HOUSE PRICES IN MADRID: INSIGHTS FOR MARKET ANALYSIS AND INVESTOR DECISION MAKING

TEAM 60, Section C

Tim Gong, Caterina Mora, Fariha Tasnim Roza, Koko Wan

1. BUSINESS UNDERSTANDING

Madrid's (Spain) real estate market is very complex, with various factors other than property size influencing housing prices.This project aims to:

1) <u>Understand Market Pricing</u>**:** We are interested in learning how houses are valued in the Madrid region and which variables contribute to higher prices.
2) <u>Real estate investor insights</u>**:** Real estate investors often conduct business by buying existing properties, remodeling them, and re-selling them for profit. The key challenge is finding properties that are undervalued or have a high potential for growth, and accurately estimating all the total expenditure. Therefore, it is crucial for investors to understand the local real estate market- in our case, Madrid- in order to identify what features or characteristics can increase a house’s value and lead to a higher selling price and return on investment.

Our goal in this project is to leverage analytics to assist the investor in making informed decisions regarding their investment. By analyzing Madrid's housing data, conducting exploratory analysis and modeling, we can predict house prices based on different property features.

Our approach will reveal the specific impact of characteristics such as square footage, number of bathrooms, and energy certificates on property valuations, and be able to predict the price you sell a house for based on the characteristics of the house. Then, we will compare the actual buy price of a property with its potential selling price predicted by our model, helping to identify undervalued houses. Furthermore, it will provide a framework for the design specifications of the houses before engaging construction firms, optimizing potential revenue for the investor.

2. DATA UNDERSTANDING

The dataset used to solve our problem was extracted from the “Kaggle” Website, and uploaded in 2020. Our dataset is about the Madrid real estate market, and provides a total of 21742 real estate listings in Madrid from popular internet portals, and 58 different variables related to house features. Examples of the variables we find in this dataset are: street name, if a renewal for the house is needed, energy certificate, and house orientation.

3. DATA PREPARATION

In the data preparation step, we performed observatory analysis and then proceeded to clean our dataset. First, we identified the variable types, their meaning, and any potential problems with the quality of the data that would have an impact on the accuracy of the model. Then, we performed data cleaning where we corrected, deleted, or imputed missing or erroneous data to ensure that the data is accurate and complete for our next step, modeling. The following description explains our approach to managing data inconsistencies:

- Missing values: Several variables contained missing values (NAs), and we used different strategies depending on the context. For categorical variables, we replaced "NA" for "Unknown" if the NAs were missing data or had no clear meaning. In some cases, we replaced NAs with the designation ‘False’ given the context of the boolean nature of the variable. In certain cases, we discovered variables for which more than 80% of the values would be NAs, and as a result, removed these variables from our dataset because they wouldn't be relevant to our model.
- Recording and/or augmenting existing variables if needed was a big part of our data cleaning process. We had to re-level and factor multiple variables and divide them into different factors so they are easier to understand.

- Removing variables if needed: We removed a certain amount of variables that we considered to be insignificant for our model. Specifically, we removed variables that were largely NAs or that had more than 90% of the observations belonging to just one factor and thus would not offer significant insights into my model.

After data cleaning, we performed exploratory data analysis (EDA) to explore how different variables interact , and identify highly correlated variables. More information about EDA can be found in the Appendix.

4. MODELING AND EVALUATION

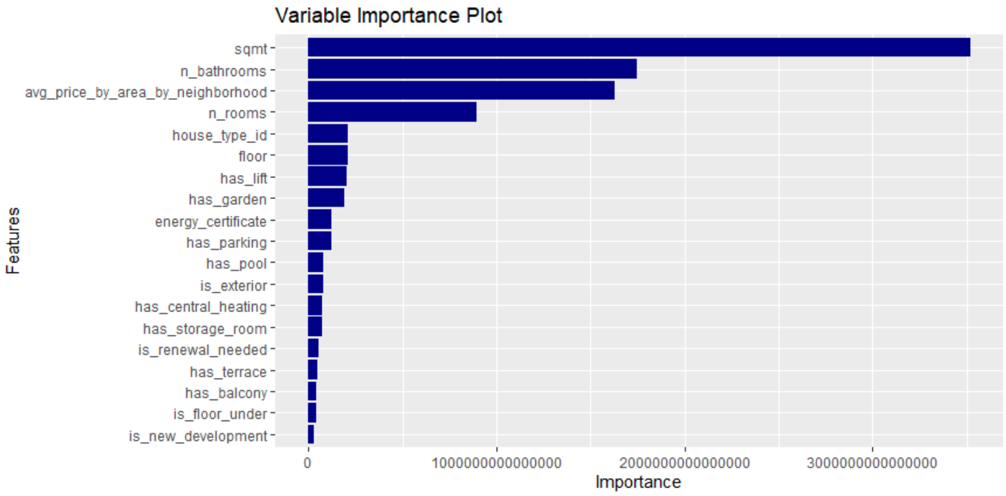
In order to predict house prices in Madrid, we built multiple models, including linear regression, Lasso, Post-Lasso, decision trees, and random forest. By using different models, we could get different predictions and choose the model that makes the most accurate predictions.

After data cleaning and EDA, we splitted the cleaned data into a training set (80%) and a testing set (20%) using stratified sampling to ensure all neighborhoods were represented in both sets. The resulting training set had 17257 obs, and the test set had 4248 obs. We used the training set to build, evaluate, and optimize models. Next, in simple words, we made the models learn the relationships within the data. Afterwards, we applied the selected model to the test set to see how well it could make predictions on unseen data.

More specifically, first, we ran a NULL model as the baseline and the other five models (linear regression, Lasso, Post-Lasso, decision trees, and random forest) on the training set. We compared their in-sample performance based on RMSE and R^2. The random forest model has best RMSE and R-squared. We use these two metrics, R^2 and RMSE, to evaluate the performance of each metric and decide on what model gives us the most accurate predictions. Nevertheless, we want to focus on the out of sample performance to make sure we choose a model that will perform well on unseen data. See Appendix Table 1 for the In-Sample R^2 and In-Sample RMSE.

A 10-fold cross validation was conducted on all models to estimate out of sample (OOS) performance. The OOS provides a better assessment for our predictions since it avoids considering models that are overfitting or don't perform well on unknown data. We compared all the models on both R^2 and TMAPE. Based on the OOS performance, random forest has the best OOS R^2 and TMAPE. Therefore, we chose the random forest for the use of running prediction on the test set.

The Random Forest model provides us with the most accurate pricing predictions, and signals the variables that are the most significant on the buy price, and their impact. The graph below shows the factors that contribute most to the variation in prices. In our case, the most significant predictors include square meter, the number of rooms and bathrooms, and the neighborhood. The prediction made by our Random Forest model will be crucial to evaluate if a property is undervalued or overvalued.



Lastly, we used the random forest model to predict price on the test set. On the test set, the random forest generates a RMSE of 230270.4 and a R^2 of 0.895. This result means that the random forest model can explain approximately 90% of the variability in house prices in the test set.

5. DEPLOYMENT

As explained in the previous section, we have used the Random Forest model to predict the value of the property based on its features. This model plays a critical role in identifying whether a property is undervalued or overvalued by comparing the predicted price with the actual price a house might be listed for. When comparing these two results, we have two possible scenarios:

Scenario 1:

$Prediction \prec Buy\ Price\ Listed$: the model predicts that the property is priced at a higher level than its expected market value, suggesting that it is overpriced and might not be a good investment.

Scenario 2:

$Prediction > Buy\ Price\ Listed$: the model predicts that the property is priced at a lower level than its expected market value, indicating that it might be a good opportunity for an investor to buy and resell for gaining potential profit.

Our random forest model will give us an output for every observation: 1 if it’s a good investment (the property is undervalued) or 0 if it’s not a good investment (the property is overvalued). This is how we identify and select what properties to buy for the investment.

Once identified the undervalued property to buy, the investor must also account for the additional buying costs. By calculating the Expected Cost, the investor knows the full financial commitment to acquire the property, which allows for a more precise planning of renovation and holding costs. Based on research, these are the additional costs for the Comunidad de Madrid, Spain:

- Property Transfer Tax (ITP): 6%
- Notary and Land Registry Fees: 0.5%
- Community Fees: 200€

The total expected cost for buying a property can be calculated as:

$$Expected\ Cost\text{: } Buy\ Price\ \times\ (1\ +\ 6\%\ +\ 0.5\%)\ +\ 200\ +\ Renovation\ Cost$$

This formula accounts for the price that the investor buys the property for, the taxes associated with the buying transaction and the renovation costs. Note that renovation cost can vary for each property. While some properties might need a big renovation (change floors, renovate kitchen, etc) some properties might just need minor additions (adding an AC unit, etc). Therefore, the renovation costs are hard to estimate and necessary to consider when purchasing the property. Investors should ensure that their renovations are cost-effective and add value to their properties. This will enable them to sell their properties for a higher price than when they were originally purchased and thus earning a profit margin.

On the other hand, when selling a property we also have to consider:

- Capital Gains Tax: 19% on the amount of profit
- Agency Fees: 3%

To consolidate, the final formula for the expected profit for this process is:

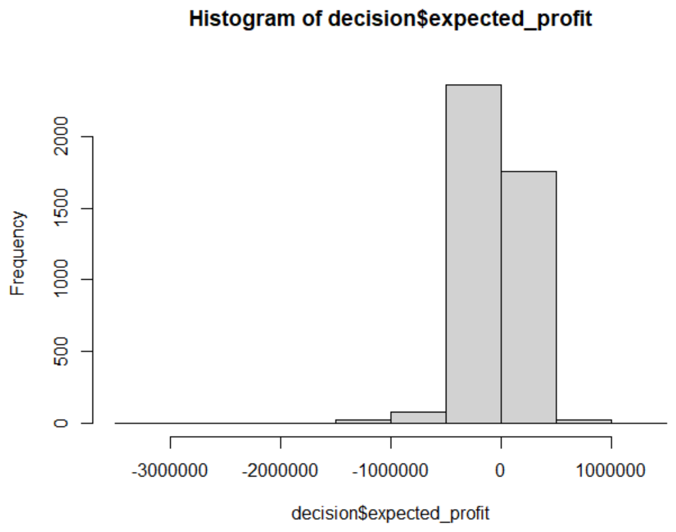
$$Expected\ Profit\text{: } (Final\ Buy\ Price\ -\ Expected\ Cost)\ \times\ (1\ -\ 19\%\ -\ 3\%)$$

After purchasing the undervalued property we identifies with our model, the investor can sell the property at a price ideally above the predicted price from our model by adding value with the renovation. The profit from this sale will be maximized by:

1) Minimizing purchase costs though finding the most undervalued properties (biggest gap between predicted price and buy price)
2) Focusing on high return features in the renovation expenses: adding features that will enhance the value of the house.

In summary, the investor should focus on being able to 1) identify properties that are undervalued using our model to target which properties are undervalued, 2) using the formulas to understand the investment before committing, 3) adding value in a cost effective way though remodeling and selling at the right time to maximize the final sale price.

To better understand our use of the prediction, the graph below demonstrates that less than half of the properties in the test set have the potential to generate profit and are worth investing in. The idea is to apply our model on unseen data, where we have information about the property buy price and the features.



We also want to consider potential limitations for our project. The first limitation is the data availability. Our dataset contains 58 different variables with information regarding the property. While this is a considerable amount of features, there might be more characteristics needed in order to predict the price of the house, as for example, the quality of the installations. In addition, we didn’t account for any external factors that affect house prices, such as proximity to cultural centers or areas of interest, supply and demand, interest rates, etc.

Furthermore, it is important to acknowledge that estimating renovating costs is challenging. These costs depend on multiple factors including the type of renovation, price of materials, construction companies and their corresponding fees, etc. We suggest that investors prioritize cost-effective renovations, ensuring that the expenses don’t outweigh potential profits upon resale.