Brazil Rent Discussion

The Brazil Houses to Rent dataset includes multiple features such as; bathrooms number, floor number, amount of rooms, taxes, etc. and shows the total cost of renting. This cost is the simple addition of the following features; housing association tax (hoa), rent amount, property tax, and fire insurance.

But could an ML model predict the total amount even without the rent amount, which was shown to be the largest contributor to the total amount?

Souce: https://www.kaggle.com/rubenssjr/brasilian-houses-to-rent

Data Analysis

Sao Paulo had considerably more records than any of the other 4 cities, in fact 55% of the records came from Sau Paulo. Furthermore, most renting prices were at the very low levels, around 2000.

Coefficients were shown amongst each other, but when plotting extremely high rent amounts, 7 very large outliers were found and removed. The new coefficients shown varied significantly from the previous ones.

It was found that the rent amount and fire insurance held an almost perfect correlation with the total cost. This makes sense for the rent amount, but what about fire insurance? It was shown that fire insurance is essentially a perfect predictor of rent amount – it must be calculated based on a percentage of the rent amount. As such, neither of these should be included in the final model.

Models

Data cleaning: The floor area had a “-“ symbol which was converted to 0 as it followed the data trend. The city column had to be one-hot encoded with the first column being dropped to prevent collinearity. And lastly, the animal and furniture columns were simply set to either 1 or 0.

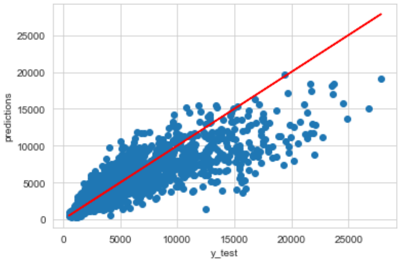
Four models were created:

The first custom sequential tensorflow model was used as a baseline. Its MAE was 1438, but with a mean total rent of 5198, this would represent quite a difference when predicting each total rent. The third model using LinearRegression, performed very similarly – both did not include the total rent or the fire insurance.

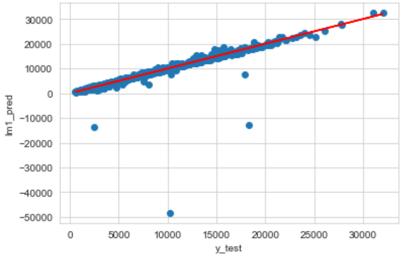
The second model, also using LinearRegression, included the fire insurance for comparison – as you can see from the graphs, it performed significantly better and only had an MAE of 260.

Unfortunately, its RMSE was 1533, which represented some very extreme outliers – in practice these would not be a problem as most of the outliers are somehow negative!

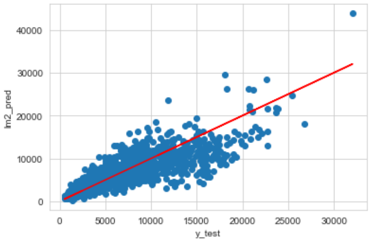
1 Sequential base



2 LinearRegression(fire inc.)



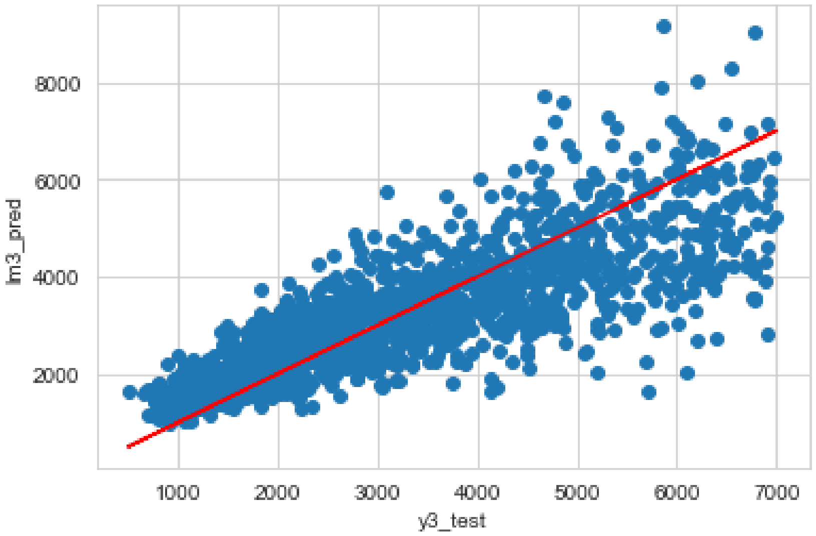
3 LinearRegression



Further Improvements – Model 4

The model could be improved by setting a cut-off at the higher total end, so that we only focus on the majority of records which fall between the ranges 0-7,000. There were only 2560 above this amount whilst 8125 records within – the upper range was 7,001 to 46,335, with a mean of 11,873.

As expected, this model had a lower absolute error of 683, although its explained variance score was lower at 0.67. On the other hand, the mean of this subset was 3094, meaning the prediction is only 5.7% better than the previous model. (1438/5198 = 27.7% [base mode] ---- 683/3094 = 22% [subset model])


lm3_pred
y3_test