Mei Tan, lt528@live.mdx.ac.uk

Property prices prediction

i) Problem definition and goals

Property values forecasting is of great interest of various real estate stakeholders such as investors, agents, house owners and buyers in the market due to the outcome would help them to make more informed decision. The goal of this assignment is to forecast the property values in London. A model with hypothesises will be built and it will learn from the existing dataset by teaching itself to refine its parameters and make data driven forecasting.

The challenges of this prediction is lack of census data which can be jointly used to measure the correlation relationship with the existing properties and point of interest(POI) dataset. Census data holds the information of population, diversity and the characteristics of neighbourhoods which would significantly affecting the price of the properties in a specific location. Prediction will be more accurate with the help of census data.

ii) Hypothesis

The assumption or hypothesis for this assignment is: H1, the price of property in a given location is rely on the distance in between the property to the attribute of high level category of Point of Interest (POI). The closest or shorter the distance between the property to the high level category of POI the higher the price of the property which is an inversely proportional relationship. Neighbourhoods with more high level category of POI like food, induction and transportation will be relatively higher population in the location. It is likely that they have higher purchasing power. Thus, the house in the location will cost higher due to increase in demand.

Abovementioned hypothesises are falsifiable as there can be proven based on the existing dataset of London properties *<id, price, size, bedrooms, type, built_date, historical_building, location & lan,lon>* and also data set of Point of Interest *<location, lat, lon, category, high_level_category>*. Prediction problem can be defined based on these measurements of property's price and distance between two points by building a model to test the hypothesis. If the prediction problem shows a good fit, hypothesis will be true otherwise will show false.

iii) Data Processing

Dataset should be process for a better sense of understanding. But due to the two datasets, London properties and Point of Interest are not accessible at the moment, data cleaning stage will be impossible to be apply.

However, metrics of the hypothesis will still be tackle by calculating the *price per square meters (price per m²)* via division of price over size of each type of the property in each location. In addition, the *distance (dis)* between the property and the high level category POI will be measured based on the data of latitude and longitude of the geographical coordinates. Later, original dataset will be reshaped to a final format by undergoing data wrangling which including the new metrics as below and table 1:

< price per m², type, built_date, historic_build, location, dis_food, dis_instruction, dis_transportation>.

price_per m²	type	built_date	historic_build	location	dis_food	dis_instruction	dis_transportation
2359	flat	2000	FALSE	Belmonth	1.3	1.55	2.83
3010	detached house	2010	FALSE	Brent Cross	1.01	1.08	1.2
5998	detached house	1949	TRUE	Brent Cross	0.23	0.38	0.1
4027	semi-detached	2005	FALSE	Greenwich	0.88	0.5	0.69

Table 1. Final format of dataset.

iv) Algorithms

In general, regression is a machine learning tool which assists in making forecasting by learning from current statistical data. This machine learning tool study the relationships between the target parameter (price per m²) which is the output and a set of other parameters as the input and then implement this function of relationship on to the real observed data in future for prediction. As stated above definition, a house's value will be dependent on parameters such as location, built date, size of property and distance from high level categories. This machine learning will be able to calculate the estimated prices of property in a given geographical area upon we apply the parameters.

v) Data Understanding

Investigation have to be carry out from the output upon we run the regression model. This is to investigate if the output show prediction power features to forecast the price of house. Data shown evidence that hypothesis is true if prediction power is high. Otherwise, new hypothesis may need to be formulated if marginal prediction power, this is to improve the existing hypothesis. Else if the hypothesis tested is false, new investigation have to be carry out.

There are some limitations of this model, factors which may need to be consider like new industrial or residential project in the location may affect the price of house. Other than that, bias of the dataset might happened for Point of interest. These data were uploaded by social media user based on the popularity of the check-in point or location.