Predicting and explaining Airbnb prices in Lisbon: Machine Learning Approach

Madalena Nunes

Dissertation written under the supervision of Professor Ana Guedes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa,

January 2023.

Abstract

Title: Predicting and explaining Airbnb prices in Lisbon: Machine Learning Approach

Author: Madalena Ribeiro dos Santos Pais Nunes

Airbnb is an online platform that provides listing and arrangement for short-term local home renting services. Since its establishment in 2008, it has offered 7 million homes and rooms in more than 81,000 cities throughout 191 countries. Airbnb price prediction is a valuable and important task both for guests and hosts. Overall, for practical applications, these models can give a host an optimal price they should charge for their new listing. On the consumer side, this will help travellers determine whether the listing price they see is fair. Much research has been done in this field; however, the longitude and latitude of Airbnb listings are often disregarded.

This project focuses on Airbnb price prediction using the most recent (Sep 2021) Airbnb data in Lisbon. Using Google Maps API, the original dataset was enriched with information on the number of ATMs, metro stations, bars and discos within a maximum radius of 1 km. Also, using the geodesic distance, the distance to the airport and the nearest attraction were computed for each listing. A Linear Regression and a Gradient Boosting algorithm were compared based on the original Airbnb dataset and the extended dataset to examine the impact of new features that have been identified. According to the results, all models perform better when the new features are included. The best results are achieved with the Gradient Boosting with the extended data, with an MAE of 0. 3102 and an adjusted R-squared of 0.4633.

Keywords: Airbnb, machine learning, price prediction, xAI, regression, Gradient Boosting

Resumo

Título: Previsão e explicação dos preços do Airbnb em Lisboa: abordagem de machine learning
Autor: Madalena Ribeiro dos Santos Pais Nunes

O Airbnb é uma plataforma online que fornece alojamento de curto prazo. Desde a sua criação em 2008, já ofereceu 7 milhões de residências e quartos em mais de 81.000 cidades, em 191 países. A previsão de preços do Aibnb é uma tarefa valiosa tanto para hóspedes como para anfitriões. No geral, estes modelos de previsão podem oferecer ao anfitrião o preço ideal que deve ser cobrado pelo alojamento. Do lado do consumidor, ajudará os hóspedes a determinar se o preço do anúncio é justo. Muitos estudos já abordaram este tema, no entanto, a longitude e a latitude são frequentemente desconsideradas.
Este projeto foca-se na previsão de preços do Airbnb em Lisboa usando os dados mais recentes (setembro de 2021). Usando a API do Google Maps, o conjunto de dados original foi ampliado adicionando colunas com o número de ATMs, estações de metro, bares e discotecas num raio máximo de 1 km. Além disso, usando a distância geodésica, a distância até o aeroporto e até à atração mais próxima foram calculadas.
Os resultados de uma regressão linear e de um Gradient Boosting, com base no conjunto de dados original do Airbnb e no conjunto de dados alargado são comparados para examinar o impacto das novas variáveis. De acordo com os resultados, todos os modelos apresentam melhor desempenho quando as novas variáveis são incluídas. Os melhores resultados são obtidos com o Gradient Boosting com os dados alargados, com um MAE 0,3102 e um adjusted R-squared de 0,4633.

Palavras-chave: Airbnb, machine learning, previsão de preços, xAI, regressão, Gradient Boosting

Acknowledgements

First, I would like to thank my parents, who always supported me in ways I could never repay. To my brother, who annoys me just as much as he supports me. To my grandparents, who are always there to comfort me with the warmest hug and the wisest words. I strive to be a source of pride and honour for all of you.

A heartfelt appreciation goes to all my friends for always being present, putting things into perspective and making me enjoy all the steps along the way. You are a constant source of joy and motivation in my life.

Last but (surely) not least, I would like to extend my most profound appreciation to my advisor Ana Guedes for her thoughtful comments and invaluable suggestions. I am very grateful for your responsiveness, ideas, and readiness to help, crucial in developing and completing this thesis.

I am grateful to have such a supportive network of people in my life. This accomplishment would not have been possible without all your help and encouragement.

Table of Contents

List of tables

List of figures

List of Appendices

List of abbreviations

API- Application Programming Interface

ATM- Automated Teller Machine

CC- Collaborative Consumption

GBM- Gradient Boosting Model

GLM- General Linear Model

GWR- Geographically Weighted Regression

LR- Linear Regression

MAE- Mean Absolute Error

MSE- Mean Squared Error

OLS- Ordinary Least Squares

RQ- Research Question

SHAP- SHapley Additive exPlanations

SVM- Support Vector Machine

xAI- Explainable artificial intelligence

1.Introduction

1.1. Context and problem definition

Since its establishment in 2008, millions of users have joined Airbnb, a peer-to-peer market that progressively has become a feasible alternative to conventional accommodation services. Airbnb is an online platform that joins two customer groups under the same marketplace, charging a fee for transactions: people who want to rent out their houses and people looking for accommodation. Its offer ranges from community bedrooms to whole appartements, whether for short or long periods of time (Jean Folger 2022). For business trips or vacations, choosing the right listing usually begins with choosing the location, but it is also important to find out how much the guest will be willing to spend. Establishing a fair price is not easy, especially because Airbnb does not intervene in the amount charged, which leads to uncertainty among the users on how to appropriately evaluate their properties at the most profitable value. On the other hand, this pricing dilemma also affects the consumer side which is unsure if the listing price they see is fair. Thus, analysing reasonable forecasts and reasonable suggestions for Airbnb listings can have tremendous real-life benefits and can also be generalized to other applications.

1.2. Motivation and research questions

Among other factors, location is found to be crucial in shaping Airbnb prices (Wang and Nicolau 2017; Cheng and Jin 2019; Gibbs et al. 2018a). The location alone (latitude and longitude, provided in the original dataset) is not interpretable in an immediate approach. For the sake of explainability, in this thesis, the aim is to enrich the original dataset from InsideAirnb (“Get the Data,” 2022.)with distance to important locations, to the most celebrated attractions, and with variables that reflect accessibility and available facilities such as the number of ATMS, metro stations, and bars within the radius of kilometer with the help of Google APIs. Using this dataset and applying Machine Learning, I will use intrinsically explainable models, such Linear Regression, together with black-box models - Gradient Boosting– proposing ways to extract insights, using approaches like SHAP. Altogether, the goal is to predict and explain the prices with the help of Machine Learning, breaking the trade-off between accuracy and explainability.

Motivated by the desire to help predict and explain Airbnb prices in Lisbon, this study addresses the following research question: "*Are geospatial information (latitude and longitude) and derived features relevant to predict prices of Airbnb*?" To answer this research question, it can be divided into the following sub-research questions:

1.*What are the determinants of Airbnb prices in Lisbon?*

2. *Is location one of the most relevant drivers of price in Lisbon?*

3. *Does enriching the dataset with relevant columns (e.g., distance to important locations) improves the power of explanation of the models trained on the original dataset?*

4. *Does every newly identified feature improve the predictions?*

Throughout the thesis, the methodology used to approach these questions is described, and a detailed answer is given in chapter 5.

1.3. Outline of chapters

In the thesis at hand, to guide and answer the reader into the research questions, the organization followed is explained.

First, a comprehensive review of current publications in the area related to Airbnb is examined in section 2. In this section, the several areas of study of Airbnb are discussed, followed by a sub-section with the different Airbnb price determinants and closing with the dissection of the machine learning algorithms used to approach the problem of predicting Airbnb prices.

In section 3, the methodology used to answer the research questions is presented, specifically the data cleaning and pre-processing needed, the exploratory data analysis conducted and its conclusions, and the algorithms applied, along with the error metrics used to evaluated them.

In section 4, the results are described, namely the performance of the algorithms with, and without the extra features, the analysis of the Linear regression (LR) coefficients and the feature importance method of Gradient Boosting Model (GBM). The GBM is explained through a SHAP analysis and the sources of the error of the best model are investigated.

After, in chapter 5 a discussion of the results is conducted and the answers to the research questions are given.

Finally, in chapter 6, the conclusions are taken, the limitations are discussed, and there are also recommendations for future work.

2. Literature review

In recent years, since Airbnb has gained growing popularity, several studies have already addressed not only the problem of predicting Airbnb prices, taking advantage of the publicly available datasets from InsideAirbnb.com, but also several other topics related to this sharing economy. The following sections of this chapter will explore how this past work helps answer the main research questions, their methods, their limitations, and their conclusions. Also, the factors driving Airbnb prices will be explored, and finally, it will be explained how this work differentiates itself from others.

2.1. Airbnb

Sharing economy and online peer-to-peer marketplaces are thriving at an impressive pace, mainly due to technology innovations and supply-side flexibility (Zervas, Proserpio, and Byers 2017) , becoming one of the main alternatives to the hotel industry. Thus, researchers have given much attention to the impact of Airbnb, a pioneer in sharing accommodation, on the traditional hotel sector (Zervas, Proserpio, and Byers 2017). They found that the entry of Airbnb into the Texas market harmed the hotel industry. On the other hand, Mody, Suess, and Dogru (2017) observed that Airbnb and hotel industry demand are essentially different (they target different customer segments). Thus, any impact on the traditional accommodation system is, at best, negligible.

Additionally, with the growth of peer-to-peer accommodations, concern about an increase in rental prices worldwide due to a lack of long-term habitation is also an issue (Barron, Kung, and Proserpio, 2018). The results show that an increase in Airbnb listings affects rental rates and house prices, making this effect slightly higher on the latter. This finding suggests that collaborative consumption has increased homeowners' value of their accommodations due to the occupation of the spare capacity.

Given all these effects and their magnitude, an area of research related to Airbnb that researchers cannot neglect is why people enthusiastically participate in Collaborative

Consumption (CC). Results show that many factors, such as sustainability, enjoyment, and economic gains, drive motivation to participate in CC. Sustainability was found to be only a significant factor for those who value ecological consumption, whereas enjoyment drives people to participate in CC simply because they appreciate engaging with others. Economic gains translate into saving money, a premise that justifies itself (Hamari, Sjöklint, and Ukkonen 2016). Also, on the supply side, the Airbnb platform has near zero marginal cost, in the sense that any room can be added (or removed) with marginal overhead. Most importantly, it can expand supply wherever buildings and apartments already exist. On the consumer side, CC offers a much broader offer than the traditional hotel industry, from rooms to luxury houses (Zervas, Proserpio, and Byers 2017).

Lastly, another significant field of investigation is the prediction of Airbnb prices and price determinants, which will be the focus of this study. This research is valuable not only for the supply side, which can infer a fair and optimal price to charge but also for the consumer side, which can evaluate if the price they see on the platform is reasonable. Furthermore, it is crucial to ensure a price in line with historic data so booking activities can be facilitated and ensure the health of an e-commerce environment. Thus, studying fair predictions and suggestions of prices of Airbnb listings can have real impacts and may also generalise to other areas.

2.2. Airbnb price determinants

As in any prediction problem, one of the first things to access is the possible determinants of the problem at hand. Researchers have enthusiastically investigated Airbnb and hotel price determinants(Christer Thrane 2007; Becerra, Santaló, and Silva 2013; Gibbs et al. 2018; Wang and Nicolau 2017). Hotel price determinants must be looked at carefully but always taken into consideration, given the similarities between both markets.

Wang and Nicolau (2017) investigated the price determinants of sharing accommodation through an OLS analysis of 25 variables about 33 cities and concluded that characteristics related to amenities, location, and accommodation type significantly affect the price charged per night on Airbnb. Concerning amenities, properties with free parking, Wi-Fi, and real beds are evaluated for a higher price.

Gibbs et al. (2018) used a hedonic pricing model (OLS Regression) applied to 15,716 Airbnb listings located in Canada and concluded that location, size of the accommodation and host

characteristics place a massive role in the evaluation of Airbnb. Regarding amenities, Airbnb users highly value Fitness services as a price driver.

Cheng and Jin (2019) adopted a strategy of sentiment analysis and text mining. They found that location and household amenities significantly impact Airbnb experience and satisfaction, thus positively impacting the price rental of a night.

Zhang et al. 2017 used both a General Linear Model (GLM) and a Geographically Weighted Regression (GWR) applied to data from Airbnb listings in Metro Nashville. GWR is different from GLM in the sense that the estimated parameters can vary spatially, which makes sense since the price charged per night in the Airbnb platform and its drivers can vary from location to location. A comparison of both models shows that GWR achieves a higher adjusted R-squared and that some factors were not significant in predicting price in GLM but were significant in GWR, legitimating the use of GWR to the detriment of GLM. Curiously they found that number of reviews and rating scores negatively affect the price. Age was found to positively affect Airbnb listing price, potentially due to more hosts' experience, who are also aware that consumers value past experiences in terms of safety and quality. As in the latter studies, location was found to be a primary driver of price, and distance to the convention centre had a significant negative impact on Airbnb price.

(Lawani et al. 2019) also used sentiment analysis to depict the quality of the listing along with a hedonic spatial autoregressive model applied to rental room prices on Airbnb in Boston to predict which determinants mostly affect hosts' prices per night. Results show that price is mainly affected by review ratings, type of room/ accommodation and especially neighbourhood features, such as distance to the city centre. Interestingly, the distance to the nearest train station and the closest high criminal index area do not appear significant.

2.3. Airbnb price prediction algorithms

Predicting the price of Airbnb is a regression problem that can be solved with the help of supervised machine-learning algorithms. There are several algorithms that we can use, and their performance depends highly on the task at hand. The common course of work is to try different algorithms, compare their performances, and consider previous work.

Tang and Sangani, (n.d.) applied a Support Vector Machine (SVM) with a linear kernel to predict not the exact price but price ranges. In this paper, they employed an extensive analysis of the pictures provided by the hosts (visual features), of the reviews and the description of the accommodations (text sentiment features) and the amenities provided (listing information features).

Liu (2021) also took most of the dataset of September 2021 for San Francisco provided by InsideAirbnb.com and used not only the numerical data available but also the listing reviews as a predictor in the model using Textblob, a sentiment analysis technique to score and classify sentiments implied in those reviews. In this analysis, sentiments were classified as neutral, positive or negative. Six models were implemented, and their performances were compared based on MSE (Mean Squared Error) with and without sentiment analysis. Between the six (Linear regression- Ridge and Lasso, Neural Network, SVM and Regression Tree,) the best performance was achieved with the Regression Tree with sentiment analysis with the least MSE of 0.287 and the highest R-squared of 0.481. The performance of all models is improved when reviews sentiment analysis is included, except for Linear Lasso.

Kalehbasti, Nikolenko, and Rezaei (2019) used the public Airbnb dataset for New York City and applied a similar strategy as the latter work using a Linear Regression as a baseline and comparing its performance to a Ridge Regression, Gradient Boosting, SVM, and Neural Network, using sentiment analysis and feature selection. SVM turned out to be the best-performing model with an MSE of 0.1067 and an R-squared of 0.7768.

Luo, Zhou, and Zhou (2019) also aimed to predict Airbnb prices in three different cities (NY, Paris and Berlin) using three different datasets. They also compared the performances of five different algorithms (Linear regression, K-nearest neighbours, Random Forest, Neural Network and XGBoost), and the neural network yielded the best results with the highest R-squared values for both the train and test set. (Cai, Han, and Wu (n.d.) .

Airbnb price prediction was a problem also studied in Melbourne, where Linear regression, Support Vector Machine, Random Forest regression, and Gradient Boosting algorithms were implemented. Besides traditional machine learning algorithms using tabular data, text analysis (reviews and descriptions) and feature engineering were also performed. Gradient Boosting performed the best among all, followed by random forest.(Cai, Han, and Wu, n.d).

Recently, Thakur et al. (2022) contrasted the results yielded by the most used Machine Learning Algorithms for this problem to the results produced by a Deep Learning algorithm- a four-layer Neural Network. With an R-squared of 0.8104 and an MSE of 0.2356, it outperformed the abovementioned models. Furthermore, this study also differentiates itself from the others since instead of giving a particular price for an Airbnb listing, it outputs a range of prices since pricing is volatile, according to the circumstances. Thus, this is a more realistic study.

2.4. Current research gaps and importance of this study

In previous research, much attention is given to sentiment analysis (Liu 2021; Kalehbasti, Nikolenko, and Rezaei 2019; Lawani et al. 2019; Cheng and Jin 2019) of costumer's reviews of Airbnb listings and the descriptions of Airbnb, together with tabular data. However, geolocation information is often disregarded, ironically ignoring the main findings in section 2.2, which shed light on the fact that location and neighbourhood are fundamental drivers of the price of Airbnb. The focus of this study is to enrich longitude and latitude information with features derived from Google Maps APIs to find valuable and attractive locations nearby such as ATMs, metro stations or bars, and discos and thus enriching the original dataset with information relevant for guests which might lead to performance improvement.

Despite numerous essential factors common to all cities, their idiosyncrasies and peculiarities are often overlooked when determining prices – that is why I found it particularly interesting studying my closest case- Lisbon.

Methodologically, this dissertation contributes by illustrating how machine learning, together with geospatial features and information derived from Google APIS can be used and interpreted visually and realistically in tourism.

3. Methodology

3.1. Data collection and description

In this study, Lisbon was chosen as the study site. Thus, it widens the geographical scope of research on the sharing economy beyond typical studies conducted in the United States. The

dataset was obtained from InsideAirb ("Get the Data," nd). Inside Airbnb is built on publicly available information from the Airbnb website. Several steps have been taken to analyse, cleanse, and aggregate the data to facilitate public discourse. It was scrapped on September 10, 2022. This file contains 74 columns, comprising features or a collection of features, being one of the prices charged per night by the host. Each row corresponds to a different listing. A basic description can be found in Appendix 1.

The source code for the cleaning, pre-processing and modelling developed in this study is available on GitHub at https://github.com/mrspnunes/tese.

The data cleaning and pre-processing for this study involved severalf steps. Due to missing or irrelevant information, some columns were dropped (*id* and *host_id*). Also, columns with high cardinality or categorical variables with only one category, not providing any additional information, were dropped.

Missing values in *host_response_rate* and *host-response_time* represented about 25% of each column. These properties are primarily those that still need to complete a booking (most likely those that have yet to be booked, although they may also include those currently being booked). However, despite making up a considerable portion of the dataset, these listings will still be included in the data as they are legitimate properties with advertised price. They continue to serve as a comparative market when determining what price your Airbnb listing should be listed at. Nonetheless, if the dataset used was based on the average price paid, these rows would have no value since they have yet to be booked.

It should be noted, however, that this group of listings probably contains quite a few 'inactive' listings as well. These are Airbnb listings which are either not bookable (so no stays can be booked) or have significantly higher prices than similar properties nearby. Since *host_response_time* is unknown for about one fourth of the listings, it will be held as its own category: unknown.

A similar story is true for *host_response_rate*, with about a third of null values. The null values for this variable were also kept as their category after grouping other values into meaningful groups (i.e., transforming this into a categorical feature rather than a numerical one). Because about 75% of hosts respond 100% of the time, this was kept as its own category, and other values were grouped into bins.

Referring to neighbourhood, the dataset contains five different columns (*host_location, neighbourhood_group_cleansed, host_neighbourhood, neighbourhood, neighbourhood_group_cleansed*), but only *neighborhood_cleansed* was kept since it had no missing data and the information was concise.

Some cleaning of property types was required as there were many categories with only a few listings. The categories Apartment, House and Other were used, as most properties can be classified as either apartments or house. To avoid odd fractions, missing values in *beds* and *bedrooms* were replaced by the median value.

A new column was added, *host_days_active*, that resulted of converting *host_since,* a datetime column, into a measure of the number of days that a host has been on the platform, measured from the date that the data was scraped (September 10, 2022).

Datetime columns *first_review* and *last_review* had 14% missing values, meaning that 14% of the listings did not have a review written for them. Dropping the columns was not a solution since reviews are crucial in the decision to book or not, and therefore for the price advertised. Also, this was a too large fraction of the dataset to fill with mean/median since it would skew the distribution considerably. Additionally, these missing values are not meaningless or the result of misplaced information- they tell us that these listings are recent in the platform and therefore have not had reviews yet. This information should be included for our models to predict the price of any accommodation correctly. Thus, these variables will be one-hot encoded and therefore treated as categorical and missing values will encompass the 'unknown' category.

In the text column "amenities", which refers to utilities in the apartment, such as Wi-Fi, balcony, tv or coffee machine, only the most frequent amenities were selected and one-hot encoded, as they are significant predictors of prices. One-hot encoding was also applied to categorical variables.

Variables like *price*, *host_response_rate* and *host_acceptance_rate* were strings, and thus the euro and percentage symbols were removed, and only the numeric part was kept. After analysing the distribution of the numeric variables, which can be found in the Appendix 2, there are two main takeaways. First*, longitude*, *accommodates*, *availability_90* and *latitude* are the only attributes that are not severely skewed. Second, the target variable is highly positively skewed and needed attention. After analysing the price distribution, one sees that advertised

prices range from 0 to 12000€. The host's inability to use the advertised prices correctly results in the extreme ends of the range. The advertised prices can be set to any arbitrary amount, offered when no dates are specified. Once one enters the dates he wants to occupy the property, prices can vary greatly. It is hard to believe that a nightly price of an Airbnb would go all the way up to 12000. In short, there are many outliers in price that could throw off the predictions. Only listings with prices above 25€ up to 1000€ were considered, resulting in the 5,42% loss of the original data. The distribution of prices can be seen in Figure 1. The distribution of prices after removing outliers can be seen in Figure 2.

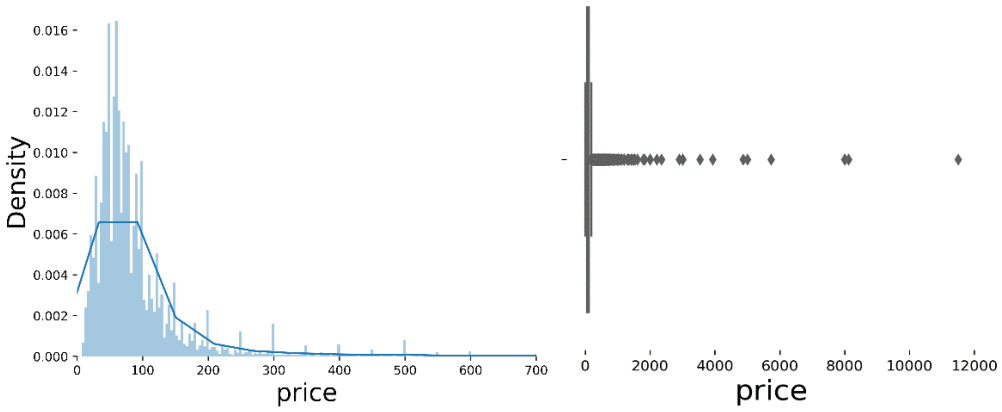


Figure 1 – Distribution of prices before removing outliers: wider variabilty

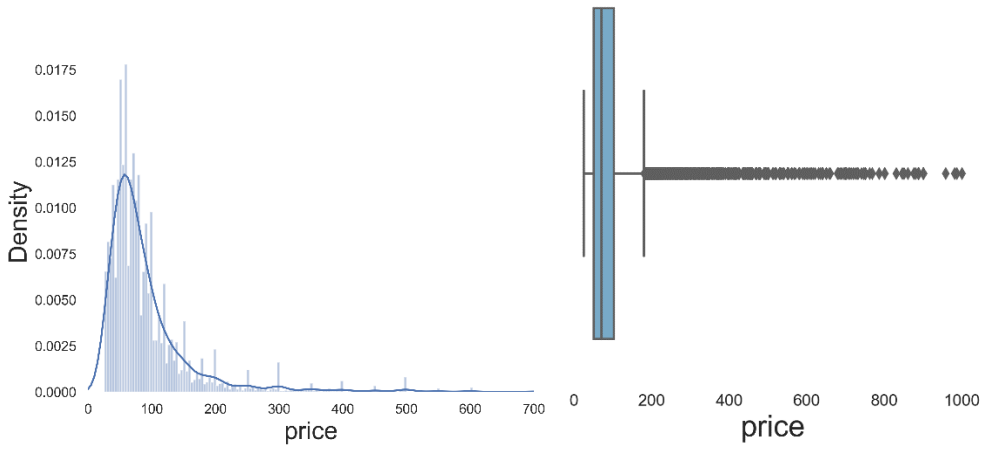


Figure 2 – Distribution of prices after removing outliers: smaller variability

Therefore, this paper's final number of listings is 17 521 rows. Multicollinearity was tested with the help of a correlation heat map (Appendix 3), and columns suspected of multicollinearity were dropped leaving 73 columns in total. A list of the features can be found in Appendix 1.

3.2. Exploratory Data Analysis

After cleaning the data, an exploratory data analysis must be conducted to understand better the data we will be working with. The main libraries used for this purpose were Matplotlib and Plotly Express.

The target variable, the price, was the first focus of attention. Regarding price changes over time, the average price per night for Airbnb listings in Lisbon has increased over the last ten years, with an exception in 2012 (Figure 3).

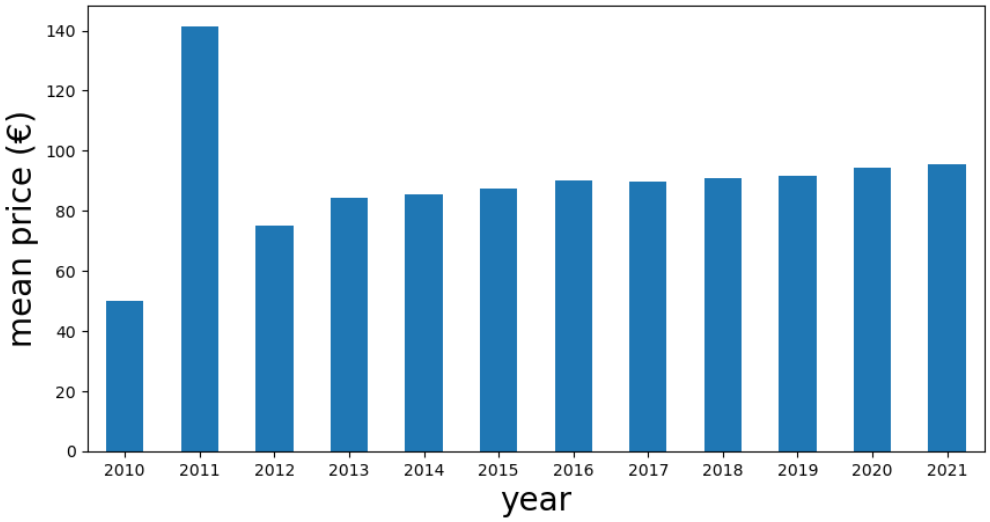


Figure 3- Mean price of listings in Lisbon over the years

In particular, the top end of property prices has increased, resulting in a larger increase in the mean price compared to the median. The mean and the median price in 2010 were 50€, where as the mean price in 2021 (the last complete year of data) was 95.47€ and the median was 72.0 €. This evolution throughout the years can be seen in Figure 4.

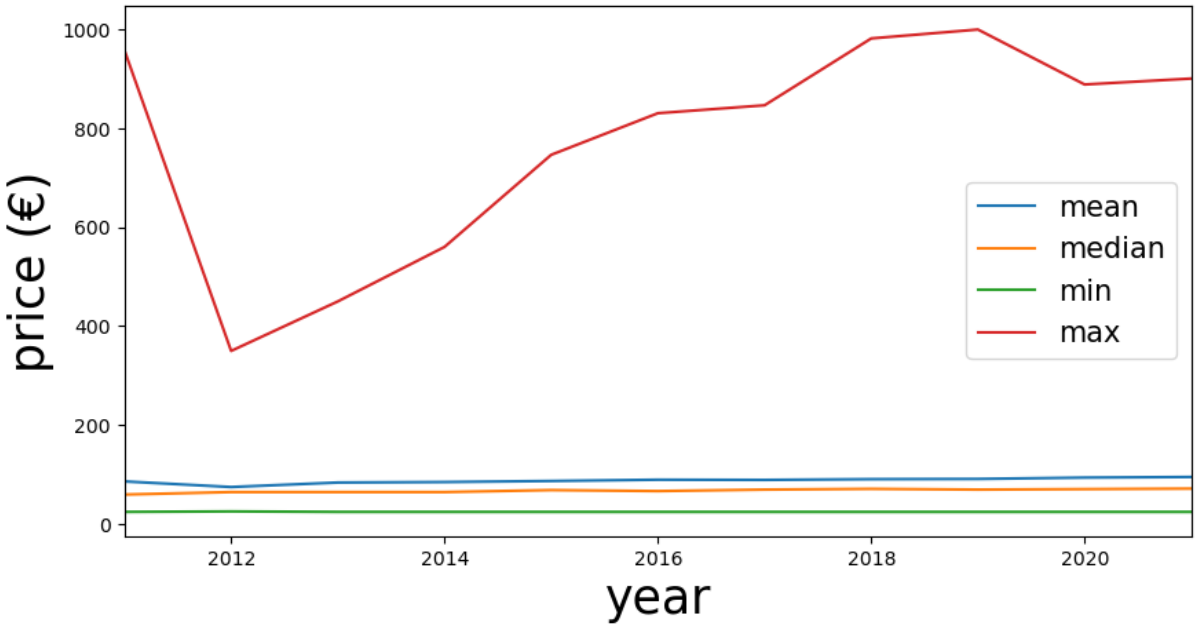


Figure 4- Evolution of minimum, maximum, median and mean price over the years

Also, plotting the prices on a map shows that the most expensive listings are located near the seaside (Figure 5). Additionally, listings far from the city centre usually provide space for more accommodates, therefore, tend to have higher advertised prices. Likewise, the most expensive neighbourhoods are Arrunha dos Vinhos and Vila Franca de Xira with a mean price of 130.33€ and 121.51€, respectively. The most budget-friendly neighbourhoods are Odivelas and Amadora, with a mean price of 67.50€ and 63.49€, in that order.

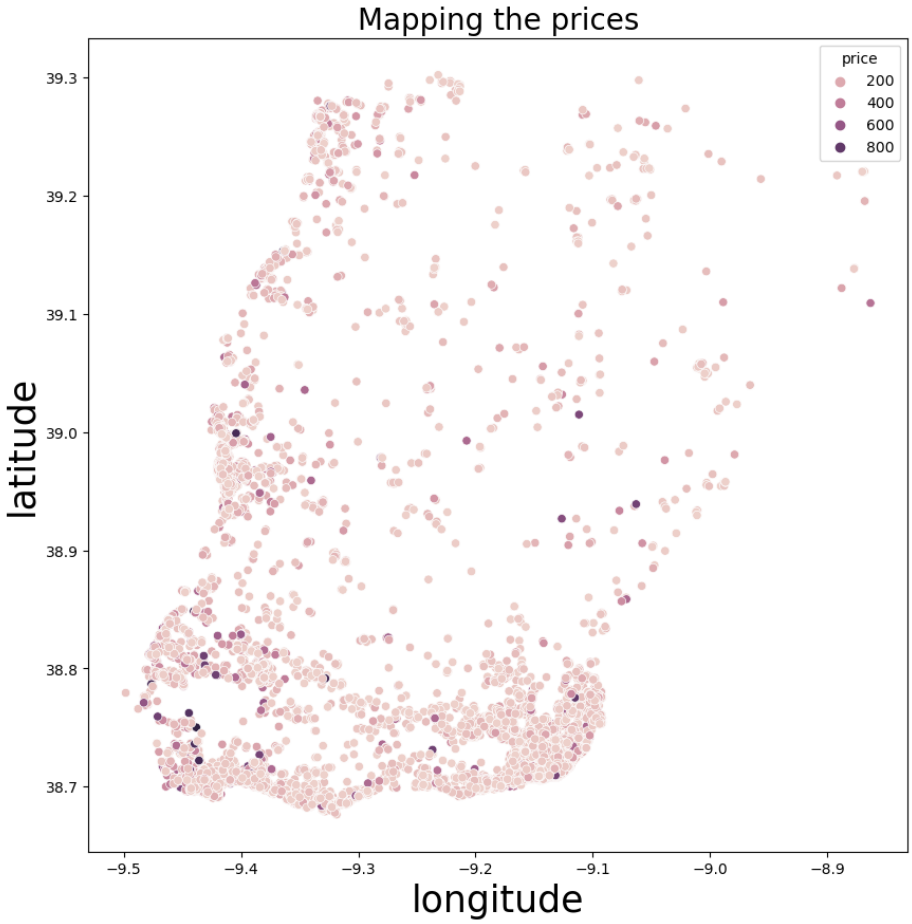


Figure 5- Distribution of the ranges of prices in Lisbon area

Another important pattern to observe is the number of listings per owner/host. Several professional Airbnb management companies host many listings under a single host profile. However, there is no consistent upwards trend in the average number of properties managed by each host (Figure 6). The median number of listings that the host of each listing (*host_listings_count*) has is 3. The mean is higher (20.74) due to some hosts managing some (very) large numbers of listings, as seen below in Figure 6. For example, the host with the highest number of accommodations under the same *id* is 2249 listings. About 21% of listings are from hosts with one listing.

The two main difficulties in discerning how many listings hosts have on average are:

- 1. Hosts with more listings will appear more frequently in the dataset since this number is only known at the listing level. (e.g., a host with ten listings may be represented up to 10 times in the dataset).

- 2. A host's other listings may not be in Lisbon so some multi-listing hosts may appear multiple times in the dataset, and others may appear only once.

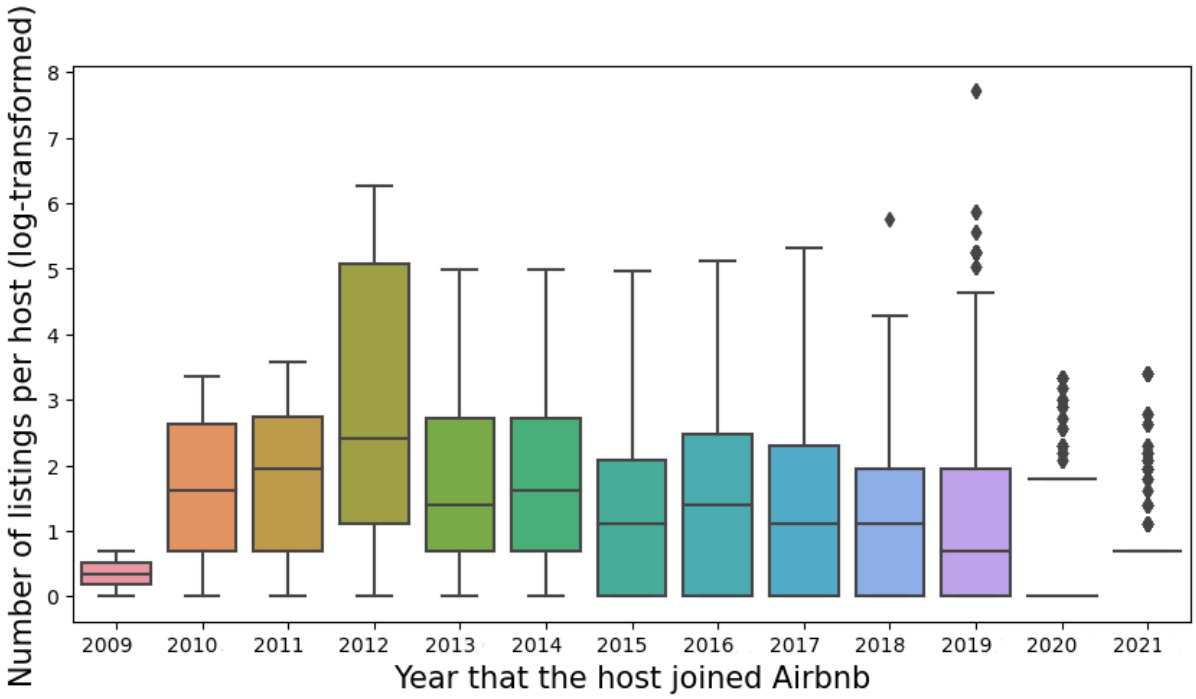


Figure 6- Change per year in the nightly price of Airbnb listings in Lisbon

Unsurprisingly, properties that accommodate more people achieve noticeably higher rates per night, with diminishing returns coming after about 11 people (Figure 7).

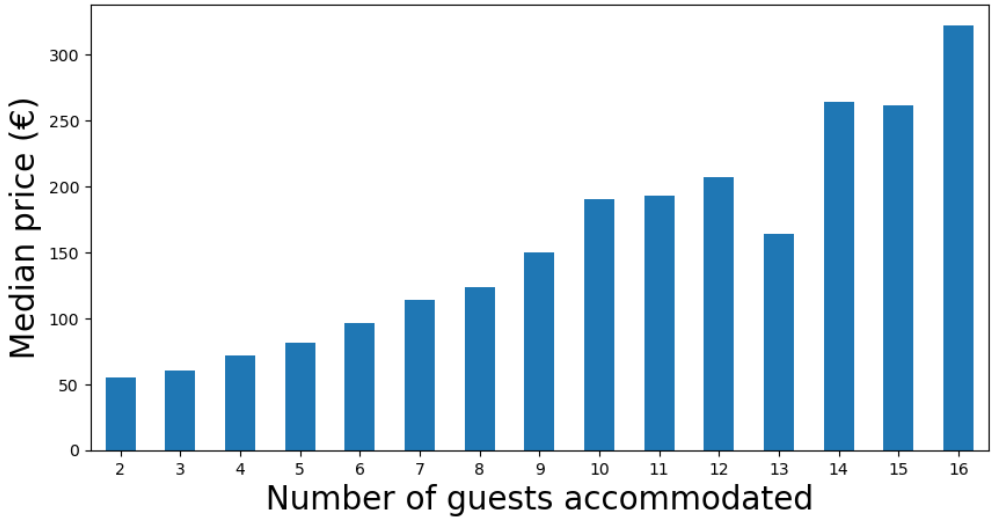


Figure 7- Median price per number of guests accommodated

Concerning the location/ geospatial features, it is possible to use latitude and longitude as indicators of the neighbourhood. However, we will stick with the neighbourhood in this subchapter to understand the story behind the dataset. Yet, in modelling, latitude and longitude will be used with the purpose of reducing the complexity of the model, since one-hot encoding of the *neighbourhood* would result in 128 extra columns. The top 3 neighbourhoods with the highest number of listings are Santa Maria Maior, Misericórdia and Arroios. A better understanding of the density of Airbnb listings in Lisbon can be achieved with Appendix 4.

About 68% of properties are houses. The remainder is apartments or more uncommon property types (e.g. bed and breakfast, boat, windmill, yurt).

About 79% of listings are entire homes (i.e., you are renting the entire property yourself). Most of the remainder are private rooms (i.e., you are renting a bedroom and possibly a bathroom, but there will be other people in the property). Fewer than 1% are shared rooms (i.e., you are sharing a room with either the property owner or other guests).

For every review category, most listings that have had a review have received a 5/5 rating for that category (or 95-100/100 overall). Ratings of 4 or below are rare. Guests are most positive about communication, check-ins, and accuracy. The most common period in which currently live Airbnb listings had their first review is 2-3 years. This means that a lot of listings on the site have been active for at least a couple of years. Relatively few have been active for more than four years, however.

The most common category for the time since a listing received its last review is 1+ years. This means that a lot of listings have not been reviewed recently. In most of these listings, the calendars of the properties are not open, so they are not available for booking, although they technically are live on the site. These are sometimes called 'inactive' listings.

Demand has been increasing since 2011, with a peak in 2022.

Of the Airbnb hosts that are still listed on the site, the first joined on April 22 2009, and the latest joined on December 08 2021. As seen in appendix 5, there is a clear seasonality. Every year, one sees a peak towards hosts joining around the middle of the year (summer), and the lowest points are the beginning and the end of each year. There was a big peak in the number

of hosts joining Airbnb in 2013, 2015 and 2017. 2015 was the year when Airbnb became increasingly popular for short-term leases to get around local legislation and taxation.

3.3. Modelling

In this sub-section, the machine learning algorithms applied are described, as well as the data preparation needed prior.

First, categorical variables were one-hot encoded, meaning we get dummies for the categorical variables to prepare the dataset for multicollinearity analysis. Multicollinearity was then analysed again, and some columns were dropped. At this point, there were still some reasonably strong correlations between highly rated properties of different review categories - i.e., if a property gets a 10/10 for one category, it is likely to get a 10/10 for other categories. However, these were left in so that they could be experimented with later to see if removing them improves the model.

As a second step, the distribution of the numeric features was plotted again to check if any could benefit from a logarithmic transformation. In fact, this conversion significantly improved both the *price* distribution and the distribution of *host_listings_count, bathrooms* and *number_of_reviews.* The models below described (especially Linear Regression) can highly benefit from the log application on price by improving the linearity between the dependent variables. Figure 8 below illustrates the distribution of prices.

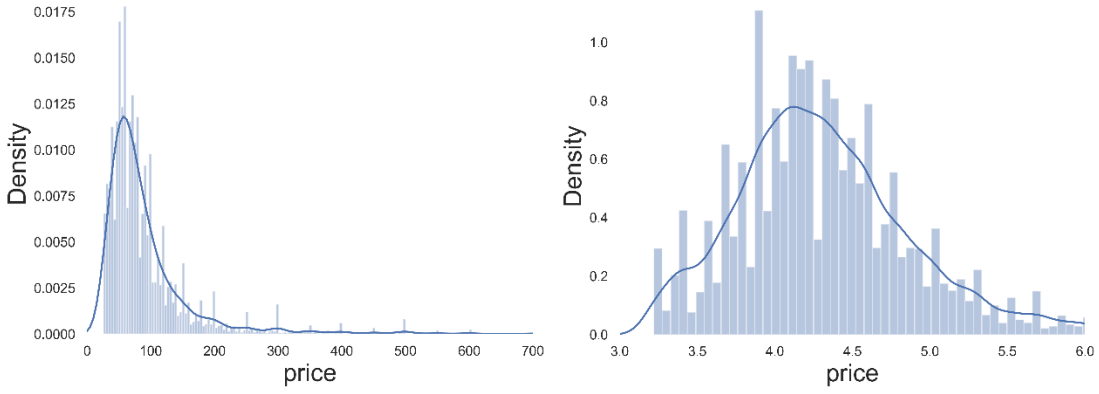


Figure 8- Distribution of prices before(left) and after(right) log transformation

Finally, the dataset was sampled and split into 80/20 train and testing data and the predictive features X and the target feature y were separated. Only then was X scaled to avoid data leakage.

StandardScaler from sklearn was used, which standardises a feature by subtracting the mean and then scaling to unit variance.

After preparing the data for modelling, we can start applying different Supervised Machine Learning models. These algorithms used the data and various assumptions to predict the output variable, the price charged per night of accommodation on the Airbnb platform. For this part, two models of each algorithm were implemented, initially with the original dataset and after with the dataset augmented with new features. A first evaluation was made with the dataset including '*review'* categories, which were included, despite some evident correlations between them, to check if they could lead to some important conclusion, as stated above. Being a sensitive algorithm, with very rigid assumptions, Linear Regression had a poor performance on the validation set compared to Gradient Boosting, having a negative R-squared, i.e., predicting worse than the model that predicts the mean of the values. The feature importance graph of Gradient Boosting was then plotted, which suggested 'review' columns were of null importance to the model, as well as the 'time_since' variables. Therefore, some 'review' columns and all 'time_since' variables were dropped, and this smaller dataset was considered in the development of the project's results (Airbnb dataset).

For this thesis, the two different algorithms used contrast in their explainability: first, A Spatial Hedonic Price Model (OLS Regression) was applied, with the LinearRegression package from the Scikit-Learn library, a model that does not require any additional explanation algorithms for interpretation. Second, the Gradient Boosting model was employed, using the GradientBoostingRegressor from the *sklearn* package, a black-box model that is not intrinsically explainable. In cases where deploying a more complex alternative is not likely to result in significant benefits, simpler (and more interpretable) models are generally preferred. Nevertheless, the predictions of complex models tend to be more accurate when used in specific applications, for example, in computer vision and natural language processing. All Linear Regression models were then improved via Ridge regularisation. Ridge regularization is a technique used in machine learning to prevent overfitting of a model by adding a penalty term to the objective function being optimized. The penalty term is the sum of the squares of the model coefficients, multiplied by a regularization parameter.

3.3.1. Linear Regression

Multiple linear regression is a statistical tool that aims to find a linear relationship between the outcome variable (the variable we want to predict) and its predictors, predicting each explanatory variable's relative importance. Multiple linear regression follows the formula:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta nxn + \varepsilon$$

Where y is the dependent variable, β_0 is the y-intercept, $\beta_1 \ldots \beta n$ are the regression coefficients, $x_1 \ldots x_n$ are the independent variables, and ε is the error term.

This simple algorithm is incredibly fast to train, making it a good baseline for common regression problems. Most notably, the most interpretable regression models are those trained with Linear Regression. However, there are five assumptions that we should satisfy to ensure that the interpretation is valid. These are:

- Linearity: assumes a linear relationship exists between the predictors and the dependent variable.
- Normality of the error terms: it assumes that the model's residuals follow a normal distribution.
- No multicollinearity among predictors: this assumes that the independent variables included in the model are not correlated with each other.
- No autocorrelation of the error terms: the presence of autocorrelation is typically an indication that we are missing some important information.
- Homoscedasticity: assumes an even variance across the error terms.

All these assumptions were tested. Specifically, the absence of multicollinearity was tested previous to fitting the Linear Regression in the data preparation section. All the others were tested after initialising the model with the help of the appropriate plots and statistical tests.

3.3.2. Gradient Boosting

Gradient Boosting is one of the variants of ensemble methods where multiple weak models are created and combined to get better performance, also standing out for its accuracy and speed in

predicting. Fundamentally, the algorithm builds models sequentially and reduces each previous model's errors. We build a new model by using the error or residuals from the previous model as the dependent variable. The objective is to minimise this loss function by adding weak learners using gradient descent. Since it is based on the loss function, we will have different loss functions like mean squared error (**MSE**) or Mean Absolute error (**MAE**) for regression tasks.

The GradientBoostingRegressor from the *sklearn* package was used to build this model. In addition, the default loss function was employed, enabling the optimisation of the squared error of the regression.

3.4. Evaluation metrics

As a measure of goodness of fit adjusted R-squared will be used in place of R-squared because the number of variables differs from model to model. The logic behind this is that R-squared always increases when the number of variables increases, meaning this metric will always penalize the model with fewer variables. The formulas are below, where N is the number of rows and M is the number of columns.

$$R^2 = 1 - \frac{\sum_i (yi - \hat{y}i)^2}{\sum_i (yi - \bar{y}i)^2}$$

$$R_{adj}^2 = 1 - (1 - R^2)\frac{N-1}{N-M-1}$$

For regression problems, two main error metrics are used: MAE and MSE. MAE is the average of all absolute errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |yi - \hat{y}i|,$$

while MSE measures the average of the squares of the errors:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (yi - \hat{y}i)^2$$

The biggest difference is that MAE fails to punish large errors in predictions, whereas MSE accentuates bigger errors and neglects minor errors. Another distinction one should consider is that MAE preserves the original units of the predicted variable, making interpretation easier.

Concerning the focus of this thesis is explaining and extrapolating insights from the models, it is crucial that every result can be interpreted. Thus, MAE will be used for the evaluation purpose.

3.5. Extra data collection

A particularly significant part of this project is the collection of data to enrich the original dataset from InsideAirbnb.com. Despite being a complete and eclectic dataset, some brand-new geospatial features are necessary to answer the main research question. Hence, one can compare the performances and power of explanation of different models: models with and without extra location features.

The first feature identified was the distance to the closest attraction. Lisbon has plenty of attractions, monuments and sightseeing places. However, a brief look at some tourist itineraries showcases Castelo de São Jorge, Terreiro do Paço, Torre de Belém, Jerónimos, Padrão dos Descobrimentos, Time-out Market and Miradouro da Graça as the most longed for and visited tourist sites. These were the attractions considered when computing the geodesic distance, in kilometres, between each of the Airbnb listings and the nearest tourist point using the *geodesic* package from the *geopy* module. Since the original dataset includes the latitude and longitude of each listing, only the coordinates of the attractions mentioned above were searched via Google Maps. Therefore, a new column, *dist_nearist_attraction,* was included.

Direct application of the procedure described above was used to calculate the distance to the airport of each Airbnb listing, resulting in a new variable, *dist_aeroporto*.

Additionally, with the goal of identifying extra geospatial features, Google Maps API was used. Its search mechanism allows one to find information about places based on various categories, such as establishments, prominent points of interest, and geographical locations. One can search for places either by proximity or a text string. When one conducts a Place Search he will receive a list of places along with a summary of each place's features. Concretely, the places considered were ATMs, metro stations and clubs and bars, as some of the essential neighbourhood characteristics that might affect the price of the listing. Metro stations seem important as tourists usually seek places from where they can easily travel around the city. The function *places_nearby* from the *Googlemaps* module in python was used and the output was the number of each feature found in a radius of 1 km since ATMs, metro or bars further away probably do not impact the price charged per night in the Airbnb platform. This application of Google Places API resulted in the identification of three new features: *metro_no, atm_no* and *bars_and_discos_no.* The distribution of the newly identified features can be seen below in Figure 9, as well as the corresponding summary statistics in Table 1.

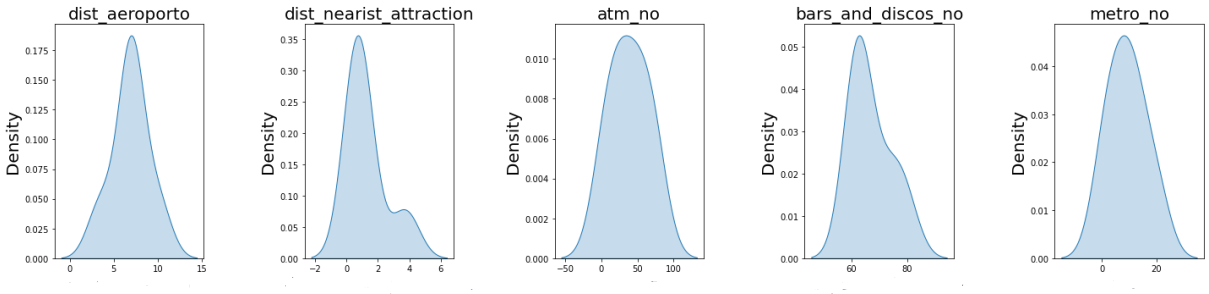


Figure 9- Distribution of the newly identified features

	dist_aeroporto	dist_nearist_attraction	atm_no	bars_and_discos_no	metro_no
count	17443.00	17443.00	17443.00	17443.00	17443.00
mean	6.89	1.25	38.33	67.0	9.176
std	2.09	1.26	27.38	7.29	7.08
min	3.45	0.41	2.0	61.0	0.00
25%	6.80	0.60	20.25	62.25	4.75
50%	6.91	0.72	37.5	63.5	9.0
75%	7.17	1.19	57.0	70.75	12.5
max	10.05	3.75	75.00	79.00	20.0

Table 1- Summary statistics of the newly identified features

3.6. Hyperparameter tuning

Hyperparameter tuning is the process of finding the ideal hyperparameters of a model. Despite the three primary methods to achieve this, Grid Search, Random Search and Bayesian Optimization, the former was selected. Grid Search creates a model for every combination of the hyperparameter to find the best values. It is computationally and time-consuming to perform Grid Search, but ultimately it would yield the best combination of parameters

Hyperparameter tuning was applied to the best model to improve the predictions and the following explainability of how the model is predicting the prices under the hood. The number of folds chosen was ten. Once each fold has been validated, the remaining nine folds are used as training folds. The list of all hyperparameters and their description can be found in Appendix 6. Here, concrete hyperparameters tested with the GridSearchCV (a function that comes in Scikit-learn's package) for Gradient Boosting are shown in Table 2.

Hyperparameters – GBM	Parameters tested	Best setting
n_estimators	500, 1000, 2000	1000
learning_rate	.001, 0.01, .1	0.1
max_depth	1, 2, 4	4
subsample	.5, .75, 1	0.75

Table 2- Hyperparameters optimized with Grid Search for Gradient Boosting

The algorithm will choose a different combination of features on each iteration. Altogether, there are 3* 3 * 3 * 3 * = 81 settings.

3.7. SHAP Analysis

Explainable AI (XAI) aims to find explanations for complex models that humans cannot understand, like the Gradient Boosting used to predict the prices of Airbnb in Lisbon. The applications range from local answers ("*Why is the price of my Airbnb so low*?") to global explanations ("*What is the factor that has the biggest impact on the prices?* ").

SHAP (Shapley Additive exPlanations) is a state of art method in xAI, aiming to explain black-box algorithms,those whose internal processes are not transparent or accessible, despite the input and output being known. SHAP allows breaking the trade-off between accuracy and explainability, often found in these more complex algorithms.

SHAP develops explanations for every dependent variable and every prediction: "By how much and in which direction does this prediction shift when this feature is deleted from the model?" In other words, they quantify a feature's impact on a prediction in terms of its magnitude and direction (positive or negative). *SHAP values,* founded on Shapley values, answer these questions.

Thus, SHAP was applied to improve the transparency and interpretability behind the decisions Gradient Boosting is making when predicting the prices.

4. Results

This section presents the results of all experiments. Detailed performance analysis of the algorithms listed in section 3 is provided for both the original and extended datasets. Next, the Linear Regression coefficients are interpreted and the results from the feature importance method are provided. After, results of the best variable setting using Nested models with new features are provided. Lastly, the results of the SHAP analysis are provided, along with an error exploration of the Gradient Boosting model.

4.1. Linear Regression and Gradient Boosting's performance

To assess the performance of each model, Mean Absolute Error (MAE) and Adjusted R-squared are compared. All the results refer to the validation set that provides an unbiased estimation of the predictive performance compared to the training set, used to train the model.

The baseline model, Linear regression, achieved an MAE of 0.3394 and an adjusted R-squared of 0.3694, while Gradient Boosting performed with an MAE of 0.3161 and an adjusted R-squared of 0.4424. With the addition of the extra features, Linear regression's adjusted R-squared increased to 0.3737 and Gradient Boosting to 0.4633. A summary table is presented below (Table 3).

Algorithm	Dataset	MAE	Adjusted R-squared
Linear Regression	Airbnb	0.3394	0.3677
	Airbnb+ new features	0.3383	0.3737
Gradient Boosting	Airbnb	0.3161	0.4424
	Airbnb+ new features	0.3102	0.4633
Hypertuned GBM	Airbnb+ new features	0.2857	0.5221

Table 3 – Comparison of performance between models

It is evident that each model performs better when new features are added than without them, so the predictions have been improved. Hypertuning the parameters also led to an enhancement

in performance, with a MAE of 0.2857 and an Adjusted R-squared of 0.5221. These results mean that Gradient Boosting was able to explain 52.21% of the variation in the outcome variable and that, for every prediction, on average, the error is 1.33€ after the exponential transformation.

4.2. Linear Regression coefficient analysis

The most significant advantage of Linear Regression is its transparency and simplicity of interpretation, a consequence of the first OLS assumption, linearity. Moreover, as we can see in the equation in section 3.3.1, the effects of each feature (the betas) are addictive and, thus, easy to separate.

Using the data without the newly identified features, the coefficient with the greatest magnitude is '*accommodates'*, with a 0.2657 coefficient after the e^{β} -1 transformation, meaning that if a given listing can accommodate one more person, its price increases by 0.27€, on average, keeping everything else constant. *Number_of_reviews* has a coefficient of -0.1709, meaning that if a listing has one extra review, the mean price decreases by 0.1709€, all else constant. Likewise, *room_type_Entirehome/apt* has a coefficient of 0.1299, indicating that an Entire home is, on average, 0.13€ more expensive than if the Airbnb was only a private room, the omitted variable.

On the other hand, when Linear Regression is implemented with the extended dataset, the highest coefficient belongs to *atm_no*, with a value of 116.00, this is, an increase of 1 ATM in the radius of 1 kilometre of a listing has a positive impact on the price of 116€, on average, all else constant. This number clearly stands out and will be discussed further. It is also important to notice that all five newly identified features appear in the top ten coefficients with the highest magnitude, shedding light, once more, on the importance of these location variables in explaining the prices of Airbnb. Below are summary tables, with the ten highest coefficients, in module, in descending order of magnitude, with equivalent interpretations.

Variable	Coefficient
accommodates	0.265760
number_of_reviews	-0.17087
room_type_Entirehome	0.129898
hot_tub_sauna_or_pool	0.086687
air_conditining	0.077929
latitude	-0.059353
bathrooms	0.058830
availability_360	0.032390
minimum_nights	-0.029945

Table 4– Top nine higher Linear Regression Coefficients with the original dataset (in magnitude)

Variable	Coefficient
atm_no	116.00000
reviewscores_rating_no	-1.00000
metro_no	-0.989994
accommodates	0.265367
number_of_reviews	-0.229749
bars_and_discos_no	-0.196391
dist_nearist_attraction	0.141575
room_type_Entirehome	0.134665
dist_aeroporto	0.108214

Table 5– Top nine higher Linear Regression Coefficients with the extended dataset (in magnitude)

4.3. Gradient Boosting feature importance

A very significant part of the explainability of our black-box models is the feature importance, the process of calculating a score for all the input features for a given model based on their effectiveness in predicting the outcome. This score reduces the importance of a feature to a number, which can then be compared to the importance of other features. The higher the value, the more important the feature. Below we can see the ten more important features of Gradient Boosting, with and without extra features.

The most important feature is how many people the property accommodates, which is in accordance with the economic theory: if a listing can accommodate more people, the price will be higher. Based on the large difference in importance between the top two features, providing more space for more people may be better overall. It is also essential to note that latitude and longitude appeared as the seventh and eighth most important features, respectively, before the

data extension, highlighting the importance of location features in the prediction of Airbnb prices.

On the other hand, in Table 7, three of the five new variables are represented in the ten most important features, namely the number of ATMs, the number of metro stations and the number of bars and clubs in the radius of 1 kilometre, shedding light on the importance of these features in the prediction accuracy.

Feature	Weight
accommodates	0.46339
bathrooms	0.11648
number_of_reviews	0.05916
air_conditioning	0.05812
room_type_Entire home/apt	0.05806
hot_tub_sauna_or_pool	0.04166
latitude	0.03045
longitude	0.02748
availability_365	0.02745
host_days_active	0.02276

Table 6- Top ten more important features and their weights according to Gradient Boosting- before extension of data

Feature	Weight
accommodates	0.43192
bathrooms	0.10877
number_of_reviews	0.06051
room_type_Entire home/apt	0.05933
air_conditioning	0.05529
hot_tub_sauna_or_pool	0.03902
atm_no	0.03099
metro_no	0.02903
bars_and_discos_no	0.02883
availability_365	0.02303

Table 7- Top ten more important features and their weights according to Gradient Boosting- after extension of data

4.4. Additional analysis of the new variables' contribution

In section 4.1, we observed that the dataset with the extra five features improved the predictions overall, adding important information to the models. However, we are interested to know if all five variables really improve the models. The results of Linear Regression and Gradient Boosting, with subsets of the entire extended dataset, are presented below to answer this question.

Variables included	MAE	Adjusted R-squared
(1) airbnb	0.3394	0.3677
(2) airbnb+dist_nearest_attraction	0.3394	0.3697
(3) airbnb+dist_nearest_attraction+dist_aeroporto	0.3391	0.3709
(4) airbnb+dist_nearest_attraction+dist_aeroporto+ atm_no	0.3385	0.3716
(5) airbnb+dist_nearest_attraction+dist_aeroporto+ atm_no+ metro_no	0.3383	0.3737
(6) airbnb+dist_nearest_attraction+dist_aeroporto+ atm_no+ metro_no + bars_and_disco_no	0.3383	0.3741

Table 8 – Results of Linear Regression nested model

For Linear Regression, despite being minor, there is an improvement of performance with the addition of every variable.

Variables included	MAE	Adjusted R-squared
(1) airbnb	0.3161	0.4424
(2) airbnb+dist_nearest_attraction	0.3394	0.4430
(3) airbnb+dist_nearest_attraction+dist_aeroporto	0.3162	0.4457
(4) airbnb+dist_nearest_attraction+dist_aeroporto+ atm_no	0.3102	0.4636
(5) airbnb+dist_nearest_attraction+dist_aeroporto+ atm_no+ metro_no	0.3102	0.4593
(6) airbnb+dist_nearest_attraction+dist_aeroporto+ atm_no+ metro_no + bars_and_disco_no	0.3102	0.4633

Table 9 – Results of Gradient Boosting nested model

In the combination number (2) of Table 9 of variables added, *dist_nearest_attraction* occupies the eighth place in the feature importance scale. In the dataset with the variables in (3), *dist_nearest_attraction,* however, occupies the tenth position and *dist_aeroporto* the seventeenth place. More interestingly, in the combination (4) *atm_no* occupies the third position in the feature importance table.

Overall, all features have approximately the same effect on improving or deteriorating model performance, with emphasis on the number of ATMs that adds a considerable amount of information to the model.

4.5. SHAP analysis

As explained in section 3.6, SHAP aims to justify, and thus validate, the prediction of a variable x by calculating the impact of each variable on the outputted prediction, making it an essential tool to make sense of the outcome of black-box models, like Gradient Boosting.

For instance, linear models can measure a feature's overall importance using its coefficients. However, these are scaled with their scale, which can lead to distortions or misunderstandings. Additionally, the coefficients do not consider the local importance of the feature and its changes as values increase or decrease. The same occurs in a tree-based model's feature importance, which is why SHAP is helpful for model interpretation.

To understand the main features globally impacting the predictions yield by the Gradient Boosting , the *beeswarm* plot is a great tool, ordering features by their effect on the prediction and depicting, at the same time, how lower and higher values affect the final price estimate. On the horizontal axis, we have the SHAP value. In contrast, on the vertical axis, we have the colour of the point to determine whether that observation had a higher or a lower value when compared with others. For example, in Figure 10 below higher values of *accommodates, bathrooms* and *availability_365* produce higher SHAP values. On the other hand, lower values of latitude lead to higher SHAP values.

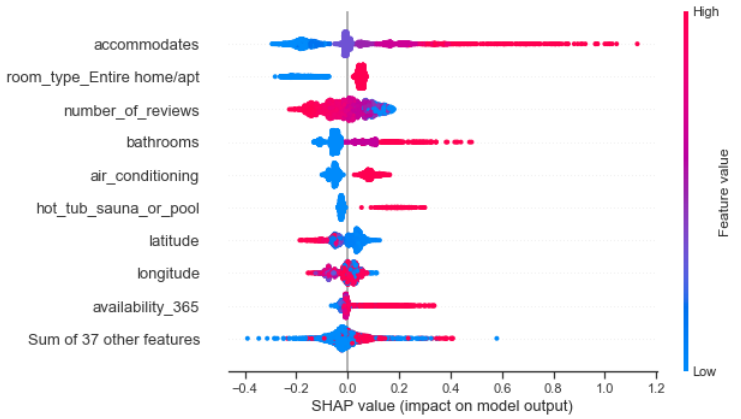


Figure 10- Summary of how the top features in the original dataset impact the GBM's output, showing accommodates has the highest positive impact on the price

SHAP analysis relies heavily on SHAP values and the plots based on them. Below in Figure 11, we have a force plot of the Gradient Boosting with the original dataset that allows an understanding of the contribution of each feature to the prediction for a given listing. In other words, it allows for a local analysis. As if competing with one another, the positive SHAP values appear on the left and the negative on the right sides of this plot. The value in bold represents the predicted price for this given listing, the 100th in the dataset, corresponding to 76.71€ after the exponential transformation. Each SHAP value expresses the marginal effect that the observed level of a variable for a listing has on the final predicted price for that listing. Overall, *accommodates*, *bathrooms*,*number_of_reviews, air_conditioning* and *room_type_entire_home/apt* are the variables contributing the most to the final predicted price for this given observation.

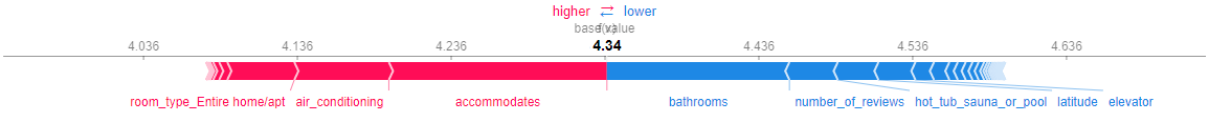


Figure 11- Summary of how the top features in the dataset impact the GBM's output for the 100th listing in the original dataset

The Figure 12 below shows how high values of *dist_aerporto* negatively affect the predicted price and how low values of *dist_nearist_attraction* negatively impact SHAP values.

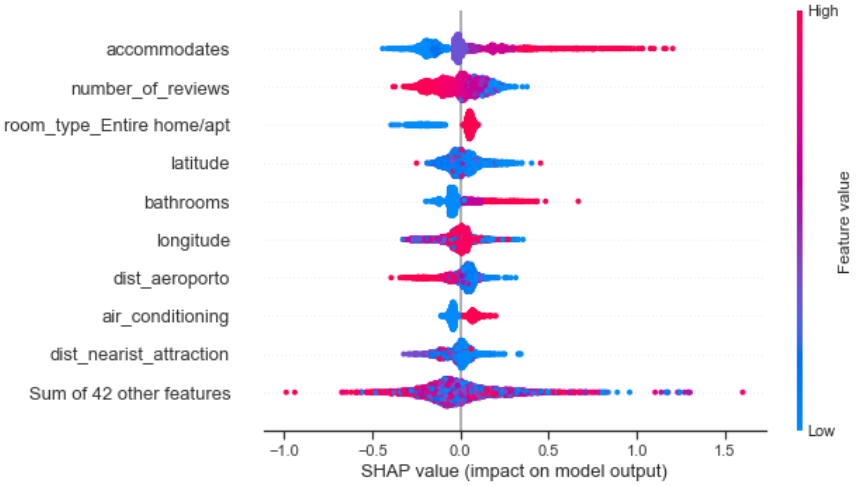


Figure 12- Summary of how the top features in the extended dataset impact the GBM's output, showing accommodates has the highest positive impact on the price

The same analysis was performed on the Gradient Boosting with the extended dataset. For the same 100th observation, the force plot now portrays *accommodates*, *atm_no and dist_aeroporto*

as the main features positively affecting the price. On the other hand, *dist_nearist_attraction*, *bathrooms* and *number_of_reviews* are negatively impacting the price, for this given listing. With the extended dataset, GBM predicts a price of 74,44€ for the same listing.

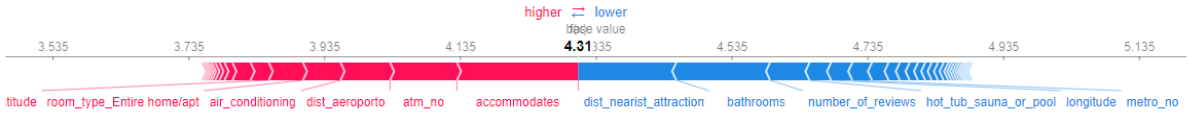


Figure 13- Summary of how the top features in the dataset impact the GBM's output for the 100th listing in the extended dataset

It is also interesting to perform this local analysis in listings in different locations, to see how SHAP explanations change in different geospatial conditions, specifically to check if the newly identified features are still relevant.

The SHAP results in Figure 13 refer to a listing in Sintra, somewhat far from the metropolis. The Figure 14 below is referent to a listing in Campo de Ourique, in the city center. It is interesting to notice that none of the newly identified features appear as the main variables affecting the price. As before, *accommodates* positively affects the price, as well as *longitude* and *high_end_eloctronics.* However,*Bathrooms*, *number_of_reviews* and *air_conditioning* negatively affect the price for this listing.

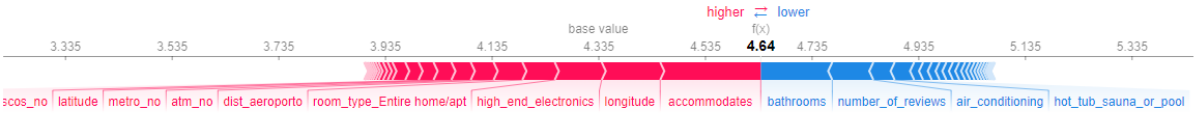


Figure 14- SHAP force plot for a listing in the city center

4.6. Error analysis of GBM with the extended data

When we reduce a model's performance to one number, it is oversimplifying and ignoring the source of the errors and how different they are from each other. To do a fair and accurate evaluation of the model's performance, we care about the local analysis, i.e., we want to know where the model does not perform so well. We can train the model in the different dimensions we care about and where error may be more represented, like different neighbourhoods, different price ranges or different properties. However, these represent many conditions, apart from those one would not even consider. To accelerate this analysis,the ErrorAnalysisDashboard tool from the raiwidgets package was used, which will be explained in this section.

As a first step, error analysis identifies the data cohorts with a higher error rate than the overall benchmark. Next, a dashboard is outputted, including an error heatmap and a decision tree allowing error exploration.

Based on the benchmark data, the decision tree finds interpretable subgroups with surprisingly high or low error rates. For example, as shown in Figure 15 below, the right side of the three comprehends 78.07% of the total error of the model. This cohort is about listings whose distance to the airport is below 14.73 km and the host's days active in the platform are above 1379 days, corresponding to 3 years and nine months, after performing the operation inverse of standardisation.

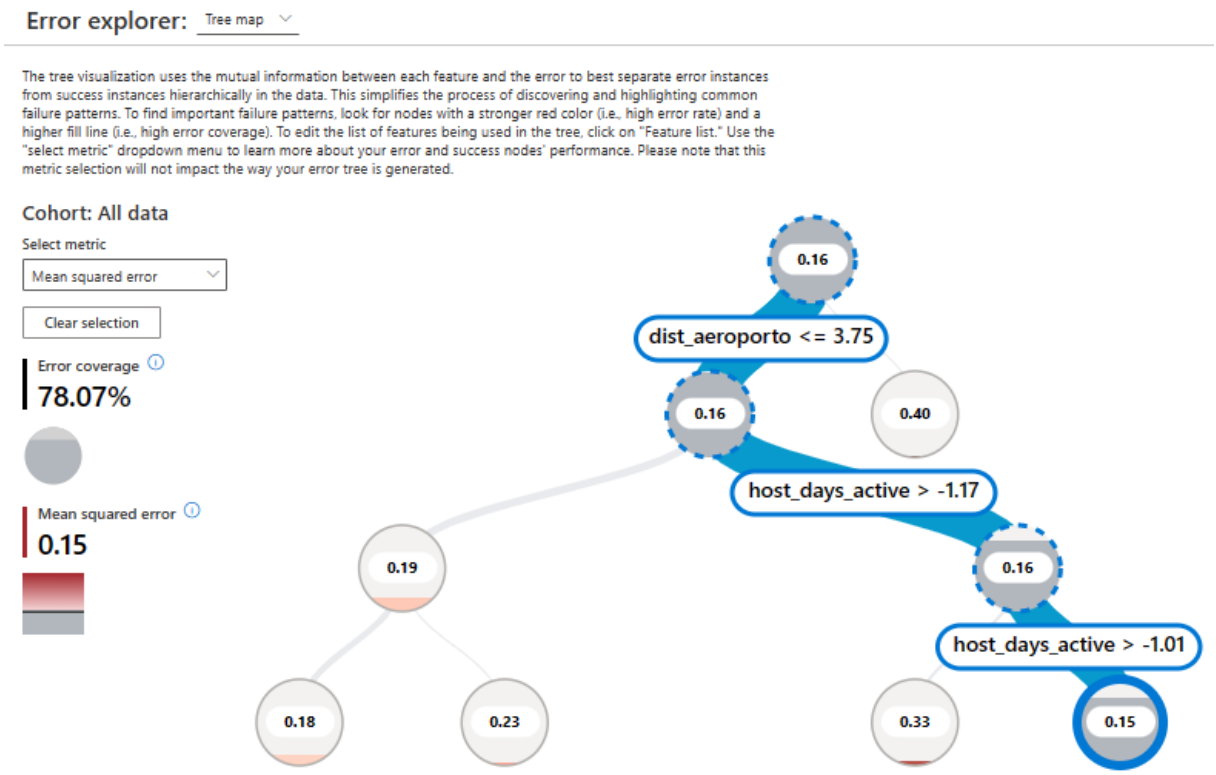


Figure 15- Error analysis dashboard of GBM

The first branch, dist_aeroporto< 3.75, oversees 99.02% of the total error, whereas the branch dist_aeroporto >3.75 only represents 0.91% of the error But why? The 'Explanation' tab allows us better to explore the reasoning behind the genesis of the error. Taking this into account, two cohorts were created, dist_aeroporto >3.75 ('big_dist') and dist_aeroporto< 3.75 ('small_dist') and analysed with the 'Explanation' tab. Below, Figure 16 depicts the top five most important features in each cohort. The biggest difference between the two is that in the 'small_dist' cohort the model leverages less important features to predict the output. In contrast, in the big_dist

cohort the model relies on latitude, which, as we have seen, is a fundamental feature in predicting Airbnb prices.

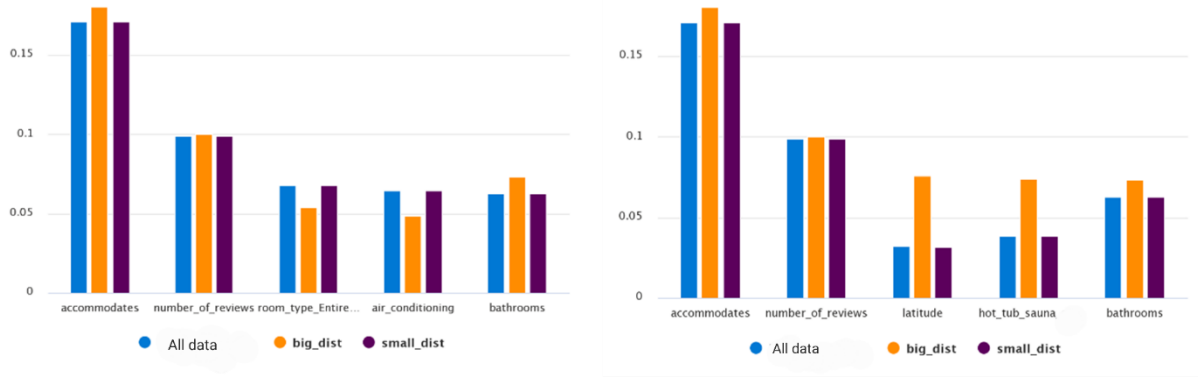


Figure 16- Top 5 most important features in the 'small_dist' (left) and 'big_dist (right) cohort

Analysing the errors by geographic area, property type and room type is also important to understand where the model fails more, so we can look at those predictions with caution.

Figure 17- Distribution of the errors of GBM in Lisbon area depicting the coastline as the area with higher error rate

In figure 17 above, we can see that the error is greater near the coast, with more frequency of darker points, that diminishes towards the centre.

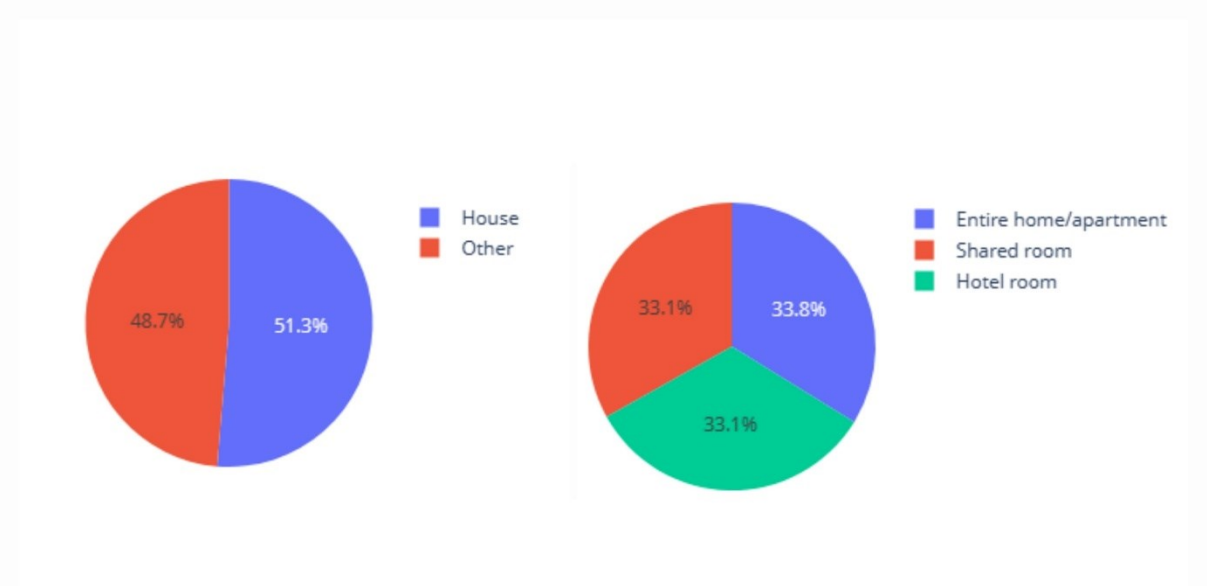


Figure 18-Pie chart of the error rate per house(left) and room (right) type

Above, in Figures 18, we have the distribution of the error of the model per room and property type. However, across the most relevant segmentation (property type and room type) error rates are very similar, suggesting that the used modelling approaches are not underperforming or overperforming in any of these segments.

5. Discussion

In this section, the results presented in section 4 are discussed and interpreted to give an answer to the research questions. Also, the limitations of the present study are described.

5.1. Findings

The present study has as its primary goal answering the research question " *Are geospatial information (latitude and longitude) and derived features relevant to predict prices of Airbnb?".* We need first to answer the sub-research questions, so we can be in conditions of answering the main research question.

1. *What are the determinants of Airbnb prices in Lisbon?*

The goal of this RQ is to give insights into the importance of this study and the main RQ, so the original dataset, without the extra features, will be considered when answering it.
To have the first answer to this question, we looked at the correlation matrix of price vs all the features where *accommodates, latitude, availability_365 and review_scores_checkin* are depicted as the numeric variables more correlated with price. For the categorical variables, several amenities appear on the top variables correlated with price, such as hot_tub_sauna_or_pool, tv, balcony and child_friendly. The Linear Regression implemented came somehow confirm this. As seen in section 4.2, the highest coefficients, in absolute number, belong to accommodates, number_of_reviews, room_type_Entirehome/apt, latitude and bathrooms. Amenities like hot_tub_sauna_or_pool and air_conditioning also stood off. Gradient Boosting feature importance leads us to the same findings, adding longitude as a key driver of price.

2. Is location one of the most relevant drivers of the price of Airbnb in Lisbon?

This sub-research question has the same purpose as the former. As seen above, latitude and longitude are very relevant drivers of the price of Airbnb in Lisbon. Despite being universal and objective indicators of location, these are hardly interpretable. However, these allow for extracting useful neighbourhood characteristics, already interpretable. Also, it corroborates and validates the main RQ and the methodology used.

3. Does enriching the dataset with relevant columns (e.g., distance to important locations) improve the power of explanation of the models trained on the original dataset?

With the purpose of answering the third research question, Linear Regression as a baseline and Gradient Boosting were applied both to the original dataset and to the extended dataset with the new geospatial features. Both the Linear Regression and the Gradient Boosting trained with the extended dataset outperformed the baseline, as evident in the comparison of results in section 4.1. The results indicate Gradient Boosting with the extended dataset as the best-performing

model with an adjusted R-squared of 0.4633 and an MAE of 0.2857. These results are in accordance with the expectations, even outperforming the results obtained by Liu (2021), who compared several machine learning algorithms, including sentiment analysis and achieved the best model with a Regression Tree with the highest R-squared of 0.481 and the least MAE of 0.385.

Looking at these results, it becomes evident that enriching the dataset with relevant location columns improves the predictions, answering the third sub-research question.

However, as seen in figure 14, none of the newly identified features appear to be relevant in the city centre. Despite being a local analysis and consequently, the results can change, the fact that the extra features do not appear relevant in this city centre listing may be due to the fact that the offer of metro stations, ATMs and bars is much superior and thus not so valued. Also, the distance to the nearest attraction may not be as relevant, since the proximity is higher overall.

4. *Does every newly identified feature improve the predictions?*

To answer this sub-research question, we will look at the coefficients of the Linear Regression, the feature importance method, SHAP analysis and the results of nested models with Linear Regression and Gradient Boosting.

When considering the coefficients of the Linear Regression, all five new features are valuable for predicting the price, emerging in the top ten highest coefficients, with emphasis on the number of ATMs. The coefficient of *atm_no* is 116, undoubtedly standing out and even nonsensical in the optic of the economic theory. All potential roots of the issue were analysed and no problem was identified. The variable was scaled and when using the function *places_nearby* all the possible filters were used. Hence, further analysis is needed before making conclusions about this number. Nevertheless, by inspecting the ten most important features of Gradient Boosting, the number of ATMs showed as the most significant variable from the new features, supporting the findings of the LR. Additionally, the distance to the airport and the distance to the nearest attraction seem irrelevant.

It is curious how two of the new features, the distance to the airport and the distance to the nearest attraction, that were computed in the same way, are relevant in the LR and negligible in the GBM. Lisbon airport can be considered located in the central part of the capital, as well

as the attractions considered to compute *dist_nearest_attraction*. Therefore, this type of information can be linked to neighbourhood information that appears to be already included by other variables, despite the lack of specific relationships between existing variables in the multicollinearity matrix. Thus, these features did not affect GBM, a decision-tree-based algorithm, since they are robust to multicollinearity. At the same time, LR highlights the slight importance of these features.

From the global analysis performed by SHAP, only *dist_aeroporto* and *dist_nearest_attraction* from the new features appear as the top features affecting the price

All in all, only some of the five newly identified features improve the models' predictions. However, the features computed resorting to Google Maps APIs, namely the number of ATMs, number of metro stations and bars and clubs, appear to add important information to the models, improving their power of explanation. The fact that the improvements are minor may indicate they are redundant with other features already in the dataset.

5.2. Limitations

This analysis has different limitations. To begin with, rather than the actual realised price in the marketplace, it used the listing price advertised on the Airbnb platform as a dependent variable. Hedonic pricing literature refers to this as typically being the case with accommodations. However, no claims can be made about the accuracy of the hosts' perceptions or the effectiveness of their pricing strategies. Moreover, the research only considered the price of each listing at one point in time, so it was not possible to account for seasonal changes in the contribution of certain attributes (for instance, swimming pools may play a bigger role in the summer) or seasonal changes in prices (e.g., in December, with Christmas markets, some cities centres may be more demanded). Finally, on the consumer side, more than the price at different points in the same year, it is important to have the price of the listing at the same time in past years. If, for instance, the price is constantly decreasing, customers expect a lower price, so it might be helpful to look at historical prices as well.

6. Conclusions

This thesis answers the main RQ "*Are geospatial information (latitude and longitude) and derived features relevant to predict prices of Airbnb?".*

The original dataset from InsideAirbnb.com was enriched to answer this RQ with five new features. The distance to the airport and the distance to the nearest attraction were calculated using the geodesic distance for each listing. The number of ATMs, metro stations and bars within a radius of one kilometre was computed by resourcing to Google Maps API. The Linear Regression and Gradient Boosting results using the original and the extended dataset with the newly identified features were then compared. For both models, including the extra variables improved the performance and power of explanation.

The best model was obtained with the Gradient Boosting trained with the extended data, with an MAE of 0.2857 and adjusted R-squared of 0.5221. These results mean that Gradient Boosting was able to explain 52.21% of the variation in the outcome variable and that, for every prediction, on average, the error is 1.33€ after the exponential transformation.

As part of its contribution to existing literature, this thesis identifies new features for each listing in the Airbnb database. Also, a lot of attention was given to explaining and interpreting the results, a field injudiciously disregarded. After all, we want more than just an algorithm that tells us what the price of an Airbnb should be. We also want to explain why it is the predicted price and which factors we should look for if we want to increase the price of our advertised listing or a lower price for the listing we are looking for. Moreover, analysing the roots of the models' errors and how they are distributed geographically is an area, to the best of my knowledge, not explored in the existing literature. Therefore, this research distinguishes itself from the existing research with the use of the most recent packages for Python. As a whole, this study identifies new factors that affect Airbnb price prediction.

In future work, it would be interesting to apply Natural Language Processing to the reviews and the listings descriptions so we could include these features in the models, thus improving their performance. The most exciting ways to implement this would be to conduct sentiment analysis on the reviews or to extract the most common words in the listings' descriptions.

Furthermore, it is also possible to access the website and analyse the images of properties provided in the dataset using a convolutional neural network model.

As mentioned in the data source section, this thesis' methodologies can be expanded to other cities, which can also be newsworthy.

References:

Barron, Kyle, Edward Kung, and Davide Proserpio. n.d. “The Effect of Home-Sharing on House Prices and Rents: Evidence from Airbnb.” https://ssrn.com/abstract=3006832.

Becerra, Manuel, Juan Santaló, and Rosario Silva. 2013. “Being Better vs. Being Different: Differentiation, Competition, and Pricing Strategies in the Spanish Hotel Industry.” *Tourism Management* 34: 71–79. https://doi.org/10.1016/j.tourman.2012.03.014.

Cai, Tiancheng, Kevin Han, and Han Wu. n.d. “Melbourne Airbnb Price Prediction.” https://drive.google.com/open?id=1D32jVpSfvEYCDCoVt6FYxS98KVFhp3Fm.

Cheng, Mingming, and Xin Jin. 2019. “What Do Airbnb Users Care about? An Analysis of Online Review Comments.” *International Journal of Hospitality Management* 76 (January): 58–70. https://doi.org/10.1016/j.ijhm.2018.04.004.

Christer Thrane. 2007. “Examining the Determinants of Room Rates for Hotels in Capital Cities: The Oslo Experience.” *Journal of Revenue and Pricing Management* .

“Get the Data.” n.d. Http://Insideairbnb.Com/Get-the-Data/.

Gibbs, Chris, Daniel Guttentag, Ulrike Gretzel, Jym Morton, and Alasdair Goodwill. 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.

Graciela Carrillo. 2019. “Predicting Airbnb Prices with Machine Learning and Location Data.” Https://Github.Com/Gracecarrillo/Predicting-Airbnb-Prices-with-Machine-Learning-and-Location-Data/Blob/Gh-Pages/Exploring_Edinburgh_Graciela_Carrillo.Ipynb. 2019.

Hamari, Juho, Mimmi Sjöklint, and Antti Ukkonen. 2016. “The Sharing Economy: Why People Participate in Collaborative Consumption.” *Journal of the Association for Information Science and Technology* 67 (9): 2047–59. https://doi.org/10.1002/asi.23552.

Kalehbasti, Pouya Rezazadeh, Liubov Nikolenko, and Hoormazd Rezaei. 2019. “Airbnb Price Prediction Using Machine Learning and Sentiment Analysis,” July. https://doi.org/10.1007/978-3-030-84060-0_11.

Lawani, Abdelaziz, Michael R. Reed, Tyler Mark, and Yuqing Zheng. 2019. “Reviews and Price on Online Platforms: Evidence from Sentiment Analysis of Airbnb Reviews in Boston.” *Regional Science and Urban Economics* 75 (March): 22–34. https://doi.org/10.1016/j.regsciurbeco.2018.11.003.

Liu, Peilu. 2021. “Airbnb Price Prediction with Sentiment Classification.” San Jose, CA, USA: San Jose State University. https://doi.org/10.31979/etd.cfxc-m67z.

Luo, Yuanhang, Xuanyu Zhou, and Yulian Zhou. 2019. “Predicting Airbnb Listing Price Across Different Cities.”

Mody, Makarand, Courtney Suess, and Tarik Dogru. 2017. “Comparing Apples and Oranges? Examining the Impacts of Airbnb on Hotel Performance in Boston.” www.bu.edu/bhr.

Tang, Emily, and Kunal Sangani. n.d. “Neighborhood and Price Prediction for San Francisco Airbnb Listings.”

Thakur, Narina, Rachna Jain, Ashee Mahajan, and Sardar M.N. Islam. 2022. “Deep Neural Network Based Data Analysis and Price Prediction Framework for Rio de Janeiro Airbnb.” In *2022 IEEE 7th International Conference for Convergence in Technology, I2CT 2022*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/I2CT54291.2022.9824383.

Wang, Dan, and Juan L. Nicolau. 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 (April): 120–31. https://doi.org/10.1016/j.ijhm.2016.12.007.

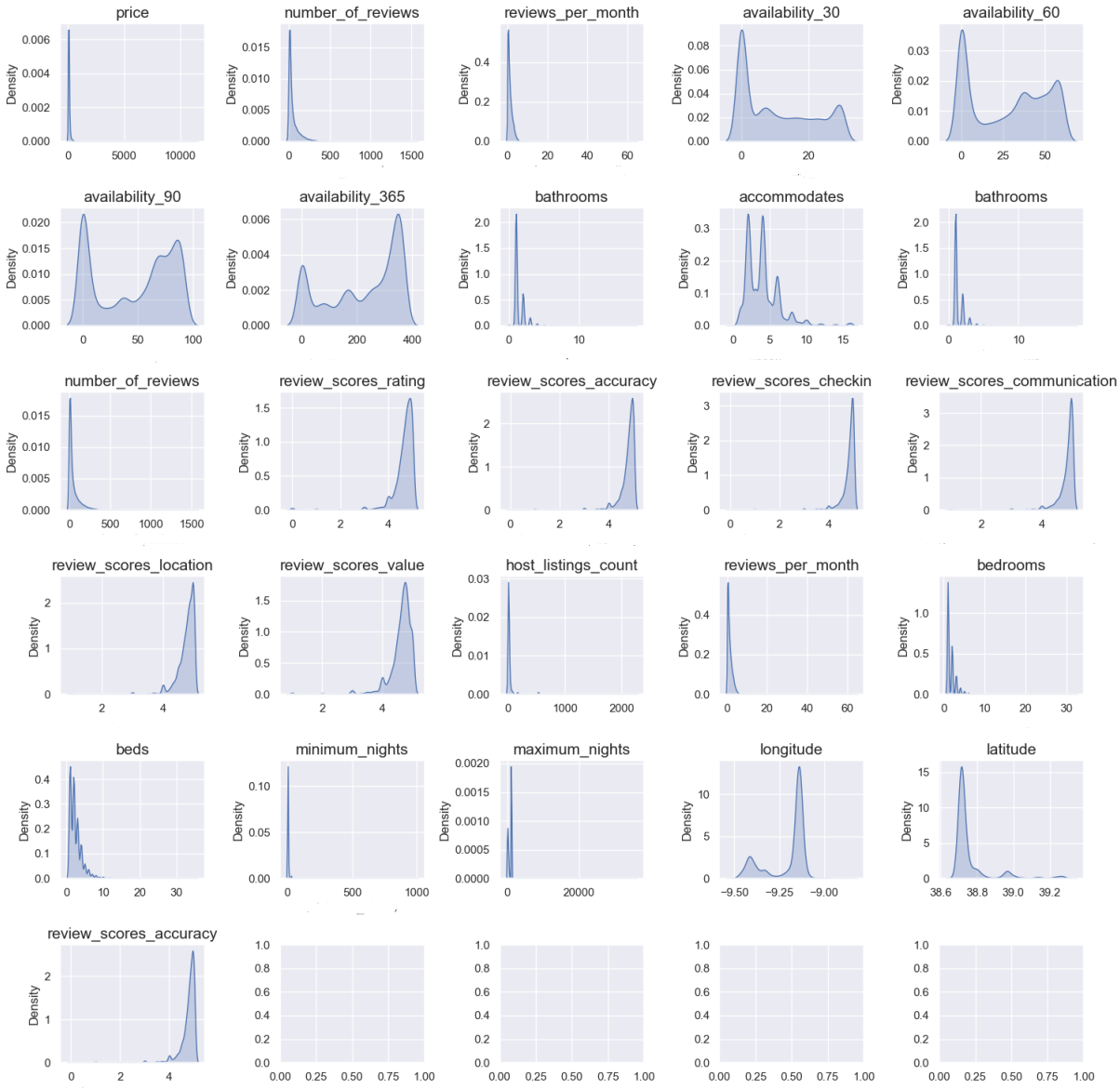
Zervas, Georgios, Davide Proserpio, and John W Byers. 2017. “The Rise of the Sharing Economy: Estimating the Impact of Airbnb on the Hotel Industry *.” https://ssrn.com/abstract=2366898.

Zhang, Zhihua, Rachel J.C. Chen, Lee D. Han, and Lu Yang. 2017. “Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach.” *Sustainability (Switzerland)* 9 (9). https://doi.org/10.3390/su9091635.

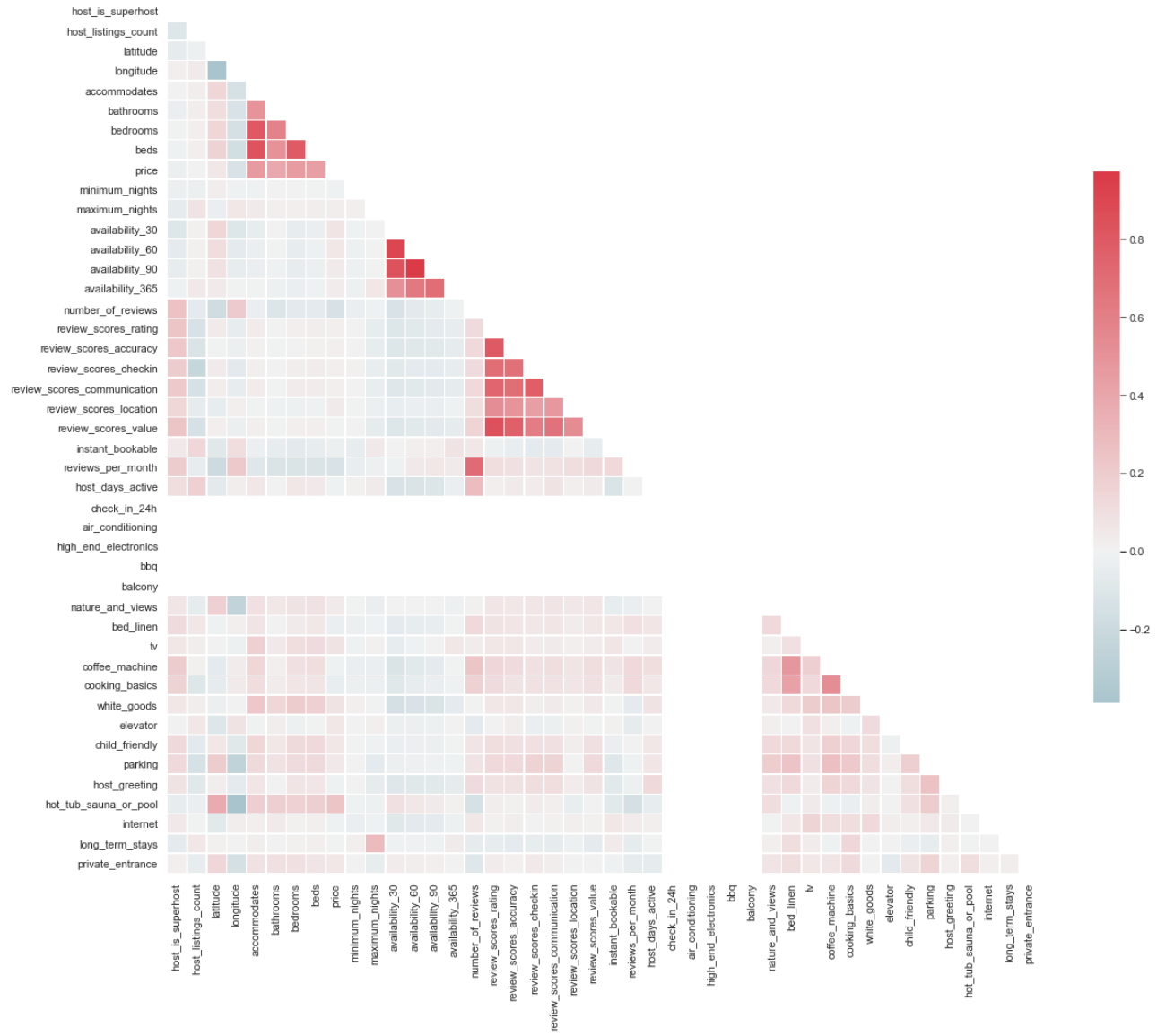
Appendix:

Variable	Description
id	unique id for each listing
host_is_superhost	whether or not the host has been verified with id
latitude	latitude of the listing
longitude	longitude of the listing
accommodates	how many people the property accommodates
price	advertised price of Airbnb listing
bathrooms	number of bathrooms
bedrooms	number of bedrooms
minimum_nights	minimum length of stay
maximum_nights	maximum length of stay
availability_365	how many nights are available to be booked in next 365 days
number_of_reviews	number of reviews left for the property
calculated_host_listings_count	how many listings host offers
check_in_24h	amenities and services
air_conditioning	amenities and services
balcony	amenities and services
tv	amenities and services
coffee_machine	amenities and services
cooking_basics	amenities and services
elevator	amenities and services
child_friendly	amenities and services
parking	amenities and services
hot_tub_sauna_or_pool	amenities and services
Internet	amenities and services
pets_allowed	amenities and services
secure	amenities and services
self_check_in	amenities and services
property_type_House	is room/house/apartment
property_type_Other	is alternative (boats, tree houses etc.)
room_type_Entire home/apt	is an entire house or aparment
room_type_Shared room	is shared room
dist_nearest_attraction	distance to the nearest attraction
dist_aeroporto	distance to the airport
atm_no	number of ATMs in the radius of 1 km
metro_no	number of metro stations in the radius of 1 km
bars_and_discos_no	number of bars and discos in the radius of 1 km
picture_url	URL to the Airbnb hosted regular sized image for the listing
host_since	the date the host/user was created. For hosts that are Airbnb guests this could be the date they registered as a guest.
neighbourhood	the neighbourhood as geocoded using the latitude and longitude against neighborhoods as defined by open or public digital shapefiles.
first_review	the date of the first/oldest review
last_review	the date of the last/newest review

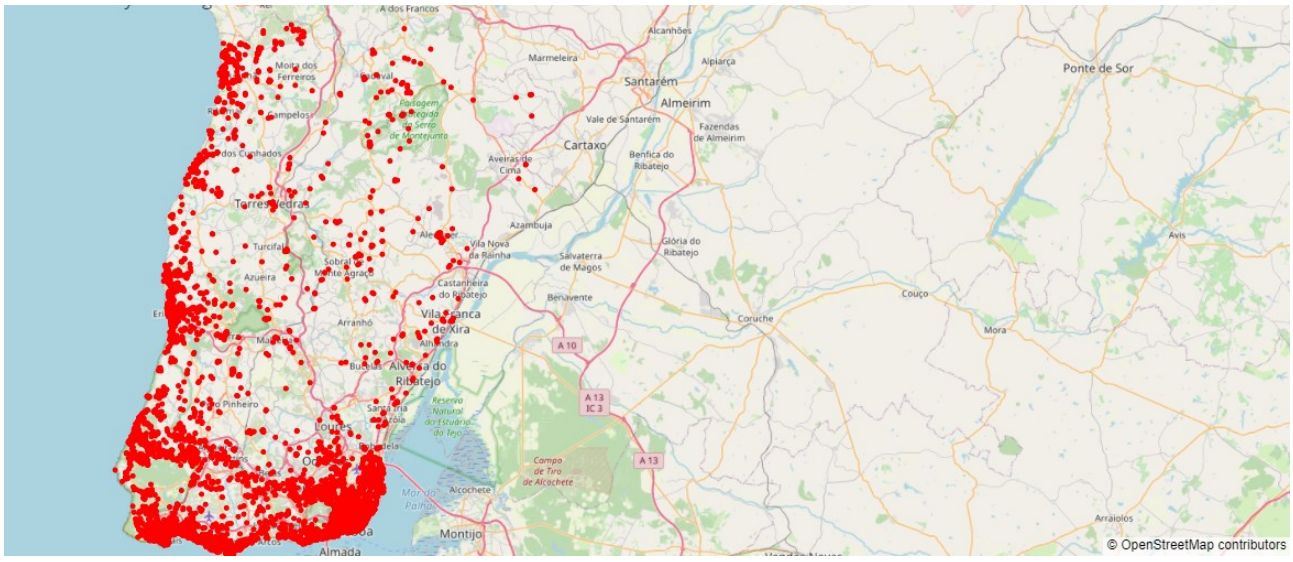
Appendix 1: Data Dictionary



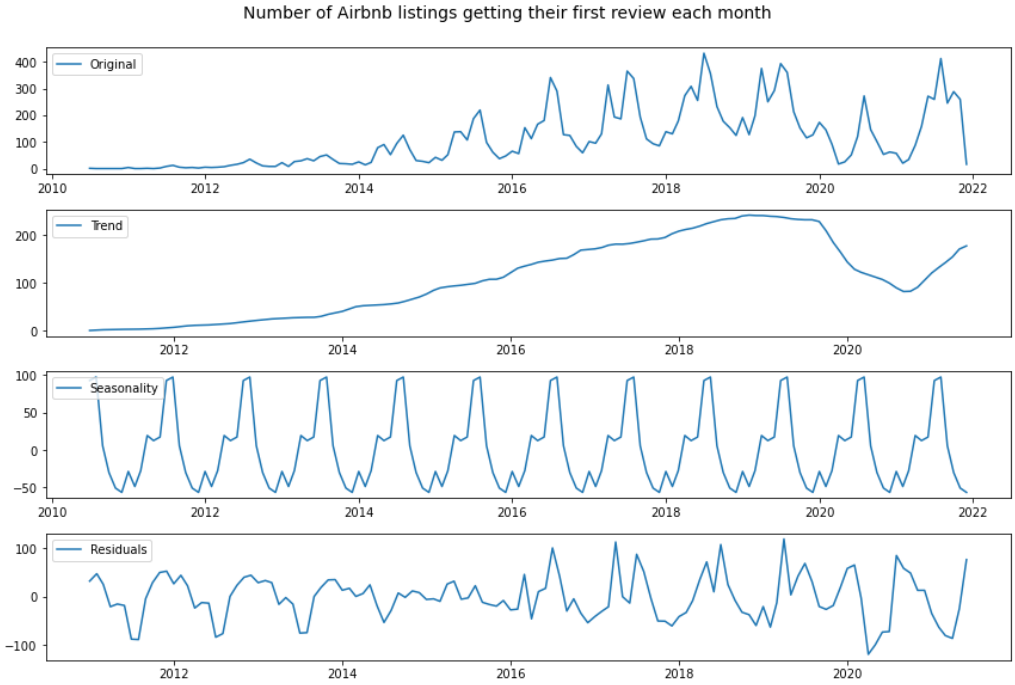
Appendix 2: Distribution of the numeric variables



Appendix 3: Correlation heat map



Appendix 4: Density of Airbnb Listings in Lisbon



Appendix 5: Price time series decomposition

Hyperparameter	Description
n_estimators	Number of gradient boosted trees
learning_rate	Scale the magnitude of parameter updates
max_depth	Maximum depth of a tree
subsample	Subsample ratio of the training instances

Appendix 6: Hyperparameters description