Lisbon Housing Market Study: Price analysis and forecast

Diogo Marques (20231428@novaims.unl.pt, 20231428); Diogo Estevens (20231424@novaims.unl.pt, 20231424); João Mineiro (20231426@novaims.unl.pt, 20231426); Renato Poirier (20231427@novaims.unl.pt, 20231427); Sérgio Monteiro (20231430@novaims.unl.pt, 20231430).

Group. B3 **TP.**2

Abstract:In the last couple of years, the housing market prices in Lisbon have suffered a significant increase, being of the most expensive cities in Europe to buy a house. Several factors, such as inflation, foreign investment and market speculation have played their role in this sudden spike in prices. Some consider it temporary; others say it is permanent. Either way, due to the current market conditions, it is important to have accurate predictive models to understand housing prices in the current market. For the creation of an accurate predictive model, a real estate listing from a prominent housing website, Supercasa, will be used as our primary database. This dataset will have detailed information about the houses for sale in the Lisbon region, where we will have the number of rooms, area, location, and the type of building. Beyond the primary information, extras will also be analyzed; if the house has elevator, garage or even a swimming pool. The impact of these variables in the house pricing will be considered. Exterior factors, such as the proximity to educational, cultural sites and metro stations will also be considered since these variables can have an impact on the final price of a house. To answer our main question, machine learning algorithms will be applied. Web-scrapping, data preprocessing, data exploration and regression techniques will be covered in this report to build a predictive model capable of estimating a house price accurately, via predictive modelling, considering different intrinsic and exterior factors.

Keywords: Predictive Modelling; Web-scrapping; Housing Prices

Introduction

The Lisbon housing market prices have been one of the main topics of discussion in Portugal. Everyone from the average person to policymakers have been discussing the topic since the prices have risen to a level where an average person in Portugal is not able to afford a house in the Lisbon region. The average total salary in Portugal, in 2023, is of 1.505€ [1], and with the increase in mortgage interest rates from the ECB, the average Portuguese inhabitant must pay the value of 362€ per month to pay their mortgage [2] . These values show the weight of the house mortgage in the average Portuguese person's budget. So, to avoid the impact of possible additional speculation in the housing market, an accurate predictive model of the housing prices, considering several internal and external factors, is an important tool to understand if the pricing is correct and adapted to the current market. As shown in the next figure, according to the data obtained through the Supercasa dataset, which will be explained on the next chapters, the median square meter in Lisbon costs, on average, around 6000€, while the Portuguese median is less than 2000€ [3] , according to the Portuguese Statistics Institute (fig. 1).

To create such model, a database of the current houses on sale in the Lisbon area is needed. Since there are no official databases available from the Portuguese State Institutions, regarding the prices of current houses on the market, we need to extract our own database. For the acquisition of such database, web-scrapping techniques will be used to obtain the necessary data of the house listings on the Supercasa website. Main information such as area, number of rooms, price, location, and other factors such as having a garage, elevator or swimming pools will be at our disposal. The same technique will also be applied to extract information on cultural sites, metro stations and schools, to understand if the pricing of a house is related to the proximity to these sites. Since we are applying web-scrapping techniques to obtain our main database and several others, the raw data will not be 100% accurate. Meaning that a lot of data processing and exploration will be needed. Several different techniques will be used to clean and process our data, so it can be analysed correctly, since machine learning algorithm are very sensitive to bad data. After the data processing and exploration part, which is the most time-consuming part of the report, we are ready to start to apply machine learning algorithms and predict, with a certain level of certainty, which is the fair price of a house in the Lisbon region considering several factors, some usually not accounted when pricing a house. Several regression models will be tested, such as Multiple Linear Regression or Ridge Regression Model, and compared to each other to understand which model is more accurate in predicting the correct price for each house

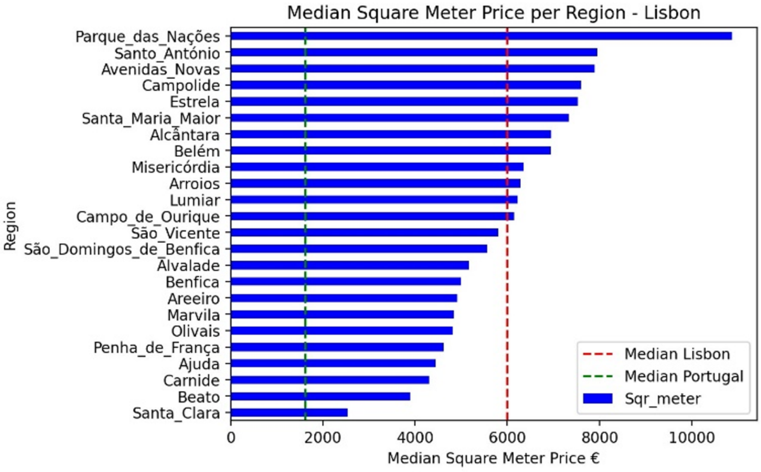

Figure 1: Median Square Meter Price by Lisbon Region, Lisbon (red line) and Portugal (green line).

Data and Methods

The main database will be extracted from the Supercasa website, via Web-scrapping. To acquire this dataset a Python program was created to scrape the data directly from the website. The main libraires used for this scrape were the following: Requests, Beautiful Soup, JSON. The Request library allows to create HTTP request, so it can take HTML content from webpages. The Beautiful Soup (bs4) is a library that pulls data out of HTML and XML files. Additionally, it can parse the content obtained from the website. Finally, the JSON library allows to decode JSON string into Python strings, allowing to decode the information obtained in the Supercasa website. So, after running the Python script, we obtained a dataset with 1309 properties for sale in the Lisbon region and with 9 descriptive columns. The following information was extracted for each property: “ID”, “Title”, “price”, “num_rooms”, “total_area”, “latitude”, “longitude”, “region” and “extras”. As expected, the main dataset obtained had several errors and inconsistencies, so data preprocessing and exploration had to be made. Firstly, duplicated houses were excluded from our dataset, since we verified that the same house could be announced on different pages. Afterwards, data quality checks were run in each individual column, and inconsistencies were found. Some of the houses had the area information in the “num_rooms” column and in the “total_area” column other types information were presented. We verified that these inconsistencies were found on T0 houses (0 rooms), which had no number of rooms information and distorted the extraction. From here a rule was created: if the house had the string “T0” on its “Title” column, the value of 0 would be created in the “num_rooms” column and the “total_area” would be corrected. After this correction, we excluded houses which did not have the correct number of rooms or the correct total area. Afterwards, there were some rows with missing values on the “total_area” column. One of the methods to handle with missing values is the K Nearest Neighbors Imputer. Firstly, the ideal number of neighbours had to be calculated, in this case was 1, which delivered the higher accuracy. All columns were transformed into the correct type since some integer variables had to be transformed to float and so forth. An additional column, which represents the real total of rooms in the house was created, “Total_N_Rooms”, since we had house being sold as having 3 rooms but in the announcement the property had 3+1 rooms, being 4 rooms. Finally, since the main goal of this report is to create a predictive model for the houses price, several dummy variables were created for each region, property type (apartment, studios) and extras (garage, swimming pool). After all these transformations, we ended up with a dataset of 1205 properties and with 94 columns, having six objects variables, two float variables (latitude and longitude) and the rest being integers. For the other datasets, the web-scrapping technique was also applied. For the Metro dataset, the information was extracted for the Wikipedia page. For the Cultural dataset, information was scraped from museums, theatres, cinemas, auditoriums sites. Afterwards these individual datasets were put together to form an individual dataset. For the Educational dataset, information from pre-schools, 1st/2nd,3rd cycle schools, high schools and universities were extracted and put all together. The data quality check and necessary transformations were applied to all datasets, like in the first dataset. For all datasets, two columns were very important to obtain and transform, which were the "latitude" and "longitude" columns, since those columns are used to calculate distance between houses in the main dataset to the locations on the other datasets. Transformations were applied in all datasets to assure that the latitude and longitude variables were as float, to apply the Haversine function, which allows to calculate the distance between two points having their coordinates. Finally, after having the distance between the houses and the rows on the other datasets, we applied distance thresholds for each dataset. After joining the datasets, three new columns were added to the main dataset for the regression models: "Stations_within_0.5km," "Cultural_fac_within_1.5km," and "Edu_within_1.5km," indicating the number of nearby sites for each house, by distance threshold. This information enabled the precise location of each house, as well as the nearby cultural and educational facilities, as illustrated in figure 2.

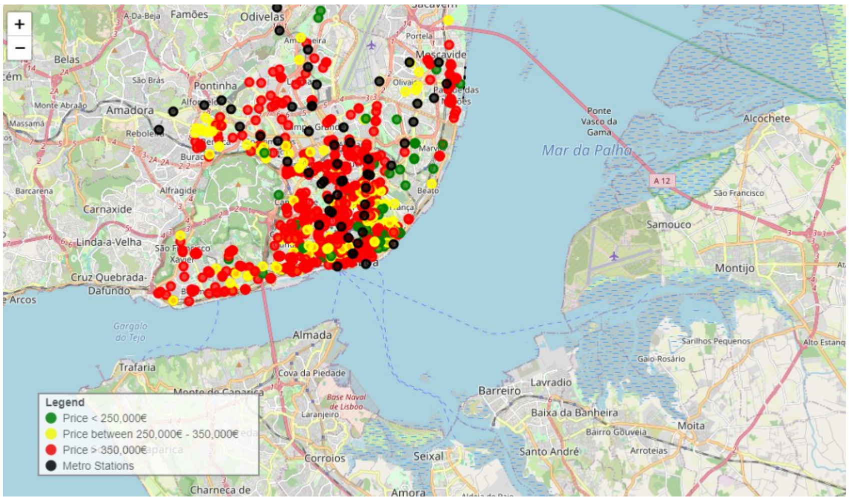

Figure 2: Geospatial Distribution of House Prices in the Lisbon Metropolitan Area and locations of metro stations.

Results and Discussion

The feature selection process was a critical step in our modeling pipeline, aimed at identifying the most relevant predictors of property prices. After initial data preprocessing, which involved cleaning and transforming the dataset, several statistical tests and machine learning techniques were applied to refine our feature set, in line with the most suited ML algorithms used to predict housing prices [4]. The Chi-square test was used to assess the importance of categorical features. This test evaluates whether there is a significant association between each feature and the target variable, enabling us to identify which categorical variables were most impactful on property prices. Features such as "Rés_do_chão", "Com_elevador", "Com_garagem", "Último_andar", and some location-based variables were found to be insignificant and thus excluded from the final model. For numerical features, we evaluated their variance to eliminate those with excessively low variance, which could introduce noise into the model. Variables with high variance can often distort the results, leading to overfitting. Therefore, "Extra_Rooms" and "N_Extras" were removed at this stage to streamline the dataset and enhance model performance. To further refine our feature set, we employed a Decision Tree classifier. This technique helped us retain only those features that significantly contributed to the model's predictive power. Key features selected included 'Total_N_Rooms', 'Area_m2', 'São_Domingos_de_Benfica', 'Apartamento', 'Stations_within_0.5km', 'Cultural_fac_within_1.5km', and 'Edu_within_1.5km'.

With a well-defined feature set, we proceeded to apply various regression models to predict property prices. Multiple Linear Regression (MLR) was the first model tested, providing a baseline for comparison with more complex models. The MLR model yielded an R² score of 0.5627 on the training set and 0.5598 on the testing set, indicating that the model could explain about 56% of the variance in property prices. The Root Mean Squared Error (RMSE) was approximately 538,749 for the training set and 498,587 for the testing set. These results suggest that while MLR captures general trends in the data, it still has significant prediction errors for individual properties. Ridge Regression, designed to handle multicollinearity among features by adding a regularization term, was next evaluated. This model yielded an R² score of 0.5173 on the training set and 0.5159 on the testing set, with RMSE values of 566,008 for the training set and 523,621 for the testing set. The slightly lower performance of Ridge Regression compared to MLR indicates that multicollinearity was not a major issue in our dataset, and the added regularization did not significantly enhance model accuracy. Lasso Regression, which combines variable selection with regularization, was then applied. This model achieved an R² score of 0.5590 on the training set and 0.5584 on the testing set, with RMSE values of approximately 539,117 for the training set and 498,955 for the testing set. Lasso Regression performed similarly to MLR but had the added advantage of simplifying the model by eliminating less important features, making it more interpretable. Elastic-Net Regression, which blends the regularization properties of both Lasso and Ridge, was tested to see if it could provide a balanced approach. The R² score for Elastic-Net was 0.5487 on the training set and 0.5470 on the testing set, with RMSE values of 545,390 for the training set and 504,726 for the testing set. While the performance was slightly lower than that of Lasso and MLR, Elastic-Net effectively managed the complexity of the model by balancing between L1 and L2 regularization terms. The results of the regression models applied are summarized in the figure 3.

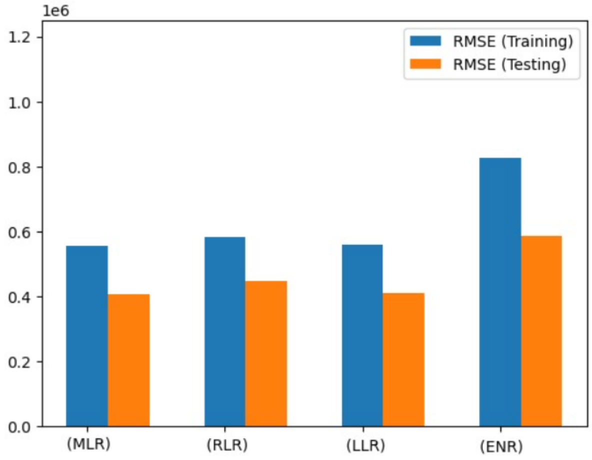

Figure 3: Root Mean Squared Error (RMSE) for training and testing datasets across four different regression models: Multiple Linear Regression (MLR), Ridge Linear Regression (RLR), Lasso Linear Regression (LLR), and Elastic-Net Regression (ENR).

The Decision Tree Regression model, despite its high R² score of 0.9914 on the training set, showed significant overfitting, with an R² score of only 0.6409 on the testing set. Overfitting occurs when a model learns the noise and fluctuations in the training data too well, capturing details that do not generalize to new data. This high variance between training and testing scores indicated that the Decision Tree model, although powerful on the training data, lacked the generalizability required for practical application.

Conclusions

This study successfully developed predictive models for Lisbon property prices using rigorous feature selection and various regression techniques. The application of multiple regression techniques provided a comparative analysis that highlighted the strengths and limitations of each model with MLR and Lasso Regression proving to be the most effective, offering robust and interpretable tools for real estate market analysis. The comprehensive feature selection process ensured that only the most relevant variables were included, enhancing the model's efficiency and interpretability.

For real estate professionals and policymakers, MLR and Lasso Regression provide valuable tools for predicting property prices and identifying key influencing factors, aiding in better decision-making and strategic planning. These models can help understand the impact of various features on property prices, facilitating informed decisions regarding property investments and market interventions. The insights gained from this study can guide pricing strategies, investment decisions, and policy formulations aimed at stabilizing the housing market.

However, it is important to note that the accuracy of our predictive models is limited by the relatively small dataset of 1,309 properties. This limited sample size can increase the margin of error and reduce the generalizability of our findings. The supercasa website only allow the navigation of twelve web pages limiting the dataset size to 1,309 properties when there are more than twelve thousand properties listed. With a larger dataset, the models could potentially achieve higher accuracy and robustness, providing more reliable predictions.

Future studies could expand the feature set by including economic indicators or property age and explore advanced machine learning techniques such as ensemble methods to improve performance and robustness. Incorporating macroeconomic variables like GDP growth, unemployment rates, and interest rates could enhance the predictive accuracy of the models. Additionally, exploring ensemble methods like Random Forests and Gradient Boosting could provide improved performance by combining the strengths of multiple base models.

Further research could also investigate the temporal dynamics of the housing market by incorporating time series analysis. This approach could capture trends and seasonal patterns in property prices, providing a more comprehensive understanding of market fluctuations. Additionally, collaboration with governmental and private institutions to access more granular and official data could enhance the reliability and validity of future predictive models.

In conclusion, this project successfully demonstrated the application of multiple regression techniques to predict property prices in Lisbon. MLR and Lasso Regression emerged as the most effective models, providing a solid foundation for further research and practical applications in the real estate domain. These insights can guide future developments in predictive modeling, contributing to more accurate and informed real estate market analysis. By continually refining and expanding the scope of predictive models, stakeholders can better navigate the complexities of the housing market, ensuring more equitable and efficient outcomes.

References

[1]Portuguese Government Statistics - Average Salary 2023
[2] The implicit interest rate on housing credit in Portugal, INE - February 2024.
[3] Housing Price Statistics at the Local Level, INE 2024.
[4] M Yasser H., Housing Price Prediction Best ML Algorithms, 2022.