LONDON HOUSE PRICES

1. Introduction

In 1995, London was just coming out of the recession-before-last. The average home costed £79,000. Today prices have jumped almost 600% to an average £514,097. Wages have failed to keep pace with this leap. In 1995 the average Londoner earned £22,487, compared to an average £48,023 today, a 113% increase. Which means property prices have risen more than 5 times as fast as incomes, locking many Londoners off the property ladder.

In 1995 just 15 per cent of Londoners rented — it was what you did for a few years before you bought a flat. Now, around one in three of the capital's residents belong to "Generation Rent" and many believe they will never own a home. In 1995 an average first-time buyer borrowed £55,575 and earned £21,575. Today's first timers earn an average £51,000 and borrow £174,000. Their average deposit is £96,000, which is a tenfold increase on 20 years.

Although house prices in London rose well above the general inflation, see Figure 1, the growth rate of the UK economy and average income, house prices have "crashed" twice in the last 20 years, between 1990 and 1992, and more recently between 2007 and 2010. With the recent uncertainty looming over British and London economy due to Brexit, how are the house prices in London changing? In one hand there is a slowdown in investments and fears that a great number of lucrative jobs will eventually have to leave London for greener pastures elsewhere in Europe. However, the weakness of the pound post Brexit has made UK property more attractive to foreign investors however, which is likely to increase demand from abroad. In addition, there are several initiatives the government has taken to slow down the rise in property prices. Multiple stamp duty changes have occurred in recent years, affecting the property market as a whole. For example an additional 3% stamp duty surcharge was introduced in April 2016 for additional residential properties, and this change has affected the prime London market in particular.

The purpose of this exercise is to build an estimation engine to guide investment decisions in London house market. You will first build machine learning algorithms (and tune them) to estimate the house prices given variety of information about each property. Then, using your algorithm, you will choose 200 houses to invest out of (about) 2000 houses on the market for sale now.

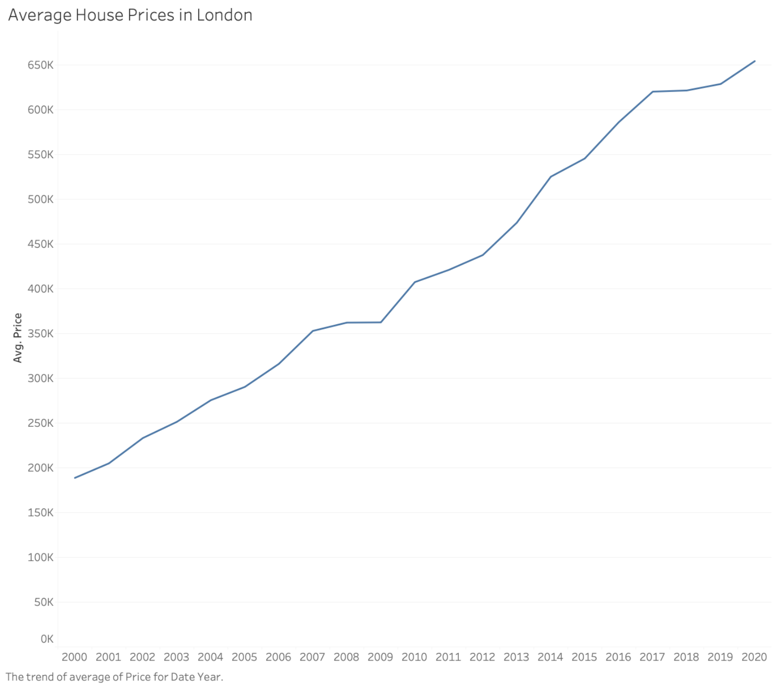


Figure 1: Average London House prices over the years

2. Data sources

Publicly available (https://www.gov.uk/government/collections/price-paid-data) HM

Land Registry's Price Paid Data tracks the property sales in England and Wales that are submitted to them for registration. The sales data is available from 1995. See Section 3 below for the data dictionary and details of the types of properties included in the data set. We only consider the data for London.

We merge this data set with (publicly available) Energy Performance Certificate (EPC) data. This database contains more information about each property including size, number of bedrooms, and the energy ratings. However, this data set only runs from 2008 to 2020 and does not contain all the properties in the HM Land Registry's database. Finally, we merged the resulting data set with data about each post code.

Third, we added the public transport information to the dataset. You can find the nearest station, walking distance to this station, and the number of lines passes through this station, for each property.

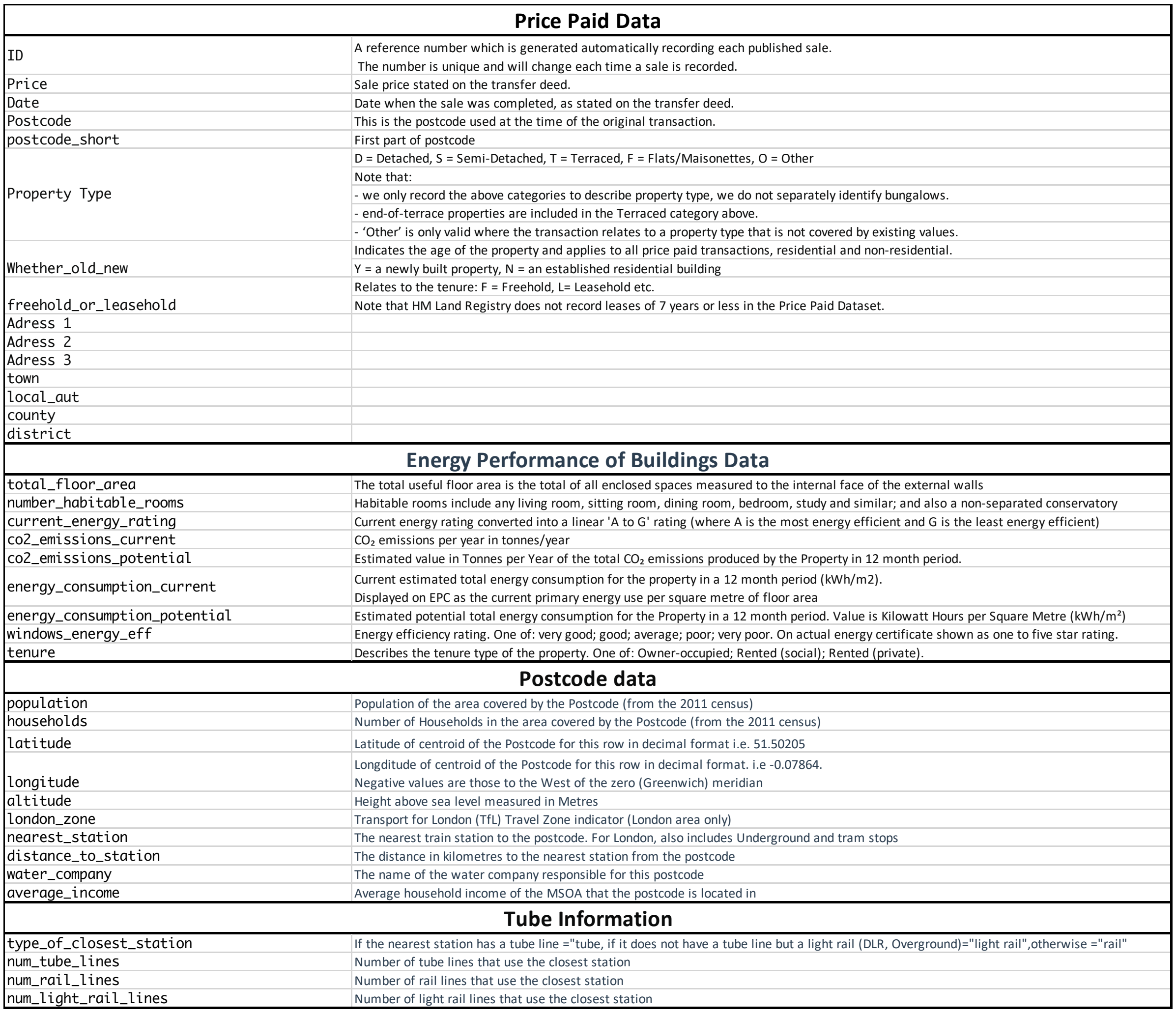
Data Cleaning: I cleaned the data to the best of my ability. If you see any potential problems, feel free to filter out the problematic data.

3. Data Dictionary

You are given two data sets: training and (out of sample) testing. In the training data you have access to the price paid of each property. In testing data, you only see the asking price but not the final price paid.

This data is a little richer than one provided for the visualization assignment, but you only have access to information about 14,000 properties sold during 2019. Please find the data below or use the excel file in this link.

Please note that the following information is missing from the out of sample test data: , *date, price, all columns between (including) address1: county*



Price Paid Data	
ID	A reference number which is generated automatically recording each published sale. The number is unique and will change each time a sale is recorded.
Price	Sale price stated on the transfer deed.
Date	Date when the sale was completed, as stated on the transfer deed.
Postcode	This is the postcode used at the time of the original transaction.
postcode_short	First part of postcode
Property Type	D = Detached, S = Semi-Detached, T = Terraced, F = Flats/Maisonettes, O = Other Note that: - we only record the above categories to describe property type, we do not separately identify bungalows. - end-of-terrace properties are included in the Terraced category above. - 'Other' is only valid where the transaction relates to a property type that is not covered by existing values.
Whether_old_new	Indicates the age of the property and applies to all price paid transactions, residential and non-residential. Y = a newly built property, N = an established residential building
freehold_or_leasehold	Relates to the tenure: F = Freehold, L= Leasehold etc. Note that HM Land Registry does not record leases of 7 years or less in the Price Paid Dataset.
Adress 1	
Adress 2	
Adress 3	
town	
local_aut	
county	
district	
Energy Performance of Buildings Data	
total_floor_area	The total useful floor area is the total of all enclosed spaces measured to the internal face of the external walls
number_habitable_rooms	Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar; and also a non-separated conservatory
current_energy_rating	Current energy rating converted into a linear 'A to G' rating (where A is the most energy efficient and G is the least energy efficient)
co2_emissions_current	CO_2 emissions per year in tonnes/year
co2_emissions_potential	Estimated value in Tonnes per Year of the total CO_2 emissions produced by the Property in 12 month period.
energy_consumption_current	Current estimated total energy consumption for the property in a 12 month period (kWh/m2). Displayed on EPC as the current primary energy use per square metre of floor area
energy_consumption_potential	Estimated potential total energy consumption for the Property in a 12 month period. Value is Kilowatt Hours per Square Metre (kWh/m^2)
windows_energy_eff	Energy efficiency rating. One of: very good; good; average; poor; very poor. On actual energy certificate shown as one to five star rating.
tenure	Describes the tenure type of the property. One of: Owner-occupied; Rented (social); Rented (private).
Postcode data	
population	Population of the area covered by the Postcode (from the 2011 census)
households	Number of Households in the area covered by the Postcode (from the 2011 census)
latitude	Latitude of centroid of the Postcode for this row in decimal format i.e. 51.50205
longitude	Longditude of centroid of the Postcode for this row in decimal format. i.e -0.07864. Negative values are those to the West of the zero (Greenwich) meridian
altitude	Height above sea level measured in Metres
london_zone	Transport for London (TfL) Travel Zone indicator (London area only)
nearest_station	The nearest train station to the postcode. For London, also includes Underground and tram stops
distance_to_station	The distance in kilometres to the nearest station from the postcode
water_company	The name of the water company responsible for this postcode
average_income	Average household income of the MSOA that the postcode is located in
Tube Information	
type_of_closest_station	If the nearest station has a tube line ="tube, if it does not have a tube line but a light rail (DLR, Overground)="light rail",otherwise ="rail"
num_tube_lines	Number of tube lines that use the closest station
num_rail_lines	Number of rail lines that use the closest station
num_light_rail_lines	Number of light rail lines that use the closest station

Figure 2: Data Dictionary

4. Assignment

You can download the data sets and the rmd file here

i) Training Data

ii) Out of sample testing Data

iii) You also have access to an initial rmd file to help you get started

You should first build machine learning algorithms using algorithms we covered in AM04 and tune them (try at least 4 algorithms) and *ensemble* them with stacking. Then pick the one that performs best.

Next, use the best algorithm to price the houses in the (out of sample) testing data and pick 200 of them that you think are good investments. Here assume that asking price is the price you will have to pay to buy each property. To measure your performance, I will compare the asking price and the actual price (which is not available to you, but I have access to) of each house you pick and then find the percentage of your profit for each property you invested; i.e.

Profit from a property = (actual price- asking price)/ asking price.

Finally, the average percentage profit you make across all 200 houses will be the final performance measure of your algorithm.

For the sake of simplicity, ignore the time value of money in your calculations.

Plagiarism declaration: You might be able to find additional information about the houses on the testing data. This is strictly prohibited. Any additional information used about these properties will be considered as breaching the plagiarism rules of LBS and you will fail this assignment.

This is an individual assignment. Although you are allowed to share your thoughts with your classmates (except where it is clearly stated not to), any form of sharing code and reports will be considered as breaching the plagiarism rules of LBS and you will fail this assignment, perhaps the course.

5. Deliverables

i) Submit your self-contained and clearly commented rmd file and knitted html file. Your file should also show clearly how you chose 200 houses from the testing data. (Comment out the parts of your code that does not contribute to your final algorithm.)

ii) Submit a csv file of the testing data by adding a column “buy” which is equal to 1 if you choose to buy the property, equals 0 otherwise (keep all the other columns). Name the csv file “[yourlastname_yourfirstname].csv”. You should choose exactly 200 properties. If you fail to follow these instructions, we will not grade your submission -- NO EXCEPTIONS.

iii) Also write a technical report explaining your approach following the technical guidelines (see the file here) up to 3000 words including appendix, following the writing guidelines here. Submit your report as a Microsoft word file. Any write-up longer than 3000 words will not be graded -- NO EXCEPTIONS. (See below for more details)

6. Elizabeth line (Crossrail)

Elizabeth line (or formerly known as Crossrail, which I prefer), is a new tube line which will traverse London from west to east. (See link for more information.) It is the most significant line added to London transport network in decades costing close to £20 billion. It is expected to reduce the commute times to Central London from many different neighborhoods. All Crossrail stations are already under construction and we have access to their location information (see link for more information about stations).

Crossrail will (or already did) have a significant impact on house prices on certain neighborhoods. However, it is not clear if its impact is completely "priced in". Please read the following article for more information.

It's Not Too Late For Investors to Get in on a London Crossrail Boon

(here is the pdf version in case URL does not open)

Assuming you can find the distance from each property to the closest Crossrail station, *how can you use your estimation engine to check if there is still a profit opportunity to invest in neighborhoods that will be served by Crossrail?* For simplicity, you can assume that Crossrail will open in the very near future (e.g. tomorrow), so that you don't have to worry about the time value of money. To answer this question, just explain the procedure to you will employ to find the impact of Crossrail on house prices, you do not need to implement your idea.

You are not allowed to discuss your ideas regarding this question with your classmates at all. If you fail to follow this guideline, the grade for the whole assignment will be affected.

7. Your report

Your report should follow the guidelines provided in the technical writing guidelines document. Also make sure it includes the following information.

- Clearly explain each step of your analysis, especially how you chose the 200 houses to invest in.

- Explain the additional features you engineer (if any) and explain why these features can help you predict the prices better.
- Explain what other (reasonable) information can be useful to predict the prices better.
- Clearly explain the additional assumptions you make during your analysis.
- Also explain which variables seem to be most important in predicting house prices and intuitively explain if these variables are what you would expect to be most important.
- Answer the question above about the potential impact of Crossrail on house prices in a separate section in the main body.