Model, prediction, and Evaluation: Land Registry Data and House Prices from 2016

Sofia Dunlosky[1]

Abstract

Multiple (increasing number of) Random forest classifiers are applied to the `HM Land Registry Price Paid Data' data-set to predict the price of houses from 2016 and onward. The additional data and algorithms should be able to keep improving the performance of the prediction. It is reasonable to expect that an accurate prediction will be generated.

Algorithm

The algorithm can be divided into two parts: the **Data generator** which loads data from csv file 5000 rows /time and generates matrices 250000 row /unit. One hot encoding is used for the categorical columns to emphasis the non-numerical property of the database. **Main training algorithm**: A Random Forest Algorithm with 10 trees are generated for each data unit. The model is integrated after training with increasing number of trees. The algorithm is selected because it is efficient in dealing with categorical data. The results of trees are combined together if the additional trees are proved to be meaningful for improving (reducing mean square error) the result in cross-validation. The training terminates at the point where the performance is not improving in 5 successive data branches.

Result

The result does not improving at all. The mean square error is 3.65 for the first branch and the predictions are ridiculously far away from the correct answers. The models and algorithms are enclosed in the zip file.

Further improvement

Starting from the most recent data (2015) might provide quicker improvement in accuracy, different algorithms and hyperparameter might be beneficial as well.

[1] University of Edinburgh, email address: sofia.dunlosky@gmail.com, cell: +44 7419840233