NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Housing Market Segmentation in Portugal

Assessing Its Impact on Price Prediction

Ana Rita Pereira Viseu

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

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

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

HOUSING MARKET SEGMENTATION IN PORTUGAL

Assessing Its Impact on Price Prediction

by

Ana Rita Pereira Viseu

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Miguel de Castro Neto, PhD, NOVA Information Management School

Co-supervised by

Bruno Jardim, PhD, NOVA Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 15th July 2024

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to my supervisors, Professor Bruno Jardim and Professor Miguel de Castro Neto, for their invaluable guidance during this research. I would also like to thank the Confidencial Imobiliário team for providing the necessary data to achieve this work.

Secondly, I want to take this opportunity to thank my family, for I am incredibly grateful for the support I have received from them. This thesis is dedicated to my parents, without whom I could have never achieved it. I am forever thankful for all the love, patience, and encouragement you have continuously shown me throughout my life. Lastly, I would like to thank my dear friends for their companionship and understanding. Your friendship has made this journey even more enjoyable.

ABSTRACT

The real estate market is vital from an economic and social perspective. Predicting housing prices is necessary for buyers, real estate agents, economists, and policy makers, representing an active area of research. It is well recognized that housing market segmentation can significantly enhance the performance of housing price prediction. In this research, four different market segmentation methods are developed. Spatially contiguous submarkets are created a priori, based on the five geographic regions of Portugal, and with a clustering algorithm based on the latitude and longitude coordinates of the properties. Moreover, spatially noncontiguous submarkets are constructed using clustering techniques. One uses property characteristics as inputs and the other mixes property characteristics, latitude and longitude values, and sociodemographic data. Two prediction models, Linear Regression and LightGBM, are used to evaluate the performance of the overall housing market and the created submarkets. The results show that spatially noncontiguous submarkets improve Linear Regression results, although not very significantly. However, the best prediction accuracy is achieved by estimating a LightGBM model for the overall Portuguese housing market.

KEYWORDS

Real Estate; Housing Market Segmentation; Housing Submarkets; Housing Price Prediction; Machine Learning; Clustering

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
CI	Confidencial Imobiliário
CRISP-DM	Cross Industry Standard Process for Data Mining
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
EFB	Exclusive Feature Bundling
FCM	Fuzzy C-Means
GDP	Gross Domestic Product
GOSS	Gradient One-Side Sampling
GTWR	Geographically and Temporal Weighted Regression
HAC	Hierarchical Agglomerative Clustering
IQR	Inter-Quartile Range
KNN	K-Nearest Neighbors
LASSO	Least Absolute Shrinkage and Selection Operator
LightGBM	Light Gradient Boosting Machine
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
REDCAP	Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning

RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
SKATER	Spatial Kluster Analysis by Tree Edge Removal
SSE	Sum of Squared Errors
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

1.1. MOTIVATION

The housing market is crucial not only from an economic standpoint, but also from a social perspective (Samadani & Costa, 2021). It provides the fundamental human need for shelter and has evolved to act as a tool for long-term investment. It is therefore seen not only as a means of meeting physiological needs but also as a significant asset class for many individuals (Çılgın & Gökçen, 2023). Housing can influence the physical, psychological, and socio-cultural environment in which it is located, shaping the overall well-being of individuals and society (Çılgın & Gökçen, 2023). From an economic angle, the housing sector has emerged as a pivotal industry in many countries and regions (Fang et al., 2022). Changes in real estate markets have extensive implications on the economy and on financial stability (Rodrigues et al., 2022).

Particularly in Portugal, which is the subject matter of this study, the cycles of the housing market have shown a strong correlation with the business cycles observed in the last two decades (Rodrigues et al., 2022). Real estate is typically the most important asset in Portuguese households, having accounted for around 20% of the country's Gross Domestic Product (GDP) in 2017 (Rodrigues et al., 2022). In the recent past, Portugal has displayed a sharp rise in housing prices, with no signs of deceleration (Lourenço & Rodrigues, 2017). In 2022, the median price of family dwellings in Portugal stood at €1,484 per square meter, indicating a 14.4% increase from the preceding year (Instituto Nacional de Estatística, 2023a). More recently, in the second quarter of 2023, the median price of dwellings sales was €1,629/m2 (Instituto Nacional de Estatística, 2023b).

Given the utmost significance of the housing market, systems capable of providing accurate predictions for prospective property acquisitions are essential. The reliability of real estate investments is largely dependent on accurate pricing (Özöğür Akyüz et al., 2022). Predicting housing prices is crucial in assisting buyers, real estate agents, economists, policy makers, banks and other financial institutions make better decisions (Cekic et al., 2022). Hence, this area of research remains active, having attracted studies in the field of Artificial Intelligence (AI), most popularly in Machine Learning (ML) (Gude, 2023).

1.1.1. Housing Market Segmentation

More specifically, this topic has drawn attention into ways to improve housing price prediction. The complexity and the heterogeneity of the housing market often affects the modelling of prices (Usman et al., 2020). Consequently, the division of the housing market into multiple submarkets has been wildly applied to enhance the performance of prediction models (C. Wu et al., 2018). This delineation or disaggregation of the market into unique submarkets is referred to as housing market segmentation (Usman et al., 2020). Accounting for submarkets can improve price prediction by dis-aggravating problems like aggregation bias

and heteroscedasticity (Keskin, 2022; Usman et al., 2020). These can be incorporated as variables in the prediction models or by estimating submarket-specific models (Keskin, 2022). Despite the consensus on the existence of submarkets, there is no agreement on how to delineate them (Keskin, 2022). There is also a disagreement in literature on whether submarkets should be spatially contiguous, meaning, whether geographic proximity should be considered when defining submarkets.

Thus, this thesis focuses on the segmentation of the Portuguese housing market for improved price prediction. Most literature in housing market segmentation or in housing price prediction does not consider a whole country, often being limited to one city or region. Moreover, the housing market in Portugal lacks research in this topic. Therefore, it is worthy of investigating the existence of submarkets in this country and its impact on housing price prediction. The data used in this study, consisting of transaction records regarding residential property sales, is provided by Confidencial Imobiliário (CI). CI is an independent data bank, being one of the main sources of statistical data about real transaction prices in Portugal (Confidencial Imobiliário, 2024). With the mission of providing information for all market players, CI could improve its price prediction models by implementing housing market segmentation.

1.2. Objectives, Methodology and Contributions

The first objective of this study is to apply and compare different ways of delineating submarkets in the Portuguese housing market. Secondly, it aims to incorporate these submarkets in price prediction in two different ways: by developing predictive models for each submarket and by adding dummy variables representing them. Lastly, the study seeks to assess whether the accuracy of price prediction can be improved by the integration of housing submarkets. Additionally, it seeks to determine the optimal housing market segmentation method. With these objectives, the following questions are aimed to be answered:

1. Does the delineation of housing submarkets improve price prediction models?
2. Which housing market segmentation method improves model performance the most?

The data provided by CI, containing property characteristics, is enriched with census data from 2021 and economic data, as these are known to affect housing prices. Following the Cross Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al., 1999), this research consists of two main phases. In the first phase, different housing market segmentation methods are applied to define submarkets. In the second phase, two ML algorithms are used to build models for each submarket and for the aggregated market with and without submarket dummy variables. Lastly, the models' performances are compared to determine the optimal housing market segmentation and price prediction framework.

This research contributes to existing literature by exploring and comparing different housing market segmentation methods in Portugal and their respective effect on price prediction accuracy. Additionally, the existence of housing submarkets is investigated

considering a whole country, as opposed to a single region or city. Finally, the methodologies used, and the prediction models developed can provide CI with a solution to improve housing price prediction. By leveraging the findings of this study, CI can enhance its price prediction models, thus offering more accurate and effective pricing guidance to customers.

The remainder of this study is organized as follows. **Literature Review** discusses previous studies on housing market segmentation and housing price prediction. **Methodology** presents the housing market segmentation methods that were developed, as well as the price prediction models implemented, and the evaluation metrics utilized to measure their performance. **Results and Discussion** reports and discusses the results achieved. Finally, **Conclusions and Future Works** summarizes and concludes the work developed, addresses limitations of this research, and proposes ideas that could be developed in the future.

2. LITERATURE REVIEW

The literature review was conducted using Scopus as a primary source, supplemented by Google Scholar. Search terms such as “housing market segmentation”, “housing submarkets”, “housing submarkets price prediction”, “housing price prediction”, “real estate price prediction” or “real estate valuation” were utilized. The searches were limited between the years 2013 and 2023 to avoid older articles, and the results were sorted by relevance. Afterwards, some studies referenced within these articles were also included, regardless of publication year. Finally, to specifically address the context of the Portuguese housing market, the term “Portugal” was integrated into the existing search terms.

2.1. HOUSING MARKET SEGMENTATION

Housing market segmentation is the division of the housing market into uniquely distinct segments, known as submarkets, which are internally homogeneous and externally heterogeneous (Hu et al., 2022). Some authors define housing submarkets as spatially contiguous and stress the importance of geographic location. For example, Goodman & Thibodeau (2007) defined a housing submarket as a price constant geographic area. An alternative definition considers that properties of the same submarket are similar in terms of housing characteristics and different in comparison with characteristics of properties from other submarkets (Skovajsa, 2023). The discrepancy in the definition of submarkets also contributes to different ways of segmenting the housing market.

In general, housing market segmentation methods can be distinguished between a priori segmentation and data-driven segmentation (Usman et al., 2020). A priori segmentation is based on subjective and pre-defined criteria (Skovajsa, 2023), such as geographical boundaries, boundaries defined by real estate experts or based on certain characteristics, such as property type. In contrast, data-driven segmentation makes use of statistical tools to delineate submarkets, often being an application of clustering algorithms (Skovajsa, 2023).

2.1.1. A Priori Segmentation

A priori segmentation was the first method used in the earlier studies on housing market segmentation. The submarkets are defined using criteria that is determined by a prior view of what is important (Bourassa et al., 1999). Most studies that use this method argue that spatial proximity of properties plays a more significant role than physical housing characteristics. Therefore, most a priori segmentation methods are based on geographical boundaries, resulting in spatially contiguous submarkets. Spatially contiguous submarkets are geographically contiguous, whereas non-contiguous submarkets are accepted as geographically dispersed. Nevertheless, non-geographical criteria can also be considered. For example, UB & Saxena (2023) trained three separate ML models to predict the prices of three different types of property: apartment, villa, and independent house. The authors argued the existence of distinct price factors based on property type.

As for segmentation based on geographical boundaries, these can include submarkets based on postal codes, street segments, socioeconomic characteristics, or school districts. For instance, Basu & Thibodeau (1998) used major highways to divide the metropolitan area of Dallas, Texas to capture essential differences between housing submarkets. In a different study, Goodman & Thibodeau (2003) examined whether delineating submarkets improved hedonic estimates of housing prices in Dallas. Three approaches were employed: one using postal codes, one with census tracts and another using elementary school zones. The study found that all submarkets performed better than the aggregated market, with elementary school zones being the best method and postal codes the worst. Similarly, Feng & Jones (2016) evaluated two alternative approaches using Greater London housing sales. The first method was based on postal code areas and considers spatial contiguity. The second was based on census tracts, considering not only spatial contiguity, but also social homogeneity. Contrary to the previous study, it was revealed that housing market segmentation using postal codes performed better than census tracts in terms of model prediction accuracy, model fit and at capturing unexplained housing price variations. The authors argued that the different criteria, approaches, scales of analysis, data, and performance measures all contribute to inconsistent findings in literature.

Another approach was taken by Bangura & Lee (2019), who grouped local government areas into five major regions of Sydney, Australia, which were in turn regrouped into high-priced and low-priced submarkets. The study concluded that the two submarkets converge to the same type of submarket overtime. In another study, Alas (2020) compared the performance of three methods in Istanbul, Turkey. Ordinary Least Squares (OLS) regressions were estimated for submarkets based on districts, neighbourhoods, and geographic areas that used the same street for transportation. Each approach represented different sized submarkets. The last method yielded the best performance and results showed that the smaller the housing submarkets, the more OLS regression accuracy increases.

To determine the best way to incorporate a priori housing submarkets in price prediction, Usman et al. (2021) segmented the commercial property market in Malaysia by considering each district as a submarket. Then, OLS regressions were estimated for the aggregated market, for each submarket separately, and for the aggregated market with submarket dummy variables. Results showed that both ways of accounting for submarkets improved price prediction. The best approach was to model each submarket separately, which increased the R^2 Score by 7% and reduced the Root Mean Squared Error (RMSE) by 10.8%.

This type of housing market segmentation has been criticized mostly due to its subjectivity and arbitrariness, as well as its non-scalability (Skovajsa, 2023). Furthermore, these approaches do not necessarily reflect the underlying processes between buyers and sellers that generate housing submarkets, such as consumers' behaviours and preferences (Tomal & Helbich, 2022). For example, homebuyers do not automatically limit themselves to looking for similar nearby properties (M. Chen et al., 2023).

A summary of the studies described in this section can be seen in Table 2.1.

Table 2.1 – Overview of a priori housing market segmentation studies.

Study	Location and Sample Period	Methods	Results and Findings
Basu & Thibodeau (1998)	Dallas, Texas (1991:4 to 1993:1)	Major highways	There is spatial autocorrelation in prices within submarkets
Goodman & Thibodeau (2003)	Dallas, Texas (1995:1 to 1997:1)	Postal codes; Census tracts; Elementary school zones	All methods improved predictions (best method is elementary school zones)
Feng & Jones (2016)	London (2011 to 2014)	Postal codes; Census tracts	Postal codes were better than census tracts in model prediction
Bangura & Lee (2019)	Sydney, Australia (1991 to 2016)	Local government areas grouped into relative high-priced and low-priced submarkets	High-priced and low-priced submarkets converge to a single market overtime
Alas (2020)	Istanbul, Turkey (2016)	District boundaries; Neighbourhood boundaries; Areas that use same street for transportation	The smaller the submarkets, the more OLS regression prediction accuracy increased
Usman et al. (2021)	Kuala-Lumpur and Selangor State, Malaysia (2000 to 2012)	Districts (modeled as submarket dummies and separately for each submarket)	The best approach is a model for each submarket, which increased price prediction
UB & Saxena (2023)	Bangalore, India	Property type (apartment, independent house, villa)	XGBoost is best for apartments and independent house, Ridge Regression is best for villas

2.1.2. Data-driven Segmentation

To overcome the limitations of a priori segmentation, data-driven approaches were introduced, since they allowed the underlying structures of the data to determine housing submarkets empirically (J.-H. Chen et al., 2021; Usman et al., 2020). Data-driven segmentation can be based on various types of criteria, such as physical, neighborhood and spatial characteristics, or a combination of these (Usman et al., 2020). The generated submarkets typically achieve a higher degree of internal homogeneity and external heterogeneity (M. Chen et al., 2023) compared to a priori submarkets. These approaches are generally based upon statistical methods, such as cluster analysis or spatial statistics.

For example, Bourassa et al. (1999) applied a statistical method to define housing submarkets in Sydney and Melbourne, Australia. First, Principal Component Analysis (PCA)

was used to extract a reduced set of orthogonal factors. Then, scores were calculated for the most important factors, which were used as inputs for Hierarchical Agglomerative Clustering (HAC) and K-means clustering. Hedonic regressions were estimated for each city, for a priori submarkets and for submarkets defined by both clustering algorithms. Comparing the Mean Squared Error (MSE) of each model, all submarket classifications performed better than the aggregated market, with the data-driven submarkets being generally superior to the a priori ones. Wiersma et al. (2022) also used PCA as a preceding step for clustering analysis. The most important principal components were retained for 80 cities in Germany, which were used in a combination of hierarchical clustering using Ward's method and K-means to produce submarkets. Results revealed that the German residential market can best be segmented into four submarkets and, although geographic contiguity plays a role, it is not a main factor.

An innovative clustering approach was employed by Shi et al. (2015) to improve housing price prediction based on Fuzzy C-Means (FCM) clustering. As a result, ten submarkets were created using floor size, building date and basement size as inputs. The prediction models based on the submarkets yielded better results. Azimlu et al. (2021) conducted a comparative study of clustering algorithms to improve housing price prediction in King County, Washington. The data was clustered using K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and HAC. Then, various models were estimated for each submarket. It was found that HAC was not successful in improving accuracy compared to applying prediction models to the whole data, showing that not all clustering techniques are effective. Finally, DBSCAN enhanced prediction for certain models, while K-means was the most successful technique. The study concluded that taking advantage of different models for each submarket can reduce the prediction error. Alternatively, to deal with the heteroscedastic nature of housing data, Özöğür Akyüz et al. (2022) proposed a hybrid algorithm using a dataset from the Kadıköy district of Istanbul and another from Ames, Iowa. The algorithm involved applying Linear Regression to the training set and obtaining the prediction errors. Then, K-means clustering was applied to the residual errors to form clusters. Afterwards, K-Nearest Neighbors (KNN) was used on the training set to predict the class labels on the test set. Lastly, a series of ML algorithms were applied to each cluster separately. As a result, most hybrid models achieved a better performance than the standalone models.

The studies above did not consider any geographical boundaries, resulting in spatially fragmented housing submarkets. This ignores the age-old principle that location is very important. Thus, many studies used data-driven approaches that are spatially constrained. Hence, data-driven segmentation can be distinguished between approaches that consider spatial proximity, delineating spatially contiguous submarkets and the ones that do not. For instance, Bourassa et al. (2010) included geographic coordinates in hierarchical clustering. C. Wu & Sharma (2012) imposed spatial constraints by using census blocks as a base unit instead of individual properties. The study also incorporated relative location (accessibility to facilities) and absolute location (coordinates of census blocks' centroids) in PCA, followed by K-means clustering to delineate submarkets. As another example, Y. Wu et al. (2020) combined K-

means, HAC and DBSCAN to delineate housing submarkets in Salt Lake County, Utah. Property characteristics, socioeconomic variables, locational indicators, and residuals of an OLS regression model as inputs were used as inputs. The submarket-based models improved prediction accuracy.

However, these methods treat geographic features similarly to other non-spatial attributes, lacking flexibility to accommodate different weights between spatial and non-spatial features (M. Chen et al., 2023). Additionally, as the dimensionality of input variables increases, spatial attributes might have less weight (M. Chen et al., 2023). Therefore, specific spatial models have been used to impose spatial constraints. For example, Tomal & Helbich (2022) analysed housing submarkets in Cracow, Poland. Geographically and Temporal Weighted Regression (GTWR) was first applied, followed by PCA to reduce the dimensionality of the GTWR parameter estimates. Lastly, a clustering algorithm of the family of the Regionalization with Dynamically Constrained Agglomerative clustering and Partitioning (REDCAP) algorithm was used to partition the principal components into submarkets. It was revealed that the closer to the city center, the more submarkets were found. Sisman & Aydinoglu (2022) attempted to improve the performance of housing price prediction through spatially constrained clustering analysis. The study used socioeconomic variables related to the levels of income and education, age of the population, marital status, and socioeconomic development as inputs for the Spatial Kluster Analysis by Tree Edge Removal (SKATER) clustering algorithm. All prediction models except for one submarket were better than the aggregated market model in terms of Mean Absolute Error (MAE), MSE and RMSE. M. Chen et al. (2023) proposed a spatially constrained data-driven approach to identify spatial housing submarkets in Franklin County, Ohio using 19 years of transaction data. Hedonic regressions were estimated for the aggregated market, for a priori submarkets (school districts) and for three spatially constrained clustering algorithms: SKATER, REDCAP and ClustGeo. The results demonstrated the superiority of the REDCAP and ClustGeo algorithms for housing price prediction.

Although data-driven segmentation is more prominent in current research due to its objectivity (J.-H. Chen et al., 2021), there is no universal method considered superior (Skovajsa, 2023). There are also studies revealing the superiority of a priori housing submarkets. For example, Keskin (2022) compared three approaches to housing market segmentation in 32 districts of Istanbul. Multilevel models were applied to predict housing prices using a priori submarkets (administrative neighbourhood boundaries), submarkets based on real estate experts' delineation, and submarkets derived by HAC (using inputs like housing characteristics and environmental attributes, such as the quality of the neighbourhood). Results showed that models with a priori submarket dummy variables can predict prices more accurately than models with expert-based or data-driven submarkets.

Thus, the current literature shows insufficient agreement on which type of housing market segmentation should be used (Kryvobokov, 2023). Moreover, there is no single

definition of a housing submarket, also leading to a disagreement as to the relative importance of spatial and property features (Keskin, 2022) and whether submarkets should be spatially contiguous.

An overview of the studies described throughout this section can be found below in Table 2.2.

Table 2.2 – Overview of data-driven housing market segmentation studies.

Study	Location and Sample Period	Methods	Results and Findings
Not Spatially Constrained			
Bourassa et al. (1999)	Sydney and Melbourne, Australia (1991)	PCA (using socioeconomic variables) followed by HAC (using PCA scores as inputs)	In terms of MSE, all submarket classifications performed better than the aggregated market, with data-driven submarkets being superior to the a priori submarkets
Shi et al. (2015)	Louisville, Kentucky, USA (2003 to 2007)	FCM (using floor size, building date and basement size as inputs)	The prediction models based on the submarkets yielded better results
Azimlu et al. (2021)	King County, Washington, USA	Random smaller datasets; K-means; DBSCAN; HAC (using physical attributes, location and price as inputs)	Random smaller datasets increased the error in all models; Hierarchical clustering was not successful in improving accuracy; DBSCAN enhanced prediction for certain models; K-means was the most successful technique
Wiersma et al. (2022)	80 cities in Germany	PCA (using socioeconomic indicators) followed by Hierarchical clustering and K-means	German market can best be segmented into 4 submarkets and geographic contiguity plays a role, but it is not a main factor
Özöğür Akyüz et al. (2022)	Istanbul, Turkey (2014 to 2015); Ames, Iowa (2006 to 2010)	K-means (using Linear Regression residual errors as input)	Most hybrid models achieved a better performance than the standalone models
Keskin (2022)	Istanbul, Turkey (November 2006 to April 2007)	Administrative neighbourhood boundaries; Real estate experts' delineation; HAC (using physical and socioeconomic attributes)	Models with a priori submarket dummy variables predicted prices more accurately

Spatially Constrained			
Bourassa et al. (2010)	Louisville, Kentucky, USA (1999)	Hierarchical clustering (using physical attributes and coordinates)	Adding submarket dummies improved accuracy by 3.8% and increasing the number of submarkets improved results
C. Wu & Sharma (2012)	Milwaukee, Winsconsin, USA (2005)	PCA (using physical attributes, socioeconomic variables, and accessibility to facilities), followed by K-means, conducted at housing level and at the census block level (incorporating coordinates of blocks' centroids)	Spatially constrained submarket hedonic model was 22% more accurate in terms of Weighted MSE compared with the best performed a priori method
Y. Wu et al. (2020)	Salt Lake County, Utah (2011)	K-means, HAC and DBSCAN (using physical attributes, socioeconomic variables, distance to facilities and residuals of an OLS regression as input)	The submarket-based models improved prediction accuracy
Tomal & Helbich (2022)	Cracow, Poland (2020:1 to 2021:1)	GTWR (using physical attributes and distances to facilities) followed by PCA and Full-Order-CLK	COVID-19 led to an increase in the number of residential submarkets, especially in the city center
Sisman & Aydinoglu (2022)	Istanbul, Turkey (January 2019 to March 2020)	SKATER (using socioeconomic variables as input)	All models except for in one submarket were better than the aggregated market model in terms of MAE, MSE and RMSE
M. Chen et al. (2023)	Franklin County, Ohio (2001 to 2019)	School districts; SKATER; REDCAP; ClustGeo (using price, area, age and coordinates as inputs)	REDCAP and ClustGeo improved model performance the most

2.1.3. Housing Market Segmentation in Portugal

The existence of housing submarkets has been studied in certain regions of Portugal. For instance, Bhattacharjee et al. (2012) examined the role of space in housing markets, applying a spatial hedonic model to the city of Aveiro. The study was based on an a priori definition of housing submarkets. First, parishes were considered, then submarkets were constructed considering a combination of administrative boundaries, urban structure, and demographic features. In another study, Marques et al. (2012) presented six different approaches to define housing submarkets also in Aveiro. The first was an a priori method based on five criteria: administrative boundaries, urban structure, demographic and historical

characteristics, and urban development. The second method used the housing price per square meter. The third method was based on physical and location characteristics, which were reduced to five factors using PCA. The fourth and fifth approaches considered the implicit prices of housing resulting from the estimation of a hedonic model. In one case, the spatial structure of the housing market is considered, while in the other, physical and location characteristics are used. Finally, the last approach combines the five factors from PCA, coefficients of the regression model, and the price per square meter. In a different study, Bhattacharjee et al. (2016) started by reducing the number of physical housing attributes using factor analysis and estimated a hedonic model. Then, HAC with Ward's method using the living area and the estimated regression coefficients was applied to create six submarkets. A summary of the studies mentioned in this section can be seen in Table 2.3.

Despite the growing importance of housing market segmentation for more accurate price prediction, research on this topic within the context of Portugal remains scarce. Existing studies predominantly focus on a single city, failing to capture the nuances of the housing market dynamics at a national level. This emphasizes the need for comprehensive investigations into housing market segmentation methodologies across Portugal.

Table 2.3 – Overview of Portuguese housing market segmentation studies.

Study	Location and Sample Period	Methods	Results and Findings
Bhattacharjee et al. (2012)	Aveiro, Portugal (2000 to 2010)	Parishes; A priori criteria (administrative boundaries, urban structure, demographic features)	Predictions using the spatial model provided a lower MSE
Marques et al. (2012)	Aveiro, Portugal (2000 to 2010)	A priori approach; Price per square meter; PCA applied to physical and location characteristics; Regression coefficients; Combination of prior methods	The definition of housing submarkets increases the prediction accuracy of the estimated hedonic model
Bhattacharjee et al. (2016)	Aveiro, Portugal (2000 to 2010)	Factor analysis; Regression coefficients; HAC	Clustering based on housing characteristics and shadow prices produces substitutable submarkets

2.2. HOUSING PRICE PREDICTION

Hedonic pricing models are traditionally the most wildly applied methods in housing price prediction, relying mostly on Multiple Linear Regression (MLR) (Geerts et al., 2023). The prices of properties are modeled as linear combinations of property characteristics and other features. Among these methods, regularization techniques such as Ridge and Least Absolute Shrinkage and Selection Operator (LASSO) are commonly used (Çılgın & Gökçen, 2023; Sheng

& Yu, 2022). However, traditional methods do not capture nonlinear relationships between housing attributes and their price (Sheng & Yu, 2022). Thus, advanced approaches have been explored more recently, including ML and Deep Learning techniques. Besides, unlike traditional models, ML algorithms do not require the training data to be normally distributed (Yağmur et al., 2022). Nevertheless, Linear Regression is still commonly used as a baseline model. It is also the typical method to assess the performance of housing market segmentation in price prediction.

Popular ML algorithms in housing price prediction include models such as Decision Trees, Support Vector Machine (SVM) and Artificial Neural Networks (ANN). For instance, Yağmur et al. (2022) compared MLR, SVM and ANN to predict prices in Antalya, Turkey. ANN was the best model in terms of MSE, Mean Absolute Percentage Error (MAPE) and R^2 Score. Ensemble models such as Random Forest, Gradient Boosting and Extreme Gradient Boosting (XGBoost) are also commonly used and are known to achieve very good performances. For example, Sheng & Yu (2022) observed that XGBoost and Light Gradient Boosting Machine (LightGBM) had better performance compared to regression methods. Other ensembles have also been proposed by researchers. For instance, Alfaro-Navarro et al. (2020) used Decision Trees as base learners for bagging, boosting and Random Forest ensembles. Lakshmi et al. (2022) also proposed a series of ensemble models based on Linear Regression, KNN and Random Forest. Xiong et al. (2020) explored the performance of stacking models by combining different base learners, showing that stacking improves performance in terms of RMSE and MAE. Although not as frequently, time series models have also been employed to forecast housing prices. For example, the ARMAX model was used by Kaboudan & Sarkar (2008), which combined OLS regression with Autoregressive Moving Average for the residuals.

In terms of input data in housing price prediction models, studies generally include features related to the properties, socioeconomic features, environmental features (related to the quality of the neighborhoods) and spatial data (Geerts et al., 2023). Considering spatial data, geographic coordinates are commonly used, which can also be utilized to obtain features such as distances to or the presence of certain facilities. For example, Ecevit et al. (2022) extracted six spatial features from latitude and longitude values, concluding that they significantly improved prediction performance when added to the models. A summary of the studies mentioned in this section can be found below in Table 2.4.

Table 2.4 – Overview of housing price prediction studies.

Study	Location and Sample Period	Data	Methods	Results and Findings
Kaboudan & Sarkar (2008)	4 cities in Southern California, USA (2000 to 2004)	Physical attributes; Spatial attributes; Mortgage rates	OLS Regression; ARMAX	Neighborhood average price equations may be better than city aggregated equations

Alfaro-Navarro et al. (2020)	Spain (2018)	Physical attributes; Coordinates	Bagging and Boosting (using Decision Trees); Random Forest	Bagging and Random Forest had the best results
Xiong et al. (2020)	Melbourne (2015 to 2019); Victoria, Australia	Physical attributes; Distance to city center; Number of houses in the suburb; Median price of the suburb	LASSO and Elastic Net Regressions; Random Forest; Gradient Boosting; Extra Tree; XGBoost; Stacked models	Stacking improved prediction accuracy
Y. Chen et al. (2021)	Ames, Iowa, USA	Physical attributes	Linear Regression; Naïve Bayes; Backpropagation and Deep Neural Networks; SVM	Naïve Bayes, Backpropagation Neural Network and SVM were more suitable in price prediction
Ecevit et al. (2022)	Istanbul, Turkey	Physical attributes; Coordinates; Spatial features (like distance to nearest neighborhood)	Linear Regression; Decision Tree; Random Forest	Random Forest was the best model and spatial features significantly improve prediction
Lakshmi et al. (2022)	California, USA	Census data	Linear Regression; KNN; Random Forest; Ensembles (RF-LR, RF-KNN, RF-LR-KNN)	RF-LR-KNN had the best performance in terms of MSE, RMSE and MAE
Yağmur et al. (2022)	Antalya, Turkey (August 2022)	Physical attributes; Neighborhood; Presence of social facilities	MLR; SVM; ANN	ANN was the best model in terms of MSE, MAPE and R^2 Score
Sheng & Yu (2022)	Ames, Iowa, USA	Physical attributes	Ridge and LASSO Regressions; LightGBM; XGBoost	LightGBM and XGBoost had better performance in terms of RMSE and R^2 Score
Çılgın & Gökçen (2023)	Ankara, Turkey (June and July 2021)	Physical and locational (access to facilities) attributes	Linear, Ridge and LASSO Regressions; XGBoost; ANN	XGBoost was best in terms of MAE, MSE, RMSE and MAPE, followed by ANN

3. METHODOLOGY

Following the studies addressed in the previous section, this part explains the methods and techniques developed and employed in this study. The CRISP-DM methodology (Chapman et al., 1999) was followed, as it is the most popular and frequently used framework in data science (Saltz, 2021). This section starts by providing a description of the data used in this study (section 3.1). Then, the steps taken to prepare the data for modelling are presented (section 3.2). Afterwards, the different methods used to construct the submarkets are explained (section 3.3). Finally, section 3.4 describes the housing price prediction models developed, followed by section 3.5, which explains the metrics used to evaluate them. The sequence of these steps was not rigid, and it was often necessary to backtrack and repeat certain steps, which is typical of the cyclical nature of the CRISP-DM process model. An overview of the pipeline followed in this study is presented below in Figure 3.1.

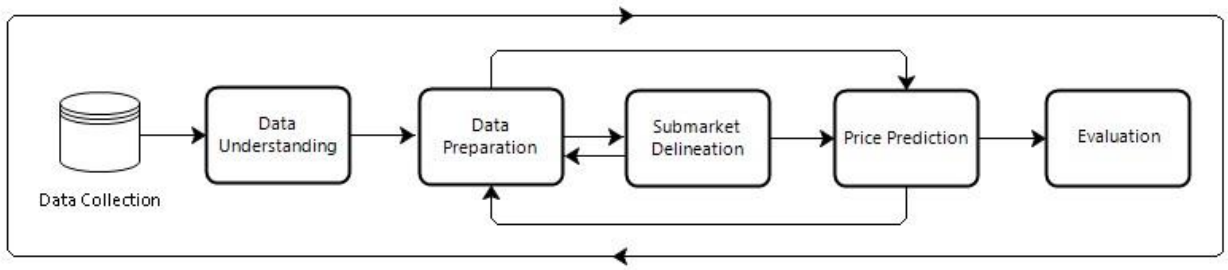

Figure 3.1 – Overview of the study's pipeline.

3.1. DATA AND STUDY AREA

The study was carried out using data collected from the different sources described in the next points. The study area covers 272 municipalities all over Portugal, excluding the autonomous regions of Azores and Madeira, throughout six years (2017-2022). With a population of 10.5 million inhabitants, Portugal is an interesting case study in terms of real estate, since it has experienced a significant increase in housing prices over the years, exhibiting very low levels of affordability (Fragoso Januário et al., 2023).

Table 3.1 displays the distribution of the sold properties per region and per year. More than half of the properties belong to the Metropolitan Area of Lisbon, followed by the north of Portugal, which contains around 21% of the total number of observations. As seen in Figure 3.2, the center of Portugal, as well as the Algarve and Alentejo regions, are less well represented. Naturally, most properties are located in the districts of Lisbon, Setúbal and Porto, with all other districts representing less than 7% of the total number of observations. In terms of selling price, the average price per square meter of a property across the six years of this study is approximately €2,013, with its standard deviation being €1,337. As seen in Figure 3.3, the Metropolitan Area of Lisbon is by far the region where the average selling price per square meter is higher, followed by the Algarve region. The center and the Alentejo

regions contain the least expensive properties. Throughout the years, it is apparent that the selling price has been rising all over Portugal, with the exceptions of the years 2019 and 2020, most likely due to the COVID-19 pandemic.

Table 3.1 – Distribution (%) of sold properties per selling year and per region.

Year	%	Region	%
2017	1.56	North	21.42
2018	15.66	Center	9.66
2019	23.18	Metropolitan Area of Lisbon	58.71
2020	23.09	Alentejo	4.17
2021	28.81	Algarve	6.04
2022	7.70		

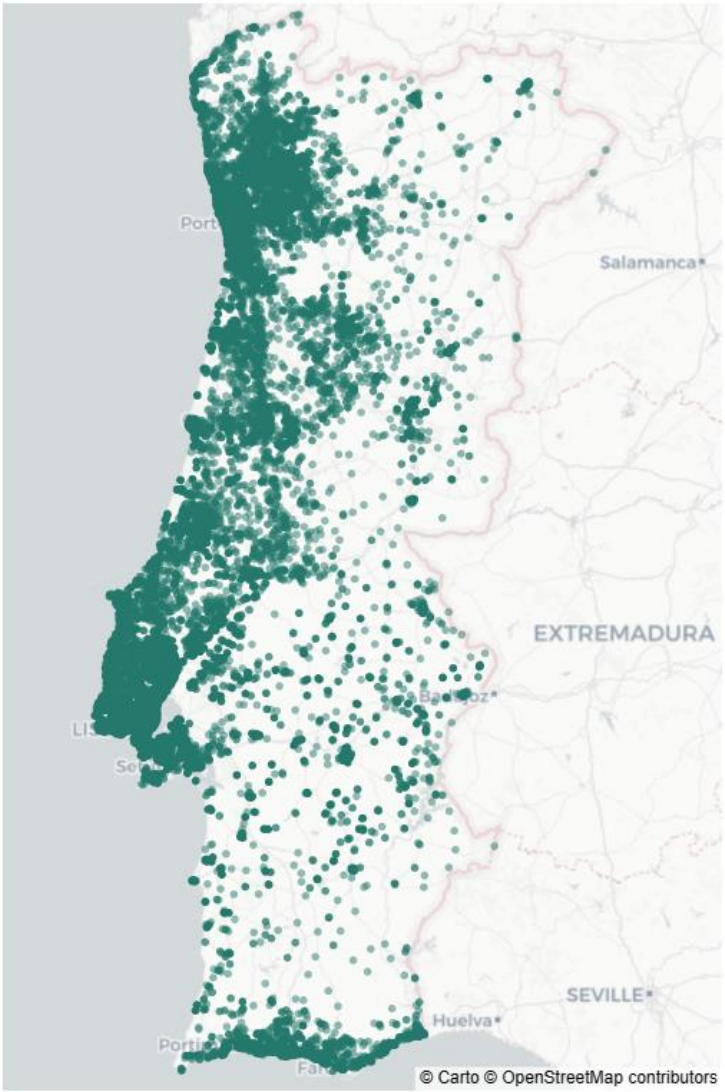

Figure 3.2 – Distribution of sold properties across Portugal.

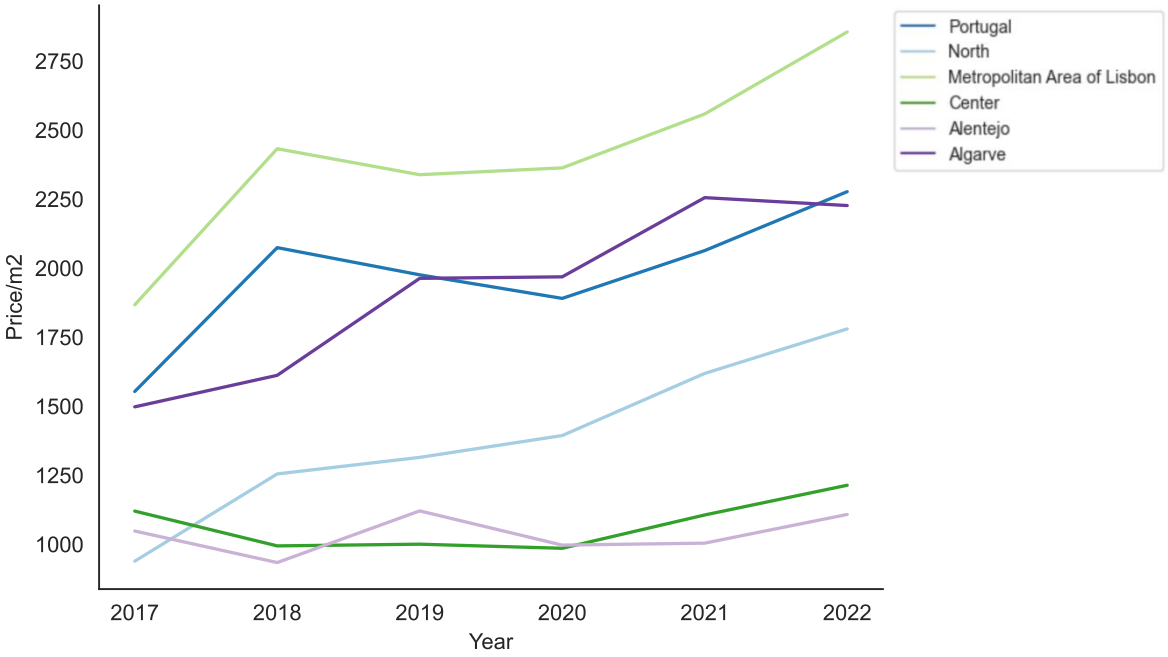

Figure 3.3 – Average price per square meter (in euros) per year for each region.

3.1.1. Property Data

The initial dataset, provided by CI, contains physical and transaction data. Physical data refers to information about the property itself, such as structural characteristics and information regarding its location. Transaction data corresponds to information involving the purchase, such as the selling price or the date of the transaction. Each row represents a residential property. From the variables present in the initial dataset, new variables were also extracted to allow a better analysis of the data or to potentially be included in the modelling phase. For instance, the year of the initial offer and the selling year were extracted from date variables. Furthermore, features representing the region, district and municipality of each property were added. A description of these features can be seen below in Table 3.2.

Table 3.2 – Description of property features.

Feature	Description
PropertyType	Indication if the property is an apartment (1) or a house (0)
NumberRooms	Number of rooms in the property
GPA	Gross private area of the property
BuildYear	Year when the property was built
Garage	Indication if the property has a garage (1) or not (0)
Terrace	Indication if the property has a terrace (1) or not (0)
Pool	Indication if the property has a pool (1) or not (0)
Patio	Indication if the property has a patio (1) or not (0)
ConservationState	Conservation state of the property (used or new)

ConservationStateValidation	Indication if the conservation state information was obtained through customers (0) or photographs (1)
EPC	Energy performance certificate on a scale from A+ to G
Latitude	Latitude of the postal code's centroid
Longitude	Longitude of the postal code's centroid
Parish	Parish where the property is located
Municipality	Municipality where the property is located
District	District where the property is located
Region	Region where the property is located
InitialOfferDate	Date when the property was first recorded on CI's database
LastRecordDate	Date when the property was last recorded on CI's database
InitialOfferYear	Year of the initial offer of the property
SellingYear	Year the property was sold (considered as the year extracted from the last record date)
TimeOnMarket	Number of days between the initial offer date and the last record date
Price	Observed selling price in €

3.1.2. Census Data

In order to enrich the dataset provided by CI, census data was collected from Instituto Nacional de Estatística (2022). The most recent census data corresponds to the first semester of 2021 and contains statistical indicators about demographic, socioeconomic and housing characteristics at a level of parish disaggregation. The indicators were selected based on the types of variables that have been utilized in previous studies on housing market segmentation, as well as housing price prediction. As a result, 25 new variables were added to the dataset, regarding information on the population, housing, employment, and educational qualifications of each parish. A brief explanation of these variables is in Table 3.3.

Table 3.3 – Description of census data.

Category	Features/Description
Population	Information regarding characteristics of the resident population (13 features)
Housing	Information regarding the residential properties and buildings (7 features)
Employment	Unemployment rate (1 feature)
Education	Educational qualifications of the resident population (3 features)
Foreign Population	Proportion of the resident population with foreign nationality (1 feature)

3.1.3. Economic Data

Economic data has also been widely used in previous studies, as housing prices and the state of the economy are deeply correlated. Therefore, yearly economic data was gathered from World Bank (2023) and PORDATA (2023). Consequently, four features were added, which are described in Table 3.4. These features were included in the dataset, corresponding to the year when the property was last recorded on CI's database, assuming it represents the selling year of each property.

Table 3.4 – Description of economic features.

Feature	Description
GDP	Gross Domestic Product (current LCU)
CPI	Consumer Price Index (2010=100)
Inflation	Inflation as measured by the Consumer Price Index
MortgageRate	Interest rate charged for a home loan

3.2. DATA PREPARATION

The collection of the described data resulted in a dataset containing 51 features. Once this data was aggregated, several preprocessing steps were taken to ensure that it was clean and ready for the subsequent modelling. Firstly, multiple properties did not have a selling price, meaning either they had not yet been sold, or the selling price was not observed. These observations were removed, since they would not be useful for modelling, leaving 119,191 rows. It was also observed that there were several observations with the same value for every variable. However, it was confirmed that these corresponded to different properties. This is possible because the latitude and longitude coordinates refer to the centroid of the postal code, so different properties can have the same values in these variables, as well as in all the remaining features. Thus, these observations were not removed. Afterwards, the data was checked for any incoherences, and data types were corrected. Then, using the hold-out method, the data was split into a training set (80%) and a test set (20%). This method was chosen because, considering the dataset was large, the resulting test error is likely a reliable estimate of the true error of the model (Berrar, 2019).

After that, missing data was imputed. Two of the variables were from the census data and, since they had very few missing values, the mean of the municipality where each parish belonged to was used to impute these values. The build year of the property contained around 86% of missing data in both the training and the test sets, so this variable was dropped. As for the remaining features, they consisted of four binary variables (Garage, Terrace, Pool, and Patio) and one ordinal variable (EPC). These features contained an extensive number of missing values, so different data imputation methods were experimented with, namely KNN Imputation and Iterative Imputation using Linear Regression and Random Forest. KNN

Imputation was chosen for all variables. Using the imputer trained on the training set, the missing values were also imputed on the test set.

Afterwards, the categorical variables were encoded. The conservation state of the properties contained four categories: used, new, good, and bad. The last two categories represented less than 5% of the data, so they were included in the first category. Thus, this feature became binary. Then, the region each property is in was one-hot encoded, resulting in four dummy variables. For the district and the municipality features, since these had a higher cardinality, all categories that consisted of less than 5% of the data were encoded as "Other". Next, both features were one-hot encoded as well. Lastly, to create the dependent variable, the total price of each property was divided by its gross private area (GPA). Therefore, the price per square meter was used as the target. However, this feature had a highly skewed distribution, as seen in Figure 3.4. Therefore, its natural logarithm was used as the dependent variable, since it was observed that it improved model performance.

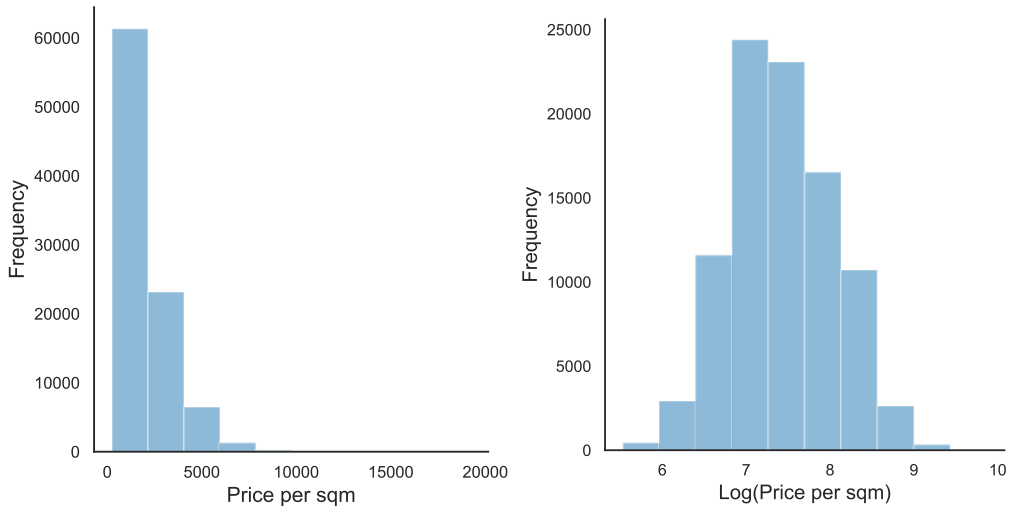

Figure 3.4 – Distribution of the price per square meter before and after log transformation.

The detection and removal of outliers is an essential task since they can skew the training of the models and result in less accurate predictions (Fernández et al., 2022). Multiple outlier detection methods were tested, such as manually removing outliers, Inter-Quartile Range (IQR), Z-score method, DBSCAN and Isolation Forest. Based on visualizations (boxplots and histograms) and on the amount of data removed by each method, DBSCAN was first chosen for outlier removal. Then, some remaining outliers were manually removed. As a result, 97.24% of the training data was kept, with 92,935 observations remaining.

Finally, unnecessary features, such as date variables, were removed. To further determine which variables to include in the prediction models, feature selection techniques were employed, since removing features of low importance improves accuracy and reduces model complexity. First, the Spearman correlations (Spearman, 1904) between all metric features and with the dependent variable were computed. For the pairs of explanatory variables that were highly correlated with each other, the one that had the lowest correlation

with the target was removed. For the categorical variables, the Kendall's correlation (Kendall, 1938) with the target was used, as well as the Analysis of Variance (ANOVA) test (Fisher, 1925). ANOVA is a statistical test which checks whether the means of two or more features are significantly different from each other (Radhika & Syed Masood, 2022). If they are significantly different, then the feature has an impact on the target and should be considered for modelling. Lastly, for all variables, Recursive Feature Elimination (RFE) (Guyon et al., 2002) with cross validation was used to identify the most important features. RFE is a feature selection algorithm that fits a model to the data, starting with all variables and repeatedly removing the less important ones (Mathew, 2019). The chosen model was Random Forest. Combining these methods, 33 features were selected as inputs for the prediction models, which are described in Table A.1 of Appendix A.

3.3. SUBMARKET DELINEATION

In this section, the construction of the housing submarkets is explained. According to the works previously mentioned in this study, different housing market segmentation methods were used, and submarkets were delineated under two different assumptions.

3.3.1. Spatially Contiguous Submarkets

Firstly, spatially contiguous submarkets were constructed, following the assumption that housing submarkets can be delineated by identifying geographically contiguous segments. Two types of housing market segmentation methods were utilized: a priori and data-driven segmentation. The a priori submarkets were constructed considering the five regions of Portugal: North, Center, Metropolitan Area of Lisbon, Alentejo and Algarve. On the other hand, data-driven submarkets were constructed using the K-means clustering algorithm (Macqueen, 1967), including the coordinates of properties (latitude and longitude) as inputs.

The K-means clustering algorithm is a partitional method and it is the most commonly used approach to evaluate the dissimilarity between two observations (Song et al., 2022). The algorithm requires the specification of the number of clusters, *K*, in advance. Once the number of clusters is specified, *K* data points $(c_1, c_2, \dots, c_k)$ of the dataset are selected as the initial centroids and each remaining observation is then assigned to the closest centroid, forming *K* initial clusters. The closest centroid of each observation was calculated based on the Euclidean distance, represented by the following equation:

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2} \tag{3.1}$$

Afterwards, the centroids of each cluster are recalculated by computing the mean value of the points in each cluster. Then, each observation is reassociated with its nearest centroid, forming new clusters. These two steps are repeated at each iteration. The algorithm stops when it reaches a specified maximum number of iterations or when the centroids of the

newly formed clusters cease to change. The main weaknesses of K-means clustering are that it is very sensitive to the initial centroids, and it requires the specification of the number of clusters in advance (Ikotun et al., 2023). As for the sensitivity to the initial centroids, this problem can be minimized by using multiple forms of initialization and to re-initialize the algorithm several times. To determine the optimal number of clusters, the Elbow Method was utilized. This method works by calculating the Sum of Squared Errors (SSE), defined by the equation below, for multiple numbers of clusters. The error decreases when the value of *K* increases (Syakur et al., 2018), and the optimal value of K corresponds to the point where the error decreases rapidly, creating an elbow shape.

$$SSE = \sum_{k=1}^{K} \sum_{x_i \in c_k} |x_i - c_k|^2 \tag{3.2}$$

3.3.2. Spatially Noncontiguous Submarkets

Contrary to the previous assumption, spatially noncontiguous submarkets were constructed following the assumption that housing submarkets do not have to be geographically contiguous, and thus, housing market segmentation can ignore the location of properties. Two methods were used to generate submarkets.

The first method used only property data as inputs, creating submarkets based on physical characteristics of the properties. The K-prototypes algorithm (Huang, 1998) was chosen, since it is capable of handling both numerical and categorical data. It is an extension of K-means, computing the Euclidean distance between numerical variables as well. However, the algorithm also calculates the distance between categorical variables using a dissimilarity measure based on the number of matching categories, defined as follows:

$$d(p,q) = \sum_i \delta(p_i, q_i) \ \ where \ \delta(p_i, q_i) = \begin{cases} 0 \ (p_i = q_i) \\ 1 \ (p_i \neq q_i) \end{cases} \tag{3.3}$$

The second method also used the K-means algorithm, mixing property, location and census variables as inputs, which were selected considering the Spearman correlations with the target. The final features used as inputs for both algorithms are presented in Table 3.5.

Table 3.5 – Features used as inputs for the clustering models.

K-prototypes	K-means
PropertyType; NumberRooms; EPC; GPA; Garage; Terrace; Pool, Patio; ConservationState; ConservationState Validation	GPA; Latitude; Longitude; AverageHouseholdSize; AverageMonthlyRent; DwellingDensity; AveragePurchaseMonthlyCharges; %HigherEducation; %Illiteracy; %ForeignPopulation

3.4. PRICE PREDICTION

Once the housing submarkets were constructed, predictive models were developed to evaluate the performance of the different housing market segmentation methods. Thus, multiple ML models were applied for five cases: the whole dataset, a priori submarkets, data-driven spatially contiguous submarkets, and two spatially noncontiguous submarkets (K-prototypes and K-means). The models were trained on each submarket separately, as well as with dummy variables representing each submarket. Since the dependent variable is continuous, two supervised regression models, which are explained in the next sections, were trained for each case. Each model was trained on the train dataset using 10-fold cross validation and was then applied on the test dataset. Once all models were trained, their hyperparameters were tuned using scikit-learn's function GridSearchCV (Scikit-learn, 2024). The best models were chosen based on the evaluation metrics. To compare all segmentation methods, the metrics of each submarket specific model were averaged.

3.4.1. Linear Regression

Linear Regression (Galton, 1889) is a simple algorithm, typically serving as a baseline for assessing the performance of housing market segmentation methods. It works under the assumption that there is a linear relationship between the dependent variable and the explanatory ones. Typically, property prices are modeled as a function of property, neighborhood and locational characteristics (Geerts et al., 2023), following the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{3.4}$$

Where y is the dependent variable, β_0 is the y-intercept (constant term), $\beta_1 \dots \beta_n$ are the regression coefficients, $x_1 \dots x_n$ are the independent variables, and ε is the error term. The most common regression method, used in this study, is the OLS regression, which aims to minimize the sum of squared residuals (residual is the difference between the observed and the predicted values), calculated as follows:

$$Sum\ of\ Squared\ Residuals = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.5}$$

However, Linear Regression requires a few assumptions:

- Linearity: there is a linear relationship between the predictors and the dependent variable.
- Normality: the residuals follow a normal distribution.
- No multicollinearity: the independent variables are not highly correlated with each other.
- Homoscedasticity: there is a similar variance of the error terms across the values of the independent variables.

3.4.2. LightGBM

LightGBM (Ke et al., 2017) is an ensemble method based on decision trees. It is an advanced implementation of the Gradient Boosting Decision Tree algorithm, specifically designed to enhance the efficiency of model training. Thus, this algorithm can often outperform other gradient boosting techniques in terms of computational speed and memory consumption.

At each iteration, LightGBM trains a new decision tree that fits the residuals from the previous tree. In the end, the results from each iteration are aggregated, achieving a stronger model. Each decision tree is built by splitting observations, using a histogram-based algorithm, where a histogram is created for each feature by dividing them into discrete bins. The best split is the one that results in the highest information gain. What sets LightGBM apart from other gradient boosting algorithms are its two optimization techniques: Gradient One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). The GOSS method ranks training instances according to the absolute value of their gradient (also known as residual error). If an instance is associated with a small gradient, the training error for this instance is small and it is already well-trained. On the contrary, if an instance has a larger gradient, the error associated with it is large so it will provide more information gain. Since discarding instances with small gradients would change the data distribution too much, GOSS keeps all instances with large gradients and performs random sampling on the instances with small gradients. To compensate for the influence on the data distribution, when calculating the information gain, the algorithm multiplies the training instances with small gradients by a constant. This way, GOSS allows the model to focus more on the under-trained instances without affecting the original data distribution too much. EFB is a novel method that effectively reduces the number of features. If a dataset has sparse features, these are usually mutually exclusive, meaning they do not have non-zero values simultaneously. EFB uses a greedy algorithm to combine these mutually exclusive features into a single feature, reducing the dimensionality.

3.5. EVALUATION METRICS

To assess the models' predictive power, the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), R^2 Score and Adjusted R^2 Score were utilized.

The MAE simply measures the average absolute difference between the actual and the predicted values, as defined in the equation below. The smaller its value, the better the performance of the model. An advantage of this metric is that it is more robust to outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.6}$$

The MAPE calculates the average percentage errors of predictions. Thus, a smaller MAPE indicates a better performance. This metric is calculated as follows:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3.7}$$

The RMSE is measured by the squared root of the squared difference between the actual and the predicted values. A smaller RMSE also indicates a better model performance. Contrary to the MAE, this metric is not as robust to outliers. The RMSE is defined by the equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.8}$$

R^2 Score is a statistical measure that indicates the proportion of variance of the dependent variable that can be predicted by an independent variable. In other words, it shows how well the target is explained by the model. The value of R^2 ranges from 0 to 1, with a value closer to 1 indicating a better model fit. It is represented by the following equation:

$$R^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y}_i)^2} \tag{3.9}$$

The major weakness of R^2 is that it keeps increasing when more variables are added to the model, even if redundant variables are being added. To deal with this problem, the Adjusted R^2 penalizes unnecessary variables, increasing only if the newly added features improve the model's performance. Similarly to R^2, a higher value signifies a better model fit. This metric is computed as follows:

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1} \tag{3.10}$$

Where:

$$y = observed\ value$$
$$\hat{y} = predicted\ value$$
$$n = number\ of\ data\ points$$
$$p = number\ of\ independent\ variables$$

4. RESULTS AND DISCUSSION

This section presents and discusses the results of this study. First, section 4.1 shows and analyses the submarkets resulting from each segmentation method. Then, in section 4.2, the performances of the prediction models are presented and discussed.

4.1. HOUSING MARKET SEGMENTATION RESULTS

The distributions of the spatially contiguous submarkets are presented in Figure 4.1. For the a priori submarkets, five groups were delineated as previously mentioned. For the K-means algorithm using the properties' coordinates as inputs, three clusters were constructed. Table 4.1 contains the number of training observations of each submarket, as well as its average price per square meter. In the a priori submarkets, the Metropolitan Area of Lisbon contains the highest number of properties, followed by the North. The remaining submarkets have a considerably lower number of observations. Regarding the average selling price, the Metropolitan Area of Lisbon has a significantly higher value, followed by the Algarve region, with a selling price near the average. The Center and the Alentejo submarkets have the lowest average prices. As for the second housing market segmentation, submarket 2 (dark green) is notably bigger and more expensive, which is natural, considering it comprises all of the Metropolitan Area of Lisbon. In contrast, submarket 1 (orange) contains a much lower number of properties. The third submarket (grey) has the lowest average price per square meter.

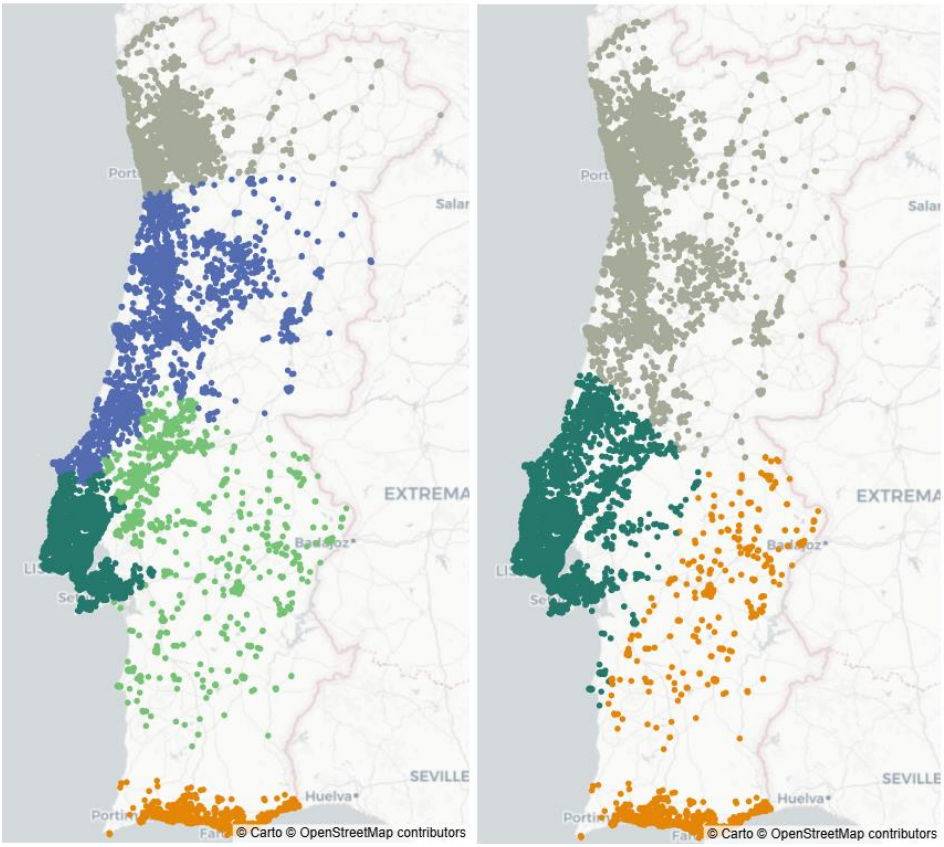

Figure 4.1 – Spatially contiguous submarkets: a priori (left) and K-means (right).

Table 4.1 – Number of training observations and average price per square meter of the spatially contiguous submarkets.

A priori			K-means		
Submarket	**Number of Observations**	**Average Price per Sqm**	**Submarket**	**Number of Observations**	**Average Price per Sqm**
Algarve (orange)	5522	2008	Submarket 1 (orange)	6757	1830
Alentejo (light green)	3435	1055	Submarket 2 (green)	60496	2347
Metropolitan Area of Lisbon (dark green)	55641	2456	Submarket 3 (grey)	25470	1378
Center (blue)	8506	1085			
North (grey)	19619	1471			

As for the noncontiguous submarkets, the K-prototypes algorithm with property characteristics resulted in four clusters, while the K-means algorithm using property characteristics, location and socioeconomic features resulted in three clusters. The geographic distributions of these submarkets are presented below in Figure 4.2 and the number of training observations and the average price per square meter of each submarket is displayed in Table 4.2. The submarkets created using the K-prototypes algorithm are not geographically clustered and are almost evenly distributed across the country. Submarkets 1 (dark green) and 3 (light green) have nearly double the number of observations of the other two submarkets. In terms of price per square meter, submarket 1 has the most expensive properties, while submarket 4 (grey) has the cheapest. Submarkets 2 (orange) and 3 have similar average prices. For the submarkets constructed with the K-means algorithm, since the location was included as an input, they are clearly geographically constrained, although not contiguous. Submarket 3 (grey) is the largest, while the remaining submarkets are nearly the same size. The second submarket (dark green) has a significantly higher average price per square meter than the rest.

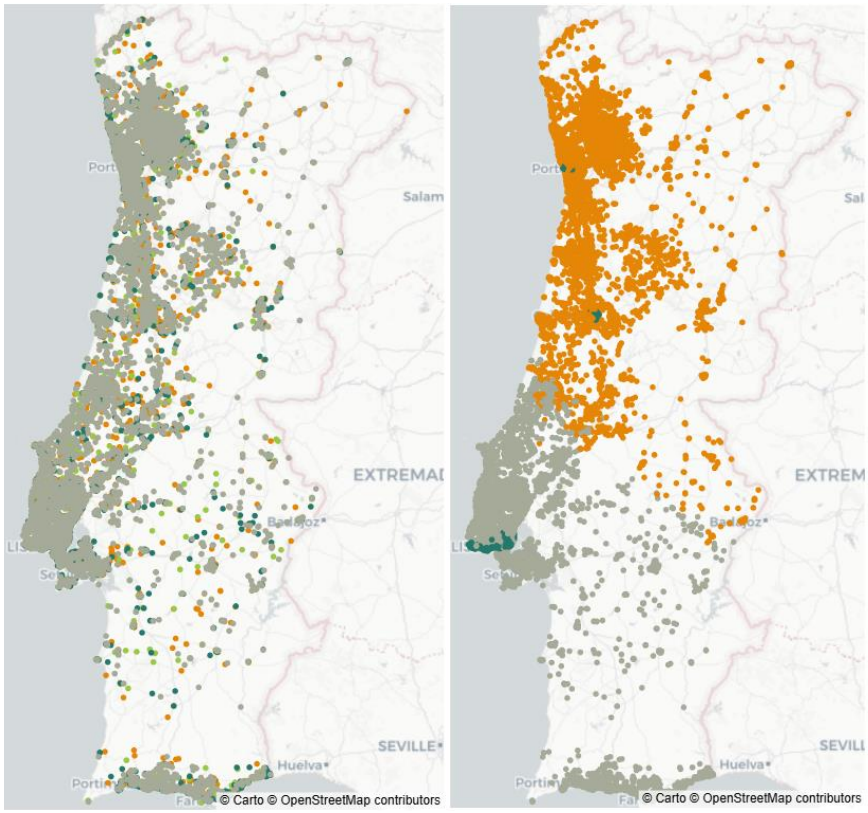

Figure 4.2 – Spatially noncontiguous submarkets: K-Prototypes (left) and K-Means (right).

Table 4.2 – Number of observations and average price per square meter of the spatially noncontiguous submarkets.

K-prototypes			K-means		
Submarket	**Number of Observations**	**Average Price per Sqm**	**Submarket**	**Number of Observations**	**Average Price per Sqm**
Submarket 1 (dark green)	31811	2284	Submarket 1 (orange)	23556	1210
Submarket 2 (orange)	17518	1947	Submarket 2 (dark green)	26664	3459
Submarket 3 (light green)	30938	2002	Submarket 3 (grey)	42503	1617
Submarket 4 (grey)	12456	1666			

4.2. Price Prediction Results

4.2.1. Comparison of Price Prediction in the Whole and Segmented Market

The first research question of this study is whether the delineation of housing submarkets improves price prediction. To answer this question, the evaluation metrics of the

whole market and of the market segmentation methods for both Linear Regression and LightGBM models are compared. These results are compiled in the following tables.

The evaluation metrics of the overall market are presented in Table 4.3. As observed, LightGBM has a higher accuracy in both the training and testing datasets, despite being more prone to overfitting, as seen by the difference between training and testing results.

Table 4.3 – Models' performances (train/validation/test) under the whole market.

Metric	**Linear Regression**			**LightGBM**		
	Train	Val	Test	Train	Val	Test
MAE	0.237	0.237	**0.239**	0.171	0.189	**0.189**
MAPE	0.032	0.032	**0.033**	0.023	0.026	**0.026**
RMSE	0.314	0.314	**0.318**	0.229	0.255	**0.256**
R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
Adjusted R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**

Table 4.4 and Table 4.5 contain the results of the averaged evaluation metrics of spatially contiguous and noncontiguous submarkets, respectively. For both prediction models, the a priori submarkets demonstrate worse prediction accuracy when compared to the overall market, as well as model fit. The average LightGBM model fit is even worse than the fit of Linear Regression on the overall market. This discrepancy of performance compared to the whole market is more evident in the LightGBM model, except in terms of R^2 and Adjusted R^2. These submarkets also display a higher tendency for overfitting when compared to the overall market. On the other hand, the submarkets constructed using the K-means algorithm with properties' coordinates as inputs are significantly better than the a priori submarkets in terms of price prediction. However, their performance is still worse than the overall housing market. Similarly to the a priori submarkets, the discrepancy of performance is more evident in the LightGBM model, except in terms of R^2 and Adjusted R^2. Contrary to the a priori segmentation, the models do not exhibit a higher tendency for overfitting.

As for the spatially noncontiguous submarkets, both methods show better accuracy than the overall market in the Linear Regression model, despite having a worse model fit. Yet, LightGBM showed worse performance in all evaluation metrics. In general, the K-means algorithm was slightly better than K-prototypes, except in terms of R^2 and Adjusted R^2, where K-prototypes reveals a better model fit.

Table 4.4 – Models' performances (train/validation/test) under spatially contiguous submarkets.

	Metric	**Linear Regression**			**LightGBM**		
		Train	Val	Test	Train	Val	Test
A priori Submarkets	MAE	0.238	0.240	**0.250**	0.194	0.208	**0.215**
	MAPE	0.034	0.034	**0.035**	0.027	0.029	**0.030**
	RMSE	0.314	0.315	**0.329**	0.257	0.277	**0.286**
	R^2	0.59	0.58	**0.55**	0.72	0.68	**0.66**
	Adjusted R^2	0.59	0.56	**0.54**	0.72	0.66	**0.65**
K-means	MAE	0.240	0.241	**0.248**	0.182	0.200	**0.203**
	MAPE	0.033	0.033	**0.034**	0.025	0.027	**0.028**
	RMSE	0.319	0.320	**0.327**	0.243	0.268	**0.273**
	R^2	0.65	0.64	**0.64**	0.79	0.75	**0.75**
	Adjusted R^2	0.64	0.63	**0.64**	0.79	0.74	**0.75**

Table 4.5 – Models' performances (train/validation/test) under spatially noncontiguous submarkets.

	Metric	**Linear Regression**			**LightGBM**		
		Train	Val	Test	Train	Val	Test
K-Prototypes	MAE	0.234	0.235	**0.236**	0.178	0.199	**0.199**
	MAPE	0.032	0.032	**0.032**	0.024	0.027	**0.028**
	RMSE	0.309	0.310	**0.312**	0.238	0.266	**0.266**
	R^2	0.73	0.73	**0.73**	0.84	0.80	**0.80**
	Adjusted R^2	0.73	0.72	**0.73**	0.84	0.80	**0.80**
K-means	MAE	0.226	0.226	**0.228**	0.184	0.195	**0.195**
	MAPE	0.031	0.031	**0.031**	0.025	0.027	**0.027**
	RMSE	0.300	0.301	**0.304**	0.245	0.261	**0.262**
	R^2	0.55	0.55	**0.55**	0.70	0.66	**0.66**
	Adjusted R^2	0.55	0.55	**0.54**	0.70	0.66	**0.66**

Additionally, the results of the models with submarket dummy variables for all segmentation methods are compiled in Table 4.6. In terms of the Linear Regression, adding dummy variables slightly improves model performance in all cases. However, the improvement is only apparent in some metrics, and it is not very significant. In the case of LightGBM, the results remain the same despite the addition of these variables.

Table 4.6 – Models' performances (train/validation/test) with submarket dummy variables.

	Metric	**Linear Regression**			**LightGBM**		
		Train	Val	Test	Train	Val	Test
A priori Submarkets	MAE	0.237	0.237	**0.238**	0.170	0.189	**0.189**
	MAPE	0.032	0.032	**0.033**	0.023	0.026	**0.026**
	RMSE	0.313	0.313	**0.317**	0.228	0.255	**0.256**
	R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
	Adjusted R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
Spatially Contiguous Submarkets (K-means)	MAE	0.237	0.237	**0.239**	0.170	0.189	**0.189**
	MAPE	0.032	0.032	**0.033**	0.023	0.026	**0.026**
	RMSE	0.314	0.314	**0.317**	0.228	0.255	**0.256**
	R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
	Adjusted R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
Spatially Noncontiguous Submarkets (K-Prototypes)	MAE	0.237	0.237	**0.239**	0.170	0.189	**0.189**
	MAPE	0.032	0.032	**0.033**	0.023	0.026	**0.026**
	RMSE	0.314	0.314	**0.317**	0.229	0.255	**0.256**
	R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
	Adjusted R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
Spatially Noncontiguous Submarkets (K-means)	MAE	0.237	0.237	**0.239**	0.170	0.189	**0.189**
	MAPE	0.032	0.032	**0.033**	0.023	0.026	**0.026**
	RMSE	0.314	0.314	**0.317**	0.228	0.255	**0.257**
	R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**
	Adjusted R^2	0.74	0.74	**0.74**	0.86	0.83	**0.83**

In terms of both prediction models, LightGBM always predicts properties' prices per square meter more accurately, confirming that Gradient Boosting algorithms generally achieve a better performance than Linear Regression models in housing price prediction. This result corroborates previous research such as Sheng & Yu (2022), Ecevit et al. (2022) and Çılgın & Gökçen (2023).

Overall, training a model for each submarket separately only improves price prediction when estimating Linear Regression models for spatially noncontiguous submarkets. Additionally, prediction results also improve when adding submarket dummy variables to a Linear Regression model, although very insignificantly. All other housing market segmentation methods display worse performance when compared to the whole market. For the LightGBM model, adding submarkets, whether by separately modelling them or by adding dummy variables, does not enhance prediction accuracy in any case. Separately applying a LightGBM model to each submarket provides worse results than the overall market for all four segmentation methods. In turn, adding dummy variables does not affect performance.

Considering the literature previously discussed in this study, these results are not in accordance with most of the referenced studies, where housing submarkets generally improve prediction accuracy. In Azimlu et al. (2021), although submarkets constructed with a K-means algorithm were successful in improving model performance, other clustering methods used in the study were not. For example, DBSCAN only displayed better performance for certain models, while hierarchical clustering did not improve any results. This shows that the chosen clustering algorithm has a strong influence on the performance of housing market segmentation, proving the need to experiment with different clustering techniques. Although K-means has been wildly used in previous literature, it does not provide improvements in this study, suggesting that the Portuguese housing market may not be heterogeneous enough to justify the delineation of submarkets for price prediction. However, it is important to consider that there are many other ways of delineating housing submarkets that could be explored. On the other hand, the fact that adding dummy variables slightly improves performance in Linear Regression but not in LightGBM suggests that more advanced ML models can already capture the differences and nuances of housing markets without needing to segment them, at least in the case of Portugal.

4.2.2. Comparison of Housing Market Segmentation Methods

The second research question is to determine which housing market segmentation method improves model performance the most. Even though the best prediction accuracy is achieved by applying a LightGBM model to the whole data, it is still relevant to compare the results of the different housing market segmentation methods, as it can provide a reference for future studies on the segmentation of the Portuguese housing market.

In terms of adding submarket dummy variables, the results are overall the same for every housing market segmentation method. However, separately modelling submarkets provides distinct results depending on the segmentation technique. The method that displays the worst model performance is the a priori segmentation, followed by the K-means algorithm with coordinates as inputs. Thus, results show that spatially noncontiguous submarkets are superior to the contiguous ones. The best algorithm is the K-means algorithm with property, location and socioeconomic features. This suggests that the best way to segment the Portuguese housing market might be by considering some kind of spatial constraint. The submarkets do not need to be spatially contiguous, but the location of properties should be considered when segmenting the market.

Further analysing the performance of each submarket specific model also provides some insights into other possible ways to segment the Portuguese housing market. For example, all a priori submarkets have worse results than the overall market, except for the Metropolitan Area of Lisbon. Table 4.7 displays the evaluation metrics for this area, where the results are better than when applying models to the whole data (Table 4.3), suggesting it could be beneficial to estimate a separate model for the Metropolitan Area of Lisbon and another one for the remaining areas of the country. For the K-means algorithm using coordinates as

inputs, a similar conclusion can be reached, as seen by the results of submarket 2 in Table 4.8. However, the performance is lower for this submarket compared to the Metropolitan Area of Lisbon in the a priori segmentation.

Table 4.7 – Models' performances (train/validation/test) for the Metropolitan Area of Lisbon.

Metric	**Linear Regression**			**LightGBM**		
	Train	Val	Test	Train	Val	Test
MAE	0.220	0.220	**0.216**	0.162	0.181	**0.177**
MAPE	0.029	0.029	**0.029**	0.021	0.024	**0.023**
RMSE	0.293	0.293	**0.290**	0.219	0.246	**0.242**
R^2	0.73	0.73	**0.74**	0.85	0.81	**0.82**
Adjusted R^2	0.73	0.73	**0.74**	0.85	0.81	**0.82**

Table 4.8 – Models' performances (train/validation/test) for submarket 2 of the spatially contiguous K-means segmentation.

Metric	**Linear Regression**			**LightGBM**		
	Train	Val	Test	Train	Val	Test
MAE	0.227	0.227	**0.224**	0.160	0.183	**0.181**
MAPE	0.030	0.030	**0.030**	0.021	0.024	**0.024**
RMSE	0.301	0.301	**0.299**	0.216	0.249	**0.246**
R^2	0.75	0.75	**0.75**	0.87	0.83	**0.83**
Adjusted R^2	0.75	0.75	**0.75**	0.87	0.□	**0.83**

5. CONCLUSIONS AND FUTURE WORKS

Throughout this research, the effect of housing market segmentation on price prediction accuracy was explored. The Portuguese housing market was segmented in four different ways, each representing a distinct type of housing market segmentation typically found in literature. Thus, two spatially contiguous and two noncontiguous groups of submarkets were delineated. The spatially contiguous submarkets were created a priori, based on the five geographic regions of Portugal, and with a clustering algorithm based on the latitude and longitude coordinates of the properties. In turn, the spatially noncontiguous submarkets were both constructed using clustering techniques. One used property characteristics as inputs and the other mixed property characteristics, latitude and longitude values, and sociodemographic data. The resulting submarkets were then incorporated into price prediction by estimating models for each segment separately and by estimating one model for the whole data containing dummy variables representing each submarket. Two models were used to predict the price per square meter of each property: Linear Regression and LightGBM. The performances of these housing market segmentation methods were then assessed by comparing different evaluation metrics with the results of a model estimated for the whole market, not taking any segments into consideration.

The results of this study have identified that spatially noncontiguous submarkets improve price prediction accuracy when estimating Linear Regression models. However, for the LightGBM model, the delineation of submarkets does not improve price prediction. These findings suggest that more advanced ML algorithms can already capture the differences found within the Portuguese housing market. Nevertheless, results show that spatially noncontiguous submarkets are more beneficial for the task of predicting housing prices compared to spatially contiguous submarkets. The results also demonstrate that when estimating a model for the Metropolitan Area of Lisbon alone, the performance is superior when compared to the entire market. This suggests that a separate model for the Metropolitan Area of Lisbon could be a way of improving price prediction. Thus, this research also provides insights into other possible ways of segmenting the Portuguese housing market.

However, this study has limitations that may have affected the results achieved. Firstly, the data that was used lacks information regarding the characteristics of properties. Having more variables that provide information about the property itself could not only allow for a better housing market segmentation, but also improve model performance. Some of the information was also not very specific, such as the properties' coordinates, which correspond to the centroid of the postal codes of each area. Having access to the exact location of the properties could provide a more rigorous segmentation. Moreover, a few of the variables regarding property characteristics were very unbalanced, not allowing the clustering algorithms to differentiate submarkets as well. Additionally, some regions of the country, mainly Algarve and Alentejo, had fewer observations, not being very well represented.

For future research, there is a need to experiment with other types of clustering algorithms, as this has a strong impact on the resulting submarkets and, consequently, on their performance in price prediction. Moreover, using different variables for submarket construction could also originate new results and findings. It would also be beneficial to extend the data that was used in this study and incorporate more information on property characteristics or take advantage of the location of the properties and add data regarding the proximity of certain amenities, such as schools or public transportation. Finally, another area of interest would be to take advantage of time series data and explore if the influence of housing market segmentation on price prediction is robust in time.

BIBLIOGRAPHICAL REFERENCES

Alas, B. (2020). A multilevel analysis of housing submarkets defined by the municipal boundaries and by the street connections in the metropolitan area: Istanbul. *Journal of Housing and the Built Environment*, *35*(4), 1201–1217. https://doi.org/10.1007/s10901-020-09735-7

Alfaro-Navarro, J.-L., Cano, E. L., Alfaro-Cortés, E., García, N., Gámez, M., & Larraz, B. (2020). A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems. *Complexity*, *2020*(1), 1–12. https://doi.org/10.1155/2020/5287263

Azimlu, F., Rahnamayan, S., & Makrehchi, M. (2021). House Price Prediction Using Clustering and Genetic Programming along with Conducting a Comparative Study. *GECCO 2021: Proceedings of the 2021 Genetic and Evolutionary Computation Conference Companion*, 1809–1816. https://doi.org/10.1145/3449726.3463141

Bangura, M., & Lee, C. L. (2019). House price diffusion of housing submarkets in Greater Sydney. *Housing Studies*, *35*(6), 1110–1141. https://doi.org/10.1080/02673037.2019.1648772

Basu, S., & Thibodeau, T. G. (1998). Analysis of Spatial Autocorrelation in House Prices. *Journal of Real Estate Finance and Economics*, *17*(1), 61–85. https://doi.org/https://doi.org/10.1023/A:1007703229507

Berrar, D. (2019). Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vol. 1, pp. 542–545). https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bhattacharjee, A., Castro, E., Maiti, T., & Marques, J. (2016). Endogenous Spatial Regression and Delineation of Submarkets: A New Framework with Application to Housing Markets. *Journal of Applied Econometrics*, *31*(1), 32–57. https://doi.org/10.1002/jae.2478

Bhattacharjee, A., Castro, E., & Marques, J. (2012). Spatial Interactions in Hedonic Pricing Models: The Urban Housing Market of Aveiro, Portugal. *Spatial Economic Analysis*, *7*(1), 133–167. https://doi.org/10.1080/17421772.2011.647058

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods. *Journal of Real Estate Research*, *32*(2), 139–160. https://doi.org/10.1080/10835547.2010.12091276

Bourassa, S. C., Hamelink, F., Hoesli, M., & Macgregor, B. D. (1999). Defining Housing Submarkets. *Journal of Housing Economics*, *8*(2), 160–183. https://doi.org/https://doi.org/10.1006/jhec.1999.0246

Cekic, M., Korkmaz, K. N., Müküs, H., Hameed, A. A., Jamil, A., & Soleimani, F. (2022). Artificial Intelligence Approach for Modeling House Price Prediction. *2022 2nd International Conference on Computing and Machine Intelligence (ICMI)*, 1–5. https://doi.org/10.1109/ICMI55296.2022.9873784

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T., & Wirth, R. (1999). *The CRISP-DM Process Model*.

Chen, J.-H., Ji, T., Su, M.-C., Wei, H.-H., Azzizi, V. T., & Hsu, S.-C. (2021). Swarm-inspired data-driven approach for housing market segmentation: a case study of Taipei city. *Journal of Housing and the Built Environment*, *36*(4), 1787–1811. https://doi.org/10.1007/s10901-021-09824-1

Chen, M., Chun, Y., & Griffith, D. A. (2023). Delineating Housing Submarkets Using Space–Time House Sales Data: Spatially Constrained Data-Driven Approaches. *Journal of Risk and Financial Management*, *16*(6), 291. https://doi.org/10.3390/jrfm16060291

Chen, Y., Xue, R., & Zhang, Y. (2021). House price prediction based on machine learning and deep learning methods. *2021 International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 699–702. https://doi.org/10.1109/EIECS53707.2021.9587907

Çılgın, C., & Gökçen, H. (2023). Machine learning methods for prediction real estate sales prices in Turkey. *Revista de La Construcción*, *22*(1), 163–177. https://doi.org/10.7764/RDLC.22.1.163

Confidencial Imobiliário. (2024). *Quem Somos*. https://www.confidencialimobiliario.com/en/quem-somos/

Ecevit, M. İ., Erdem, Z., & Dağ, H. (2022). Reviewing the Effects of Spatial Features on Price Prediction for Real Estate Market: Istanbul Case. *2022 7th International Conference on Computer Science and Engineering (UBMK)*, 490–495. https://doi.org/10.1109/UBMK55850.2022.9919540

Fang, Y., Li, T., & Zhao, H. (2022). Random Forest Model for the House Price Forecasting. *2022 14th International Conference on Computer Research and Development (ICCRD)*, 140–143. https://doi.org/10.1109/ICCRD54409.2022.9730190

Feng, Y., & Jones, K. (2016). Comparing two neighbourhood classifications: A multilevel analysis of London property prices 2011-2014. *Pacific Rim Real Estate Conference*. https://www.researchgate.net/publication/298896834

Fernández, Á., Bella, J., & Dorronsoro, J. R. (2022). Supervised outlier detection for classification and regression. *Neurocomputing*, *486*, 77–92. https://doi.org/10.1016/j.neucom.2022.02.047

Fisher, R. A. (1925). *Statistical Methods for Research Workers* (11th ed.). Oliver and Boyd.

Fragoso Januário, J., Oliveira Cruz, C., Varum, H., & Faria e Sousa, V. (2023). Is housing becoming less affordable? A study of affordability in the Portuguese housing market. *Property Management*, *41*(5), 698–728. https://doi.org/10.1108/PM-08-2022-0059

Galton, F. (1889). *Natural Inheritance*. Macmillan.

Geerts, M., vanden Broucke, S., & De Weerdt, J. (2023). A Survey of Methods and Input Data Types for House Price Prediction. *ISPRS International Journal of Geo-Information*, *12*(5), 200. https://doi.org/10.3390/ijgi12050200

Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics*, *12*(3), 181–201. https://doi.org/10.1016/S1051-1377(03)00031-7

Goodman, A. C., & Thibodeau, T. G. (2007). The Spatial Proximity of Metropolitan Area Housing Submarkets. *Real Estate Economics*, *35*(2), 209–232. https://doi.org/10.1111/j.1540-6229.2007.00188.x

Gude, V. (2023). A multi-level modeling approach for predicting real-estate dynamics. *International Journal of Housing Markets and Analysis*, *17*(1), 48–59. https://doi.org/10.1108/IJHMA-02-2023-0024

Guyon, I., Weston, J., & Barnhill, S. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, *46*, 389–422. https://doi.org/https://doi.org/10.1023/A:1012487302797

Hu, L., He, S., & Su, S. (2022). A novel approach to examining urban housing market segmentation: Comparing the dynamics between sales submarkets and rental submarkets. *Computers, Environment and Urban Systems*, *94*. https://doi.org/10.1016/j.compenvurbsys.2022.101775

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://doi.org/https://doi.org/10.1023/A:1009769707641

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Instituto Nacional de Estatística. (2022). *Censos 2021*. https://tabulador.ine.pt/censos2021/

Instituto Nacional de Estatística. (2023a). *Estatísticas da Construção e Habitação - 2022*. https://www.ine.pt/xurl/pub/280978640

Instituto Nacional de Estatística. (2023b). *House prices deceleration in 17 of the 24 most populous municipalities - 2nd Quarter 2023 [Press release]*. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=593987801&DESTAQUESmodo=2&xlang=en

Kaboudan, M., & Sarkar, A. (2008). Forecasting prices of single family homes using GIS-defined neighborhoods. *Journal of Geographical Systems*, *10*(1), 23–45. https://doi.org/10.1007/s10109-007-0054-0

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 3146–3154. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/https://doi.org/10.1093/biomet/30.1-2.81

Keskin, B. (2022). Multilevel approach to the analysis of housing submarkets. *Regional Studies, Regional Science*, *9*(1), 264–279. https://doi.org/10.1080/21681376.2022.2067005

Kryvobokov, M. (2023). Housing rental submarkets in hedonic regression: econometric arguments and practical application. *Journal of Housing and the Built Environment*, *38*(2), 951–978. https://doi.org/10.1007/s10901-022-09972-y

Lakshmi, K., Narayana, P. V., Bhavani, P. D., Madhavacharyulu, V. V. S., Lavanya, N., & Pousia, S. (2022). An Enhanced Regression Technique for House Price Prediction. *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 457–462. https://doi.org/10.1109/IC3I56241.2022.10072476

Lourenço, R. F., & Rodrigues, P. M. M. (2017). House prices in Portugal - what happened since the crisis? *Banco de Portugal Economic Studies*, *3*(4), 41–57. https://www.bportugal.pt/sites/default/files/anexos/papers/re201713_e.pdf

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

Marques, J. L., Castro, E., Bhattacharjee, A., & Batista, P. (2012). Spatial heterogeneity across housing sub-markets in an urban area of Portugal. *52nd Congress of the European Regional Science Association*. https://hdl.handle.net/10419/120764

Mathew, T. E. (2019). A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. *International Journal on Emerging Technologies*, *10*(3), 55–63.

Özöğür Akyüz, S., Eygi Erdogan, B., Yıldız, Ö., & Karadayı Ataş, P. (2022). A Novel Hybrid House Price Prediction Model. *Computational Economics*, *62*(3), 1215–1232. https://doi.org/10.1007/s10614-022-10298-8

PORDATA. (2023). *Interest rates on new loan operations (annual average) to individuals: total and by purpose type*. https://www.pordata.pt/en/Portugal/Interest+rates+on+new+loan+operations+(annual+average)+to+individuals+total+and+by+purpose+type-2845

Radhika, A., & Syed Masood, M. (2022). Crop Yield Prediction by Integrating Et-DP Dimensionality Reduction and ABP-XGBOOST Technique. *Journal of Internet Services and Information Security (JISIS)*, *12*(4), 177–196. https://doi.org/10.58346/JISIS.2022.I4.013

Rodrigues, P. M. M., Aguiar-Conraria, L., Barros, V. G., Batista, P., Brinca, P., Castro, E. A., Duarte, J. B., Gonçalves, D., Huget, R., Lourenço, R. F., Marques, J. L., Peralta, S., Reis, V., Santos, J. P. dos, & Soares, M. J. (2022). *The real estate market in Portugal: Prices, rents, tourism and accessibility*. Fundação Francisco Manuel dos Santos.

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*, 2337–2344. https://doi.org/10.1109/BigData52589.2021.9671634

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

Scikit-learn. (2024). *GridSearchCV*. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Sheng, C., & Yu, H. (2022). An optimized prediction algorithm based on XGBoost. *2022 International Conference on Networking and Network Applications (NaNA)*, 1–6. https://doi.org/10.1109/NaNA56854.2022.00082

Shi, D., Guan, J., Zurada, J., & Levitan, A. S. (2015). An Innovative Clustering Approach to Market Segmentation for Improved Price Prediction. *Journal of International Technology and Information Management*, *24*(1), 15–32. https://doi.org/10.58729/1941-6679.1033

Sisman, S., & Aydinoglu, A. C. (2022). Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis. *Land Use Policy*, *119*. https://doi.org/10.1016/j.landusepol.2022.106167

Skovajsa, Š. (2023). Review of Clustering Methods Used in Data-Driven Housing Market Segmentation. *Real Estate Management and Valuation*, *31*(3), 67–74. https://doi.org/10.2478/remav-2023-0022

Song, Z., Wilhelmsson, M., & Yang, Z. (2022). Constructing segmented rental housing indices: evidence from Beijing, China. *Property Management*, *40*(3), 409–436. https://doi.org/10.1108/PM-07-2021-0052

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. https://doi.org/https://doi.org/10.2307/1412159

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Tomal, M., & Helbich, M. (2022). The private rental housing market before and during the COVID-19 pandemic: A submarket analysis in Cracow, Poland. *Environment and Planning B: Urban Analytics and City Science*, *49*(6), 1646–1662. https://doi.org/10.1177/23998083211062907

UB, D., & Saxena, S. (2023). Real Estate Property Price Estimator Using Machine Learning. *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, 895–900. https://doi.org/10.1109/CISES58720.2023.10183626

Usman, H., Lizam, M., & Adekunle, M. U. (2020). Property Price Modelling, Market Segmentation and Submarket Classifications: A Review. *Real Estate Management and Valuation*, *28*(3), 24–35. https://doi.org/10.1515/remav-2020-0021

Usman, H., Lizam, M., & Burhan, B. (2021). A Priori Spatial Segmentation of Commercial Property Market using Hedonic Price Modelling. *Real Estate Management and Valuation*, *29*(2), 16–28. https://doi.org/10.2478/remav-2021-0010

Wiersma, S., Just, T., & Heinrich, M. (2022). Segmenting German housing markets using principal component and cluster analyses. *International Journal of Housing Markets and Analysis*, *15*(3), 548–578. https://doi.org/10.1108/IJHMA-01-2021-0006

World Bank. (2023). *World Bank Open Data*. https://data.worldbank.org/

Wu, C., & Sharma, R. (2012). Housing submarket classification: The role of spatial contiguity. *Applied Geography*, *32*(2), 746–756. https://doi.org/10.1016/j.apgeog.2011.08.011

Wu, C., Ye, X., Ren, F., & Du, Q. (2018). Modified Data-Driven Framework for Housing Market Segmentation. *Journal of Urban Planning and Development*, *144*(4). https://doi.org/10.1061/(asce)up.1943-5444.0000473

Wu, Y., Wei, Y. D., & Li, H. (2020). Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets. *Applied Spatial Analysis and Policy*, *13*(1), 223–256. https://doi.org/10.1007/s12061-019-09301-x

Xiong, S., Sun, Q., & Zhou, A. (2020). Improve the House Price Prediction Accuracy with a Stacked Generalization Ensemble Model. *Internet of Vehicles. Technologies and Services Toward Smart Cities*, *11894 LNCS*, 382–389. https://doi.org/10.1007/978-3-030-38651-1_32

Yağmur, A., Kayakuş, M., & Terzioğlu, M. (2022). House price prediction modeling using machine learning techniques: a comparative study. *Aestimum*, *81*, 39–51. https://doi.org/10.36253/aestim-13703

APPENDIX A

Table A.1 – Features used as inputs for the prediction models.

Feature	Description
PropertyType	Indication if the property is an apartment (1) or a house (0)
NumberRooms	Number of rooms in the property
EPC	Energy performance certificate on a scale from A+ to G
GPA	Gross private area of the property
Latitude	Latitude of the postal code's centroid
Longitude	Longitude of the postal code's centroid
Garage	Indication if the property has a garage (1) or not (0)
Pool	Indication if the property has a pool (1) or not (0)
ConservationState	Conservation state of the property (used or new)
Population	Number of residents
AgeingIndex	Ratio between the number of people over 64 years old and the number of people under 15 years old
ActivityRate	Ratio between active and resident population
AverageCommutingTime	Average commuting time (in minutes) of the employed or student population
AverageResidenceYears	Average number of residence years in private domestic aggregates
AverageHouseholdSize	Average number of people in private domestic aggregates
PopulationChangeRate	Change rate of the resident population between 2011 and 2021
AverageMonthlyRent	Average monthly rent of classic family accommodations (€)
NumberBuildings	Number of buildings
DwellingDensity	Density of dwellings
AveragePurchaseMonthlyCharges	Average monthly charges due to the purchase of a property
%SeasonalProperties	Percentage of properties of seasonal use
UnemploymentRate	Ratio between unemployed and active population
%HigherEducation	Percentage of the resident population that higher education
%Illiteracy	Percentage of the resident population over 10 years old that is illiterate
%ForeignPopulation	Percentage of the resident population that is foreign
InitialOfferYear	Year of the initial offer of the property
TimeOnMarket	Number of days between the initial offer date and the last record date
AreaTypology_PUA	Indication if the property is in a predominantly urban area (1) or not (0)
DistrictLisboa	Indication if the property is in the district of Lisbon (1) or not (0)
DistrictSetúbal	Indication if the property is in the district of Setúbal (1) or not (0)

DistrictOther	Indication if the property is in another district (1) or not (0)
MunicipalityLisboa	Indication if the property is in the municipality of Lisbon (1) or not (0)
MunicipalityOther	Indication if the property is in another municipality (1) or not (0)

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa