Machine Learning I

Asynchronous Session 14 – Linear Regression Practice

Arturo Ford Sosa

The purpose of this practice is to analyze variables of impact on the monthly rent price of flats in the Madrid area. Given the dataset of ~2000 instances of available flats, the process to identify a pattern that could predict prices is as follows, in step order:

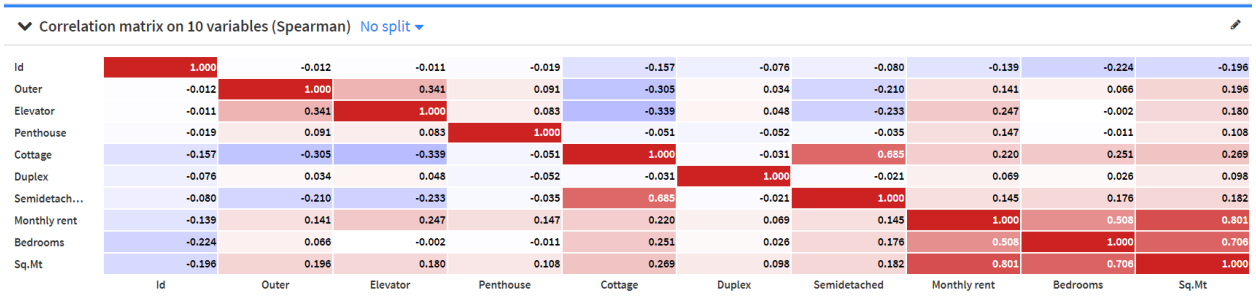
- Dataset was uploaded to Dataiku and cleansed.
  - 'Floor' had invalid values removed (negatives or impossible values such as 43039).
  - Removed outliers in 'Bedrooms', 'Sq.Mt', 'Floor'.
  - Imputed Mean value in 'Outer', 'Bedrooms', 'Elevator', 'Floor' where null values were present in order to make them less significant and avoid their full removal. The logic is to not affect other fields that correlate with null values in these fields, such as 'Cottage'.
- Cleansed dataset was run through the Stepwise Linear Regression Algorithm:

```
Add  Sq.Mt                with p-value 0.0
Add  Cottage              with p-value 5.95322e-29
Add  Floor                with p-value 5.95964e-10
Add  Bedrooms             with p-value 6.16181e-06
Add  Duplex               with p-value 0.000898983
Add  Elevator             with p-value 0.00122696
Add  Penthouse            with p-value 0.0110957
```

  The fields 'Id', 'District', 'Address', 'Number', 'Area', 'Monthly rent' were removed from the algorithm because they are text-based. 'Id' is the identifier, so it was removed, and 'Monthly rent' has been removed as well as it is the target variable. All the rest of the variables were included, and as seen above, they are all significant (p-value < 0.5). The resulting features are as follows:

```
resulting features:
['Sq.Mt', 'Cottage', 'Floor', 'Bedrooms', 'Duplex', 'Elevator', 'Penthouse']
```

- Ran all variables through a correlation matrix to estimate multicollinearity:

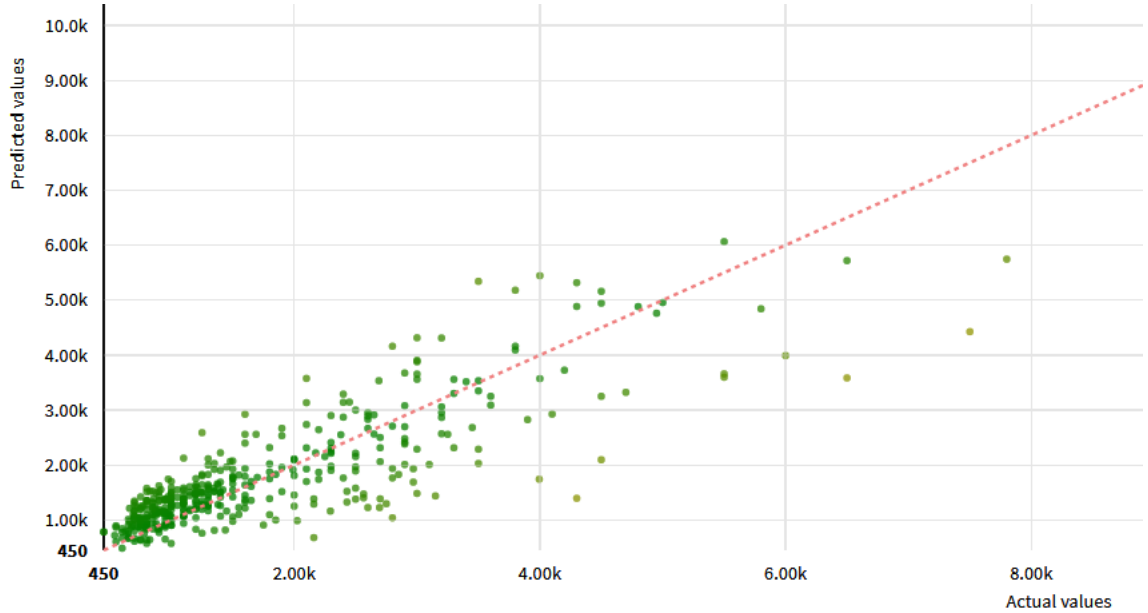


  Notable correlations include 'Cottage' with 'Semidetached', as well as 'Bedrooms' with 'Sq.Mt'. Therefore, to avoid multicollinearity, 'Bedrooms' was removed. Since 'Semidetached' was recognized as non-significant by Stepwise, it was never included so 'Cottage' stayed in.
- Ran the remaining, cleansed variables through an Ordinary Least Squares regression, where the resulting Coefficients were as follows:

Variable	Coefficient
Cottage	-1,301.8445
Duplex	-252.2748
Penthouse	228.9625
Elevator	202.4575
Outer	-62.3511
Floor	29.4078
Sq.Mt	12.8344
Intercept	100.1944

Conclusions that can be drawn from Coefficients:

- All else equal, an apartment being a 'Cottage' will be cheaper than the rest. This can be misleading, as cottages are usually more expensive because they are larger, so someone untrained to understand these numbers might assume that cottages are cheaper on average. However, they will only be cheaper in comparison if the flats compared to them are of the same size and characteristics. Happens similarly with duplexes.
- Every additional square meter will result in roughly 12.8 more Euro of monthly rent, all else equal. This makes sense in general in real estate, where a larger piece of estate will trend to increment linearly for every additional square meter.
- Interestingly, features such as 'Outer' and 'Elevator' go in different directions even though one would expect them to go in similar ways. I would suggest evaluating again with a larger sample size to see if the trends stick.



The above chart illustrates the regression found with these fields, and where every flat tested lies. The trend seems to be towards the lower end of the line, with most flats having competitive prices to one another.

As for the sub-goal of this analysis, these are all the districts in the dataset as well as their average prices and predictions:

	Avg of Monthly rent	Avg of prediction
Arganzuela	1132	1285
Barajas	949	1477
Carabanchel	772	1192
Centro	1924	1464
Chamartín	2198	1995
Chamberí	2257	2088
Ciudad Lineal	1327	1620
Fuencarral	1423	1938
Hortaleza	2312	2484
Latina	915	1356
Moncloa	2010	2136
Moratalaz	1117	1845
Puente Vallecas	782	1093
Retiro	2380	2007
Salamanca	2437	2141
San Blás	1128	1652
Tetuán	1353	1519
Usera	876	1183
Vicálvaro	707	1138
Villa de Vallecas	880	1301

Based on this data, we can estimate the top 3 districts which are over/under valued according to our regression results:

- Overvalued:
  - Centro
  - Retiro
  - Salamanca
- Undervalued:
  - Moratalaz
  - Barajas
  - San Blás