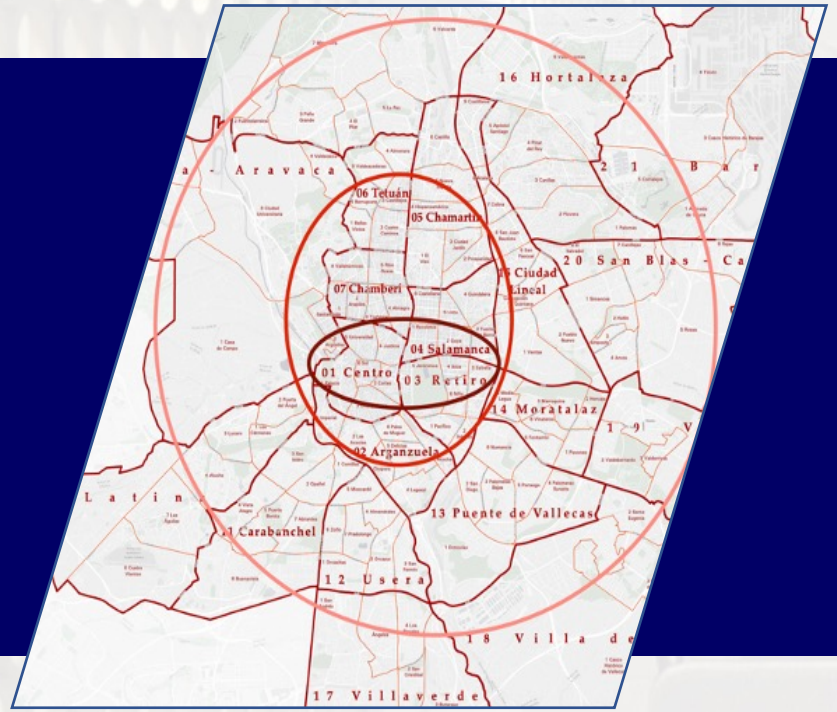
"Unlocking Rent Patterns Through Linear Regressions"

By: César Alejandro García, Ignacio Alonso López-Linares, Kangjie Yu, Lea Brunner & Morten Aas-Lyngby

Made at ie

AGENDA

1. **Executive Summary – How did we achieve this?**
   - A A product-based approach
   - B Business Recommendations
2. **Technical: Linear Regression Modeling**
   - A Preprocessing
   - B Modelling
3. **Annex & Additional Materials**

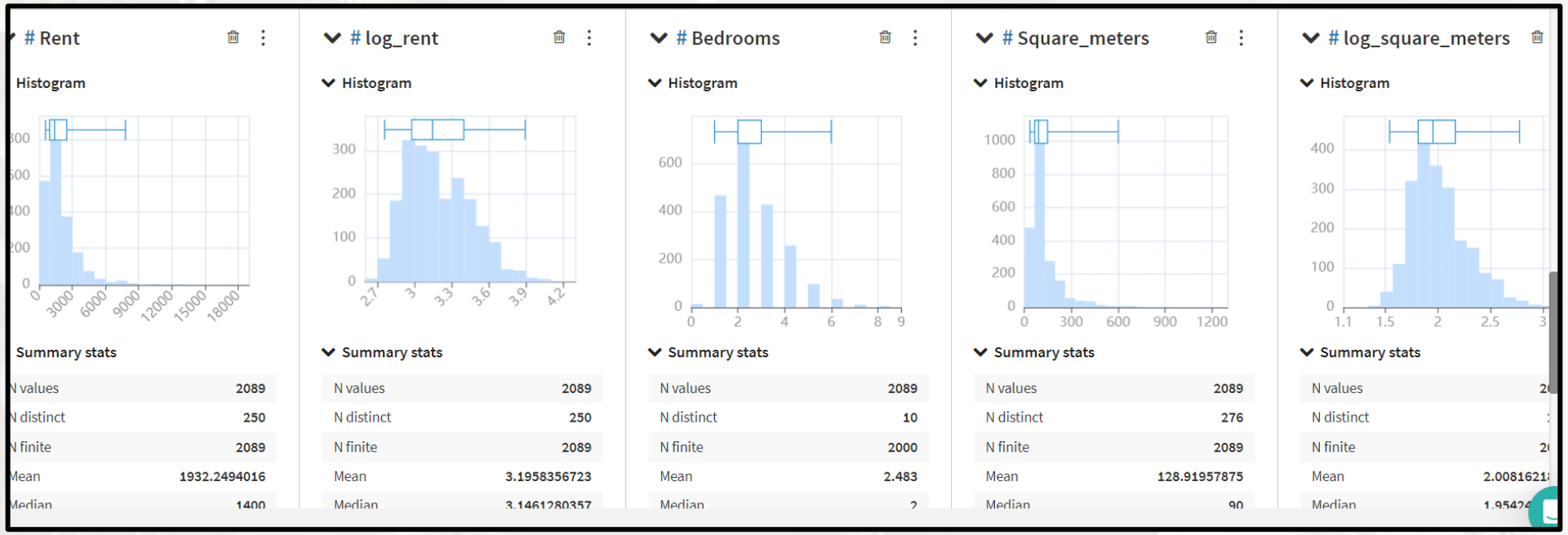


Executive Summary

- Data Understanding
- Univariate & Bivariate
- A Product Approach
- Mapping the Results

Executive Summary – Data Exploration: Univariate

A We found an outlier in the floor variable. The floor was 43039, so it must be something wrong with the input. We see its a chalet, so it doesn't really have a floor. Therefore, we delete the value of floor for this row:

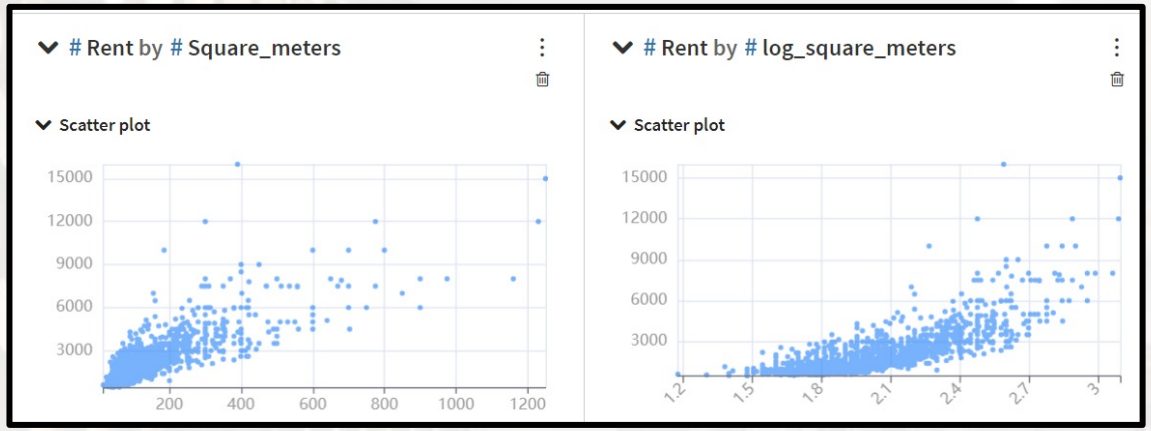


B Rent and square meters are not normally distributed. Due to this we made a variable with the log rent and log square meters and see that the distribution is more normal.

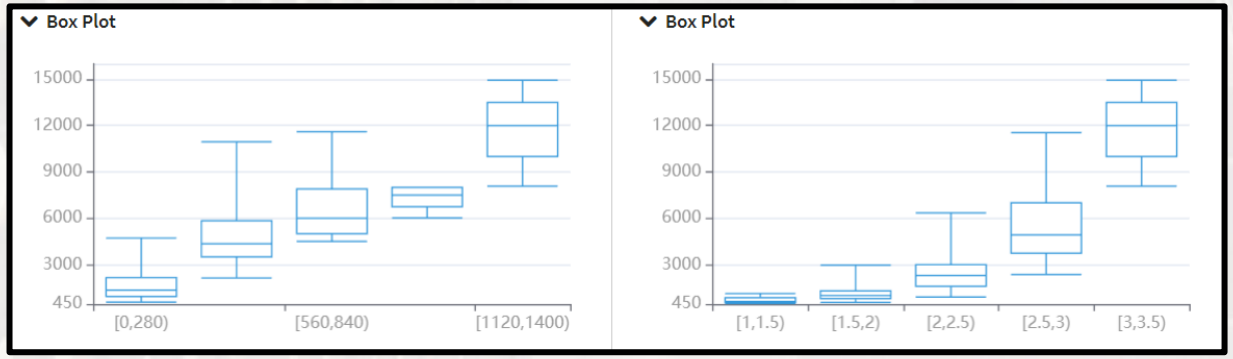
(We decided to not use them because we want to be able to extrapolate the model for business insight, and to keep interpretability for non-technical).

C **Piso is the largest group and casa, chalet and caseron are a small part of the dataset. Class imbalance!!**

Executive Summary – Data Exploration: Bivariate I



Here we see that with rent on log square meters normalizes the relationship and makes it more linear. **Still with some outliers and losing interpretability** (we would have to convert back the square meters to interpret it correctly)

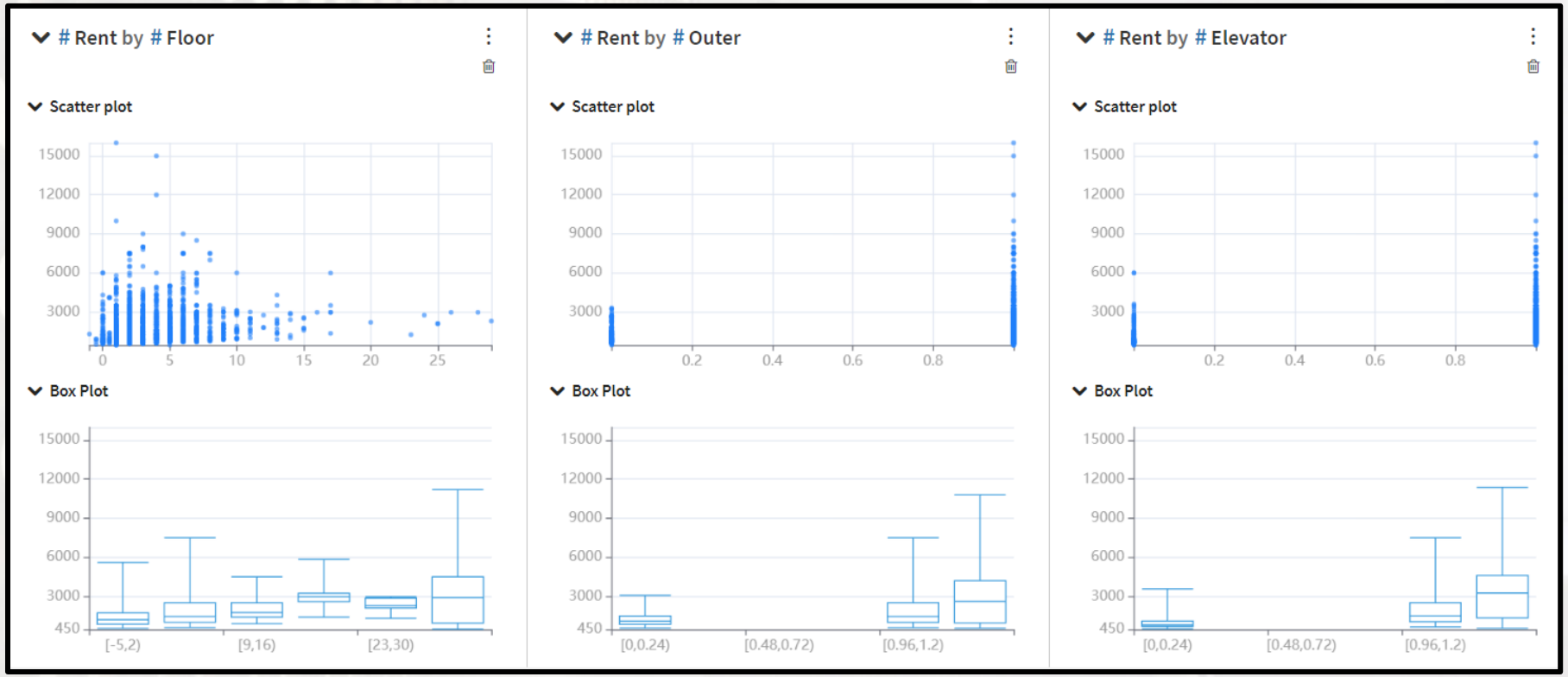


Now, we see how **the property types** are distributed on rent. Looking at prices for the different types: casa, chalet and caserón seem to be a sub-population. Since the dataset is not that big, and we have few observations for these types, **we considered oversampling for these, but this would result in overfitting which such a small sample.**

Type of property is therefore very important. Although this might be caught up in factors such as square meters, bedrooms area and district **(piso in the centre and casa in the suburbs).**

Executive Summary – Data Exploration: Bivariate II

Stepwise Ordinary Least Squares

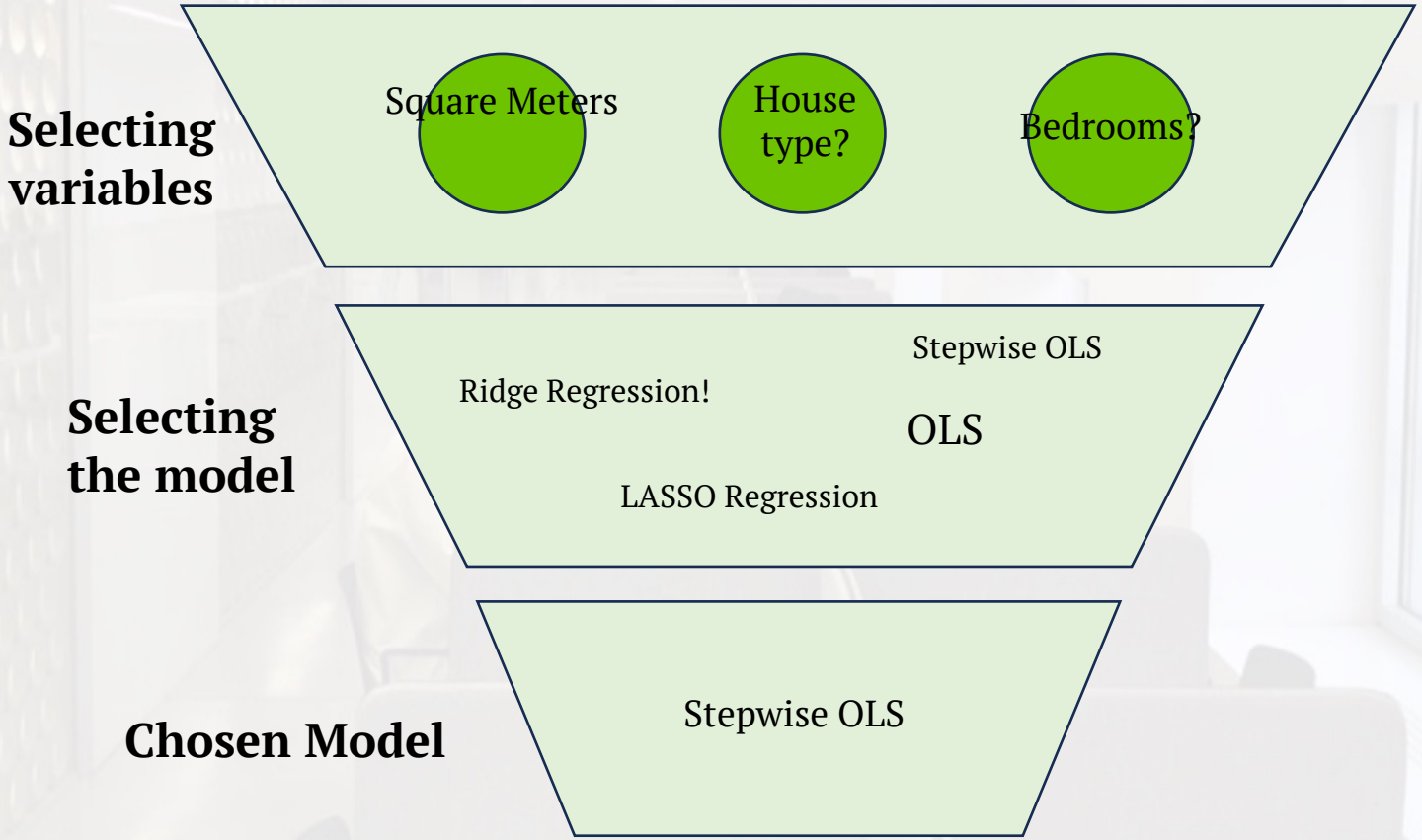


The last box is for the no value rows in elevator. It makes sense that these are the highest, because those are for the most part casa and chalet.

We tried a bivariate with log rent but did not see a substantial improvement and decided to leave it due to functionality.

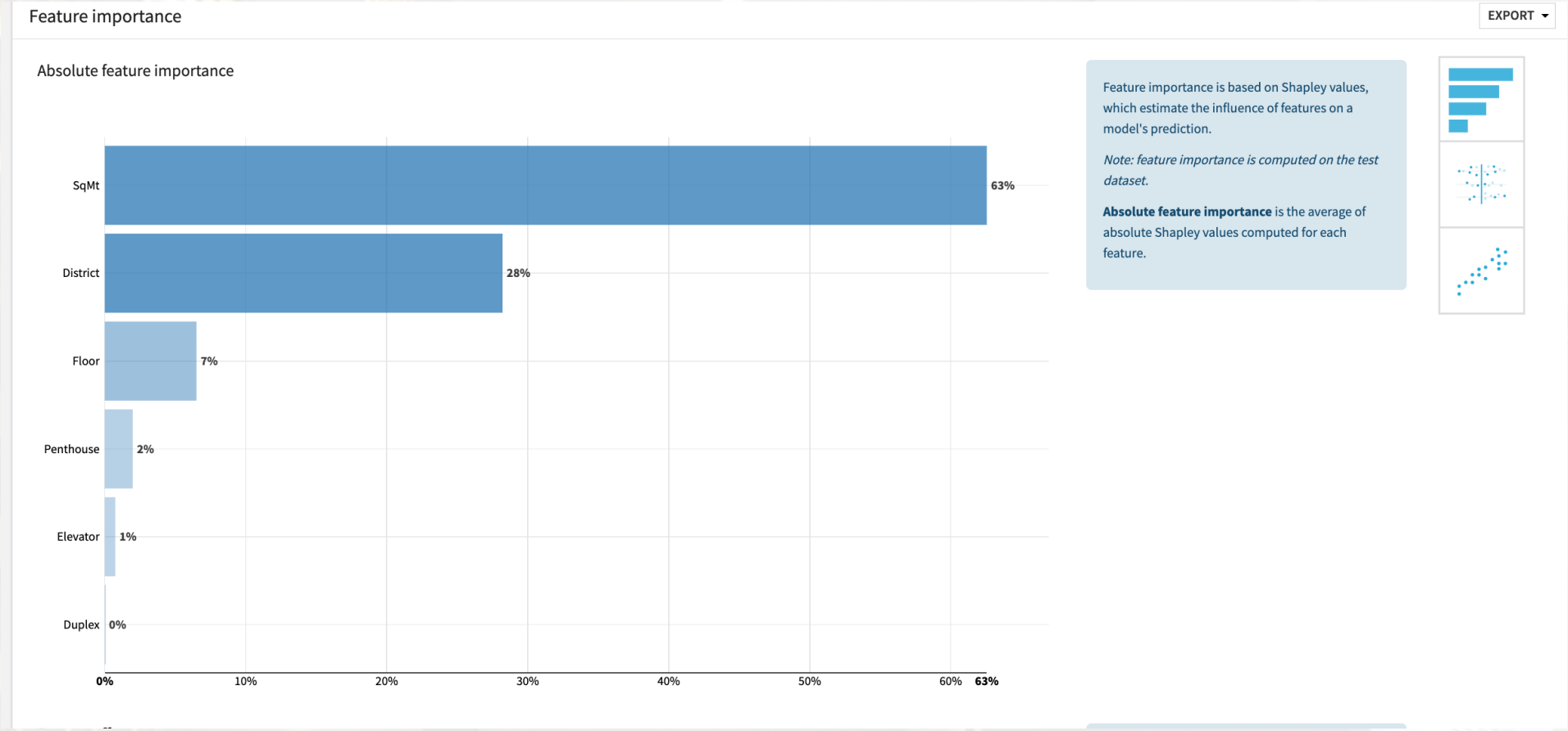
Executive Summary – A Product Approach I

With the objectives in mind, we built an interpretable product.

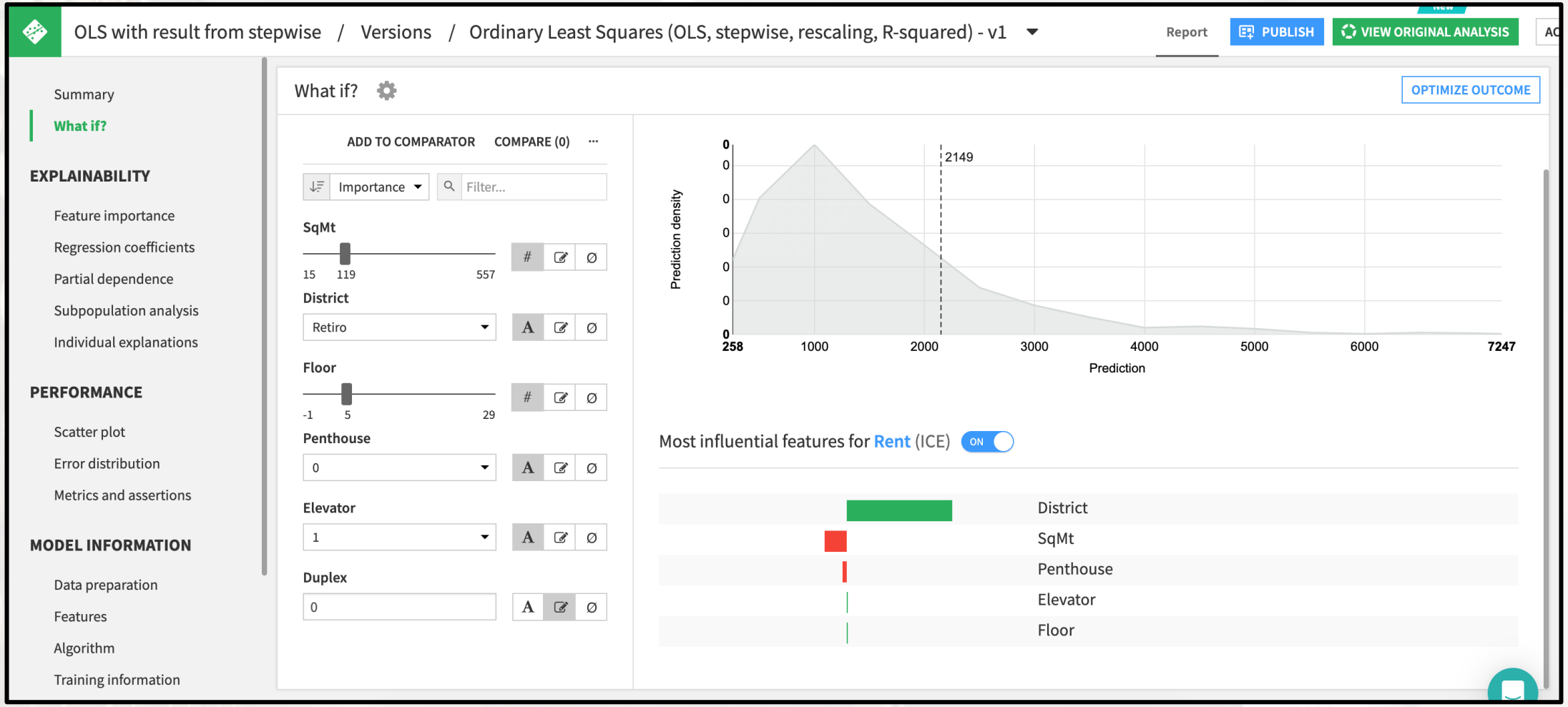


Interpretability was key to us!

Executive Summary – A Product Approach I



Executive Summary – A Product Approach for the Firm II



RISK ASSESSMENT: we built this fantastic product for calculating rent prices for the investment fund. In fact, we tried it with 3 friends that pay rent in Madrid, guessing correctly the 3 rent prices.

Executive Summary - Distribution of Residuals & Volatility

Error distribution

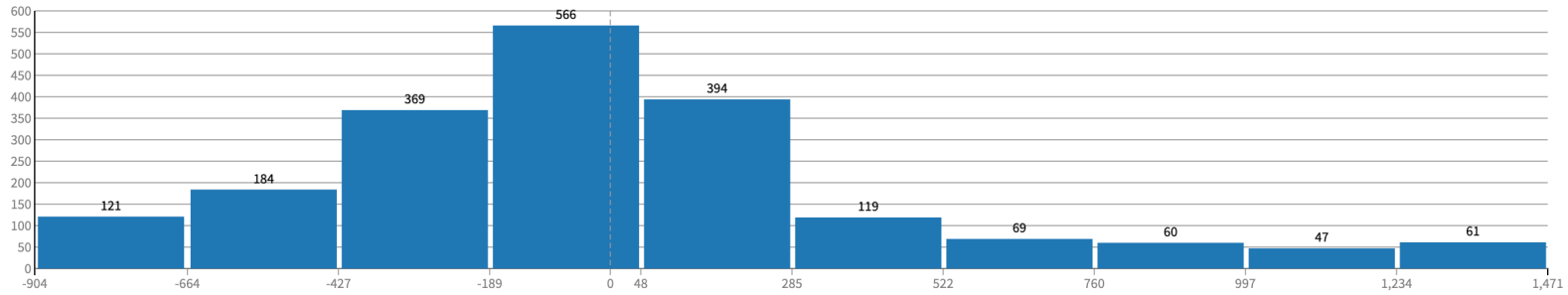
Min. (raw)	Min.	25th perc.	Median	75th perc.	90th perc.	Max.	Max. (raw)
-1993.0	-901.27	-286.19	-56.351	169.33	634.14	1471.5	4280.7

Average -4.7103 **Standard deviation** 478.58

Description

The errors (difference between predicted and actual values) should be centered around zero, and the distribution should be "narrow", i.e the spread of the error should be limited. More generally, the errors should be "normally" distributed around zero (the curve should look like a bell).

To reduce the effect of possible spurious outliers, error distribution is winsorized (clipped) at the 2nd and 98th percentiles.



Normally distributed errors/volatility = OK, some heterocedasticity due to extreme values/outliers but not something to disregard the model

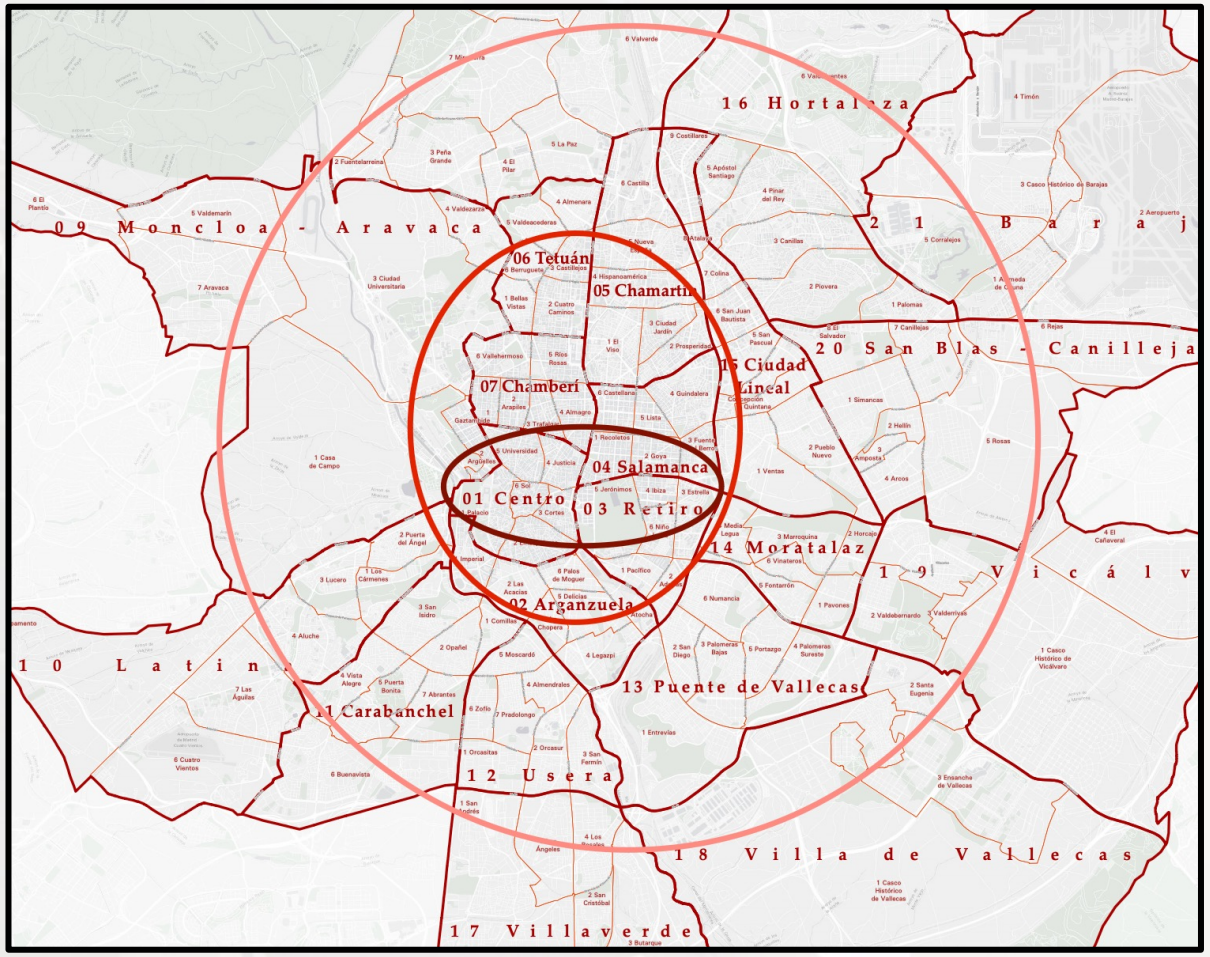
Executive Summary – Mapping the results

Interpretability:

if District is Salamanca or = 1, it will increase rent by 1071€.

If District is Chamberí or = 1, it will do it by 712€.

Variable	Coefficient
District is Salamanca	1,071.5507
District is Retiro	841.3345
District is Centro	809.4617
District is Chamberí	712.3129
District is Chamartín	661.7325
District is Moncloa	542.3748
District is Tetuán	452.6316
District is Arganzuela	339.6134
District is Hortaleza	292.2212
District is Ciudad Lineal	218.2651
District is Puente Vallecas	205.9255
District is San Blás	145.2651
District is Usera	123.3422
District is Fuencarral	113.8374
District is Carabanchel	93.5133
District is Latina	92.8663
District is Villa de Vallecas	78.4843
District is Barajas	50.0559



Executive Summary – Socio-economic factors

Investing in Retiro and Salamanca, two prominent districts in Madrid, Spain, can offer a range of advantages:

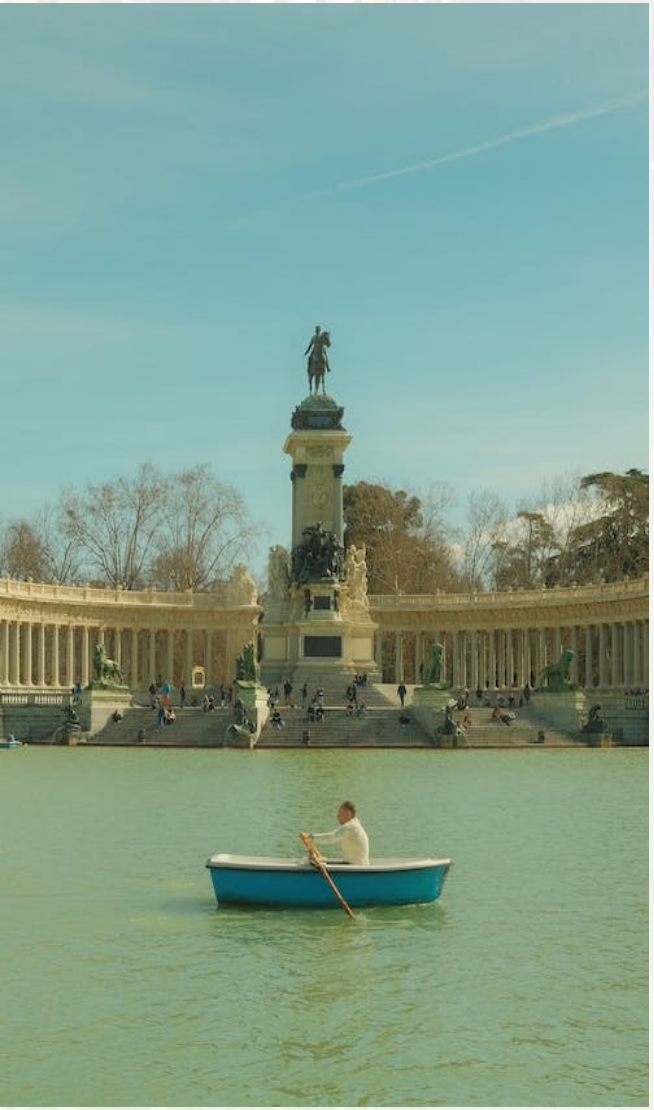
1.Prime Locations and Prestige: Both districts are renowned for their prestige and prime locations in Madrid. Retiro is known for its beautiful park and elegant residential areas, while Salamanca is famous for its luxury shopping streets, like Calle de Serrano, and high-end residential properties. This status often leads to a stable or increasing property value.

2.Attractive Real Estate Market: The real estate market in these areas is considered robust and resilient, with a consistent demand for both buying and renting. Properties in Retiro and Salamanca tend to maintain their value, offering a good opportunity for long-term investment.

3.High Rental Demand: Given their desirable locations, there's a high demand for rental properties. This can lead to potentially lucrative rental yields for investors, especially given the popularity of these areas among expatriates and well-to-do locals.

4. Cultural and Historical Significance: Both districts are rich in culture and history, enhancing their appeal to tourists and residents alike. Retiro is home to several museums and the famous Retiro Park, while Salamanca boasts historical architecture and exclusive shopping areas.

Executive Summary – Socio-economic factors

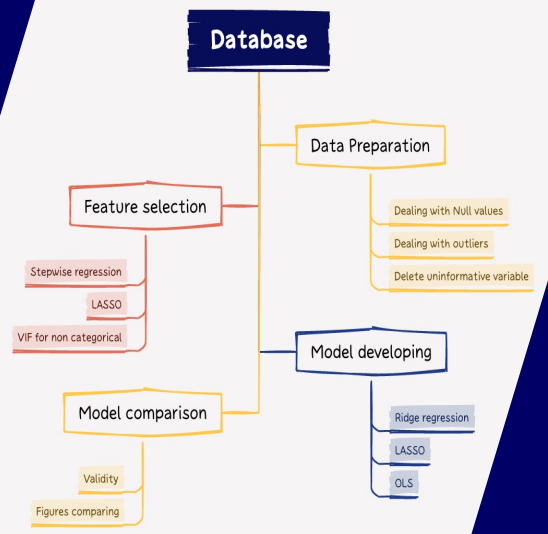


5. Quality of Life: These districts offer a high quality of life with excellent infrastructure, safety, and a range of amenities including parks, restaurants, and shops. This makes them attractive for families and professionals, ensuring a steady demand for properties.

6. Strong Economy of Madrid: Madrid's strong and diverse economy contributes to the stability and growth potential of property investments in its prime districts. As the capital city, it attracts businesses, tourists, and residents, supporting the real estate market.

7. Potential Tax Benefits: Spain offers various tax benefits for real estate investors, which can make investing in these districts more attractive financially.

8. International Appeal: Madrid, and particularly districts like Retiro and Salamanca, have a strong international appeal, attracting investors and residents from around the world. This global interest can further boost the potential for investment growth.



Technical Summary

- Preprocessing
- Feature selection
- Model developing
- Model comparison

Technical Summary – Preprocessing

Univariate

Null values
- bedroom: mostly related to estudios.
- floor: mostly related to casa and chalet.
- outer, elevator: same as floor.

Outliers
- ‘43039’ in floor column.
- Only 1 type of ‘caseron’.

Number
- Not informative.

Distribution
- Rent and square meters are not normally distributed.

Null values
- bedroom: 0 for estudios, mean for the rest.
- floor: mean value for apartments, “not relevant” for the others.
- outer, elevator: “not relevant” for casa, chalet and mode for the rest.

Outliers
- Deleted.

Number
- Deleted.

Distribution
- We created log for both variables. (not used for interpretation to non-techs)

Rent	Bedrooms	Sq.Mt	Floor
bigint	bigint	bigint	double
Integer	Integer	Integer	Decimal
7500	5	310	43039

Property_type (1 value)

Id	District	Address	Property_t...	Area	Rent	log
bigint	string	string	string	string	bigint	dou
Integer	Text	Natural lang.	Text	Text	Integer	Dec
115	Fuencarral	Caserón en calle Río Bullaque	Caserón	La Paz	2100	

Technical Summary –Preprocessing

Univariate

```
# applying mean value for apartments and not re
New = data.groupby('Type')['Floor'].mean()
data = pd.merge(data,New,how='left',on='Type')
data.loc[(data['Floor_x'].isna())&(data['Type']
data.loc[(data['Floor_x'].isna())&(data['Type']
data.loc[(data['Floor_x'].isna()),'Floor_x'] =
```

```
# checking null values
data['Floor_x'].isna().any()
```

```
# Dealing with null values in 'outer','elevator'
data.loc[data['Type'].isin(['Casa', 'Chalet', 'D
data.loc[data['Type'].isin(['Casa', 'Chalet', 'D
data.loc[data['Elevator'].isna(), 'Elevator'] =
data.loc[data['Outer'].isna(), 'Outer'] = 1
```

```
data.Outer.isna().any()
```

```
# dealing with null values in 'bedrooms',0 for estudios and mean for the rest
data.loc[(data['Bedrooms'].isna()) & (data['Type']=='Estudio'), 'Bedrooms'] = 0
New = data.groupby('Type')['Bedrooms'].mean()
data = pd.merge(data,New,how='left',on='Type')
data.loc[data['Type'].isin(['Ático', 'Dúplex']) & data['Bedrooms_x'].isna(), 'Bedrooms_x'] = data['Bedrooms_y']
```

Floor *(NOK)* 84 matching rows

Type	Area	Rent	Bedrooms	Sq.Mt	Floor	Outer	Elevator
string	string	bigint	double	bigint	string	string	string
Text	Text	Integer	Decimal	Integer	Decimal	Decimal	Decimal
Chalet	Aravaca	5000	5.0	530	not relevant	not relevant	not relevant
Casa	El Plantío	6000	6.0	700	not relevant	not relevant	not relevant
Casa	Valdemarín	8000	6.0	1160	not relevant	not relevant	not relevant
Casa	El Plantío	6000	6.0	750	not relevant	not relevant	not relevant
Chalet	Valdebebas - Valdefuentes	2600	4.0	399	not relevant	not relevant	not relevant
Casa	Conde Orgaz-Piovera	4500	5.0	300	not relevant	not relevant	not relevant
Chalet	Palomas	2100	4.0	212	not relevant	not relevant	not relevant
Casa	Aravaca	8000	6.0	900	not relevant	not relevant	not relevant
Casa	Valdemarín	3700	3.0	300	not relevant	not relevant	not relevant

Coding used for manipulating the data

Technical Summary – Feature selection

1. Manual selection

- Address and area excluded since they are too specific compared to district.
- ID excluded since it's mainly for identification.
- Type excluded because it depends strongly on other variables. For example, whether it contains a bedroom or not determines if it's an estudio or piso.

2. VIF

- Though VIF can be used to check multicollinearity, it's not very useful for categorical variables. So only 2 factors are tested.

3. Bidirectional stepwise regression

- Bidirectional stepwise regression is used to select features for the OLS regression because it provides a balanced compromise between purely automated variable selection and manual choice. This method allows us to start with a simpler model and add complexity only when a variable significantly improves the model's performance. Giving the result: Sqmt, floor, bedrooms, duplex, elevator and penthouse.

```python
# -*- coding: utf-8 -*-
import dataiku
import pandas as pd
import numpy as np
from dataiku import pandasutils as pdu
import statsmodels.api as sm

# Read recipe inputs
rent= dataiku.Dataset("Rent_prepared_python_prepared")
df= rent.get_dataframe()

pdu.audit(df)

# Stepwise function
def stepwise_selection(X, y,
                       initial_list=[],
                       threshold_in=0.05,
                       threshold_out = 0.1,
                       verbose=True):
    """ Perform a forward-backward feature selection
    based on p-value from statsmodels.api.OLS
```

SqMt	Floor	Bedrooms	Duplex	Elevator	Penthouse
string	string	string	string	string	string
Text	Text	Text	Text	Text	Text

Technical Summary – Preprocessing

Outer is not relevant	-3,198,384,082,735,243.0000
Elevator is not relevant	2,505,648,264,260,478.5000
Floor is not relevant	692,735,818,472,768.4000
Floor is N/A	-55,715,225.7629
Outer is N/A	-11,650.8477
Floor is -1.0	-1,777.6311
Floor is -0.5	-1,722.7928
Floor is 12.0	-1,589.4824
Floor is 5.0	-1,521.6127
Floor is 2.647727272727273	-1,502.9649
Floor is 0.0	-1,482.7861

However, changing the values to “not relevant” has led to a problem that since “casa”, “chalet” are much bigger and the rent are higher, the model attribute it to the appearing of “not relevant” in those rows, misunderstanding the hidden correlation.

Technical Summary – Model Developing

To have a better model, we used 3 methods to develop it, aiming to decrease the multicollinearity.

1. Ridge regression

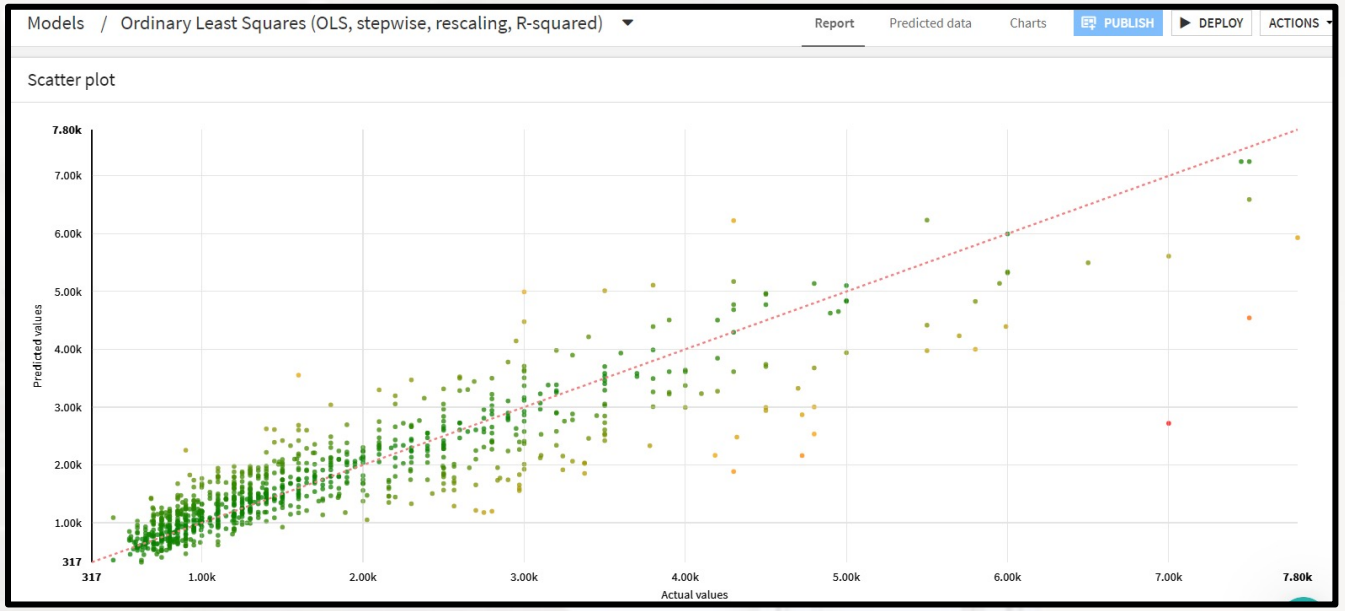
When we did the model for Ridge we got coefficient 0 for duplex, so that variable should be removed from the final model. Ridge therefore has 7 variables.

2. Lasso regression

When we did the Lasso, we got 0 for the duplex this time too, but also for elevator. The LASSO model therefore only has 6 variables.

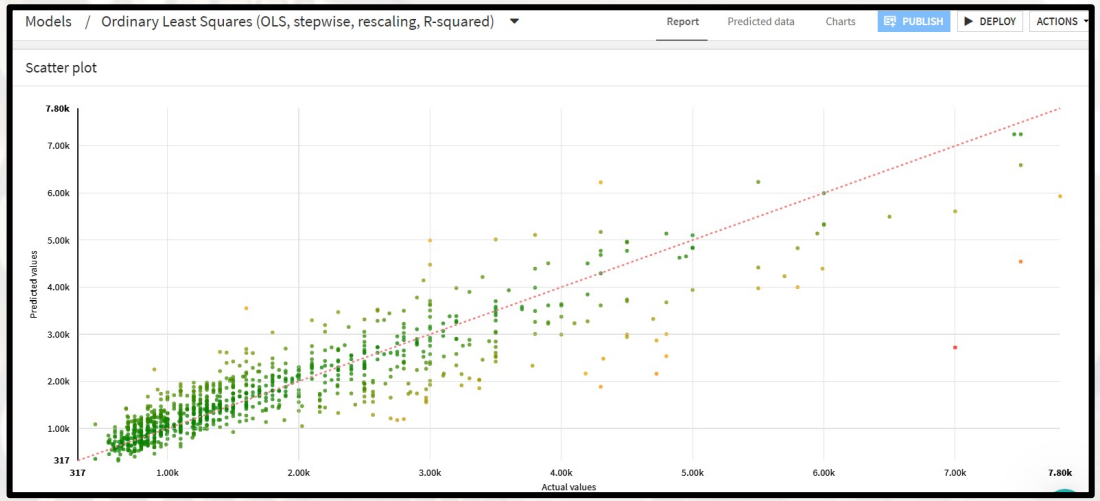
3. OLS

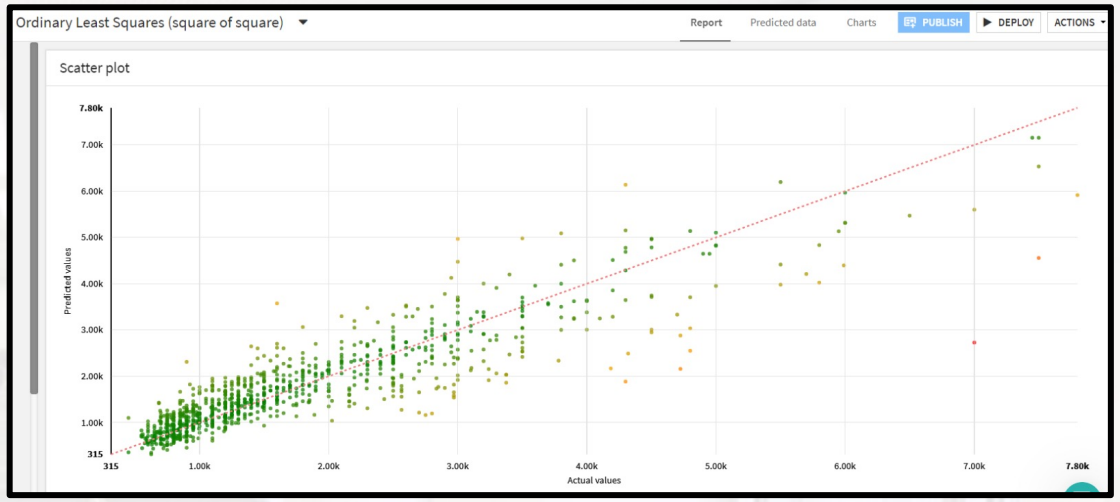
We used the features given by stepwise regression to train the model.

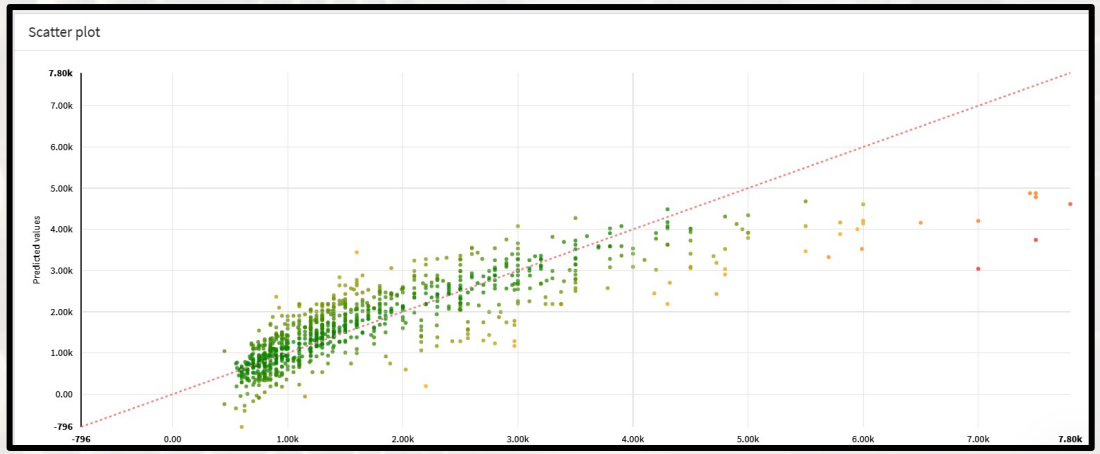


The scatter plot of the 3 models all look similar like this one.

Technical Summary – Model Developing

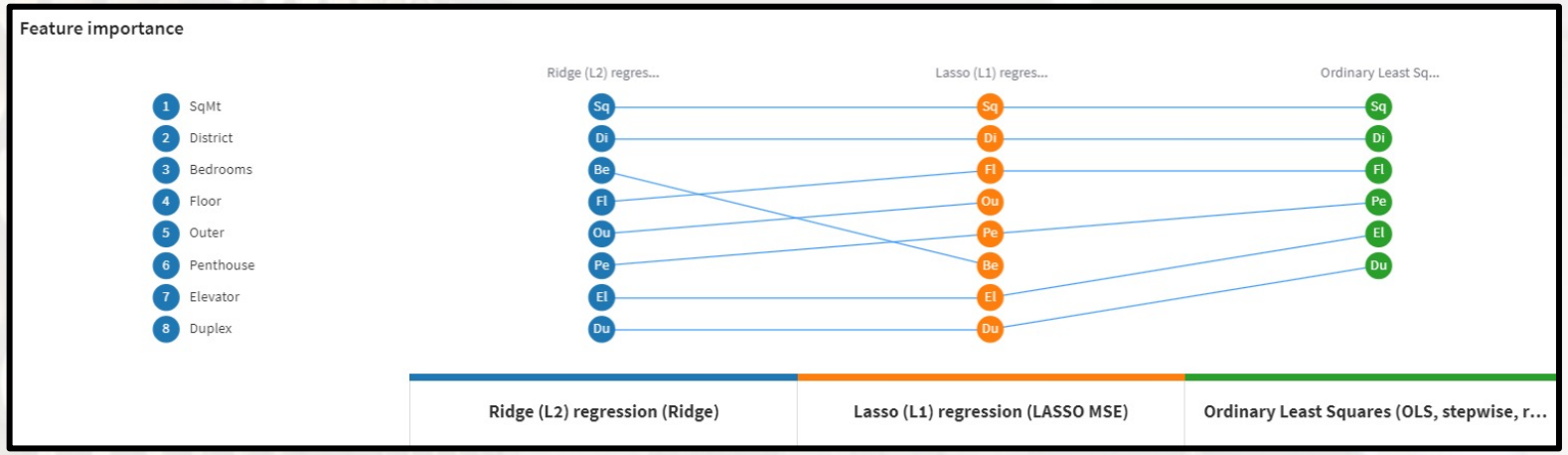






We also included model training with log_square_meters (left), log_rent and log_square_meter (upper) to try making it more accurate when predicting higher values, which did not help much but only decrease the readability.

Technical Summary – Model Selection



Model Evaluation

Name		evaluation	evaluationDataset	model	Performance metrics							
				name	MSE	R2 Score	EVS	MAPE	MAE	RMSE	RMSLE	Correlation
Ridge (L2) regression (Ridge)	OPEN			Ridge (L2) regression (Ri...	**3.07e+5**	**0.777**	0.777	22.2%	371	554	0.272	0.884
Lasso (L1) regression (LASSO ...	OPEN			Lasso (L1) regression (L...	**2.94e+5**	**0.786**	0.786	21.5%	363	542	0.268	0.887
Ordinary Least Squares (OLS, ...	OPEN			Ordinary Least Squares ...	**2.96e+5**	**0.785**	0.785	21.6%	364	544	0.268	0.886

In general, the **LASSO regression model scored the highest**, but selecting a model is not just a pursuit of getting higher scores, we also need to take into consideration the interpretability and other aspects, **for which we chose the OLS model, since it has fewer variables,** making it easier for the business users, and it is clearer at the same time.

Technical Summary – Model Selection

Limitations of the model:

We did not include the casa, chalet and caseron in the training because the sub-population skewed the model. Therefore, the **model works better for apartments** (piso, estudio, ático, dúplex), which is, according to the dataset obtained, the largest market for renting in Madrid. We consider the people that prefer a chalet or casa as another type of customer to target (more towards mortgages).

In addition**, the variance increases as rent goes up**, which is probably because, for larger houses (high square meters predict high price), the price is more affected by the age of the building and the time span from the last renovation.

Contact sheet & Team information

César Alejandro García

IE Sci-Tech – MBD Student

LinkedIn cesargarcia@student.ie.edu

Ignacio Alonso López-Linares

IE Sci-Tech – MBD Student

LinkedIn i.alonso.linares@student.ie.edu

Kangjie Yu

IE Sci-Tech – MBD Student

LinkedIn carlos.yu@student.ie.edu

Morten Aas-Lyngby

IE Sci-Tech – MBD Student

LinkedIn morten.aas.lyngby@student.ie.edu

Lea Brunner

IE Sci-Tech – MBD Student

LinkedIn lea.brunner@student.ie.edu

Thank you for listening and Happy Holidays :D