A Machine Learning Approach to Forecasting House Prices in the Portuguese Market

João Lourenço Vidal Esteves Ferreira

Dissertation

Master in Data Analytics

Supervised by
Professor Bruno Veloso
Professor João Gama

Acknowledgments

First and foremost, I would like to thank my parents for their constant support and encouragement throughout my academic journey. Their guidance and faith in me have been invaluable and critical to my career.

I am grateful to my supervisor, Professor Bruno Veloso, and co-supervisor, Professor João Gama, for their advice and assistance in preparing my dissertation. Their knowledge and thoughts have been priceless and have tremendously aided the success of our endeavor.

I'd also like to thank my partner, Inês, for her love, support, and encouragement during this journey. Her presence has been an ongoing source of encouragement and inspiration for me.

Finally, I'd like to thank everyone who contributed to this effort in some way, including coworkers, friends, and family. Their support and cooperation have been critical to the dissertation's success.

Abstract

The real estate market in Portugal is a complex system influenced by numerous factors, including location, property characteristics, and socioeconomic conditions. Accurate prediction of house prices is essential for various stakeholders, such as buyers, sellers, and investors, as it can significantly impact their decision-making processes. This study proposes a machine learning approach to forecast house prices in the Portuguese market using data scraped from online real estate advertisements with additional enrichment from socioeconomic statistics and ammenities detail. The research addresses the problem of predicting house prices by developing a model that considers various property attributes, such as typology, location, size, age, and condition, as well as external factors like socioeconomic indicators and geolocation attributes.

The solution involved employing multiple machine learning regression algorithms and feature enginnering techniques to analyze and model the collected data, with the final model being the XGBoost, as the most effective model for accurate price prediction using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results demonstrate that machine learning models can help understand and predict house prices in the Portuguese market, providing valuable insights into the key factors driving property values.

Table of Contents

List of Figures

List of Tables

1. Introduction

This chapter serves as the study's foundation, providing a comprehensive understanding of the research context, problem, and objectives. It starts with a discussion of Context and Motivation, emphasizing the significance of the study and the larger context in which it exists. Following that, the Problem Definition subsection describes the specific research problem that this dissertation aims to solve. The section on Objectives and Research Questions identifies the study's primary goals as well as the research questions that will guide the investigation. Finally, the Structure of the Document subsection provides an overview of the dissertation's organization, outlining how each chapter contributes to solving the research problem and achieving the study's goals.

1.1. Context and Motivation

The real estate market is a critical component of the global economy, influencing not only individual financial decisions but also broader economic trends. In Portugal, the real estate sector has experienced significant growth and volatility in recent years, driven by factors such as tourism, foreign investment, and economic fluctuations (Tavares et al., 2014). As property prices have soared, the need for accurate and timely valuation methods has become increasingly important for various stakeholders, including buyers, sellers, investors, and policymakers (Glaeser, 2013).

The advent of big data and advancements in machine learning present a significant opportunity to improve the accuracy and efficiency of property price predictions . By leveraging vast amounts of data from online real estate advertisements and incorporating sophisticated machine learning algorithms, it is possible to develop models that can better account for the multifaceted factors influencing property values (Bricongne et al., 2021). This approach not only enhances the precision of price predictions but also provides valuable insights into the underlying drivers of market trends.

The findings of this study could have substantial ramifications for real estate professionals, investors, and individuals intending to buy or sell properties in the Portuguese market. The

model could assist these stakeholders in making more informed decisions and possibly achieve better results by offering effective estimates of house prices and capturing the price definition dynamics. The findings from this study may also guide future research on the application of machine learning algorithms for forecasting home prices in different markets.

1.2. Problem Definition

Traditionally, predicting house prices relied on manual analysis of properties, economic and financial factors, and subjective human judgment (Pagourtzi et al., 2003). This could be not only very time-consuming but also prone to errors, with the final result not fully capturing the complex dynamics and nuance of the real estate market as well as the rapid change in prices that occur in a market with strong volatility and speculation (Glaeser, 2013). On the other hand, machine learning algorithms have the potential to automate the prediction process, find hidden patterns in the data, and improve the forecast's effectiveness. In fact, machine learning models may learn to recognize patterns and relationships in the data that may not be obvious to people by employing vast datasets and complex algorithms (Worzala et al., 1990).

In light of the foregoing, the purpose of this dissertation is to apply machine learning regression methods to forecast house values in the Portuguese market using information from web scraped real estate advertisements leveraging the use of more granular and real-time data available (Bricongne et al., 2021) . The main goal is to create a model that can reliably forecast home prices based on a variety of characteristics of a given property, including typology, location, size, age, and condition, taking into consideration additional insights such as socioeconomic variables (Kang et al., 2021) and geolocation attributes (Kim et al., 2020). The findings will then be assessed using the proper metrics, and the model's accuracy will be tested using a hold-out dataset.

1.3. Objectives and Research Questions

The primary goal of this dissertation is to create a machine learning model that can accurately predict house prices in the Portuguese real estate market. This goal is motivated by the desire to improve the accuracy of property valuations, which are currently hampered by the limitations of traditional methods. By using data scraped from online real estate advertisements and incorporating various property attributes and socioeconomic factors, the study hopes to create a model that not only forecasts prices more accurately but also provides insights into the key factors influencing these prices.

To achieve this primary objective, the following specific objectives are set:

1. **Data Collection and Preprocessing**: To gather a comprehensive dataset from online real estate advertisements and preprocess it to ensure the accuracy and consistency of the data.
2. **Feature Selection and Engineering**: To identify and select relevant features that significantly contribute to the accuracy of house price predictions, and to engineer new features that may enhance model performance.
3. **Model Development**: To experiment with various machine learning regression algorithms to determine the most effective model for predicting house prices in Portugal.
4. **Model Evaluation**: To rigorously evaluate the performance of the developed model using appropriate metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and to compare its performance with existing models.
5. **Insight Generation**: To analyze the model's outputs to derive insights into the key determinants of house prices in the Portuguese market, thereby contributing to the understanding of market dynamics.

Based on these objectives, the following research questions are formulated:

1. What are the most significant factors influencing house prices in the Portuguese real estate market?

2. What role do geographical factors (proximity to amenities, neighborhood crime rates, school districts) play in influencing house prices in Portugal?
3. What preprocessing and feature engineering techniques are most effective in enhancing the performance of house price prediction models?
4. Which machine learning regression algorithm offers the best performance for house price prediction in the Portuguese market?
5. Which features most significantly impact house price predictions in the Portuguese real estate market and how do these features interact to influence model predictions?

These research questions direct the investigation and ensure that the study remains focused on the key issues concerning property valuation in Portugal.

1.4. Structure of the Document

This dissertation begins with an Introduction that lays the groundwork, introducing the context, motivation, and key research questions.

Furthermore, the Literature Review extensively examines crucial areas relevant to the thesis, starting with real estate valuation and progressing into various aspects of machine learning such as regression algorithms, feature selection, and hyper-parameter optimization. It also assesses the techniques used for web scraping and the application of machine learning in predicting house prices. In the Methodology section, a detailed description of the research approach is provided, encompassing everything from business and data understanding to data preparation and modeling.

Additionally, results from these efforts are presented in the Results section, which discusses the final pipeline and hyperparameter tuning and delves into model explainability with methods like XGBoost feature importance, permutation importance, and SHAP analysis, as well as a discussion section with a reflection on the results, discussing their implications, limitations, and practical applications.

2. Literature Review

The Literature Review section is designed to provide a thorough examination of the key themes related to the integration of machine learning into real estate valuation. It begins with a look at Real Estate Valuation, which covers the fundamental approaches and economic factors that influence property assessments. Following that, the review delves into an analysis of Machine Learning, breaking down the specific methodologies used, such as regression algorithms, feature selection, feature engineering, and hyper-parameter optimization, techniques aimed at improving the precision and effectiveness of prediction models. The section also includes a focused discussion of Web Scraping, a data collection technique required for model training. The review concludes with a detailed examination of how these machine learning methodologies were specifically applied to house price prediction in previous works, demonstrating some practical potential/limitations of these technologies in the real estate market. This structured review not only elaborates on the technical and theoretical aspects, but also contextualizes their importance in the rapidly changing landscape of real estate analytics.

2.1. Real Estate Valuation

The interests, benefits, rights, and encumbrances inherent in the ownership of tangible real estate are referred to as real property, which includes land as well as any improvements or amenities permanently attached to it. Real estate valuation is the process of determining the value of a piece of real estate and is necessary to provide a quantitative measure of the benefits and liabilities associated with owning it. Property valuations are often required for various purposes, such as purchase and sale, transfer, tax assessment, expropriation, inheritance or estate settlement, investment, and financing, and may be carried out by professionals such as real estate agents, investors and fund managers, lenders, market researchers and analysts (Pagourtzi et al., 2003) .

Furthermore, valuation is the process of determining the value of a piece of property, and for a valuation to be valid, it must accurately estimate the market price of the property, which should reflect the market conditions at the time of the valuation and be based on the underlying fundamentals of the market. A valuation is frequently regarded as the best estimate of a certain property's trading price in the real estate market (Clayton et al., 2009).

Market value is a term used to describe the estimated price at which a property would sell in the open market and might be estimated by applying various valuation methods and procedures that are designed to reflect these factors, which can range from more traditional approaches to more technological approaches (Pagourtzi et al., 2003). Additionally, the past transaction price of an apartment may not be a reliable indicator of its current value after a certain period of time has passed. As a result, professional appraisers frequently use the comparable sales method to determine the current price of an apartment.(Kim et al., 2020).

In this context, the comparable method is a common technique used in real estate valuation to determine the market value of a property. The value of the subject property being appraised is compared to the selling prices of similar properties within the same market area that have recently been sold. The appraiser selects several comparable properties and adjusts the selling prices of these properties to account for differences between the subject property and the comparable, such as size, age, quality of construction, selling date, and location. The accuracy and completeness of sale transaction data is important for the success of this approach, and the process of finding comparables involves using a measure of comparability called "distance" to weight the differences between the subject property and the comparable and calculate a weighted estimate of value (Pagourtzi et al., 2003).

Furthermore, spatial econometrics and similar techniques have emerged as a way for valuation researchers to include the impact of location more accurately in pricing models (L. Krause & Bitter, 2012). Krause & Bitter argue that there are many opportunities for researchers in the field of real estate valuation to make a significant contribution by incorporating spatially explicit methods. The availability and quality of public data is increasing, and there is growing demand for valuation studies (L. Krause & Bitter, 2012). In terms of spatial attributes there are distance features which are characteristics of a property that are influenced by external factors, like its distance from public facilities like subway stations, schools, supermarkets, and intrinsic features that characterize the estate itself, including the number of households, location, and heating method and the typology of the house (Kim et al., 2020).

Finally, research shows that the real estate market was subject to chronic speculation, with rising prices being most strongly associated with optimistic expectations, and credit market conditions playing a supporting role, which reinforces not only the challenges of the heavy

speculation that affects this market but also the impact that economic cycles and market conditions can have on the property prices (Glaeser, 2013).

2.2. Machine Learning

This section aims to introduce various fundamental concepts to lay a solid theoretical foundation for the project being developed. Even though many definitions for machine learning have been published in the last decades, one of the most prevalent was published by Tom Mitchell in 1997, in which the author defines the process of ML as "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E." (Mitchell T., 1997).

Furthermore, machine learning is a rapidly growing field that aims to build computers that can improve over time through experience. It combines computer science and statistics and is at the heart of artificial intelligence and data science. The recent advancements in this field are due to both the development of new learning algorithms and the abundance of data and low-cost computation. Machine learning methods are now being used in various industries such as healthcare, manufacturing, education, finance, policing, and marketing to improve decision-making and find insights in data (Jordan & Mitchell, 2015).

The most extensively used machine learning method is supervised learning, which includes systems such as spam classifiers, face recognizers, and medical diagnosis systems (Burkov, 2019). They are all based on the function approximation problem, in which the goal is to generate a prediction in relation to information from a set of (x,y) pairings. Inputs x can be simple vectors or more complicated things like papers, photos, DNA sequences, or graphs. In a multiclass or multilabel classification issue, the result y can have one of two values or numerous labels. Predictions are made using a learned mapping f(x), which can be decision trees, support vector machines, kernel machines, neural networks, and Bayesian classifiers (Jordan & Mitchell, 2015).

Additionally, unsupervised learning is the study of unlabeled data while making assumptions about the structural features of the data. Dimension reduction approaches include principal component analysis, manifold learning, factor analysis, random projections, and autoencoders. Another example of unsupervised learning is clustering, which is the problem

of finding a partition of the data without explicit labels. A wide range of clustering procedures have been developed. In both clustering and dimension reduction, computational complexity is an important consideration, as the goal is to exploit large data sets without labeled data (Jordan & Mitchell, 2015). Finally, reinforcement learning is a third key machine learning paradigm in which the information provided in the training data is split between supervised and unsupervised learning. Unlike supervised learning, where the proper output is provided for a particular input, the training data in reinforcement learning merely shows whether an action is correct or not. If an action is erroneous, the job is to find the proper action (Jordan & Mitchell, 2015).

Another important feature that separates algorithms is the assumptions made about data distribution. Linear regression and logistic regression are two examples of models that make assumptions about the underlying probability distribution of the data and are characterized by a defined number of parameters. They are straightforward to build and analyze, but they might be sensitive to assumptions about the underlying data distribution. Non-parametric models, such as decision trees, K-nearest neighbors, Random Forest, and Support Vector Machines, make no assumptions about the underlying probability distribution of the data and are characterized by a flexible collection of parameters that can vary as the data set grows in size. They are more versatile and resilient in general, but they can be computationally more costly (Géron, 2022).

Finally, even though machine learning models can be extremely advantageous and insightful, the implementation of this algorithms can pose several challenges and issues. Overfitting is a typical problem that happens when a model is overly sophisticated and performs well on training data but badly on new or unknown data. Overfitting can be induced by a large number of features or a complex model, and it can be mitigated by applying regularization techniques or simpler models (Burkov, 2019). Underfitting, on the other hand, is the inverse of overfitting: it occurs when the model is too simplistic to understand the underlying structure of the data, thus generating erroneous predictions even on training data (Géron, 2022).

Another consideration is explainability, or the ability to comprehend and justify the model's predictions. Some models, such as decision trees and random forests, are simpler to understand than others, such as neural networks, which can be more complex.

Interpretability is more crucial than accuracy in some applications, such as healthcare, where a model's predictions must be understandable by medical specialists. In such cases, simpler models, such as linear regression, may be preferred over more complex models, such as neural networks. Furthermore, Explainability is not merely a tool for debugging models in some cases, it might be a legal requirement, such as when a system decides whether or not to grant a loan, to guarantee that there are no ethical and discriminatory issues embedded in the model (Géron, 2022).

2.2.1. **Regression Algorithms**

Because the purpose of this dissertation is to investigate and leverage the use of regression algorithms, this section will go into multiple regression algorithms that may be utilized, as well as some particular instances in which they can be employed, as well as certain strengths and limitations.

Linear regression (LR) is a frequently used approach that uses a linear equation to represent the connection between a dependent variable and one or more independent variables. It is frequently used in applications such as forecasting, time series analysis, and determining the link between variables to predict a continuous outcome variable (Mitchell, 1997).

Furthermore, Polynomial regression (PR) is an extension of linear regression that uses a polynomial equation to represent the connection between a dependent variable and one or more independent variables. It is beneficial when the connection between the variables is non-linear. It may be used to model complex relationships and is frequently utilized in non-linear time series analysis (Ostertagová, 2012).

In order to prevent the overfitting that might occur when using a LR model, the Ridge regression (RR) is a sort of regularized linear regression that prevents overfitting by adding a parameter to the cost function, hence increasing the model's stability and interpretability (Hoerl & Kennard, 1994).

Regarding nonparametric methods, the K-Nearest Neighbors (KNN) uses the "k" closest training examples to make predictions about a new test example. It is used for both classification and regression problems. KNN is simple to implement and understand, but it

can be computationally expensive and sensitive to the choice of the value of k, since it is an instance based model (Géron, 2022).

The Decision Trees algorithm (DT) is another non-parametric approach that recursively divides data into subsets depending on the value of one attribute at a time, using measures such as the Gini impurity or Entropy to define the best split. Gini impurity is an excellent default since it is somewhat faster to compute. When they vary, Gini impurity isolates the most common class in its own branch of the tree, whereas entropy produces somewhat more balanced trees (Géron, 2022). This algorithm is applied to both categorical and continuous dependent variables. It is straightforward to interpret and analyze, but if the tree is deep, it is prone to overfitting. To make judgments, an acyclic graph will be employed, and at each branching node of the graph, a specific feature j of the feature vector will be considered. If the value of the characteristic is less than a certain threshold, the left branch is taken; otherwise, the right branch is taken (Burkov, 2019).

Another powerful tree-based model is the Random Forest (RF). Instead of using just one tree, this algorithm creates multiple decision trees and averages their predictions, improving not only the stability of the model but also reducing the overfitting since it averages the predictions of many decision trees (Breiman, 2001).

Gradient Boosting (GB) is an iterative method that creates multiple decision trees and combines their predictions. It is used to improve the stability and interpretability of the model and to reduce overfitting by combining the predictions of many decision trees (Friedman, 2001). Even though there are similarities to the RF algorithm, there are many key points that differ. Regarding the building process the RF builds multiple decision trees independently, using a random subset of the features for each tree, while the GB builds the trees sequentially, where each tree is built to correct the errors made by the previous tree. Additionally, RF tends to have a lower variance and higher bias than Gradient Boosting, so RF is less likely to overfit on the one hand but on the other hand may not capture the underlying relationship as well as Gradient Boosting. Because the averaging of multiple decision trees in Random Forest reduces the effect of outliers, the GB tends to be more sensitive to outliers because it builds the trees sequentially, and outliers may have a large impact on the errors made by the previous tree. Finally, RF is computationally cheaper than Gradient Boosting since it builds multiple decision trees independently, while Gradient Boosting builds the trees sequentially and requires more computation.

The Extreme Gradient Boosting algorithm (XGBoost) is an advanced gradient boosting algorithm that is used to improve the stability and interpretability of the model. It is an efficient and scalable algorithm that can handle large datasets, also including several enhancements such as regularization, which helps to prevent overfitting, and parallel and distributed computing, which helps to speed up the training process. It is often used in applications such as computer vision, natural language processing and bioinformatics, due to the need necessity of handling a huge volume of data (Chen & He, 2017).

In the discipline of machine learning, these are some of the most often used regression methods. Each algorithm has its own set of strengths and weaknesses, and the algorithm chosen will be determined by the specific problem at hand as well as the available data, always evaluating and comparing the performance of different algorithms using the appropriate metrics to determine which algorithm is best suited for a given problem.

2.2.2. Feature Selection and Feature Engineering

There are critical processes and tools that must be employed before feeding data to the algorithms in order to acquire the best prediction results and avoid typical pitfalls when modeling data. Feature selection in this sense refers to the process of picking a subset of the most relevant characteristics from a given dataset to utilize in a model. It is a crucial phase in the modeling process since it reduces the dimensionality of the data, enhancing model interpretability and minimizing computational costs. Forward selection, backward elimination, and recursive feature elimination are all common feature selection strategies (Burkov, 2019).

Furthermore, the practice of developing new features or altering existing features to improve model performance is known as feature engineering. It entails applying domain knowledge and problem understanding to generate new features that are more informative and relevant to the problem. Polynomial expansion, interaction terms, and log transformations are all common feature engineering strategies (Géron, 2022).

Another important step is feature scaling, which might alter the algorithm's performance. Scaling is a method of standardizing features to a common range. It is possible to perform

this calculation by subtracting the mean and dividing by the standard deviation, or by normalizing the features to a range between 0 and 1 (Burkov, 2019).

2.2.3. Hyper-Parameter Optimization

Tuning hyper-parameters is a critical step in the implementation of machine learning models. Hyper-parameters are parameters that are established before training begins rather than learned during training. Parameters like as the learning rate, the number of trees, or the regularization term in a linear regression are examples of these model parameters. The optimal values of these hyper-parameters can have a considerable impact on the model's performance, and they should be defined based on the dataset and the unique situation.
There are several methods for hyper-parameter tuning, such as grid search and random search, which involve specifying a range of possible values for each hyper-parameter and training models with different combinations of hyper-parameters (Géron, 2022).

2.3. Web Scraping

Web scraping, also known as web harvesting or web extraction, refers to a method of taking data from the World Wide Web (WWW) and saving it to a file system or suitable database for subsequent study or consultation (Zhao, 2017). The process of web scraping can be divided into two steps: acquiring web resources and extracting desired information from the acquired data and can be performed through the use of bots or web crawlers. To acquire web resources, a web scraping program sends an HTTP request to a targeted website, which can be formatted as either a URL with a GET query or an HTTP message with a POST query. Once the request is received and processed by the targeted website, the requested resource is sent back to the web scraping program. The resource can be in several formats, including HTML web pages, XML or JSON data, or multimedia data such as images, audio, or video files. To extract information from the raw data, web scraping programs use modules such as Urllib2 or Selenium to handle HTTP requests and modules such as Beautiful Soup to parse and extract information from the raw HTML code (Glez-Peña et al., 2013).

Web scraping is extensively used to capture big data from the WWW since a massive amount of data is available in online resources such as web pages, newsgroup messages, and online news databases (Adomavicius & Tuzhilin, 2005). Web Scraping has also become increasingly

automated, with state-of-the-art tools capable of even simulating human browsing behavior and integrating with computer vision and natural language processing (Yi et al., 2003).

Although web scraping is a powerful technique in collecting large data sets, it has been controversial and may raise legal and ethical issues related to copyright, terms of service, and privacy. In terms of copyright, Web Scraping may involve copying and distributing content that is protected by copyright laws. In terms of terms of service, Web Scraping may involve accessing and collecting data from websites in a way that violates the terms of service of those websites. In terms of privacy, Web Scraping may involve collecting and potentially sharing personal information without the knowledge or consent of the individuals concerned (Krotov, 2018) .

Regarding the tools for performing the data extraction, there are several instruments available that can be extremely powerful implemented in several programming languages including Python, the programming language that will be used to develop this project for web scraping. In this context, Beautiful Soup is a widely used tool for web scraping designed to extract information from HTML and XML documents. It offers a range of functions for navigating, searching, and modifying parse trees, and is able to use either lxml or html5lib to extract the desired data. Additionally, Beautiful Soup can automatically detect the encoding of the document being parsed and convert it to a format that is easily readable by the client (Zhao, 2017).

Additionally, Selenium is a tool for automating web browsers allowing the developer to control a web browser and execute commands such as navigating to a website, clicking on buttons, and filling out forms, and can also be very powerful when performing Web Scraping routines, by providing efficiency on data extraction due to its web drivers compatibility and its human mimicking web automation methodology (Chaulagain et al., 2017).

2.4. Machine Learning applied to House Price Prediction

In the paper "Understanding house price appreciation using multi-source big geo-data and machine learning" by Yuhao Kang, Fan Zhang, Wenzhe Peng, Song Gao, Jinmeng Rao, Fabio Duarte, and Carlo Ratti, the authors propose a data-fusion approach to investigate

how well the potential for house price rise may be forecasted by merging numerous data sources (Kang et al., 2021). To improve their understanding of real estate appreciation and predictive modeling, they used data sets such as house structural qualities, house photographs, locational amenities, street view images, transportation accessibility, visitation patterns, and socioeconomic attributes of neighborhoods.

In practice, the authors examine the geographical dependency of house price appreciations, influential variables, and their relationships using a case study of over 20 000 residences in the Greater Boston Area. They employ a deep learning model to extract deep characteristics from street view photographs and house photos, as well as merge features from multi-source data and model house price appreciation using machine learning models, such as the Gradient Boosting Machine, in addition to the spatially weighted regression at two spatial scales: fine-scale point level and aggregated neighborhood level, using RMSE and *R*2 with k-fold cross-validations to evaluate the model performance. The results show that the proposed framework can model the house price appreciation rate with high accuracy, and the results provide insights into how multi-source big geo-data can be used in machine learning frameworks to characterize real estate price trends and help policymakers understand human settlements.

Furthermore, in the paper "House Price Prediction Using Machine Learning and Neural Networks" authored by Ayush Varma, Sagar Doshi, Abhijit Sarma, and Rohini Nair, it is presented a research project that aims to predict housing prices in Mumbai using various machine learning techniques. The authors use a dataset that contains diverse parameters such as square footage, number of bedrooms, bathrooms, flooring type, lift availability, parking availability, and furnishings condition to make their predictions. In order to achieve accurate predictions, the authors used a combination of multiple regression techniques, specifically linear regression, forest regression, and boosted regression. They also incorporate neural networks to further improve the accuracy of the predictions.

The authors used various test runs and concluded that using a combination of multiple algorithms instead of using a single algorithm yields better results. Additionally, the paper concludes that the actual real estate value also depends on nearby local amenities such as railway stations, supermarkets, schools, hospitals, temples, and parks. To address this, the

paper proposes a unique approach that utilizes the Google Maps API. By using locality search, the paper is able to narrow down on a radius of 0.5 km and take into account the proximity of these amenities in the predictions. They also suggest that this system can be used by customers to prevent the risk of investing in the wrong house, and that additional features can be added in the future to further benefit customers (Varma & Doshi, 2018).

Additionally, the paper "A Comparative Study on House Price Prediction" published in the International Journal for Modern Trends in Science and Technology, compares three machine learning algorithms for predicting house prices: Multivariable Linear Regression, Decision Tree Regression, and Random Forest Regression. The authors used a real dataset from the Indian state of Bangalore, which had 9 columns and around 13,000 rows, 8 independent variables and one target variable (price). In this paper, the authors used machine learning algorithms to predict house prices. They pre-processed the data and fed it to different regression models to determine the best model. Through the use of both K-Fold Cross-Validation and Root Mean Squared Error (RMSE) the authors were able to determine the accuracy of each model. The results showed that the Multivariable Linear Regression model was the most effective, with the highest accuracy and the lowest RMSE. The authors suggest that this model can be used as a reliable method for predicting house prices, which would benefit both buyers and sellers (Akash Dagar and Shreya Kapoor, 2020).

Moreover, in the paper "Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review", the authors conduct a systematic review of previous studies that have used machine learning (ML) techniques for property price prediction and valuation, with the aim to explore optimal solutions for predicting property price indices and identifying the best model for predicting property values based on characteristics such as location, land size, and number of rooms. In order to conduct the review, the authors searched for articles on real estate price prediction and valuation using ML techniques using electronic databases.

The findings of the review indicate that most researchers applied supervised ML techniques in their studies, with popular techniques including Support Vector Regression, Ensemble Method, Decision Tree, Linear Regression, K-Nearest Network, Random Forest, Neural Network, Support Vector Machine, Discriminant Classification, K-Nearest Neighbour and Naive Bayes. The authors found that the Random Forest algorithm was the most successful

among those used, having better compatibility with the data situation. The authors also note that the use of ML in the real estate field is increasing and suggest that their findings may be beneficial for policymakers in assessing the overall economic situation. Overall, the paper provides a comprehensive overview of the current state of research on using ML for property price prediction and valuation, highlighting the most popular techniques and the best-performing algorithms (Shahirah et al., 2021).

Another insightful work in the area is the "Review on House Price Prediction using Machine Learning" by Monika Sahu and Awantika Singh, in which the authors aim to develop a machine learning-based framework for predicting future house costs. The authors argue that predicting house prices can assist both buyers and sellers in making informed decisions and focuses on three factors that influence the cost of a house: physical condition, concept, and location. The authors use regression as their model due to its adaptable and probabilistic methodology on model selection.

The study utilizes machine learning algorithms and historical property transactions data to develop a housing price prediction model using a number of techniques to find the best among them- The results of the study indicate that the approach used in the paper is successful and produces predictions that are comparable to other house cost prediction models. Additionally, the study evaluates the performance of the model using different evaluation metrics. The results of the study prove that the proposed model yields minimum error and maximum accuracy than individual algorithm applied. That authors conclude that the proposed model can assist vendors to estimate the selling cost of a house and assist people to predict the best time to buy a house. The authors also suggest that their research has the potential to benefit property investors in understanding the trend of housing prices in a certain location (Sahu, 2021).

Moreover, Jean-Charles Bricongne, Baptiste Meunier and Sylvain Pouget (2021) discuss the importance of using publicly available data from real estate websites to create real-time indicators for the housing market, specifically focusing on the UK. The authors use web scraping to collect information on 1.5 million offers (for sale and to rent) per day, including details on price, location, area, number of rooms, description, type of offers, and type of dwelling. They argue that this data can provide a more rapid and granular understanding of

the housing market than official statistics, as it offers a seller's perspective and can be used to track changes in prices and new offers over time. The authors used this data to track the UK housing market during the Covid-19 crisis and found that there was a 80% decline in the number of new offers during the first lockdown, with sellers refraining from adjusting prices. They also find large discrepancies in price changes and buyers"negotiation margins across different regions, with London being the most affected by the virus and having the lowest buyers"negotiation margin (Bricongne et al., 2021).

In the paper "Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times" the authors developed a model to predict house prices in a Spanish city, and to quantify the impact of the COVID-19 pandemic on house prices. The paper uses machine learning algorithms to analyze and identify patterns in large structured and georeferenced datasets, collected from the internet using web scraping. The methodology covers data preparation, feature engineering, hyperparameter tuning and optimization, model evaluation and selection, and model interpretation with several algorithms being explored, ranging from Ensemble learning algorithms such as boosting (Extreme Gradient Boosting, Gradient Boosting Regressor and Light Gradient Boosting Machine) and bagging (random forest and extra-trees regressor). The model was created by combining georeferenced data with information from other sources such as socioeconomic indicators and satellite photos. The findings revealed that machine learning algorithms outperform classic linear models because they are better adaptable to the nonlinearities of complex data, such as real estate market data. The researchers also concluded that boosting algorithms outperformed bagging techniques in terms of performance and overfitting.(Mora-Garcia et al., 2022).

In the paper "Big Data in Real Estate? From Manual Appraisal to Automated Valuation", the authors discuss the challenges in determining the value of commercial real estate and the problems that arise as a result, arguing that appraisals, which are typically based on the capitalization of the net income of an asset using a yield inferred from comparable transactions, are frequently incorrect since comparable property deals are difficult to compare in terms of time or building attributes. Appraisals also tend to lag the market and produce smoothed approximations of current market prices, which can be unduly low in bull markets and high in negative markets. This can present issues for investors, lenders, and banks who do not have access to correct pricing data to underwrite assets. The study used

84,305 observations from California, Florida, and Texas between 2011 and 2016 to assess the performance of several machine learning algorithms. Ordinary least squares regression (OLS), random forest (RF), gradient boosting regression (GBR), and extreme gradient boosting machine are among the techniques compared (XGBM). Overall, the results showed that XGBM performed the best. The study used 84,305 observations from California, Florida, and Texas between 2011 and 2016 to assess the performance of several machine learning algorithms. Ordinary least squares regression (OLS), gradient boosting regression (GBR), random forest (RF) and extreme gradient boosting machine are among the techniques compared (XGBM). Overall, the results showed that XGBM performed the best.(Kok et al., 2017).

The following table summarizes the papers the main key points mentioned in this chapter:

Author	Paper	Task	Models	Dataset	Accuracy Measures
Kang et al. (2021)	*"Understanding house price appreciation using multi-source big geo-data and machine learning"*	Investigating the potential for house price rise forecasted by merging data sources	Gradient Boosting Machine, Spatially Weighted Regression	House structural qualities, house photographs, locational amenities, street view images, transportation accessibility, visitation patterns, and socioeconomic attributes of neighborhoods in the Greater Boston Area	RMSE, R2 with k-fold cross-validation
Ayush Varma, Sagar Doshi, Abhijit Sarma, and Rohini Nair (2018)	*"House Price Prediction Using Machine Learning and Neural Networks"*	Predicting housing prices in Mumbai	Multiple regression (linear regression, forest regression, and boosted regression) and Neural Networks	Parameters such as square footage, number of bedrooms, bathrooms, flooring type, lift availability, parking availability, and furnishings condition.Locality search with a radius of 0.5 km.	Accuracy
Akash Dagar and Shreya Kapoor (2020)	*"A Comparative Study on House Price Prediction"*	House Price Prediction	Multivariable Linear Regression, Decision Tree Regression, Random Forest Regression	Real dataset from Bangalore, India (9 columns, 13,000 rows, 8 independent variables and 1 target variable (price))	K-Fold Cross-Validation, Root Mean Squared Error (RMSE)
Shahirah et al. (2021)	*"Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review"*	Property Price Prediction and Valuation	Support Vector Regression, Ensemble Method, Decision Tree, Linear Regression, K-Nearest Network, Random Forest, Neural Network, Support Vector Machine, Discriminant Classification, K-Nearest Neighbour, Naive Bayes	N/A (review)	N/A (review)
Monika Sahu, Awantika Singh (2021)	*"Review on House Price Prediction using Machine Learning"*	Predicting future house costs	Linear Regression	The dataset contains 7 columns and 5000 rows of houses in New York city.	Minimum error, Maximum accuracy

Jean-Charles Bricongne, Baptiste Meunier and Sylvain Pouget (2021)	*"Web Scraping Housing Prices in Real-time: the Covid-19 Crisis in the UK"*	Track changes in the UK housing market during the Covid-19 crisis	K-Nearest-Neighbours	1.5 million offers per day including details on price, location, area, number of rooms, description, type of offers, and type of dwelling	N/A
Mora-Garcia et al. (2022)	*"Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times"*	Predict house prices in a Spanish city, quantify impact of COVID-19	Ensemble learning (Extreme Gradient Boosting, Gradient Boosting Regressor, Light Gradient Boosting Machine), Bagging (Random Forest, Extra-trees Regressor)	Structured and georeferenced datasets collected from the internet using web scraping combined with socioeconomic indicators and satellite photos	R2, Mean absolute error (MAE). Mean squared error (MSE), Root Mean squared error (RMSE)
Kok et al. (2017)	*"Big Data in Real Estate? From Manual Appraisal to Automated Valuation"*	Determine value of commercial real estate	OLS, Random Forest, Gradient Boosting, XGBoost	84,305 observations from California, Florida, and Texas (2011-2016)	R2, Mean squared error (MSE), Mean Absolute Percentage Error (MAPE)

Table 1- Previous Works Comparison

2.5. Literature Review Summary

This section explored the main topics of the dissertation from real estate valuation to machine learning techniques. Starting with a foundational understanding of real estate valuation, examining the key methodologies and economic principles that govern property assessments, followed by a discussion about the integration of machine learning, dissecting how these technologies refine predictive models through regression algorithms, feature selection, feature engineering, and hyper-parameter optimization.

A focused examination of web scraping revealed its critical role in data acquisition, which underpins the effectiveness of machine learning applications in real estate, culminating in an overview and review of the application of these methodologies to house price prediction in previous works, demonstrating their practical implications and the significant benefits they can bring to the real estate market. Finally, a summary table was presented summarizing the main key points such as the main tasks, algorithms, and accuracy measures used in the papers mentioned, which are extremely important in guiding this dissertation.

3. Methodology

This dissertation employs the CRISP-DM methodology to guide the research and analysis processes, ensuring a systematic approach to developing a predictive model for real estate valuation using machine learning techniques. This implementation offers a thorough, repeatable approach to data mining, with clear paths from data collection to real-world application, not only improving research rigor but also offering a framework for future studies to build on.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a strong, widely used methodology for carrying out data mining projects. It provides a structured way to plan and execute projects that entail extracting information and insights from data. This approach is broken down into six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase focuses on a specific set of actions required to advance through the project lifecycle, from initial conceptualization to solution deployment. (Wirth & Hipp, 2000).

The following subsection provide an overview of how each relevant CRISP-DM phase is applied in the context of this research:

3.1. Business Understanding

In the Business Understanding section its important to important an overview of the Portuguese real estate market, its significance, key trends, and factors influencing valuation. The Portuguese real estate market has been a significant component of the national economy, serving as a key indicator of economic performance. The market's importance was reinforced by the global economic turndown between 2007 and 2009, which highlighted the real estate sector's integral role in both the economy and monetary policy. Housing markets are especially relevant, not only as a large share of household expenditure but also as a source of collateral for bank loans. In Portugal, home ownership is high, with about 80% of households owning property (Tavares et al., 2014).

The valuation of real estate in Portugal has traditionally been conducted using the comparative method, which is the most commonly used approach. Other methods, such as

the income method, are applied in certain contexts, especially when estimating the market value of investment properties. Property valuation in Portugal is highly influenced by factors such as Location (with regions like Lisbon, Porto and Algarve having higher property values to demand from both local and foreign buyers), property typology and economic confidence (Tavares et al., 2014).

3.2. Data Understanding

The Data Understanding section of this dissertation delves into the various components of the dataset that are critical for analyzing real estate market trends. It begins by explaining how data is extracted from various sources, including real estate advertisements detailing property features and pricing, geolocation data providing spatial contexts, and municipal socioeconomic information. This is followed by a thorough examination of the datasets to determine statistics, underlying trends, and relationships, as well as assessments of missing values, correlations, and outliers in the data.

3.2.1. Data Extraction

One of the most crucial parts and problems of this dissertation is that all of the data was collected from the internet while taking into account the prior works discussed in the literature review. This posed a significant risk and importance for this stage of the project, as failure to extract the essential data would prevent progress to the next stages. All of the details and approaches will be explained in this section, but in summary, Python was used to web scrape all of the necessary data, which was then stored in a SQL Server database using the Pyodbc library to connect the Jupyter notebook to the local SQL Server database.

3.2.1.1. Real Estate Advertisement Data

The first challenge of the project was to extract from the web the main data regarding real estate advertisements, not only getting the most recent data available at the time but also capturing all the attributes possible to enrich the model. With this in mind, one of the most popular websites for selling properties in Portugal was chosen, and the next challenge would be to web scrape not only all the houses being sold at the moment on the platform but also

be able to map all the non-structured attributes (text) and extract them to a standardized structured format.

To start this web scraping challenge, the Python libraries Beautiful Soup and Requests were used. Even though the terms and conditions of this specific website state that web scraping is only prohibited for commercial purposes, when performing multiple requests from the same IP address, even with a generous and random waiting period between requests, the website would block the requests, which turned out to be an additional and unexpected challenge.

After exploring several options, such as other web scrapping tools and proxy rotation, the final approach was to use Selenium and the Undetected Chromedriver, which is an optimized version of the standard ChromeDriver designed to bypass the detection mechanisms of most anti-bot solutions like DataDome, Perimeterx, and Cloudflare, in addition to random waiting periods between requests.

After solving this challenge, the next task was to comprehend the structure of the website and develop a script to navigate through all the advertisements published at the moment. One important aspect of the website is that it attributes a given ID to each advertisement, so the challenge would be to first get all the ID's and then perform the requests for each advertisement ID. To do so, the first attempt was to use the main search page to get the ID of each advertisement listed; however, it would only show up to 60 pages with 30 ads each, which would only lead to a total extraction of 1800 ads. To address this limitation, the first step of the script is to perform a search in segments of prices for properties within a given range of prices, increasing that range by 10000€ each time. So, for instance, the script will start by using URL filters to search for properties with prices ranging from 50000€ to 60000€ and so on. Then, for each of these intervals, we will select the total number of properties for sale, which is presented as a header in the HTML page. Furthermore, we will divide that number of properties by 30 and round up in order to calculate the total number of pages of ad listings that will need to be extracted. For instance, if there are 430 properties for sale in that range, then the script will need to navigate from 1 to 15 pages and get each one of the IDs being advertised. All these URLs were then stored in the SQL Server table [AdsPagination] that was used in the next step of the script:

RowNum	URL	MinPrice	MaxPrice	PageNumber	CreationDate	WasScraped
1	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	10000	19999	1	2023-03-18 22:27:33.130	1
2	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-2?ordem=precos-asc	10000	19999	2	2023-03-18 22:27:33.133	1
3	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-3?ordem=precos-asc	10000	19999	3	2023-03-18 22:27:33.133	1
4	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-4?ordem=precos-asc	10000	19999	4	2023-03-18 22:27:33.133	1
5	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-1?ordem=precos-asc	20000	29999	1	2023-03-18 22:27:33.133	1
6	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-2?ordem=precos-asc	20000	29999	2	2023-03-18 22:27:33.133	1
7	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-3?ordem=precos-asc	20000	29999	3	2023-03-18 22:27:33.133	1
8	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-4?ordem=precos-asc	20000	29999	4	2023-03-18 22:27:33.133	1
9	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-5?ordem=precos-asc	20000	29999	5	2023-03-18 22:27:33.133	1
10	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-6?ordem=precos-asc	20000	29999	6	2023-03-18 22:27:33.133	1
11	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_29999,preco-min_20000/portugal/pagina-7?ordem=precos-asc	20000	29999	7	2023-03-18 22:27:33.133	1
12	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_39999,preco-min_30000/portugal/pagina-1?ordem=precos-asc	30000	39999	1	2023-03-18 22:27:33.137	1
13	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_39999,preco-min_30000/portugal/pagina-2?ordem=precos-asc	30000	39999	2	2023-03-18 22:27:33.137	1

Figure 1- Pagination Table

The next step was for each of these URLs to extract the IDs of all the advertisements listed and also store them in another SQL Server table *[AdsToExtract]* in order to keep track of all the information necessary and structure it:

RowNum	RealEstateAdKey	URL	WasScraped	CreationDate
1	32039492	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
2	30935355	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
3	30463020	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
4	32499544	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
5	32329256	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
6	32504683	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
7	32386401	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
8	28706822	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.313
9	31524574	https://www.platform.pt/pesquisa/comprar-casas/com-preco-max_19999,preco-min_10000/portugal/pagina-1?ordem=precos-asc	1	2023-03-18 22:40:50.317

Figure 2- Advertisements to Extract

As can be seen above, for each of the URL of the previous table, in this table we have all the RealEstateAdKey's that were extracted for each one of them, and at the end, 26837 IDs were extracted.

Finally, the next challenge was to parse the HTML content of each advertisement, identify all the possible attributes, to extract and store them in a structured and standardized manner. After developing the necessary script and iterating over each of the IDs already identified the following dataset was obtained and data was stored in the table *[RealEstateAds]*:

Column	Description	Data Type
RowNum	Row number (an identifier for each row in the dataset).	Integer
RealEstateAdKey	Key/identifier for the real estate advertisement.	Integer
URL	The URL link to the real estate advertisement.	String
Title	The title of the real estate property in the advertisement.	String
Price	The price of the property listed in the advertisement.	Decimal
Location	The location or address of the property.	String

Area	The total area of the property in square meters (m²).	String
Rooms	The number of rooms in the property (e.g., T7 - 7 bedrooms).	Decimal
Floors	The number of floors in the property (if available).	Integer
Energy	Energy efficiency rating of the property (if available).	String
BuiltIn	The year the property was built (if available).	Integer
FullArea	The full area of the property (if available).	Decimal
Orientation	The orientation of the property (e.g., North, South, East, West) if available.	String
Parking	Information about parking availability, e.g., "Parking space included in the price."	Binary
FittedWardrobes	Indicates if the property has fitted wardrobes (if available).	Binary
Terraced	Indicates if the property is terraced (if available).	Binary
Bathrooms	The number of bathrooms in the property.	Integer
Balcony	Indicates if the property has a balcony (if available).	Binary
Street	The street name where the property is located.	String
Parish	The parish or neighborhood where the property is located.	String
County	The county or administrative region where the property is located.	String
District	The district or larger region where the property is located (if available).	String
Latitude	The latitude coordinates of the property's location.	Decimal
Longitude	The longitude coordinates of the property's location.	Decimal
Others	Additional information about the property or its features	String
Creationdate	The date and time when the advertisement was extracted.	Datetime

Table 2- Real Estate Ads Dataset Description

3.2.1.2. Geolocation Data

As previously emphasized in the literature review, the amenities of a given property play an important role in terms of its value. With this in mind, the next challenge was to be able to

determine the main points of interest around each one of the properties in the extracted dataset.

To do so, and in order to extract the most up to date data available at the time, a Python script was developed taking advantage of the Google Maps API, which provides an endpoint that allows for nearby searches around a given pair of coordinates, however, another challenge came up. Since the coordinates are an option field to fill in the advertisement's platform, only approximately 21% of the properties have the coordinates filled by the advertiser. In order to overcome this issue, another approach was used to get an approximate pair of coordinates for each of the houses listed for sale that were missing coordinates. Google Maps API also has an endpoint that will output the coordinates of a given address. Taking advantage of this endpoint it was possible to leverage the geographical attributes extracted previously and get the approximate coordinates for the respective houses. Unfortunately, since the address is also an optional field to fill, many houses have no specific address associated with them, only the municipality for instance. Nevertheless, an approximate pair of coordinates was obtained for these cases and a new attribute was created to catalog the origin of the coordinates:

GeolocationType	**Count**
Municipal Coordinates	16269 (61%)
Original Coordinates	5548 (21%)
Address Coordinates	5020 (19%)

Table 3- Distribution of the Coordinates Type

Finally, after obtaining coordinates for all the properties, another script was developed to search for points of interest within a radius of 1 km. Regarding the points of interest, taking into consideration previous works, the following ones were considered:

- Schools
- Supermarkets
- Bus Stops
- Subway Stations
- Hospitals
- Pharmacies

- Restaurants
- Stores

By iterating over each one of the points of interest for each distinct pair of coordinates in the dataset the script performed over 50 000 requests to the Google Maps API and stored all the data using pyodbc in the SQL Server Table *[GeoLocationDetail]* as can be seen below:

RealEstateAdKey	GeolocationType	Number_Schools	Number_Supermarkets	Number_BusStops	Number_SubwayStations	Number_Hospitals	Number_Pharmacies	Number_Restaurants	Number_Stores
31515017	Coordenadas Gerais	0	0	9	0	0	0	1	1
31515860	Coordenadas pela morada	20	20	2	0	0	15	20	20
31516922	Coordenadas Gerais	0	0	0	0	0	0	1	0
31517971	Coordenadas pela morada	3	4	0	0	0	3	5	4
31519010	Coordenadas Gerais	10	20	20	0	1	4	20	17
31519121	Coordenadas Gerais	3	6	0	0	0	1	3	12
31519916	Coordenadas Originais	0	0	0	0	0	0	1	0
31520082	Coordenadas Gerais	0	1	0	0	0	0	0	5
31520380	Coordenadas pela morada	6	10	0	0	0	3	18	20
31520740	Coordenadas Gerais	2	2	15	0	0	1	10	3

Table 4- Geolocation Detail

3.2.1.3. Municipal Socioeconomic Data

Finally, as explored in the literature review, the socioeconomic data from the municipal area also brings insights into the value of real estate. In order to complement the dataset with this crucial data, the PorData website appeared as the best source of broad social and economic information. To extract the data from the website, another approach was also tested by using automation logic to replicate human behavior interacting with the website.

Instead of scraping and interpreting the HTML that was changing when selecting different districts, the option was used to export to Excel the table displayed on the website. Since there was no option to download all the districts in one Excel file, it would be necessary to select each district, download individual Excel files, and then merge them into a single dataset. The challenge was precisely to develop a script to do it in an automated way, effectively minimizing the manual effort and potential errors, thereby ensuring the integrity and efficiency of data processing for the dissertation research.

The code is segmented into several specialized functions, each dedicated to executing specific tasks that cumulatively enhance the efficiency and organization of the data handling process.

The first part of the code consists of function definitions. The get_mouse_location function is designed to continuously monitor and display the current mouse position, a vital feature for identifying coordinates for automated clicking in subsequent steps. Following this, the move_and_click function takes charge of automating mouse movements to specified coordinates and executing a click action. This function is pivotal in simulating user interactions during the automated data extraction process. Additionally, the extract_excel_pordata function automates the extraction of data from the Pordata website. It adeptly manages a series of automated mouse movements and clicks to download Excel files and then navigates through the website to select and download data for different districts. Lastly, the rename_excel_files function is employed. After downloading, this function streamlines the process by renaming the Excel files, removing unnecessary parts from their original names, and thereby standardizing the file names for smoother processing.

The subsequent segment of the script focuses on creating a data frame. The script initializes an empty DataFrame and then loops through each Excel file in the specified directory, ensuring that only files in Excel formats are included. For each file, specific rows and columns are selected, and the columns are renamed for clarity, creating a structured and understandable dataset.

Further, the script applies several transformations to the data. These include cleaning and standardizing the "metric column by removing unwanted characters and patterns, converting the 'valuecolumn to a float data type for numerical analysis, and aggregating the data. Following this, the DataFrame undergoes a pivotal restructuring using the pivot method, enhancing the representation of the data. The columns are also thoughtfully renamed to enhance readability and understanding of the dataset.

After developing the necessary script and iterating over each of the IDs already identified, the following dataset was obtained, and the data was stored in a SQL Server table. The final dataset is extremely rich and has several columns regarding crucial social and economic aspects, as can be seen in the list of the columns below grouped by categories:

1. **Housing and Accommodation:**
    a. *Alojamentos familiares clássicos* (Classic Family Dwellings)
    b. *Alojamentos turísticos* (Tourist Accommodations)
    c. *Edifícios novos concluídos para habitação familiar* (New Buildings Completed for Family Housing)
2. **Educational Facilities:**
    a. *Alunos do ensino não superior* (Non-Higher Education Students)
    b. *Alunos do ensino superior* (Higher Education Students)
    c. *Estabelecimentos do 1º, 2º, 3º ciclo do ensino básico e ensino pré-escolar* (Educational Establishments for 1st, 2nd, 3rd Cycles of Basic Education and Pre-school)
    d. *Estabelecimentos do ensino secundário e superior* (Secondary and Higher Education Establishments)
3. **Financial Infrastructure:**
    a. *Bancos Caixas Económicas* (Banks and Savings Banks)
    b. *Caixas automáticas multibanco* (ATMs)
    c. *Caixas de Crédito Agrícola Mútuo* (Agricultural Credit Union Boxes)
4. **Social and Cultural Indicators:**
    a. *Beneficiários do Rendimento Social de Inserção* (RSI) (Beneficiaries of Social Insertion Income)
    b. *Museus* (Museums)
    c. *Ecrãs de cinema* (Cinema Screens)
    d. *Sessões de espectáculos ao vivo* (Live Show Sessions)
    e. *Casamentos* (Marriages)
    f. *Divórcios* (Divorces)
5. **Demographics and Population Data:**
    a. *Densidade populacional* (Population Density)
    b. *Idosos* (%) (Elderly Population Percentage)
    c. *Jovens* (%) (Youth Population Percentage)
    d. *População em idade activa* (%) (Active Age Population Percentage)
    e. *População estrangeira e em % da população residente* (Foreign Population and Percentage of Resident Population)
    f. *População residente* (Resident Population)
    g. *Índice de envelhecimento* (Aging Index)

h. *Saldo natural* (Natural Balance)

6. **Economic and Employment Indicators:**
    a. *Desempregados inscritos nos centros de emprego e em % da população residente* (Unemployed Registered at Employment Centers and as a Percentage of the Resident Population)
    b. *Ganho médio mensal dos trabalhadores por conta de outrem* (Average Monthly Earnings of Employees)
    c. *Trabalhadores da Administração Pública Local* (Local Public Administration Workers)
    d. *Volume de negócios das quatro maiores empresas do município* (%) (Turnover of the Four Largest Companies in the Municipality)
7. **Municipal and Environmental Data:**
    a. *Despesas da Câmara Municipal e em cultura, desporto, ambiente* (%) (Municipal Expenses in Culture, Sports, Environment)
    b. *Saldo financeiro da Câmara Municipal* (Municipal Financial Balance)
    c. *Transferências recebidas no total das receitas da Câmara Municipal* (%) (Transfers Received in the Total Revenue of the Municipal)
    d. *Resíduos urbanos recolhidos selectivamente por habitante* (kg) (Urban Waste Collected Selectively per Inhabitant)
    e. *Consumo de energia eléctrica por habitante* (kWh) (Electricity Consumption per Inhabitant)
    f. *Superfície em km2* (Area in km²)
    g. *Taxa de mortalidade infantil* (Infant Mortality Rate)

In the final stages of data processing, unnecessary columns are removed from the data frame, ensuring that only the relevant data remains, eliminating any potential clutter that might detract from the analysis. The culmination of this extensive data preparation process is marked by the export of the processed and transformed data to a CSV file. This file is created using a specific encoding and delimiter, ensuring that the data is ready for further analysis or presentation, thus marking the successful completion of a vital phase in the dissertation research.

3.2.2. Data Exploration

After successfully integrating the entire dataset, which includes data taken from numerous sources, it is critical to undertake a thorough analysis of the data. This exploration is critical not only for realizing the dataset's full potential but also for choosing the most effective techniques for future stages. These stages include the use of preprocessing methods and the implementation of machine learning algorithms for predictive analytics. This careful method guarantees that the data is used to its greatest potential, allowing for accurate and insightful predictions.

3.2.2.1. Global Data Analysis

The dataset used for this study consists of 26,837 observations and 79 variables. As previously stated, the dataset includes information from three different sources and is composed of variables including information about the real estate advertisements, such as the price, area, number of rooms, and number of floors; amenities detail; as well as demographic and socioeconomic variables for the respective municipalities where the properties are located.

Before delving into the analysis of the main statistics of the dataset, some cleaning was performed on the data, and several transformations were done in order to clean the dataset from:

- For the *BuiltIn* column, the script replaces the string *"Built in"* with an empty string.
- For the *Bathrooms* column, the script replaces the strings *"bathrooms"* and *"bathroom"* with an empty string. It also replaces the string "No" with "0".
- For the Orientation column, the script replaces the string "Orientation" with an empty string.
- For the Rooms column, the script replaces the characters "T"and ","with an empty string.
- For the Floors column, the script replaces the strings *"floors"*and *"floor"*with an empty string.

- For the Energy column, the script first removes any "+"or "-"signs using regex, then converts all letters to lowercase, and finally maps each letter to a corresponding number using a dictionary.
- For certain columns (*Parking, FittedWardrobes, Terraced, and Balcony*), the script converts non-null values to 1 and null values to 0.
- For the Orientation column, the script creates four new columns (*Orientation_North, Orientation_South, Orientation_West, and Orientation_East*) and sets the value to 1 if the corresponding orientation is present in the Orientation column, otherwise sets the value to 0. Finally, the script drops the Orientation column from the dataset.

After these transformations, the dataset will have undergone multiple changes and be ready for additional analysis or modeling. Finally, to arrange the data exploration section, the dataset will be examined independently for each data source.

3.2.2.2. Advertisements Data Analysis

Feature	count	mean	std	min	25%	50%	75%	max
Price	26837.0	544416.4	7214037.0	6100.0	190000.0	330000.0	590000.0	995000.0
Area	26428.0	191.4	144.1	11.0	96.0	146.0	240.0	999.0
Energy	21584.0	4.0	1.6	0.0	3.0	4.0	6.0	6.0
Parking	26837.0	0.4	0.5	0.0	0.0	0.0	1.0	1.0
FittedWardrobes	26837.0	0.3	0.5	0.0	0.0	0.0	1.0	1.0
Terraced	26837.0	0.2	0.4	0.0	0.0	0.0	0.0	1.0
Balcony	26837.0	0.2	0.4	0.0	0.0	0.0	0.0	1.0
Orient_North	26837.0	0.1	0.2	0.0	0.0	0.0	0.0	1.0
Orient_South	26837.0	0.1	0.3	0.0	0.0	0.0	0.0	1.0
Orient_West	26837.0	0.1	0.3	0.0	0.0	0.0	0.0	1.0
Orient_East	26837.0	0.1	0.3	0.0	0.0	0.0	0.0	1.0

Table 5- Advertisement Summary Statistics

Regarding the data originally obtained from the advertisement web site, the most important column is the Price which is the target attribute of the dataset. Regarding the Price, the mean value of the column is 544,416.4 and the standard deviation is 721,437.0. The minimum price is 6,100 and the maximum price is 9,950,000. The 25th percentile of the data is 190,000, the median is 330,000, and the 75th percentile is 590,000, so it can be concluded that the data is highly positively skewed, with most of the observations falling in the lower range of prices. Additionally, it is possible to identify outliers at the higher end of the data range, which indicates the presence of a few high-priced luxurious houses in the dataset.

Another very important column in the dataset is the Area. The mean of the Area column is 191.4 and the standard deviation is 144.1. The minimum area is 11 and the maximum area is 999. The 25th percentile of the data is 96, the median is 146, and the 75th percentile is 240. From this analysis, it can be concluded that the data is moderately skewed, with most of the observations falling in the mid-range of areas in contrast to what was previously inferred in the Price column.

Furthermore, the typology of the houses is also a very important characteristic. The mean of the Rooms column is 3.7 and the standard deviation is 19.3. The minimum number of rooms is 0 and the maximum number of rooms is 1750. The 25th percentile of the data is 2, the median is 3, and the 75th percentile is 4. Based on this analysis, we can conclude that the data is highly skewed, with most of the observations falling in the lower range of room numbers and the presence of outliers.

The *"BuiltIn"* column has 13,267 non-null values, with a mean of 1994.2 and a standard deviation of 29.2. The earliest year that a house was built is 1700 and the latest year is 2025, which are houses that are still being constructed. The 25th percentile of years is 1982, the median is 2001, and the 75th percentile is 2020. The data shows a relatively small variation in years of construction, with most houses being built in the late 20th or early 21st century.

The *"Energy"* column has 21,584 non-null values with a mean of 4.0 (C rating) and a standard deviation of 1.6. The minimum value in the column is 0.0, the maximum value is 6.0, and the quartile values are 3.0 (D rating) for the 25th percentile, 4.0 (C rating) for the 50th percentile, and 6.0 (A rating) for the 75th percentile.

Analyzing the boolean columns, the *"Parking"* variable has a mean of 0.4 and a standard deviation of 0.5. This indicates that 40% of the observations have a value of 1 (True) for this variable, meaning that the properties have parking available, while 60% of the observations have a value of 0 (False). The standard deviation of 0.5 suggests that the values for this variable are quite spread out from the mean.

The *"FittedWardrobes"* variable has a mean of 0.3 and a standard deviation of 0.5. Even though it might not that intuitive to analyze the mean of a binary variable, the mean in this case is representative of proportion, in this case indicating that 30% of the observations have a value of 1 (True) for this variable, meaning that the properties have fitted wardrobes, while 70% of the observations have a value of 0 (False). The standard deviation of 0.5 suggests that the values for this variable are quite spread out from the mean.

The "*Terraced"* variable has a mean of 0.2 and a standard deviation of 0.4. This indicates that 20% of the observations have a value of 1 (True) for this variable, meaning that the properties are terraced, while 80% of the observations have a value of 0 (False). The standard deviation of 0.4 suggests that the values for this variable are also spread out from the mean.

The *"Balcony"* variable has a mean of 0.2 and a standard deviation of 0.4. This indicates that 20% of the observations have a value of 1 (True) for this variable, meaning that the properties have a balcony, while 80% of the observations have a value of 0 (False). The standard deviation of 0.4 suggests that the values for this variable are also spread out from the mean.

The *"Orientation_North", "Orientation_South", "Orientation_West"*, and *"Orientation_East"* variables indicate the orientation of the properties. Each of these variables has a mean of 0.1 or less, with standard deviations of 0.2 or less. This indicates that the majority of the observations have a value of 0 (False) for each of these variables, meaning that the properties do not face that particular direction. For example, the *"Orientation_North"* variable has a mean of 0.1, indicating that 10% of the properties face North, while the remaining 90% do not.

3.2.2.3. Geolocation Data Analysis

The analysis of geolocation data in the study revealed three distinct types of geolocation methods employed in the dataset, as can be seen in the table below:

GeolocationType	Count
Coordenadas Gerais	16269
Coordenadas Originais	5548
Coordenadas pela Morada	5020

Table 6 - Distribution of Geolocation Typology

The most common method, *"Coordenadas Gerais"* accounted for a total of 16,269 instances, indicating its widespread usage. The second most prevalent method, *"Coordenadas Originais"*, comprised 5,548 instances, suggesting a significant but comparatively lower utilization. Interestingly, the analysis also identified *"Coordenadas pela morada"* as a distinct geolocation type, with 5,020 occurrences. These findings shed light on the diversity of geolocation methods utilized in the dataset, providing valuable insights for understanding the spatial characteristics of the data under investigation.

Attribute	count	mean	std	min	0,25	0,5	0,75	max
Number_Supermarkets	26768.0	11.5	8.0	0.0	3.0	12.0	20.0	20.0
Number_BusStops	26768.0	10.7	9.1	0.0	0.0	12.0	20.0	20.0
Number_SubwayStations	26768.0	0.0	0.1	0.0	0.0	0.0	0.0	2.0
Number_Hospitals	26768.0	1.2	2.2	0.0	0.0	0.0	1.0	20.0
Number_Pharmacies	26768.0	7.0	7.0	0.0	1.0	4.0	12.0	20.0
Number_Restaurants	26768.0	14.7	7.5	0.0	8.0	20.0	20.0	20.0
Number_Stores	26768.0	13.6	7.0	0.0	8.0	17.0	20.0	20.0

Table 7- Geolocation Summary Statistics

The descriptive statistics obtained from the transposed output of the describe function provide valuable insights into various features of the dataset related to properties amenities. For instance, the column *"Number_Supermarkets"* has a mean value of 11.5 which suggests that, on average, there are approximately 11.5 supermarkets near each property. The standard deviation of 8.0 indicates a moderate degree of variability in the data, with some observations having significantly more or fewer supermarkets. The minimum and maximum values of 0.0 and 20.0, respectively, highlight the range of the variable.

Similarly, the *"Number_BusStops"* has a mean value of 10.7 suggesting that, on average, there are approximately 10.7 bus stops in the proximity of the houses for sale. The standard deviation of 9.1 indicates a relatively higher degree of variability compared to the *"Number_Supermarkets"* column. The presence of minimum and maximum values of 0.0 and 20.0 demonstrates the range of bus stop counts.

Moving on to *"Number_SubwayStations",* it is apparent that this feature has a mean value close to 0.0, suggesting that subway stations are rarely present in the dataset. The minimum and maximum values of 0.0 and 2.0 indicate that, in some observations, there may be up to two subway stations. The low standard deviation of 0.1 implies that there is little variability in the subway station counts across the dataset.

Analyzing *"Number_Hospitals"*on average each observation has approximately 1.2 hospitals in their proximity. However, the standard deviation of 2.2 indicates a significant degree of variability in the hospital counts, with some observations having no hospitals. The presence of the 75th percentile value of 1.0 suggests that most observations have one or fewer hospitals.

Moving to *"Number_Pharmacies*", we observe that, on average, there are approximately 7.0 pharmacies in each observation. The standard deviation of 7.0 indicates a moderate degree of variability in the pharmacy counts. The minimum value of 0.0 indicates that some observations may not have any pharmacies, while the maximum value of 20.0 suggests that certain observations may have a substantial number of pharmacies.

The *"Number_Restaurants"* column reveals that, on average, each observation contains approximately 14.7 restaurants. The standard deviation of 7.5 suggests a moderate degree of variability in the restaurant counts. The presence of the maximum value of 20.0 indicates that some observations may have the maximum number of restaurants.

Lastly, the *"Number_Stores"* column indicates that, on average, each observation has approximately 13.6 stores. The standard deviation of 7.0 suggests a moderate degree of variability in the store counts. The minimum and maximum values of 0.0 and 20.0, respectively, indicate the range of store counts across the dataset.

These detailed analyses of the various features provide a comprehensive understanding of the dataset's distribution and characteristics, allowing researchers to learn about the availability and distribution of supermarkets, bus stops, subway stations, hospitals, pharmacies, restaurants, and stores in the study area.

3.2.2.4. Municipal Data Analysis

In this part of the thesis, key summary statistics that shed light on various aspects of our study area will be examined. These statistics cover important areas like housing, education, banking, and demographics.

By looking at these numbers, a better understanding of the social and economic conditions of the region will be gained. This analysis will help in identifying trends and patterns, which are crucial for the overall understanding and conclusions of the study. The focus will be on some of the most significant variables to provide a clear and concise overview of the region's current state.

Starting with *"Alojamentos familiares clássicos"*(Classic Family Dwellings), we observe a mean of 69,378.8 and a high standard deviation of 80,645.2. This variability and the substantial range between the minimum (1,018) and maximum (324,554) values suggest significant diversity in family dwellings across the study area. The quartile values reinforce this diversity, with a median significantly lower than the mean, indicating a skew towards lower-value dwellings.

In contrast, *"Alojamentos turísticos"*(Tourist Accommodations) have a much lower mean of 81.1, with a standard deviation of 131.7. The range here is narrower, from 0 to 487, indicating a more uniform distribution but with potential outliers, as suggested by the high standard deviation relative to the mean.

The statistics for *"Alunos do ensino não superior"*(Non-Higher Education Students) and *"Alunos do ensino superior"*(Higher Education Students) highlight the educational landscape. The mean for non-higher education students is 21,527.7, with a larger standard deviation of 29,151, suggesting considerable variability in student populations across different regions. For higher education students, the mean is lower at 13,396.9, but the standard deviation is significantly higher at 32,948.1, indicating even greater variability and possibly a more concentrated presence in specific areas.

The *"Bancos Caixas Económicas"* (Banks and Savings Banks) and *"Caixas automáticas multibanco"*(ATMs) indicate financial infrastructure distribution. Banks have a lower mean (57.6) with a high standard deviation (110.9), suggesting an uneven distribution. ATMs show a similar trend but with a higher mean of 190.1 and a standard deviation of 310.3, implying a broader but still uneven spread.

"Beneficiários do Rendimento Social de Inserção (RSI)" (Beneficiaries of Social Insertion Income) data, with a mean of 3,891.4 and a standard deviation of 5,868.3, reflect socio-economic disparities. The range and quartile values suggest a concentration of beneficiaries in certain areas.

The *"Idosos (%)"* (Elderly Population Percentage) and *"Jovens (%)"* (Youth Population Percentage) provide insights into demographic profiles. The elderly population shows a higher mean percentage (24.2%) with a smaller standard deviation (4.5), indicating a consistently aging population. In contrast, the youth population has a lower mean percentage (13.0%) and a smaller standard deviation (1.6), which suggests a more uniform distribution of the younger population across regions.

"Densidade populacional" (Population Density) is another crucial variable, with a mean of 1,224.9 and a high standard deviation of 1,814.2, indicating significant differences in

population density across regions. The range from a minimum of 4.4 to a maximum of 7,262 per square kilometer highlights the vast disparity in urban and rural settlement patterns.

Lastly, *"Óbitos"* (Deaths) statistics, with a mean of 1,482.7 and a standard deviation of 1,913.8, mirror the varying mortality rates across different regions. The considerable range in these values underscores the diversity in health and longevity patterns among the population.

Overall, these statistics provide a comprehensive picture of the socioeconomic and demographic landscape, highlighting significant disparities and areas that may require targeted policies and interventions.

3.2.2.5. Missing Value Analysis

The dataset contains missing values in multiple columns, with "Floors" having the most missing values (23270). In some of these columns, the absence of a value indicates a lack of feature, for instance, if there is no information regarding the parking then it means there is no parking included (or at least it wasn't included in the advertisement). The same rational can also be applied to other features such as *"Floors"* or *"Balcony"*, in which the lack of web scraped value indicates the absence of the characteristic.

The missing values of other columns such *as "Orientation", "Latitude"* and "*Longitude*" on the other hand, are cases in which the advertiser didn't include that specific information about the property.

In addition, there are 69 missing values in various columns related to local amenities such as supermarkets and bus stops, which meets the results previously stated that the script was unable to find the amenities for 69 advertisements. Finally, there are also 65 missing values in columns related to municipality data, such as population demographics and government spending, which are related to the houses that the algorithm was incapable of attributing to the municipality.

In conclusion, there are different meanings for the missing values in this dataset, depending on the column. The dataset will require transformations and the imputation of missing values in the feature engineering phase of the project accordingly.

3.2.2.6. Correlation Analysis

Furthermore, a correlation analysis was conducted to try to understand how various factors relate to house prices. Note that in order to perform this analysis, the missing values previously mentioned were imputed with the average of each column, ensuring completeness for the analysis.

Correlation with Price	Correlation
Price	1.000000
Area	0.457304
População estrangeira em % da população residente	0.299776
Crimes registados pelas polícias por mil habitantes	0.231846
População estrangeira	0.226495
Alojamentos turísticos	0.206814
Estabelecimentos do 2 º ciclo do ensino básico	0.205644
Alojamentos familiares clássicos	0.198735
Trabalhadores da Administração Pública Local	0.198315
Alunos do ensino não superior	0.194933

Table 8 - Correlation to Price

The findings revealed that the size of the property, measured by area, had the strongest connection with house prices, with a correlation coefficient of 0.457, which indicates that generally, larger properties tend to be priced higher. Additionally, the percentage of foreign residents in an area also showed a significant positive correlation with house prices (0.300) suggesting that neighborhoods with higher percentages of foreign residents might be more desirable, potentially driving up property values with the increased demand.

Another notable result was the positive correlation between house prices and recorded crime rates per thousand residents (0.232). This might reflect higher property values in urban areas, which often experience higher crime rates than rural areas.

Other characteristics, such as the availability of tourist lodgings and schools, also revealed positive associations with house prices, implying that regions with convenient amenities and educational institutions are likely to have higher property values.

To summarize, this correlation research assisted in identifying the most relevant factors influencing home prices in Portugal, emphasizing not just traditional criteria such as property size and location, but also the impact of socioeconomic issues on the housing market.

3.2.2.7. Outlier Analysis

Additionally, an analysis of the outliers present in the dataset was performed, since the findings obtained might prove to be essential for the preprocessing phase of modeling, as they can highlight areas where the data may need to be treated or cleaned to improve model accuracy.

A quantitative method was applied to identify outliers calculating the interquartile range (IQR) for each variable, and defining outliers as those data points that fell below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR. This approach is widely used in statistical analyses to detect unusually high or low values that may affect the overall analysis.

The results indicated a substantial number of outliers across a range of variables; for instance, there were notable outliers in *"Alojamentos turísticos*" and *"Alunos do ensino superior"*, suggesting significant variability in these areas across different regions. Such disparities could indicate regional differences in economic activity and educational infrastructure, potentially influencing local housing markets.

Conversely, variables such as *"Ganho médio mensal dos trabalhadores por conta de outrem*" showed no outliers, implying a more uniform distribution. This could suggest consistent income levels or similar economic conditions across Portugal.

This analysis suggests that, while some locations demonstrate consistency, others exhibit large variations, which could be critical in understanding the factors influencing property

pricing. Such insights are crucial for improving predictive models and ensuring their accuracy and generalizability across different regions of Portugal.

3.3. Data Preparation

Following a thorough analysis of the data, which revealed patterns, trends, and preliminary correlations, the feature engineering phase is the next critical step in achieving the best prediction outcomes, with this phase being critical in refining the dataset to a form that is most suited for predictive modeling. (Géron, 2022). In this section, several feature transformation procedures will be used in order to not only clean the data but also extract and construct new features. Additionally, feature selection will be also covered to reduce the dataset to the attributes that bring value to the model and help improve its performance. Finally, *scikit-learn* pipelines will also be introduced as the framework used to apply the transformations consistently across the training and testing datasets, maintaining the integrity and reproducibility of the model's input data.

3.3.1. Feature Transformation

With the focus of optimizing the dataset, feature engineering procedures were performed on the dataset to enhance and improve the input for the machine learning model aimed at forecasting housing values in the Portuguese market. These procedures included a variety of transformation and preparation strategies aimed at improving the model's capacity to discover meaningful patterns from the data.

3.3.1.1. Transformation and Cleaning of Data

The feature engineering phase commenced with basic transformations to standardize and clean the dataset. This included replacing placeholder values like "NA" with actual *NaN* values, which are more readily handled during analysis. Property-specific fields such as *"BuiltIn"*, *"Bathrooms"*, *"Rooms"*, and *"Floors*" underwent transformations to strip non-numeric characters and convert them into float types for mathematical operations. This step is crucial as it ensures that all numerical data is in a uniform format, facilitating more accurate computations and predictions.

3.3.1.2. Extracting and Constructing New Features

One of the more complex transformations involved extracting two key metrics from the *"FullArea*" field: "*BuiltArea*" and "*FloorArea"*". This was achieved using regular expressions to parse and separate the built and floor areas into distinct features before dropping the original *"FullArea*" column. Such detailed breakdowns allow the model to more precisely evaluate property values based on specific area metrics.

Additionally, the energy efficiency ratings were converted from alphabetic to numeric scales. This mapping simplifies the model's understanding by quantifying energy efficiency, which is a significant factor in property valuation. The transformation of textual descriptions into binary flags for features such as *"Parking"*, *"FittedWardrobes"*, *"Terraced"*, and *"Balcony*" was also performed. These binary flags are particularly useful for machine learning models as they provide clear, actionable signals about the presence or absence of these attributes.

Additionally, the generation of a new feature, *"Age"*, was calculated by subtracting the year properties were built from the current year. This feature provides a direct measure of property age, which can be a determinant of value due to factors like architectural style, wear and tear, and historical value.

3.3.1.3. Orientation and Geolocation Adjustments

A different approach was taken to handle the *"Orientation*" of properties by creating separate binary columns for each cardinal direction. This allows the model to capture the effects of property orientation on pricing, which can be significant due to factors like sunlight exposure and prevailing winds. The removal of *"Latitude*" and *"Longitude*" simplifies the model while focusing on more impactful location-based features encoded elsewhere.

3.3.2. Feature Selection

Additionally feature selection was performed which was critical for separating the most important variables and rejecting those that may dilute the model's predictive ability. By focusing on the most influential features, we ensure that the machine learning model is both efficient and resistant to overfitting. This selection is critical for improving the model's

performance and interpretability, laying the groundwork for the succeeding modeling stages. (Venkatesh & Anuradha, 2019).

3.3.2.1. Feature Selection Approach

To streamline the feature selection process, the Recursive Feature Elimination with Cross-Validation (RFECV) method was employed, using an XGBRegressor as the estimator. RFECV is a robust technique that merges the capabilities of recursive feature elimination and cross-validation, making it highly effective for pinpointing the most relevant features in predictive modeling (Misra & Yadav, 2020).

The process begins by initializing the XGBRegressor, chosen for its ability to handle diverse data types and model complex relationships. Then, the RFECV is instantiated with a 5-fold cross-validation to ensure a thorough evaluation of the model's performance across different subsets of the dataset. This approach not only helps in validating the model robustly but also in fine-tuning the accuracy of the predictions. The scoring metric used was the negative mean absolute percentage error, which focuses on minimizing the errors relative to the actual values—an approach especially critical in the precise estimation of house prices.

Initially, the model is trained with all available features to establish a baseline performance. It then assesses each feature's importance, typically derived from the internal feature importance metrics of the XGBRegressor. Features deemed least important are recursively eliminated, and the model is retrained on the pared-down set of features. This cycle of training, evaluation, and elimination continues, with each iteration undergoing cross-validation to assess the impact of the reduced feature set on model performance (Géron, 2022).

This iterative feature elimination continues until the optimal subset is established, which is defined as no further improvement or even a potential drop in performance when additional characteristics are removed. The final collection of selected features includes the most effective predictors, which are a balanced combination of property-specific qualities and broader socioeconomic factors.

In the practical implementation of the feature selection strategy, Python was used, leveraging the powerful machine learning library, scikit-learn. The core of the feature selection process

was encapsulated through the use of the RFECV class from the *sklearn.feature_selection* module, with XGBRegressor from XGBoost serving as the underlying estimator.

3.3.2.2. Feature Selection Results

Out of the extensive list of initial features, which included a variety of demographic, economic, and property-specific attributes, the RFECV process selected a substantial subset as being most predictive of house prices. Specifically, of the total 79 initial set of variables, RFECV identified 58 features that significantly contribute to the predictive power of the model.

These selected features encompass a diverse range of indicators, from socio-economic variables such as *"Alunos do ensino superior*" (higher education students) and *"Desempregados inscritos nos centros de emprego*" (registered unemployed), to more direct property-related factors like *"Area"*, *"Bathrooms"*, and *"Floors"*. This selection underscores the multifaceted nature of real estate pricing, which is influenced by a combination of local amenities, economic conditions, and property characteristics.

However, this also implies that a significant number of features were omitted from the final model. These eliminated qualities, which did not make the cut, were likely deemed less influential due to redundancy, low volatility, or insufficient association with the target variable, house prices.

The deliberate selection of variables guarantees that the forecasting model is both manageable and strong, emphasizing data inputs that provide the most significant insights into the dynamics of the housing market.

3.3.3. Pipeline Integration and Application

All these transformations were integrated into a cohesive pipeline using Scikit-Learn's Pipeline and *ColumnTransformer* frameworks using Python. This setup ensures that all transformations are applied consistently across the training and testing datasets, maintaining the integrity and reproducibility of the model's input data. Numeric and categorical data were handled separately to preserve the intrinsic properties of each data type, with categorical variables undergoing one-hot encoding to transform them into a machine-readable format.

This comprehensive approach to feature engineering not only prepares the dataset in a form that is optimal for modeling but also embeds domain-specific knowledge into the preprocessing steps, enhancing the predictive power and accuracy of the model dedicated to forecasting house prices in the Portuguese market.

3.4. Modeling

Following the initial feature engineering stages mentioned in the preceding section, a lengthy experimental phase was carried out to assess not just various feature engineering methodologies but also different regression models. This phase aims to determine the most effective combination of feature engineering methodologies and algorithms for increasing the accuracy of house price estimations in the Portuguese market.

The experiment will involve iteratively applying several feature engineering strategies, with an emphasis on three major areas: missing value imputation, feature scaling, and outlier handling. Each technique will be tested in combination to determine its overall impact on the performance of the various regression models, which were chosen based on previous research as mentioned in the literature review. The various feature engineering techniques and regression models employed are summarized as follows:

Feature Engineering Techniques Explored:

- Missing Value Imputation:
    - Median Imputation
    - K-Nearest Neighbors Algorithm Imputation (KNN)
- Feature Scaling:
    - Standardization
    - Normalization
    - Robust Scaling
- Outlier Handling:
    - Interquartile Range (IQR)
    - None

Regression Models Explored:

- Decision Tree Regressor
- Random Forest Regressor
- K-Neighbors Regressor
- Gradient Boosting Regressor
- Support Vector Machines
- Cat Boost Regressor
- XGBoost Regressor

The experimental setup uses a combination of the feature engineering techniques and regression algorithms mentioned, generating a total of 84 unique combinations of feature engineering techniques and models in your experiment.

Regarding the evaluation protocol, to assess the performance of the model the dataset was initially split into training and testing sets in Python using the *train_test_split* function from scikit-learn, allocating 80% for training and 20% for testing set, using a random state to ensure reproducibility of the split, assessing the models performance based on several metrics, including fit time, score time, Root Mean Squared Error (RMSE), the coefficient of determination (R^2) and Mean Absolute Percentage Error (MAPE).. Only the training dataset was used in this section to not introduce bias in the model and evaluate its final performance correctly.

Additionally, for each combination of techniques Stratified 5-fold Cross-Validation was employed to the training dataset to ensure robustness of the results, training and testing each model five times, once for each separate fold, to ensure a comprehensive assessment of performance across different segments of the dataset, consequently, yielding a total of 420 distinct results.

In addition to the feature engineering and model evaluation strategies that will be described, properties with prices below €50,000 and above €1,000,000 were excluded from the analysis and scope of the model and thesis. This filtering ensured that the dataset focused on a more standardized range of house prices, reducing the impact of extreme values that could skew the model's learning and predictive accuracy.

3.4.1. Feature Engineering Experimentation

The following section explores the results obtained from the different combinations of techniques and regression models on average.

3.4.1.1. Missing Value Imputation

As mentioned, the experimentation included two primary methods for handling missing data. The first method, Median Imputation, involves replacing missing values with the median of the respective feature, which is beneficial due to its resistance to outliers and simplicity in application. The second method, KNN Imputation, uses the K-Nearest Neighbors algorithm to impute missing values based on the proximity to similar data points, which, while more computationally demanding, can yield more accurate estimations, especially in datasets with complex correlations (Zhang, 2012). After all the iterations were completed, the average results were as follows:

Model	Fit Time (s)	Score Time (s)	RMSE	R^2	MAPE
KNN	22.3	2.5	142,768.1	0.55	34.66%
Median	21.6	2.5	142,607.4	0.55	34.71%

Table 9 - Missing Value Metrics Comparison

In the evaluation of the two distinct approaches for imputing missing values, all the measures were very similar on average. Additionally, the Mean Absolute Percentage Error (MAPE) provides insightful metrics on their predictive accuracy, indicating a marginal difference between the two methodologies: the KNN approach yields a MAPE of 34.66%, while the Median imputation method results in a slightly higher MAPE of 34.71%. To better evaluate the performance of the KNN approach, further experimentation was performed with different numbers of neighbors.

3.4.1.1.1. Experiment with KNN Imputation Values

The experiment involved varying the number of neighbors in the KNN Imputer from 1 to 15. This range was chosen to understand the effects of both smaller and larger neighborhood sizes on imputation quality and subsequent model performance. Each configuration was

systematically applied to preprocess the data. The following average results were obtained after running all the iterations:

KNN (k)	MAPE (%)	Fit Time (s)
1	35.01	21.4
3	34.68	21.7
5	34.77	21.6
7	34.83	21.9
10	34.74	21.8
15	34.87	22.1

Table 10- Impact of Neighbor Count on KNN Imputation Performance

In this extended analysis of the K-Nearest Neighbors (KNN) algorithm for missing value imputation, varying the number of neighbors demonstrates subtle but noteworthy differences in predictive accuracy, as measured by the Mean Absolute Percentage Error (MAPE). The results show that the smallest k value (k=1) resulted in the highest MAPE at 35.01%, suggesting that using a single neighbor may lead to overfitting and less accurate imputations. Conversely, a moderate increase in the number of neighbors tends to improve accuracy, with k=3 achieving the lowest MAPE of 34.68%.

3.4.1.2. Feature Scaling Methods

To standardize the range of independent variables, three scaling methods were applied iteratively. Standardization adjusts features to have a zero mean and unit variance, commonly used to bring all variables into a comparable range without distorting differences in the ranges of values. Normalization scales data into a bounded interval, which is often essential when features are measured on different scales. Lastly, Robust Scaling uses more resilient statistics that are not influenced as much by outliers, making it suitable for datasets where outliers might skew the representation of genuine data distribution (Ahsan et al., 2021). The following results were obtained on average after running all the iterations:

Preprocessing Technique	Fit Time (s)	Score Time (s)	RMSE	R^2	MAPE

Normalization	21.8	2.5	143,522.3	0.55	34.90%
Robust	22.2	2.5	142,952.4	0.55	34.91%
Standardization	21.8	2.5	141,588.4	0.56	34.24%

Table 11 - Results Comparison Feature Scaling Techniques

The results show a slightly higher MAPE for both normalization (34.90%) and robust scaler (34.91%), suggesting minimal impact on the accuracy of these techniques when compared to each other. However, standardization stands out with a notably lower MAPE of 34.24%, indicating superior performance in reducing prediction errors. This disparity demonstrates the importance of standardization in improving model dependability by changing feature scales to a common metric, decreasing the impact of outliers and scale disparities. According to the analysis, standardization among the assessed strategies may provide a minor boost in prediction model accuracy.

3.4.1.3. Outlier Handling Approaches

Two approaches were considered to manage outliers in the dataset. The IQR Method utilizes the Interquartile Range to detect outliers, either removing them or adjusting them to the nearest acceptable range defined by the IQR thresholds. This method helps in mitigating the impact of extreme values on the model's training process (Aguinis et al., 2013). The alternative approach was to perform no outlier handling at all, allowing the models to train on the full, unmodified dataset, which helps in assessing the effect of naturally occurring outlier values on model performance. Following the completion of every loop, the average outcomes were as follows:

Outlier Handling Method	Fit Time (s)	Score Time (s)	RMSE	R^2	MAPE
IQR	21.5	2.5	143,022.3	0.55	34.71%
None	22.4	2.5	142,353.1	0.55	34.66%

Table 12- Outlier Handling Methods Results

The results of the outlier handling techniques focusing on the Interquartile Range (IQR) method versus no outlier treatment, reveal subtle differences in model performance. The use of the IQR method results in a slightly higher Mean Absolute Percentage Error (MAPE) of 34.71% compared to 34.66% when no outliers are removed and very similar fit and score time. It is also important to note that, even though the IQR performance was slightly worse on average than no outlier treatment, the Robust Scaler was also being tested to treat the outliers when used.

3.4.2. Regression Models Experimentation

The performance of various regression models is explored to determine their efficacy in predicting outcomes based on historical data. The results from this experimentation helped identify the most effective model in terms of accuracy, efficiency, and suitability for the specific dataset used, and were summarized in the table below:

Model	Fit Time (s)	Score Time (s)	RMSE	R^2	MAPE
CatBoost	26.9	0.0	113,280.4	0.73	27.77%
Decision Tree	0.9	0.0	159,858.5	0.48	33.16%
Gradient Boosting	12.3	0.0	128,360.0	0.66	32.31%
K-Nearest Neighbors	0.0	1.5	140,020.4	0.60	33.46%
Random Forest	51.8	0.2	114,151.1	0.72	27.53%
Support Vector Machine	52.1	15.8	227,239.5	-0.06	64.18%
XGBoost	9.7	0.0	110,904.2	0.74	27.38%

Table 13 - Comparative Performance of Regression Models

The experimentation with various regression models demonstrates significant differences in their performance, pinpointing which models are most suited for predictive tasks with the dataset at hand. XGBoost stood out as the top performer with the highest R^2 value of 0.74 and a Mean Absolute Percentage Error (MAPE) of 27.38%, indicating its superior predictive accuracy and consistency. Its moderate fit time of 9.7 seconds suggests a good balance

between efficiency and performance. Close behind, the Random Forest model showed a robust R^2 of 0.72 and a MAPE of 27.53%, but with a notably longer fit time of 51.8 seconds, which might be a consideration in time-sensitive environments.

Conversely, the Support Vector Machine (SVM) model drastically underperformed, with the highest MAPE at 64.18% and a negative R^2 of -0.06, indicating poor fit to the data. Not only was its predictive performance lacking, but the SVM also had the longest score time at 15.8 seconds, significantly higher than any other model tested, which could further disadvantage it in scenarios where rapid predictions are crucial.

3.4.3. Experimentation Summary

Additionaly, below is the summary table with the results obtained for each execution of the combinations of techniques and models:

Row Labels	Avg Fit Time	Avg Score Time	Avg RMSE	Avg R2	Avg MAPE
CB	26.9	0.0	113280.4	0.73	27.77%
DT	0.9	0.0	159858.5	0.48	33.16%
GB	12.3	0.0	128360.0	0.66	32.31%
KNN	0.0	1.5	140020.4	0.6	33.46%
RF	51.8	0.2	114151.1	0.72	27.53%
SVM	52.1	15.8	227239.5	-0.06	64.18%
XGB	9.7	0.0	110904.2	0.74	27.38%
knn (Missing Value Imputation)	22.3	2.5	142768.1	0.55	34.66%
median (Missing Value Imputation)	21.6	2.5	142607.4	0.55	34.71%
Grand Total (Missing Value Imputation)	21.9	2.5	142687.7	0.55	34.68%

normalization (Feature Scaling)	21.8	2.5	143522.3	0.55	34.90%
robust (Feature Scaling)	22.2	2.5	142952.4	0.55	34.91%
standardization (Feature Scaling)	21.8	2.5	141588.4	0.56	34.24%
Grand Total (Feature Scaling)	21.9	2.5	142687.7	0.55	34.68%
IQR (Outlier Scaling)	21.5	2.5	143022.3	0.55	34.71%
None (Outlier Scaling)	22.4	2.5	142353.1	0.55	34.66%
Grand Total (Outlier Scaling)	21.9	2.5	142687.7	0.55	34.68%

Table 14- Results Summary

4. Results

In this section, the final pipeline will be developed based on the works of the previous sections followed by the tuning of the parameters of the model to improve its results, with the results being presented. Additionally, explainability techniques were also used providing insights into the model's decision-making processes enhancing its transparency, as well as a final section with a discussion about the limitations of the results obtained and considerations about future work.

Based on the objectives, the results will provide insights regarding the significant factors that influence house prices in the Portuguese real estate market while uncovering insights about the dynamics within the Portuguese real estate market, determine how machine learning algorithms can predict the house prices, identify the most effective machine learning regression algorithm for this purpose and explore the best preprocessing and feature engineering techniques to enhance model performance.

Finally, to assess the performance of the model the test dataset that resulted from the initial split was used and the Mean Absolute Percentage Error (MAPE) as the measures to evaluate the model accuracy.

4.1. Final Pipeline

Building on the foundations established in the feature engineering and modeling phases, the deployment of the final pipeline constitutes the final step in optimizing the dataset for modeling and predicting the data. The final pipeline, crafted using Scikit-Learn's pipeline functionalities, is designed to handle various aspects of data preprocessing systematically:

- **Initial Cleansing Operations**: The cleanse_pipeline initiates the transformation process by performing global transformations such as replacing placeholders, normalizing textual data, and extracting numerical values from mixed-type columns.
- **Handling Categorical Variables**: The *cat_pipeline* specifically addresses categorical data by employing One Hot Encoding. This transformation is crucial for handling

categorical variables by creating binary columns for each category, thereby eliminating the model's bias towards ordinal relationships where none exists.

- **Handling Numerical Variables**: The *num_pipeline* applies robust scaling through *RobustScaler()* – both scaling the data and dealing with the outliers - and missing value imputation using *KNNImputation* with 3 neighbors, following the best results obtained in the modeling phase.

Additionally, the regression algorithm chosen was the XGBoost Regressor since it was the one with the best results on average in the modeling phase and was also quite efficient in terms of training and scoring times.

4.2. Hyperparameter Tuning

Hyperparameter tuning is a critical phase in the optimization of machine learning models, adjusting the model parameters in the training process can greatly influence the model's performance therefore, the emphasis was on refining the XGBoost regressor.

For the tuning process Repeated Randomized Search with Grid Search was used which is basicly taking the most out of two main strategies were used: Randomized Search and Grid Search, exploring a broad range of parameter combinations and then refining the search around the most promising configurations (Bergstra & Bengio, 2012).

The first stage employs *RandomizedSearchCV* from Scikit-Learn, which samples a given number of parameter combinations, from specified distributions. This method is particularly useful when the parameter space is large, offering a more efficient alternative to exhaustive grid search. The param_grid_random was defined as follows, reflecting a diverse range of values to explore for the parameters of XGBoost (Chen & Guestrin, 2016):

- ***n_estimators***: specifies the number of gradient boosted trees in the model, with options ranging from 400 to 1000. Increasing the number of trees can improve model accuracy but may also lead to overfitting if not balanced with other parameters.
- ***learning_rate***: controls the step size shrinkage used to enhance model performance and prevent overfitting. The values vary from 0.01 to 0.5, allowing the model to adjust the contribution of each tree added to the ensemble.

- ***max_depth***: sets the maximum depth of each tree, selected from the set {8, 9, 10, 15}. Deeper trees can model more complex patterns but may capture noise, leading to overfitting.
- ***subsample***: determines the fraction of samples used for fitting each tree, chosen from [0.7, 0.8, 0.9, 1.0]. A subsample value less than 1.0 can lead to a reduction in variance and an increase in bias, aiding in the prevention of overfitting.
- ***colsample_bytree***: specifies the fraction of features to use per tree, with possible values of 0.8, 0.9, and 1.0. Like subsample, this parameter helps in managing the model's complexity and overfitting by altering feature space in each iteration.

Multiple runs of randomized search were carried out to reduce the randomness inherent in the search process, ensuring that the observed optimal parameters were resilient and reproducible.

Following the randomized search, the best hyperparameters were chosen to define a more specific *param_grid_grid* for the grid search. This step is conducting a more in-depth search with *GridSearchCV*, focusing on the proximity of the best results from the randomized phase. To fine-tune the model, the grid search parameter grid moved each parameter marginally closer to the best-found value.

The process was repeated 10 times, and the overall best configuration across all runs was identified to include a *colsample_bytree* of 0.8, a *learning_rate* of 0.1, a *max_depth* of 7, *n_estimators* of 500, and a subsample of 0.9.

To quantify the impact of hyperparameter tuning, the model's performance before and after tuning was compared, obtaining a Mean Absolute Percentage Error (MAPE). Utilizing the base model with default parameters resulted in a MAPE of 27.4%, which serves as the baseline for performance comparison. After applying the optimal parameters identified through the tuning process, the MAPE improved to 26.08%. This improvement in MAPE illustrates the effectiveness of hyperparameter tuning, clearly demonstrating its value in enhancing the model's accuracy and predictive reliability in regression tasks.

The results obtained for the final tuned pipeline are summarized in the table below:

Row Labels	Avg Fit Time (s)	Avg Score Time (s)	Avg RMSE	Avg R2	Avg MAPE
XGB	9.5	0.0	110834.5	0.76	26.08%

Table 15- Final Results Summary

4.3. Explainability

The use of explainability techniques in machine learning models not only provides insight into the model's decision-making processes but also enhances trust and transparency, particularly in critical applications such as real estate valuation (Linardatos et al., 2020). One of the core components of explainability in this context is feature importance, which helps in understanding the influence of different variables on the predictive outcomes of the model.

4.3.1. XGBoost Feature Importance

The XGBoost algorithm used in this dissertation includes a technique for estimating feature importance, which is an important part of model interpretation. This approach largely uses Gini importance, also known as mean decrease impurity, to estimate the significance of each attribute. The Gini significance assesses how each feature contributes to the uniformity of the nodes and leaves in the model's decision trees (Chen & Guestrin, 2016). The chart below visualizes the top 10 features determined by XGBoost's feature importance from the model:

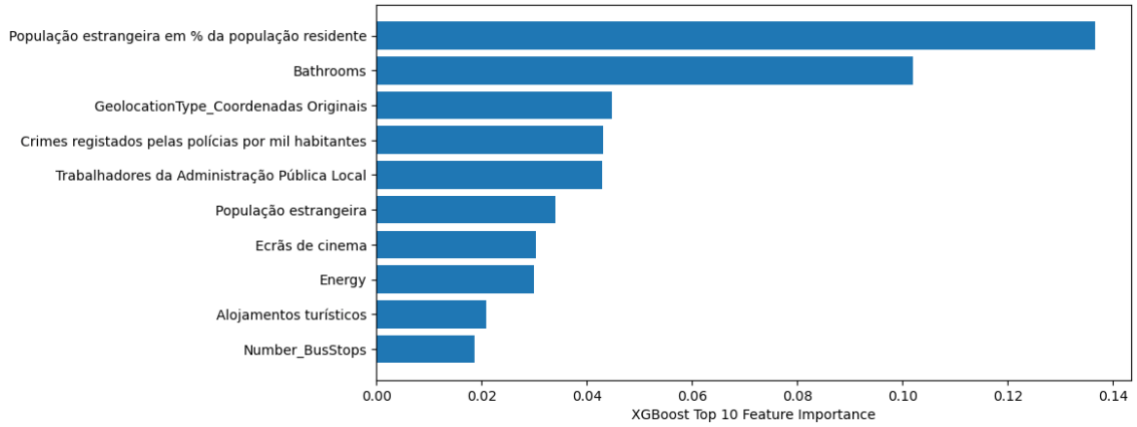


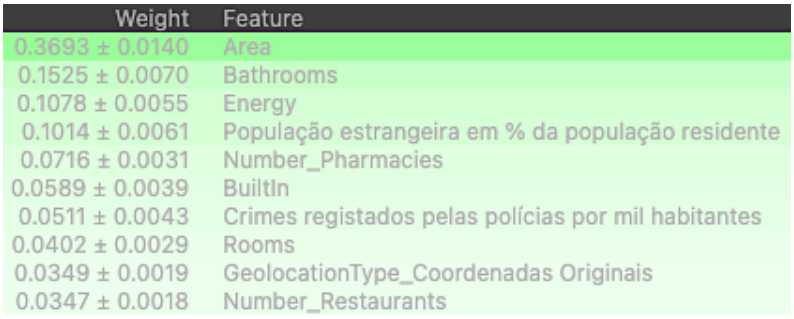
Figure 3- XGBoost Top 10 Feature Importance

The most significant feature as depicted is "*População estrangeira em % da população residente*" indicating a strong influence of the percentage of foreign population in a given area on property values. This feature is followed by other important predictors such as "Bathrooms," "*GeolocationType_Coordenadas Originais*" and "*Crimes registados pela polícia por mil habitants*" each contributing differently to the model's predictions.

4.3.2. Permutation Importance

Permutation importance offers a compelling alternative to the standard feature importance metrics employed by tree-based models like XGBoost. Permutation importance is a model-agnostic metric that measures the increase in the model's prediction error after permuting the feature by randomly shuffling each predictor variable in the validation dataset and observing the effect on the model's accuracy, as opposed to impurity-based importance, which can be biased toward features with higher cardinality or more categories. This procedure fractures the relationship between the feature and the outcome; thus, the change in model performance reflects the feature's predictive capability, with a significant drop in model performance following shuffling implying a high importance of the particular feature. (Altmann et al., 2010).

This technique was implemented in the context of this project by taking advantage of the *PermutationImportance* module from *eli5*, a Python library for debugging machine learning classifiers and explaining their predictions. The following results were obtained:



Weight	Feature
0.3693 ± 0.0140	Area
0.1525 ± 0.0070	Bathrooms
0.1078 ± 0.0055	Energy
0.1014 ± 0.0061	População estrangeira em % da população residente
0.0716 ± 0.0031	Number_Pharmacies
0.0589 ± 0.0039	BuiltIn
0.0511 ± 0.0043	Crimes registados pelas polícias por mil habitantes
0.0402 ± 0.0029	Rooms
0.0349 ± 0.0019	GeolocationType_Coordenadas Originais
0.0347 ± 0.0018	Number_Restaurants

Figure 4- Permutation Importance Results

As can be seen in the figure, the permutation importance revealed a hierarchy of features that significantly affect the predictive accuracy of the real estate pricing model with the most crucial feature, "*Area*" with a weight of 0.3693 ± 0.0140, notably impacting the model's error rate when altered, underscoring its dominant role in property valuation. Following closely are "*Bathrooms*" and "*Energy*" with importance weights of 0.1525 ± 0.0070 and 0.1078 ± 0.0055, respectively. These features reflect the property's size, luxury level, and energy efficiency, which are crucial considerations in the real estate market. Other features like "*GeolocationType_Coordenadas Originais*" and "*Crimes registados pelas polícias por mil habitantes*" also emerged as important, though less so than in impurity-based rankings as seen in the previous chapter, illustrating the value of permutation importance in offering a more balanced and less biased assessment of feature significance.

4.3.3. SHAP Analysis

SHAP (*SHapley Additive exPlanations*) values provide a powerful and intuitive way to understand the contribution of each feature to individual predictions made by a model. This method decomposes a prediction to show the impact of each feature, offering a detailed explanation of how factors influence the model's output and enhancing transparency and trust in the model's decisions (Ekanayake et al., 2022).

For the real estate pricing model, SHAP values were computed to analyze the influence of features on a specific prediction. The process began by selecting a single prediction instance from the test dataset representing a typical data point within the dataset used for testing the

model, and then a *TreeExplainer* object was created using the optimized XGBoost model previously obtained. The following results were obtained:

Figure 5- SHAP Analysis Example

The SHAP force plot for the selected instance from the test dataset provides a detailed perspective of how various characteristics influence the anticipated property value. The size of the property has a significant positive influence, demonstrating that larger properties typically attract higher prices. Similarly, the number of bathrooms increases the property's anticipated price greatly, reflecting both luxury and practicality. Furthermore, having more stores nearby increases the property's desirability, which raises its value.

Conversely, some attributes have a negative impact on property value, such as lower energy ratings, which detract from the property's worth because the market prefers more energy-efficient or modern buildings. Furthermore, having more restaurants nearby, while potentially indicating a bustling neighborhood, somewhat lowers property value, which could be linked to worries about noise or congestion that are common in densely populated regions.

Additionally, a SHAP summary was performed, providing a comprehensive view of feature impacts across the entire dataset, revealing how each feature contributes to the model's prediction, which can be very useful in identifying patterns and trends that may not be apparent from individual predictions. The following results were obtained:

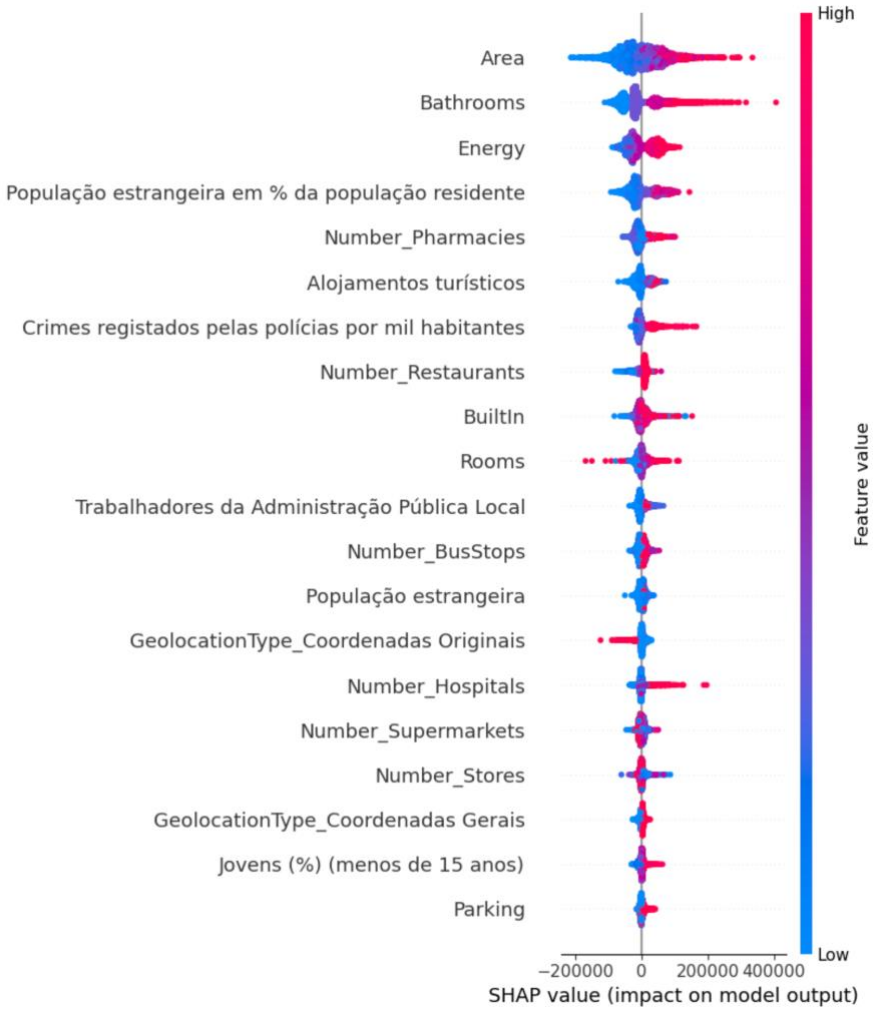


Figure 6- SHAP Summary Results

The SHAP summary figure above depicts the impact of numerous attributes on predictions, with each point indicating a feature's SHAP value across multiple forecasts. The plot is designed to provide a clear picture of feature influence by grouping the dots vertically based on the feature they represent and aligning all SHAP values for a given feature along a horizontal line. The color of the dots varies—red for high feature values and blue for low ones—making it easier to see how varied levels of feature values affect predictions. The arrangement of these dots along the horizontal axis reflects their influence on the model's output; dots to the right indicate a positive impact, increasing property values, while dots to the left indicate a drop.

With this in mind, the SHAP summary plot reveals several insights into how features affected house prices. For instance, features such as "*Area*" and "*Bathrooms*" prominently display red dots shifted far to the right, underscoring their substantial positive influence on property values—a reflection of the basic market understanding that larger spaces and more amenities boost property prices. Meanwhile, features like "*Energy*" and "*Number of Pharmacies*" also impact predictions, but more subtly; their dots, while centered, lean slightly towards higher values, suggesting a nuanced effect on property valuation. Conversely, features like "*Number of Restaurants*" and "*População estrangeira em % da população residente*" predominantly show blue dots on the left, indicating that higher instances of these features may slightly depress property values, possibly reflecting concerns such as noise or social preferences.

4.4. Discussion

This objective of this subsection is to directly address the research questions outlined at the beginning of the dissertation.

Regarding the first question of what the most significant factors are influencing house prices in the Portuguese real estate market, from a comprehensive analysis that included data extraction from online real estate advertisements, geolocation details, and municipal socioeconomic data, the correlation analysis highlighted that the area of a property is the most significant predictor, with larger properties commanding higher prices. Moreover, regions with a higher percentage of foreign residents tend to have elevated property values, suggesting a desirability factor that influences pricing. Other factors, such as the availability of tourist accommodations and proximity to educational institutions, also positively impact house prices. This suggests that amenities and local infrastructure significantly contribute to real estate valuation, with properties in well-serviced areas attracting higher prices. This nuanced understanding, gained through data-driven analysis, underscores the complex interplay of various factors shaping the housing market in Portugal.

In response to the second question, regarding the role of geographical factors, as evidenced by the comprehensive data analysis conducted in this dissertation, these factors play a pivotal role in influencing house prices in Portugal. Proximity to amenities such as schools, supermarkets, and healthcare facilities significantly enhances property values, reflecting the premium that buyers place on convenience and accessibility. Additionally, the correlation

analysis highlights that neighborhood crime rates also impact house prices, albeit in a complex manner. Surprisingly, areas with higher crime rates, often urban centers, also tend to have higher property values, possibly due to their central location and the concentration of resources. School districts further compound this effect; properties located within the catchment areas of well-regarded schools command higher prices, underscoring the importance of educational opportunities in real estate valuation. These geographical factors collectively shape the desirability and, consequently, the pricing of properties in various regions across Portugal, demonstrating their significant influence on the real estate market.

Additionally, regarding the most effective techniques for preprocessing and feature engineering for enhancing the performance of house price prediction models, in the evaluation section, after some initial data cleansing and data standardization several techniques were experimented and the results showed that in the context of the dataset the most powerful technique regarding missing value imputation was KNN slightly outperforming its peers, with the best results being obtained with 3 neighbours. Regarding the outlier handling and feature scaling techniques, the results were very close and the robust scaler was used dealing with both at the same time.

In determining the most effective machine learning regression algorithm for predicting house prices in the Portuguese market, multiple regression algorithms were tested and in the context of this dissertation the one that proved to be superior was the XGBoost Regressor. This algorithm was chosen based on its superior performance in preliminary modeling phases, where it consistently delivered accurate predictions and maintained efficient training and scoring times, and by the inclusion of hyperparameter tuning, specifically using a combination of Randomized and Grid Search methods, further refined the XGBoost model, leading to an improvement in the Mean Absolute Percentage Error (MAPE) from to 26.08%.

Finally, regarding the features that most significantly impact house price predictions in the Portuguese real estate market and how do these features interact to influence model predictions, several explainability techniques were used in order to highlight which features significantly impact house price predictions and how they interact to influence model predictions. The XGBoost model, employing Gini importance, identified the percentage of foreign population as a pivotal factor, influencing property values significantly, along with

other critical features like "Bathrooms" and "Crimes registados pela polícia por mil habitantes", while the permutation importance technique further refined our understanding by emphasizing the dominant role of property "Area" in valuation, significantly affecting the model's error rate when altered. Meanwhile, SHAP analysis offered a granular view of how individual attributes such as the number of bathrooms and the proximity of stores positively affect house prices, while aspects like lower energy ratings negatively impact valuations. This analysis not only assists in highlighting key features that command house prices but also in understanding their interaction, thereby enhancing the predictive reliability and transparency of the model.

4.5. Final Considerations

The accuracy of the predicted results, while positive, are somewhat underwhelming, most likely due to the dataset's insufficient size and diversity to fully capture the intricate dynamics of the housing market, particularly in Portugal, where prices change quickly and speculatively. To produce better predictions, a more comprehensive and diverse dataset that captures these rapidly changing trends and oscillations is required.

In the future, it would be desirable to create and orchestrate a data pipeline that runs in real-time or near-real-time over time, built to extract data from housing sales platforms on a continual basis, conducting ETL to a SQL Server database or similar technology and regularly refreshing and increasing the dataset to improve its size and richness.

Investment in improved data collection and processing infrastructure would be required to handle the increasing data volume and velocity, ensuring that the dataset remains representative of current market conditions and is resilient enough to enable more advanced analytical models. This strategic update would increase the model's ability to predict market trends and price movements with more accuracy, helping stakeholders that rely on timely and precise market projections. On the other side, this improvement would result in greater infrastructure expenditures, not only to enable real-time web scraping advertisements and database integration, but also to cover the price of the Google Maps API for the enrichment of amenities via coordinates.

Furthermore, this data extraction over time would result in a dataset with a temporal component, which would incorporate time series analysis to better understand and predict trends over time, potentially including valuable insights such as tracking price fluctuations, seasonal variations, and economic cycles that affect the real estate market.

Furthermore, while the experiment concentrated on extracting property listings from a single platform, expanding data gathering to include multiple advertisement platforms has the potential to dramatically improve the dataset's comprehensiveness and accuracy. By aggregating data from multiple sources, the project can capture a broader spectrum of the market, including a wider range of property types, price points, and geographic locations, potentially mitigating biases and limitations inherent in single-source data collection and providing a more comprehensive view of market dynamics. While this can improve the model's predictive power, when expanding data collection efforts, it is critical to adhere to legal and ethical compliance, particularly when it comes to web scraping practices, to ensure that the expansion of data sources remains legal and ethical.

In the future, it would be desirable to extract and save the photographs posted in housing sale adverts in a data lake. This approach would allow for the use of machine learning models specifically designed to analyze these images and assess the visual appeal and potential desirability of the properties, by implementing models that can evaluate factors such as design aesthetics, room layout, and overall condition from images, giving stakeholders the opportunity to gain deeper insights into how these elements influence buyer preferences and pricing, enabling a more nuanced understanding.

The predictive model's adoption would also benefit greatly from the inclusion of a user-friendly interface for stakeholders, as well as automated alarm systems and feedback mechanisms. This interface would work as a central point for users, ranging from real estate brokers and investors to homebuyers, to engage with the model by entering specific criteria and receiving personalized predictions and insights. The automated alert system would improve this capability by informing consumers of significant market changes or new listings that match their preferences, ensuring they receive timely and relevant information. Furthermore, including a feedback mechanism would be critical, since it would allow users to report anomalies or make recommendations based on their real-world experiences and

outcomes. Together, these elements would not only make the model more accessible and responsive to stakeholder needs but would also encourage a more interactive and user-centered approach to real estate market analysis.

5. Conclusion

This dissertation demonstrated how machine learning models can forecast house prices in the Portuguese real estate market using web-scraped data from multiple sources. The comprehensive analysis made use of a variety of sophisticated machine learning techniques to handle, process, and model the complex dynamics of real estate pricing. The findings not only improve our understanding of the factors that influence house prices, but they also contribute to academic research by combining multiple data sources and advanced machine learning techniques.

The models tested demonstrated that machine learning can capture the effects of various property characteristics, socioeconomic factors, and geographic details on house prices. Among the methods used, the XGBoost regressor was the most effective, providing stakeholders with a tool for forecasting real estate values, which can be especially useful for investors, policymakers, and real estate professionals who need precise market analysis to make sound decisions.

Furthermore, the use of this model in a real-world scenario was discussed, with an emphasis on the practical implications, potential applications of this research, and future improvements that could be made. The methodology, findings, and considerations presented provide a road map for future work, including potential studies and improvements in predictive modeling in the real estate market.

6. Annex

6.1. Overview of the Tools Used

6.1.1. Python

Python is a popular programming language known for its simplicity, adaptability, and substantial library support. Python's open-source nature, combined with a strong community support system and continuous development efforts, has contributed to its evolution. Programmers have readily embraced this language due to its simplicity and readability, establishing it as one of the most widely used programming languages today. (Matthes, 2023).

Python's diverse uses are especially prevalent in data science and machine learning. Python's robust ecosystem of libraries and frameworks, such as NumPy, Pandas, and Matplotlib, provides powerful tools for data manipulation, analysis, and visualization. In addition to the integration with popular machine learning frameworks like scikit-learn and TensorFlow, has further strengthened Python's position in the data science realm, enabling the development and deployment of sophisticated predictive models. (McKinney, 2022) .

The advent of the Jupyter Notebook environment has significantly transformed scientific research, providing an interactive platform for data exploration, experimentation, and collaboration (Johnson, 2020).

6.1.2. SQL Server

SQL Server, developed by Microsoft, is a relational database management system (RDBMS) known for its robustness, scalability, and extensive feature set. Its continuous development and regular updates have contributed to its popularity in the database management space.

It is widely used for managing and storing structured data in various applications. The RDBMS's robustness, security features, and scalability make it suitable for enterprise-level data management with core functionalities such as data querying, manipulation, and transaction management enabling efficient data operations also supporting stored procedures, triggers and views for the final user (Mistry & Misner, 2014).

SQL Server's Business Intelligence (BI) capabilities have gained significant recognition in the industry. The Integration Services (SSIS) component enables Extract, Transform and Load (ETL) processes for data integration from diverse sources. SQL Server Analysis Services (SSAS) provides multidimensional and tabular models for data analysis and reporting. Additionally, SQL Server Reporting Services (SSRS) offers powerful reporting features, allowing the creation of interactive and visually appealing reports. SQL Server's integration with Power BI extends its BI capabilities, enabling advanced analytics and data visualization (Hancock & Toren, 2006).

6.2. Summary Statistics

	mean	std	min	0,25	0,5	0,75	max
Alojamentos familiares clássicos	69378.8	80645.2	1018.0	21 07 0.0	42 74 1.0	827 63. 0	324 554 .0
Alojamentos turísticos	81.1	131.7	0.0	11. 0	28. 0	61. 0	487 .0
Alunos do ensino não superior	21527.7	29151.0	3.0	38 56. 0	10 91 2.0	259 97. 0	115 294 .0
Alunos do ensino superior	13396.9	32948.1	0.0	0.0	0.0	522 4.0	122 554 .0
Bancos Caixas Económicas	57.6	110.9	1.0	8.0	18. 0	49. 0	436 .0
Beneficiários do Rendimento Social de Inserção (RSI)	3891.4	5868.3	22.0	48 0.0	12 81. 0	378 1.0	206 58. 0
Caixas automáticas multibanco	190.1	310.3	2.0	39. 0	86. 0	183 .0	124 0.0
Caixas de Crédito Agrícola Mútuo	3.3	2.4	0.0	1.0	3.0	5.0	16. 0
Casamentos	394.2	579.1	0.0	56. 0	19 1.0	380 .0	221 8.0
Consumo de energia eléctrica por habitante (kWh)	5065.9	4603.6	1863.9	32 02. 3	41 00. 4	513 6.9	841 92. 6

Crimes registados pelas polícias por mil habitantes	30.5	10.4	10.7	23.2	27.3	36.7	64.6
Densidade populacional (número médio de indivíduos por km2)	1224.9	1814.2	4.4	104.2	325.8	1756.9	7262.0
Desempregados inscritos nos centros de emprego	4840.3	6192.9	0.0	781.0	2448.0	5493.0	22812.0
Desempregados inscritos nos centros de emprego em % da população residente	5.8	2.3	0.0	5.0	6.0	7.0	15.0
Despesas da Câmara Municipal	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Despesas da Câmara Municipal em cultura e desporto (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Despesas do município em ambiente (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Diferença entre os nascimentos e os óbitos (saldo natural)	-443.3	587.9	-2295.0	-494.0	-237.0	-1117.0	1222.0
Divórcios	196.8	200.6	1.0	46.0	126.0	291.0	732.0
Ecrãs de cinema	10.4	17.1	0.0	0.0	5.0	14.0	66.0
Edifícios novos concluídos para habitação familiar	71.0	72.4	0.0	24.0	46.0	85.0	323.0
Estabelecimentos do 1 º ciclo do ensino básico	42.8	47.1	1.0	13.0	27.0	51.0	185.0
Estabelecimentos do 2 º ciclo do ensino básico	15.2	22.5	1.0	4.0	7.0	16.0	89.0

Estabelecimentos do 3 º ciclo do ensino básico	18.2	25.9	0.0	4.0	9.0	18.0	102.0
Estabelecimentos do ensino pré-escolar	63.4	71.9	1.0	15.0	37.0	81.0	282.0
Estabelecimentos do ensino secundário	13.0	20.5	0.0	2.0	5.0	11.0	77.0
Estabelecimentos do ensino superior	7.9	18.1	0.0	0.0	0.0	4.0	67.0
Farmácias	40.5	63.6	0.0	8.0	16.0	36.0	253.0
Ganho médio mensal dos trabalhadores por conta de outrem €	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hospitais	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Idosos (%) (65 e mais anos)	24.2	4.5	10.8	21.4	23.4	25.9	47.1
Jovens (%) (menos de 15 anos)	13.0	1.6	5.8	11.9	13.3	14.0	19.1
Museus	5.7	11.3	0.0	1.0	2.0	4.0	44.0
Nascimentos	1039.4	1358.9	7.0	209.0	580.0	1432.0	5398.0
Pensões da Caixa Geral de Aposentações	11808.5	23432.6	97.0	1316.0	3028.0	12730.0	93125.0
Pensões da Segurança Social (velhice invalidez e sobrevivência)	32728.5	38124.7	0.0	7647.0	1780 3.0	42506.0	148485.0
Pensões da Segurança Social e da CGA em % da população residente	39.4	7.0	0.0	35.0	38.0	44.0	66.0
Pessoal ao serviço nas empresas não financeiras	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Pessoal ao serviço nas quatro maiores	0.0	0.0	0.0	0.0	0.0	0.0	0.0

empresas do município (%)							
População em idade activa (%) (15 aos 64 anos)	62.8	3.3	45.6	61.3	62.9	64.8	70.4
População estrangeira	14231.1	27403.4	17.0	10	44	127	108
População estrangeira em % da população residente	9.0	8.3	0.5	94.0	60.0	18.0	653.0
População residente	121977.4	140180.4	1407.0	3.5	6.3	11.9	41.1
Receitas da Câmara Municipal	0.0	0.0	0.0	29	67	166	545
Resíduos urbanos recolhidos selectivamente por habitante (kg)	0.0	0.0	0.0	94	58	682	813
Saldo financeiro da Câmara Municipal (€ milhares)	0.0	0.0	0.0	3.0	2.0	.0	.0
Sessões de espectáculos ao vivo	591.6	1338.8	0.0	0.0	0.0	0.0	0.0
Superfície em km2	262.8	268.3	8.0	0.0	0.0	0.0	0.0
Taxa de mortalidade infantil (‰) (óbitos de crianças com menos de 1 ano de idade por cada 1000 nascimentos)	2.2	3.2	0.0	37.0	98.0	384.0	521 1.0
Trabalhadores da Administração Pública Local	1732.5	2396.3	57.0	97.0	16 8.0	319.0	172 1.0
Transferências recebidas no total das receitas	0.0	0.0	0.0	0.0	1.7	3.3	83.3
				44 1.0	96 3.0	195 1.0	978 6.0
				0.0	0.0	0.0	0.0

da Câmara Municipal (%)							
Volume de negócios das quatro maiores empresas do município (%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Índice de envelhecimento (idosos por cada 100 jovens)	194.3	72.1	56.0	15 4.0	17 5.0	217 .0	797 .0
Óbitos	1482.7	1913.8	25.0	38 0.0	80 9.0	171 1.0	769 3.0

7. References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. In *IEEE Transactions on Knowledge and Data Engineering* (Vol. 17, Issue 6, pp. 734–749). https://doi.org/10.1109/TKDE.2005.99

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301.

Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

Akash Dagar and Shreya Kapoor. (2020). A Comparative Study on House Price Prediction. *International Journal for Modern Trends in Science and Technology*, *6*(12), 103–107. https://doi.org/10.46501/ijmtst061220

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Breiman, L. (2001). *Random Forests* (Vol. 45).

Bricongne, J.-C., Meunier, B., & Pouget, S. (2021). *Web Scraping Housing Prices in Real-time: the Covid-19 Crisis in the UK*.

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1). Andriy Burkov Quebec City, QC, Canada.

Chaulagain, R. S., Pandey, S., Basnet, S. R., & Shakya, S. (2017). Cloud Based Web Scraping for Big Data Applications. *Proceedings - 2nd IEEE International Conference on Smart Cloud, SmartCloud 2017*, 138–143. https://doi.org/10.1109/SmartCloud.2017.28

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, T., & He, T. (2017). *xgboost: eXtreme Gradient Boosting*.

Clayton, J., Ling, D. C., & Naranjo, A. (2009). Commercial real estate valuation: Fundamentals versus investor sentiment. *Journal of Real Estate Finance and Economics*, *38*(1), 5–37. https://doi.org/10.1007/s11146-008-9130-6

Ekanayake, I. U., Meddage, D. P. P., & Rathnayake, U. (2022). A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, *16*, e01059.

Friedman, J. H. (2001). 999 REITZ LECTURE GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE 1. In *The Annals of Statistics* (Vol. 29, Issue 5).

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

Glaeser, E. L. (2013). A nation of gamblers: Real estate speculation and American history. *American Economic Review*, *103*(3), 1–42. https://doi.org/10.1257/aer.103.3.1

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., & Fdez-Riverola, F. (2013). Web scraping technologies in an API world. *Briefings in Bioinformatics*, *15*(5), 788–797. https://doi.org/10.1093/bib/bbt026

Hancock, J. C., & Toren, R. (2006). *Practical Business Intelligence with SQL Server 2005*. Pearson Education.

Hoerl, A. E., & Kennard, R. W. (1994). *Ridge Regression: Biased Estimation for Nonorthogonal Problems*.

Johnson, J. W. (2020). Benefits and pitfalls of jupyter notebooks in the classroom. *Proceedings of the 21st Annual Conference on Information Technology Education*, 32–37.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, *111*. https://doi.org/10.1016/j.landusepol.2020.104919

Kim, Y., Choi, S., & Yi, M. Y. (2020). Applying comparable sales method to the automated estimation of real estate prices. *Sustainability (Switzerland)*, *12*(14). https://doi.org/10.3390/su12145679

Kok, N., Koponen, E., & Adriana Martínez-barbosa, C. (2017). *Big Data in Real Estate? From Manual Appraisal to Automated Valuation*.

Krotov, V. (2018). *Legality and Ethics of Web Scraping*. https://www.researchgate.net/publication/324907302

L. Krause, A., & Bitter, C. (2012). Spatial econometrics, land values and sustainability: Trends in real estate valuation research. *Cities*, *29*(SUPPL.2). https://doi.org/10.1016/j.cities.2012.06.006

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Matthes, E. (2023). *Python crash course: A hands-on, project-based introduction to programming*. no starch press.

McKinney, W. (2022). *Python for data analysis*. " O'Reilly Media, Inc."

Misra, P., & Yadav, A. S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. *Int. J. Emerg. Technol*, *11*(3), 659–665.

Mistry, R., & Misner, S. (2014). *Introducing Microsoft SQL Server 2014*. Microsoft Press.

Mitchell, T. M. (1997). *Machine learning* (Vol. 1, Issue 9). McGraw-hill New York.

Mora-Garcia, R.-T., Cespedes-Lopez, M.-F., & Perez-Sanchez, V. R. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land*, *11*(11), 2100. https://doi.org/10.3390/land11112100

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering*, *48*, 500–506. https://doi.org/10.1016/j.proeng.2012.09.545

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: A review of valuation methods. In *Journal of Property Investment & Finance* (Vol. 21, Issue 4, pp. 383–401). https://doi.org/10.1108/14635780310483656

Sahu, M. (2021). Review on House Price Prediction using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *9*(5), 2178–2182. https://doi.org/10.22214/ijraset.2021.34804

Shahirah, N., Afar, J. ', Mohamad, J., & Ismail, S. (2021). MACHINE LEARNING FOR PROPERTY PRICE PREDICTION AND PRICE VALUATION: A SYSTEMATIC LITERATURE REVIEW. In *Journal of the Malaysian Institute of Planners* (Vol. 19).

Tavares, de O. F. A., Pereira, E. T., & Carrizo, M. A. (2014). The Portuguese residential real estate market: An evaluation of the last decade. *Panoeconomicus*, *61*(6), 739–757.

Varma, A., & Doshi, S. (2018). *House Price Prediction Using Machine Learning And Neural Networks*.

Venkatesh, B., & Anuradha, J. (2019). A review of feature selection and its methods. *Cybernetics and Information Technologies*, *19*(1), 3–26.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, *1*, 29–39.

Worzala, E., Lenk, M., & Silva, A. (1990). *Introduction THE JOURNAL OF REAL ESTATE RESEARCH 1 185 An Exploration of Neural Networks and Its Application to Real Estate Valuation*. Trippi and Turban.

Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 427–434. https://doi.org/10.1109/icdm.2003.1250949

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, *85*(11), 2541–2552.

Zhao, B. (2017). Web Scraping. In *Encyclopedia of Big Data* (pp. 1–3). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_483-1