uc3m | Universidad Carlos III de Madrid

Master Degree in Statistics for Data Science
2024-2025

Regression Models

Modelling Competition: House prices 2024

OVERFITTERS

Marcos Álvarez Martín
Nicolás Carrizosa Arias
Ángel Pellitero García
Simon Schmetz

María Luz Durbán Reguera
Madrid-Puerta de Toledo, November 2024

CONTENTS

1. INTRODUCTION

As part of the course "Regression Models" of the Master in Statistics for Data Science at the Universidad Carlos III de Madrid, a Modeling competition is held to predict the $€/m^2$ price of residential real estate offers across the Madrid Metropolitan Area. The dataset, consisting in 1000 instances, contains a wide array of variables with respect to property type, location, and living quality. The goal of this work is to create a model to predict the price in $€/m^2$ of the properties with the maximum possible accuracy. This model is then used to predict the unknown prices for a test subset, the results of which will be handed in for scoring to evaluate the best model within the participating groups.

The project was performed with an agile approach with the goal of incremental advances. To do so, the main steps of the modeling process, namely the exploratory data analysis, the pre-processing, the modeling and the evaluation were continually improved over iterations (Figure 1.1). The following documentation is structured in the same way, with sections corresponding to chapters of this workflow.

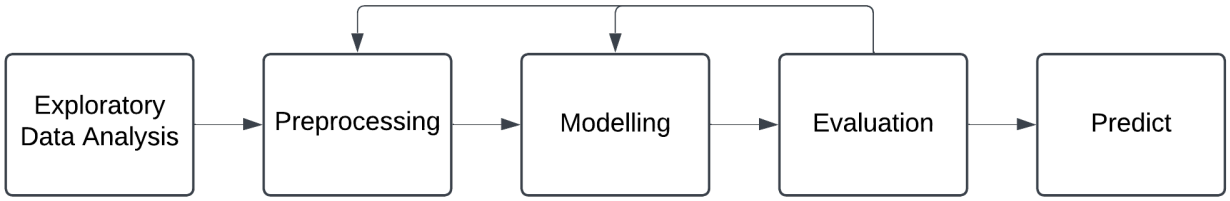

Figure 1.1: Workflow of the project

Reflecting this workflow, a forecasting pipeline was built in parallel where the improvements in preprocessing, modelling and evaluation were integrated, while model development was done in separate exploratory notebooks to provide sufficient flexibility. Models developed in these notebooks could then be loaded into the pipeline to be evaluated and to predict for the test set.

All code used in this project is stored in the public git repository GitHub.

2. EXPLORATORY DATA ANALYSIS

The exploratory data analysis is designed to gain a deeper understanding of the underlying data characteristics and design the prepossessing (section 3) and modeling (section 4) accordingly. Due to limited space available in the documentation of this project, the following is limited to giving a brief overview of the available data and detail the most important findings from an extensive Exploratory Data Analysis.

The variables in the dataset can be grouped by variable type (continuous, categorical, binary - as shown in Table 6.1 and through co-correlations to create a contextual grouping. When sorting the correlation matrix of the continuous variables as shown in Figure 2.1, multiple groups emerge that show noticeable inter-correlation.

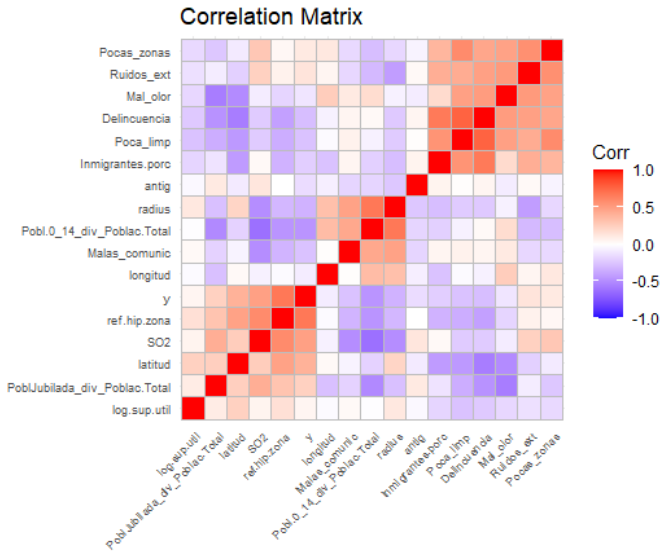

Figure 2.1: Correlations of continuous Variables. y is the objective variable

Looking at the previous figure, along with a combination of domain knowledge gathered from the information as shown in the variable explanation in Table 6.1, we can identify the following groups: "Location" is the group of variables that contains information regarding location of the object, "Object Properties" are the variables giving information on how the object looks like, like how many dormitories or baths the object has, "Quality of Life in Area" gives information on how livable the surrounding area is, "Air Quality" are all variables that contain measurements on pollutants in the area and "Area demographics" are the variables regarding demographic composition of the area.

Table 2.1: Contextually Grouped Variables in Data Set

Group	**Variables**
Location	barrio_name, barrio_code, distrito_name, distrito_code, longitud, latitud, casco.historico, M.30
Object Properties	sup.const, sup.util, tipo.house, inter.exter, ascensor, estado, antig, comercial, ref.hip.zona
Quality of Life in Area	Ruidos_ext, Mal_olor, Poca_limp, Malas_comunic, Pocas_zonas, Delincuencia
Air Quality	CO, NO2, Nox, O3, SO2, PM10
Area Demographics	Pobl.0_14_div_Poblac.Total, PoblJubilada_div_Poblac.Total, Inmigrantes.porc

The target variable is "precio.house.m2", being the sale price in $€/m^2$. The following subsections go into more detail on individual variables, with the sections being structured via the variable type. The train dataset consists of 736 data points and was generated by combining data from multiple sources of data, with the real estate data coming from the Spanish real estate website "Idealista".

2.1. Continuous Variables

The continuous variables of the dataset as shown in the Appendix in Table 6.1 are now investigated, with the most correlated variables and notable observations discussed. The ultimate goal of this project is to develop a predictive model for property prices based on their characteristics. Therefore, understanding which variables

are most strongly associated with the target variable is crucial. To facilitate the analysis and focus on the most relevant predictors, we will only consider those continuous variables with a higher degree of linear association with property prices. Table 2.2 showcases the most relevant linear predictors of the property price in order of importance. To gain deeper insights into the distribution of the most significant variables, we will include a map visualizing their disposition across Madrid.

Table 2.2: Continuous predictors with the highest correlation with property price.

	ref.hip.zona	**SO2**	**Pobl. 0-14**	**latitud**	**CO**	**PM10**	**Malas_comunic**
ρ	0,674	0,494	-0,444	0,346	0,330	-0.274	-0,266

Figure 2.2 shows a significant correlation between the mortgage reference of the area and the property price. As observed, the central districts and certain parts of the north exhibit higher mortgage reference values, which aligns with their position as the most expensive and desirable areas of the city.

In contrast, the southern parts of the city show a notable decline in both mortgage reference and property prices. This suggests that these areas may lack infrastructure, suffer from reduced accessibility to urban centers, or economic disparities.

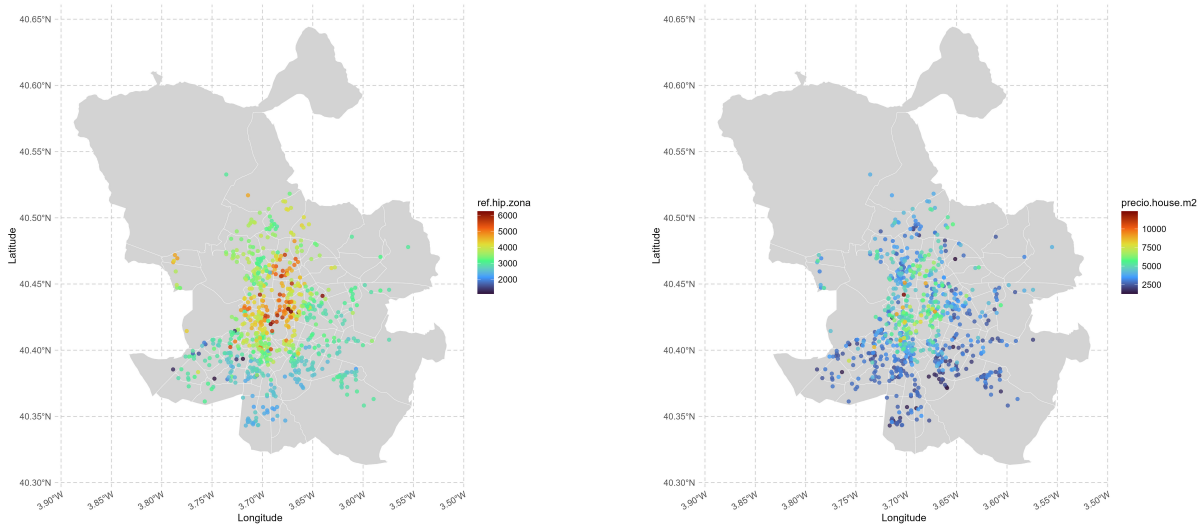

Figure 2.2: Map of Mortgage Reference of Madrid.

In Figure 2.3 (a), we observe that SO_2 concentration levels are highest in the central districts of Madrid, decreasing gradually towards the outskirts of the city. This pattern is likely due to the dense urban activity in the city center, including high traffic volumes. Figure 2.3 (b) illustrates the percentage of children aged 0–14 across the districts of Madrid. The city center has a significantly lower proportion of children, which increases as we move towards the suburbs. This pattern likely reflects two key factors:

- **Housing and affordability constraints:** The central districts are characterized by smaller and more expensive flats, making them less suitable for families with children. This economic pressure often pushes families to relocate to the suburbs, where larger and more affordable housing options are available.
- **Socio-economic dynamics:** As seen in Figure 2.2, the outskirts of Madrid overlap with lower-income areas. It is well-documented that impoverished communities tend to have higher birth rates, which could explain the higher concentration of children in these zones.

Figure 2.3 (c) shows the distribution of how well-communicated properties are across Madrid. As expected, the city center is the best-connected area, benefiting from the highest density of public transportation options.

Interestingly, the disparity between the northern and southern districts, which is evident in other metrics like property prices, is less pronounced in this case. This pattern suggests that the level of connectivity within the city is most strongly influenced by the proximity of individual properties to public transportation infrastructure. For instance, properties located further from subway lines, regardless of their district, tend to exhibit worse connectivity.

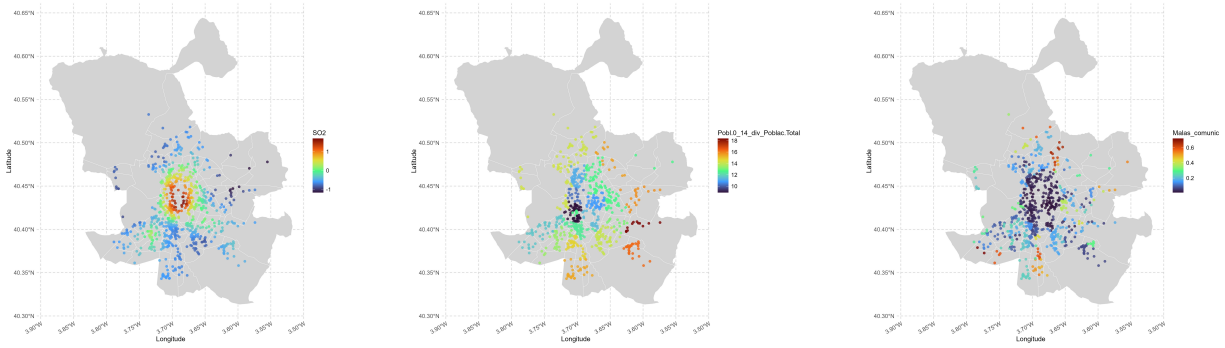

Figure 2.3: Maps of Madrid showing (a) SO_2 levels, (b) percentage of children and (c) poorly communicated areas.

Within the investigation of these relationships, some notable observations were made that are relevant for prepossessing and selection. The observations are listed below:

`precio.house.m2`	In frequency analysis, shows heavy right skewness of the distribution.
`CO`, NO_2, `NOx`, O_3, SO_2, PM_{10}	Air pollution shows a strong correlation to densely populated areas or areas with heavy traffic. For example, SO_2 shows a strong correlation with price, stemming partly from its strong correlation with central locations. Central locations are the actual reason for high prices, so the real correlation between price and SO_2 should be significantly lower.
`sup.const`, `sup.util`	Built area and usable area show an unsurprisingly strong correlation of around 0.85 both equally low correlation to price of around 0.17.
`Pobl.0_14`, `PoblJubilada`, `Inmigrantes.porc`	All three of these population composition measures are defined on the district level, e.g., all data points in a given district have the same values.
`Ruidos_ext`, `Mal_olor`, `Poca_limp`, `Malas_comunic`, `Pocas_zonas`, `Delincuencia`	All of these variables are measured on the district level, e.g., they have equal values for all data points in a given district. Nonetheless, we will consider them to be continuous because the variable in itself could take a non-countable set of values.

2.2. Categorical Variables

The categorical variables shown in Appendix Table 6.1 are investigated for their correlation and for relevant behaviors. We are interested in those features that present the bigger differences in the distribution of property prices between levels. To evaluate the impact of categorical covariates on the objective variable, we employ

the *Eta Squared coefficient* (η^2). This statistical measure quantifies the proportion of variance in a continuous variable that can be attributed to a categorical variable, serving as an effect size indicator commonly used in ANOVA. Higher values of η^2 imply a greater influence of the categorical variable, indicating stronger differences in the distribution of the continuous variable across its levels.

Table 2.4: Categorical predictors with the highest η^2 with property price.

	Barrio	**Distrito**	**Dorm**	**Baños**	**Tipo Casa**	**Estado**
η^2	0.544	0.410	0.012	0.056	0.050	0.027

Table 2.4 displays the relation that exists between the categorical predictors of the property price in order of importance. Bearing in mind that the variables *barrio/district* and *cod_barrio/cod_district* contain the same information we only included the first pair.

As we did in the previous part, we will include a map visualizing their disposition across Madrid.

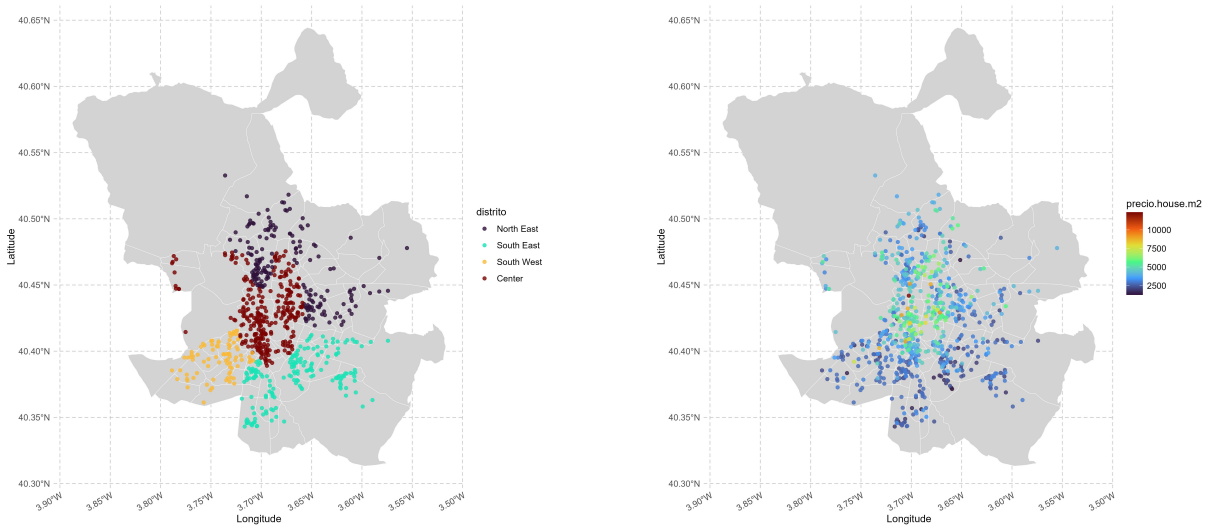

Figure 2.4: Map of Madrid's districts.

Within the categorical variables, only one relevant observation was made and is listed below:

`barrio/cod_barrio, distrito/cod_distrito`	For both Barrio and District, there are two equivalent columns in the data, one for the name and one for a numeric code. Furthermore, barrio is a very fine separation with 118 levels over 736 data points, some barrios only with one data point.

2.3. Binary Variables

As for the previous categories of variables, the binary variables as shown in Appendix Table 6.1 are investigated. We will also analyze the importance of binary variables in explaining the distribution of property prices. As we did for the categorical variables, the *Eta Squared coefficient* (η^2) will be employed as a metric to quantify their influence.

Table 2.6 includes the *Eta Squared coefficient* for the binary variables with the property price on decreasing order. Furthermore, the following observations have been made.

Table 2.6: Binary predictors with the highest η^2 with property price.

	Comercial	**Casco Hitórico**	**M.30**	**Ascensor**	**Interior-Exterior**
η^2	0.330	0.274	0.135	0.105	0.014

`casco.historico, comercial, M.30`

The Figure 2.5 includes a plot of the historic center in the left, the commercial area in the center and the properties of Madrid included inside the M.30 highway in the right. We observe that all of the variables seem to indicate the same thing, proximity of the properties to the city center. It is also clear from the plots the high correlation that exists between the variables, especially between the variables *casco.historico* and *comercial* where all objects in the historic center also lie in a commercial area, while 54 out of 309 objects that lie in commercial areas do not lie in the historic core.

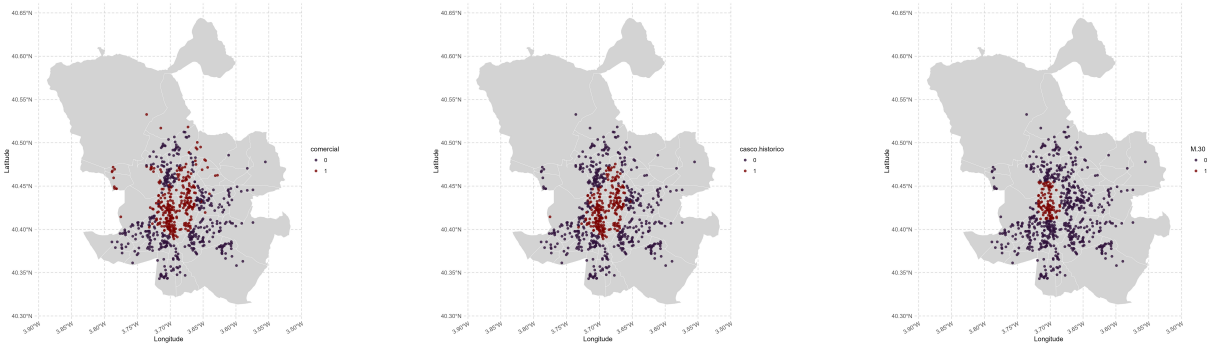

Figure 2.5: Map of the historic center, commercial areas and M.30 of Madrid

3. PREPROCESSING

After an extensive exploration of the dataset in the previous chapter, some of the observations like distribution skewness have to be taken into account in the prepossessing step. This step modifies the input data to the models that are then used to predict the price per m^2. It is expected that these preprocessing steps improve the overall interpretability and performance of the regression models.

3.1. Continuous Variables

First of all, the predictor numerical variables in the dataset will be standardized to ensure that differences in scale and variability do not disproportionately influence the regression model coefficients. Standardization transforms variables so they have a mean value of zero and a standard deviation of one, allowing them to contribute equally to the model. Additionally, this step is needed to perform some techniques like Ridge regression. The distribution of the variables *precio.house.m2* and *sup.util* reveals a noticeable right skewness, deviating from normality. This deviation may impact some of the regression model assumptions, such as normality and homoscedasticity. To address this, the **logarithmic transformation** of the variables will be applied as a corrective measure before the normalization.

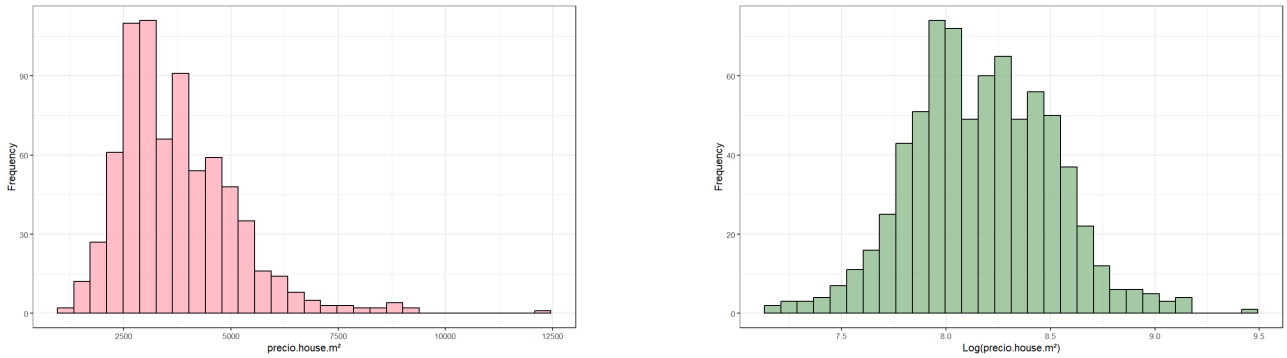

Figure 3.1: Histogram of *precio.house.m2* before and after logarithmic transformation

In addition, the variables *sup.const* and *sup.util* exhibit a high degree of collinearity, which can lead to instability in the regression model and affect the reliability of the estimated coefficients. To address this, *sup.const* will be removed from the analysis as *sup.util* is considered to provide more directly relevant information about the usable space of the property, making it a more meaningful predictor for our purposes.

To simplify the spatial information about each property that the dataset contains in the form of latitude and longitude while retaining as much relevant information as possible, the variable ***radius*** has been created to represent the distance (in kilometers) between each property and the geographic center of Madrid, defined as Puerta del Sol. Apart from reducing dimensionality, this approach provides also a direct and easily interpretable metric from the distance to the city center which often plays a significant role in real estate price. Due to a high degree of multicollinearity within the variables describing air pollution (CO, NO2, Nox, O3, SO2, PM10) as identified in the exploratory data analysis, a decision was made to proceed to modelling only with SO2 as the remaining variable from that group. This decision was taken, in fact not to include the effects pollution has on real estate prices as the isolation of this effect is exceedingly difficult, but because as a byproduct of fossil fuel-powered cars, SO2 appears to be a good indicator of the most urban and central areas of Madrid. SO2 is therefore used to indicate both location and air quality is assumed to be not significant for real estate prices.

3.2. Categorical Variables

Since the variables *cod_barrio* and *cod_district* give the same information as *barrio* and *district* coded in different formats the variables *cod_barrio* and *cod_district* will be removed.

Additionally, the variable *barrio* will also be removed due to the high amount of factors it contains. Rendering is impossible to interpret, while presenting similar information to *district*. Furthermore, the dataset contains some categorical variables with an excessive number of levels. This makes model interpretability too complex. Moreover, a high number of categories can result in insufficient data for meaningful interactions with continuous variables, hence reducing the significance of the results. To address these issues, a recategorization of the variables has been implemented grouping those levels that present some rational similarities, ensuring that the number of categories remains reasonable.

Table 3.1: Comparison of the original and recategorized **room** distributions.

dorm	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Count	9	130	226	254	74	25	11
%	1.04	15.06	26.14	29.39	8.56	2.89	1.27

dorm	**0-1**	**2**	**3**	**≥4**
Count	139	226	254	117
%	16.11	26.14	29.39	13.56

Table 3.2: Comparison of the original and recategorized **bathroom** distributions.

banos	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Count	459	215	39	14	5	0	4
%	62.36	29.21	5.29	1.90	0.67	0.00	0.54

banos	**1**	**2**	**≥3**
Count	459	215	62
%	62.36	29.21	8.42

Table 3.3: Comparison of the original and recategorized **type of property** distributions.

tipo.casa	**ático**	**chalet**	**dúplex**	**studio**	**otro**	**piso**
Count	42	19	17	20	1	637
%	5.71	2.58	2.31	2.71	0.14	96.55

tipo.casa	**Piso**	**Ático/Estudio**	**Chalet/Dúplex**
Count	638	62	36
%	87,97	8,54	4,96

Table 3.4: Comparison of the original and recategorized **state of the property** distributions.

estado	**a-reformar**	**bueno**	**excelente**	**nuevo**	**reformado**	**malo**	**segunda-mano**
Count	94	559	5	20	32	2	24
%	12.77	75.95	0.68	2.72	4.35	0.27	3.26

estado	**Malo**	**Medio**	**Alto**
Count	96	583	57
%	13.74	79.21	7.74

The final categorical variable to be recategorized is *distrito*. The districts will be grouped into broader geographical zones as it was shown in Figure 2.4, this approach aligns with the observed distribution of property prices, providing a meaningful and interpretable grouping.

3.3. Binary Variables

As the last step in the preprocessing, the features *M.30*, *comercial*, and *casco.historico* refer to similar areas, overlapping in the information they provide. To simplify the model and avoid colinearity, we will keep only the variable *comercial*, as it covers a broader range of locations.

4. MODELING

The following Chapter contains the documentation of the modelling process as well as the evaluation of the chosen models. To do so, first, the evaluation methods used are defined before the three model types: Linear models, Ridge Regression and General Additive Models are presented and interpreted. The last section contains the selection of the best model which will then be used for predicting the test set of the competition.

4.1. Evaluation Methods

The evaluation process is explained in the following subsections.

4.1.1. Cross Validation: K-Fold Evaluation of RMSE and R_a^2

In order to test the model's efficiency and ensure that they are not overfitting to the training data a k-fold cross-validation procedure will be implemented. The dataset will be divided into $k = 4$ subsets, with each subset serving as the testing set once during the iterative process. In each iteration, the model will be trained on the remaining three folds. The chosen error metric will be calculated for each iteration, and the average of these values will be used to evaluate the overall performance of the model. To ensure that the folds are

properly balanced (meaning they contain a similar proportion of instances for each level of the categorical variables) a fixed seed has been set in the model code. This guarantees reproducibility and consistency in the cross-validation process. The choice of this seed was determined through the following process: a loop was implemented to test various seeds, calculating the proportions of each categorical variable across the folds, until coherent enough balance was achieved. As a result, the seed selected was **248**. All RMSE in this work are calculated through this procedure, as well as the adjusted coefficient of determination, R_a^2.

4.1.2. Condition Number & Variance Inflation Factor

Both the Condition Number as well as the Variance Inflation Factor are measures for the colinearity of variables of the model The Variance Inflation Factor (VIF) measures the increase in Variance of a regression coefficient as a result of multicollinearity. The VIF of a variable j is calculated as

$$\text{VIF}(X_j) = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination in the prediction of the given variable and the rest of the selected ones. $VIF \geq 10$ indicates multicolinearity between variables. The conditional number on the other hand, measures the overall correlation of numeric variables in the whole model/subset of the data. It is either based on the eigenvalues of the model matrix or in those of its correlation matrix, and it is calculated as follows

$$\text{C.N} = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \quad \text{where } \lambda_{max}, \lambda_{min} \in \sigma(X).$$

A C.N < 10 indicates negligible, 10 ≤C.N< 30 moderate and C.N > 30 severe multicolinearity.

It is relevant to note that we will calculate these measures after the selection of predictors in a given model, so as to assess if its resulting metrics are trustworthy.

4.2. Linear Models

To explore different approaches, we will construct two linear models: a sparse model that prioritizes interpretability by including only the most relevant predictors, and a non-sparse model aimed at maximizing performance by incorporating a larger set of variables. These models will then be evaluated and compared based on their RMSE, selecting the model yielding the smallest RMSE.

- For the sparse model, we begin by including all variables without accounting for any interactions between them. Since such a model would be impossible to interpret, we then apply a combination of forward and backward stepwise selection techniques, guided by the Bayesian Information Criterion (BIC). This will help us identify the most relevant variables while maintaining a sparse model structure. Once the key variables are selected, we add to them the variables *sup.util* and *tipo.casa*, because we consider them to be significant. Afterwards, we analyze the interactions between the selected categorical and continuous variables, analysis from which we include a single interaction: *sup.util* with *comercial*.

- For the non-sparse model, we begin with a full model that includes all variables with no interactions as before. From this initial model, we used a combination of forward and backward stepwise selection techniques based on the Akaike Information Criterion (AIC) to identify the most significant variables, resulting in a reduced model. Next, we directly incorporate all interactions between the selected variables . We apply stepwise selection using the Bayesian Information Criterion (BIC), further refining the interactions included and finally, we add the variable *dorm* to the model, as it is considered relevant.

The tables below shows the corresponding coefficients and their p-values for both of these models.

Table 4.1: Sparse linear model Coefficients

Variable	Estimate	p-value
Intercept	8.2710	$< 2e{-}16$
latitud	0.0582	$2.18e{-}08$
ref.hip.zona	0.0887	$5.96e{-}11$
Poca_limp	-0.0512	$2.66e{-}05$
Pocas_zonas	0.0325	0.0079
radius	-0.0256	$5.45e{-}10$
log.sup.util	-0.0554	0.0068
dorm3	-0.0748	$5.11e{-}06$
dorm0-1	0.0736	0.0060
dorm4+	-0.0512	0.0004
banos1	-0.1038	< 0.0001
banos3+	0.1145	0.0037
ascensorsi	0.0704	$1.47e{-}08$
estadoBajo	-0.1171	$7.28e{-}06$
estadoAlto	0.0763	0.0027
comercial1	0.0719	$2.14e{-}06$
tipo.casaAtico/Estudio	0.0625	0.0348
tipo.casaChalet/Duplex	-0.1038	0.0020
log.sup.util:comercial1	0.0564	0.0012

Table 4.2: Non-sparse linear model Coefficients

Variable	Estimate	p-value
Intercept	8.1713	$< 2e{-}16$
latitud	0.0644	$1.77e{-}08$
ref.hip.zona	0.0798	$1.92e{-}09$
antig	-0.0282	0.0044
Poca_limp	-0.0370	0.0001
Pobl.0_14_div_Poblac.Total	-0.0007	0.9618
radius	-0.0174	0.0009
log.sup.util	-0.0097	0.7854
banos1	-0.0607	0.0147
banos3+	0.0992	0.1539
tipo.casaAtico/Estudio	0.0505	0.0885
tipo.casaChalet/Duplex	-0.1074	0.0079
ascensorsi	0.0757	0.0001
estadoBajo	-0.1127	$6.55e{-}06$
estadoAlto	0.0674	0.0211
dorm3	-0.0846	0.0009
dorm0-1	0.0312	0.2667
dorm4+	-0.1167	0.0012
comercial1	0.1398	$9.13e{-}09$
log.sup.util:banos1	-0.0919	0.0048
log.sup.util:banos3+	0.1041	0.0345
Pobl.0_14:casaAtico/Estudio	-0.2569	$2.65e{-}06$
Pobl.0_14:casaChalet/Duplex	-0.0643	0.0006
log.sup.util:comercial1	0.0460	0.0084

In Table 4.3 no significant improvement in prediction power is shown from the non-sparse model with respect to the sparse one, only 6.69 €/m^2 in the linear scale and 0.003 points in the logarithmic one. This improvement is not enough to justify the increase in complexity in interpretability, so therefore we will only consider the sparse model described in Table 4.1 for the rest of the project.

Table 4.3: Comparison of Sparse and Non-Sparse Models via K-fold

Model	$\mathbf{RMSE_y}$	$\mathbf{RMSE_{log(y)}}$	R_a^2
Sparse	871,438	0.212	0.567
Non-Sparse	864,748	0.209	0.559

The interpretability of linear models is one of their greatest strengths, so to conclude this section, we will provide a brief interpretation of the Sparse model outlined in Table 4.1. The main conclusions are as follows:

- The state and condition of the property are key factors for their pricing. Properties that are newer or have undergone significant renovations tend to hold higher value compared to older or poorly maintained ones. Also having an elevator increases the price.
- Given that the objective variable is the price per m^2, smaller properties such as flats or studios tend to have a higher price per m^2 compared to larger ones like chalets. This trend is likely due to smaller properties often located in high-demand areas, even though they may lack certain amenities. An exception to this pattern can be observed in luxury real estate, where features like having more than three bathrooms (*banos3+*) can significantly increase the price per m^2.

- The zone of the city seems to be of great significance when pricing a property as can be seen by the inclusion of the variable *ref.hip.zona* in the model. As well, there seems to be a strong tendency that properties with higher latitude (North parts of the city) are more expensive, but the further the property is form the city center, the lower its expected value.
- The abscense of certain predictors which we thought to be relevant is noticeable, such as *distrito* or *antig* which, even though they were tested, they supposed no significant increase in the predictive power of the model, potentially due to the multicollinearity they may have with several of the selected variables, so they add no extra information to the model.
- The linear model has, as objective variable *log(precio.house.m2)*, so we have to keep in mind that the effect of the covariates in the non-transformed objective is multiplicative.

Before using a linear regression model, it is important to diagnose the model by means of a residual analysis. If we look at the subplots of Figure 4.1, we could say that the linearity assumption is met, that there is no exaggerated heterocedastic behavior and that the residuals deviate from normality, mainly if we look at the tails. No point exceeds the Cook's distance, so there are no influential points, but there are some leverage points (Which may very well be the cause for the slight deviation from normality at the end-points of the QQ-plot). Although the residual plots are not perfect, we can assume that they meet the assumptions of the model reasonably well. Hence, we will opt not to eliminate observations, primarily because some may belong to sparse categories in one or more predictors, and their absence could limit the range of properties needed for accurate predictions.

As well, the selected subset of predictors show no sign of multicolinearity issues, with C.N. = 9.21 and all VIF values among 1 and 4.5.

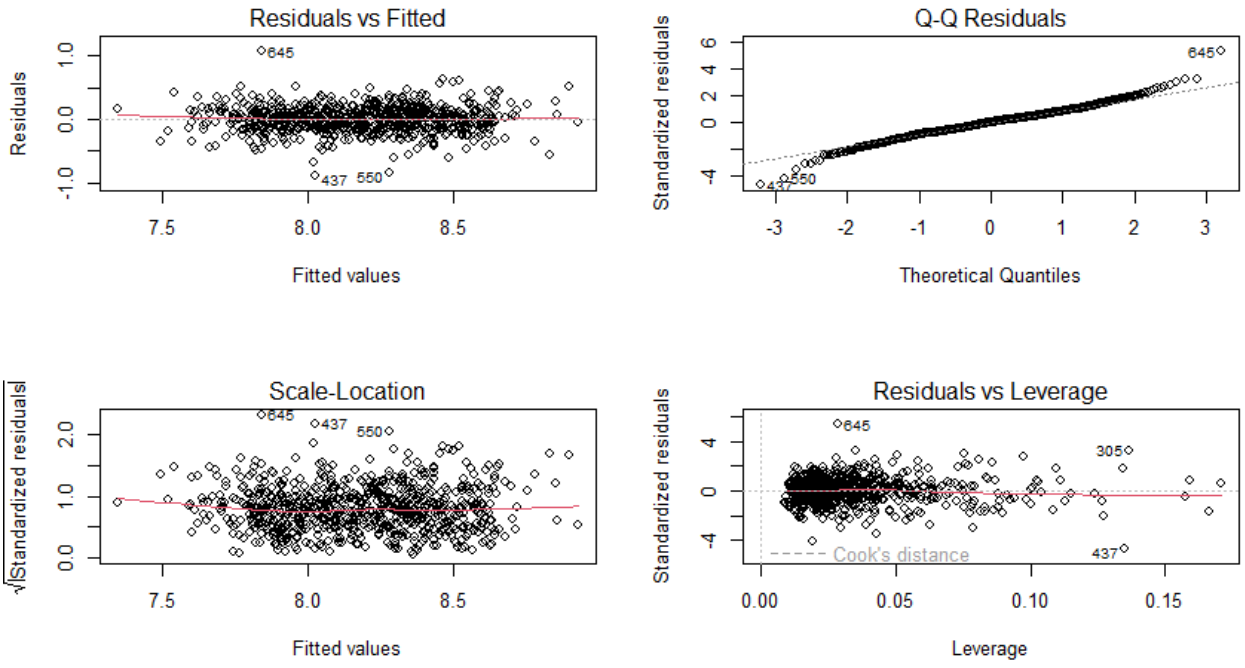

Figure 4.1: Sparse Linear Model - Evaluation Plots

4.3. Ridge Regression

In case of multicollinearity, an inherent issue in the sample, a good option to address it is Ridge regression. Ridge regression is a regularization method that adds a penalty term (λ) to the ordinary least squares problem. This penalty discourages large coefficients by shrinking them, particularly for variables that contribute to multicollinearity. As a result, Ridge regression reduces the influence of highly correlated variables and improves the model's stability.

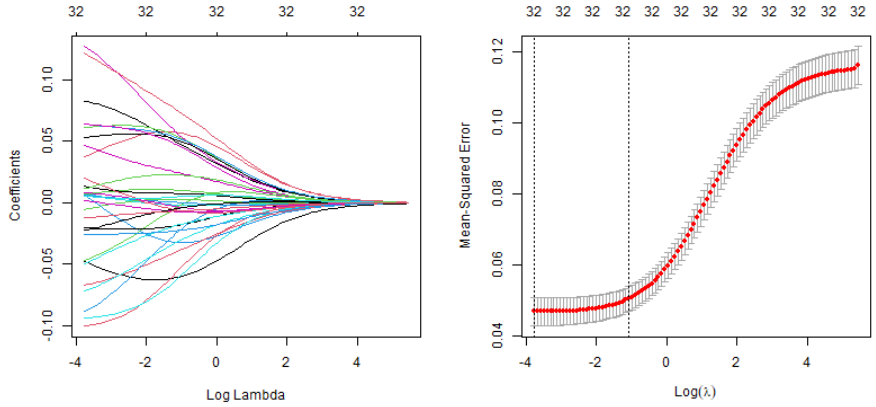

Figure 4.2: Ridge model: Selection of λ and its effect in the predictors' importance in the model

Using techniques such as cross-validation, the program evaluates different values of the hyperparameter λ. The objective is to find the optimal λ that minimizes the error in the validation dataset.

Once this optimal λ is identified, it is fixed and the final model is trained using all available data. Thus, the hyperparameter is no longer subsequently adjusted. Table 4.4 shows the K-fold cross-validation metrics the Ridge model showcases.

Table 4.4: Comparison between the two Ridge best models with interactions

RMSE (Log)	**RMSE (Original)**	R_a^2 **(Log)**	**Optimal** λ
0.214	886.294	0.521	0.0277

It is noticeable in Figure 4.2 how, in the first plot, no clear order of predictors converge to 0 (Even though several of them already start close to it, probably because of their lack of relation with the objective) and most of then do at similar values of λ. This may have primarily been caused, as we have already commented, on the fact that most of these predictors are location-driven, so a big portion of information is shared among them.

As well, this model is worse than both the linear models considered, probably because, since we have previously dealt with multicollinearity in them, Ridge doesn't really imply an improvement in comparison.

4.4. General Additive Model: P-Splines

Given the importance the location seems to have in the analyzed linear model, as well as the difficult-to-identify structure/relationship of several covariates with the objective variable (mainly due to the real-world nature of the data), it could be an interesting idea to fit a GAM Model. This way, we can identify non-linear relationships not-so-easily-observed between the objective and covariates, which may improve the predictive power of the model.

The methodology we will use is as follows: Assuming that the selected covariates have an additive effect on the objective function and starting from the two linear models we analyzed, we will, for each model, build two GAM's:

- Fit a GAM with all the linear model's predictors, taking the numerical ones as smoothing terms.
- Fit a second GAM, similar to the previous but excluding any additional location covariates other than the tensorial product of *longitud* and *latitud*, which we believe is the most informative way to incorporate location data into the model (as shown in Figure 2.4, a surface, which is what the tensorial product does, seems to be the best approach).

As the smoothing terms, we will use P-Splines of degree $m = 2$ and 15 or 20 knots, depending on the distribution of the given covariate. The optimal smoothing parameter is automatically selected by the program.

Afterwards, for each of the four constructed models, we analyze the effective degrees of freedom of each smooth term to acknowledge if it should be included as linear ($edf < 2$) or not. Then, the models are refit taking the previous idea into consideration and finally we compare the cross-validated RMSE and R_a^2.

The resulting GAM model is the following:

Table 4.5: Model Coefficients

Variable	**Estimate**	**p-value**
Intercept	8.2626	$< 2e{-}16$
latitud	0.0664	$4.24e{-}09$
Poca_limp	-0.0514	$4.00e{-}05$
radius	-0.025	$2.11e{-}09$
log.sup.util	-0.0609	0.0028
dorm3	-0.0722	0.00058
dorm0-1	0.0683	0.0100
dorm4+	-0.1090	0.0014
banos1	-0.1038	$3.71e{-}06$
banos3+	0.1115	0.0046
ascensorsi	0.0893	$4.22e{-}06$
estadoBajo	-0.1128	$5.11e{-}06$
estadoAlto	0.0735	0.01336
comercial1	0.1396	$3.57e{-}08$
tipo.casaAtico/Estudio	0.0575	0.0508
tipo.casaChalet/Duplex	-0.0928	0.0236
log.sup.util:comercial1	0.0587	0.0006
Smooth Terms	**edf**	**p-value**
s(ref.hip.zona)	2.328	$< 2e{-}16$
s(Pocas_zonas)	2.396	0.0022

Notably, the only two terms which were admisable as non-linear were *ref.hip.zona* and *Pocas_zonas*, although their effective degrees of freedom don't show great non-linearity. This can as well be seen in Figure 4.3, in which one can't visually distinguish non-linearity and the main deviations from linearity happen at the extreme points of the data, which may suggest an overly influential effect of those points.

As well, it is remarkable that the best GAM model was, against our intuition, one without the tensor product. One could argue that the information given in that term in potentially contained already in some other non-location covariates, but a deeper analysis is needed to conclude that.

The model shows the following cross-validation metrics, where the adjusted determination coefficient is calculated as

$$R_a^2 = 1 - \frac{\text{SSE}/\text{EDF}_{\text{residual}}}{\text{SST}/(n-1)}.$$

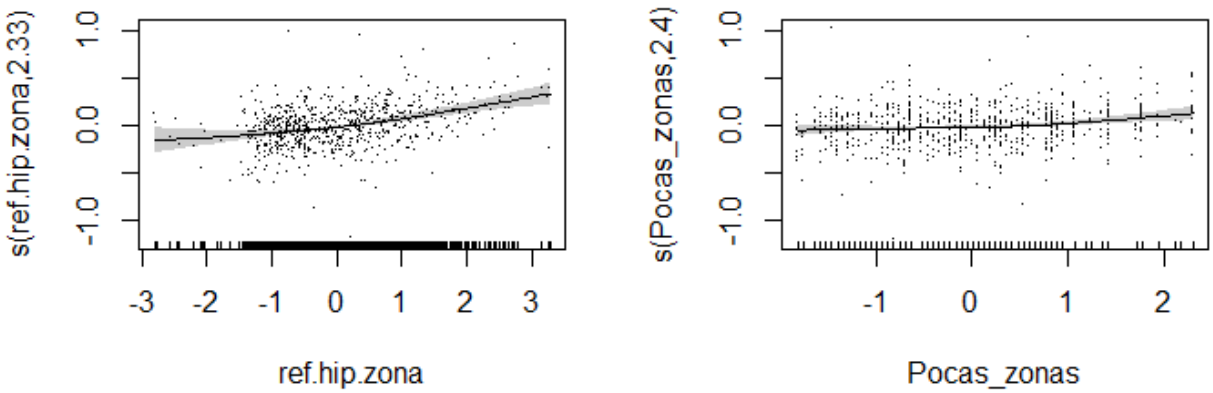

Figure 4.3: Individualized fits of the smooth terms

Table 4.6: Error Metrics for the GAM model via K-fold

RMSE (Log)	**RMSE (Original)**	R_a^2 **(Log)**
0.211	869.171	0.615

To conclude this section, we can argue that as observed in Tables 4.3 and 4.6, the inclusion of the smooth terms doesn't seem to have a significant impact in the model performance compared with the linear one. As such, the linear behavior of the selected covariates is evident.

4.5. Model Performance Comparison

The comparison between model performances will be made on the basis of the k-fold cross validation results (the performance of a model on different partitions of the data is evaluated) for each of the models trained in this task. As mentioned above, the evaluation metric is RMSE, so the model with the lowest RMSE is the model of choice. Although this is the canonical way to proceed, it is also important to take into account model complexity, or interpretability.

Table 4.7: Comparison of all models via K-fold

Model	$\mathbf{RMSE_y}$	$\mathbf{RMSE_{log(y)}}$	R_a^2	$\sigma(\mathbf{MSE_y})$
Sparse Lm	871,438	0.212	0.567	151.82
Non-Sparse Lm	**864,748**	**0.209**	0.559	**130.05**
Ridge	886.294	0.214	0.521	151.85
GAM	869.171	0.211	**0.615**	167.13

Table 4.7 evinces that the Non-Sparse linear model is both the most accurate and stable (least RMSE_y and $\sigma(\text{MSE}_\text{y})$). Although it doesn't show the highest R_a^2 among the models, we will use it following the cross-validation methodology we have developed.

5. CONCLUSIONS

After extensive exploratory data analysis, preprocessing, and model-fitting with techniques such as linear regression, ridge regression, and generalized additive models (GAMs) with P-splines, results reaffirm that the best-performing model is not necessarily the most complex, but an intelligently-built linear model. The main challenges arose in preprocessing and variable selection due to high multicollinearity and redundancy in the raw data. Particularly, we faced difficulties isolating correlations—e.g., the effect of air pollution on real estate prices was confounded by strong correlations with location (urban centers).

Despite these challenges, effective preprocessing decisions enabled fairly well-performing models, with the non-sparse linear model delivering the highest predictive power and the highest consistency cross-validation. Preprocessing had the greatest impact on model performance, suggesting further improvements in generating better location-based variables (e.g., "radius" was only marginally significant due to Madrid's multicentric nature). Additionally, isolating the effects of air pollution remains a challenge for future work.

To address high multicollinearity, dimension reduction techniques like Principal Component Analysis (PCA) could improve preprocessing before applying further models. While current models achieved substantial predictive power (the selected model is capable of explaining around 56% of the data variability), we could argue that there is a need for more information (potentially non-location related) to further explain the price per m^2 of the real estate properties in Madrid.

Nonetheless, we have reached the following conclusions via the interpretation of the linear models: Estimated mortgage value of the property (*ref.hip.zona*) seems to be the most influential variable in the real estate price, since it directly represents a metric for its value. Further, the location overall information of the property seems to be also of utmost importance, but this information is disseminated throughout several of the studied variables. As well, the state and overall quality of the house is relevant, although seems less important than the previous comments.

6. APPENDIX

Table 6.1: Variables in the Data Set

Variable	Description
Continuous Variables	
precio.house.m2	Property sale price in euros per square meter
antig	Age of the property (years)
Ruidos_ext	External noise (%)
Mal_olor	Pollution or bad odors (%)
Poca_limp	Lack of street cleanliness (%)
Malas_comunic	Poor communications (%)
Pocas_zonas	Few green spaces (%)
Delincuencia	Crime (%)
CO	Level of CO in the air at the property's coordinates (standardized values)
NO2	Level of NO2 in the air at the property's coordinates (standardized values)
Nox	Level of NOx in the air at the property's coordinates (standardized values)
O3	Level of O3 in the air at the property's coordinates (standardized values)
SO2	Level of SO2 in the air at the property's coordinates (standardized values)
PM10	Level of PM10 in the air at the property's coordinates (standardized values)
Pobl.0_14_div_Poblac.Total	Percentage of children between 0 and 14 years in the district
PoblJubilada_div_Poblac.Total	Percentage of retired population in the district
Inmigrantes.porc	Percentage of immigrant population in the district
sup.const	Built area of the property
sup.util	Usable area of the property
ref.hip.zona	Mortgage reference of the area
longitud	Geographical coordinates of the property (longitude and latitude)
latitud	Geographical coordinates of the property (longitude and latitude)
Categorical Variables	
barrio	Name of the neighborhood in the city of Madrid
cod_barrio	Code of the neighborhood in the city of Madrid
distrito	Name of the district in the city of Madrid
cod_distrito	Code of the district in the city of Madrid
dorm	Number of bedrooms
banos	Number of bathrooms
tipo.casa	Type of property
estado	Condition of the property

Continued on next page

Variable	Description
Binary Variables	
inter.exter	Interior or exterior design of the property
ascensor	Elevator availability in the building
comercial	Indicates if the property is located in a commercial area
casco.historico	Indicates if the property is in Madrid's historic center
M.30	Indicates if the property is within the M-30