Consumer Data Research Centre

An ESRC Data Investment

Knight Frank

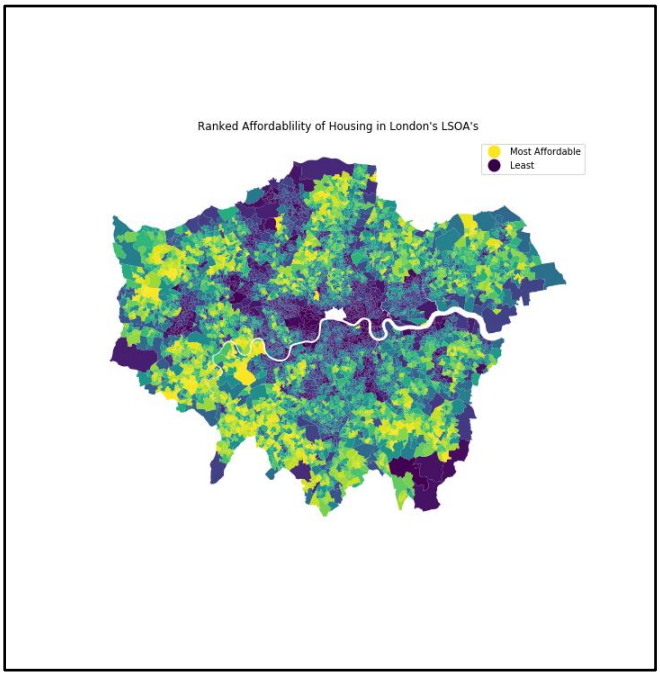
Creating & Applying a Multivariate Index of London's Housing Market using Machine Learning

James Pryor [1], Naru Shiode [1], Faisal Durrani[2] and James Culley[2]
[1]King's College London, [2] Knight Frank

Project Background

London's housing market regularly ranks as one of the least affordable markets globally; curating a dataset that represents the characteristics of small-scale areas of the capital can help us to assess what factors drive these high-levels of unaffordability, and with this information, evaluate and match these areas to various buyer profiles seeking to live in the city.

Such an index can also be used in predicting the growth potential of property values/ rents, given any planned or theoretical modifications to an area, for example, the creation (or destruction) of greenspaces, or the expansion of the city's transportation network.



Data and Methods

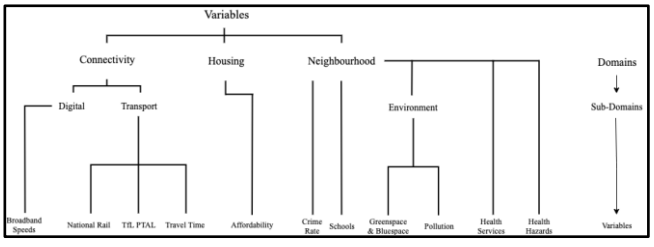
To measure London's property market, this study employed 13 variables relating to the infrastructure, environment and affordability of the 4,829 LSOA's in the city (excluding the City of London) (Figure 1). For these variables, each was ranked and standardised to allow for cumulating into an overall score and domain-specific scores. The index (~125,000 data points) was then used to produce graphs for each of the standardised variables (see Figure 2 for affordability), to be used in a raster analysis – whereby the requirement for each layer (affordability, travel time etc.) were modified to reflect the requirements of specific types of buyer (e.g. first-time, downsizers) and suggest markets for these buyers.

For the final part of this project, a Random Forest Regressor was built to predict the quantitative change to a property's value, given a change in one of the 13 variables. Random Forest was chosen as it corrects the tendency for decision trees to overfit the model on their training set by iterating over samples of the training data.

Key Findings

The index provided a huge insight into the geography of the target markets for different buyer-groups, highlighting areas that represent the demands of each profile. For instance, first-time buyers are directed towards the periphery of Zones 2 & 3 on the tube map in South-Eastern postcodes – where accessibility to the urban core remains strong, but house prices lower than their Northern counterparts. By contrast, for buyers that have typically progressed their careers – and now with substantial incomes – West London (Kensington, Westminster and Paddington), up through to Hampstead Heath are the most attractive areas given the quality of education, access to greenspace and commutability to the various business districts.

From the results of the machine learning, it is evident that connectivity through transportation commands a great control over the value of properties, as does greater access to (larger) greenspaces and above average performing schools. Adversely, a deteriorating physical environment (e.g. more pollution) has an opposite outcome. For variables such as broadband speed, it was found these have not yet fed-in to the price making of a property but remain particularly relevant to the development of commercial and mixed-use buildings.



Value of the Research

The output from this study represents an ongoing stride in the data-isation of the real estate sector and can be used to identify target markets for both buyers and developers. Furthermore, the application of Machine Learning methods has provided quantitative insights as to the benefits (or costs) of modifications to an area that can be utilised by policy makers, market speculators or homeowners/ buyers.