Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

MANAGEMENT INFORMATION SYSTEMS

MASTER'S FINAL WORK

DISSERTATION

ESTIMATING HOUSE MARKET PRICES: A DATA-DRIVEN APPROACH ON A PORTUGUESE DATA SET

MYKE DYLAN VAN INGEN PALMA

SUPERVISION:

RUI PEREIRA

DECEMBER - 2023

ABSTRACT

At the time of this research, the real estate market was experiencing an unprecedented surge in value, marked by consistent exponential growth with only a few notable exceptions. Predicting the value of a property could potentially serve as an initial step toward understanding this market, offering substantial advantages to prospective home buyers and sellers alike. This study aimed to estimate property market values through the application of data science techniques, utilizing both internal and external data sources associated with real estate properties.

In order to achieve the aforementioned objective, a meticulous process of data selection and collection was undertaken. A comprehensive and up-to-date real estate data source was leveraged, although it should be noted that it had certain data limitations. To ensure a consistent and structured approach to project progression, the Cross-industry standard process for data mining (CRISP-DM) model was applied.

Drawing insights from existing literature and the acquired data set, efforts were made to enhance the understanding of the real estate domain. Subsequently, data exploration was conducted by adopting Power BI, enabling a thorough examination of the data set. All data underwent aggregation and preprocessing using Python to facilitate the development and testing of various predictive models. Ultimately, the model demonstrating the highest accuracy was selected and integrated into a Flask web application, enabling user interactions with the developed predictive model.

This study, centered on the real estate market in Portugal, highlights both internal and external factors exert a significant influence on property values. Internal variables, such as the number of rooms, bathrooms, property type, size, and location, alongside external variables, including the Euribor rate and inflation, were found to have direct correlations with house prices.

Through rigorous model testing, the findings reinforced existing research: the Multilayer Perceptron algorithm emerged as the most effective model for estimating property values, demonstrating a remarkably low error rate, ranging from as little as 2,3% to a maximum of 16,3% in the worst-case scenario tested. Furthermore, it was observed that the input variables within this particular model exhibited interdependence, signifying that changes in one variable directly influenced the impact of other variables in the predictive model result.

KEYWORDS: Portuguese Housing Market, Machine learning, Neural Network

RESUMO

Aquando deste estudo, o mercado imobiliário estava a passar por um aumento de valor sem precedentes, marcado por um consistente crescimento exponencial, com apenas algumas exceções notáveis. Prever o valor de uma propriedade poderia servir como um passo inicial para entender este mercado, beneficiando substancialmente indivíduos interessados em comprar ou vender casa. Este estudo teve como objetivo estimar valores do mercado imobiliário através da aplicação de técnicas de *Data Science*, utilizando fontes de dados internas e externas associadas a imóveis.

De forma a atingir o objetivo acima mencionado, foi realizado um processo meticuloso de seleção e recolha de dados. Para isso uma fonte de dados de imobiliários recente e rica foi utilizada, embora deva ser salientado que, a mesma tinha a algumas limitações. Para garantir uma abordagem consistente e estruturada no desenvolvimento do projeto, foi aplicado o modelo *Cross-industry standard process for data mining* (CRISP-DM).

Tentou-se aprofundar a compreensão do domínio do mercado imobiliário, extraindo informações da literatura existente e do conjunto de dados adquirido. Subsequentemente, os dados foram explorados com recurso a *Power BI*, possibilitando assim, uma análise minuciosa deste conjunto de dados. Utilizando *Python*, todos os dados foram agregados e pré-processados de forma a facilitar o desenvolvimento e teste de vários modelos preditivos. Por fim, o modelo que demonstrou maior precisão foi selecionado e integrado numa aplicação web *Flask*, permitindo interações de utilizadores com o modelo preditivo desenvolvido.

Este estudo, que foi focado no mercado imobiliário português, revela que fatores internos e externos exercem uma influência significativa no valor dos imóveis. Verificou-se que variáveis internas como o número de quartos, casas de banho, o tipo de imóvel, dimensão e localização, em conjunto com variáveis externas, como a taxa Euribor e a inflação, têm correlações diretas com os preços das casas.

Resultados de testes rigorosos ao modelo corroboraram a literatura existente: o algoritmo *Multi-layer Perceptron* revelou ser o modelo mais eficaz para estimar o valor de imóveis, apresentando uma taxa de erro notavelmente baixa, variando apenas de 2,3% a um máximo de 16,3% no pior cenário testado. Além disso, observou-se que variáveis de *input* neste modelo em particular exibiam interdependência, portanto mudanças numa variável influenciavam diretamente o impacto de outras variáveis no resultado do modelo.

KEYWORDS: Mercado Imobiliário Português, Machine learning, Redes Neuronais

List of Figures

LIST OF TABLES

ACKNOWLEDGMENTS

Embarking on this academic odyssey and reaching the culmination of this dissertation has been a journey filled with challenges and triumphs. I am deeply grateful to the many individuals and organizations who have made this endeavor a reality.

I extend my gratitude to my supervisor, Rui Pereira, whose support, wisdom, and advice have guided this research. His mentoring has enhanced my growth during academic classes and inspired me to develop this study.

To the Lisbon School of Economics and Management, I offer my sincere appreciation for providing me with access to resources that have been essential in completing this dissertation.

A big thanks to all the data sources that provided data for this study. Without them, achieving the goals and findings would be impossible.

To my family, my profound appreciation for their unyielding support, patience, and love. Your belief in me has been my unwavering foundation, and your support has allowed me to focus on my studies.

I extend my deepest thanks to my friends and colleagues who have stood by me throughout this academic journey, offering their unwavering encouragement and reminding me of the importance of balance and resilience.

A special thanks to my colleague and friend Bruno Gonçalves, who was my primary motivation to never give up on pursuing the best output I could deliver.

1 Introduction

1.1 Context

As per data from INE (2023a), in the year 2022, there were 167,900 property transactions in Portugal, marking a 1.3% increase compared to the previous year. Concomitant with this upward trajectory in new transactions, it was observed that house prices surged by 13.1% within the same year. This surge in demand for new properties can also be attributed, at least in some part, to the impact of the COVID-19 pandemic (or post-pandemic), as it prompted a growing need for remote work and new houses (Mckinsey, 2023).

Nonetheless, it's important to note that the inflation rate has been on the rise since the beginning of 2022 (oecd, 2023), and there has been a noticeable uptick in the Euribor rate, commencing around July 2022 (Euribor, 2023b). These factors have collectively led to a reduction in the purchasing power of the Portuguese population. Given this economic and social context, it became imperative to comprehensively examine the conditions that could influence fluctuations in housing values and draw comparisons with historical instances to gain a deeper understanding of the evolving dynamics in the real estate market. This could help to predict whether we are getting closer to a new economic cycle known as a Bubble, which might soon burst in the real estate market and lead to a quick decrease in value, known as a "crash" or "bubble burst".

> "A bubble may be defined loosely as a sharp rise in the price of an asset or a range of assets in a continuous process, with the initial rise generating expectations of further rises and attracting new buyers" (Siegel, 2003).

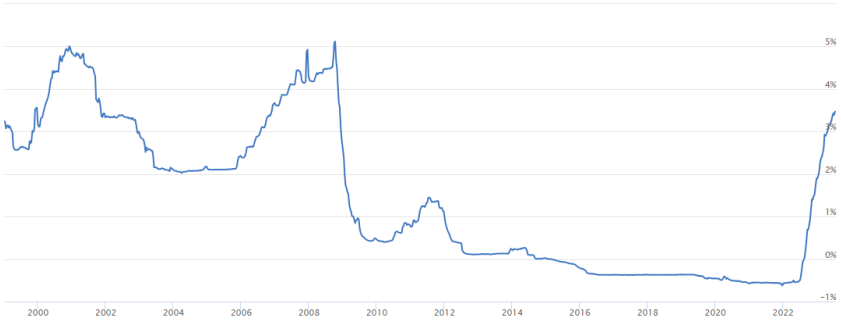


Figure 1: Euribor 1-month Tax rate from 1999 to 2023, (Euribor, 2023a)

As previously mentioned, the Euribor rate was at the time of writing of this report experiencing a significant increase, surpassing the rates seen during the recessions of 2000

and 2008. While it took about a year to rise to 2.61% in 2000 and three years to reach 3.26% in 2008, it has risen 3.8% in just one year, from -0.546 on 13 June 2022 to 3.35 on 15 June 2023 (Euribor, 2023b). However, despite this massive increase, many factors that caused previous recessions were not at play in the United States of America at the time of this study, which usually influences the rest of the world markets, such as subprime mortgages, low loan underwriting requirements, and exaggerated credit ratings. The banking sector is also better capitalized, which reduces the likelihood of a real estate market crisis (Afxentiou et al., 2022). Nevertheless, potential buyers should still consider factors such as the Euribor rate, Inflation, average Portuguese salary, and the characteristics of the houses, which can impact their value.

1.2 Motivation and Main Goals

Towards the end of 2022, I began contemplating the idea of moving out and potentially purchasing a house. However, my understanding of the housing market was limited at the time. As I searched for a better way to comprehend the market, I realized that no solution was available to help me toward my goal. No comprehensive system could accurately correlate and seek the impact of all variables available. It would be extremely useful for me and those interested in purchasing or selling a house if a software component could collect and use available data to estimate a house's market value.

The primary objective of this master's thesis was to comprehend the real estate market situation at the time of writing this report, mainly in Portugal, and generate a model that could estimate house market values with high accuracy. I aimed to create a prototype solution that utilizes available data and considers various variables to assist buyers and sellers in predicting market trends. Although my approach can be applied to any region in Portugal, I mainly focused my research efforts on the Lisbon metropolitan area. This decision was based on larger data availability and familiarity with the region, as I reside nearby. The Cross-industry standard process for data mining was used to have a more consistent procedure to achieve the goal.

1.3 Dissertation Structure

After the introduction, this report contains four more chapters.

In Chapter 2, the major themes and concepts relevant to the research are introduced, context to understanding the literature that follows is provided, and the existing research related to the dissertation topic is summarized and critically analyzed.

Additionally, in Chapter 3, the overall approach and design of the study are described.

The solution implemented is presented, the methods used to gather data are detailed, and the methods used to analyze the data collected during the study are described.

Chapter 4 is used to present the results of the data analysis in a clear and organized manner, using tables, graphs, or charts to illustrate key findings. Then, they were analyzed, and the results were interpreted in the context of the research objectives and literature review and discussed any unexpected or significant findings. At last, it explores the created model with real case scenarios and reveals its statistical accuracy.

Finally, in Chapter 5, the main research questions and objectives are recapitulated, the findings that address the research questions are summarized, the practical implications of the study's findings are discussed, and how they can be applied in real-world scenarios. The work that can be developed in the future is also put in perspective. Besides that, it also provided a concise summary of the dissertation's key contributions and significance.

2 LITERATURE REVIEW

The real estate market is somewhat volatile and influences most other markets. Understanding the existing literature would help fulfill the purpose of this study. Therefore, that was the first action to be taken. This chapter will explain the reality of the real estate market in Portugal and introduce some important concepts to understand the methodology and outputs of this study.

Returning to the past, it is possible to see some spikes in house market values (Shiller, 2007). However, these have gone down again. In 1950, caused by fears of war or terrorism. In 1970, fears related to the destruction of the environment (and the increase of agricultural land), and more recently, in 2008, mainly provoked by subprime mortgages (Ackermann, 2008). The correlations between house prices and the mentioned events served as a starting point for the study. Recent events such as the COVID-19 pandemic and the war in Ukraine might indicate a drop in the real estate market.

The final price of a house is mainly affected by the supply versus demand ratio, which is fairly stable in the short-term (Jacobsen and Naug, 2005). Nevertheless, an effect can be observed on the final house price when comparing it with the physical qualities and location of the house (Rahadi et al., 2015). An article by Hu et al. (2013) studied how properties like the number of rooms, number of bathrooms, and size, among other properties affected the final house price, and it concluded that size, bathroom count, having natural gas or a fireplace did not have a significant effect on the house price in linear regression models.

In Portugal, a considerable gap exists between demand and supply, resulting in unbalanced high property prices. There has been an oversupply of housing noticed as early as the year 2000, and in contrast to other countries, Portugal didn't have a collapse of house prices due to the growth of banking and construction sectors (Santos et al., 2015). According to Branco and Alves (2020), there is a lack of real estate regularization and rehabilitation policies in place. Employing non-market housing into the current market would boost supply, balancing the demand and supply ratio. An article by Mendes (2022) explored the governmental policies in place and concluded they are an obstacle to the modernization of the rental sector. Families with low income who can't afford to buy a house cannot afford rent and have no support. Policies in place take away house owners' confidence to rent their homes.

A qualitative analysis by Jover and Cocola-Gant (2023) explored the main causes of the increase in short-term rentals in Portuguese touristic areas. The increase in tourist demand (INE, 2023b), the disruption of the market caused by the COVID-19 Pandemic,

innovative apps like Airbnb and Booking.com, and governmental policies that favored house renting ended up increasing the rental sector in Portugal by 12,8%, a figure 5,3% higher than the average European growth (OECD, 2023).

To better understand the following sections of this dissertation, it is essential first to understand what a predictive model is. A predictive model is the process of creating a mathematical model that estimates the value or quantity of something of interest (Eduardo, 2023). There is a wide variety of predictive model types (indeed, 2023). The linear regression model is known as the simplest model and generates a predictive function with a linear combination of predictors (Su et al., 2012). On the other hand, the (artificial) Neural network model is one of the most complicated predictive models. Its complexity is usually compared with the neurons in a brain (Kriegeskorte and Golan, 2019), consisting of a mixture of node layers of 3 different types (input layer, hidden layers, and output layer), each with an arbitrarily selected amount of nodes (also known as artificial neurons) (IBM, 2023), illustrated in Figure 2. These two models are used with multiple inputs and a single output.

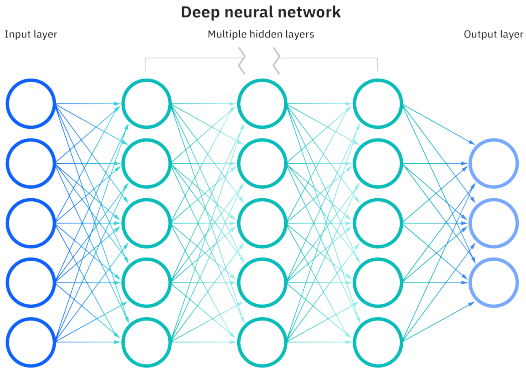


Figure 2: Neural Network (IBM, 2023)

Multiple studies have been conducted to try to find the best machine learning model to predict the real estate market. An article by Zulkifley et al. (2020), which aggregated data from multiple studies related to using machine learning to predict house prices, found that a house's location and structural attributes are the most important when predicting its price. This article also found that the most precise algorithms were the Artificial Neural Network, Support Vector Regression, and XGBoost.

Mohd et al. (2020) used multiple machine-learning models to find if using green materials was a relevant attribute to raise the value of a house. It concluded that it wasn't a relevant variable.

A study by Madhuri et al. (2019) compared regression techniques that help predict

the price per square meter of the real estate market and found that the most accurate algorithm was the gradient boosting algorithm. Varma et al. (2022) used a mixture of machine learning algorithms and neural networks to increase the efficiency of a house price prediction algorithm and understand if it is a good investment.

Another study realized by Zhang (2023) to find the most accurate model for predicting the house market value of a house in Chicago concluded that the best algorithm for this purpose was a Neural Network (Multi-layer Perceptron regressor). The output obtained was identical when using data sets from Hong Kong's (Abidoye et al., 2019) or Italian (Rampini and Cecconi, 2021) housing markets when using Artificial Neural Networks, Elastic net, XGBoost, and Support Vector Machine models for predicting property prices.

The Multi-layer Perceptron regressor (scikit learn, 2023a) is a supervised learning model that creates a neural network to predict its outputs (Taud and Mas, 2017). The number of layers and perceptions per layer is arbitrarily selected, affecting its precision. Being able to fiddle with the number of layers and perceptrons makes it extremely versatile for most practical problems (Murtagh, 1991).

A research was conducted by Shanker et al. (1996) to understand the effects of statistical standardizing of data when using a neural network model. The conclusion was that data standardization increased the model's accuracy, although this method's benefits decreased with larger networks and sample sizes.

Following a consistent process model is imperative to realize a data science project. The Cross-industry standard process for data mining (CRISP-DM) is the industry standard process model for applying data mining projects (Schröer et al., 2021). It consists of 6 steps:

1. Business Understanding - Before starting a project, evaluate the business situation and identify the data mining objective. Create a comprehensive project plan to ensure a smooth execution.

2. Data Understanding - Collect various data sources, analyze, and verify their quality. Use statistical analysis to identify attributes and correlations for the data mining objective defined in the Business Understanding phase.

3. Data Preparation - Data quality can be enhanced by setting standards for data selection and utilizing effective cleaning techniques.

4. Modeling - When faced with a business problem and a set of available data, selecting a data modeling technique best suited for the task is essential. Once that

decision has been made, it is crucial to establish specific parameters for the model and evaluate it against a set of criteria to determine the optimal approach.

5. Evaluation - Results are checked against business objectives during the evaluation phase. Interpretation and action planning follow. The entire previous process must also be reviewed.

6. Deployment - The deployment phase could involve either a final report or a software component.

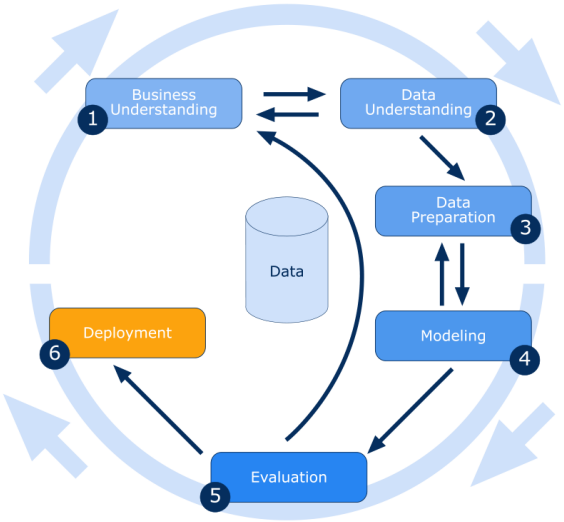


Figure 3: CRISP-DM process (linkedin, 2023)

3 METHODOLOGY

As previously mentioned, the main goal of this project was to develop a predictive model that would estimate the market value of a house. To achieve this goal, a need to understand the market rose. It was required to get data about real houses, and respective variables that could influence their market value were a must. After contacting several data sources, one seemed to fit this purpose well. This Chapter will describe the steps taken to reach the final goal.

The CRISP-DM process model helped to create smaller steps needed to reach the end goal of this study and ended up being a crucial part of reaching the goal. A quantitative research approach was used to analyze the accuracy of the project's output and assess its performance through statistical analysis (Sukamolson, 2007).

3.1 Data

The first step of the journey was to understand the real estate market. The existing literature was insufficient to get clear outputs of how the market worked. If it were possible to get data from different data sources, it would expand the knowledge from the literature and gain a deeper understanding of where the market stood at the moment of research.

To make that data gathering possible, a search for data source providers began. Relevant data for this project was collected from several known data source providers, such as:

- Pordata - Provides data of multiple areas of society, for Portugal and its municipalities, and the European countries;
- OurWorldIndata - Provides data related to changing quality of life conditions around the world;
- OECD Data - Provides data on several different areas regarding Organisation for Economic Co-Operation and Development countries (but not only).

To make it possible to further dig into their data, they were reached via email to explain the core of this project and request the use of their data (public or not) and, therefore, for them to provide access to their APIs. As a result of these contacts, they provided access to data that was available on their websites, and they forwarded some research related to the housing market topic that was useful in understanding its picture at the time.

Data related to historical average house value and Inflation, Euribor rates, and average Portuguese salary, among others, were collected from OECD Data, the Euribor website, and Pordata, respectively. This data had a monthly and yearly granularity and served as a way to understand the historical evolution of the housing market.

The study entirely depended on getting actual data about houses. Several house-selling platforms were contacted explaining the purpose of the study, but it took a long time to receive any answers. The most exciting response was from Idealista. Idealista is a known online house-selling platform in Portugal, and after requesting resources through a form on their developers' website (Idealista, 2023b), access to their API was granted. This opened the door to access a data set that was probably a good fit for the need.

For the reasons mentioned before stating their relevance, similarly to the study of Zhang (2023), the chosen variables to feed into the model were:

- House location (latitude and longitude)
- Room count
- Bathroom count
- Property type
- House size
- Euribor tax rates
- Inflation

3.2 Instruments

In the pursuit of fulfilling the objectives outlined in this thesis, the acquired data was employed among the following set of tools that were discovered as a result of the research itself, professional work, or introduced as part of the Management Information Systems Master's:

- Power BI - A data visualization solution owned by Microsoft.
- Visual Studio 2022 - An integrated development environment.
- SQL Server Management Studio - A software application developed by Microsoft to administer SQL Server instances.
- Jupyter Notebook - A web-based interactive computing platform.
- Visual Studio Code - A source-code editor.

3.3 Method

Historical data was explored using Power BI. This allowed better visualization of the collected data and, most importantly, revealed relationships between the collected variables. In the first subsection of Chapter 4, there is a focus on exploring the past through analysis and research, showing the most relevant outputs, and giving the reader a brief understanding of the current state of the real estate market. Although, at first sight, there seemed to be a correlation between Inflation, the Euribor rate, and the housing market itself, this wasn't enough to create a model that could predict house pricing. The need for actual house prices arose to be able to study and understand the properties that may increase the value of a house.

After being given access to Idealista data it was confirmed that it would be a good fit for the need. They provided some documentation on using their API and the available data. The API access they offered was limited to 100 calls a month. Idealista's API was protected with OAuth 2.0 protocol (Auth0, 2023), and it was required to get a bearer token (Idealista, 2023a) to call the other endpoints. This token had a limited validity period. Therefore, it was necessary to get new tokens regularly to continue using their API. They had a GET endpoint where they made available information regarding houses for sale and for rent on their website (Idealista, 2023a). The information provided by this endpoint seemed reasonable for this study (an example of the data output is attached on page 37 of the attachments).

It was essential to clean the data before storing it because some data would only add noise (iguazio, 2023) to this study's purpose. Therefore, only some values were selected to be studied, such as House location (latitude and longitude), Room count, Bathroom count, Property type, House size, Price, and Price By Area.

Each call to the endpoint retrieved the last 50 houses at most, ordered by the date when the announcement was published. This data could be filtered by choosing geographical points (latitude and longitude coordinates), and it would return houses in a range of 1500m around that point, never overlapping each other as exemplified in Figure 4. To never exceed the limitation of the maximum number of API calls per month, 18 geographical points in Portugal were chosen (Figure 5) and were called every week. It's important to note that this data was actual data from that moment, and it wasn't possible to get historical data from earlier dates.

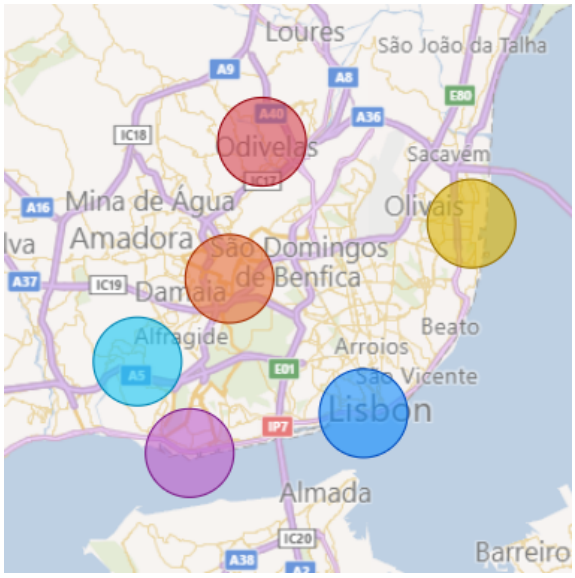


Figure 4: Data collection coverage area around geographical points

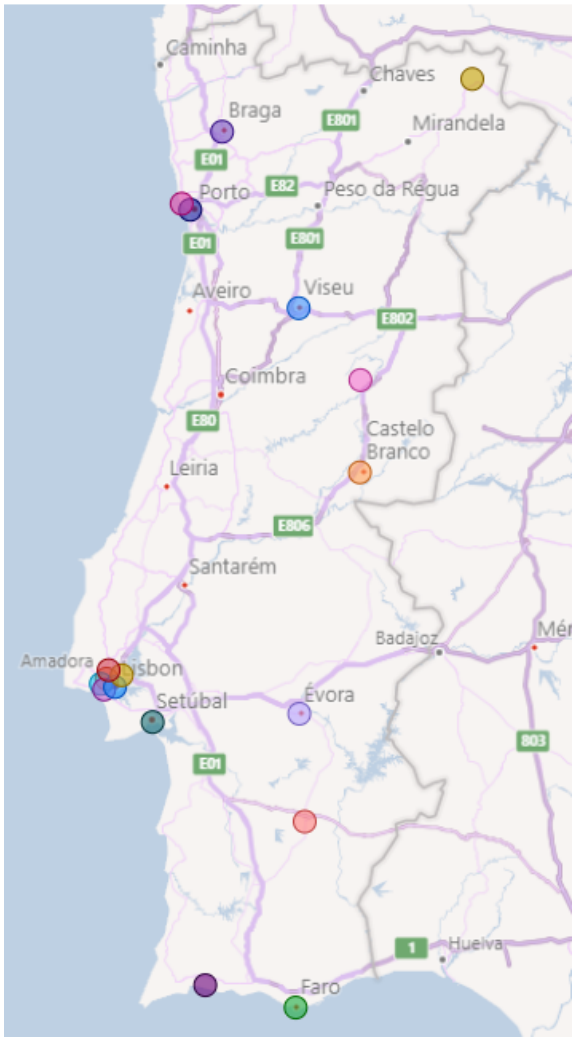


Figure 5: Geographical points requested

A solution was idealized to consistently obtain and store the data provided by that endpoint for some time to make it possible to understand the evolution of house pricing and overcome Idealista's API limitations. A Hosted Service was built with .NET 6 (Code, 2023) to make this task less manual. This framework was used for this purpose because it is fast, reliable, already used on many open-source projects, and was well-documented (Microsoft, 2023a).

> ".NET is a free, cross-platform, open-source developer platform for building many kinds of applications. .NET is built on a high-performance runtime that is used in production by many high-scale apps." (Microsoft, 2023b)

The version chosen was 6 because it was the Long-Term Support at the time. The purpose of this application was to call Idealista's GET endpoint using HTTP requests and save the retrieved data on a database. SQL Server relational database with the ORM Entity Framework Core was used because of the consistency of the data requests (Sahatqija et al., 2018). The database schema is presented in Figure 6.

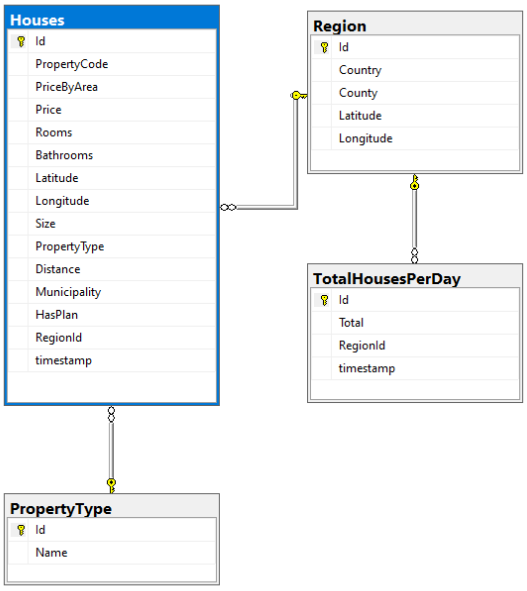


Figure 6: Database schema

Automating this process would simplify data collection within a considerable period of time, which was required for the study to have a more significant and meaningful dataset. With the automated process described before, data was gathered during 3 months, starting April 2023 and ending July 2023.

After that data-gathering period, it was time to explore and understand that data (to complete the Data Understanding step of CRISP-DM). Using Power BI, several charts were created that allowed the visualization of each variable's direct impact on the final price of a house and revealed patterns related to this influence. These outputs served as a guide to understand if the later generated model was consistent with these patterns.

After this, a Python Jupyter Notebook (jupyter, 2023) was created to explore the available data further and later create the predictive models.

To prepare the data (Data Preparation step in CRISP-DM), all acquired data had to be aggregated by date, which meant for each house for sale announcement, there was the need to know the Euribor and Inflation rate at that time. Dataframes from Pandas Library (pandas, 2023) facilitated this process. Then charts were generated with Seaborn (seaborn, 2023) and Matplotlib (matplotlib, 2023) Libraries to understand the correlation between each variable, as exemplified in Figure 24.

Machine learning models consist of statistical models, which meant numerical data needed to be fed to create a model, so all numerical data gathered was used to train the model (Modeling step in CRISP-DM). Of all the collected data, the input parameters that seemed more relevant were chosen to feed the model. Those were: the number of bathrooms, number of rooms, property type, house location (latitude and longitude coordinates), house square meter size, house price, Euribor rate, and Inflation rate. Standardizing the data resulted in better precision and faster processing of most models.

Data was randomly distributed into test and train groups in a 25% to 75% ratio. Then, using Sklearn (scikit learn, 2023b) Library Pipelines, it was possible to retrieve each of the model's accuracy in predicting house prices. The following table presents the precision of each model tested with the same train/test data sets:

Table I: Evaluation of the models' precision in predicting house prices

Model	**Accuracy on training data**	**Accuracy on test data**
Linear Regression	0.341	0.348
Ridge Regression	0.341	0.348
Lasso Regression	0.341	0.348
BayesianRidge Regression	0.341	0.348
svm Regression	-0.075	-0.090
SGDRegressor	-1080710783.223	-86050707.779
KNeighborsRegressor	0.811	0.672
GradientBoostingClassifier	0.037	0.032
MLPRegressor	**0.873**	**0.769**
RandomForest	0.577	0.585

Linear and non-linear models were tested, but this project focused on the non-linear Deep learning model Multi-layer Perceptron regressor. (Evaluation step in CRISP-DM) The main reason for selecting this model was related to the unpredictability of the housing market and the multiple variables that can affect it. It wasn't expected to have a linear evolution, which excluded using all linear models. Also, this model allows us to tweak the number of perceptions per layer to have a more precise model for the available data at the testing time. A Python code function was created to be able to find the best combination of perceptrons count per layer in the model for best accuracy. The same method was applied for the Random Forest algorithm by using a different Python function to find the most accurate number of estimators. Nevertheless, it wasn't as accurate as the one chosen.

To understand if the used model was accurate, the analysis of the impact of a variable in the final house price had to be used and compared side to side with its impact on the model output. To comprehend each variable's weight in predicting a house's value, charts that represented the variable's influence on the total value of a house were created. This was done by getting a house example from the data set, then tweaking only one of the variables at the time and, retrieving the final price of the house that the model estimated, then using this data to create a chart that could be used for the comparison.

Finally, Flask Framework was used to create a web page that possible house buyers or sellers could access to estimate their property's value (Deployment step in CRISP-DM). The users just needed to insert the variables, as shown in the attached Figure 25, the application connected to the generated model, and they would get an estimated house value. A complete diagram of the project's architecture can be found in Figure 7.

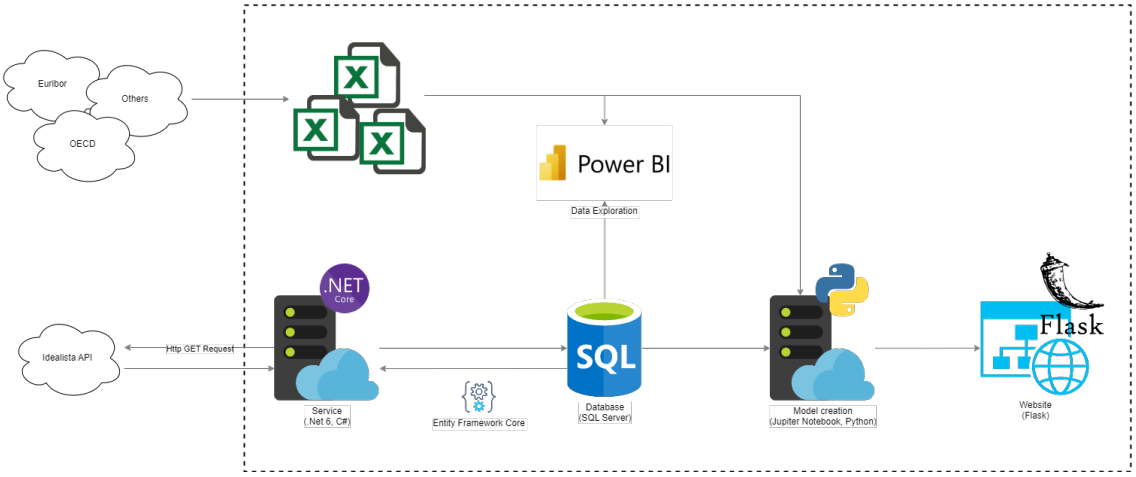


Figure 7: System Architecture

This Flask web app was then used with values from Idealista's API that weren't used to train the model. By comparing the real house price with the model's estimation, it was possible to verify the output of the model's accuracy in Table I.

4 Main Findings

As shown in the previous chapter, this study's best machine learning model was the Multi-layer Perceptron (Neural Network). However, the reasoning for choosing the variables used to feed the model has not been explained yet. This chapter will present the investigation done with the data in Power BI and Python regarding the real estate market and related data. Each data variable will be compared to the predictions the generated model makes. Finally, the chapter ends with real-case scenario tests to check the models' precision in estimating house sale price values.

4.1 A glimpse into the past

By exploring the world's past data (Inflation, Housing pricing, and Euribor % By Year) with Power BI, it was possible to create the following graph:

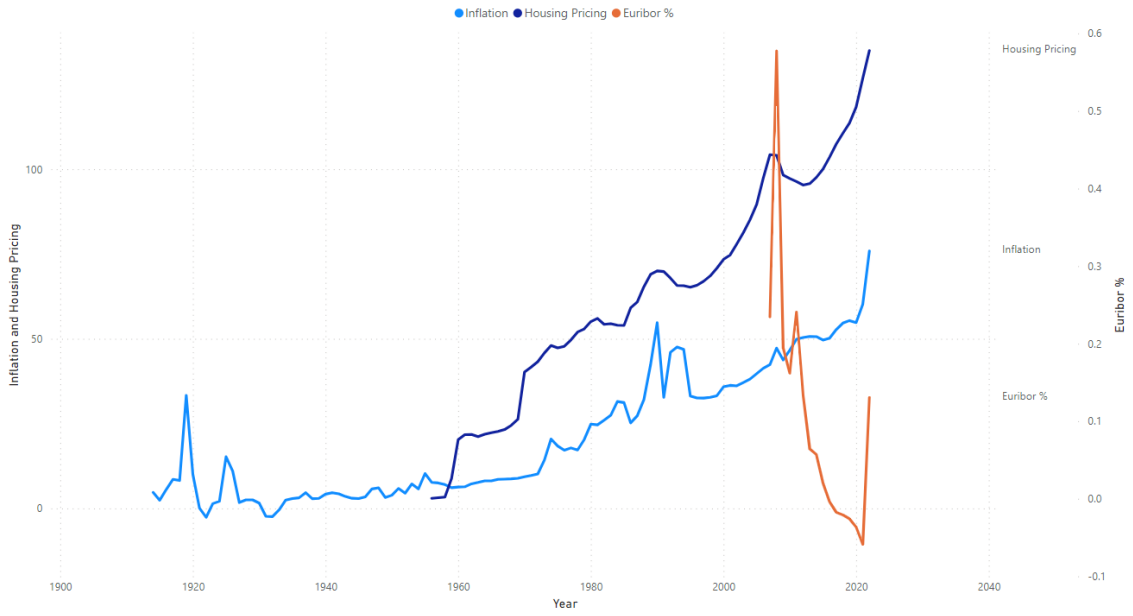


Figure 8: Inflation, Housing pricing, and Euribor % By Year

As shown in Figure 8, the housing pricing value and Inflation have been growing since 1980, with some minor downturns.

From this graph, it is possible to see that there seems to be a direct relation between the drop in the global housing market value of 2008 and the growth of the Euribor tax. This can be explained by the higher price of house mortgages and people being unable to pay for them and having to sell the houses. With more availability, the price tends to drop (Santos et al., 2015).

From a closer perspective, the Euribor tax has increased since July 2022. The last time a spike was seen like this, there was a massive drop in the housing market value - the famous subprime market crash of 2008. Another drop in the housing market is possible in the next couple of years. Nevertheless, since the start of the Euribor tax (1999), there has been only one drop, which doesn't provide enough data to predict if the next drop will happen shortly or if this tax will cause it. Also, there seems to be a slight relationship between inflation and the housing market value, where both tend to go up together. The article by Rehman et al. (2019) studied the effect inflation had on the real estate market in the United Kingdom, Canada, and the United States of America and found that there was an asymmetric relationship between Inflation and property prices in each country. Over the past few years, inflation has risen, so part of the house price increment may be caused by this.

The Portuguese newspaper Jornal Económico reported that the real estate market in Portugal slowed down in 2023. Fewer houses were purchased or sold, and prices rose due to increasing mortgage interest rates (Económico, 2023). Diário de Notícias's article states that house market prices grew 95.5% in the second trimester of 2010 to 2023, more than other countries in comparison, like Spain and France having 33% and 46.7% increases (de Notícias, 2023). It is possible to corroborate this finding by exploring the Portuguese market on INE's data source, confirming that the real estate market's value doubled over the last decade (INE, 2023c). The historical data on the average Portuguese salary shows an average growth of 3.98% per year over the same time frame (Pordata, 2023), which does not cover the amount that house prices have risen.

A study by Rodrigues (2022) expected that the COVID-19 pandemic would negatively affect all markets, but this study had different results. The house market value was at its highest value ever in history and hadn't been affected negatively by the COVID-19 pandemic and the war in Ukraine yet, a conclusion that is reinforced by the real estate market trends article from blogiad (2023). On the other side, an article by idealista (2023) presents the state of the rental sector in Portugal, and it is clear that this sector has had an impact, as the supply of houses for rent has decreased by 30%.

4.2 Understanding the data

Based on the information available in the houses data source, the number of houses listed for sale appears nearly ten times greater than those available for rent.

To choose which of the available variables to later feed into the model, it was necessary to try to understand which ones could have a direct relationship with the houses' selling prices. Lisbon's real estate market value has been rising for an extended period,

and understanding the difference between the most and least expensive zones gives a better perspective on how this market works. With real data taken from Idealista during the months of this study, it was possible to understand some patterns.

Regarding the houses for sale geographical distribution in Portugal, Figure 9 shows that between April and June 2023, most announcements were for houses located in coastal regions. Considering that, it was essential to understand if the data could have volatility over time. The attached Figure 26 has a zoomed-in perspective into Lisbon.

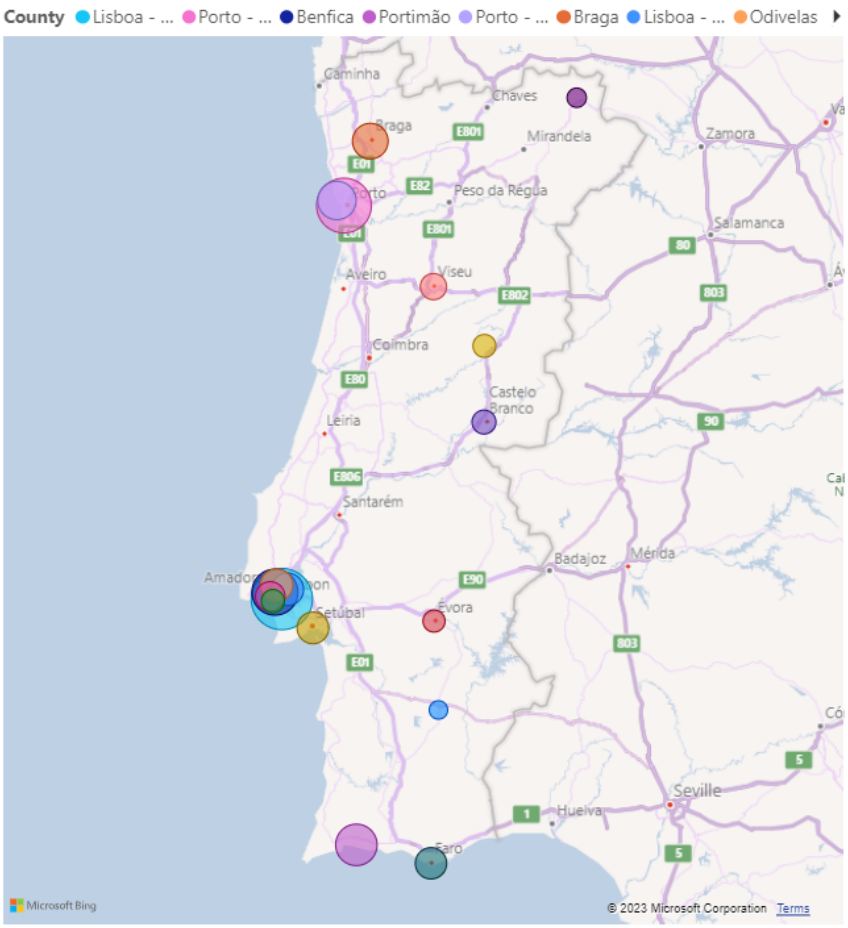


Figure 9: Houses for sale in Portugal in June 2023

The number of houses available for sale on the website during the data collection period seemed stable, with only slight fluctuations as shown in Figure 10. This could mean the supply and demand balanced each other during this period.

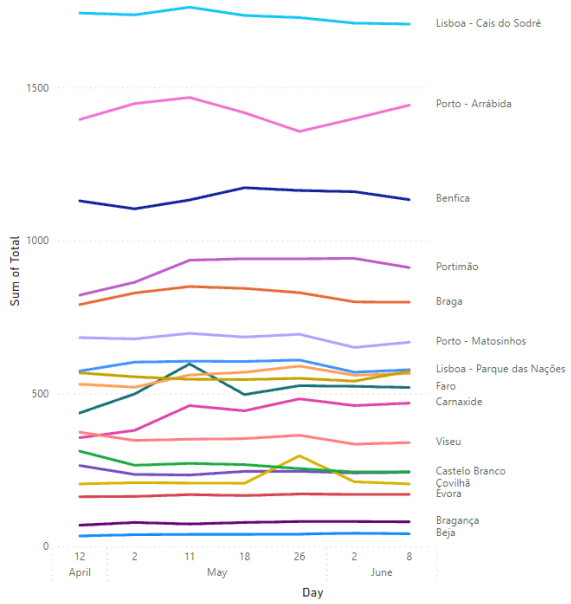


Figure 10: Total houses for sale by county

As seen in Figure 11, the average house prices considerably vary per region. This average price was calculated by selecting all houses of each specific region collected from the Idealista website and by getting the average price value of those houses.

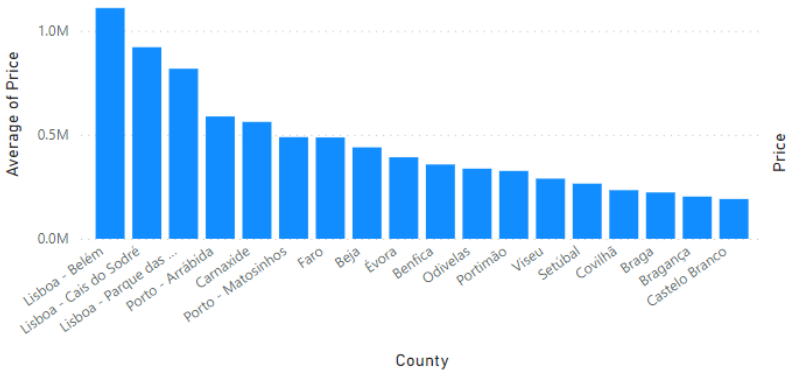


Figure 11: Average house prices per region in Portugal

Since the research focus was Lisbon, conducting an in-depth analysis of house prices was necessary. Figure 12 digs deeper into the prices of houses per region in Lisbon. Belem seemed to be the most expensive region studied, where the average cost of a house was more than a million euros. In contrast, Benfica and Odivelas were the least costly regions in this study. That said, house location seemed to be a significant metric to estimate a house's value in the created model.

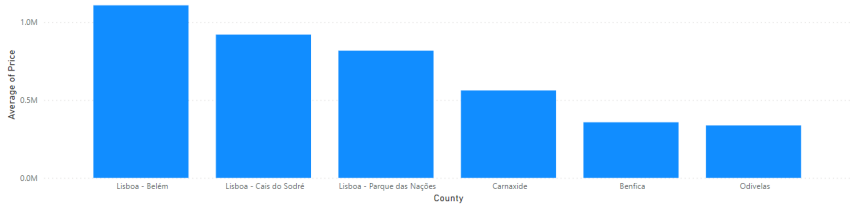


Figure 12: Average house prices per region in Lisbon

Six different property types were found when collecting data from houses in sale using Idealista's API. As expected, each property type had its own average house price, as shown in Figure 13. The most expensive being country Houses and the cheapest being studio.

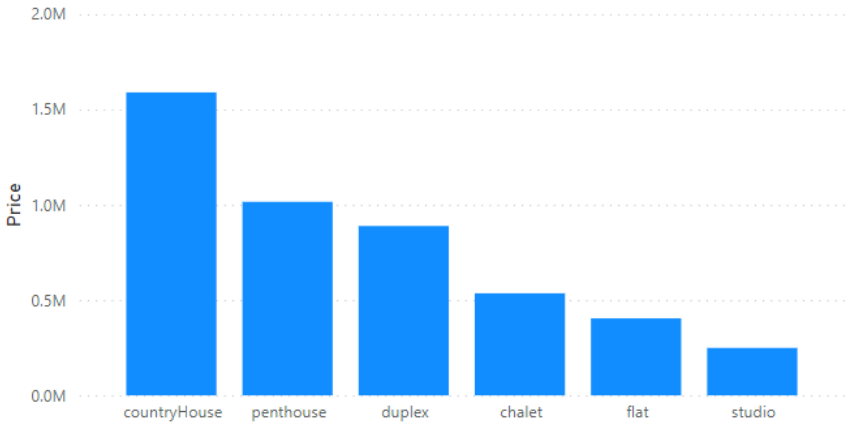
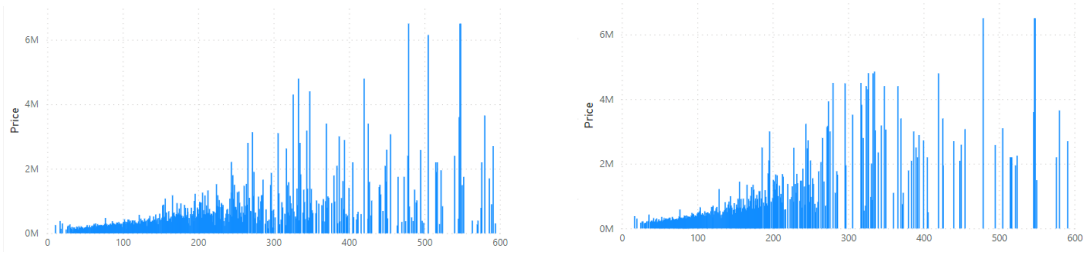


Figure 13: Average house price by property type

The size of a property has almost a linear evolution. The bigger the house, the more it costs. Figure 14 compares average Portugal values versus Lisbon's values. Both patterns are similar.



(a) Average house price per size in Portugal (b) Average house price per size in Lisbon

Figure 14: Average house price per house size in Portugal vs. Lisbon

Regarding the number of rooms per house, the value seems to rise until it reaches a

constant value of around 2.5 million euros. There was potential in using this variable to feed into the created model. Figure 15 presents the relation between the average price and the number of rooms in a house in Lisbon.

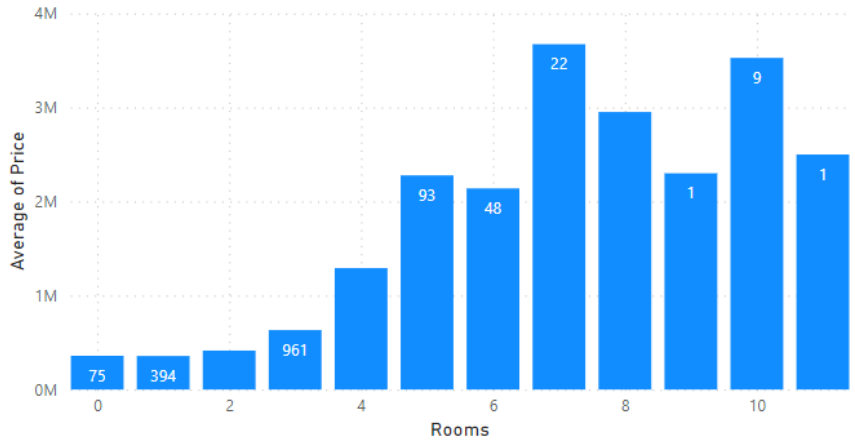


Figure 15: Average house price per room count

Lastly, there seemed to be a direct relation between the average price of a house and the number of bathrooms, where the number of bathrooms trends up the average price, as shown in Figure 16. Therefore, this was another variable chosen to feed the model.

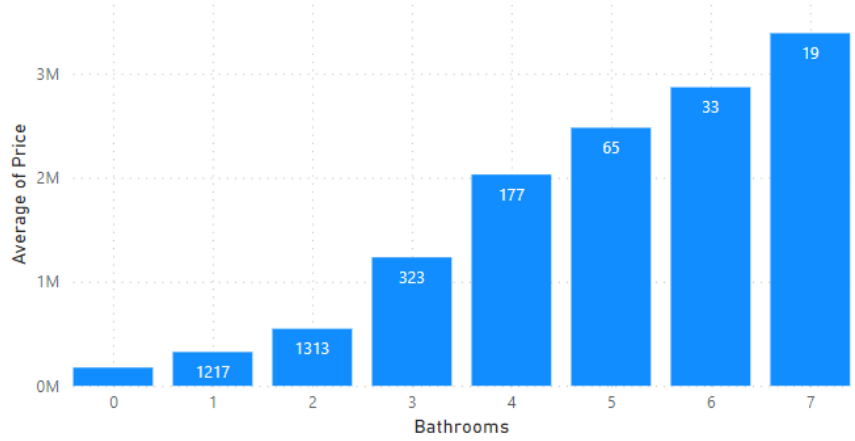


Figure 16: Average house price per bathroom count

4.3 Collected data versus model overview

In this section, a deeper analysis of each of the chosen variables will be presented. For each variable, there will be shown three charts:

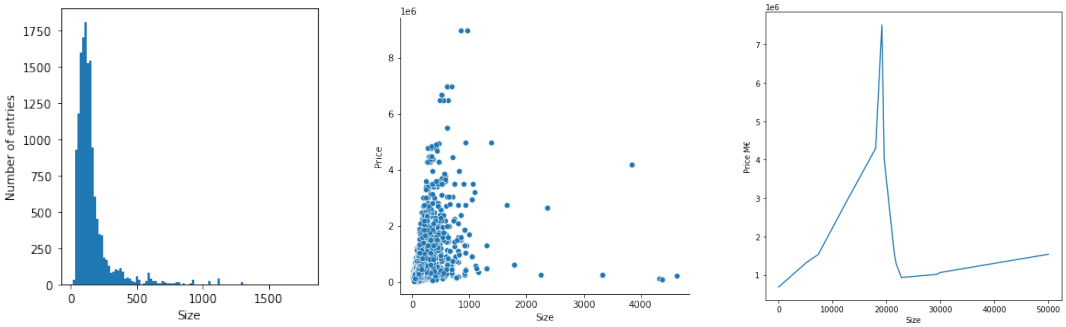
- The chart to the left presents the variable histogram, which allows the understanding of the data distribution for the variable.

- The middle chart presents the variable values distribution versus their relative prices in the real-case scenario.
- The final chart presents the influence the selected variable has on the created model when predicting the house price. This was made possible by changing the selected variable while keeping all other variables constant.

A randomly chosen house was used to create the charts on the right. The house base values were the following: rooms = 1, bathrooms = 1, latitude = 38.7157204, longitude = -9.1494941, size = 100, property type = 5, Euribor (1 month) = 0.0339, Inflation = 117.5722

4.3.1 House size

Figure 17a shows that most houses in the database are below 250 square meters. Figure 17b compares price and house size, which doesn't look linear because bigger houses in the countryside cost less than smaller houses in the big cities. Finally, the created model predicts a significant spike in value around 20,000 square meters after it gets very low and gradually grows again, probably due to a lower sample data.

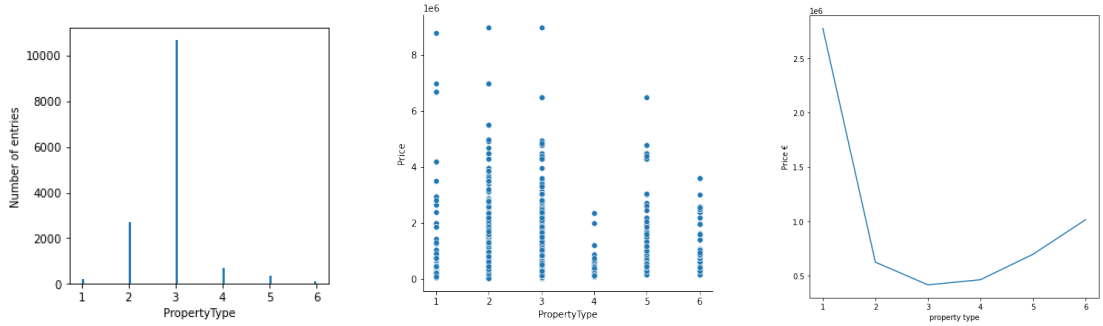


(a) House size Histogram (b) House size vs Price (c) House size Model

Figure 17: Variable weight in the model output: House size

4.3.2 Property type

Figure 18a shows that the distribution by property type of the collected data was: 210 Country Houses, 2685 Chalets, 10669 Flats, 721 Studios, 341 Duplexes, and 136 Penthouses. There doesn't seem to be a relevant pattern in price regarding property type in Figure 18b. Lastly, the model predicts that country houses are the most expensive, while flats are the cheapest.



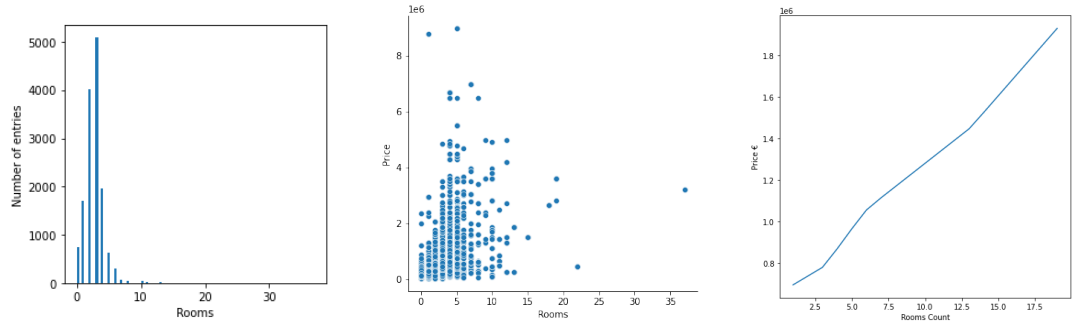
(a) Property type Histogram (b) Property type vs Price (c) Property Type Model

Types: 1 - Country House, 2 - Chalet, 3 - Flat, 4 - Studio, 5 - Duplex, 6 - Penthouse

Figure 18: Variable weight in the model output: Property type

4.3.3 Rooms count

From the collected data, a total of 9105 houses had 2 or 3 rooms, 4013 had two rooms, and 5092 houses had three rooms, and they represent 61,6% of all houses, represented in Figure 19a. No clear outputs could be taken from the house price per room count (figure 19), but the generated model seems to have a close to linear evolution: the more rooms, the higher the house cost.

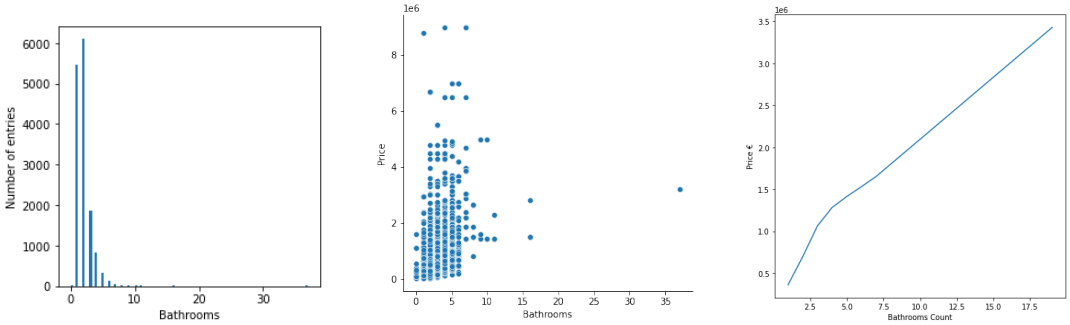


(a) Rooms Histogram (b) Rooms vs Price (c) Rooms Model

Figure 19: Variable weight in the model output: Room count

4.3.4 Bathrooms count

Regarding the number of bathrooms, a total of 11570 houses had one or two bathrooms, 5465 houses had one bathroom, and 6105 had two bathrooms, and they represented 78,3% of all houses, which can be observed in Figure 20a. From Figure 20b there didn't seem to be a pattern, but the generated model appears to have almost a linear evolution, starting with a steeper slope until around 3 bathrooms and then having a close to linear growth.

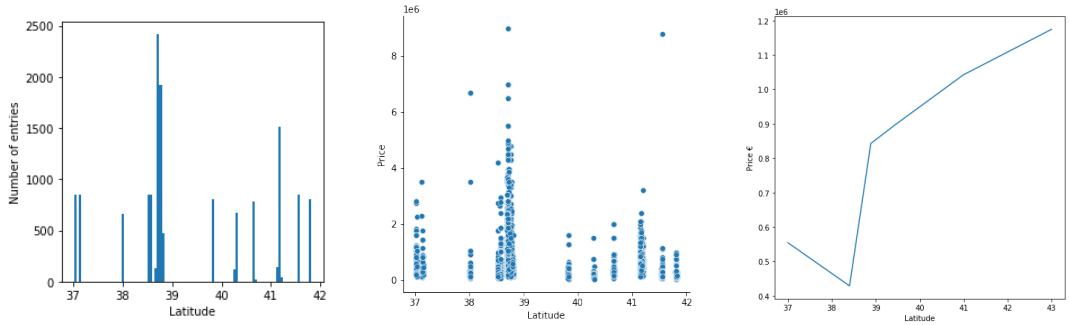


(a) Bathrooms Histogram (b) Bathrooms vs Price (c) Bathrooms Model

Figure 20: Variable weight in the model output: Bathroom count

4.3.5 Latitude

No clear output could be taken from the Latitude Histogram chart (Figure 21a). But when comparing the house price for latitude, there seemed to exist a spike in house prices around 38.5 and 39 degrees, representing Lisbon, Setúbal, and Évora regions. The model predicted otherwise: a growth in the house price the higher the latitude, which meant the house price rose the further north it was.

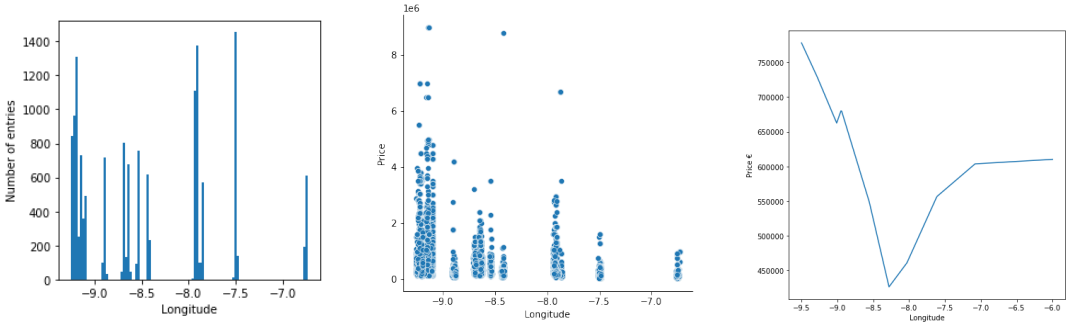


(a) Latitude Histogram (b) Latitude vs Price (c) Latitude Model

Figure 21: Variable weight in the model output: Latitude

4.3.6 Longitude

The Longitude Histogram (Figure 22a) shows that most houses for sale were under -9 degrees, which represents the left coastal region of Portugal, corroborating the conclusions taken from Figure 9. The priciest houses were in the coastal region when comparing longitude and price. The generated model also represented this, falling until around -8.3 degrees, which is the middle of the country, going back up until becoming almost stable when getting closer to the east border of Portugal.

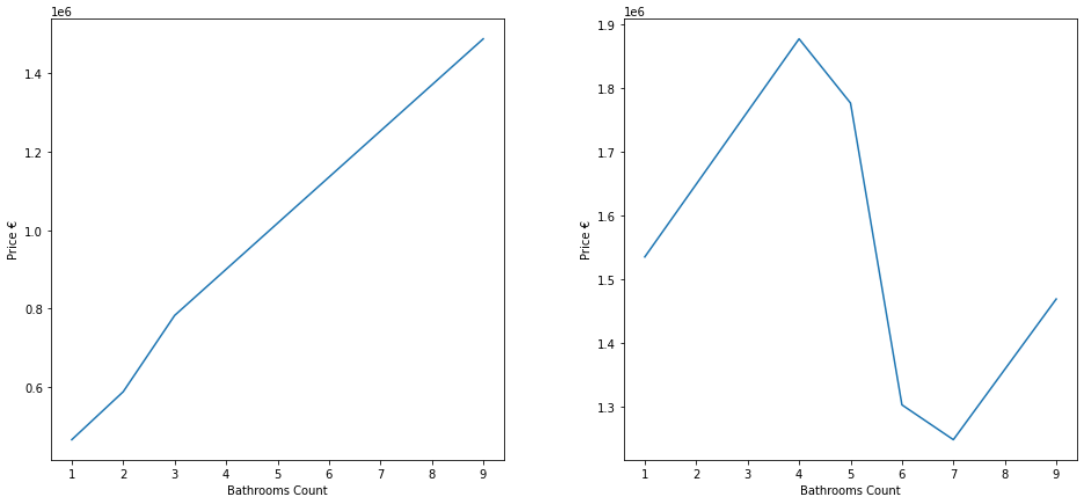


(a) Longitude Histogram (b) Longitude vs price (c) Longitude Model

Figure 22: Variable weight in the model output: Longitude

4.3.7 Variable interrelationships

This analysis of the influence of variables in the model versus the actual data showed some interesting facts related to the model's expected predictions. For example, the more rooms, the higher the house cost. However, the base values of the tested house mentioned at the start of this section do influence the model predictions of each variable. To try this, the same effort to understand the bathroom count impact on the model was made, but changing the room count and testing with one and fifteen Rooms while keeping all other variables constant. Represented in Figure 23, a considerable difference in the charts can be seen. Although it is an exaggerated and non-real-case scenario, it confirms that each variable affects the other variables.



(a) Houses with 1 Room (b) Houses with 15 Rooms

Figure 23: Room count weight on Bathroom count influence on house price estimation

4.4 Real case scenario testing

This subsection analyzed some real case scenarios using the created Flask web app. These tests were made to try to understand if the desired goals were achieved and if it was possible to, by providing some input parameters, have an accurate estimation of the actual sale value of a house.

That being said, after the solution was built and the model created, four random houses for sale were chosen by calling Idealista's API (on 13 September 2023) in three different regions, and the last houses posted in those regions were used. It was verified that the model wasn't fed with these houses before to avoid biased outputs. The test's output value is not representative of the model accuracy because a more significant amount of tests would need to be performed to guarantee that the accuracy is really the expected. Nevertheless, this could give some insight into how close to reality the model's predictions were. The idea was to compare the real sale value and the model prediction of the same house. The four test case results are shown in Table II, which contains the location, the model estimated value of the house, the sale price on Idealista's website, and the relative difference.

Table II: Model real case scenarios

Location	**Model estimation**	**Real price**	**Relative difference**
Lisbon (1)	1,378,214€	1,280,000€	107,7%
Lisbon (2)	2,427,930€	2,900,000€	83,7%
Évora	215,849€	189,900€	113,7%
Vila Nova de Milfontes	307,740€	315,000€	97,7%

The error ratio was calculated as follows:

$$RelativeDifference = EstimatedPrice/HousePrice * 100$$

Two big findings were drawn from these real test case scenarios. The most evident one is that it is possible to create a machine learning algorithm with high accuracy when estimating real estate market prices, as the generated model had missed, at worst, only 16,3% of the real price of the house, and doing this with a data set of only three months of data collection. Secondly, although only having a test for this scenario is not proof enough, predicting a house value in a region without data had a surprisingly good estimation. In the test case from Vila Nova de Milfontes, the model could predict, with a small error rate of 2,3%, a house located far away from any of the collected house regions.

5 Conclusions, Limitations and Future work

Although many results and discussions were presented in the last chapter, this final one will try to conclude if the primary goal set at the beginning of this study was achieved and why. A subsection will be followed to present the many limitations encountered during this study's development. Finally, possible solutions will be mentioned to help surpass some limitations and develop future work.

5.1 Conclusions

At the inception of this study, the real estate market had become a prominent and widely discussed subject. A pervasive sense of uncertainty prevailed due to the escalating nature of various financial factors, particularly the Euribor and mortgage interest rates, which, in turn, resulted in a significant increase in housing mortgage expenses. This led to some homeowners being unable to pay their new mortgages and being forced to put their properties for sale. This impacted the rental sector, causing a drop of 30% in the supply of houses for rent in the first trimester of 2023 (idealista, 2023). Which theoretically would make the sale prices drop due to the law of demand and supply, but this wasn't the case.

During that period, a viable solution for accurately predicting the selling price of a house based on its specific characteristics had not yet been established. Consequently, one of the principal objectives of this research was to create a system capable of not only aggregating available information but also learning from it, ultimately enabling predictions of the potential selling price of a given property.

To be able to develop this system, a lot of essential tools and technologies had to be used, such as Power BI, C# and .NET, Python (and Flask), HTML and CSS, and SQL, most of which were taught while taking the Management Information Systems Master's degree at ISEG.

The final result of all the practical work developed was a system capable of using a machine learning model that could estimate the value of a house with high accuracy (between 2,3% and 16,3% error rate), as presented in the last section of Chapter 4. Having this accessible to the public could bring significant benefits to anyone who wants to sell or buy a house because they would be presented with a reasonable estimate of the house sale value. That being said, it's possible to state that the main goal of building a system capable of accurate estimations regarding house sale prices was attained.

Nevertheless, this was more of a proof of concept, and some limitations and future work would need to be addressed before delivering a final solution ready for the end users. These subjects will be discussed below.

5.2 Limitations

As would be expected, this study had internal and external limitations. Many obstacles were faced and decisions had to be made to surpass them.

The first limitation was to obtain meaningful data to use in the study. It was not easy to find good data sources and to be able to get authorization to access that data. That proved to be a complex task. Although Idealista had a big enough data source for the purpose of this study, ideally, more data sources should have been used. Collecting real data from a single source presented a risk of having a somewhat biased model. If given the chance, data could have been taken from other sources like Imovirtual (Imovirtual, 2023) or Supercasa (Supercasa, 2023).

Idealista was kind enough to provide access to their API, but it had some limitations, as mentioned before. The API access they offered was limited to 100 calls a month. Each call to the endpoint would retrieve, at most, the last 50 houses. To overcome this, the API was called for a single property type (houses) on fewer regions than desired, only once a week.

As a result of time constraints, the data collection time range was a limitation that couldn't have been overcome for this study. Considering this, the three months (from April to July) of data collection were a reasonable time range to get a working prototype.

Regarding Euribor tax data, it was only available since 1999, given this was the year it was created. Therefore, it wasn't possible to confirm if this variable significantly impacted the real estate market drops and if it affected the 2008 drop.

5.3 Future Work

Although this study proved that it is possible to create a predictive model to estimate the market value of a house with good precision, improvements could be made regarding the system built. That being said, several steps could be considered.

A more extensive and diversified data set could help improve the solution because that would make the model more accurate. Besides that, it could also help to have other house characteristics that usually increase the value of houses, such as pools, garages, and elevators. Real estate data could be retrieved from multiple sources more frequently than once a week. Idealista's API limitation would need to be surpassed, and/or other new high-quality data sources would need to be found. The model would also need to be fed regularly with this new data.

After gathering a large enough data set, with a higher time range, it could be possible

to extend the generated model to be able to predict the future trend of the real estate market.

This solution was also built thinking about the Portuguese scenario since the data retrieved was from Portugal. Nevertheless, if good data sources from other countries were found, the system built could be easily updated to consider different countries. Besides all the current variables, the model would also need to be fed with the country where the house is being sold.

The most remarkable improvement after acquiring more real estate data sources would be to get different variables to feed into the model. Variables related to the house's region could also be considered if available. Data include pollution, population count, natural disasters, and regional conflicts.

This study tested multiple models, but only the most accurate was used to predict house market values. Additional machine learning algorithms could also be considered. Also, the Multi-layer Perceptron algorithm has endless combinations of layers and perceptrons per layer to improve the current model. This process requires a lot of computational resources.

Regarding the model's accuracy in estimating the house market value, more tests would need to be performed to ensure the best predictions.

In short, despite all the limitations and obstacles found while doing this work and the fact that there was still a lot of room for improvement, the core objective of this study was successfully achieved.

REFERENCES

[1] Abidoye, R. B., Chan, A. P., Abidoye, F. A., and Oshodi, O. S. (2019). Predicting property price index using artificial intelligence techniques: Evidence from hong kong. *International Journal of Housing Markets and Analysis*.

[2] Ackermann, J. (2008). The subprime crisis and its consequences. *Journal of Financial Stability*, 4:329–337.

[3] Afxentiou, D., Harris, P., and Kutasovic, P. (2022). The covid-19 housing boom: Is a 2007–2009-type crisis on the horizon? *J. Risk Financial Manag*, 15:371.

[4] Auth0 (2023). auth0.com. (Accessed August 27th, 2023).

[5] blogiad (2023). blog.iadportugal.pt. (Accessed October 6th, 2023).

[6] Branco, R. and Alves, S. (2020). Urban rehabilitation, governance, and housing affordability: lessons from portugal. *Urban Research & Practice*, 13:157–179.

[7] Code, R. T. (2023). www.roundthecode.com. (Accessed August 29th, 2023).

[8] de Notícias, D. (2023). www.dn.pt. (Accessed October 9th, 2023).

[9] Económico, J. (2023). jornaleconomico.pt. (Accessed October 6th, 2023).

[10] Eduardo, G.-P. (2023). Notes for predictive modeling.

[11] Euribor (2023a). www.euribor-rates.eu. (Accessed July 26th, 2023).

[12] Euribor (2023b). www.euribor-rates.eu. (Accessed July 26th, 2023).

[13] Hu, G., Wang, J., and Feng, W. (2013). Multivariate regression modeling for home value estimates with evaluation using maximum information coefficient. *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2012*, pages 69–681.

[14] IBM (2023). www.ibm.com. (Accessed September 23th, 2023).

[15] Idealista (2023a). api.idealista.com. (Accessed August 27th, 2023).

[16] Idealista (2023b). developers.idealista.com. (Accessed August 27th, 2023).

[17] idealista (2023). www.idealista.pt. (Accessed October 5th, 2023).

[18] iguazio (2023). www.iguazio.com. (Accessed October 6th, 2023).

[19] Imovirtual (2023). www.imovirtual.com. (Accessed September 15th, 2023).

[20] indeed (2023). www.indeed.com. (Accessed September 23th, 2023).

[21] INE (2023a). Instituto nacional de estatística. (Accessed July 26th, 2023).

[22] INE (2023b). www.ine.pt. (Accessed October 12th, 2023).

[23] INE (2023c). www.ine.pt. (Accessed October 6th, 2023).

[24] Jacobsen, D. H. and Naug, B. E. (2005). What drives house prices? *Economic Bulletin*, 5.

[25] Jover, J. and Cocola-Gant, A. (2023). The political economy of housing investment in the short-term rental market: Insights from urban portugal. *Antipode*, 55:134–155.

[26] jupyter (2023). jupyter.org. (Accessed October 6th, 2023).

[27] Kriegeskorte, N. and Golan, T. (2019). Neural network models and deep learning. *Current Biology*, 29:231–236.

[28] linkedin (2023). www.linkedin.com. (Accessed September 30th, 2023).

[29] Madhuri, C. R., Anuradha, G., and Pujitha, M. V. (2019). House price prediction using regression techniques: A comparative study. *IEEE 6th ICSSS 2019*.

[30] matplotlib (2023). matplotlib.org.

[31] Mckinsey (2023). empty-spaces-and-hybrid-places. (Accessed July 26th, 2023).

[32] Mendes, L. (2022). The dysfunctional rental market in portugal a policy review. *Land*.

[33] Microsoft (2023a). dotnet.microsoft.com. (Accessed August 25th, 2023).

[34] Microsoft (2023b). learn.microsoft.com. (Accessed September 23th, 2023).

[35] Mohd, T., Jamil, S., and Masrom, S. (2020). Machine learning building price prediction with green building determinant. *International Journal of Artificial Intelligence*.

[36] Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, page 183–197.

[37] OECD (2023). data.oecd.org. (Accessed October 12th, 2023).

[38] oecd (2023). Organisation for economic co-operation and development. (Accessed July 26th, 2023).

[39] pandas (2023). pandas.pydata.org.

[40] Pordata (2023). www.pordata.pt. (Accessed September 23th, 2023).

[41] Rahadi, R. A., Wiryono, S. K., Koesrindartoto, D. P., and Syamwil, I. B. (2015). Factors influencing the price of housing in indonesia. *International Journal of Housing Markets and Analysis*, 8:169–188.

[42] Rampini, L. and Cecconi, F. R. (2021). Artificial intelligence algorithms to predict italian real estate market prices. *Journal of Property Investment & Finance*.

[43] Rehman, M. U., Ali, S., and Shahzad, S. J. H. (2019). Asymmetric nonlinear impact of oil prices and inflation on residential property prices: a case of US, UK and canada. *The Journal of Real Estate Finance and Economics*, 61:39–54.

[44] Rodrigues, P. M. M. (2022). The real estate market in portugal. *Estudos da fundação*.

[45] Sahatqija, K., Ajdari, J., Zenuni, X., Raufi, B., and Ismaili, F. (2018). Comparison between relational and nosql databases. *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0216–0221.

[46] Santos, A., Serra, N., and Teles, N. (2015). Finance and housing provision in portugal. *FESSUD Working Paper Series*, pages 1–58.

[47] Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

[48] scikit learn (2023a). scikit-learn.org. (Accessed September 21th, 2023).

[49] scikit learn (2023b). scikit-learn.org. (Accessed August 27th, 2023).

[50] seaborn (2023). seaborn.pydata.org.

[51] Shanker, M., Hu, M. Y., and Hung, M. S. (1996). Effect of data standardization on neural network training. *Omega*, 24:385–397.

[52] Shiller, R. J. (2007). Understanding recent trends in house prices and home ownership. *NBER Working Paper Series*.

[53] Siegel, J. J. (2003). What is an asset price bubble? an operational definition. *European Financial Management*, 9:11–24.

[54] Su, X., Yan, X., and Tsai, C.-L. (2012). Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4:275–294.

[55] Sukamolson, S. (2007). Fundamentals of quantitative research. *Language Institute Chulalongkorn University*, 1:1–20.

[56] Supercasa (2023). supercasa.pt. (Accessed September 15th, 2023).

[57] Taud, H. and Mas, J. (2017). Multilayer perceptron (mlp). *Lecture Notes in Geoinformation and Cartography*, page 451–455.

[58] Varma, A., Sarma, A., Doshi, S., and Nair, R. (2022). House price prediction using machine learning and neural networks. *IEEE Xplore*.

[59] Zhang, R. (2023). Determining the best model among candidate machine learning models for chicago suburb house price data. *ICACTIC 2023*.

[60] Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., and Ibrahim, I. (2020). House price prediction using a machine learning model: A survey of literature. *Modern Education and Computer Science*.

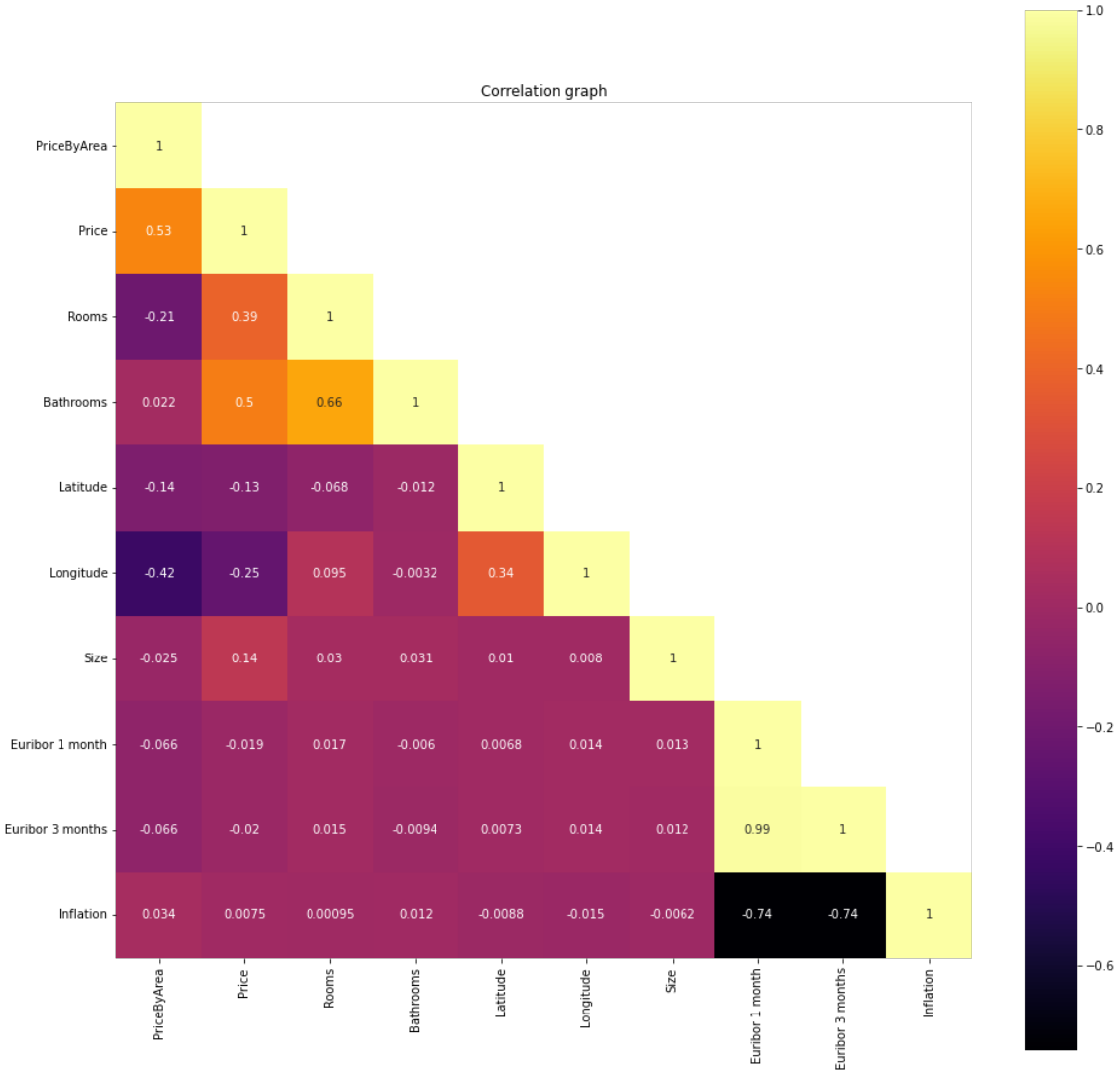


Figure 24: Variables Correlation

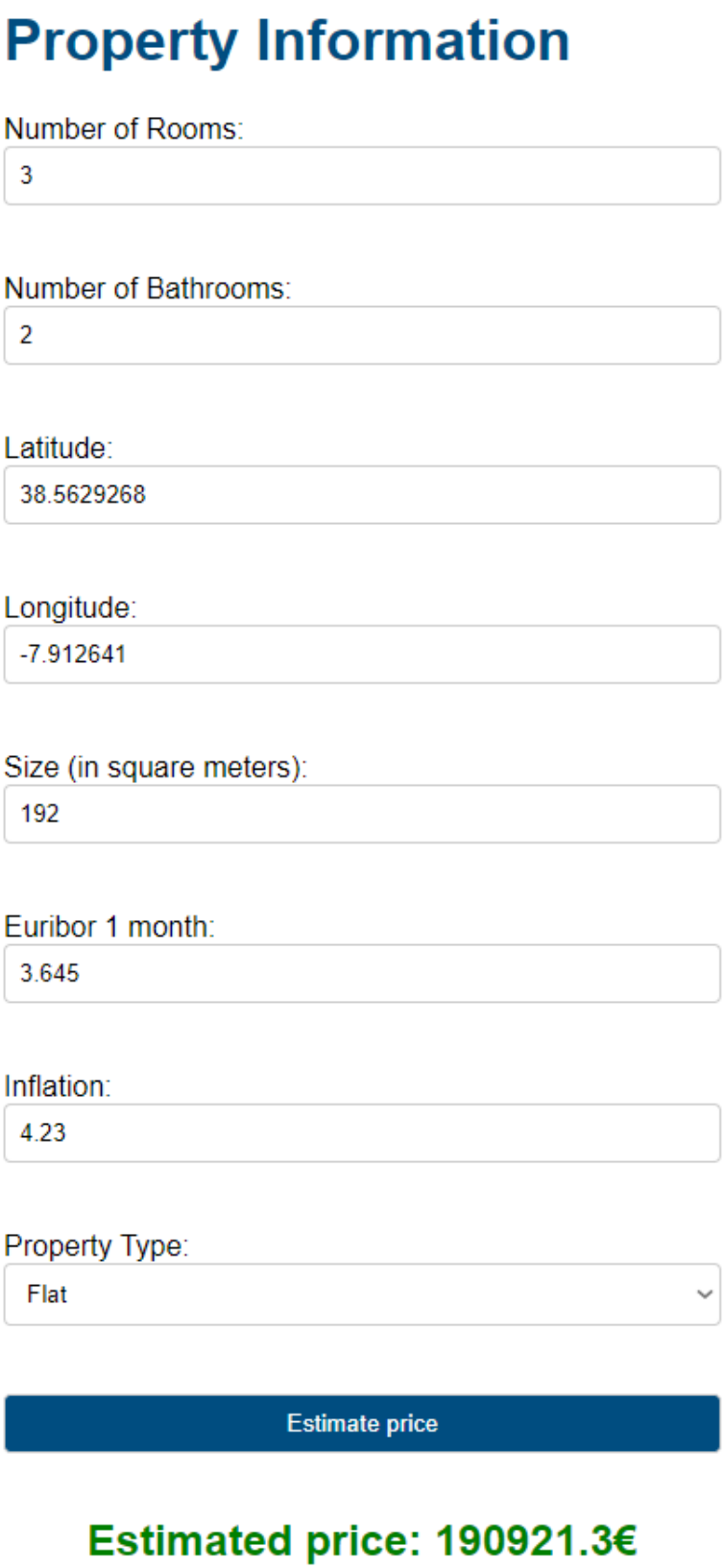


Figure 25: Created Website

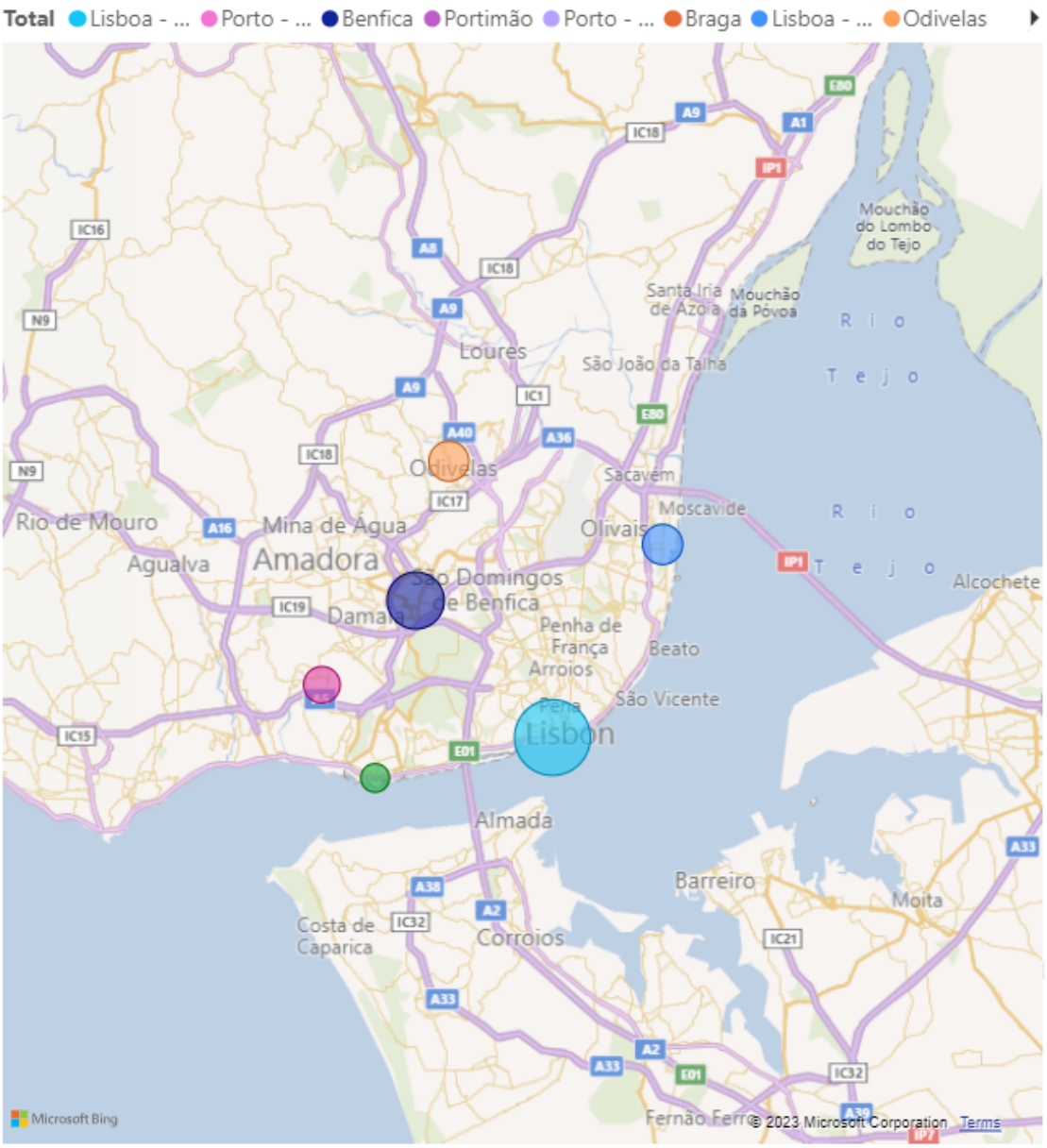


Figure 26: Houses for sale in Lisbon in June 2023

Idealista Search Endpoint Output Example:


```
{
    "elementList": [
        {
            "propertyCode": "32778360",
            "thumbnail": "URL",
            "numPhotos": 32,
            "floor": "2",
            "price": 189900.0,
            "propertyType": "flat",
            "operation": "sale",
            "size": 140.0,
            "rooms": 3,
            "bathrooms": 2,
            "address": "Rua Horta das Figueiras, 4",
            "province": "Évora",
            "municipality": "municipality",
            "country": "pt",
            "latitude": 38.5629268,
            "longitude": -7.912641,
            "showAddress": true,
            "url": "URL",
            "distance": "378",
            "description": "Description",
            "hasVideo": false,
            "status": "good",
            "newDevelopment": false,
            "hasLift": false,
            "priceByArea": 1356.0,
            "detailedType": {
                "typology": "flat"
            },
            "suggestedTexts": {
                "subtitle": "subtitle",
                "title": "title"
            },
            "hasPlan": true,
```

```
            "has3DTour": false,
            "has360": true,
            "hasStaging": false,
            "superTopHighlight": false,
            "topNewDevelopment": false
        }
        // Removed 49 results for simplicity
    ],
    "total": 27,
    "totalPages": 1,
    "actualPage": 1,
    "itemsPerPage": 50,
    "numPaginations": 0,
    "summary": [
        "Comprar",
        "Viviendas",
        "Malagueira e Horta das Figueiras, Évora",
        "Todos los precios",
        "Todos los tamaños"
    ],
    "alertName": "alertName",
    "totalAppliedFilters": 0,
    "lowerRangePosition": 0,
    "upperRangePosition": 27,
    "paginable": false
}
```

C# code snippet - Get the Houses from Idealista API

```
public async Task<IdealistaSearchObject> GetHouses(
    string bearerToken,
    decimal latitude,
    decimal longitude)
{
    string endpoint = $"{path}3.5/pt/search?
        operation=sale&
        propertyType=homes&
        center={latitude},{longitude}&
        country=pt&
        distance=1500&
        maxItems=50&
        order=publicationDate&
        sort=desc";

    _client.DefaultRequestHeaders.Authorization =
        new AuthenticationHeaderValue(
            "Bearer", bearerToken);

    HttpResponseMessage response =
        await _client.PostAsync(endpoint, null);

    string responseContentString =
        await response.Content.ReadAsStringAsync();

    IdealistaSearchObject idealistaSearchObject =
        _converters.DeserialiseIdealistaApiData(
            responseContentString);

    return idealistaSearchObject;
}
```

Python code snippet - Get houses data

```
con_string = 'DRIVER={SQL Server};SERVER=localhost;
DATABASE=HousingMarket'

connection = pyodbc.connect(con_string)

houses_query = """select [PropertyCode]
     ,max([PriceByArea]) AS [PriceByArea]
     ,max([Price]) AS [Price]
     ,max([Rooms]) AS [Rooms]
     ,max([Bathrooms]) AS [Bathrooms]
     ,max([Latitude]) AS [Latitude]
     ,max([Longitude]) AS [Longitude]
     ,max([Size]) AS [Size]
     ,max([PropertyType]) AS [PropertyType]
     ,max([Distance]) AS [Distance]
     ,max([RegionId]) AS [RegionId]
     ,truedate
     from  (SELECT *, CONVERT(DATE, [timestamp]) as truedate
         FROM [HousingMarket].[dbo].[Houses]) AS housedata
group by [PropertyCode], truedate"""

houses = pd.read_sql(houses_query, connection)
```

Python code snippet - Get Inflation and Euribor data

```
#Read data From CSV
euribor =
pd.read_excel("../3. HistoricalData/Euribor.xlsx")

housesWithEuribor = pd.merge(houses.assign(grouper=
    houses['truedate'].astype(str).str[:7]),
    euribor.assign(grouper=euribor['Data']
        .astype(str).str[:4] + '-'
    + euribor['Data'].astype(str).str[8:]),
                how='left', on='grouper')

housesWithEuribor =
housesWithEuribor.drop(columns=['grouper','Data'])

inflation =
pd.read_csv("../3. HistoricalData/Inflation.csv")

housesWithData = pd.merge(
housesWithEuribor
    .assign(
        grouper=housesWithEuribor['truedate']
            .astype(str).str[:7]),
inflation[['Value','TIME']]
    .assign(grouper=inflation['TIME']
        .astype(str).str[:7]),
how='left', on='grouper')

housesWithData =
    housesWithData.drop(columns=['grouper','TIME'])
```

Python code snippet - Data standardization

```
variables = [
    'Rooms',
    'Bathrooms',
    'Latitude',
    'Longitude',
    'Size',
    'Euribor 1 week',
    'Inflation',
    'PropertyType']

scaler = StandardScaler()

standarizedHouses =
    scaler.fit_transform(housesWithData[variables])
standarizedHouses = pd.DataFrame(standarizedHouses,
columns = housesWithData[variables].columns)
```

Python code snippet - Test/Train data splitting

```
X_train, X_test, y_train, y_test =
train_test_split(X, Y, random_state = 1, test_size= 0.25)
```

Python code snippet - Pipeline creation and result printing

```
pipelines = []
pipelines.append(Pipeline([('linear',
    linear_model.LinearRegression())]))
pipelines.append(Pipeline([('Ridge',
    linear_model.Ridge())]))
pipelines.append(Pipeline([('Lasso',
    linear_model.Lasso(alpha = .5))]))
pipelines.append(Pipeline([('BayesianRidge',
    linear_model.BayesianRidge())]))
pipelines.append(Pipeline(([('svm', svm.SVR())])))
pipelines.append(Pipeline([('SGDRegressor',
    linear_model.SGDRegressor())]))
pipelines.append(Pipeline([('KNeighborsRegressor',
    neighbors.KNeighborsRegressor())]))
pipelines.append(Pipeline([('GradientBoostingClassifier',
GradientBoostingClassifier(
    n_estimators=100,
    learning_rate=1.0, max_depth=1,
    random_state=0))]))
pipelines.append(Pipeline([('MLPRegressor',
    MLPRegressor(random_state=1,
        hidden_layer_sizes = (10,10,9,6),
        activation='relu',
        max_iter=5000, solver='lbfgs'))]))
pipelines.append(Pipeline([('RandomForest',
    RandomForestRegressor(n_estimators=4,
    max_depth=3, random_state=0))]))

for pipeline in pipelines:
    pipeline = pipeline.fit(X_train, y_train)
    print('model: {} - Accuracy on the training subset:
{:.3f}, test subset: {:.3f}'
    .format(
        pipeline.steps[0][0],
        pipeline.score(X_train, y_train),
        pipeline.score(X_test, y_test)))
```

Python code snippet - Model creation

```
print('Creating Model')
mlp = MLPRegressor(random_state=1,
    hidden_layer_sizes =
    (10,8,9,8),
    activation='relu',
    max_iter=50000,
    solver='lbfgs')

mlp.fit(X,Y)
print('Model Created')
```

Python code snippet - Model testing

```
rooms = 2
bathrooms = 2
latitude = 41.18386
longitude = -8.69622
size = 1
property_type = 5
timestampStr = '2023-07-30T13:54'
euribor1month = 0.0339
Inflation = 117.5722

new_data_scaled =
    scaleData(
        rooms,
        bathrooms,
        latitude,
        longitude,
        size,
        euribor1month,
        Inflation,
        property_type)

PredictedPrice =
    mlp.predict(new_data_scaled[variables])

print('Predicted value: {0}'
    .format(PredictedPrice[0]))
```

Python code snippet - Model testing

```
from random import randint
from flask import Flask, render_template, request

#random Port to be able to rerun
port = randint(0,99999)
app = Flask(__name__)

@app.route("/")
@app.route("/HousingMarketModel", methods=['GET'])
def show_form():
    return render_template(
        "/HousingMarketModel.html",
        rooms = 1,
        bathrooms = 1,
        latitude = 41.18386,
        longitude = -8.69622,
        size = 120,
        euribor1month = 0.0339,
        Inflation = 117.5722,
        PredictedPrice = 'To Be Calculed',
        propertyType = "3")

@app.route("/HousingMarketModel", methods=['POST'])
def TestModel():
    rooms = request.form['rooms']
    bathrooms = request.form['bathrooms']
    latitude = request.form['latitude']
    longitude = request.form['longitude']
    size = request.form['size']
    euribor1month = request.form['euribor1month']
    Inflation = request.form['Inflation']
    propertyType = request.form['propertyType']

    new_data_scaled = scaleData(
        rooms,
        bathrooms,
```

```
            latitude,
            longitude,
            size,
            euribor1month,
            Inflation,
            propertyType)

        PredictedPrice = mlp.predict(
            new_data_scaled[variables])

        return render_template(
            "/HousingMarketModel.html",
            rooms = rooms,
            bathrooms = bathrooms,
            latitude = latitude,
            longitude = longitude,
            size = size,
            euribor1month = euribor1month,
            Inflation = Inflation,
            PredictedPrice =
                str(round(PredictedPrice[0])) + ' euros',
            propertyType = propertyType)

if __name__ == "__main__":
    app.run(port=port)
```