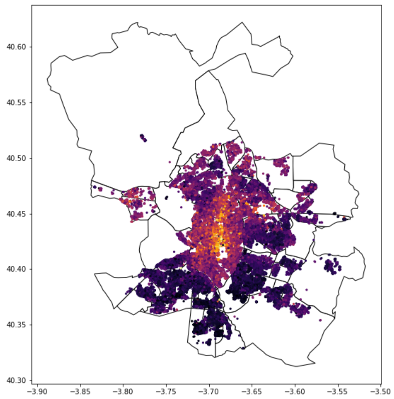
1,772 × 1

Application of Data Science in Housing Valuation in Madrid Spain

The Valley
Digital Business School

Idealista Challenge - MDS



Master's Final Project

Autors

Leonardo Velásquez Estupiñan
Alejandro González López

Director: David Rey, Chief Data Officer of Idealista

Summary

This document presents the project "Application of Data Science in Housing Valuation in Madrid Spain", a housing valuation model developed as part of the Master of Data Science and Big Data Analytics at The Valley Business School in collaboration with Idealista as a solution to the challenge launched by Idealista named "How much is my house worth?". Using a large database of housing listings in Madrid in 2018 in the Ideaslita platform, the project provides a detailed approach, results obtained, and details on implementing the model in a cloud environment.

Keywords

Hedonic Regression, Machine Learning, Deep Learning, Random Forest Regression, Spatial Regression, XGBoost, KNN, Neural Networks, Bayesian Search, Streamlit, Geopandas

Chapter 1: Introducing the Idealista Challenge and Proposed Approach

Abstract

This document presents the data product developed as part of the challenge "How much is my house worth?" launched by the company Idealista as a final project of the Master in Data Science and Big Data Analytics of The Valley Business School. The main objective of this project is to create a valuation model of homes in the buying and selling market that allows estimating the price at which an advertisement of a house published on the Idealista portal would be found.

The challenge is based on a large database that includes 94,815 advertisements of properties for sale in Madrid, Spain, in 2018. This database provides detailed information on the characteristics of the properties, such as the number of square meters built, their distribution, and more. In addition, complementary datasets have been made available to enrich the variables used to train the model, including geographic data, distances to points of interest, and cadastral information.

This document details the approach and results achieved by the authors of the final project, Alejandro López and Leonardo Velásquez, under the direction of David Rey, CDO of Idealista. It explains the development phases of the model, from its conception to its implementation, and presents the results obtained. It also describes how the model has been deployed in a cloud environment for its use and evaluation.

Available data sources

As data resources, the Idealista challenge made available a variety of valuable sources that were used in this project:

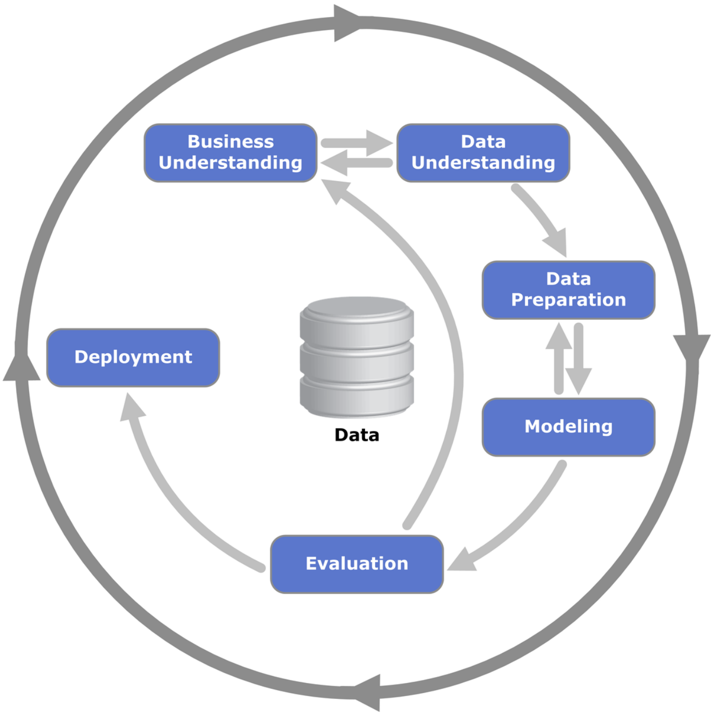
- Idealista database: this database contains property sales listings in the city of Madrid, Spain, during 2018. The data given by Idealista was previously anonymized by Idealista due to legal requirements introducing a small random perturbation in the total price and exact position.
- Polygon database: used to divide the space of the city of Madrid into specific geographic zones, such as neighborhoods and areas of interest within Idealista.
- POIs (Points of Interest): Points of interest were collected in Madrid, which are key elements within this metropolis, ranging from locations such as pharmacies and restaurants to relevant urban facilities.
- OpenStreetMaps (OSM) database: OpenStreetMaps points of interest, such as pharmacies, restaurants, and other urban facilities, were included to enrich the geographic information further.
- INE (Instituto Nacional de Estadística) database: This data source contains INE census sections as well as population and housing census data collected in 2011. Each census

section is identified by a ten-character code called CUSEC. Despite its age, this dataset provides descriptive information of high value.

Using these aforementioned data resources, a final working dataset was created, which is further detailed in the variable engineering chapter of this paper.

Approach

We followed the development cycle of a standard artificial intelligence project, which comprises the following stages: 1) Understand the business context, 2) explore and thoroughly understand the available data resources, 3) prepare the data thoroughly, 4) create and tune the model, 5) evaluate and finally 6) deploy the model in a production environment.



Our objective is to generate a housing value prediction model that achieves accuracy standards comparable or superior to the reference models, prioritizing the model that has demonstrated an optimal overall performance, as well as evaluating the performance of the model in terms of error and accuracy in each of the neighborhoods of Madrid to study the granularity of the model and its accuracy by area.

A fundamental aspect of our project is to guarantee the integrity and consistency of the variables used in the model training since the dataset has real data, and its evaluation will be contrasted with real use cases of the Idealista platform. This was achieved through a rigorous process of data cleaning and treatment based on statistical models and business knowledge.

Proposed use case

According to data from Idealista[1], Madrid holds a prominent position in the national real estate scene and leads the ranking of cities with the highest number of sale and purchase transactions. Some 25% of the properties are sold in less than a week, 22% are sold in a period ranging from one week to one month, another 25% require from one to three months to sell, 22% take from three months to a year to find a buyer, and 5% require more than a year to complete the transaction.

With this scenario, we decided to address the challenge from the perspective of users who own properties in Madrid and want to estimate the value of their properties to advertise them on Idealista later or connect with a real estate agency that will manage the commercial management of the property, streamlining an agile and meaningful estimation of their properties based on market data.

As a final result, we proposed producing a fictitious functional digital product deployed in the cloud. This product would be available to Idealista users, allowing them to make predictions about the value of their property in a few minutes and access valuable property information by simply providing their property data. The predicted value can be used as a reference value to contrast with appraisals made in situ, to contrast with the value of other real estate listings in the same areas, or simply to estimate the value of users' properties agilely.

For Idealista, this digital product would represent an opportunity to activate and acquire more advertisers on the platform and referrals to the real estate agencies with which Idealista has agreements, and therefore strengthen its business lines related to the appraisal, the listing of properties on its platform and its commercial agreements with real estate agencies.

Theoretical Framework

The theoretical framework of this project is based on several fundamental concepts, among which the hedonic regression model in housing research and Machine Learning models with tree structure are highlighted. In addition, other areas of relevance that contribute to accurate property valuation are incorporated. The five frameworks within which this project was developed are described below:

1. **Hedonic Regression Model in Home Valuation:** Hedonic regression is an essential pillar in home valuation. This approach allows the price of a property to be broken down according to its specific characteristics. This leads to the identification and quantification of how aspects such as size, location, condition of the property and other characteristics influence its market price.

2. **Importance of Location and Proximity to Points of Interest:** We explore how geographic location and the availability of nearby services impact the price of homes, paying attention to variables such as distance to the city center, metro stations and the Paseo La Castellana on the basis that the location of a property and its proximity to points of interest play a crucial role in the valuation of homes.

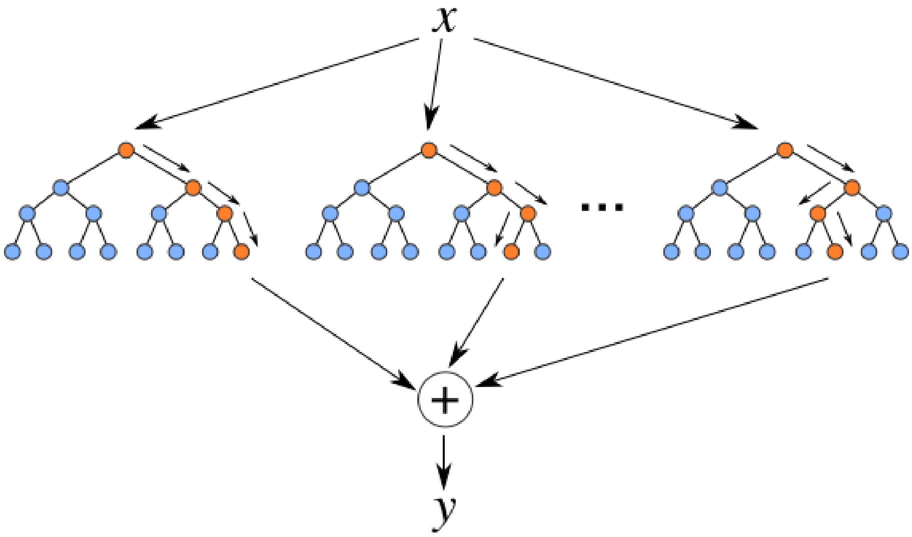
3. **Statistical Analysis and Features Engineering:** A statistical analysis of the variables relevant to the project was carried out such as the correlation between variables and their influence on the value of the homes, using metrics such as the P-Value to identify the attributes that have a greater relevance in the valuation as well as the exploration of the

data distribution, identification of outliers and assessment of normality. These variables have also been treated so that they can be used in the model with methods such as target encoding, dummification and logarithmic adjustment.

4. **Machine Learning and Deep Learning Models:** Machine learning and deep learning techniques such as KNN, XGboost, Random Forest and Linear Regression among others were incorporated in the project to find the best possible performance for housing price prediction. Emphasis was given to regression models based on tree-like structures, such as Random Forest Regression, which have shown outstanding performance compared to the other methods explored.

5. **Random Forest Regression Model:** Random Forest Regression is a machine learning technique that combines multiple decision tree regression models into a Bagging-type ensemble, hence the name "forest", to make more accurate and robust predictions. During the modeling phase, this model type showed superior performance compared to other models, such as linear regression and K-Nearest Neighbors (KNN) clustering models. As evaluation metrics for this type of model, it is common to use RMSE (root mean square error) because it penalizes the results in which the deviation of the prediction is very high, MAPE (mean percentage error), and Points for quality of fit R² calculated as variance error/variance of the validation set. These metrics were used during the evaluation and optimization phase.

Esquema gráfico de un modelo Random Forest



Taken together, this theoretical framework provides a solid conceptual foundation that combines traditional real estate valuation methods with advanced artificial intelligence techniques to estimate the price of homes on the Idealista portal accurately.

State of the art and references

Several resources have been identified and consulted to enrich understanding in the field of home valuation. These resources include

- The "Boston House Prices-Advanced Regression Techniques" dataset that provides regression techniques applied to house prices.

- The dataset "A geo-referenced micro-data set of real estate listings for Spain's three largest cities" developed by the Idealista team, which provides a detailed view of the real estate market in Spain's major cities and enriches the local context as well as the understanding of the variables in the Idealista challenge dataset.
- Statistical precedents of Hedonic Pricing Models, including parametric and non-parametric approaches to housing valuation.
- Recent trends and advances in the field, such as the application of spatial regression models and the use of natural language processing (NLP) and deep learning (LLM) techniques in the analysis of housing descriptions in online advertisements.

These resources have significantly enriched the approach to the project and have contributed to a more complete understanding of the subject, placing the project in a referenced context of real estate valuation.

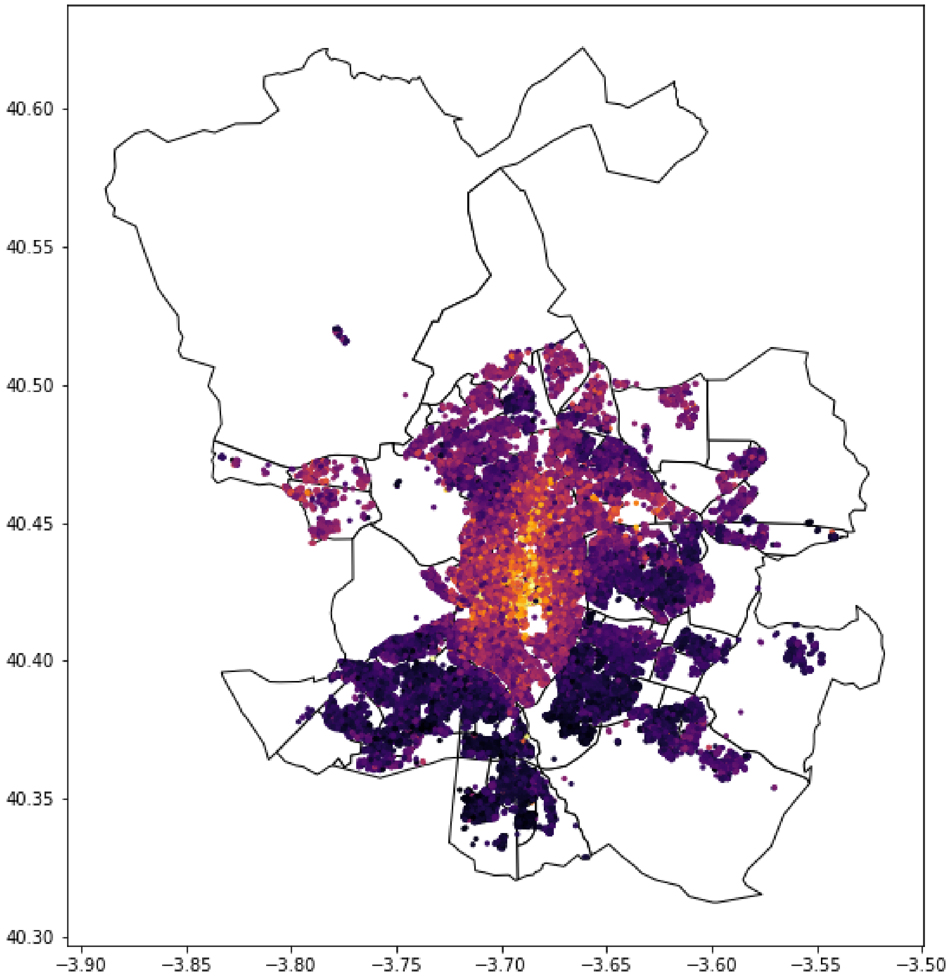
Chapter 2: Exploratory Analysis and Features Engineering

Summary Statistics

The original data set consists of 94,815 rows and 46 variables, making it a complete resource with a low presence of null values. Only three variables have missing data, the most relevant of which is "CONSTRUCTIONYEAR". However, we have identified a solution to address this lack of data by replacing them with the values available in "CADCONSTRUCTIONYEAR", which comes from external sources and is more reliable.

After resolving the null values, we identified duplicates in the data set. Using the unique identifier "ASSETID", we found that there are initially 75,804 unique results under this column. However, in line with a strategic business decision, we have chosen to retain only the most recent record, applying a sorting of the assets from most recent to oldest. This ensures our dataset reflects the most up-to-date information relevant to our analysis and objectives.

Price distribution in Madrid



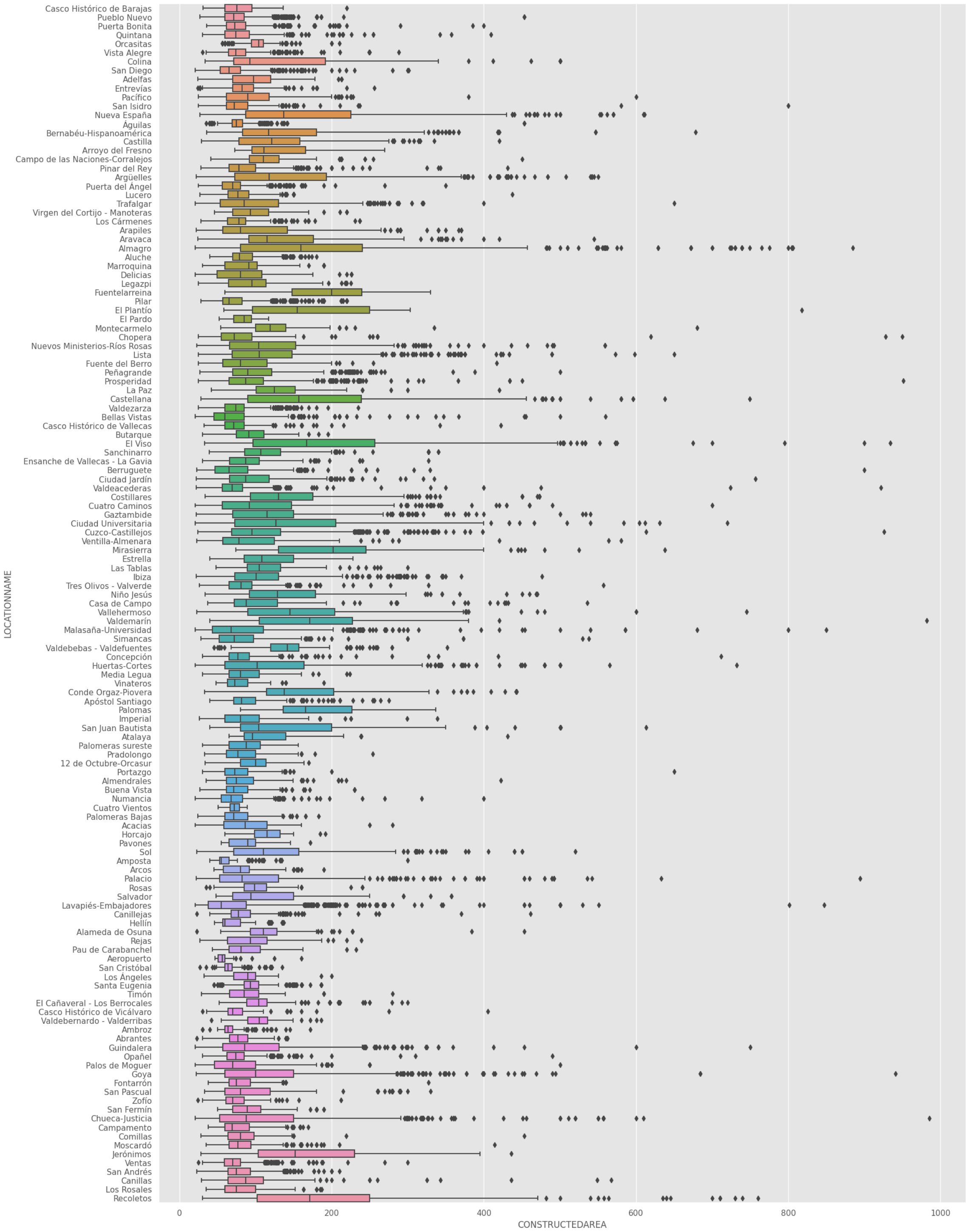
We begin this process by visually plotting all the properties contained in our dataset, using their geographic coordinates. This provides us with a first perspective of the spatial distribution of the properties.

To obtain aggregated and contextualized information, it was necessary to integrate our dataset with another dataset that provides data on the neighborhoods of Madrid. We used the Geopandas library to carry out this data union. In this way, we were able to assign each property to a specific neighborhood in Madrid, allowing us to place each property within a defined and closed geographic set.

In the image, we can see how the coordinates of the properties are linked to the polygons of the neighborhoods in Madrid. This step is crucial in order to be able to carry out subsequent analyses that depend on the geographical location of the properties in relation to their surroundings.

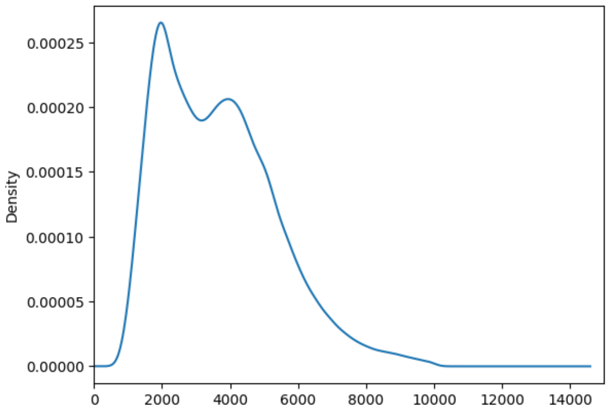
Once the neighborhoods have been assigned, we can see how the properties are distributed within each zone according to the built-up area:

Distribution of constructed area by neighborhood



The distribution suggests that there are multiple groups or patterns within the data. This could indicate the existence of subsets of data with distinct characteristics. It is crucial to recognize and deal with these differences, as they can significantly influence the effectiveness of the model.

Unit distribution

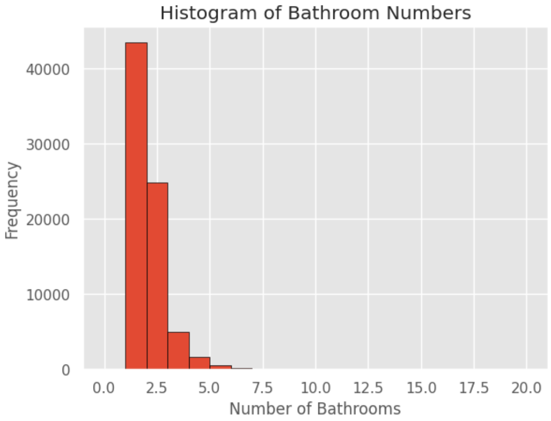


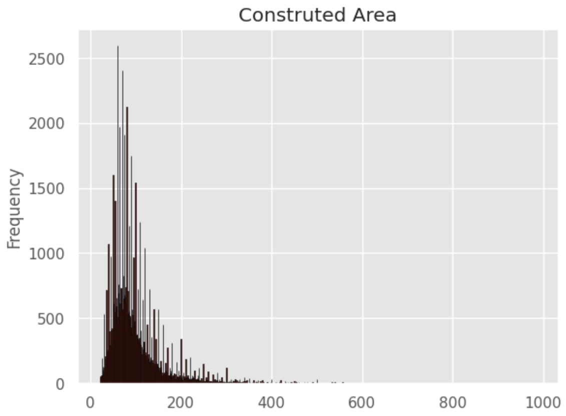
Skewness in the distribution of values means that the data are not equally distributed on both sides of the mean. This can affect the ability of the model to generalize correctly. It may be necessary to apply transformations to the data or consider more robust models that can handle this skewness.

The wide range of values in the data indicates that some variables may have extremely high or low values. This can lead to challenges in scaling the variables and interpreting the model. Normalizing or standardizing the variables may be necessary to prevent some features from dominating over others in the modeling process.

Outliers and data correction

In our process of searching for outliers in the real estate data, we first eliminate those records that do not have information on bathrooms. We then perform a thorough analysis to detect any inconsistencies in the relationship between the number of bathrooms and floor area. Despite finding no apparent anomalous situations, we continue to carefully evaluate the data to ensure the integrity of our data set.





Focusing now on the relationship between the number of rooms and the built area, we have observed that this relationship exhibits an asymmetric and extremely wide distribution. To address this situation, we have implemented the following steps:

1. **We applied Linear Regression:** We used linear regression to better understand the relationship between the number of rooms and floor area in the data set.

2. **Neighborhood Analysis:** We conduct neighborhood-specific analyses to compare properties with their surroundings. This allows us to evaluate how this relationship behaves in different geographic areas.

3. **Non-Consistent Data Filtering:** As part of our strategy, we identify and eliminate those records that do not show consistency in the relationship between the built area and the number of rooms within each neighborhood. This process helps us maintain the integrity of our data and ensure that our conclusions are robust and representative.

In summary, our focus at this stage of the project is on understanding and optimizing the relationship between the number of rooms and floor area, taking into account the geographic variability of the properties. This will allow us to make more informed and accurate decisions in the data analysis process.

As a result of the search for outliers, the cleaning process, and the elimination of duplicates in the initial dataset containing 94,815 properties, we have obtained a final dataset consisting of 74,213 properties. This cleaned and optimized dataset will be the main input for the construction and training of our model. The reduction of the dataset to 74,213 properties ensures that we are working with high-quality data that is relevant to our analytical and modeling objectives.

Target Encoding - Locationname (MEAN)

We applied the method "Target Encoding" to the categorical variable "Locationname". This technique involves assigning to each category of "Locationname" the average value of the target attribute "unit price" corresponding to that category. The purpose of this technique is to represent the categorical variable "Locationname" numerically in a way that captures the relationships between the location and the target attribute "unit price". This resulting numerical representation is used in machine learning models to improve predictive capability, especially when the location (neighborhood) is an important factor in the outcomes that the project seeks to predict.

Next steps in features engineering

In the next iterations of variable engineering, we are continuing to improve the quality of our data and enriching our feature set. Here is a summary of the key steps we will be taking:

1. **Incorporation of Spatial Variables:** We will continue to assign spatial variables, such as the "Census Section" and the "Homogeneous Cadastral Value Zones". These data will provide us with statistical information on the population in each area, as well as details on sales and purchases, housing homogeneity, and prices. This will enrich our data set and improve our understanding of the factors affecting property values.

2. **Extending Linear Regressions:** Extending the linear regression analysis beyond "Floor Area", "Bathrooms" and "Number of Bedrooms", we will work on applying this approach to other variables. The goal is to obtain a cleaner and more reliable dataset by better understanding the relationships between the different characteristics.

3. **Correlation Analysis and Advanced Variable Engineering:** We will perform a more detailed correlation analysis to identify complex relationships and patterns between variables. In addition, we will continue to explore more advanced variable engineering opportunities to improve the accuracy of our models.

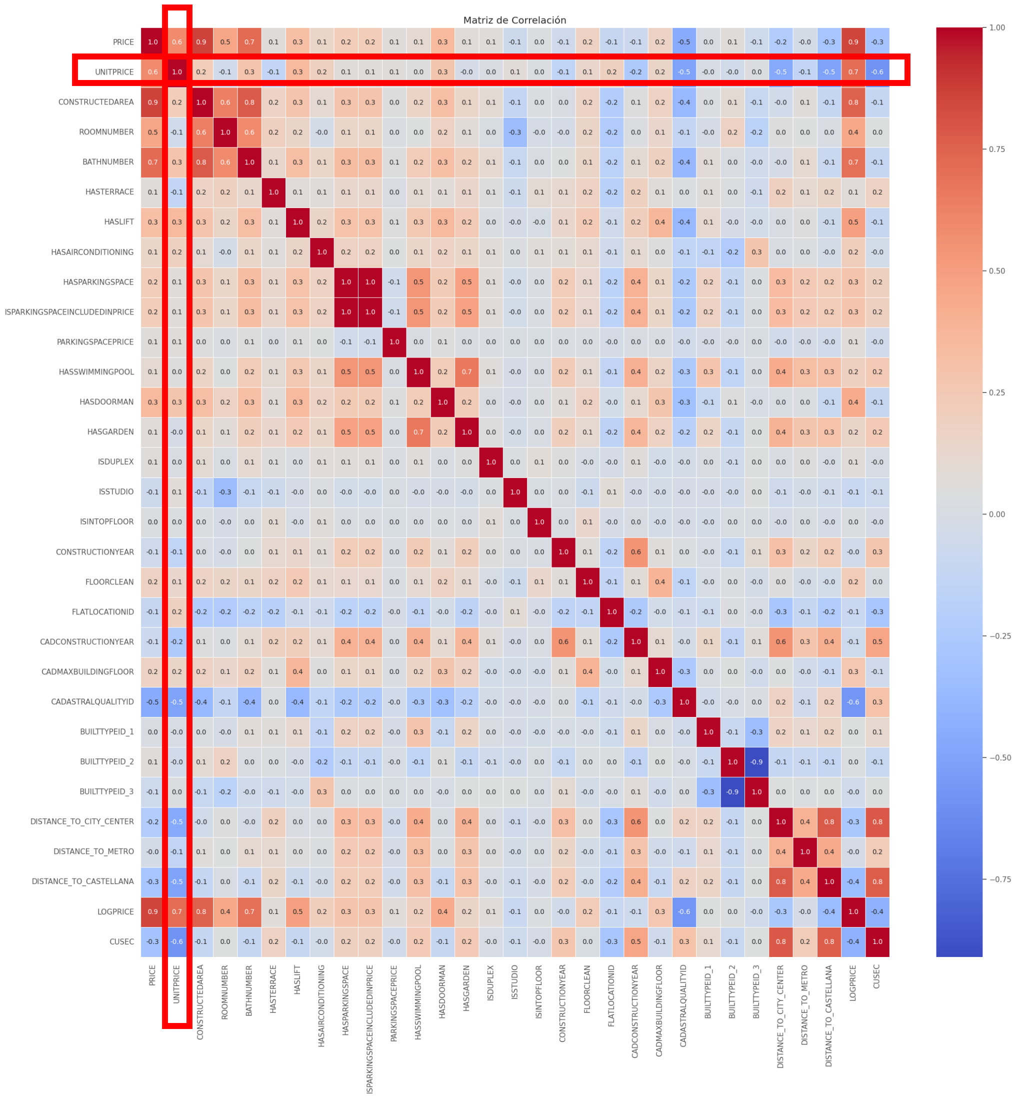
4. **Testing with Discarded Variables:** We will revisit previously discarded variables to assess whether any of them may be useful after all or if they can be meaningfully transformed for inclusion in the model.

5. **Homogeneity of Urban Areas:** We will continue to investigate the homogeneity of urban areas to understand how it affects the property market and how we can reflect this in our characteristics.

These steps will lead us towards a more complete and sophisticated data set, which in turn will improve the accuracy of our models and allow us to make more informed project decisions.

Correlation Matrix

Once the dataset is clean we apply a distribution matrix so that we can see how each variable influences the others. Bearing in mind that our target variable will be the unit value of the real estate:



The dispersion is very large, and there are no variables that are significantly related to the unit value once those variables are directly dependent on it, such as the price or the logarithmic price, are obviated.

In summary, to build an effective model under these conditions, it is essential to understand and address multimodality, skewness, and range of values. Techniques for data preprocessing, feature selection, and choice of models that are suitable for handling these specific data characteristics must be explored. In addition, it is critical to perform detailed exploratory analysis to discover patterns and relationships within the data that can effectively guide model building.

Scaling the variables

In order to equalize the scale of the target variable and those of the input variables of the model, we decided to apply the logarithmic scale to both the unit value and the average unit value of the neighborhoods so that, when introducing the variables in the model, they would be more homogeneous.

Final variables

Once the redundant variables with the target variable had been eliminated, we filter those that either have a high correlation in the matrix or those that, due to business experience, are relevant when carrying out a real estate valuation model.

The variables "ASSETID" and "LOCATIONNAME" were retained for later statistical purposes.

```
#   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   ASSETID                        74213 non-null  object
 1   CONSTRUCTEDAREA                74213 non-null  int64
 2   ROOMNUMBER                     74213 non-null  int64
 3   BATHNUMBER                     74213 non-null  int64
 4   HASLIFT                        74213 non-null  int64
 5   ISDUPLEX                       74213 non-null  int64
 6   CADCONSTRUCTIONYEAR            74213 non-null  int64
 7   BUILTTYPEID_1                  74213 non-null  int64
 8   BUILTTYPEID_2                  74213 non-null  int64
 9   BUILTTYPEID_3                  74213 non-null  int64
 10  DISTANCE_TO_CITY_CENTER        74213 non-null  float64
 11  DISTANCE_TO_METRO              74213 non-null  float64
 12  DISTANCE_TO_CASTELLANA         74213 non-null  float64
 13  LOCATIONNAME                   74213 non-null  object
 14  UNIT_PRICE_LOG                 74213 non-null  float64
 15  LOCATION_MEAN_UNITPRICE_log    74213 non-null  float64
```

Chapter 3: Modeling

We started by prioritizing the type of model that, according to the literature consulted, could offer the best performance with these data. We start with Random Forest and compare its performance with two other methods as part of our evaluation: Linear Regression and the XGBoost algorithm.

1. Random Forest Regression

To improve the performance of our Random Forest model, we performed the hyperparameter tunning using two different approaches. This will allow us to find the optimal configuration that maximizes the accuracy of our predictions.

→ Random Forest cross-validation K-Folds + Random Search

Description: In this hyperparameter tuning stage, we use the Random Search method to find the optimal hyperparameter configuration for our Random Forest model with K-Folds Cross Validation. The process consists of the user setting a number of iterations to test the model. At each iteration, the algorithm randomly selects a set of hyperparameters from a predefined grid (without replacement) and computes the corresponding loss metrics. The more iterations performed, the higher the probability of finding the best hyperparameters, but this will also increase the algorithm run time.

Advantages:

- Relative Efficiency: This method is faster compared to other exhaustive hyperparameter search approaches, making it suitable for fitting models in a reasonable time.
- Parallelization: It allows parallelization, which means that several sets of hyperparameters can be tested simultaneously, further speeding up the process.
- Random Exploration: Randomization in hyperparameter selection can lead to the exploration of a wider space, which increases the chances of finding effective solutions.

Disadvantages:

- Suboptimal with Few Iterations: With an insufficient number of iterations, there is a possibility that the algorithm will select suboptimal hyperparameters and fail to find the best configuration.
- No Guarantee of Global Optimum: Unlike exhaustive search, Random Search is not guaranteed to find the optimal hyperparameter configuration, as it depends on randomness in the selection.

This approach combines the efficiency of Random Search with K-Folds cross-validation to efficiently tune the hyperparameters of our Random Forest model, thus improving its performance on our specific dataset.

Results:

```
{'n_estimators': 200,
'min_samples_split': 2,
'min_samples_leaf': 4,
'max_features': 'sqrt',
'max_depth': None,
'bootstrap': False}
```

General Metrics:

```
MAE (Test): 467.17 €
R2 (Test): 0.83
MPE_mediana: 9,7 %
```

→ Random forest cross validation K-Folds + Bayesian Search

Description: In this modeling stage, we implement a Bayesian search strategy to optimize the hyperparameters of our Random Forest model. Bayesian search uses information from previous iterations to make informed decisions about which hyperparameter values to explore next. This accelerates convergence to an optimal solution.

Advantages:

- Aprendizaje de Iteraciones Anteriores: La búsqueda bayesiana se beneficia de las lecciones aprendidas en iteraciones anteriores, lo que permite una convergencia más rápida hacia soluciones óptimas.

Disadvantages:

- Limitations on Parallelization: In general, Bayesian search does not benefit from parallelization, which can significantly increase the time required to find a solution.
- Risk of Local Maximum: As with all Bayesian methods, there is a risk of converging to a local maximum instead of finding the global maximum, which could limit the quality of the final solution.
- This combination of Random Forest with K-Folds Cross Validation and Bayesian Search allows us to obtain a more accurate and efficient model by optimizing the Random Forest hyperparameters for our specific data set.

Results:

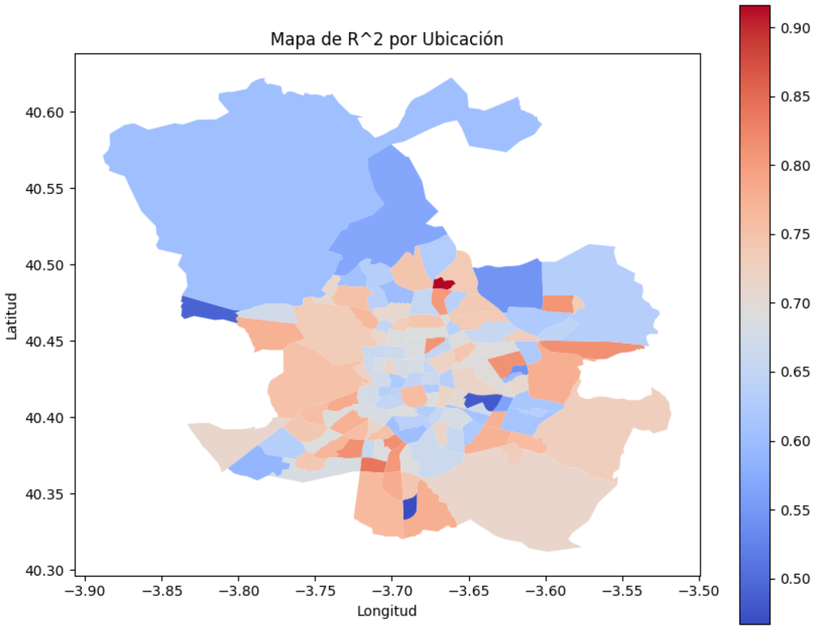
```
'bootstrap', False),
('max_depth', 34),
('max_features', 0.6857565539493793),
('min_samples_leaf', 4),
('min_samples_split', 2),
('n_estimators', 300)])
```

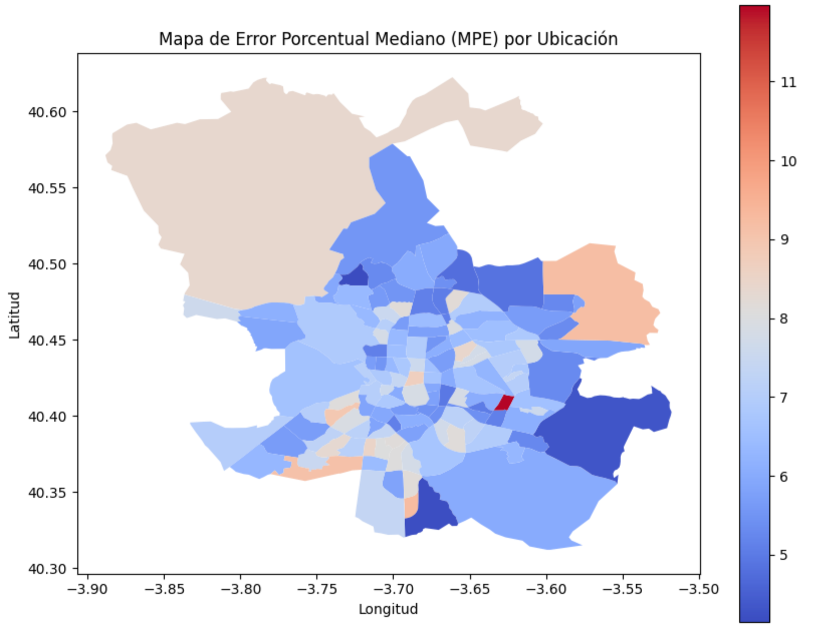
General Metrics

```
MAE (Test): 467.17 €
```

```
R2 (Test): 0.83
MPE_mediana: 9,39 %
```

To check the consistency of the model in each of the neighborhoods and given the high segmentation of the data, a particularized metrics check is made for each of the areas:





2. Linear Regression

Linear regression is a fundamental technique in the field of statistics and machine learning used to understand and model the relationship between independent variables and a target variable. It attempts to establish a linear relationship between these variables, which means that it seeks to find a straight line that best fits the observed data.

General Metrics:

```
MSE del modelo: 0.0548
R^2 del modelo: 0.7725
Media del Error Porcentual: 2.17%
```

3. XGBoost

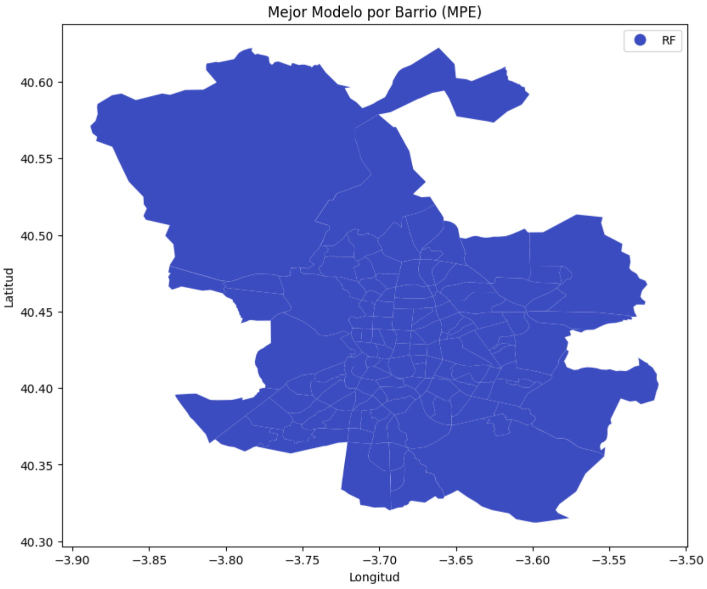
For the XGBoost model, a Bayesian Search has been launched in order to locate the hyperparameters that best fit this dataset, giving us as a result:

```
([('colsample_bytree', 0.8311563895216271),
('gamma', 1),
('learning_rate', 0.6020666067370067),
('max_depth', 40),
('min_child_weight', 6),
('n_estimators', 74),
```

General Metrics:

```
MAE (Test): 549.58 €
R2 (Test): 0.79
MPE_mediana: 11.9%
```

Once the three models have been launched, and although Random Forest apparently works better in general, a comparison is made at the neighborhood level to see which model works better in each of the zones. (ver documento)



Final model selection

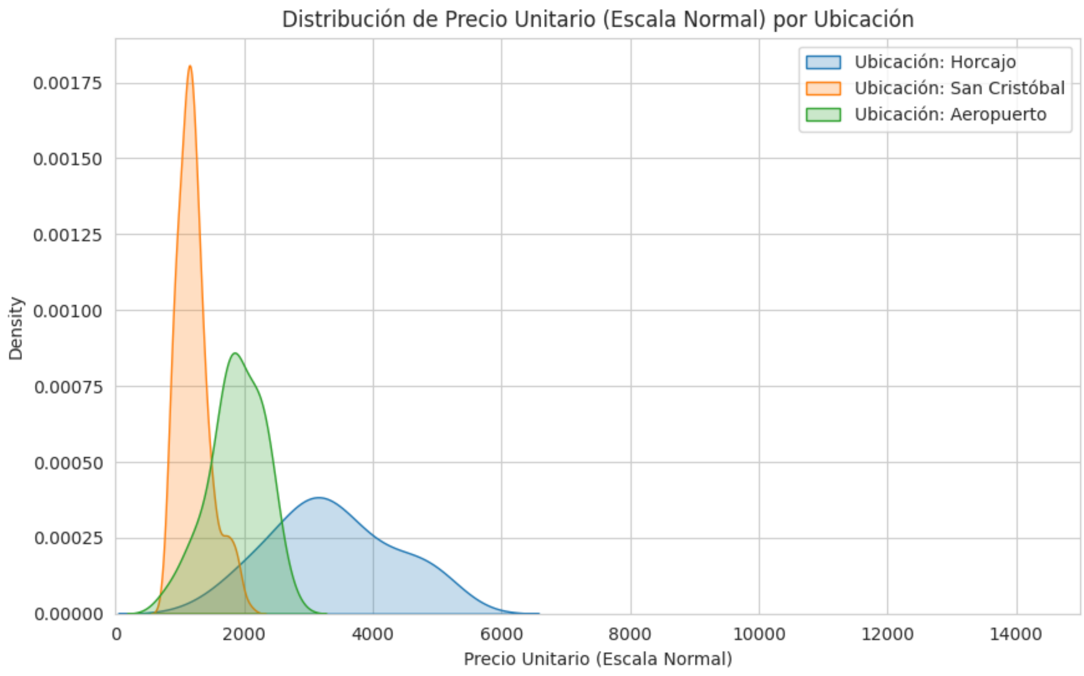
The Random Forest model is ratified as the most suitable model for estimating the value of housing in all neighborhoods, therefore it is taken to production.

Discussion

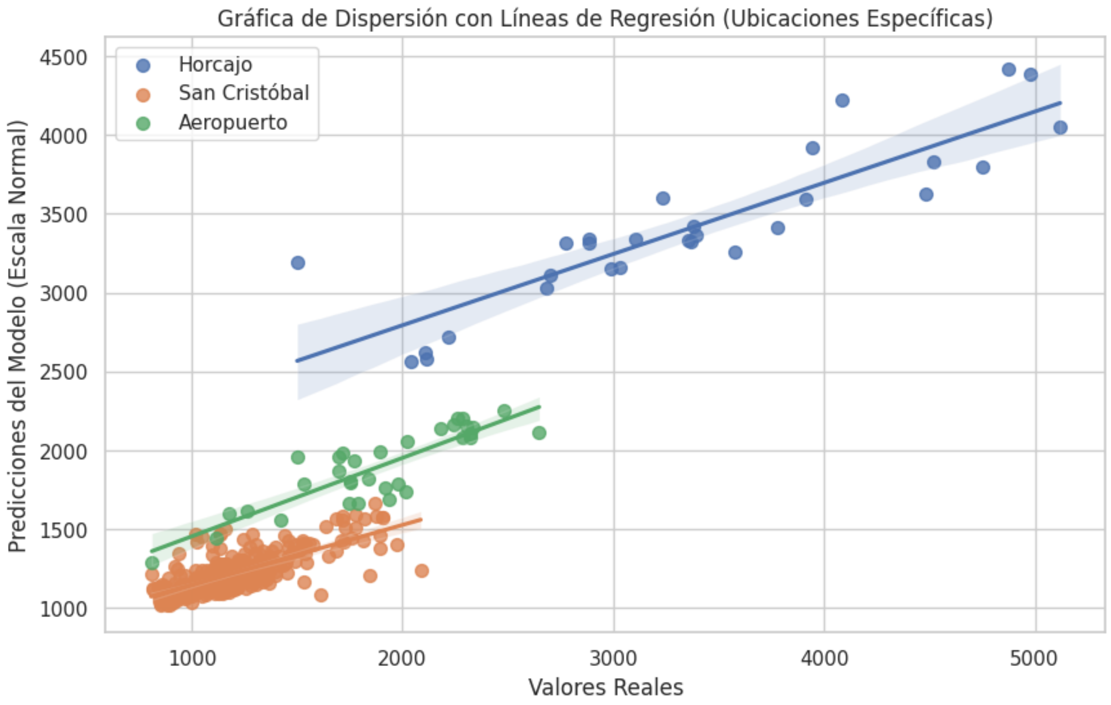
The Random Forest model works well in general, but when we enter the distribution by neighborhoods we see that there are areas where the model degrades quite a lot. (See the notebook)

```
Ubicación: Horcajo
R^2: 0.5632
MPE: 11.97%
Población de Muestras: 29

Ubicación: San Cristóbal
R^2: 0.4670
MPE: 9.28%
Población de Muestras: 263

Ubicación: Aeropuerto
R^2: 0.6320
MPE: 9.21%
Población de Muestras: 33
```

And these are their Unitary distributions:



As well as scatter plots of actual values versus values predicted by the model.



We have in common the small population of the samples, we should think in the next iteration to make a model of urban twins in order to increase the number of samples even in different environments, with similar characteristics, besides the point clouds are very dispersed.

We will try in the next iteration to make specific models for each of the neighborhoods so that the hyperparameters are specific for each of the zonings and thus combat the disparity of the samples from the training.

Modeling 2.0

As part of the continuous product improvement process, we have started the second iteration of modeling. In this phase, we are implementing specific approaches for each of the neighborhoods in which our system operates.

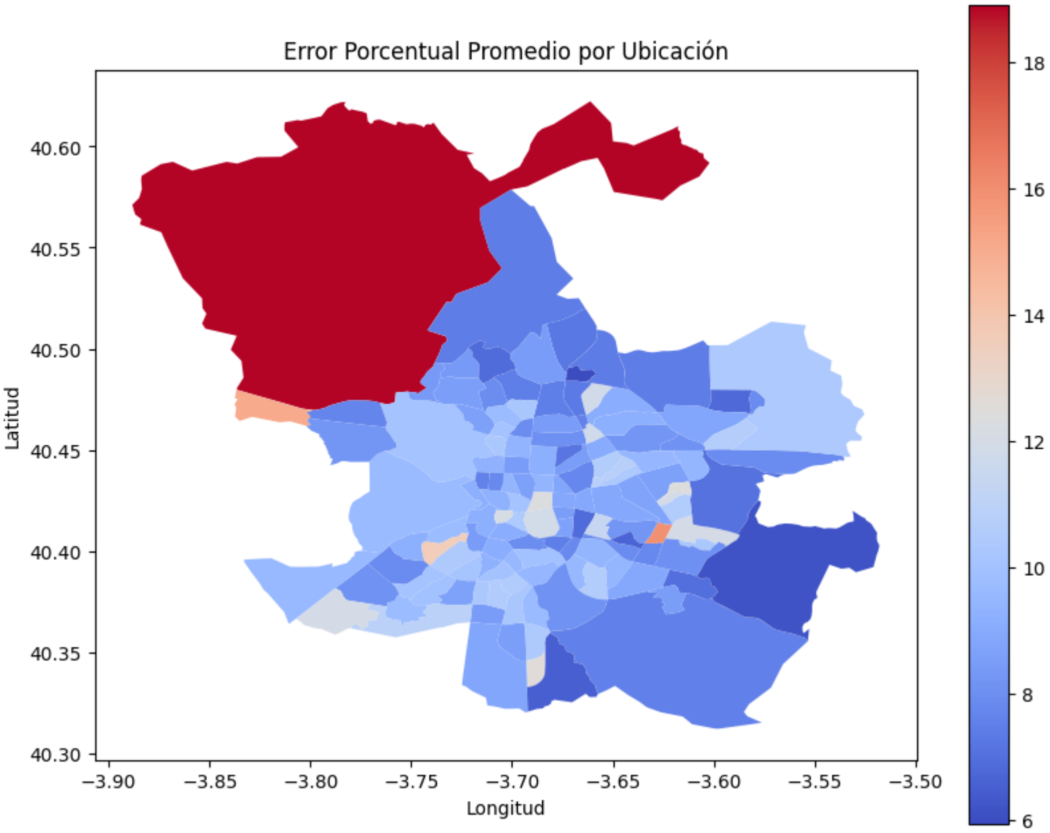
- **Neighborhood-Specific Models:** To carry out this iteration, we are generating models individually for each of the neighborhoods that make up our area of operation. This approach will allow us to better adapt the system to the particularities of each location and ultimately improve the quality of our predictions and recommendations.

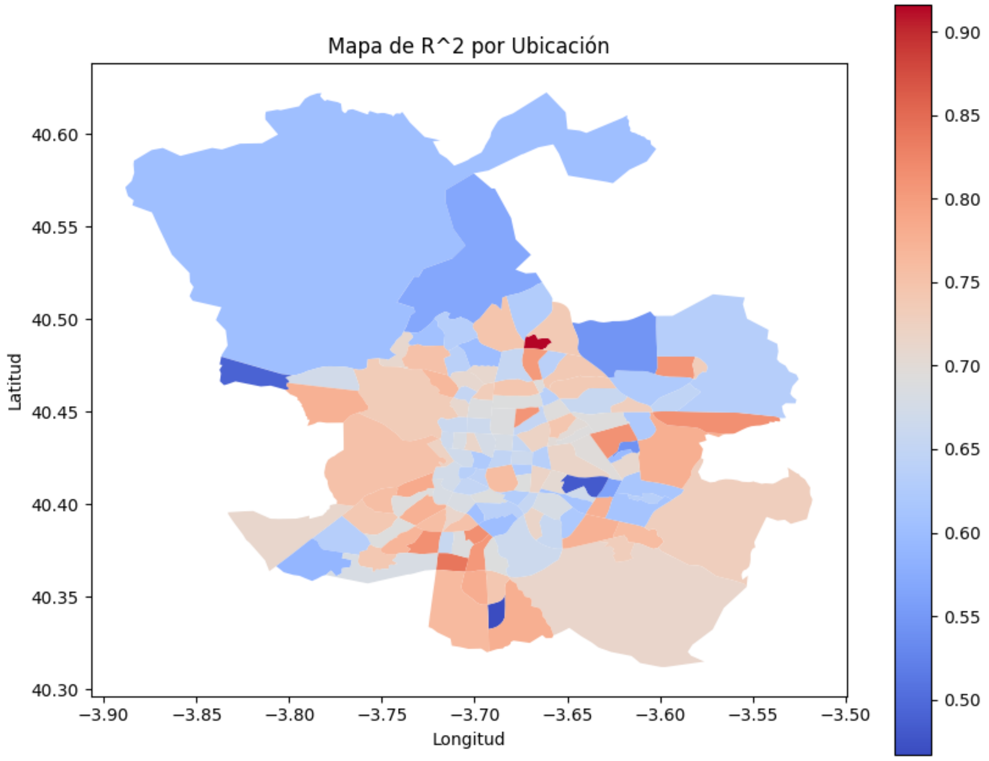
- **Bayesian Search for Each Location:** As part of the process, we have re-initiated the Bayesian hyperparameter search process. However, unlike the previous iteration, this time we are performing the hyperparameter search independently for each of the neighborhoods. This strategy gives us the ability to tune the models more precisely and specifically to the needs of each location.

- **Training with the Best Hyperparameters per Neighborhood:** Once we have identified the optimal hyperparameters for each neighborhood through Bayesian search, we proceed to train the models using these custom settings. This ensures that each model is optimized for its specific environment and maximizes its predictive capability.

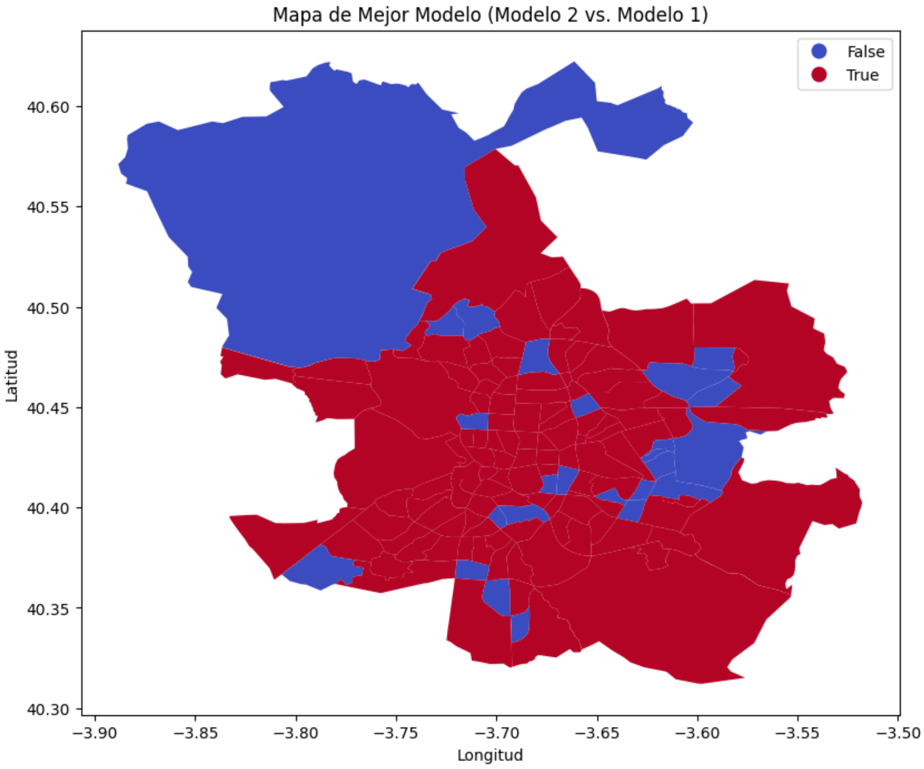
- **Metrics Evaluation and Comparison:** Finally, we evaluate the performance of each model in terms of relevant metrics for each neighborhood. Then, we compare the results obtained in this iteration with those of the original Random Forest-based model. This comparison will allow us to measure the impact of our improvements and determine whether we have made significant progress in the quality of the predictions.

Here we have the results of Mean Percentage Error and R2 in each of the neighborhoods:





By analyzing the R2 we can determine that in general, the second model performs better than the first one, but in some cases, the original model is more accurate, so both models will have to be merged in order to improve the results.



Chapter 4: Production

AI Idealista - App

After completing the Modeling and Evaluation phase, we have developed a web application for Idealista users. This digital tool aims to bring the developed model closer to the users, allowing them to use it effectively. Users can provide information about their properties and run the model in real-time to obtain an estimate of the value of their homes.

The application features a sidebar where users enter their property details by answering a series of questions using input widgets from the Streamlit library. These questions include information such as square footage, number of bedrooms, number of bathrooms, elevator availability, whether it is a duplex, year built, whether it is new construction, and other relevant features.

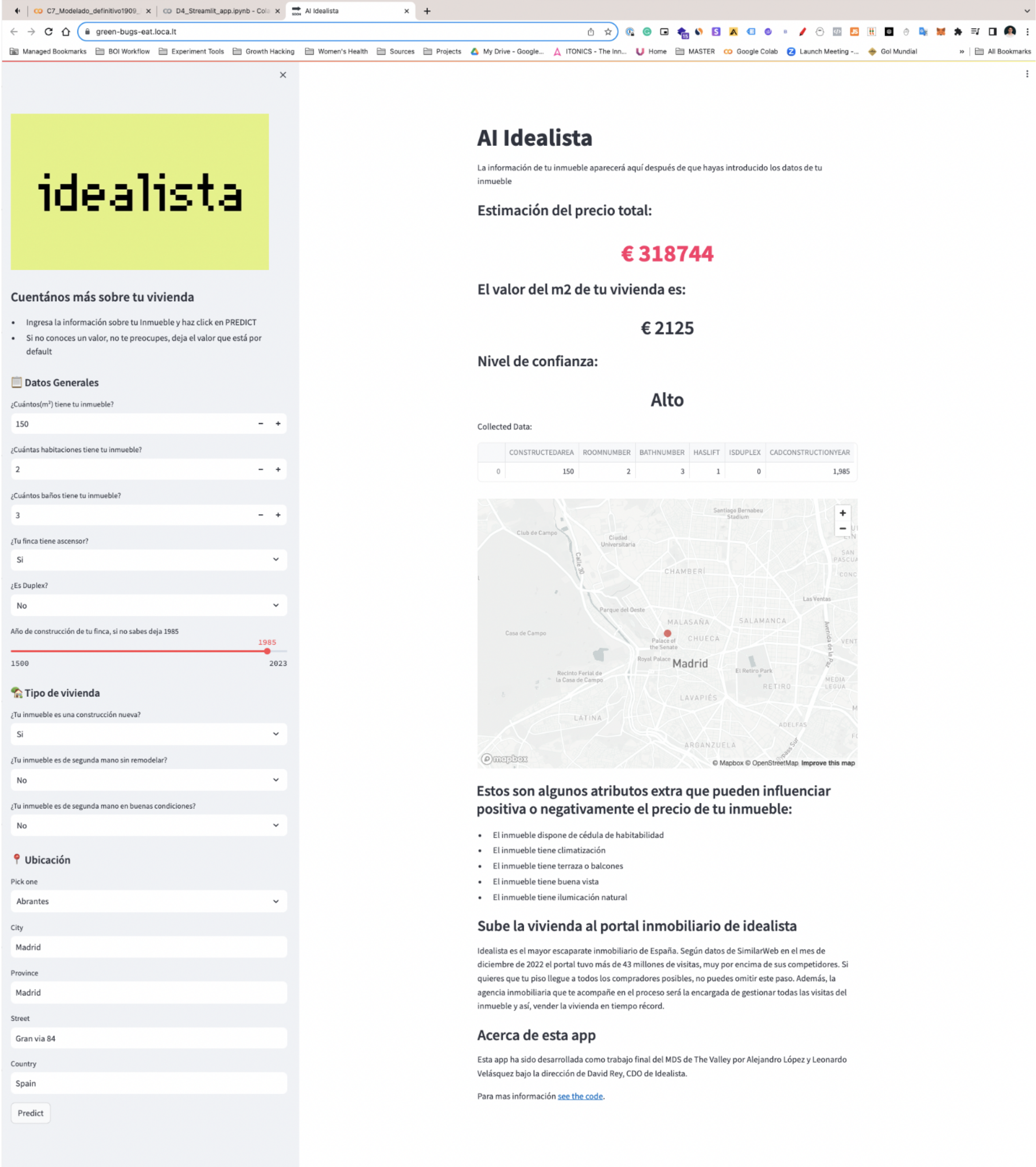
These user-supplied variables are essential, as they correspond to 7 of the 10 variables necessary for the model to make an accurate prediction about the value of the property. The remaining three variables will be automatically inferred by the application code: the distance to the city center, the distance to the subway, and the distance to Castellana G, as well as the average unit price of the location. We will expand on this later.

Once the user has completed the data entry and clicks on the "Predict" button, the application provides the user with:

- An estimate of the total price of the property
- An estimate of the price per square meter of the property.
- An indicator of confidence in the prediction (Low, Medium, High).
- A visualization of the entered data.
- The geolocation of the property.
- A list of additional attributes that may influence the price of the property, such as occupancy certificate, air conditioning, terrace or balconies, views, and natural light.
- Information about the advantages of listing the property on the Idealista platform.
- A brief explanation of how the application works.

In summary, the user can obtain all this information in approximately 2 minutes from the time he/she enters the application until he/she receives the final results.

App AI-Idealista



Cuentános más sobre tu vivienda

- Ingresa la información sobre tu Inmueble y haz click en PREDICT
- Si no conoces un valor, no te preocupes, deja el valor que está por default

Datos Generales

¿Cuántos(m²) tiene tu inmueble?
150

¿Cuántas habitaciones tiene tu inmueble?
2

¿Cuántos baños tiene tu inmueble?
3

¿Tu finca tiene ascensor?
Si

¿Es Duplex?
No

Año de construcción de tu finca, si no sabes deja 1985
1985
1500 2023

Tipo de vivienda

¿Tu inmueble es una construcción nueva?
Si

¿Tu inmueble es de segunda mano sin remodelar?
No

¿Tu inmueble es de segunda mano en buenas condiciones?
No

Ubicación

Pick one
Abrantes

City
Madrid

Province
Madrid

Street
Gran via 84

Country
Spain

Predict

AI Idealista

La información de tu inmueble aparecerá aquí después de que hayas introducido los datos de tu inmueble

Estimación del precio total:

€ 318744

El valor del m2 de tu vivienda es:

€ 2125

Nivel de confianza:

Alto

Collected Data:

	CONSTRUCTEDAREA	ROOMNUMBER	BATHNUMBER	HASLIFT	ISDUPLEX	CADCONSTRUCTIONYEAR
0	150	2	3	1	0	1,985

Estos son algunos atributos extra que pueden influenciar positiva o negativamente el precio de tu inmueble:

- El inmueble dispone de cédula de habitabilidad
- El inmueble tiene climatización
- El inmueble tiene terraza o balcones
- El inmueble tiene buena vista
- El inmueble tiene iluminación natural

Sube la vivienda al portal inmobiliario de idealista

Idealista es el mayor escaparate inmobiliario de España. Según datos de SimilarWeb en el mes de diciembre de 2022 el portal tuvo más de 43 millones de visitas, muy por encima de sus competidores. Si quieres que tu piso llegue a todos los compradores posibles, no puedes omitir este paso. Además, la agencia inmobiliaria que te acompañe en el proceso será la encargada de gestionar todas las visitas del inmueble y así, vender la vivienda en tiempo récord.

Acerca de esta app

Esta app ha sido desarrollada como trabajo final del MDS de The Valley por Alejandro López y Leonardo Velásquez bajo la dirección de David Rey, CDO de Idealista.

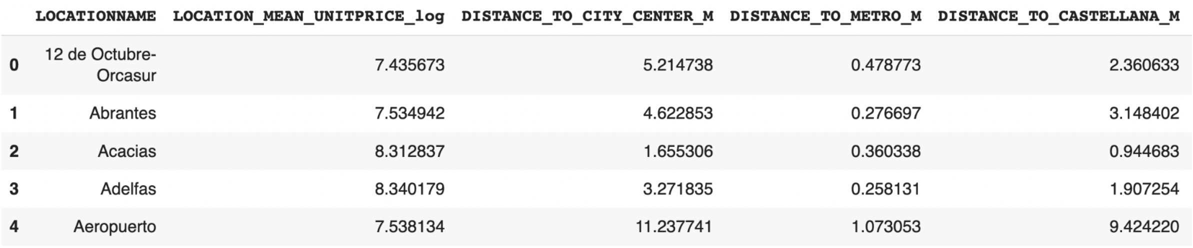
Para mas información see the code.

Automatic inference of distance variables to urban points

Given that the model requires three distance variables to key points of interest in the city of Madrid (1) Distance to Castellana, (2) Distance to Metro, and (3) Distance to Downtown, and given that users will rarely have access to these values, we had the challenge of inferring these variables.

Being geographic variables that depend on networks of points and distance lines, mainly Euclidean distance measures, developing a model that automatically calculates these variables using the location provided by the user as input was a challenge that we could not address within the timeframe foreseen for the project. Therefore, we devised an interim solution for this first iteration of the application. This solution involves the use of a database containing the average value of these variables per neighborhood. The Python script queries this database and uses the user-specified neighborhood as a reference point to provide these variables to the model.

In addition, we take advantage of this same database to supply the model with the variable LOCATION_MEAN_UNIT_PRICE, which represents the average value of the square meter on a logarithmic scale per neighborhood. This variable is also relevant to the performance of the model.



	LOCATIONNAME	LOCATION_MEAN_UNITPRICE_log	DISTANCE_TO_CITY_CENTER_M	DISTANCE_TO_METRO_M	DISTANCE_TO_CASTELLANA_M
0	12 de Octubre-Orcasur	7.435673	5.214738	0.478773	2.360633
1	Abrantes	7.534942	4.622853	0.276697	3.148402
2	Acacias	8.312837	1.655306	0.360338	0.944683
3	Adelfas	8.340179	3.271835	0.258131	1.907254
4	Aeropuerto	7.538134	11.237741	1.073053	9.424220

In future iterations of the model, we plan to calculate these geographic variables through the Google Maps API.

Streamlit

Once the best Random Forest model was trained and identified, it was exported in a file format (.pkl) using the Joblib library. This allowed the model to be loaded and activated in a web application.

To develop this digital product, the Streamlit library was used, an open-source tool that allows developers to create interactive web applications quickly and easily using Python. This makes it easy for users to use the application in real-time.

A Python script is responsible for orchestrating the application logic, including loading dependencies, the neighborhood database, loading the model, and the logic behind the Streamlit application.

You can consult the Python code here.

Transparency

Since this application is user-driven and provides a prediction of real estate value, which is sensitive data that can have a significant impact on the user, it is essential to present this prediction together with a confidence estimate. To address this, we have performed a quartile distribution of the (R^2) of the model for each neighborhood. Then, we have defined a traffic light type scale in which the prediction is considered low confidence if the R^2 is in the range of 0% to 49% of the distribution, medium confidence if it is in the range of 50% to 74%, and high confidence if it is between 75% and 100%.

This scale has been incorporated into the database that the Python script loads into the Streamlit application at startup, and the application queries it against the neighborhood provided by the user. This ensures that the predictions are accompanied by an indication of confidence that will help users interpret and use the information appropriately.

Production 2.0

These are a list of possible next key iterations to improve and optimize our application:

- **Integrate a Google Maps geolocation API call:** We will implement the ability to call a google maps API to automatically calculate distances to points of interest in the city, such as Castellana, Metro and Downtown. This will improve the accuracy of our predictions and facilitate the user experience.

- **Deliver a confidence interval calculation to the user:** We will provide the user with a confidence interval calculation along with the predictions. This will allow users to better understand the reliability of the estimates and make more informed decisions.

- **Monitoring and maintenance:** We will establish a continuous monitoring system to oversee application performance and perform proactive maintenance. This will ensure that the application is performing optimally at all times.

- **Update models with 2022-3 databases:** We will keep our models up to date by incorporating updated property data from the year 2022. This will ensure that our forecasts accurately reflect the latest real estate market conditions.

- **Reinforcement Learning:** We will explore opportunities to implement Reinforcement Learning techniques to improve the model's accuracy and ability to adapt to changes in the environment.

These steps will help us improve and evolve our application, providing users with a more reliable and valuable experience when searching for real estate value information.