NOVA IMS Information Management School

MGI

Master's Degree Program in

Information Management

Predictive Modeling of Real Estate Prices in Portugal

Integrating Property Features for Accurate Valuation

Raimundo Mujica Costa

Master Thesis

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Predictive Modeling of Real Estate Prices in Lisbon:

Integrating Property Features and External Factors for Accurate Valuation.

by

Raimundo Mujica Costa

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management.

Supervised by

Miguel de Castro Neto, PhD, NOVA Information Management School

Bruno Jardim, PhD, NOVA Information Magnamente 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.

[Nova IMS, July 13th of 2024]

DEDICATION

This thesis is dedicated to my partner in life, Paula Colomba Muñoz. Without her unwavering support and encouragement, this journey would not have been the same. As we both navigate the challenges and triumphs of completing our master's theses, her constant support has been my rock. I hope I have been able to support her as profoundly as she has supported me. This dedication serves as a reminder that together, we can achieve anything we set our minds to.

I also dedicate this work to the university and all the professors and staff at NOVA Information Management School (NOVA IMS). Your guidance and commitment to excellence have pushed me to reach my highest potential. I am proud to be a part of NOVA IMS, and I will always strive to represent our university in the best possible light.

Finally, I dedicate this thesis to the readers—whether you seek specific information or have a general interest in this topic. I hope that my research aids you in your journey, whatever it may be.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to everyone who has supported me throughout this journey.

First and foremost, I am profoundly grateful to my fiancée and partner, Paula Colomba Muñoz. Your unwavering support, understanding, and encouragement have been my pillars of strength.

I would like to extend my heartfelt thanks to my thesis supervisors at NOVA Information Management School (NOVA IMS). Your insightful guidance, continuous support, and encouragement throughout the course of this research have been crucial in shaping this thesis.

I am also grateful to all my lecture and practical professors at NOVA IMS who have imparted their knowledge and wisdom. Your teachings have provided a strong foundation for my career and now in my research, and your advice has been invaluable.

A special thanks to the institution itself, NOVA IMS. Being part of this esteemed university has been a privilege, and I am proud to represent NOVA IMS.

Lastly, I would like to thank the company Confidencial Imobiliário for granting me access to their comprehensive dataset. Your data was essential for the development and validation of the predictive models used in this research.

Thank you all for your support and contributions to this journey. This thesis would not have been possible without you.

ABSTRACT

This study is a comprehensive exploration of the application of machine learning algorithms in predicting real estate prices in Portugal. It tackles the challenges posed by the complex and dynamic nature of the real estate sector, leveraging advanced analytical tools to enhance prediction accuracy. The study develops and compares five distinct machine learning models—Multilinear Regression, Regression Trees, Random Forest, Extreme Gradient Boosting (XGBoost), and Neural Networks—to identify the most effective real estate price prediction algorithm. The methodology involves data preparation, feature engineering, and model evaluation using three widely accepted evaluation metrics: Root Mean Squared Error (RMSE), R-squared (R2), and Mean Absolute Error (MAE). The findings indicate that XGBoost outperforms other models, providing superior accuracy and generalization with a mean absolute error of approximately 47 943 euros and an R2 value of 0.76. The study demonstrates how machine learning can address the intricacies of the real estate market but also offers practical, data-driven insights to support informed decision-making, thereby contributing to the academic understanding of market dynamics.

KEYWORDS

Machine Learning; Real Estate Property Prices; Predictive Modeling; Housing Prices

Sustainable Development Goals (SDG)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

RMSE	Root Mean Squared Error
GDP	Gross Domestic Product
ANOVA	Analysis of Variance
MAE	Mean Absolute Error
CRISP-DM	Cross-Industry Standard Process for Data Mining
XGBoost	Extreme Gradient Boosting
SDG	Sustainable Development Goals
NRAU	New Urban Rental Regime
PER	Special Rehousing Program
RFE	Recursive Feature Elimination
KNN	K-nearest Neighbour
R2	R-squared
MIT	Massachusetts Institute of Technology
EU	European Union
GDPR	General Data Protection Regulation
V0	Version 0 (categorical variable
V1	Version 1 (categorical variable
CP7	Postal Code (specific to Portugal, 'CP' may stand for 'Código Postal' which means Postal Code)

1. INTRODUCTION

The real estate market in Portugal has been influenced by global economic trends, local policies, and socio-cultural changes, notably affecting housing affordability and market dynamics (Lorga et al., 2022). These shifts are the result of difficult interplays between government interventions, economic fluctuations, market liberalization efforts, and the arrival of international capital (Anacker, 2019). The landscape needs some approaches to property valuation, highlighting the importance of integrating machine learning techniques. These techniques could leverage predictive modeling and large datasets to forecast property prices with enhanced accuracy, addressing the market's increased complexity (Breiman, 2001; James et al., 2013).

1.1 RESEARCH CONTEXT

As the Portugal real estate market becomes more and more complex and volatile, the precision in property valuation becomes paramount. This research is situated against the backdrop of Portugal's real estate dynamics, where prices increase fast and a significant shift toward short-term rentals have marked recent years. Such trends underscore the need for more elaborated models capable of accurately predicting property price movements, thereby aiding stakeholders in navigating the market more effectively. (Anacker, 2019; Jakabovics et al., 2014.; Geithner, 2014).

The adoption of machine learning algorithms in real estate valuation represents an innovative departure from traditional methods (Breiman et al., 1984; James et al., 2013; Bevans, 2020; Zach, 2020; Breiman et al., 1984; Louppe, 2014). By utilizing Multilinear Regression, Regression Trees, Random Forest, XGBoost and Neural Networks algorithms, this study aims to capture the relationships affecting real estate prices, influenced by property-specific and broader external factors. These algorithms are selected for their ability to process and analyse complex datasets, finding patterns and insights beyond the reach of conventional analytical approaches.

This research works to how data analytics can be applied to decipher and address complex issues in the real estate market. The study points out the value of advanced analytical tools in enhancing our understanding and predictive capabilities regarding real estate pricing dynamics.

By focusing on predictive modelling role in real estate valuation, this study aligns technological innovation with practical market analysis. It aims to offer an understanding of Portugal's real estate market, providing a resource for them. The objective is to provide these stakeholders with a data-driven tool for anticipating and responding to the challenges of urban real estate markets, both in Portugal and potentially in comparable urban contexts worldwide.

1.2 MOTIVATIONS

The application of machine learning algorithms in predicting real estate prices has been identified as a promising research area that remains underexplored (Breiman, 2001; James et al., 2013). Studies have highlighted the potential of various algorithms, including Multilinear Regression, Regression Trees, and Random Forest, to improve prediction accuracy by analysing large datasets and identifying complex patterns. However, there is a gap in comparative analysis of these algorithms within specific real estate markets. This thesis is motivated by the opportunity to contribute to the real estate market by evaluating the performance of these algorithms in the context of Portugal's real estate market, providing insights into their respective strengths and limitations.

While the capabilities of different machine learning algorithms have been recognized, there is a need for studies that compare these algorithms side by side in specific market conditions (Breiman, 2001; James et al., 2013). Such comparative analyses are crucial for identifying the most effective algorithm for predicting housing prices in different context. This thesis seeks to address this gap by conducting an evaluation of Multilinear Regression, Regression Trees, Random Forest, XGBoost and Neural Networks algorithms, focusing on their applicability and performance in predicting housing prices in Portugal. This approach not only aligns with suggestions for future research found in existing literature but also offer practical recommendations for stakeholders interested in machine learning for real estate valuation.

The fast pace of urbanization and the associated challenges in housing markets worldwide (Gordon, 2020; Cocola-Gant and Gago, 2021; Lorga et al., 2022) call for improved analytical tools and models. As some cities in Portugal continue to experience significant changes in their real estate landscapes, the importance of developing accurate and reliable predictive models becomes more evident. This thesis is motivated by the goal of contributing to the development of tools that can help making more informed decisions in the face of complex and fast evolving urban real estate markets.

1.3 RESEARCH FOCUS

The focus of this research is on developing and comparing the efficacy of five different machine learning algorithms (Multilinear Regression, Regression Trees, Random Forest, XGBoost and Neural Networks) for predicting real estate prices in Portugal. This study aims to:

Develop Predictive Models: Construct predictive models for real estate pricing in Portugal using Multilinear Regression, Regression Trees, Random Forest, XGBoost and Neural Networks algorithms. This includes integrating property specific features affecting real estate prices to create complete models.

Evaluate and Compare Algorithm Performance: Evaluate the performance of these algorithms in the context of Portugal's real estate market. The comparison will be based on metrics, such as accuracy, reliability, and the ability to handle the market's complexity and

dynamics. This comparison looks to identify the strengths and limitations of each algorithm in accurately forecasting real estate prices.

Contribute to Methodological Advancements: By conducting a comparative analysis of the algorithms, this research aims to contribute to the methodological advancements in the real estate price prediction. It aims to provide insights into the applicability of machine learning techniques in real estate valuation, also helping stakeholders in making informed decisions.

1.4 Methodology Overview

The study will employ a quantitative research methodology, focusing on the application and comparative analysis of five machine learning algorithms: Multilinear Regression, Regression Trees, Random Forest, XGBoost and Neural Networks. The methodology includes the CRISP-DM model (Chapman et al., 2000) utilizing the housing dataset provided by Confidencial Imobiliário.

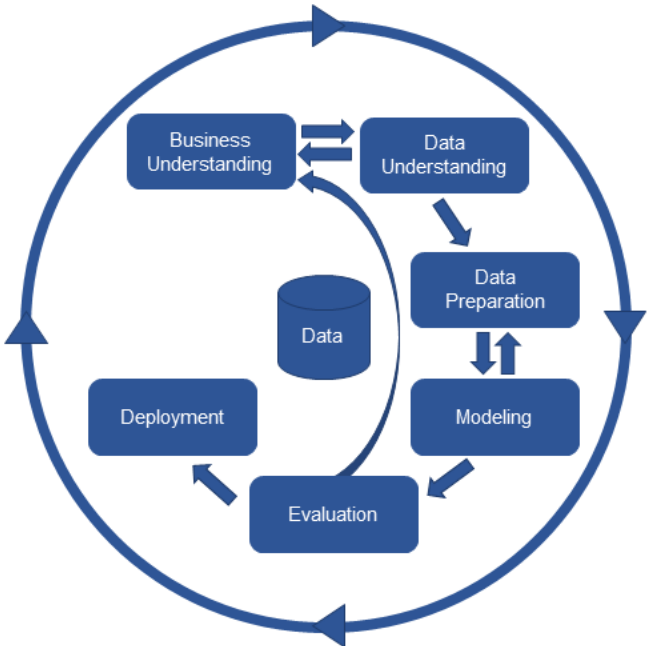

Figure 1 - Phases of the CRISP-DM

1.5 Data Sources

The datasets provided by Confidencial Imobiliário, which form the empirical foundation for this thesis on predicting real estate prices in Portugal using machine learning algorithms. These datasets offer a comprehensive overview of the real estate market in Portugal, covering property characteristics.

The research utilizes three primary datasets from Confidencial Imobiliário, each having data points important for developing and validating the predictive models. The integration of these datasets allows to analysis the factors influencing real estate prices in Portugal. The specifics of each dataset provided by Confidencial Imobiliário will be elaborated upon in the 'Data & Methodology' chapter.

In the following chapters of this research, we will develop the literature review, which will develop the most critical views, points, and research related to this project. We will also focus on the data and methodology used for this work. Then, it will develop the discussion and the research results, where it will demonstrate the alignment of our research goals with the actual results we get, ensuring the integrity of our research. Finally, the document will have a chapter with the project's conclusions and future work that could develop in the same areas of this project.

2. LITERATURE REVIEW

Housing affordability, an aspect of real estate dynamics, has experienced significant transformations over the past few decades. The relationship between rents, house prices, and household incomes has seen a notable divergence in metropolitan areas across Western countries. Anacker (2019) underscores several factors contributing to this trend, shedding light on the difficulties of the actual real estate landscape.
The escalating costs of housing can be attributed to the decreasing availability of easily developable land. Developers, over time, have change from expansive outward expansion to densification, facing challenges associated with unconventional zoning and increased construction costs (Anacker, 2019; Peterson, 2018). Planning regulations, coupled with increased construction material costs and tight lending standards post the global financial crisis, have further powered the upward trajectory of rents and house prices. (Anacker, 2019; Jakabovics et al., 2014.; Geithner, 2014).

A critical factor influencing affordability lies in the dynamics of filtering and moving chains, which, in practice, have not operated as anticipated. Developers often moves towards building or rehabilitating units for upper income households or high net worth individuals, driven by profit margins rather than regulatory guidelines (Anacker, 2019; Prevost, 2013; Harrington, 2016). Simultaneously, pulling down older, affordable units and their replacement with residences providing to higher income residents has intensified the challenge (Anacker, 2019).

Governmental shifts in social policies and austerity measures have contributed to reduced funding for affordable housing, adding to the effort on affordability (Anacker, 2019; Lennartz & Ronald, 2017). Conversely, household incomes have delayed the escalating costs of housing for various reasons. Technological advancements leading to a productivity pay gap, increased trust on credit following regulatory changes, decreased association, and a shift in employer priorities towards shareholder value over employee wellbeing are among the factors that have contributed to this income lag (Anacker, 2019; Economic Policy Institute, 2022.; De Graaf et al., 2005; Dorling, 2014; Madrick, 2011).

2.1 LOCALIZED FACTORS – PORTUGAL MARKET

Portugal, following the financial crisis in 2011, implemented austerity measures leading to income reductions. Later policies aimed at stimulating the real estate market caused an increase in housing prices, particularly in Lisbon. This rises intensified affordability challenges for the middle class, transforming the historic district of Lisbon and replacing households with medium and low incomes to peripheral areas (Lorga et al., 2022).

Historically, housing affordability challenges in Portugal affected first the low-income families. State policies, dating back to Salazar's era, focused on encouraging homeownership. The introduction of the Special Rehousing Program (PER) in the 1980s marked an important step in addressing housing challenges, facilitating large scale construction primarily in suburban areas of Lisbon and Porto (Alves & Andersen, 2015).

The global financial crisis of 2007/2008 led to transformative policies in Portugal. Initial measures targeted the most economically vulnerable, but later, policies aimed at overcoming the crisis led to the liberalization of the Portuguese banking sector and increased support for homeownership. The real estate market showed signs of overheating, especially related variable interest rates in housing loans (Banco de Portugal, 2022).

In response to the crisis, Portugal implemented new public housing policies, weakening territorial planning regulations, and introducing tax regimes favorable to real estate speculation. The Golden Visa scheme and tax regime for non-habitual residents inspired foreign investment but intensified social inequalities, increasing housing affordability issues (Cocola-Gant and Gago, 2021; Tulumello and Allegretti, 2020; Banco de Portugal, 2022).

The post crisis period seen an immediate transformation in the housing sector in Portugal. Tourist attraction, joined with neoliberal measures motivating foreign investment, resulted in a rise in short-term rentals. The commoditization of real estate, with a focus on short term rentals, created a housing market disconnected from national incomes, causing affordability problems for local buyers. The arrival of foreign investment led to the division of real estate prices from national incomes, creating a transferred demand and further contributing to housing affordability challenges (Gordon, 2020).

In response to the housing crisis, the Lisbon City Council introduced housing programs governed by the Municipal Regulation of the Right to Housing. Programs such as the Affordable Income Program (PRA) and Municipal Subsidy for Affordable Lease Program (SMAA) target middle income families unable to access the private rental market in Lisbon. However, the success of these programs is limited, with a low winning rate and subsidies granted to only a fraction of the candidates. Young people, renters, and first-time buyers are particularly affected by rising prices and dropping housing affordability, emphasizing the need for targeted interventions (Município de Lisboa, 2022; Gordon, 2020).

In the Portugal real estate market, 2024 marks a significant year with the introduction of a new rent cap. The Portuguese government has established a maximum increase in rents at 6.94% for the year, as determined by the New Urban Rental Regime (NRAU). This rate is calculated based on the average inflation rate of the preceding 12 months, excluding housing, as measured on August 31 of the previous year (Portugal Resident, 2024).

The impact of this regulation is particularly noticeable in Portugal. Data from the first half of 2023 indicated an average rent of €1,463 in the city. With the anticipated increase, this amount could escalate to about €1,564, unless the government implements additional restrictions (The Portugal News, 2023a). Furthermore, the median rent for new contracts in Lisbon's Metropolitan Area in the first quarter of 2023 was €10.26 per square meter, representing an over 9% increase compared to the same period in 2022 (The Portugal News, 2023b).

It's noteworthy that the 6.94% cap does not apply to all rental contracts, particularly “old rents”, which may see updates based on different criteria. For landlords to adjust rents in accordance with the new cap, they must provide a 30 day notice to tenants, a requirement

that comes into effect after a year since the last rent adjustment. Landlords with 'old rental' contracts are also eligible for monthly compensation starting from July 2024 (Portugal Resident, 2024).

These regulatory changes in Lisbon's housing market highlight the government's attempt to create more stability. They aim to protect the interests of both landlords and tenants in an environment characterized by inflationary tendencies. Such measures are crucial for maintaining a stable housing market and ensuring the accessibility of rental accommodations for a wider demographic.

2.2 Application of Machine Learning in Real Estate

The integration of machine learning in real estate holds significant promise. Real estate technology, though advancing, has room for improvement. Understanding the potential applications and opportunities for machine learning and artificial intelligence can improve decision-making, simplify workflows, and reduce risks. (Conway, 2018)

Jennifer Conway's work in 2018 explores the application of artificial intelligence and machine learning specifically in the real estate market. Highlighting the potential advantages, Conway discusses how machine learning can be a tool for predicting market trends, providing insights for real estate professionals.

The Portugal real estate market, like many others, has been influenced by many factors such as housing affordability, government policies, and socio-economic changes (Anacker, 2019; Lorga et al., 2022). These factors have created a complex landscape that requires analytical tools for accurate prediction and analysis. Machine Learning offers such tools, allowing the extraction of meaningful patterns from large datasets.

Historical factors, such as the austerity measures post 2011 financial crisis, have shaped the current state of the Portugal real estate market (Lorga et al., 2022). Additionally, local dynamics like short term rentals and neoliberal measures have further complicated the market (Gordon, 2020; Cocola-Gant and Gago, 2021). These elements provide rich data for ML algorithms to analyze market trends and pricing.

Machine Learning, specifically algorithms like Multilinear Regression, Regression Trees, and Random Forest, have shown significant potential in real estate price prediction (Breiman et al., 1984; James et al., 2013). While Multilinear Regression offers a linear approach to understanding market factors (Bevans, 2020; Zach, 2020), Regression Trees and Random Forest provide more nuanced analyses by considering many variables and their interactions (Breiman et al., 1984; Louppe, 2014).

Multilinear Regression, with its ability to handle multiple influencing factors simultaneously, can be crucial in understanding how different market variables collectively impact property prices in Portugal (Bevans, 2020; Zach, 2020).

Regression Trees offer an interpretable model, making them suitable for scenarios where decision-making processes are as crucial as the outcomes (James et al., 2013). On the other

hand, Random Forest, with its ensemble approach, provides robust predictions by aggregating the results of multiple decision trees, reducing the risk of overfitting, and increasing predictive accuracy (Breiman, 2001; Louppe, 2014).

Despite the advancements in ML algorithms, a critical gap exists in identifying the most effective model for predicting real estate prices in Portugal. Studies in United State have shown varying results, with some advocating for the use of Random Forest due to its robustness, while others prefer Regression Trees for their interpretability (Breiman, 2001; James et al., 2013). This disparity needs a comparative analysis of these models in the context of Portugal's unique market dynamics.

The integration of Machine Learning with real estate pricing in Portugal presents a promising avenue for advanced market analysis. However, the effectiveness of different ML models changes, and there is a need for comprehensive studies that evaluate these models in the context of Portugal's real estate market. Such research will not only increase predictive accuracy but also offer valuable insights for stakeholders in the Portugal real estate sector.

3. DATA & METHODOLOGIES

As highlighted in the methodology overview, this research harnesses the power of the CRISP-DM model. The steps of the CRISP-DM, which will be the principal sections of this chapter, play a pivotal role in our research, ensuring a comprehensive and systematic approach (Chapman et al., 2000).

This model is a fitting choice for our research's study area, which delves into the housing market in Portugal. Our theoretical and practical objective is to predict housing prices using machine learning algorithms. These algorithms, applied to the dataset provided by Confidencial Imobiliário, a real estate company in Lisbon, will provide us with the ability to make informed predictions, thereby increasing the practicality of our research.

The dataset we are working with covers information in various areas of Portugal and includes a wide range of housing factors, which will be elaborated on in the following sections of this chapter. The data is diverse regarding geographical area, adding depth to our research. Therefore, focusing on the specific factors that influence prices is crucial, as highlighted in a report by the Brookings Institution (2018), which highlight the significant price gaps between regional locations and capital cities.

In the following sections, we will analyze the steps of Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation of the CRISP-DM model to ensure the reliability and validity of our research.

3.1 BUSINESS UNDERSTANDING

This section focuses on understanding the project requirements and the objectives we want to obtain in this research. First, it is crucial to understand the key factors that could influence housing prices; as we know from previous chapters, housing affordability is more challenging than it was in the past. Also, the demand is greater than the number of houses available, and the focus on reforming previous houses to get a more valuable home is another factor that helps to increase the prices. Another fact is that the location is essential; according to the Brookings Institution (2018), regional houses are cheaper than cities with less space, and the influence of economic opportunities, infrastructure, and population density are also significant (Brookings Institution, 2018).

Our next step involves analyzing the data to identify the key factors. We aim to develop a practical and versatile predictive model that can be applied across different scenarios, aligning with our research objectives.

3.2 DATA UNDERSTANDING

The datasets provided by Confidencial Imobiliário are compile in Table 1.

Table 1 - Features Confidencial Imobiliário data.

Column Name	Description
IMOVEL_ID	Unique identifier for the property.
TIPO_IMOVEL	Type of property (1 = Apartment, 2 = House).
N_QUARTOS	Number of rooms.
CCE (Classificação do Certificado Energético)	Energy Certificate Classification (0 = no information, 1 = A+, 2 = A, ..., 9 = G).
ABP	Gross Private Area.
FREGUESIA	ID according to the General Directorate of Territory.
CP7	7-digit Postal Code.
LATITUDE	Latitude of the centroid of CP7.
LONGITUDE	Longitude of the centroid of CP7.
GARAGEM	Garage (FALSE = No, TRUE = Yes, empty = unknown).
TERRACO	Terrace (FALSE = No, TRUE = Yes, empty = unknown).
PISCINA	Swimming pool (FALSE = No, TRUE = Yes, empty = unknown).
LOGRADOURO	Yard (FALSE = No, TRUE = Yes, empty = unknown).
ESTADO_DE_CONSERVACAO	Condition (Used/New).
FLAG_ESTADO_DE_CONSERVACAO	Condition flag (V0 = Client-provided information, V1 = Information validated with photos).
ULTIMA_DATA_EM_MERCADO	Last date the property appeared in the database.
DATA_OFERTA_INICIAL	Initial offer date.
VALOR_OFERTA_INICIAL	Initial offer value.
VALOR_OFERTA_ATUAL_FINAL	Current final offer value.
PRECO_VENDA_ESTIMADO	Estimated sale price.
PRECO_VENDA_OBSERVADO	Observed sale price.
INDICE_MAIS_RECENTE	Price index for the most recent period.
INDICE_ULTIMA_DATA_EM_MERCADO	Price index on the last date the property appeared in the database.

While most of the features in the dataset provided by Confidencial Imobiliário are relatively straightforward to identify, it's crucial to note that some of them are significantly relevant to the pricing aspect of our predictive model.

Our model's dependent variable, "PRECO_VENDA_OBSERVADO," represents the value we aim to predict. However, certain features, such as "VALOR_OFERTA_INICIAL," "VALOR_OFERTA_ATUAL_FINAL," "PRECO_VENDA_ESTIMADO," "INDICE_MAIS_RECENTE," and "INDICE_ULTIMA_DATA_EM_MERCADO," are not readily available in a house and therefore pose a challenge in our prediction process.

That is because Confidencial Imobiliário creates these features to perform their predictive model; sometimes, as a real estate company, they had access to the initial value of a house and knew the actual value as a house was purchased. So, in the case of this project, we will not consider some of the features because they will interfere with the principal objective of the project that is to apply the model in a practical way and different scenarios.

Then, it is important to identify the principal statistics and outliers of the data, Figure 2 shows some visualizations representing the outliers of some of the features.

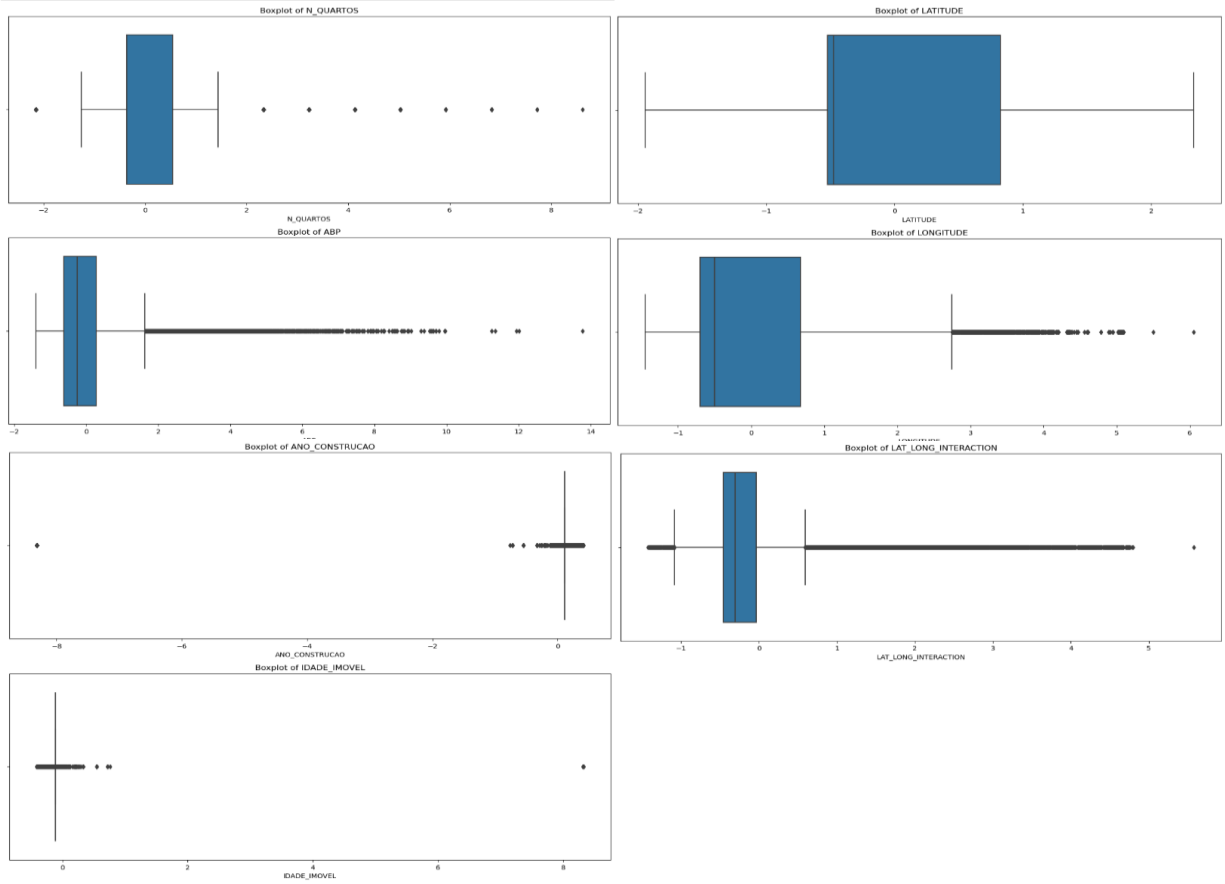

Figure 2 - Analysing Outliers Box Plot

Here, we can gain many insights regarding numerical features. Most records about the number of rooms in the Gross Floor Area and the ABP have a normal range, but there are also some outliers.

The Latitude and Longitude help to confirm that inside the data, we are analyzing only houses from Portugal.

Talking about the "GARAGEM," "TERRACO," "PISCINA," and "LOGRADOURO," we can conclude that most of the houses in Portugal do not have these amenities because it does not register many of these features.

Also, the data shows that most of the properties last dates on the market were between 2017 and 2022. The sales prices reveal a wide range from 25 500 euros to 4 950 000 euros with a mean of 209 363 euros. Additionally, most properties were built around 1966, between 1750 and 2022.

Table 2 - Categorical Statistics

	FREGUESIA	CP7	FLAG_ESTADO_DE_CONSERVACAO
unique	1976	36730	2
top	110656	1750-220	V0
freq	2945	211	89235

Regarding the categorical features, Table 2 shows the statistics of the categorical features, each property has its identifier without duplicates in the data. For "FREGUESIA," which means parish, we have 1976 different parishes across all the data; we can identify some parishes are denser than others regarding the number of properties. We have a typical wide variety in "CP7," the postal code. Finally, "FLAG_ESTADO_DE_CONSERVACAO," says that most of the records on the data only had a description of the property without a photo (V0), with 74.87% and 25.13% of the data having photos for identifying the property (V1).

3.3 DATA PREPARATION

This section will prepare the data to implement the predictive algorithms and build our model as accurately as possible. First, we need to keep in mind which techniques we will used for the data preprocessing. In the first phase we deal with the missing values, keeping in mind that we have inside the data features that are categorical and features that are numerical, we will use different techniques to deal with them.

So, one of the features is the "ANO_CONSTRUCAO" that is a numeric feature, we determine the missing actual data. Then, we replace the missing values for the feature median. According to Bhandari (2020), "The median does a better job of capturing the 'typical' value of a dataset than the mean when the data is skewed or contains outliers" (Bhandari, 2020). Also, we have more columns that had missing values "GARAGEM", "TERRACO", "PISCINA", "LOGRADOURO", as all these features are categorical values, we replace them with the value that is more repeated for each feature, the mode.

After missing values, we went forward to Feature Engineering, where we used different feature engineering techniques as date extraction, interaction terms, binning, and arithmetic

transformation; like this, we can create meaningful features that could help to have better results (Zheng & Casari, 2018). So, we created the features that are shown in Table 3.

Table 3 - New Columns

Feature	Description
ULTIMA_DATA_ANO	The year extracted from the 'ULTIMA_DATA_EM_MERCADO' column. This represents the last year the property was on the market. (Date Extraction).
ULTIMA_DATA_MES	The month extracted from the 'ULTIMA_DATA_EM_MERCADO' column. This represents the last month the property was on the market. (Date Extraction).
LAT_LONG_INTERACTION	A feature created by multiplying the 'LATITUDE' and 'LONGITUDE' values. This interaction term may capture spatial relationships in the data. (Interaction Terms).
GARAGE_TERRACE_INTERACTION	A feature created by multiplying the 'GARAGEM' and 'TERRACO' values. This interaction term might capture the combined effect of having both a garage and a terrace on property value. (Interaction Terms).
ABP_BINNED	A binned version of the 'ABP' (Area Bruta Privativa), which is the gross private area of the property. The binned values categorize the 'ABP' into specified ranges for better analysis. (Binning).
IDADE_IMOVEL	The age of the property, calculated by subtracting the 'ANO_CONSTRUCAO' (year of construction) from the current year. This represents how old the property is. (Arithmetic Transformation).
IDADE_GRUPO	A binned version of the 'IDADE_IMOVEL'. This feature categorizes the age of the property into specified ranges, allowing for a more structured analysis of property age groups. (Binning).

After Feature Engineering, we used only one type of encoding for the categorical features: Target encoding. We do not use one-hot encoding, because if we use it, we will have more dimensions and we wanted to avoid the curse of dimensionality.

Then, we do a standardization process with the numerical features. This way, we have the same scale for all features and can achieve better performance. As cited by Goodfellow, Bengio, and Courville (2016), "Features with different scales can make it difficult for the optimization algorithm to converge because it has to adjust weights for features with different magnitudes" (Goodfellow, Bengio, & Courville, 2016; Hsu et al., 2016).

Finally, we apply Feature Selection. For this step we combined different techniques. First, we use Recursive Feature Elimination (RFE) that is an iterative feature selection that fits the model and remove the weakest feature, this process is repeated until the technique finds the appropriate number of features. In each interaction the model is trained, and the importance of each feature is set.

In Guyon et al. (2002) research, they used RFE in the context of gene selection for cancer classification using support vector machines. This method has been adopted in machine learning applications due to its effectiveness in improving model performance by deleting irrelevant or redundant features.

In our case, RFE was used with Linear Regression model, and the number of features to retain was varied between 2 and 15. We evaluate the performance of each interaction with mean squared error (MSE) and the optimal number was the one that minimize the error (Guyon et al., 2002).

The other technique that was implemented was SelectKBest with ANOVA F-value. SelectKBest is a univariate feature selection method that select the top K features based on statistical procedures. In this study, SelectKBest was used with ANOVA F-value, which evaluates the correlation between each feature and the target variable. This specific technique is documented in the scikit-learn library, that is a machine learning library in Python. (Pedregosa et al., 2011).

This method selects the features that have the strongest linear relationship with the target variable. The top 15 features were selected based on their ANOVA F-values that are shown in Table 4 (Pedregosa et al., 2011).

Table 4 - Feature Selection

FEATURE	SELECTION
IMOVEL_ID	Selected
TIPO_IMOVEL	Selected
N_QUARTOS	Selected
CCE	Selected
ABP	Selected
FREGUESIA	Selected
CP7	Selected
LATITUDE	Selected
LONGITUDE	Selected
GARAGEM	Not Selected
TERRACO	Not Selected
PISCINA	Selected
LOGRADOURO	Not Selected
ESTADO_DE_CONSERVACAO	Selected
FLAG_ESTADO_DE_CONSERVACAO	Not Selected
ANO_CONSTRUCAO	Selected
IDADE_IMOVEL	Selected
ABP_BINNED	Selected
LAT_LONG_INTERACTION	Not Selected
GARAGE_TERRACE_INTERACTION	Not Selected
IDADE_GRUPO	Selected

3.4 MODELLING

Here, it is vital to highlight the model selection. We wanted to implement different model algorithms and decided to implement Random Forest, Neural Network, Multilinear Regression, XGBoost Regressor, Linear Regression, and Decision Tree. (Breiman, 2001; Rumelhart et al., 1986; Draper et al., 1998; Chen & Guestrin, 2016; Galton, 1886; Quinlan, 1986)

The reasons for selecting these algorithms were the following:

In the case of Random Forest, it is a method that builds multiple decision trees and merges them to predict the results. One of the reasons that we chose this model was because it is accurate and handles overfitting well because of the randomness in tree construction (Breiman, 2001).

For Neural Networks, this model can handle complex data relationships; it is flexible and can approximate any continuous function given enough data and computational power. For regression tasks, neural networks are decisive for learning complex mappings from inputs and outputs (Hornik et al., 1989).

XGBoost Regressor is one of the models that we chose. This model is known because of the good results that it can manage in terms of efficiency, accuracy, and scalability. XGBoost consistently outperforms other methods in various machine learning challenges because of its optimization techniques and regulation capabilities (Chen & Guestrin, 2016).

On the other hand, linear regression is a simple model that is effective in many practical problems. So, it is a good option because of its interpretability and easy implementation (Draper & Smith, 1998).

Our last model is the Decision Trees Regressor. Like linear regression, it is easy to interpret. Because of its simplicity and interpretability, it is effective for classification and regression tasks (Quinlan, 1986).

After our model selection, we test for each model different parameter, the specific parameters for each model are presented in Table 5.

Table 5 - Model Parameter

Model	Parameter	Value
XGBoost	n_estimators	[100, 300, 500]
	max_depth	[3, 5, 7]
	learning_rate	[0.01, 0.1, 0.2]
	subsample	[0.8, 1.0]
	colsample_bytree	[0.8, 1.0]
	gamma	[0, 1, 5]
	reg_lambda	[1, 2, 5]
Multilinear Regression	fit_intercept	[True, False]
Decision Tree Regression	criterion	['mse', 'friedman_mse', 'mae']
	splitter	['best', 'random']
	max_depth	[None, 10, 20, 30]
	min_samples_split	[2, 10, 20]
	min_samples_leaf	[1, 5, 10]
Neural Networks	batch_size	[10, 20]
	epochs	[50, 100]
	model_optimizer	['SGD', 'Adam']
	model_init	['uniform', 'normal']
Random Forest	n_estimators	[100, 200]
	criterion	['squared_error']
	max_depth	[None, 10, 20]
	min_samples_split	[2, 10]
	min_samples_leaf	[1, 5]
	bootstrap	[True]

We will be choosing the best parameters for each model with a technique called Grid Search, where is testing in the dataset the different combination of parameters with cross-validation, and it is selected the parameters that perform better for each model (Bergstra and Bengio, 2012).

After choosing the different parameters in each model, we created the pipeline. In the pipeline, we include the different imputations that we will use, scaling, and encoding. Finally, we train the model; in this step, we preprocess the data and evaluate the model.

3.5 EVALUATION

For Evaluation, we use cross-validation. We decided to use cross-validation because it effectively provides robust results of model performance and prevents overfitting (Kohavi, 1995). Basically, we test the model on different subsets of the data to detect and reduce the overfitting. To avoid data leakage, we used the preprocessing pipeline after splitting the data in cross-validation.

We use root mean squared error (RMSE), R2 (coefficient of determination), and mean absolute error (MAE) to measure the performance of the models.

RMSE measures the average magnitude error between the predicted values and actual values. It is sensitive to significant errors, which makes it useful for measurement purposes. RMSE is an effective measure for assessing the accuracy of model predictions, highlighting scenarios where it is important to avoid undesirable errors (Chai & Draxler, 2014).

R2, on the other hand, measures the proportion of the variance in the dependent variable that is predictable from the independent variable. It indicates how well the model explains the variability in the data (Steel & Torrie, 1980); a higher R2 means that it is a better fit.

The other measure is the MAE, which measures the average magnitude of the error in a set of predictions without considering their direction. It is less sensitive to errors than RMSE. Willmott and Matsuura (2005) emphasized that MAE is a straightforward and interpretable metric, so it helps assess model accuracy in general (Willmott & Matsuura, 2005).

These measures, together with a collection of the best parameters and performance metrics for each model, can help us determine which combination and algorithm performs better.

4. RESULTS AND DISCUSSION

4.1 RESULTS

As we highlighted in the previous chapter, we wanted to test the performance of different algorithms. For that reason, we evaluate the algorithms one by one doing similar steps.

The first step was building a Grid Search for find the best parameters for each model (Table 6). Here is testing in the entire dataset the different combination of parameters and is selected the one that performs better.

In Table 6 we highlight the selected parameters for each model, where it specifies the value that was selected for each parameter of each model.

Table 6 - Parameter Selection

Model	Parameter	Value
XGBoost	n_estimators	500
	max_depth	7
	learning_rate	0.1
	subsample	1.0
	colsample_bytree	1.0
	gamma	0
	reg_lambda	5
	random_state	42
Multilinear Regression	fit_intercept	True
Decision Tree Regression	criterion	friedman_mse
	max_depth	10
	min_samples_leaf	10
	min_samples_split	2
	splitter	best
Neural Networks	batch_size	10
	epochs	100
	init	normal
	optimizer	Adam
Random Forest	n_estimators	200
	min_samples_split	2
	min_samples_leaf	1
	max_depth	20
	criterion	squared_error
	bootstrap	True

Then, with the selected parameters in each model and one by one, we run our models in the entire dataset with cross-validation with 10 splits.

Finally, we measure the performance of each algorithm in the Train data and in the Test data, Table 7 shows the results with the measures that it was used (MAE, RMSE and R2) for each model.

Table 7 - Performance of All Models

Metric	Average Train MAE	Average Train RMSE	Average Train R2	Average Test MAE	Average Test RMSE	Average Test R2
Neural Network	49583.80	98953.25	0.77	63204.81	124935.80	0.63
Decision Tree	40699.61	76654.76	0.86	60122.40	119938.54	0.66
Multiple Linear Regression	64422.73	122380.54	0.64	67272.51	129521.55	0.60
XGBoost	29007.88	46941.54	0.95	**47943.07**	**100331.81**	**0.76**
Random Forest	**18701.92**	**33526.18**	**0.97**	50744.73	107242.32	0.72
Best Model	Random Forest	Random Forest	Random Forest	XGBoost	XGBoost	XGBoost

The RMSE metric illustrates that a lower value denotes the superior performance of a predictive model. The R^2 score, representing the proportion of variance in the dependent variable predictable from the independent variables, confirms the robustness of the predictions. If we analyze the performance of all models (Table 7), Random Forest performs exceptionally well on the training data, with the lowest MAE and RMSE and the highest R2. This indicate that the model fits the training data very well. But this could also suggest potential overfitting if the performance on test data is not equally good. For test data Extreme Gradient Boosting (XGBoost) shows the best performance on the test data with the lowest MAE and RMSE and the highest R2. This suggest that XGBoost generalizes better to unseen data compared to other models.

The results suggest that XGBoost is a better model for generalization to new data in compared with Random Forest that overfit a little bit more.

Considering the Real Estate industry, a MAE of around 47 943 euros means that, on average, the predictions are off by this amount. We can say that for luxury homes this amount might be acceptable. However, for lower-value properties could represent a significant portion of the total value.

Talking about R2, an 0.76 is good, indicating that 76% of the variance in the property prices can be explain by the model. This suggest that the model is overfitting, but have a good predictive power, which is generally favorable in real estate valuation models.

Because of XGBoost is overfitting, it is important to understand where the model is underperforming. So, it is important to analyze the residuals of the model across different dimensions.

The Figure 3 shows the residuals (difference between observed values and the values predicted by the model) vs the year of construction ("ANO_CONSTRUCAO"), the figure reveals residuals around the zero line, highlighting the properties built after 1750. So, the model performs better in more modern properties.

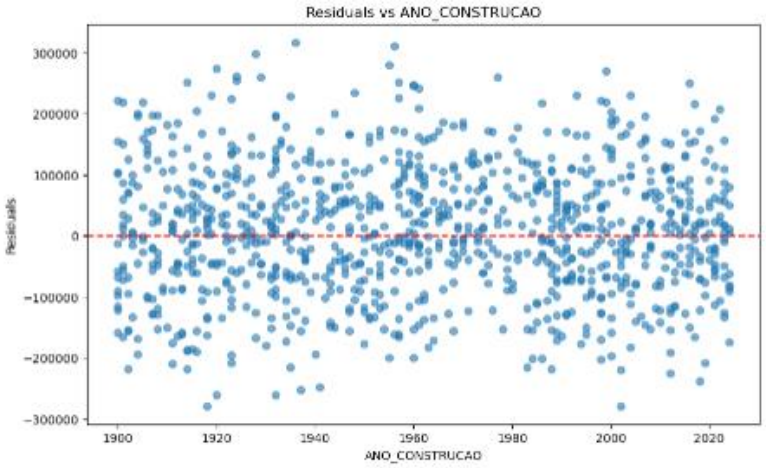

Figure 3 - Residuals vs ANO_CONSTRUCAO

Analyzing another dimension, the residuals analysis by the geography location base in the latitude and longitude shows significant errors in specific locations. A list of location with their Postal Code (CP7) shows the average prediction error in Appendix A.

Paying attention into the property size (ABP) and the residuals, we get that the model gets higher errors for properties with larger areas. Figure 4 shows the relation between these two features.

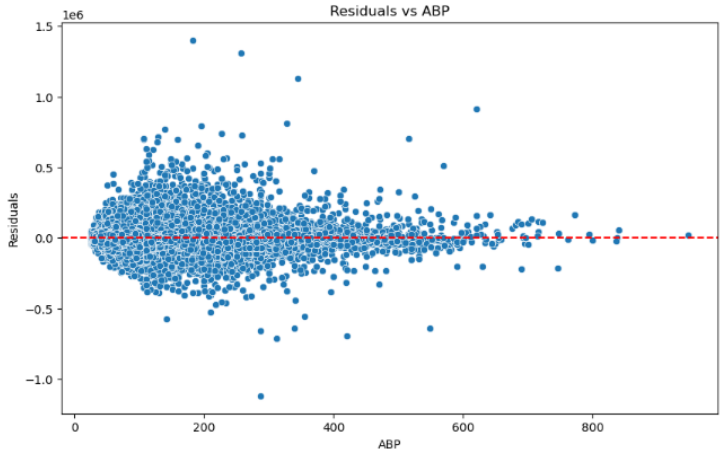

Figure 4 - Residuals vs ABP

In other hand, the model is performing well when its talking about property type, when analyzing apartment and houses we get not significant differences, indicating the model's performance is stable across property types. Figure 5 shows the relation of these features.

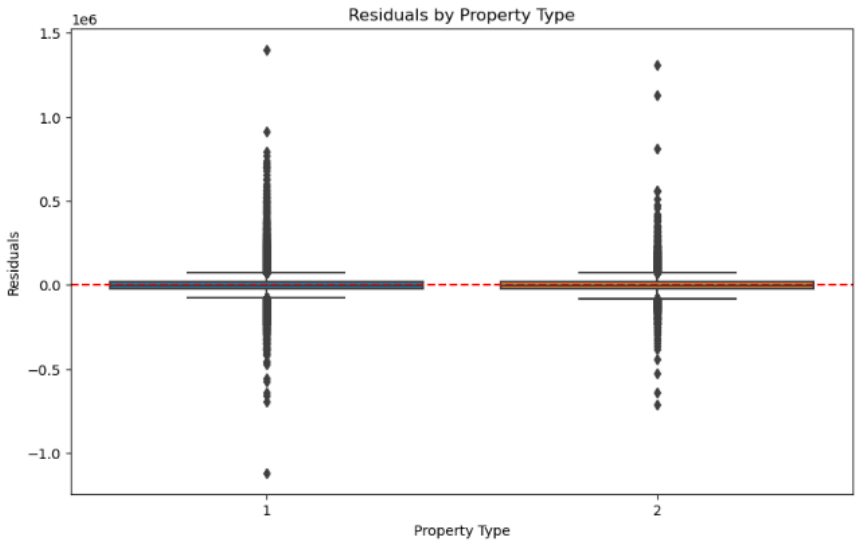

Figure 5 - Residuals by Property Type

4.2 DISCUSSION

Here, it is vital to delve into the research questions that were highlighted at the beginning of this research. The objective is to examine the research and find the initial goals contributing to understanding real estate price predictions using machine learning algorithms.

The first research goal was to develop and optimize predictive models for real estate prices in Lisbon using Multilinear Regression, Regression Trees, Random Forest, Extreme Gradient Boosting (XGBoost) and Neural Networks algorithms. The principal objective was to build models that reflected the complexity of the Portuguese market.

Our findings indicate that this objective was achieved through several steps. First, talking about the Data Preparation and Feature Engineering phase, the dataset from Confidencial Imobiliário was preprocessed to handle missing values and outliers. It was applying feature engineering techniques to create new, meaningful features, such as "LAT_LONG_INTERACTION" and "IDADE_IMOVEL," which provided additional predictive power. Also, categorical features were encoded using target encoding to avoid the curse of dimensionality.

Taking about model selection and parameter tuning, grid search was employed to find the best parameters for each model, ensuring the best combination of parameters for each algorithm. Also, was implemented cross-validation with 10 splits to evaluate the performance of each model, helping to prevent overfitting and ensuring robust results.

In addition, taking about performance evaluation, the models were evaluated using RMSE, R2 and MAE metrics. XGBoost showed the best performance on test data with the lowest MAE (47943.07 euros), RMSE (100331.81), and highest R2 (0.76), indicating its strong generalization ability. In the other hand, Random Forest performed exceptionally well on training data but showed signs of overfitting when evaluated on test data.

We successfully developed and optimized predictive models that can accurately forecast real estate prices in Portugal, demonstrating the capability of machine learning algorithms to handle complex datasets and provide valuable insights.

The second goal was to compare the performance of the selected machine learning algorithms in predicting real estate prices in the Portugal market. This comparison was based on several performance metrics and was crucial in identifying the most suitable algorithm for this specific application.

Comparing the algorithms performance, we evaluated using MAE, RMSE, and R2 metrics on training and test datasets. XGBoost outperformed other models on test data, suggesting it is the most effective algorithm for generalization and providing reliable predictions on unseen data. Random Forest, in other hand, showed excellent performance on training data but did not generalize as well to the test data, indicating potential overfitting. Neural Networks, Multilinear Regression and Regression Trees provided reasonable results, but were outperformed by XGBoost and Random Forest.

The study identified XGBoost as the most effective model for predicting real estate prices in Portugal, providing steakholders with a reliable tool for market analysis.

The third goal was to give actionable insights and recommendations from the comparative analysis of the predictive models, aimed at helping stakeholders to make informed decisions.

The high R2 value (0.76) of XGBoost indicates a strong predictive power, suggesting that it can explain a significant portion of the variance in property prices. This makes XGBoost a valuable tool for stakeholders seeking to understand market dynamics. Also, the MAE of around 47 943 euros, while acceptable for high-value properties, suggests that stakeholders should consider the price range of properties when interpreting model predictions.

The companies in the real estate market should utilize the predictive models, especially XGBoost, to forecast market trends and assess the impact of regulatory changes on property prices. Also, this model's application of advanced analytics can aid in analyzing housing affordability and market stability. By leveraging the accuracy and robustness of these models, the real estate market can better understand the potential outcomes of the regulatory decisions in some areas and create strategies to get better opportunities.

However, when analyzing the model in different dimensions, we can discover how to improve the technique. We get that in some geographic locations the model has high variability in their

results, the same happened related with the modern properties and older properties. That kind of analysis gives as the perspective that different models dedicated for different locations and time-based attributes can even help us to get better results.

In that case, these models can help investors can leverage the predictive accuracy of XGBoost to make informed investment decisions. The model's ability to accurately forecast property prices allow investors to identify lucrative opportunities and mitigate risks associated with market volatility. By using these insights, investors can optimize their portfolios, make strategic investments, and avoid potential pitfalls in the dynamic real estate market.

5. CONCLUSIONS AND FUTURE WORKS

This study set out to explore the application of machine learning algorithms in predicting real estate prices in Portugal, with a particular focus on the Lisbon market. Through the development and optimization of predictive models, the study aimed to capture the complex dynamics influencing property prices and provide actionable insights for stakeholders.

The research successfully achieved its objectives, demonstrating the efficacy of various machine learning models, including Multilinear Regression, Regression Trees, Random Forest, XGBoost, and Neural Networks. Among these, XGBoost emerged as the most effective algorithm, offering the best balance of accuracy and generalization, as evidenced by its superior performance on test data with a Mean Absolute Error (MAE) of 47 943 euros and an R2 value of 0.76.

Some of the findings of the study include are effective data preparation and feature engineering, handling missing values, outliers, and applying feature engineering techniques significantly enhanced the predictive power of the models. Model selection and optimization, employing grid search and cross-validation ensured that the best parameters were selected for each model, enhancing their performance and robustness. Performance evaluation, the comprehensive evaluation using RMSE, R2, and MAE metrics provided a clear comparison of the models, identifying XGBoost as the most effective for real estate price prediction in the Portuguese market.

While this study has made significant contributions to the field of real estate price prediction, there are several avenues for future research and limitations that could further enhance the models and their applicability.

One limitation that we found during the research was a more update dataset. The database itself has multiple features that helps to the success of the process. However, the properties from the dataset were maximum until the year 2022.

Also, we incorporate only the data provided from Confidencial Imobiliário. Considering incorporating additional data sources. Future research could integrate more granular data from additional sources, such as economic indicators, demographic trends, and environmental factors, to further improve model accuracy and provide deeper insights.

An additional point, it is to explore more advanced algorithms. Beyond the algorithms evaluated in this study, exploring other advanced machine learning techniques, such as deep learning models and ensemble methods, could yield even more accurate predictions.

One other thing that in future works could be improve is addressing lower-value properties. Given that the MAE of around 47 943 euros might be significant for lower-value properties, future work could focus on refining the models to improve accuracy across all property value ranges.

An interesting approach that could be done, it is to set different models per each district. A study by Chen, Wei, and Xu (2017) demonstrated that using divided the data into districts and perform different models for house price prediction improved the model's performance compared to a more generalized approach.

In addition, could be interesting to add real-time data integration. Developing models that can integrate real-time data could provide up-to-date predictions, which would be particularly valuable in fast-moving markets.

Finally, it could be helpful to apply this kind of research in a user-friendly interface. Creating user-friendly interfaces for these predictive models would make them more accessible to non-technical stakeholders, enhancing their practical utility in decision-making processes.

BIBLIOGRAPHICAL REFERENCES

Alves, S., & Andersen, H. T. (2015). Social housing in Portugal and Denmark: a comparative perspective - Paper presented at ENHR - European Network for Housing Research Conference 2015, Lisboa. http://hdl.handle.net/10451/19984

Anacker, K. B. (2019). Introduction: housing affordability and affordable housing. In *International Journal of Housing Policy* (Vol. 19, Issue 1, pp. 1–16). Routledge. https://doi.org/10.1080/19491247.2018.1560544

Banco de Portugal. (2022). *Boletim Económico - Dezembro 2022*. https://www.bportugal.pt/sites/default/files/anexos/pdf-boletim/be_dez22_p.pdf

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305. Retrieved from http://www.jmlr.org/papers/v13/bergstra12a.html

Bevans, R. (2020, February 20). *Multiple Linear Regression | A Quick Guide (Examples)*. https://www.scribbr.com/statistics/multiple-linear-regression/

Bhandari, P. (2020, July 30). Central Tendency | Understanding the Mean, Median & Mode. Scribbr. Retrieved from https://www.scribbr.com/statistics/central-tendency/

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. doi:10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression* Trees. (1st Edition). Routledge.

Brookings Institution. (2018). Housing in the US is too expensive, too cheap, and just right. It depends on where you live. Retrieved from https://www.brookings.edu/articles/housing-in-the-us-is-too-expensive-too-cheap-and-just-right-it-depends-on-where-you-live

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247-1250. doi:10.5194/gmd-7-1247-2014

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium. Retrieved from https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

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

Chen, X., Wei, L., & Xu, J. (2017). House price prediction using LSTM. arXiv preprint arXiv:1709.08432. Retrieved from http://www.jmlr.org/papers/v13/bergstra12a.html

Cocola-Gant, A., & Gago, A. (2021). Airbnb, buy-to-let investment and tourism-driven displacement: A case study in Lisbon. Environment and Planning A, 53(7), 1671–1688. https://doi.org/10.1177/0308518X19869012

Conway, J., & Architecture, B. A. (2018). *Artificial Intelligence and Machine Learning: Current Applications in Real Estate*.

De Graaf, J., Wann, D., & Naylor, T. H. (2002). Affluenza: The All-Consuming Epidemic. In Environmental Management and Health (Vol. 13). Emerald Group Publishing Limited. https://doi.org/10.1108/emh.2002.13.2.224.3

Dorling, D. (2014). Inequality and the 1%. Verso.

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (Vol. 326). John Wiley & Sons. doi:10.1002/9781118625590

Economic Policy Institute. (2022, October). The Productivity–Pay Gap. Https://Www.Epi.Org/Productivity-Pay-Gap/.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland, 15*, 246-263. https://doi.org/10.2307/2841583

Geithner, T. F. (2015). Stress test: Reflections on financial crises. Crown Publishers.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Retrieved from https://www.deeplearningbook.org/

Gordon, J. C. (2020). Reconnecting the housing market to the labour market: Foreign ownership and housing affordability in Urban Canada. In *Canadian Public Policy* (Vol. 46, Issue 1, pp. 1–22). University of Toronto Press Inc. https://doi.org/10.3138/cpp.2019-009

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389-422. https://doi.org/10.1023/A:1012487302797

Harrington, B. (2016). Capital without Borders: Wealth Managers and the One Percent (1st Edition). Harvard University Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural networks, 2(5), 359-366. doi:10.1016/0893-6080(89)90020-8

Hsu, C. W., Chang, C. C., & Lin, C. J. (2016). A practical guide to support vector classification. In Recommender Systems Handbook (pp. 357-376). Springer. doi:10.1007/978-0-387-85820-3_9

Jakabovics, A., Ross, L. M., Simpson, M., & Spotts, M. (2014). *Bending the cost curve: Solutions to expand the supply of affordable rentals*. Urban Land Institute. https://uli.org/wp-content/uploads/ULI-Documents/BendingCostCurve-Solutions_2014_web.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI (Vol. 14, No. 2, pp. 1137-1145).

Lennartz, C., & Ronald, R. (2017). Asset-based Welfare and Social Investment: Competing, Compatible, or Complementary Social Policy Strategies for the New Welfare State? Housing, Theory and Society, 34(2), 201–220. https://doi.org/10.1080/14036096.2016.1220422

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

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. http://arxiv.org/abs/1407.7502

Madrick, J. (2011). Age of Greed: The Triumph of Finance and the Decline of America, 1970 to the Present. Alfred A. Knopf.

Município de Lisboa. (2022, May 13). Condições de acesso aos Programas de Habitação do Município de Lisboa. Https://Habitarlisboa.Cm-Lisboa.Pt/Ords/f?P=100:14::::14::&cs=3622CFBNKledJO3FDrFy_W-51O9E

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html

Peterson, K. (2018). Backdoor revolution: The definitive guide to ADU development. Accessory Dwelling Strategies, LLC.

Prevost, L. (2013). Snob Zones: Fear, Prejudice, and Real Estate. Beacon Press.

Portugal Resident. (2024, January 10). 6.94% rent cap in 2024. Https://Www.Portugalresident.Com/6-94-Rent-Cap-in-2024/.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106. doi:10.1007/BF00116251

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536. https://doi.org/10.1038/323533a0

Steel, R. G. D., & Torrie, J. H. (1980). Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill.

The Portugal News. (2023b, December 30). Where have rental prices risen most in Portugal this year? Https://Www.Theportugalnews.Com/News/2023-12-30/Where-Have-Rental-Prices-Risen-Most-in-Portugal-This-Year/84670.

The Portugal News. (2023a, December 26). Rents up 10.5% in the third quarter. Https://Www.Theportugalnews.Com/News/2023-12-26/Rents-up-105-in-the-Third-Quarter/84506.

Tulumello, S., & Allegretti, G. (2020). Articulating urban change in Southern Europe: Gentrification, touristification and financialisation in Mouraria, Lisbon. European Urban and Regional Studies. https://doi.org/10.1177/0969776420963381

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79-82. doi:10.3354/cr030079

Zach. (2020, October 27). *Introduction to Multiple Linear Regression*. Statology.

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists. O'Reilly Media, Inc.

APPENDIX A

CP7	OFF BY €
21102	74194.02
21117	81959.69
30215	87079.34
31112	84594.67
60122	138031.08
70502	107342.50
80106	72048.81
80503	82340.10
80604	110386.84
80704	74118.06
80801	180172.51
81102	76619.84
81504	68737.16
100925	69570.61
110203	69191.26
110508	70132.11
110602	70758.38
110610	77241.27
110655	79900.16
110656	76538.87
110657	106916.96
110658	80329.65
110659	87420.15
110660	92661.14
110661	107147.83
110662	80971.41
110665	86376.67
110666	101733.89
111105	68323.06
111202	69819.26
130129	93240.42
130722	84797.77
131216	75706.70
141104	83481.46
141629	86831.83
150106	129523.63
150505	75710.87
160128	92465.08
160201	92429.75
160222	69603.61
160717	92504.33
160915	76050.44
171436	102474.03
180404	104911.95

Appendix A – Residuals by Geography

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

Universidade Nova de Lisboa