NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

House Price Prediction Model: Incorporating house locations on decision trees-based ensemble methods

André Bartolomeu Rodrigues Dias

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

House Price Prediction Model: Incorporating house locations on decision trees-based ensemble methods

by

André Bartolomeu Rodrigues Dias

Dissertation presented as a partial requirement for obtaining the master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Professor Vítor Santos

November 2023

Abstract

This work explores the dynamics of the short-term rent market, with a specific focus on building a robust price prediction model for Airbnb listings. The model is designed to encompass various dimensions, including house features, host attributes, reviews scores and external variables. It focuses on a particular set of machine learning models called Ensemble Models, in particular the Random Forest, the Gradient Boosting and the XGBoost. It investigates the impact of incorporating location-related features, including neighborhood characteristics and distances to key landmarks such as the city center, the subway stations, and the Lisbon coast, within these type of Machine Learning models. The dissertation aims to adopt a pragmatic methodology based on machine learning principals. The findings highlight that features related with the house dimensions carry the most significant weight when predicting prices, while the calculated distances features exhibit a limited influence on defining the target variable. In essence, the best model, XGBoost encounters challenges in comprehending underlying patterns and explaining the decisions behind the criteria for the nightly prices chosen by the host with the considered features. This work contributes to understanding the dynamics shaping pricing strategies and emphasizes the complexity of host decision-making processes in this context.

Keywords

Airbnb; Distances; Ensemble Models; House; Hosts; Lisbon; Listings; Location; Machine Learning; Platform; Prediction; Price.

List of Acronymous

API	Application Programming Interface
CRS	Coordinate Reference System
GBM	Gradient Boosting Model
GWR	Geographic Weighted Regression
LTM	Last 12 months
LTR	Long-Term Rent
MAE	Mean Absolute Error
MDI	Mean Decrease in Impurity
ML	Machine Learning
MSE	Mean Squared Error
P2P	Peer-to-Peer
RFM	Random Forest Model
RFE	Recursive Feature Selection
RMSE	Root Mean Square Error
SSR	Sum of Squared Residuals
STR	Short-term rental

Glossary of Terms

Demand – The number of available properties that are being booked by guests. Comparing supply against demand determines occupancy rates.

Dwelling - A separate and independent place that was built, rebuilt, enlarged, or converted to be used as a private accommodation, and that is not totally occupied for other purposes during the reference period.

Market - The geographic area a vacation rental is located in. This can be anything from a large city to a specific neighborhood. It can also be a type of market, such as "urban" or "mountain/resort."

Median value per m² of dwellings sales - Median value of prices per square meter of transactions by sale of dwellings for residential purposes with private gross area greater than or equal to 20 m².

Nightly Rate - How much a vacation rental host charges for a single night. Nightly rates often vary per night based on demand, day of the week, seasonality, etc. The terms Daily Rate and Nightly Rate are often used interchangeably.

Occupancy rate - A percentage showing how often a property is booked by comparing the number of total booked days versus total available days, often measured by the last 12 months (LTM). A key metric when assessing the success of a short-term rental (STR) property or market.

Short-term rental - A property intended for shorter stays, often less than 30 days.

Supply - Total number of available short-term rental (STR) listings.

Source: Utimate Short-Term Vacation Rental Glossary | AirDNA, 2023

List of Figures

List of Tables

Index

1 Introduction

1.1 Background

A short-term rental is a house, apartment, or room that is rented from one night to 28 days (Ding et al., 2023; Guttentag, 2019; Jiang et al., 2023; Z. Zhang et al., 2017). On average, short-term rentals are occupied for a few days at a time. It has the potential for greater earnings due to higher prices compared to contracts for long-term rentals. Seasonality, holidays, and events such as concerts and football games, attract many people monthly hoping to get a flexible and easy way of having a home during their stay. On the side of a property owner, this means flexibility to determine the nightly rates depending on this demand, which ultimately can lead to an increase in profits. Depending on the city of your rental property, you may have tax benefits according to the number of days you rent per year and be allowed to deduct expenses related to insurance, maintenance, and utilities.

The advent of the sharing economy has revolutionized the way people travel and seek accommodation (Hati et al., 2021). This study will focus on one of the most popular platforms for short-term renting, Airbnb, with more than 7 million listings throughout the world. Recent Airbnb statistics show that they are represented in more than 220 countries and regions, and they have listings in more than 100,000 cities globally according to the company website. The main goal of the company is to provide people with an authentic experience compared to their main competitors. The company states that in most countries, they have “more competitive rates than hotels”, becoming even more desirable for a diverse type of travelers. The sharing accommodation platform has fundamentally altered market segmentation within the industry.

The website operates as an online marketplace for people who are looking for accommodations and hosts to rent their properties or rooms to guests. For guests, Airbnb gives affordable temporary housing options and for hosts, it provides a convenient source of income to earn extra money. Now, the company itself makes money by being a mediator between renters and future guests: every time a guest books the host’s property, the host pays Airbnb a fee and it also adds service fees depending on the total cost of the guest’s stay. It allows individuals to monetize their spare space, transforming their homes into lucrative assets. However, as the platform has evolved, it has raised critical questions about pricing strategies, market competition, and the factors that influence property values within the Airbnb marketplace.

1.1.1 Portuguese Housing Market

A study published in the Journal of Housing Economics by Horn & Merante (2017), found that an increase in Airbnb listings in Boston (a city in Massachusetts, USA) was associated with an increase in rent prices of 0.4% which as a consequence, raises the cost of living for local renters. A similar study (Barron et al., 2021), focused on listings in the entire USA country, found a more specific conclusion: a 1% increase in Airbnb listings leads to a 0.018% increase in rents and a 0.026% increase in house prices. During the summer of 2022, 80% of the landlords in Portugal switched their properties from long-term rentals, which are aimed at residents and students, to short-term rentals. It is known that not only inflation was the main reason for their switch, but also the growth of the so-called "digital nomads" and the tourism upturn supports their decision. As in every city, Lisbon has a finite supply of housing, so this process will eventually drive-up rental rates over time. A study conducted on the correlation between short-term rental (STR) and long-term rental (LTR) shows that the mean rental price per bedroom of both Airbnb and Zoopla (Long-term rent platform) are highly correlated, with a result of 0.86. It is also observed a positive linear correlation between the number of Airbnb listings and LTR price (r of 0.70). This indicates that areas with a high Airbnb supply are mostly located in highly priced residential areas (Shabrina & Morphet, 2022). The results also suggest that professional hosts invest in specific city areas and provide accommodation at higher price points. In 2021, findings support that the uncontrolled growth of Airbnb supply may lead to stronger negative processes, including the gentrification of neighborhoods (Gyódi & Nawaro, 2021). To contain these harmful effects, governments may consider regulations that enable occasional home-sharing (Nieuwland & van Melik, 2020).

This study also arises from the premise that it is very challenging to obtain an accurate value for the nightly price since it is well-known that is susceptible to multiple factors (Yang, 2022). Thus, it is important to collect related information and transform them into data, so that it can be used in suitable models to analyze and evaluate the results in an increasingly better way. As users continue to grow on both the supply and demand side, homeowners may find it harder to properly price their property as time passes. It is known that the key to success for a real estate investment business is finding the right property in a top-market location. The same holds true for investing in an Airbnb rental property. Through extensive analyses like the examples gathered above, this study will help to understand how certain house characteristics can be combined with external factors to provide a more optimal price within a specific time. The focus will not only be on someone willing to invest in a property, but also on a customer's perspective, hence the significant study fields taken into consideration. Plus, the following exploration can also be applied to local accommodations and even hotels with the right set of changes according to the type of business.

1.2 Problem Definition – A two-side equilibrium

The problem at hand revolves around the dynamic and multifaceted domain of short-term rental pricing, specifically in the context of online platforms such as Airbnb. One of the central issues is the lack of a consistent and data-driven approach for hosts to set appropriate rental prices. While experience and intuition are valuable, they often fall short in optimizing pricing strategies. The rental property price is based on the nightly rate decided by the host, plus fees and extra costs. The calculations are mostly based on two factors: a simple market research where the property owner searches for listings in the area to find out what they charge and the fees that come with the renting process, such as the cleaning fee and the service cost (Airbnb, 2022). This approach could be seen as being naïve, because without the right set of features to complete a more elaborate analysis, hosts risk underpricing their listings, leading to missed revenue opportunities, or overpricing, potentially deterring potential guests.

This problem is intricate due to various influencing factors that impact pricing decisions. Building such model would not only benefit the individual property owners (host), but also contribute to the broader discussion on responsible data-driven decision-making in the context of the sharing economy. With the right data, a tool for the hosts can be provided with immense potential to perform accurate price predictions.

1.3 Objectives

The main goal will be to analyze and understand how the use of attributes related to house features, host attributes, reviews scores, and distances to points of interest, can be deterministic to accurately predict the optimal listing price for both the host and guests. One of the main goals is for the home providers to reach an equilibrium price that optimizes profit and affordability. It is also increasingly more relevant for people to realize what features of an Airbnb listing are most determinant when trying to find a place to stay.

This research aims to develop a predictive model that not only predicts Airbnb listing prices accurately but also uncovers the underlying patterns and features that contribute to these price differentials. The heterogeneous dataset used in this study includes essential variables such as the number of accommodations, number of reviews, neighborhood median prices, host status, distance to the city center, and many more. Through rigorous analysis and modeling, this study aims to advance the understanding of the evolving landscape of short-term accommodation market. Furthermore, the model should provide accurate predictions and also offer interpretable insights into the factors that drive pricing decisions, enabling hosts to make informed choices.

The aim of the paper will be to create a model that predicts the most optimal price for a short-term rental property according to the house attributes, host characteristics, reviews scores, and geolocation distances to points of interest.

To achieve this goal, the following objectives were defined:

1. Conduct a comprehensive literature review on the short-term house renting market and machine learning ensemble methods to understand the current state of the industry and the application of Machine Learning (ML) techniques in the field of pricing prediction.
2. Study the influence of each feature on the price set of short-term rentals. Compare the effects of these same features to identify which factors have the most significant impact.
3. Identify the influence of external factors that make fluctuate the overall nightly price rates of short-term house rentals and explore the results on decision trees-based ensemble models when adding distances to points of interest and neighborhood economic factors.
4. Propose and develop a machine learning model for price prediction based on an ensemble method with decision trees. Utilize the insights gained from feature analysis and external factors to later select an accurate and robust model for predicting short-term rental prices.

1.4 Study Impact

Tourism is a vital part of the Portuguese economy. In 2021, it contributed 16.8 billion euros to the national GDP, directly and indirectly. The value reflects 8% of the total national GDP according to data released by the National Institute of Statistics (Statistics Portugal - Web Portal, 2022) which published the preliminary estimate of the sector for the same year. Accommodation takes a big percentage of that value, so it is important that when traveling, people can have several options to stay in an affordable house based on their own budget.

The traditional hotel industry adopts relatively fixed pricing methods, but Airbnb listing prices are entirely determined by hosts (Z. Zhang et al., 2017). The rise of Airbnb has been disrupting conventional lodging providers, challenging established business models, and posing intriguing questions about the coexistence and competition between these two sectors. The appearance of sharing accommodation platforms has brought about a significant transformation in the hospitality industry as the competition intensifies between the two industries (Dogru et al., 2020). With a wider selection of accommodation choices, travelers have become more sensitive to price, and consequently, hotels have been compelled to reevaluate their pricing strategies to maintain competitiveness, potentially leading to a reduction in their average daily rates. Research suggests that hotels located in markets with a significant concentration of Airbnb listings may be particularly susceptible to adverse impacts on the average nightly rates (Zervas et al., 2017). Hotels have responded by adjusting their strategies and services, but the magnitude of their impact varies depending on factors related to location, market dynamics, and regulatory frameworks. Similar papers to the ones that are presented here regarding shared accommodation have been made in the hotel industry. An intriguing article, written by Abrate et al. (2011), applied a hedonic price technique through the analysis of quality signals to help explain price differentials among Turin's hotels in Italy. The method estimates the impact of product differentiation on price levels and has the potential to quantify the effect of each quality signal on price proposals. In another study made in Austin, where Airbnb supply is high, the causal impact on hotel revenue is in the 8%–10% range (Zervas et al., 2017). This goes to show that the application of hedonic price models is not something new in the study of problems related to pricing strategies in this type of industry.

2 Literature Review

The main purpose of the study is to develop a model capable of recommending the best price for a specific listing and reveal factors that strongly influence hosts rental strategies. Hedonic models have been widely used in real estate, tourism, and hotels (Y. Chen & Xie, 2017; Gibbs et al., 2018, Hung et al., 2010). An Airbnb accommodation listing, according to hedonic pricing theory, is therefore a bundle of elements that influence the quality of the overall product and provide consumers with value and satisfaction. Accordingly, a listing's price can be linked to the presence or absence of specific items (Gibbs et al., 2018). Until 2010, there was relatively few research on the use of price prediction models in the short-term rental property market. However, at that time, pricing was already widely acknowledged to be one of the most critical factors determining the long-term success of the accommodation industry (Hung et al., 2010). Furthermore, it was not clear how Airbnb hosts set their prices and how perceptions of consumers' willingness to pay for specific accommodation attributes potentially affect their pricing decisions.

2.1 Influencing price factors for the price accommodation

Several studies have been conducted to identify the most important features of Airbnb listings that influence guest satisfaction and are deterministic in the booking decision. These studies have used various methods such as surveys analyses, text mining techniques, and machine learning models to gather information about the data. The study introduced by Xie & Mao (2017) is one of the first to approach the effects of host quality on listing performance on the Airbnb sharing economy platform. The results show that having more reservations in the subsequent month depends in part on becoming a Superhost, increasing the host response rate, and having more experience as an Airbnb host. However, the assumption that location characteristics can have a high positive impact on the reservation does not significantly influence the reservation potential of a host's listing set. A similar study focused mainly on the "gamification design" developed by Airbnb that awards a "Superhost badge" to hosts who receive good reviews and observes how this can impact an accommodation's review volume and ratings (Liang et al., 2017). This study essentially explores the review volume, the hypothesis that an accommodation with the Superhost badge is more likely to receive reviews, but still is observed that price is negatively associated with review volume. In addition, due to the causality effect between the Superhost badge and the price, it is shown that guests are most likely to spend more for accommodations with the Superhost badge.

According to several studies (Jiang et al., 2022; Saló et al., 2014) the number of bedrooms was the main contributor to listing prices. The listing prices in Shanghai's central urban area were most sensitive to changes in the number of bedrooms. According to Saló et al. in 2014, the rental price of a holiday apartment was 10.9% less than that of a terraced house, a detached house was 13.8% higher than a terraced house, and an extra room generated an extra 13.8% in price. Similarly, it was identified that an additional room in an apartment increases the price by 20.6% (Wang & Nicolau, 2017). The minimum length of stay appeared to have a negligible effect on property prices according to Jiang et al. (2022), although excessively firm restrictions could diminish guests' willingness to stay. As stated, this factor could influence the traveler's decision to stay, since a lot of them tend to opt for a very short-term stay (1 or 2 days) in a specific area. This happens so that hosts can minimize operational costs, making it financially advantageous for hosts. Moreover, a study by Guo et al. (2020) on pricing strategies, highlights how hosts use minimum stay requirements to maximize profits. Further research into this topic, a more specific paper was conducted on several successful listings in Thessaloniki, Greece where they were analyzed using data mining. The study used customer reviews as a proxy for room reservations and introduced three dependent variables (occupancy, bookings, and revenue) to develop models that predict listings' performance and reveal factors that strongly influence customers' purchase decisions. The Random Forest method managed to outperform all the competitors being the most suitable classifier (Kirkos, 2022).

Finally, it is highly pertinent the explanation behind the search process of the platform to identify the different options that the consumers have when trying to find the right accommodation: the level one, consumers enter the location, dates, and number of guests. The presence of the number of guests field at this level makes the capacity of an Airbnb listing a primary attribute; in level two, consumers are presented with the ability to filter for room type (private, shared, or entire home) and price. A map on the side of the search interface also allows for filtering based on location. The third and final stage of consumer search via the Airbnb platform presents the ability to use filters about very specific traveler needs. Categories offered by this third level of search allow the consumer to filter for size (number of bedrooms, washrooms, and beds), booking options ("Instant Book" and host's "Superhost" status), neighborhoods, amenities (wireless Internet, kitchen, pool, etc.), property type (apartment, house, etc.), and host language.

2.2 Hosts Characteristics

The model results from a study conducted by Kirkos (2022), showed that the Superhost badge was the most deterministic factor for the increase in the occupancy of a property. In fact, the impact of the Superhost status found positive significance evidence on several analyzed cities

(Deboosere et al., 2019; Gibbs et al., 2018; Wang & Nicolau, 2017; Xie & Mao, 2017). In particular, there is a study that shows that a Superhost can earn approximately two more reservations for their listings than regular hosts, and a 1% increase in the response rate, produces five more reservations for a host's listings (Xie & Mao, 2017). In extensive literature research, only one study reported no significant coefficients to support the importance of this higher status.

According to the official Airbnb website, being a Superhost is one of the most prestigious statuses for a host to have. The company performs 4 yearly assessments to find out if the host has hit all the requirements within the past year for all their listings. If they meet all the criteria, then the "Superhost badge" is attributed to the account. To qualify, a listing owner with an account must meet the following criteria:

- Completed at least 3 reservations that total at least 100 nights.
- Maintained a 90% response rate or higher.
- Maintained a less than 1% cancellation rate.
- Maintained a 4.8 overall rating (A review counts towards Superhost status when either the guest and the Host have submitted a review, or the 14-day window for reviews is over, whichever comes first).

In many of the studies analyzed, the management of many listings appears to be a disadvantage for occupancy mainly due to the decrease in response rate. The host-related quality information cues, such as being a Superhost, having long operating experience, and maintaining a high response rate, appear to have significant effects on trust from Airbnb travelers (Xie & Mao, 2017; Gyódi & Nawaro, 2021). Both studies emphasize the relevance of ensuring that there must be a balance between quantity and quality as the host manages more listings.

2.3 Spatial Heterogeneity

Traditionally, tourists are heavily concentrated in areas where hotels are located (Shabrina & Morphet, 2022). This could be in central areas (monocentric) or agglomerated around certain geographical points such as tourist destinations, airports, hospitals, etc. Tussyadiah & Zach (2017), suggested that proximity to points of interest and the characteristics of the neighborhood are two of the most important contributing factors to a successful Airbnb location. This is supported by the findings in various papers (Gyódi & Nawaro, 2021; Jiang et al., 2022; Volgger et al., 2018) stating that guests prefer staying in areas with proximal locations to relevant touristic attractions. The keyword "accessibility" appears as the main variable when discussing location factors that affect price, where even small differences in the distance can lead to exponential

variances. This is even more strengthened by a study based on the sentiment analysis of Airbnb reviews, which found that one of the key attributes of an Airbnb experience is location, followed by amenities and hosts attributes (Cheng & Jin, 2019).

Among the studies, (Gutiérrez et al., 2017; Gyódi & Nawaro, 2021; Wegmann & Jiao, 2017) use the distance to each city center of the selected cities that were analyzed as the baseline measure. On a more general note, a set of various types of holiday-based accommodations were considered and another positive driver for price variation was the distance to the beach (Espinet et al., 2003; Saló et al., 2014). For this case, the city of Lisbon is known for the large coast that attracts many tourists, because of the popular viewpoints and the most prestigious historical monuments. This distance variable could be highly correlated to the distance to the main attractions of the capital. The paper from Shabrina & Morphet (2022) proposed a model to predict short-term rent house price patterns in the Greater London area based on 2018 annual visits to tourist attractions. Similarly, a study like the one of Gyódi & Nawaro (2021), proposes the use of *TripAdvisor* data, an online travel information platform, to add relevant dynamic variables that can explain the variability in price due to neighborhoods' attractiveness.

In terms of extensive model results comparison, an interesting study where three distinct models were employed (Linear Regression, Geographic Weighted Regression, and Random Forest) to examine influencing factors in 8,012 active Shanghai Airbnb listings through the incorporation of geographic variation in price modeling, found that individual heterogeneity is highly prominent in the shared accommodation industry. Furthermore, traditional spatial analysis methods like geographically weighted regression (GWR), which are frequently used in the traditional lodging industry, show poor performance (H. Zhang et al., 2011). The initial results reveal that traditional models face many challenges in solving such complex problems, so they proceeded to adopt a machine learning algorithm, the Random Forest. The study concluded that spatial heterogeneity plays a vital role in explaining the pricing patterns of Airbnb rentals within the city of Shanghai.

3 Methodology

The proposed approach for the research academic project follows along the line of the development of a machine learning product solution. There are many methodologies used in a ML project. In this case, since the focus is on an academic context, there is a need to do an extensive subject exploration, followed by data gathering and consecutive analyses, and finally evaluate how the different models behave in the final dataset. The methodology needs to provide a comprehensive framework for developing a robust price prediction model for short-term rentals. Rosen (1974), introduced the concept of hedonic price modeling and outlined the methodology for estimating implicit prices based on a set of relevant features. The sequence of methods that are going to be applied can be summed up in the chart process:

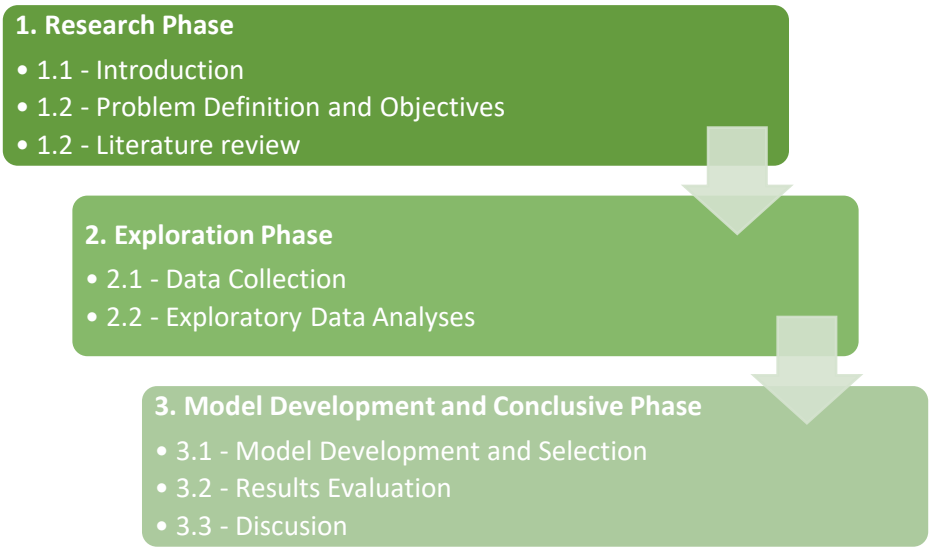

Figure 3.1 - Methodology Diagram

The Research Phase serves as the foundation for the subsequent stages, ensuring a clear problem definition, research objectives, and a comprehensive literature review of the existing papers in the field of machine learning and the subject of short-term renting. The first step involves defining the problem, specifically addressing the challenges and complexities associated with the industry. This stage aims to establish a well-defined background, considering several factors that need to be addressed. With a clear overview, research objectives are set to guide the study. The main objectives revolve around an elaborated literature review, detailed feature analysis, external factors investigation, and a theoretical background on Machine Learning Models. A comprehensive literature review was made considering the existing academic research papers

related to sharing accommodation, as this review is essential for identifying gaps about the subjects mentioned and are synthesed into a cohesive summary of existing knowledge.

3.1 Exploration Phase

The Exploration Phase encompasses the data collection stage where the main source of information about the listings was retrieved from a web scraping solution. The term "Web scraping" refers to the process of extracting data from various sources on the internet and databases (Lotfi et al., 2021). The literature review chapter enabled to choose a pertinent set of data sources to serve as deterministic input for the price prediction model. An essential step in the data preparation process is adhering to established academic norms, which include rigorous data operations, such as data cleaning, transformation, and reduction (removing inconsistencies, outliers' detection, and handling missing values) to ensure the quality of the final dataset. The consultation of academic references serves as a guide to approach feature selection, focusing on established techniques. This stage consists of selecting the most relevant variables for price prediction and transforming categorical data using academically recommended methods, such as one-hot encoding. Exploratory Data Analysis (EDA) is a critical step to gain insights into the dataset using statistical analysis, hypothesis tests, and visualizations to understand data distribution and relationships. This is the most demanding stage since there is a need to understand and gather useful insights from the previously gathered data.

3.1.1 Fundamental Data

To conduct this research, data was collected from a web scraping platform, *insideAirbnb*, which "provides data about Airbnb's impact on residential communities" as the official website mentions (Inside Airbnb: Home, 2022). In this case, the focus is on the district of Lisbon, Portugal. The overall dataset retrieved had about 19,690 listings and 75 features at the time of the scraped data of 17/12/2022. Initially, it was filtered to consider the city of Lisbon and "Entire homes/apartments" which resulted in a total of 24 Parishes and 10,666 individual properties. Airbnb has a lot of different types of accommodations, such as campers, boats, tents, and even treehouses, but these types were filtered out due to their irrelevance for city tourism.

These listings dataset is thoughtfully structured, with each entry being uniquely identified by its property ID, serving as the index. The dataset was divided into 4 different sub-datasets for specific exploratory data analysis to cover each aspect individually: house features, host attributes, reviews scores, and points of relevance. The house dataset has all the important

features related to the listing, such as the number of people that accommodates, the number of bathrooms, the number of beds, minimum and maximum nights allowed, and the occupancy rate, ranging from 30 to 365 days. The host attributes, being the positive correlated *host_response_rate* and *host_acceptance_rate*, the Boolean variable superhost status, the number of listings that each host has, and the days as a host, calculated from the date since the person created the listing, were the variables retrieved from the original dataset. The reviews dataset has the total number of reviews and the rating scores related to 5 aspects: rating, check-in, communication, and location. The last dataset has the neighborhood distinctions and the distances calculated.

3.1.2 Neighborhood

In this specific use case, a shapefile was gathered that provides the boundary coordinates for the 24 regions (parishes) of Lisbon. This shapefile serves as a dataset to manipulate and visualize the areas (Limite de Concelho - Datasets - Portal Dados Abertos, 2023) using the GeoPandas[1] Python library. The capital of Portugal, with a resident population of approximately 511,667 people, is as mentioned previously the focus of the study. To gain a comprehensive understanding of the short-term rental market in the city, there is a need to employ geodata visualizations. The horizontal bar plot displayed in Figure 3.2 illustrates the distribution of short-term rental listings within each parish.

The first two parishes, Santa Maria Maior and Misericórdia stand out with a notable disparity when compared to the third parish, Arroios, with a considerably higher number of listings, approximately 2.41 and 1.71 times higher than Arroios, respectively. This discrepancy can be attributed to the popularity among tourists, because of their proximity to the city center and the presence of numerous historical landmarks, such as Terreiro do Paço, the Alfama District, São Jorge Castle, and many more. This analysis provides valuable insights into the spatial distribution of short-term rental listings and highlights the pronounced demand for accommodations in these historically rich and centrally located parishes.

[1] Extension of Pandas, a popular python package to work with geospatial data.

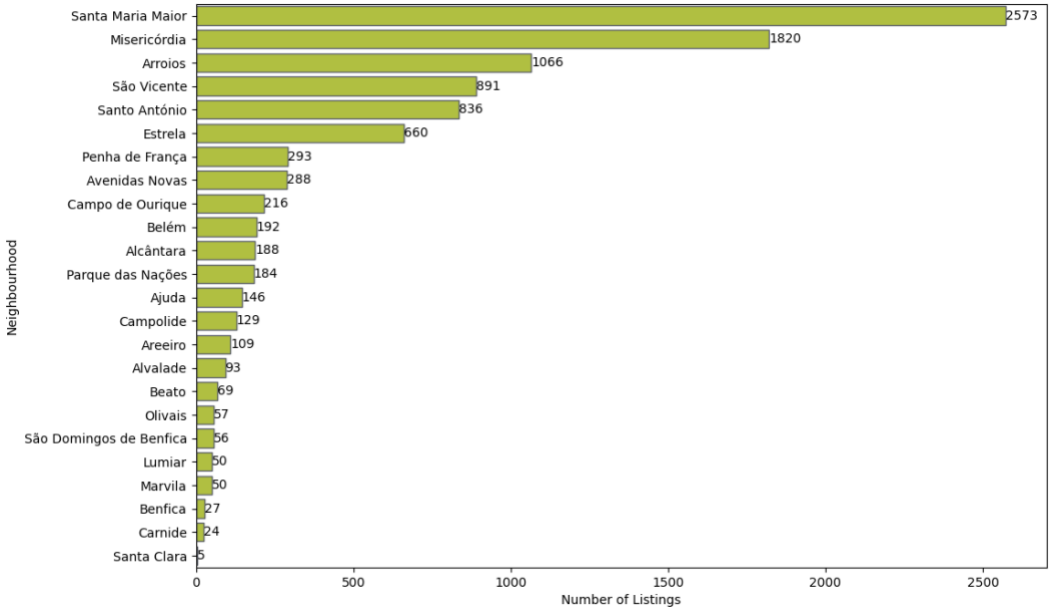

Figure 3.2 - Number of Listings per Neighborhood

The objective of calculating the average price for each parish was to gain insights into their pricing differences, resulting in a range from a minimum of 88.4 € to a maximum of 178.2 €. However, these values were not considered, because some parishes, such as Santa Clara with only 5 listings and Lumiar with 50 listings, lack sufficient data to accurately represent the true mean price. Santa Clara displayed an average of 178.2 €, while Lumiar ranked fourth with 124.0 €. To ensure more precise and reliable numerical representations of the various areas in Lisbon, it was opted to utilize median house prices.

3.1.2.1 Median House Price Values (m^2)

As stated in a few studies (Barron et al., 2021; Sheppard & Udell, 2016; Xu et al., 2020), there are significant differences in prices for each region of a city. In Portugal, most of the municipalities are subdivided into civil parishes since the creation of a democratic local administration in 1976. The Portuguese parishes are ruled by a system composed of an executive body (the Junta de Freguesia) and a deliberative body (the Assembleia de Freguesia). These entities are elected, and the policies applied to the community can vary from parish to parish. The parishes are then responsible for the preparation and implementation of certain decisions related to urban and rural planning, water supply, education, environment, well-being, etc.

To provide beneficial findings for this study, the median price per squared meters, m^2 of each parish was added to the model to have a variable to differentiate the regions in the city of Lisbon. The data was retrieved from INE – Instituto Superior de Estatística, corresponding to the median values per square meter of dwelling sales from the 1st Quarter of 2022. The Figure 3.3 presented illustrates the variations in property median values across the 24 parishes within the city of Lisbon. This range extends from a minimum of 2,556 € in Santa Clara to a maximum of 5,753 € in Santo António, situated in the heart of the city. To enhance visual clarity, five distinct intervals were dynamically created, each represented by a unique color gradient:

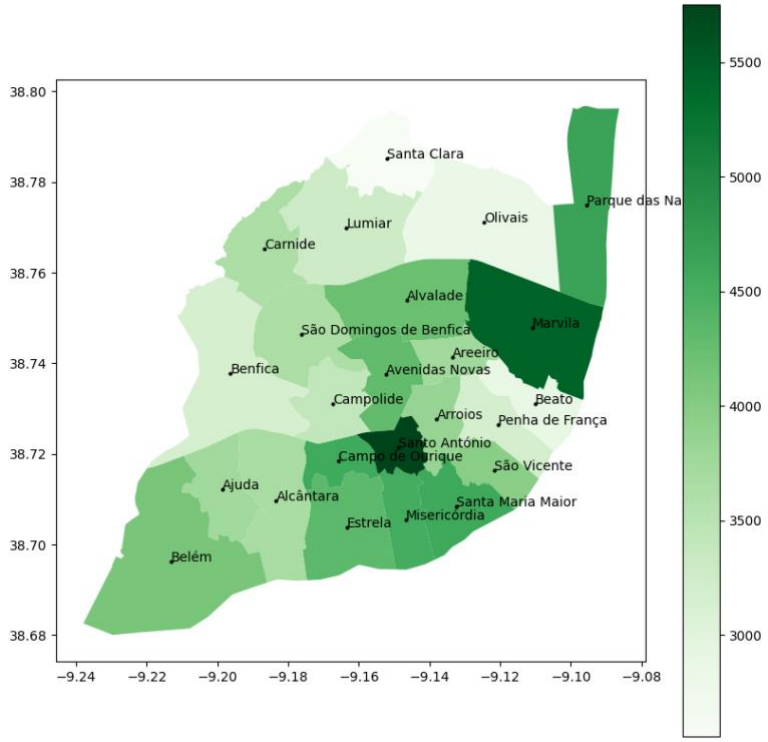

Figure 3.4 - Polygons Visualization of the Median Values per Parish

This distinction helps identify the different geographical areas based on examined values. The darker shades of green predominantly correspond to locations along the Portuguese city coast, characterized by median values exceedingly approximately 4,000 €. Notably, the city center, Santo António, and Marvila exhibit the deepest hues, indicating the highest property values. This visual analysis suggests that incorporating variables such as proximity to the city center and specific border locations may have a favorable impact on property rent prices. These insights accentuate the potential value of including such features in the dataset and help to understand the spatial distribution of the short-term rental listings.

3.1.3 Location factors

The importance of geolocation in the context of short-term rentals cannot be overstated. It starts with the assumption that the distances between listings and tourist points of interest can explain the fluctuation in prices in similar properties. There are numerous papers that explore the inherent diversity in pricing across different geographic locations within each city while measuring their impact on the short-term rent landscape (Gibbs et al., 2018; Gyódi & Nawaro, 2021; Xu et al., 2020). The majority states the hypothesis that listing prices are spatially dependent on points of interest, such as distance to the city center, transportation facilities, tourist attractions, and restaurants, which in consequence can help to understand the price fluctuation between different listings.

The GeoPandas library (GeoPandas 0.14.0 — GeoPandas Documentation, 2023) was used to calculate the distances between house locations and relevant points such as the city center, the closest subway station, and the city coast. Once the latitude and longitude coordinates were added, the data was then converted into a *GeoDataFrame* that shares a common Coordinate Reference System (CRS) for precise distance calculations. The distances were calculated using the GeoPandas *.distance()* method, which measures the projected Euclidean distance between each point and the target point. The listings population is represented in the Figure 3.4 where the distribution can be visualized on the map using the *.explore()* function from the same Python library.

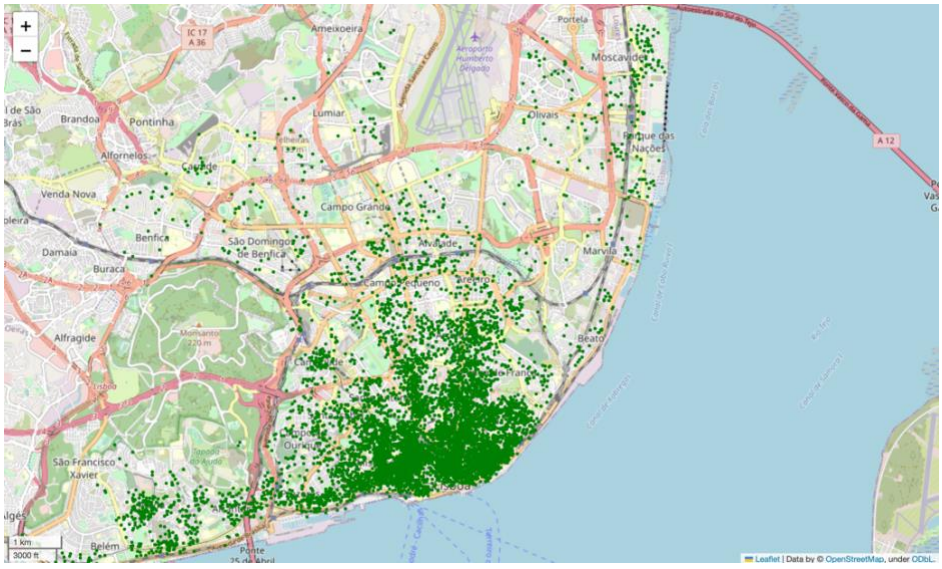

Figure 3.5 - GeoPandas Listings Map Visualization

3.1.3.1 Distance to Points: City Center and Closest Subway Station

Built between 1917 and 1934, the Marquês de Pombal monument is considered by most, the earth of Lisbon. Situated near Avenida da Liberdade, a place where several companies chose to place their headquarters, full of luxury stores, and near the famous Parque Eduardo VII, which is the city's largest greenspace. For this study, the statue's exact location was considered as the city center to be used to calculate the distance to the house listings.

The set GPS Coordinates regarding subway stations were obtained through the Lisbon government in *geojson* format (Estações de Metro - Dados.Gov.Pt - Portal de Dados Abertos Da Administração Pública, 2023). It has 58 stations in total divided into four lines: red, blue, green, and yellow with an extension of 46.5 km of transportation rails network.

3.1.3.2 Distances to Polygons: Lisbon Coast

The Lisbon Municipality provides open-source data to explore the cartography of the city through its geodata website. One of the downloaded datasets, maps the 1,086 blocks that the city has (Quarteirões | Quarteirões | Câmara Municipal Lisboa - Geodados_novo, 2023). The riverfront blocks polygons along the 20 km Lisbon coast was used to determine the distance to each house location. The assumption is that the price can increase not only because of the closeness to the water (Cohen et al., 2015), but also due to the existence of a lot of touristic sites, such as Terreiro do Paço, Torre de Belém, Oceanário, and many more.

3.1.4 Price Distribution

In the exploratory analysis of the price variable, the initial step was to visually inspect the distribution to understand its shape and characteristics. Upon visual examination, it became evident that the price distribution exhibited a right-skewed pattern. It was apparent in the histogram of the Figure 3.6 where most data points were located towards the lower end of the price scale, and a long tail extended to the right, indicating higher prices. In practical terms, there is a significant skewness of 3.89 according to the Fisher-Pearson coefficient of skewness, which means fewer high-priced properties. This imbalance can affect the model's predictive accuracy, as it may perform well in predicting lower prices but struggle with high-priced outliers.

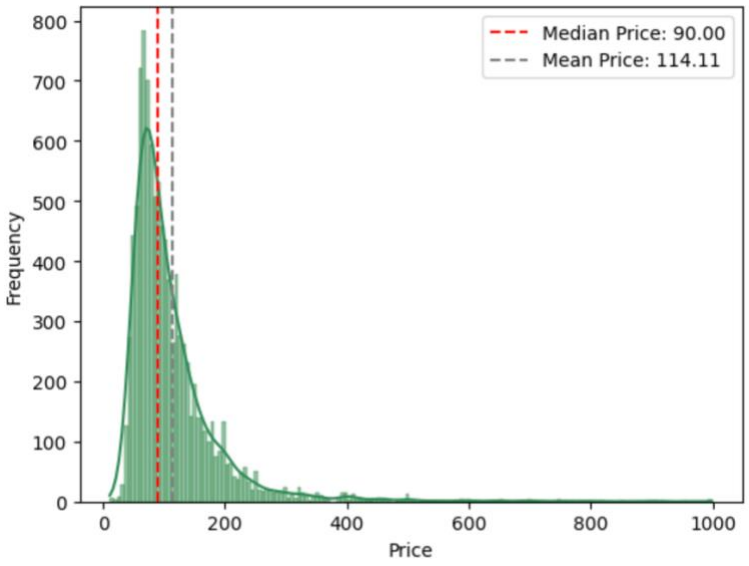

Figure 3.7 - Price Distribution: Mean and Median Comparison

Furthermore, a calculation of the central tendency metrics, such as the mean, median, and mode, proved the previously defined price distribution, since the mean of 114.11 was notably higher than the median of 90, and the mode represented the most frequently occurring value of 80. Overall, the difference between the mean and median highlighted the existence of a right-skewed distribution and the potential influence of higher-priced outliers, while the mode emphasized the prevalence of lower prices.

3.2 Model Development Phase and Conclusive Phase

In the Model Development and Conclusive Phase, the model selection is based on a thorough review of academic papers and machine learning literature, where the appropriate machine learning algorithms for price prediction is chosen. Data is then divided into training, validation, and test sets. Hyperparameter tuning and optimization is meticulously performed to ensure optimal model performance. The extensive analyses and comparison between different ensemble model results helps to determine the right set of features according to their feature importance graphs. The performance of the developed price prediction model is then assessed using the pre-determined evaluation metrics. Model performance is compared, and the significance of the findings is discussed.

In the end, the research findings are synthesized and discussed in the context of the problem definition, research objectives, and the existing literature. Insights from the model's interpretability techniques are used to explain the factors influencing short-term rental prices.

3.2.1 Bias and Variance

In supervised learning, the goal is to create a model that accurately predicts outcomes based on a set of input features. Mentioning the theory behind this trade-off is essential to understand the implications of choosing and developing a successful machine learning model. The goal is to ensure that the price prediction model is fair, transparent, and avoids potential biases.

Bias refers to the error introduced by overly simplistic assumptions in the model. It is said that a highly biased model tends to oversimplify the underlying data distribution. Such models often fail to capture the complexity of real-world relationships between features and outcomes, which in this case, the high bias can result in a model that consistently underestimates or overestimates prices, leading to poor predictions. So, if it tends to have difficult to fit the model, it means that there is a need to choose a more complex model for the data. Variance, on the other hand, represents the error of the model's sensitivity to fluctuations in the training data. Models with high variance are overly complex and can fit the training data perfectly, but struggle to generalize to new, unseen data. In price regression, high variance can result in a model that fits the training data perfectly but generalizes poorly to unseen data leading to a phenomenon called *overfitting*. The concept of *overfitting* appears when the model fits the training set well, but not the testing set. Pedro Domingos (2012), mentions this common concept in the machine learning area and manifests the idea of this trade-off, saying that “contrary to intuition, a more powerful learner is not necessarily better than a less powerful one”.

The challenge lies in striking the right balance between bias and variance, a simple model, and a complex one. A model with high bias is unlikely to perform well on both the training and test datasets because it fails to capture essential patterns. Conversely, a high-variance model may perform exceedingly well on the training data but exhibit poor generalization to new data. As shown in the Figure 3.6, the optimal model has low bias and low variance:

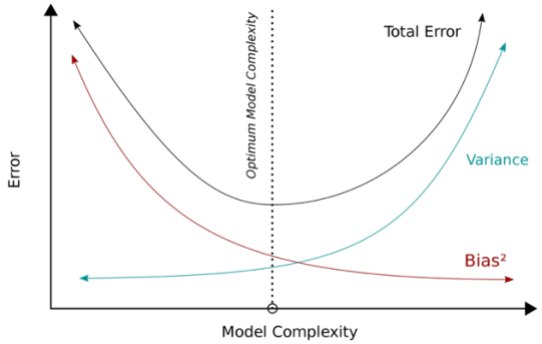

Figure 3.8 - Bias and Variance on Model Complexity (Lightner & Hagen, 2022)

One of the strategies to manage the Bias-Variance Trade-Off that researchers propose is the use of regularization techniques. The Lasso regression (Tibshirani, 1996) and Ridge regression (Hoerl & Kennard, 2000) introduce regularization terms that penalize complex models, which in consequence reduces bias. Cross-validation, more specifically the k-fold cross-validation aids in estimating a model's performance on unseen data to get valuable insights into the changes in bias and variance when training the model. Moreover, careful feature selection and a proper set of feature engineering can help mitigate bias. There are several metrics to evaluate the model fitting to the dataset, such as the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). Ensemble methods like Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) combine multiple models to reduce variance: they train multiple regressors and aggregate their predictions, leading to more stable and accurate results. These two models are the focus of this study, since according to the literature review, were the ones who got better overall performance against simpler models, such as the Linear Regression and the Geographic Weighted Regression.

3.2.2 Ensemble Methods: Random Forest and Gradient Boosted Trees

Ensemble Methods combine the outputs from individual trees (weak learners). Decision trees are recursive structures that split the feature space into regions, with each region associated with a predicted value (a leaf node). They differ in the way they are built and the way the results are combined. The paper from Hothorn et al. (2006) proposes a unified framework for recursive partitioning that embeds tree-structured regression models into a well-defined theory of conditional inference procedures. Stopping criteria based on multiple test procedures are implemented and it is shown that the predictive performance of the resulting trees is as good as the performance of established exhaustive search procedures. Several approaches are identified (Murthy et al., 1994; Quinlan, 1986).

The paper from Mohammed & Kora (2023) illustrated the recent trends in ensemble learning using quantitative analysis of several research papers. Moreover, it offers various factors that influence ensemble methods' success, including sampling the training data, training the baseline models, and the fusion techniques of the baseline models. At the core of any ensemble-based system are two techniques for training individual ensemble members: the sequential ensemble technique and the parallel ensemble technique (Quinlan, 1996). The paper reports the results of applying both techniques to a system that has decision trees as the baseline model and tests on a representative collection of datasets. It compares both ensemble techniques and emphasizes the substantially better results from the boosting method.

In the sequential ensemble technique, different learners learn sequentially because of data dependency. The introduced concept of "data dependency" refers to the extent to which the individual models or learners in the ensemble rely on the previous training data. So, the previously mislabeled data are tuned based on their weight to get the performance of the overall system improved data dependency. Thus, the errors made by the first model are sequentially corrected by the second model. Whereas in the parallel ensemble technique (Tang et al., 2020), base learners are generated simultaneously, as there is no data dependency. So, each data in the base learner is generated independently. This technique's basic advantage is exploiting the independence between base learners. Thus, the errors made by one model differ from those found in another independent model, allowing the ensemble model to calculate the average of the errors (Valle et al., 2010). The boosting implementation presents a few more challenges compared to the bagging method. The first one is related to the sequential training scalability, which normally is computationally costly and is more vulnerable to overfitting when increasing the number of iterations. Finally, it can be noted that boosting algorithms can be slower to train when compared to bagging since a large number of parameters can affect the behavior of the model.

All the papers in this chapter provide a comprehensive review of the various strategies for ensemble learning and it helps to enhance the relevance of this chapter. There is a need to know the mathematical explanation behind ensemble methods, such as the Random Forest and the Gradient Boosting Trees so that the model decisions can be understood and later improve predictions. The two algorithms require careful tuning of hyperparameters to prevent overfitting and improve the performance of the model. A variety of other ensemble methods do exist such as the AdaBoost algorithm implemented by Valle et al. (2010) for noise detection, Asbai & Amrouche (2017) for speech feature extraction and the SGB algorithm implemented by Shin (2019) for early prediction of safety accidents at construction sites. However, this study does not enter into much detail regarding these models.

3.2.3 Random Forest Regressor

Random Forest is an ensemble learning technique that employs a combination of decision trees. It uses a technique called Bootstrap Aggregating or Bagging (Breiman, 2001) to create multiple subsets (bootstrapped samples) of the original training data. Each subset is generated by randomly selecting data points with replacement. The goal of bagging is to create more diverse predictive models by adjusting a stochastic distribution of the training datasets, where small changes in the training data set lead to significant changes in the model predictions. In aggregation, the final result is achieved by majority voting of the model's predictions performed to determine the final prediction. For regression, the predictions from individual decision trees are combined through ensemble averaging. This typically involves computing the mean (average) of the predictions from all individual trees. The function of bagging is shown as follows:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} f_m(x)$$

Where $f_m(x)$ weak learners, $\frac{1}{M}$ generates bootstrapping sets.

The ensemble averaging reduces the variance of the model and leads to a more stable and accurate prediction. The drawback of bagging is that it is computationally expensive, can have high bias, and also can lead to a loss of interpretability of a model (Bühlmann & Yu, 2002). There are several challenges to implementing the bagging method that one faces when trying to define

the most optimal hyperparameters, such as the optimal number of base learners and subsets and the maximum number of bootstraps per subset. In addition, the output integration of the base classifiers from various voting methods can present diverse results (Mohammed & Kora, 2023).

3.2.4 Gradient Boost Regressor

The goal is to explain the intuition behind gradient boosting, provide visualizations for model construction, and explain one of the most popular machine learning algorithms. Jerome Friedman, in his paper "Greedy Function Approximation: A Gradient Boosting Machine" from 2001 introduced the idea of Gradient Boosting (Friedman, 2001). The method combines multiple simple models (called week learners), in this case, decision trees, into a single composite model to infer a prediction. The idea behind a gradient boost is to improve on a previous iteration of the model by correcting its predictions using another model based on the negative gradient of the loss (Johansson, 2023). Adding up a bunch of subfunctions to create a composite function that models some data points is then called additive modeling. Gradient boosting machines use additive modeling to gradually nudge an approximate model towards a better model, by adding those simple submodels. By combining the output, in theory the model not only becomes much stronger, but also in most cases does not overfit the original data.

In this context of regression, numerical predictions were made based on variables that influence price. Assuming each observation has a vector of features x, the goal is to find a scalar target value y for a bunch of (x_i, y_i) pairs. Given an example of a single feature vector x and scalar target value y for a single observation, the composite model that predicts $\hat{y}$ can be expressed as the addition of M weak models, $f_m(x)$:

$$\hat{y} = \sum_{m=1}^{M} f_m(x)$$

To build a boosted regression model, it starts with the base model $f_0(x)$ in which the predictions are based on the median value of every observation as the first result. Then, the model is "pushed" gradually towards the known target value y by adding more changes at each iteration $\Delta_m(x)$. The number of stages M highly affects the model accuracy: the more stages we have, the more accurate the model, but the more likely it can overfit:

$$\hat{y} = f_0(x) + \sum_{m=1}^{M} \Delta_m(x)$$

GBM implementation supports a hyperparameter called Learning Rate (η) to control the overall approach of $\hat{y}$ to y which in turn helps to reduce the probability of overfitting. In other words, it is a variable that controls the "step size" at which the model adapts to errors. According to the formula presented, smaller values make the learning more gradual, but in most cases make the learning process run for a longer time. The final prediction of the XGBoost model is the sum of predictions from all the trees in the ensemble.

$$F_m(x) = F_{m-1}(x) + \eta * f_m(x)$$

3.2.5 Improved Gradient Boosting: XGBoost

In the paper "XGBoost: A Scalable Tree Boosting System" authored by Tianqi Chen and published in March 2016, the XGBoost algorithm is introduced. XGBoost, short for Extreme Gradient Boosting, is a powerful ensemble learning technique, primarily designed for decision tree-based models. The algorithm introduces an innovative regularized objective function that combines a loss function with a regularization term. This combination is used to prevent overfitting and enhances the model's generalization capabilities.

One of the standout features of XGBoost is its scalability and computational efficiency. It incorporates various techniques to boost its performance, including parallel processing and tree pruning (pruning reduces the complexity of trees during model training, contributing to improved model generalization and mitigating overfitting). These optimizations make XGBoost significantly faster compared to traditional gradient boosting methods. The algorithm supports both L1 (Lasso) and L2 (Ridge) regularization terms, giving the possibility to control model complexity and sparsity. After each tree is built, XGBoost applies pruning to remove branches of the tree that do not lead to a significant reduction in the loss function. An example of the practical implementation was seen in a paper (Haumahu et al., 2021) that pretends to classify the hoaxes (fake news) based on a dataset consisting of several Indonesian news.

3.2.6 Model Evaluation Metrics

In both types of models, an objective function (also known as a loss function) is used to quantify the difference between the predicted and actual values. For regression tasks, the loss functions normally used are the mean squared error (MSE) and the mean absolute error (MAE). The loss across all N observations is just the average of all the individual's observation losses. The goal here is to find the pair $\big(y, F_M(X)\big)$ that minimizes L.

$$L\big(y, F_M(X)\big) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, F_M(x_i))$$

3.2.6.1 Root Mean Squared Error - RMSE

In $[0, \infty)$, the smaller the error, the better performance the final model has. The RMSE is calculated as the square root of the mean of the squared errors. The function measures the average magnitude of the residuals. This particular measure gives more weight to large deviations such as outliers, since large differences square become larger can give to the model.

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{|y_i - \widehat{y_i}|}{N}}$$

Where:

N is the number of observations,
y_i is the actual value for observation i,
$\widehat{y_i}$ is the predicted value for observation i.

3.2.6.2 Mean Absolute Error - MAE

In $[0, \infty)$, the same logic applies to the MAE metric, where a value closer to zero indicates a better overall fit to the data. It measures the absolute magnitude of the error. The formula is simpler, so the errors are easier to interpret, and contrary to the RMSE, it is less sensitive to outliers and all errors contribute proportionally to the overall metric.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Where:

N is the number of observations,
y_i is the actual value for observation i,
$\hat{y}_i$ is the predicted value for observation i.

4 Results

The dataset preparation for the modeling part consisted of a combination of multiple data sources so that a proper evaluation of the different results could be addressed. The original Airbnb dataset, which consists of house features, host attributes, and reviews scores, was merged with the external features, the calculated distances to the selected points of interest and neighborhood information.

According to the person correlation, the number of bathrooms and the number of people that the house accommodates are the most significant correlated variables with the price, each with a correlation coefficient of 0.63 and 0.57, respectively. It is known that one of the major factors that can explain the variance in price is the number of people a listing can accommodate. The separate analyses of the linear relationship between the two shows that for each additional person the listing can accommodate, the price increases on average by 20.26 euros. The value is not surprising for the number of guests as in practice the more people the house can have, the higher the price. However, for the number of bathrooms, this high correlation could be associated with the house dimensions. This theory is confirmed in a paper from Zietz et al. (2008), whose results describe the square footage, lot size, bathrooms, and floor type as having a greater impact as the selling price increases. The correlation between the review score rating and other review categories, including cleanliness, check-in, communication, and location, range from 0.55 to 0.84. This suggests a positive association, indicating that higher overall review ratings are likely to be followed by elevated ratings in the sub-reviews categories. Lastly, the distance to the city center presented a coefficient of - 0.0031, which in turn has few negative impacts on the pricing. In fact, there are other factors that influence the price, hence the importance of considering the combination of several variables.

4.1 Model Performance

According to the models and regression metrics phase to evaluate the model performance explored in the Methodology chapter, these were the results for the combination of the datasets:

	Model Metrics								
	Random Forest Regressor			Gradient Boosting Regressor			XGBoost		
Datasets	RMSE	MAE	R^2	RMSE	MAE	R2	RMSE	MAE	R^2
House Features + Hosts Attributes + Reviews Scores	57.42	32.18	0.56	56.33	31.50	0.58	54.00	29.24	0.61
House Features + Hosts Attributes + Reviews Scores + Neighborhood + Distances	57.37	31.48	0.57	54.83	31.06	0.60	52.77	28.06	0.63

Table 4.1 - Models Results

The reason behind the separation into two datasets was to analyze the effects when adding variables related to the house locations, such as the neighborhood and the distances previously calculated. The three models chosen present similar RMSE results with a prevalence of better overall value for the XGBoost model in both datasets. The second dataset consistently yields lower RMSE and MAE values compared to the first one which has only fundamental information about the listings. The higher R^2 values that represent the proportion of the variance for the dependent variable that is explained by the considered independent variables, denote a good fit with a value of 0.63. This result shows a strong coefficient that can help explain the dynamics influencing Airbnb listing prices. The features related to the distances to the city center, coast, and subway, as well as the median value per square meter, appear to have significantly less impact than expected since the metrics values only differ slightly from the first dataset.

4.2 Feature Importance analysis

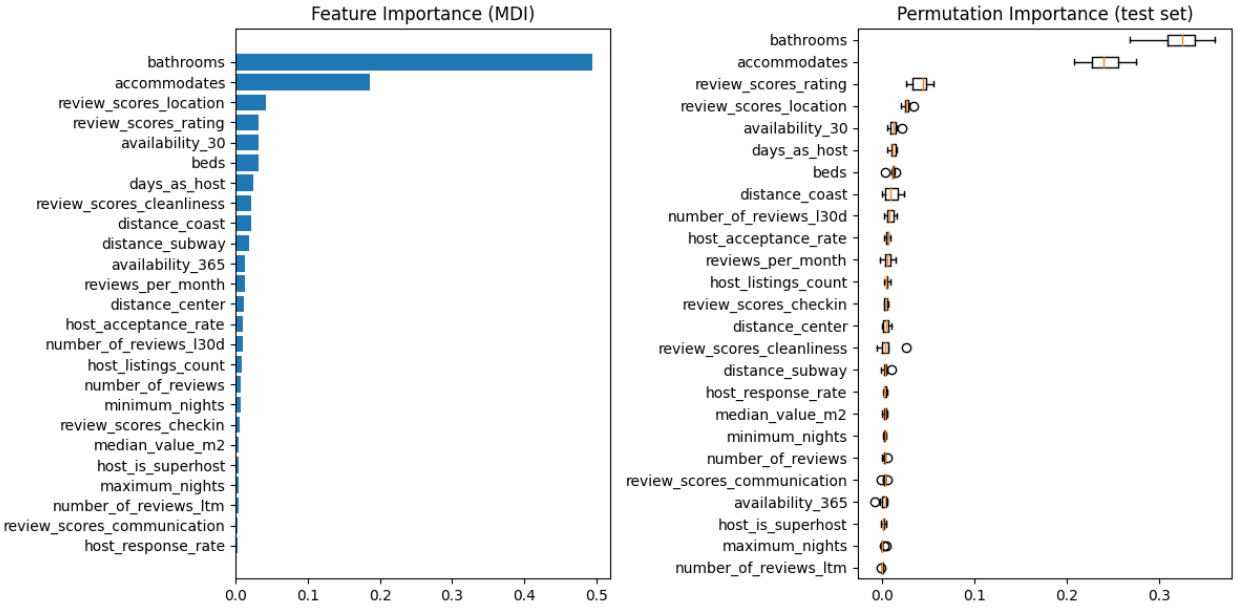

Figure 4.1 - Feature Importance and Permutation Score for XGBoost

The results of the feature importance (Mean Decrease in Impurity) graph that presents the total reduction of the criterion brought by each feature, show that the number of bathrooms and the number of people that the house can accommodate (or number of guests) emerge has the most influential features, contributing to explained variances of 0.49 and 0.19, respectively.

The findings from the permutation importance graph, as a technique to estimate the importance of each feature by measuring the change in the model's performance when a determined feature value is randomly shuffled, support the statement that the model relies heavily on the two main variables to make predictions. The reviews scores for location and overall rating score contribute with similar explained variances, 0.041 and 0.032, respectively. It is noteworthy that in the top 5 features, only the positions of the reviews score on rating and reviews score on location have interchanged compared to the rank in the feature importance graph.

5 Discussion

The data cleansing techniques applied tried to minimize the need for listings removal. Techniques were applied to remove potential outliers or misleading house prices that could come from web scraping errors or unrealistic price inputs given by the host. The cases of multicollinearity were treated by removing certain variables, such as the availability of 60 and 90 days and the number of bedrooms (this last one, also due to 5% missing values). The feature selection techniques considered were the Recursive Feature Selection (RFE) and the SelectKBest which helped to reduce dimensionality and remove irrelevant features that may negatively impact the predictions, such as the *host_identity_verified* and the *instant_bookable*.

Despite the results achieved in each model, a MAE value of 28.06 means that, on average the predictions deviate from the actual prices by approximately 28.06 euros. There are several factors that contributed to this level of error. The complexity of the short-term rental market, which is influenced by a multitude of variables, including location, property features, and market dynamics, could have been a determinant factor for the model's inability to capture all the important price influencers since key features could be missing from the dataset. Another factor is related to the existence of outliers due to errors on the web scrapped data that can highly contribute to the error increase and mislead the true nightly rate of the listing, which in consequence can lead to non-optimal split tree leaves. Furthermore, the model's predictive accuracy may have been affected by the quantity of the data used for training, since this study focuses solely on entire proprieties in the city of Lisbon, and in turn, this can cause the model to lack enough data to understand the underlying characteristics to clearly define a more accurate pricing.

5.1 Host main attribute: Superhost title

A statistical analysis was performed to test the significant influence of the variable *host_is_superhost* on the price. The first results show a T-statistic of 5.822, which suggests that there is a substantial difference in the mean price between superhosts and non-superhosts. Assuming a result higher than 2 is considered statistically significant, the value presented is relatively high, indicating a substantial difference in the mean price between the two groups. This is confirmed by the very low p-value of approximately 6.126e-09, which indicates that there is strong evidence to reject the null hypothesis.

The results are supported by the chi-squared test statistic result of approximately 731.67. This relatively high value suggests that there is a difference between the observed and expected frequencies, which indicates a significant relationship between the superhost variable and the price variable. In addition, the very low p-value (approximately 1.938e-17) confirms that the association between the variables is highly unlikely to have occurred by random chance. Taking into consideration both test results, the *host_is_superhost* variable is significantly associated with differences in price as similarly seen in the study of Liang in 2017 that states that "guests are willing to pay more for Superhost accommodations (Liang et al., 2017).

5.2 Distances and neighborhood impact

One of the main objectives is to explore the impact on decision trees-based ensemble models when adding features related to the location of points of interest and neighborhood characteristics in the hope that can provide more explainability to the predictive model. According to the previous figure about the feature importance and permutation score, the variables distance to the city center (*distance_center*), distance to the closest subway (*distance_subway*), and distance to the coast (*distance_coast*) did not significantly contribute to the predictive capacity of the model in the defined area as expected. However, the variable that corresponds to the values on location ratings (*review_scores_location*) appears on the graph as the third most important feature, which can indicate that there are external variables related to the listing location that can be significant, but this study was not able to identify them.

As a surprise, the median values per m^2 were not considered as valuable as expected for the feature importance and permutation importance graphs. This could mean that the values could have drastically changed during 2022 or the prices of the listings do not reflect the price per square meter of each parish, which in urban investment theory should be a characteristic to consider.

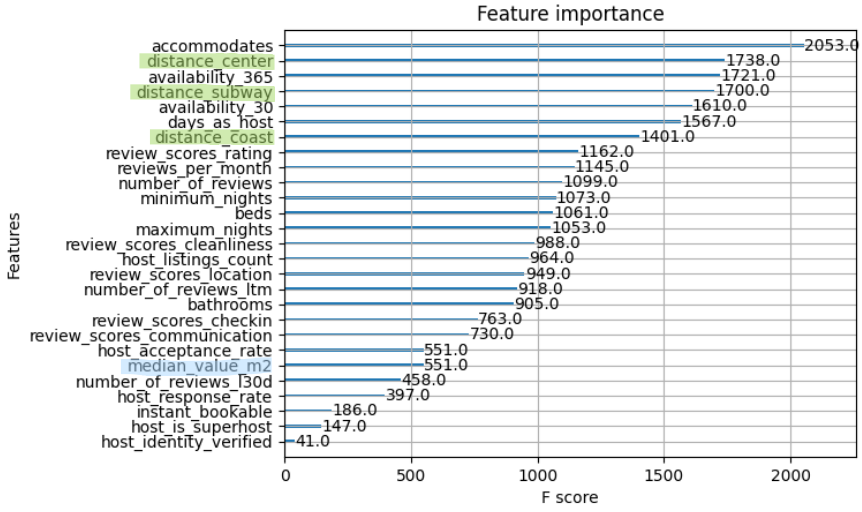

Figure 5.1 - Feature Importance based on weight score for XGBoost

Lastly, the feature importance graph presented above highlights the significant weight assigned to the distance-related features. Specifically, the three distances (in green) emerged as one of the most used features to split the data across the trees. This outcome is consistent with the theory behind this type of model stating that they are biased towards the high cardinality features, as seen in the relevant literature. Consequently, the values of these distance variables in the feature importance score are reflected in the decision-making processes of the ensemble.

5.3 In-depth analysis

A thorough analysis was performed to examine the specific instances where the model's predictions differ significantly from the actual prices to identify areas for model improvement, such as outlier handling, feature engineering, or even oversampling techniques. The results shows that 75% of the values corresponding to the difference between the real value and the prediction made with XGBoost are below 33.0 difference, which means that most of the errors are quite low. The following boxplot illustrates the distribution:

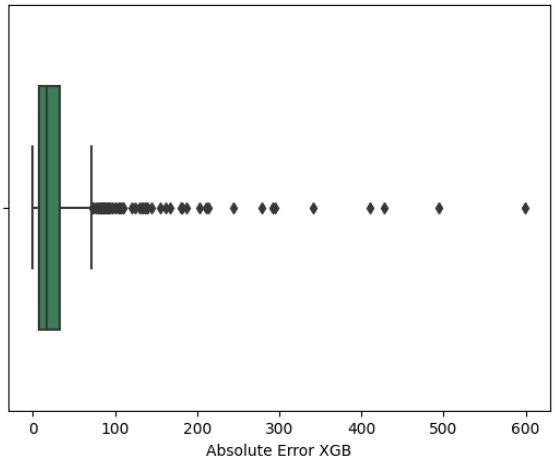

Figure 5.2 - Absolute Errors XGBoost

What impacts the most the RMSE are the potential outliers that arise from the houses that have the highest prices. This is supported by the evidence that if the absolute errors are sort descending and the focus is on the top 100 values, only 6 houses have their price value lower than 100 euros. According to these results, a few assumptions can be made:

1. The hosts are placing a price that is unrealistic considering the features related to accommodation and location.
2. The hosts are speculating a market price due to external seasonal factors or other external variables that are not addressed in this study.
3. The model is not able to capture the underlying features that differentiate a 500 euros house from a 50 euros house (assuming that the most important features are equal, such as the number of guests that the property accommodates).

Further investigation upon the first assumption, it is noted that from the same first 100 listings, only 5% of them have values higher than 10 guests. The results shown previously by the feature importance graph consider the number of accommodates to be one of the most relevant features when trying to predict the price. This means that when splitting the trees, the 95 lower values might be grouped into similar values in terms of this variable, which as consequence will increase the error when calculating the average value for the same group, leaving these houses with a high prediction result. Regarding location to the city center, the results are clustered around the interval range between 1,000 and 3,000 meters. There are only 3 values that can be considered outliers since they accommodate less than 10 people and have a distance to the center higher than 3,000 meters.

The second and third assumptions revolve around two main determinant factors in every machine learning dataset, the lack of features related to the influence of the price difference between similar listings and not enough data (in terms of the number of houses) that can help the model to understand the underline characteristics of the market pricing. One is related to the other in the sense that both are needed to properly explain the variance in the listing's prices. Similar conclusions were taken from a paper by Hong et al. (2020) that showed that there are significant market complexities that make the value determination process unable through simplified assumptions of the conventional hedonic pricing model. In this area of hedonic models, Chau & Chin (2002) stated that the models suffer from some level of biases due to missing variables.

6 Conclusion

This study is based on the context of the short-term rental market, a dynamic and rapidly evolving sector within the broader real estate industry. Notably, online platforms like Airbnb have played a transformative role, facilitating the connection between hosts and guests, streamlining bookings, and providing a platform for property owners to monetize their properties. The study included a thorough research phase, problem definition, objectives, and an extensive literature review, which provided the foundation for the subsequent analytical and conclusive phases.

Through the analytical phase, the research examined the most influential factors affecting short-term rental prices, uncovering the influence of each property feature, host characteristics, and external location factors. The reason behind the choice of this set of variables is supported by numerous studies in the field (Jiang et al., 2022; Tussyadiah & Zach, 2017), where they study in-depth the determinants of Airbnb listing prices and incorporate geographic variation in price modeling. The first example mentioned provided an analysis of the spatial dependence of price influencing factors and identified the premium capacity of variables within different price intervals and thresholds. The second one analyses online reviews of listings to extract salient attributes through extensive lexical analyses. Both authors mentioned the implemented attributes in this work, such as the location (proximity to point of interest and characteristics of the neighborhood), host (service and hospitality), and property (facilities and atmosphere).

The decision to develop a machine learning model employing ensemble methods, was supported by the premise that the combination of several individuals' models can lead to an improved prediction performance (Mohammed & Kora, 2023). The results of this thesis emphasize the complexity of short-term rental pricing and the intricate influence that different locations can have in this market. It is expected to contribute to property owners, hosts, and investors who are eager to understand the underline conclusions taken from this study. As in similar papers (Hong et al., 2020), this investigation highlighted the importance of feature analysis and geospatial insights in understanding the complex dynamics of the field. Above all, the goal was to enhance the potential of leveraging data-driven methodologies to improve decision-making in the business of short-term property rentals.

As a conclusion, the dynamics of this market are influenced by a variety of factors as explained, including property location, house features, host characteristics, and various location factors. Understanding and optimizing these variables are critical to succeed in this competitive market. The context of this study encompasses exploring, analyzing, and predicting the determinants of short-term rental prices, with the aim of providing valuable insights and tools to enhance pricing strategies. Additionally, it involves considering the broader economic and regulatory implications

of the short-term rental market, including its impact on traditional hospitality sectors and the Portuguese urban housing market.

6.1 Limitations

The study focuses mainly on developing a predictive model for house price prediction and the behavior of adding location attributes to tree-based models. However, it would be interesting to also evaluate the impact on different machine learning models, such as the SVR or a Bayesian method. The last type could provide valuable insights since allows for a probabilistic approach to modeling by adding prior events or information capable of explaining the dependent variable. In the paper from Robert Egbenta & Etuk (2019), macroeconomic metrics related to inflation in Nigeria were added to the house characteristics dataset which thereby enhanced the accuracy of the housing price estimation. The results provided in this paper are related to the static period, meaning that prices may vary significantly depending on the seasonality. In accordance with the paper, the consideration of seasonal features could have a positive impact on the model performance.

As in similar studies (Ghosh et al., 2023), the amenities available in each listing were discarded from this work. It was found that in total the original dataset presented 7,322 different features which means that by adding those to the dataset, would likely encounter problems regarding Curse of Dimensionality, meaning that as columns increase, the amount of data needed to support the model also increases exponentially. The solution then would be to apply dimensionality reduction techniques like Principal Component Analyses (PCA) or combine similar ones through the application of Natural Language Processing Models (NLP) to improve the model interpretability and provide a short list of the fundamental amenities to include in the dataset.

6.2 Future Work

Recent events have demonstrated the critical importance of flexibility in short-term rental pricing strategies. Hosts have learned that the most effective Airbnb pricing strategy is one that remains adaptable. The influence of seasonality on pricing during peak and slow seasons can help hosts identify the optimal pricing balance for occupancy and revenue. A model that would incorporate dynamic pricing, would not only consider market supply and demand, but also incorporate real-time data sources, such as local events, competitor pricing, and weather conditions. Moreover, real-time access to website views and clicks can offer significant potential for dynamic pricing strategies, since the integration of real-time analytics can be used to adjust price based on user

interactions with the listing. This approach would enable hosts to respond to immediate market changes and capture booking opportunities as they arise.

In terms of feature extraction, the application of text mining techniques and image recognition models can help extract determinant aspects like ambiance and unique property attributes. The use of Natural Language Processing Models for feature extraction from listing descriptions and reviews can significantly improve the understanding of property attributes, while Deep Learning Models (DL) for analyzing images can be determinant to automatically evaluate image quality based on criteria like resolution, lighting, and overall visual appeal. An interesting analysis would be to address the relationship between image quality and booking rates, offering insights into the impact of high-quality images on listing performance. Finally, applying customer segmentation techniques would enable hosts to create personalized offers, amenities, and experiences for various customer segments based on factors such as booking history, preferences, demographics, and behavior, ultimately enhancing guest satisfaction and loyalty.

References

Abrate, G., Capriello, A., & Fraquelli, G. (2011). When quality signals talk: Evidence from the Turin hotel industry. *Tourism Management*, *32*(4), 912–921. https://doi.org/10.1016/J.TOURMAN.2010.08.006

Asbai, N., & Amrouche, A. (2017). Boosting scores fusion approach using Front-End Diversity and adaboost Algorithm, for speaker verification. *Computers & Electrical Engineering*, *62*, 648–662. https://doi.org/10.1016/J.COMPELECENG.2017.03.022

Barron, K., Kung, E., & Proserpio, D. (2021). The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Science*, *40*(1). https://doi.org/10.1287/mksc.2020.1227

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, *30*(4), 927–961. https://doi.org/10.1214/aos/1031689014

Cekic, M., Korkmaz, K. N., Mukus, H., Hameed, A. A., Jamil, A., & Soleimani, F. (2022). Artificial Intelligence Approach for Modeling House Price Prediction. *2022 2nd International Conference on Computing and Machine Intelligence, ICMI 2022 - Proceedings*. https://doi.org/10.1109/ICMI55296.2022.9873784

Chau, K. W., & Chin, T. L. (2002). *A Critical Review of Literature on the Hedonic Price Model*. https://papers.ssrn.com/abstract=2073594

Chen, Y., & Xie, K. L. (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

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. *International Journal of Hospitality Management*, *76*, 58–70. https://doi.org/10.1016/J.IJHM.2018.04.004

Cohen, J. P., Cromley, R. G., & Banach, K. T. (2015). Are Homes Near Water Bodies and Wetlands Worth More or Less? An Analysis of Housing Prices in One Connecticut Town. *Growth and Change*, *46*(1), 114–132. https://doi.org/10.1111/GROW.12073

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

Ding, K., Niu, Y., & Choo, W. C. (2023). The evolution of Airbnb research: A systematic literature review using structural topic modeling. *Heliyon*, *9*(6), e17090. https://doi.org/10.1016/J.HELIYON.2023.E17090

Dogru, T., Hanks, L., Mody, M., Suess, C., & Sirakaya-Turk, E. (2020). The effects of Airbnb on hotel performance: Evidence from cities beyond the United States. *Tourism Management*, *79*. https://doi.org/10.1016/J.TOURMAN.2020.104090

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Espinet, J. M., Saez, M., Coenders, G., & Fluvia, M. (2003). Effect on Prices of the Attributes of Holiday Hotels: A Hedonic Prices Approach. *Http://Dx.Doi.Org/10.5367/000000003101298330*, *9*(2), 165–177. https://doi.org/10.5367/000000003101298330

Estações de Metro - dados.gov.pt - Portal de dados abertos da Administração Pública. (n.d.). Retrieved October 28, 2023, from https://dados.gov.pt/pt/datasets/estacoes-de-metro/

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Https://Doi.Org/10.1214/Aos/1013203451*, *29*(5), 1189–1232. https://doi.org/10.1214/AOS/1013203451

GeoPandas 0.14.0 — GeoPandas 0.14.0+0.g0eb2a5e.dirty documentation. (n.d.). Retrieved October 28, 2023, from https://geopandas.org/en/stable/

Ghosh, I., Jana, R. K., & Abedin, M. Z. (2023). An ensemble machine learning framework for Airbnb rental price modeling without using amenity-driven features. *International Journal of Contemporary Hospitality Management*, *35*(10), 3592–3611. https://doi.org/10.1108/IJCHM-05-2022-0562/FULL/PDF

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J., & Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. *Journal of Travel and Tourism Marketing*, *35*(1), 46–56. https://doi.org/10.1080/10548408.2017.1308292

Guo, Y., Chen, J., Zhou, Y., Geng, J., Guo, Y., Chen, J., Zhou, Y., & Geng, J. (2020). Sharing Economy Platforms' Pricing Strategies and Decision Preferences: The Example of DiDi. *Open Journal of Business and Management*, *8*(4), 1641–1656. https://doi.org/10.4236/OJBM.2020.84104

Gutiérrez, J., García-Palomares, J. C., Romanillos, G., & Salas-Olmedo, M. H. (2017). The eruption of Airbnb in tourist cities: Comparing spatial patterns of hotels and peer-to-peer accommodation in Barcelona. *Tourism Management*, *62*, 278–291. https://doi.org/10.1016/J.TOURMAN.2017.05.003

Guttentag, D. (2019). Progress on Airbnb: a literature review. *Journal of Hospitality and Tourism Technology*, *10*(3), 233–263. https://doi.org/10.1108/JHTT-08-2018-0075

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

Hati, S. R. H., Balqiah, T. E., Hananto, A., & Yuliati, E. (2021). A decade of systematic literature review on Airbnb: the sharing economy from a multiple stakeholder perspective. *Heliyon*, *7*(10), e08222. https://doi.org/10.1016/J.HELIYON.2021.E08222

Haumahu, J. P., Permana, S. D. H., & Yaddarabullah, Y. (2021). Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost). *IOP Conference Series: Materials Science and Engineering*, *1098*(5), 052081. https://doi.org/10.1088/1757-899X/1098/5/052081

Hoerl, A. E., & Kennard, R. W. (2000). *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. *42*(1), 80–86.

Hong, J., Choi, H., & Kim, W. S. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in south korea. *International Journal of Strategic Property Management*, *24*(3), 140–152. https://doi.org/10.3846/IJSPM.2020.11544

Horn, K., & Merante, M. (2017). Is home sharing driving up rents? Evidence from Airbnb in Boston. *Journal of Housing Economics*, *38*, 14–24. https://doi.org/10.1016/J.JHE.2017.08.002

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hung, W. T., Shang, J. K., & Wang, F. C. (2010). Pricing determinants in the hotel industry: Quantile regression analysis. *International Journal of Hospitality Management*, *29*(3), 378–384. https://doi.org/10.1016/J.IJHM.2009.09.001

Inside Airbnb: Home. (n.d.). Retrieved February 7, 2024, from http://insideairbnb.com/

Jiang, Y., Zhang, H., Cao, X., Wei, G., & Yang, Y. (2022). How to better incorporate geographic variation in Airbnb price modeling? *Tourism Economics*. https://doi.org/10.1177/13548166221097585

Jiang, Y., Zhang, H., Cao, X., Wei, G., & Yang, Y. (2023). How to better incorporate geographic variation in Airbnb price modeling? *Tourism Economics*, *29*(5), 1181–1203. https://doi.org/10.1177/13548166221097585/ASSET/IMAGES/LARGE/10.1177_13548166221097585-FIG5.JPEG

Johansson, R. (n.d.). *An intuitive explanation of gradient boosting*. Retrieved September 18, 2023, from http://www.cse.chalmers.se/

Kirkos, E. (2022). Airbnb listings' performance: determinants and predictive models. *European Journal of Tourism Research*, *30*. https://doi.org/10.54055/ejtr.v30i.2142

Liang, S., Schuckert, M., Law, R., & Chen, C. C. (2017a). Be a "Superhost": The importance of badge systems for peer-to-peer rental accommodations. *Tourism Management*, *60*, 454–465. https://doi.org/10.1016/J.TOURMAN.2017.01.007

Liang, S., Schuckert, M., Law, R., & Chen, C. C. (2017b). Be a "Superhost": The importance of badge systems for peer-to-peer rental accommodations. *Tourism Management*, *60*, 454–465. https://doi.org/10.1016/J.TOURMAN.2017.01.007

Limite de Concelho - Datasets - Portal Dados Abertos. (n.d.). Retrieved October 28, 2023, from https://dados.cm-lisboa.pt/en/dataset/limite-de-concelho

Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2021). Web Scraping Techniques and Applications: A Literature Review. *SCRS CONFERENCE PROCEEDINGS ON INTELLIGENT SYSTEMS*, 381–394. https://doi.org/10.52458/978-93-91842-08-6-38

Mohammed, A., & Kora, R. (2023). A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 757–774. https://doi.org/10.1016/j.jksuci.2023.01.014

Murthy, S. K., Kasif, S., & Salzberg, S. (1994). A System for Induction of Oblique Decision Trees. *Journal of Artificial Intelligence Research*, *2*, 1–32. https://doi.org/10.1613/JAIR.63

Nieuwland, S., & van Melik, R. (2020). Regulating Airbnb: how cities deal with perceived negative externalities of short-term rentals. *Current Issues in Tourism*, *23*(7), 811–825. https://doi.org/10.1080/13683500.2018.1504899

Quarteirões | Quarteirões | Câmara Municipal Lisboa - Geodados_novo. (n.d.). Retrieved October 28, 2023, from https://geodados-cml.hub.arcgis.com/datasets/CML::quarteir%C3%B5es/explore

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1986 1:1*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1996). Bagging, Boosting, and C4.5. *AAAI/IAAI, Vol. 1*.

Robert Egbenta, I., & Etuk, S. U. (2019). Prediction of House Prices Using Hedonic and Bayesian Models: An Application to Uyo Housing Market, Nigeria. *Middle-East Journal of Scientific Research*, *27*(8), 615–625. https://doi.org/10.5829/idosi.mejsr.2019.615.625

Rosen, S. (n.d.). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Source: Journal of Political Economy*, *82*(1), 34–55.

Saló, A., Garriga, A., Rigall-I-Torrent, R., Vila, M., & Fluvià, M. (2014). Do implicit prices for hotels and second homes show differences in tourists' valuation for public attributes for each type of accommodation facility? *International Journal of Hospitality Management*, *36*, 120–129. https://doi.org/10.1016/J.IJHM.2013.08.011

Shabrina, Z., & Morphet, R. (2022). Understanding patterns and competitions of short- and long-term rental markets: Evidence from London. *Transactions in GIS*, *26*(7), 2914–2931. https://doi.org/10.1111/TGIS.12989

Sheppard, S., & Udell, A. (2016). *Do Airbnb properties affect house prices?*

Shin, Y. (2019). Application of Stochastic Gradient Boosting Approach to Early Prediction of Safety Accidents at Construction Site. *Advances in Civil Engineering*, *2019*. https://doi.org/10.1155/2019/1574297

Statistics Portugal - Web Portal. (2022). Retrieved February 11, 2024, from https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=540831279&DESTAQUESmodo=2

Tang, J., Su, Q., Su, B., Fong, S., Cao, W., & Gong, X. (2020). Parallel ensemble learning of convolutional neural networks and local binary patterns for face recognition. *Computer Methods and Programs in Biomedicine*, *197*, 105622. https://doi.org/10.1016/j.cmpb.2020.105622

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tussyadiah, I. P., & Zach, F. (2017). Identifying salient attributes of peer-to-peer accommodation experience. *Journal of Travel & Tourism Marketing*, *34*(5), 636–652. https://doi.org/10.1080/10548408.2016.1209153

Utimate Short-Term Vacation Rental Glossary | AirDNA. (n.d.). Retrieved February 11, 2024, from https://www.airdna.co/short-term-rental-glossary

Valle, C., Saravia, F., Allende, H., Monge, R., & Fernández, C. (2010). Parallel approach for ensemble learning with locally coupled neural networks. *Neural Processing Letters*, *32*(3), 277–291. https://doi.org/10.1007/S11063-010-9157-6/METRICS

Volgger, M., Pforr, C., Stawinoga, A. E., Taplin, R., & Matthews, S. (2018). Who adopts the Airbnb innovation? An analysis of international visitors to Western Australia. *Tourism Recreation Research*, *43*(3), 305–320. https://doi.org/10.1080/02508281.2018.1443052

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. *International Journal of Hospitality Management*, *62*, 120–131. https://doi.org/10.1016/J.IJHM.2016.12.007

Wegmann, J., & Jiao, J. (2017). Taming Airbnb: Toward guiding principles for local regulation of urban vacation rentals based on empirical results from five US cities. *Land Use Policy*, *69*, 494–501. https://doi.org/10.1016/J.LANDUSEPOL.2017.09.025

Xie, K., & Mao, Z. (2017). The impacts of quality and quantity attributes of Airbnb hosts on listing performance. *International Journal of Contemporary Hospitality Management*, *29*(9), 2240–2260. https://doi.org/10.1108/IJCHM-07-2016-0345/FULL/PDF

Xu, F., Hu, M., La, L., Wang, J., & Huang, C. (2020). The influence of neighbourhood environment on Airbnb: a geographically weighed regression analysis. *Tourism Geographies*, *22*(1), 192–209. https://doi.org/10.1080/14616688.2019.1586987

Yang, J. (2022). Big Data Analyzing Techniques in Mathematical House Price Prediction Model. *2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms, EEBDA 2022*, 1174–1177. https://doi.org/10.1109/EEBDA53927.2022.9744970

Zervas, G., Proserpio, D., & Byers, J. W. (2017a). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, *54*(5), 687–705. https://doi.org/10.1509/JMR.15.0204/ASSET/IMAGES/LARGE/10.1509_JMR.15.0204-FIG9.JPEG

Zervas, G., Proserpio, D., & Byers, J. W. (2017b). The rise of the sharing economy: Estimating the impact of airbnb on the hotel industry. *Journal of Marketing Research*, *54*(5), 687–705. https://doi.org/10.1509/JMR.15.0204/ASSET/IMAGES/LARGE/10.1509_JMR.15.0204-FIG9.JPEG

Zhang, H., Zhang, J., Lu, S., Cheng, S., & Zhang, J. (2011). Modeling hotel room price with geographically weighted regression. *International Journal of Hospitality Management*, *30*(4), 1036–1043. https://doi.org/10.1016/J.IJHM.2011.03.010

Zhang, Z., Chen, R. J. C., Han, L. D., & Yang, L. (2017a). Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach. *Sustainability 2017, Vol. 9, Page 1635*, *9*(9), 1635. https://doi.org/10.3390/SU9091635

Zhang, Z., Chen, R. J. C., Han, L. D., & Yang, L. (2017b). Key factors affecting the price of Airbnb listings: A geographically weighted approach. *Sustainability (Switzerland)*, *9*(9). https://doi.org/10.3390/SU9091635

Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: A quantile regression approach. *Journal of Real Estate Finance and Economics*, *37*(4), 317–333. https://doi.org/10.1007/S11146-007-9053-7

Appendix

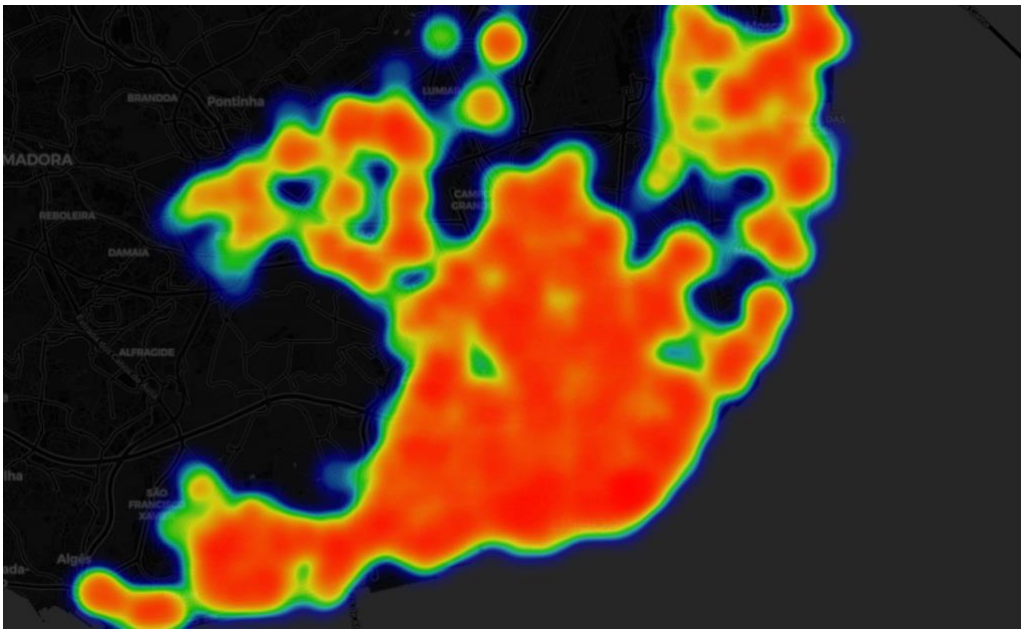

Figure A1 – Folium Plugin Heatmap: House Listings Distribution

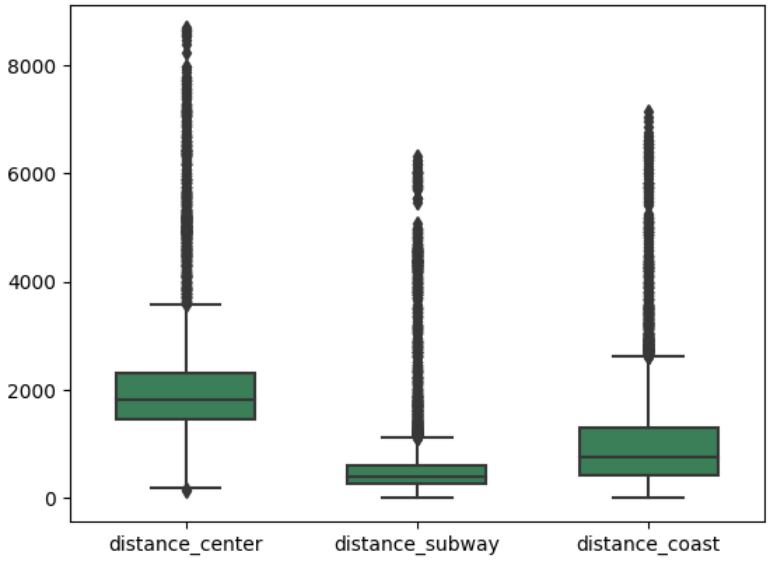

Figure A2 – Distances variables Distributions

Cod Parish	Dsg Parish	Median value m^2 (€)	Yearly growth rate median value (m^2)
170110601	Ajuda	3 750	20
170110602	Alcântara	3 667	16
170110607	Beato	2 857	5
170110608	Benfica	3 173	11
170110610	Campolide	3 398	-3
170110611	Carnide	3 631	9
170110618	Lumiar	3 288	5
170110621	Marvila	5 440	30
170110633	Olivais	2 839	2
170110639	São Domingos de Benfica	3 630	8
170110654	Alvalade	4 217	13
170110655	Areeiro	3 747	13
170110656	Arroios	3 831	15
170110657	Avenidas Novas	4 282	1
170110658	Belém	4 107	8
170110659	Campo de Ourique	4 560	20
170110660	Estrela	4 333	9
170110661	Misericórdia	4 521	17
170110662	Parque das Nações	4 642	7
170110663	Penha de França	3 185	3
170110664	Santa Clara	2 556	2
170110665	Santa Maria Maior	4 574	20
170110666	Santo António	5 753	6
170110667	São Vicente	3 965	11

Table A1 – Median House Value per Parish

	Best Hyperparameters
Random Forest	'max_depth': 20, 'max_features': 10, 'min_samples_leaf': 5, 'min_samples_split': 6, 'n_estimators': 100
Gradient Boosting	'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 1
XGBoost	'eta': 0.05, 'max_depth': 10, 'min_child_weight': 5, 'n_estimators': 100, 'subsample': 0.8

Table A2 – Set of the best hyperparameters for the models