HousePrices - Machine Learning applied to Housing Data

Andrew P. McMahon[1]

[1] *Department of Physics, Imperial College London*

(Dated: October 7, 2016)

Here I present a quick attempt at performing a regression analysis on UK land registry data in order to predict house prices.

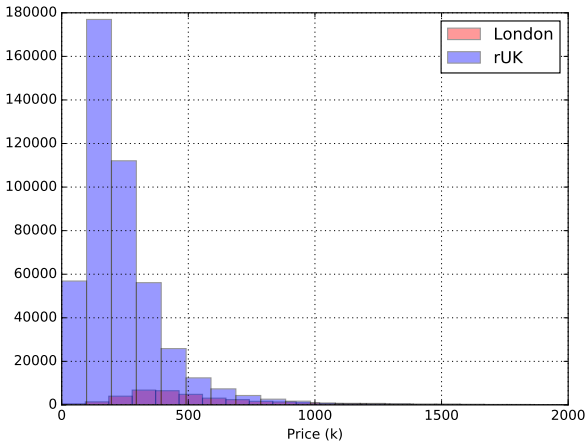

FIG. 1. Prices of houses within and without London. The data is heavily biased towards houses outside of London, however the two distributions do have different modes and seemingly different tails to their distributions, with London prices have a mode at higher prices. Only data for house prices up to £2 million are shown.

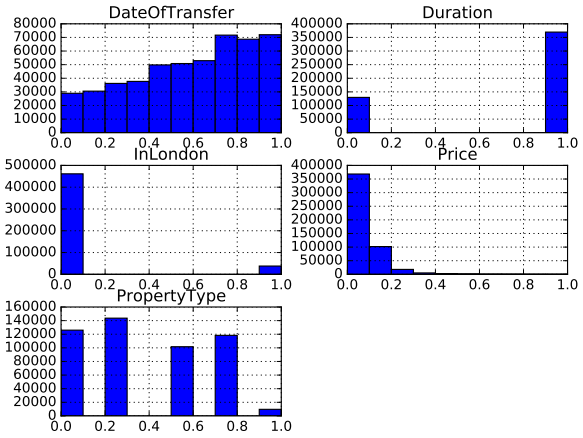

FIG. 2. Histograms of some features of the data set.

I. UK LAND REGISTRY DATA

In order to facilitate a quick analysis, only data from 2015 in the UK land registry was analysed.

II. DATA PREPARATION

The tail of the price distribution is extremely large, and looking at boxplots of the data there are a very large number of outliers. In order to make the data more amenable to analysis some outliers are removed by dropping all data where the price is greater than 3 standard deviations larger than the mean. Since the data is heavily skewed to smaller prices I did not remove values less than 3 standard deviations from the mean. Even upon this trimming there are still many outliers as can be seen in Figure 3.

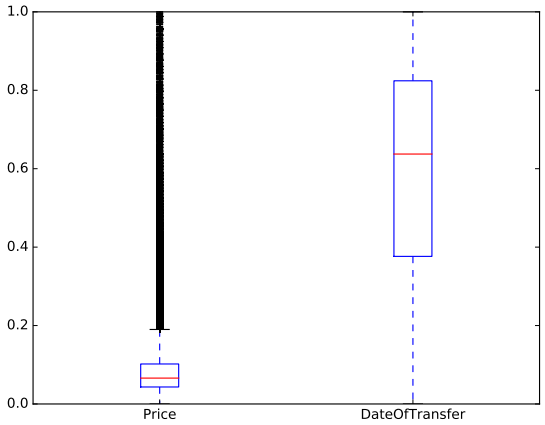

FIG. 3. Boxplots for Price and DateOfTransfer for the houses post trimming of values greater that 3 standard deviations from the mean of the price.

III. FEATURE SELECTION

Three features are suggested in the given document: lease duration, whether or not the property is in London and the property type. I have decided to see if I can find a better three features based on a preliminary linear regression with those three features plus the date of transfer for the property deeds. I will then choose the three features which have the strongest correlation from this simple regression. The 4-feature model returned the coefficients shown in Figure **??**.

Selecting the three most strongly correlated features

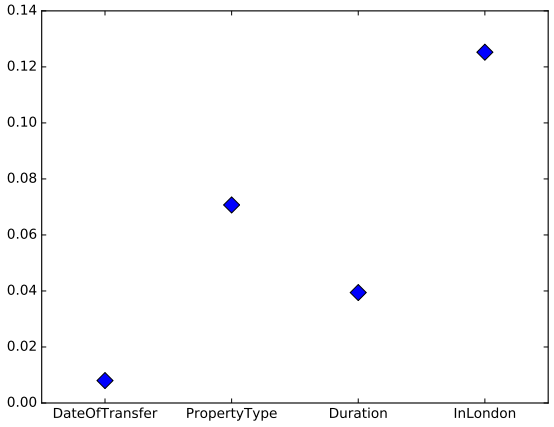

FIG. 4. Coefficients for the linear regression model based on the 4 features listed.

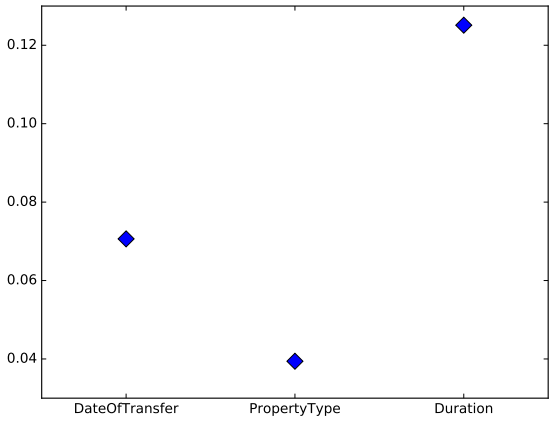

FIG. 5. Coefficients for the linear regression model based on the 2 features listed.

from this (three feature models at most were asked for in the document), we should disregard the date of transfer and focus on the remaining three.

It is clear from Figure **??** that the variables "InLondon" and "DateOfTransfer" are the most strongly correlated with the final price. The model is therefore rerun with only these as features.

IV. MODEL AND RESULTS

Due to time constraints only linear regression was performed, this is useful as a first attempt, as the coefficients of the linear regression will allow us to see which features are most strongly correlated with the house price. I also attempted Gaussian Process based regression on this smaller data set, however my laptop ran out of RAM.

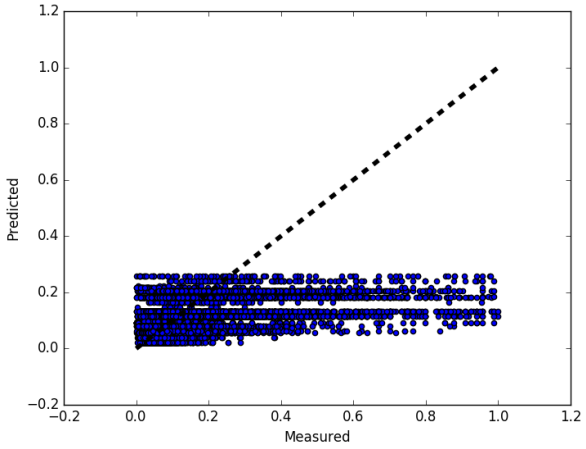

FIG. 6. Prediction versus measured test data for the price based on the 3 feature model.

Gaussian Processes is quite intensive and will scale poorly with data size, if more time were available approximation methods could have been employed to aid the scaling behaviour. More computing resources would also have helped.

Another concern is that the features selected are mainly categorical, so a linear regression is not the best algorithm. Gaussian Processes and other kernel based methods would have been able to cope with this complication, however as mentioned this was not attempted in detail due to time constraints.

In the 4-feature model (see Figure 4, the linear regression returns a coefficient of determination $R^2 = 0.231586381204$ on the training data set, which is relatively poor. This does not change when Lasso or Ridge regression, which both contain regularisation parameters, are employed.

In the 3-feature model (see Figure 5, the linear regression returns a coefficient of determination $R^2 = 0.230821416905$ on the training data set, which is again relatively poor but still comparable with the 4-feature model. This does not change when Lasso or Ridge regression, which both contain regularisation parameters, are employed.

Figure 6 shows the distribution of predicted prices vs the test data set price values, we can see that the model is indeed quite poor.