# Modelling Relationships Between Area Characteristics and Property Value

## Proposal

This project aims to develop models using ML approaches to predict how a given property with a known price may change in value as location variables change. For example if a school is built in my area, how will it impact the price of my house. Current approaches to predicting house prices focus on predicting the price of the house given some facts about it and its surrounding area, by focusing solely on the area data it should be possible to provide better predictions of house price change. This information can then be used to advise local planners. Furthermore, by only using area data we can use a larger data set which contains non-residential properties, though the finished models may be separate for different property types.

Another drawback of most approaches is that they use average sales for an area to form predictions, however there exists publicly available data sets that make it possible to create training samples based on individual sales. The land registry contains publicly available data describing every property sale going back to 1995, this would form the backbone of any data set generated during this project. Location data would be pulled from a variety of sources and will be in different formats that would need to be adapted. Such sources include census data, government websites and map services such as Google.

Technical challenges include scraping locational data from public sources, e.g. locations of amenities, and mapping each property sales distance to each amenity type. Not only must this be done accurately but also in a scalable manner given the potential size of the dataset, possibly by generating quadtrees for each amenity to be mapped.

Following the creation of the dataset the machine learning models can be applied, it is likely that linear regression will be the best model to use, though other supervised learning models should be trialled and if they are found to have reasonable accuracy then ensemble modelling can be applied.

A test set can be generated by extracting from the dataset all properties that appear at least twice, having been sold on multiple occasions since the start date of the data set. The models can then be queried with the test sets first sale prices alongside the initial and later location details, the result can then be compared with the later sale prices. This approach has some drawbacks as it primarily relies on there being a large number of resold properties with non-negligible changes in their local area over the same time frame. Furthermore, it may be skewed by a disproportionate number of certain scenarios, e.g. houses that have been renovated immediately after purchase and then sold again may be common in the test set. As such it will be necessary to further investigate how such cases can be removed and if there is another approach to testing that can be taken.

**Deliverables**

- A main data set used to generate models.
- Report on the most successful models for generating predictions from the dataset.
- Software to select a model from a given list, train it and then run queries against it.
- Software for extending the data set, e.g. by automatically adding new amenity distances using Google maps, or by taking an input file of areas and data about each area.
- Software for viewing the data in a user friendly form that highlights trends.