10.1080/09537325.2020.1719058

Ter Huurne, M., Ronteltap, A., Corten, R., & Buskens, V. (2017). Antecedents of trust in the sharing economy: A systematic review. *Journal of Consumer Behaviour*, *16*(6), 485–498. https://doi.org/10.1002/cb.1667

Wang, Q., & Huang, L. (2023). Machine Learning-Based Study on Airbnb Housing Price Prediction—Evidence from Hong Kong. *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*, 889–896. https://doi.org/10.1145/3650400.3650551

Wang, R., & Rasouli, S. (2022). Contribution of streetscape features to the hedonic pricing model using Geographically Weighted Regression: Evidence from Amsterdam. *Tourism Management*, *91*, 104523. https://doi.org/10.1016/j.tourman.2022.104523

Zhao, C., Wu, Y., Chen, Y., & Chen, G. (2023). Multiscale effects of hedonic attributes on Airbnb listing prices based on MGWR: A case study of Beijing, China. *Sustainability*, 15(2), 1703. https://doi.org/10.3390/su15021703

Zhan, W., Hu, Y., Zeng, W., Fang, X., Kang, X., & Li, D. (2024). Total least squares estimation in Hedonic house price models. ISPRS International Journal of Geo-Information, 13(5), 159. https://doi.org/10.3390/ijgi13050159

Zhao, H. (2023). Influence of built environment features on Airbnb listing price and the spatio-temporal heterogeneity: An empirical study from Copenhagen, Denmark. *Geografisk Tidsskrift-Danish Journal of Geography*, *123*(1), 42–60. https://doi.org/10.1080/00167223.2023.2275037

APPENDIX

Annex A – Variable Description

Variables	Description
id	Identifier of each listing
listing_url	URL to the listing page
scrape_id	ID for each data scrape
last_scraped	The date the data was last scraped from the website
source	The source from which the listing was gathered
name	The title of the listing
description	A description of the property
neighborhood_overview	Overview of the neighborhood where the property is located
picture_url	URL to the main picture of the listing
host_id	ID for the host
host_url	URL to the host's profile
host_name	The name of the host
host_since	The year the host first joined Airbnb
host_location	The location of the host
host_about	Description about the host
host_response_time	The time taken by the host to respond
host_response_rate	The percentage of responses from the host
host_acceptance_rate	The percentage of booking requests the host accepts
host_is_superhost	Whether the host is a "Superhost" (Yes/No), which is a mark of quality for the top-rated and most experienced hosts
host_thumbnail_url	URL to the host's profile thumbnail
host_picture_url	URL to the host's profile picture
host_neighbourhood	The neighborhood the host is located in
host_listings_count	Number of listings the host has on Airbnb
host_total_listings_count	Total number of listings the host manages
host_verifications	List of verification methods the host has undergone
host_has_profile_pic	Whether the host has a profile picture (Yes/No)
host_identity_verified	Whether the host's identity is verified (Yes/No)
neighbourhood	The neighborhood where the property is located
neighbourhood_cleansed	The cleaned name of the neighborhood
neighbourhood_group_cleansed	Neighborhoods parish
latitude	Latitude of the listing's geographical location
longitude	Longitude of the listing's geographical location
property_type	The type of property (apartment, house)
room_type	The type of room (entire home/apt, private room, shared room)

Variables	Description
accommodates	Number of people the listing can accommodate
bathrooms	Total number of bathrooms
bathrooms_text	Descriptive text about the bathrooms
bedrooms	Number of bedrooms in the listing
beds	Total number of beds
amenities	A list of amenities available in the listing (eg, Wi-Fi, parking)
price	Price per night
minimum_nights	Minimum number of nights required
maximum_nights	Maximum number of nights allowed
minimum_minimum_nights	The minimum value for minimum nights across all listings in the dataset
maximum_minimum_nights	The maximum value for minimum nights across all listings in the dataset
minimum_maximum_nights	The minimum value for maximum nights across all listings in the dataset
maximum_maximum_nights	The maximum value for maximum nights across all listings in the dataset
minimum_nights_avg_ntm	The average minimum number of nights for listings
maximum_nights_avg_ntm	The average maximum number of nights for listings
calendar_updated	Date when the calendar was last updated
has_availability	Whether the listing has availability for booking
availability_30	Number of available nights in the next 30 days
availability_60	Number of available nights in the next 60 days
availability_90	Number of available nights in the next 90 days
availability_365	Number of available nights in the next 365 days
calendar_last_scraped	Date when the calendar data was last scraped
number_of_reviews	Total number of reviews the listing
number_of_reviews_ltm	Number of reviews received in the last 12 months
number_of_reviews_l30d	Number of reviews received in the last 30 days
first_review	The date of the first review
last_review	The date of the last review
review_scores_rating	The overall rating of the listing based on reviews
review_scores_accuracy	The rating of the listing's accuracy
review_scores_cleanliness	The rating of the listing's cleanliness
review_scores_checkin	The rating of the listing's check-in experience
review_scores_communication	The rating of the listing's communication
review_scores_location	The rating of the listing's location
review_scores_value	The rating of the listing's value
license	The license under which the property is registered
instant_bookable	Whether the listing can be booked instantly (Yes/No)
calculated_host_listings_count	The number of listings the host has, calculated from the dataset

Variables	Description
calculated_host_listings_count_entire_homes	Number of entire home listings the host has
calculated_host_listings_count_private_rooms	Number of private room listings the host has
calculated_host_listings_count_shared_rooms	Number of shared room listings the host has
reviews_per_month	The average number of reviews per month for the listing

Annex B – Data Transformation

```
# Rename columns to simplify names
df1 = df1.rename(columns={'host_is_superhost': 'superhost'})
df1 = df1.rename(columns={'number_of_reviews': 'reviews'})
df1 = df1.rename(columns={'review_scores_rating': 'ratings'})

# Create a new column 'professional_host' indicating if a host has 2 or more listings (1 if yes, 0 if no)
df1['professional_host'] = df1['host_total_listings_count'].apply(lambda x: 1 if x >= 2 else 0)

# Convert 'superhost' column values: replace 'f' with 0 (not superhost) and 't' with 1 (is a superhost)
df1['superhost'] = df1['superhost'].replace({'f': 0, 't': 1})

# Clean and convert the 'price' column: remove dollar signs and commas, then convert to float type
df1['price'] = df1['price'].replace({'\$': '', ',': ''}, regex=True).astype(float)
# Convert 'bathrooms' column from float to integer type for consistency
df1['bathrooms'] = df1['bathrooms'].astype(int)

exchange_rate = 0.9683 #4th february 2025
# Converting dollar to euro
df1["price"] = df1["price"] * exchange_rate
```

Annex C – Num_accommodation variable

```
def calculate_num_accomodation(listings_df, hotels_POIs_df):

    # Convert listings and hotels POIs to GeoDataFrames
    listings_gdf = gpd.GeoDataFrame(
        listings_df,
        geometry=gpd.points_from_xy(listings_df.longitude, listings_df.latitude),
        crs="EPSG:4326"  # WGS84 lat/lon CRS
    )

    hotels_gdf = gpd.GeoDataFrame(
        hotels_POIs_df,
        geometry=gpd.points_from_xy(hotels_POIs_df.longitude, hotels_POIs_df.latitude),
        crs="EPSG:4326"
    )

    # Project to a metric CRS (e.g., EPSG:3857)
    listings_gdf = listings_gdf.to_crs(epsg=3857)
    hotels_gdf = hotels_gdf.to_crs(epsg=3857)

    # Buffer each listing by 1 km and count hotels POIs within the buffer
    listings_gdf['num_accomodation'] = listings_gdf.geometry.apply(
        lambda x: hotels_gdf[hotels_gdf.geometry.within(x.buffer(1000))].shape[0]
    )

    return listings_gdf[['latitude', 'longitude', 'num_accomodation']]

# Calculate num_accomodation
listings_with_num_accomodation = calculate_num_accomodation(df1, gdf2)
```

Annex D – Mean Price per neighbourhood

Neighbourhood	Mean Price
Santo António	163.599728
Misericórdia	162.312102
Santa Maria Maior	161.991578
Parque das Nações	157.713164
Avenidas Novas	153.814175
Santa Clara	151.746443
Arroios	149.558336
Areeiro	139.156814
Estrela	134.621578
Campo de Ourique	131.162087
São Vicente	129.705026
Marvila	129.299488
São Domingos de Benfica	128.530652
Olivais	126.325908
Penha de França	124.271622
Alcântara	122.918944
Belém	118.116726
Campolide	118.001956
Alvalade	117.473757
Lumiar	113.351619
Ajuda	111.420371
Benfica	109.560297
Beato	103.485358
Carnide	94.8934

Annex E – OLS coefficient interpretation

OLS Regression Results

Dep. Variable:	price	R-squared:	0.470
Model:	OLS	Adj. R-squared:	0.469
Method:	Least Squares	F-statistic:	611.7
Date:	Sun, 16 Feb 2025	Prob (F-statistic):	0.00
Time:	23:06:38	Log-Likelihood:	-45782.
No. Observations:	8298	AIC:	9.159e+04
Df Residuals:	8285	BIC:	9.168e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	148.2811	0.662	224.070	0.000	146.984	149.578
accommodates	32.3542	0.827	39.102	0.000	30.732	33.976
bathrooms	25.6014	0.826	30.981	0.000	23.982	27.221

reviews	-9.3381	0.717	-13.031	0.000	-10.743	-7.933
ratings	9.0960	0.729	12.485	0.000	7.668	10.524
minimum_nights	-3.6382	0.664	-5.481	0.000	-4.939	-2.337
maximum_nights	0.2145	0.669	0.320	0.749	-1.097	1.526
superhost	7.6077	0.740	10.279	0.000	6.157	9.059
professional_host	2.8949	0.666	4.345	0.000	1.589	4.201
num_sports	-3.3436	0.722	-4.631	0.000	-4.759	-1.928
num_accomodation	15.4044	0.891	17.289	0.000	13.658	17.151
dist_hotspots	-2.2072	0.866	-2.550	0.011	-3.904	-0.510
dist_metro	0.2423	0.807	0.300	0.764	-1.339	1.824

Omnibus:	3656.438	Durbin-Watson:	1.783
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30600.687
Skew:	1.912	Prob(JB):	0.00
Kurtosis:	11.596	Cond. No.	2.41

Annex F – OLS modeling

```
# Extract dependent variable (Target)
y = np.array(df1['price']).reshape((-1, 1))

# Extract independent variables
X1 = df1[['accommodates', 'bathrooms', 'reviews', 'ratings','minimum_nights', 'maximum_nights',
          'superhost', 'professional_host', 'num_sports',
          'num_accomodation', 'dist_hotspots', 'dist_metro']]

# Standardize features
scaler = StandardScaler()
X2 = pd.DataFrame(scaler.fit_transform(X1))

# Convert to numpy array
X = np.array(X2)

# Add intercept column for OLS
X = sm.add_constant(X)  # Adds a column of ones for the intercept term

# Set up cross-validation
kf = KFold(n_splits=5, shuffle=True, random_state=42)

# Store evaluation metrics
rmse_scores = []
mae_scores = []
r2_scores = []
mape_scores = []

# Perform k-fold cross-validation
for train_index, test_index in kf.split(X):
    X_train, X_test = X[train_index], X[test_index]
    y_train, y_test = y[train_index], y[test_index]

    # Train OLS model
    ols_model = sm.OLS(y_train, X_train).fit()

    # Predict on test data
    y_pred = ols_model.predict(X_test)

    # Calculate metrics
    rmse_scores.append(np.sqrt(mean_squared_error(y_test, y_pred)))
    mae_scores.append(mean_absolute_error(y_test, y_pred))
    r2_scores.append(r2_score(y_test, y_pred))
    mape_scores.append(mean_absolute_percentage_error(y_test, y_pred))
```