Madrid, Dec 24th 2018

Master in Business Analytics and Big Data

Machine Learning I

Section O-2

Group H

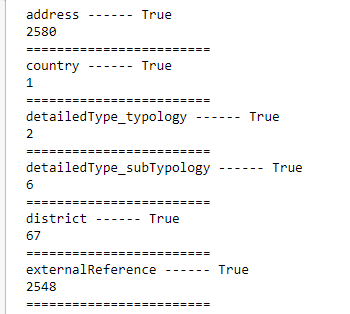
Linear Regression Process Summary

Madrid Real Estate Project

Linear Regression Code Explanation:

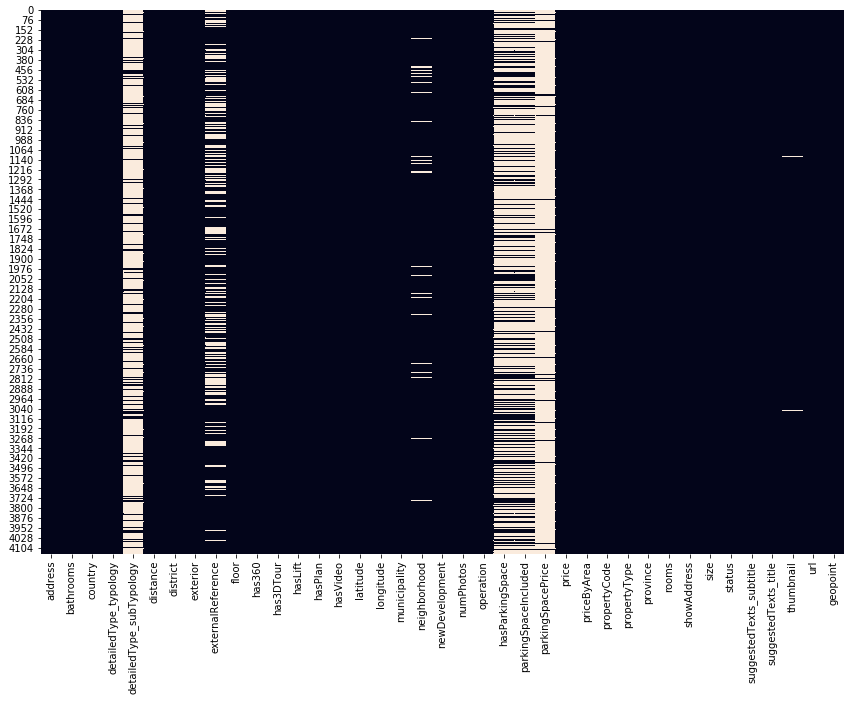
This is a summarized procedure of the attached python code to iterate through linear regression models.

First we loaded the libraries needed for this project and loaded the data we got from the Idealista API. We then checked the dataframe and saw that there were a lot of categorical variables. As those would explode into many dimensions when dummyfied, we decided to create a loop to show us how many unique values were in each category.



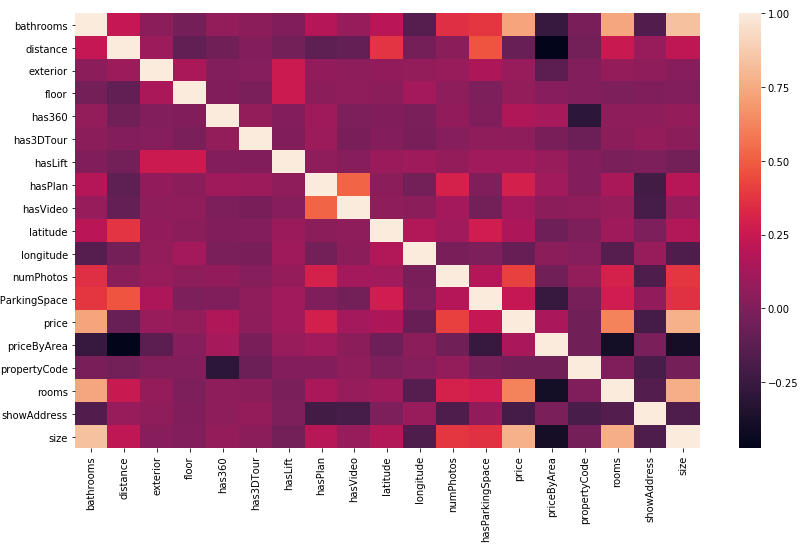
We saw that for some categories like address and external references we could not use them, so we dropped them. Some categories like district we were wary of suing but kept them just in case.

We then looked for null values using a heatmap



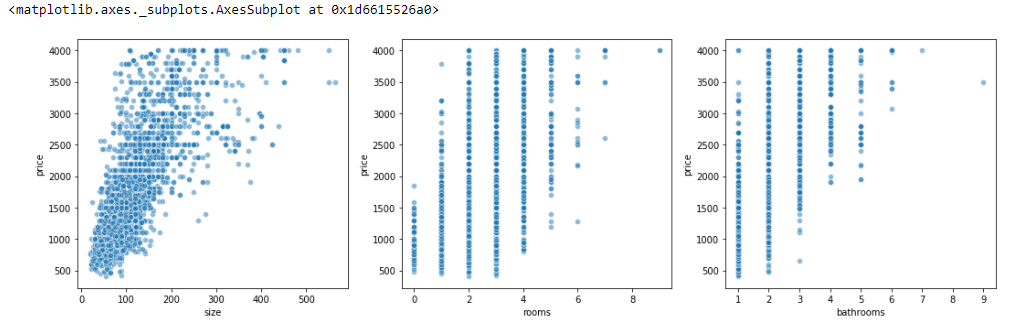
And found that some columns we had to dispose of, because they had too many null values (white in the graph). We then dropped the columns with more than 50% null values, we imputed False on the hasParkingSpace null values and dropped the useless categorical values like Address, externalReference, etc.

For our Exploratory Data Analysis first we made a graph of correlation between numerical variables



Showed us that there is some correlation between price and some variables, and that there is some correlation between bathrooms, rooms and size, which makes sense.

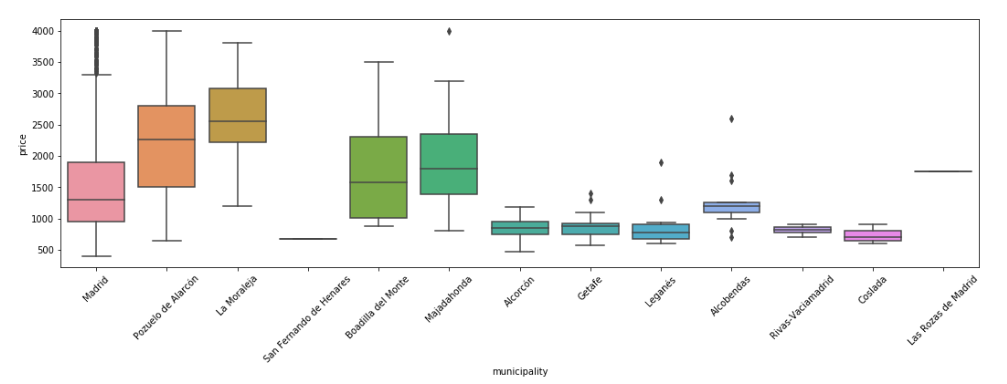
We checked the correlation between the price and those 3 variables.



And we can see that there is indeed a clear positive correlation.

After that we wanted to check if the different districts/municipalities were significant in the difference in price.

And we got the following graph

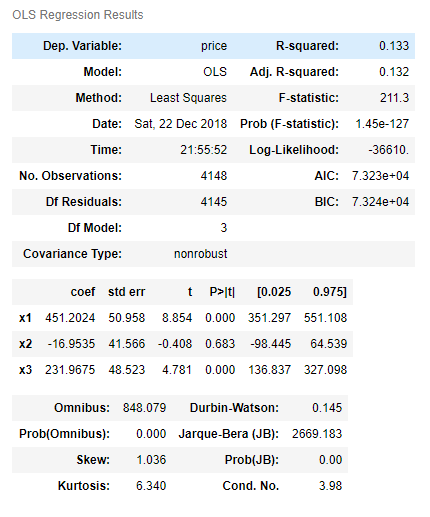


Showing us that indeed there is a difference between prices depending on the location of the property which is also logical. Some neighborhoods command a premium, while others do not.

We also checked against districts and we saw the same effect, where places like Salamanca were more expensive than others.

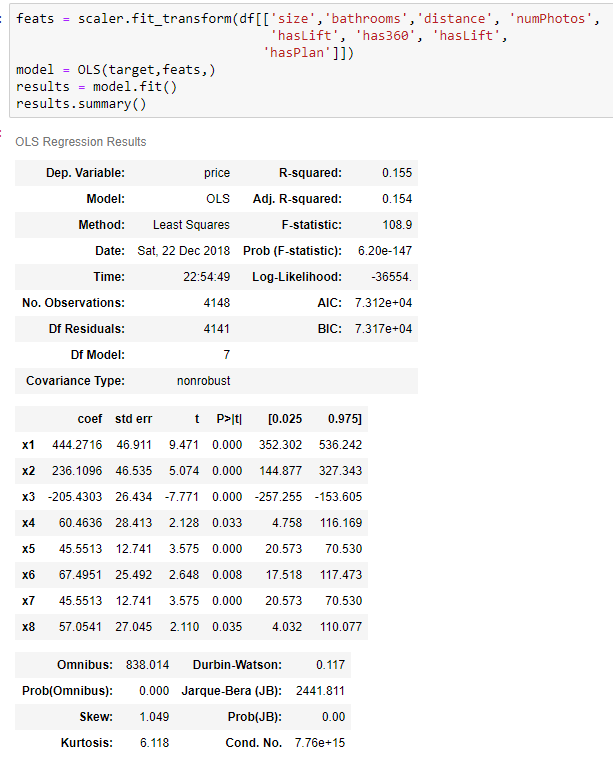
With this information we started building our model using OLS or Linear Regression.

First, we chose bathroom, rooms and size against price and found out



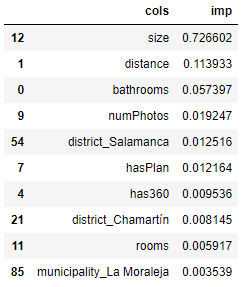
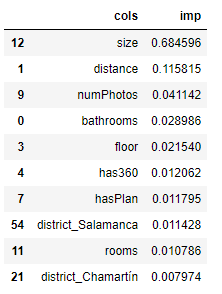
That rooms is not a significant variable in this model. However, we saw that there was a high degree or correlation between rooms and bathrooms so there seems to be some collinearity between them. So, we take rooms away from the model and rerun it a few more times adding more variables.

At the end we got the following report.



Which shows that even though every variable is significant, we are only explaining about 15.5% of the variance in price by the variance in those variables. So, we’re missing a good chunk of information on other variables. The only ones remaining are the categorical ones we saw earlier (municipalities and district).

We then proceeded to check using more advanced models for variable importance. We used Random Forest Regressor and Gradient Boosted Trees and we got the following results



Although size was by far the most important variable to predict price using these algorithms, we saw some importance given to the district where the property is.

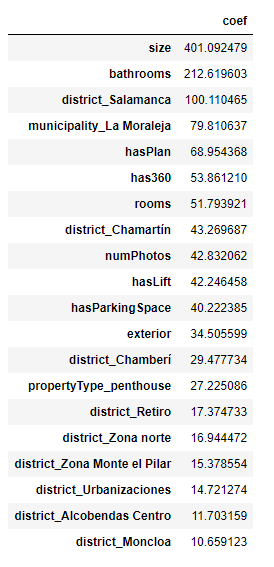
The problem is that at this point we will be introducing 80 new variables in dummy form. And that will create a lot of insignificant variables (neighborhoods that don’t command a premium nor a discount on properties) and reduce the interpretability of the model.

But as we need our model to have as much predicting power and care less about the inference, we decide to go through it. But instead start using the Machine Learning package sklearn to use different linear models like Lasso and Ridge.

As we know from the attached code document, we have a lot of variables that have little significance to our model, we decide to use models with regularization to minimize the effect of those models. And Lasso regression gave us the highest R^2 score of 0.799 with 60 variables with coefficient 0.

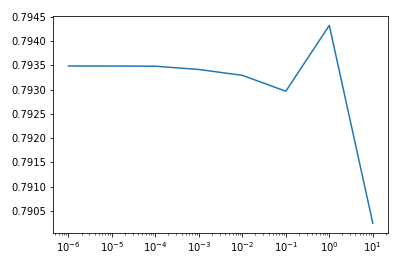
Note: Lasso regression is Linear Regression with L1 regularization. This means that the variables will little significance get minimized closer to and including 0 (L2 regularization won’t make them exactly 0). This works as a sort of automatic feature selection.

This gives the following coefficients:



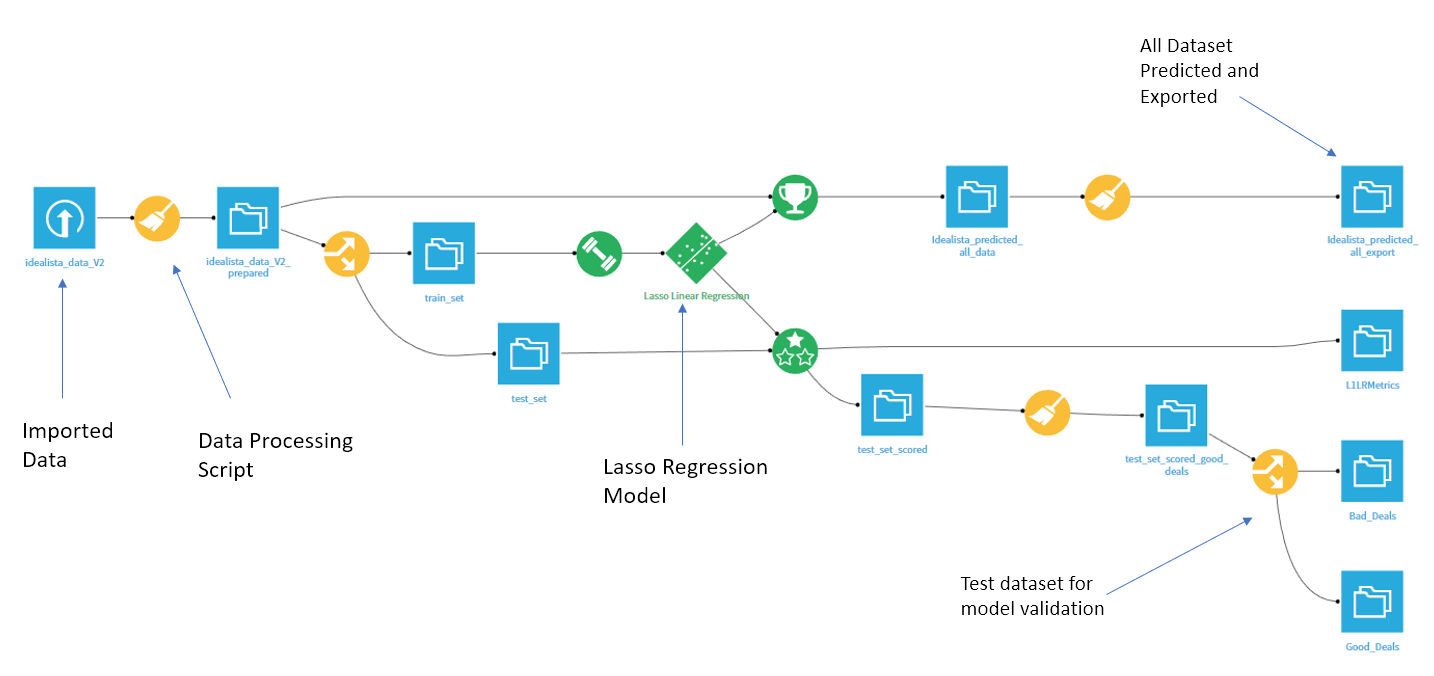
Note that the variables have been standardized, so these coefficients can’t be applied directly to get a price.

We then iterated through various cost coefficients for the Lasso Regression and found that it made minimal change to deviate from 1



So we finally decided to use a Lasso Regression with a cost coefficient of 1 in Dataiku to generate our model and predict the property prices.

The final workflow looks like:



Where we can see the imported and processed dataset, the Model and test set for validation purposes and then the predicted 100% dataset (after model validation) to make the data ready for visualization.

Any new data points can be imported into this workflow and then use the same model to score and the top right recipe to create new features and have it ready for deployment.