London House Price Prediction

Vivek Kanna Jayaprakash
MSc Data Science
City University of London
London, United Kingdom
Email : Vivek.Jayaprakash@city.ac.uk

Abstract : In this project we attempt to calculate the exact price of a Housing property in the current market rate in London based on the buyer's requirements Everybody deserves the right to know the exact market rate for the property they are setting out to buy especially in London after the pandemic of 2020.Various modelling methods are practiced to identify the best one, Visual representation of the data with the result and the most accurate model is deployed for any buyer to make use of the same for their needs.

I. INTRODUCTION

The surge in real estate prices after Covid-19 is a well-known event. The prices are said to have increased 1.65 times on an average worldwide because of the WFH culture and the need for good homes has been more predominant than ever before. Every central bank across the face of the planet has slashed the interest rates and governments cutting down on stamp duty combining with various tax benefits has lured everyone into purchasing a home for themselves. The increase in demand has therefore contributed heavily to the surge in the prices.

The property topic in the UK is one of the hugely advertised and highly politized topic since the 1700's. As discussed earlier the surge in Prices after Covid-19 is an extraordinary phenomenon because the last significant stock market crash before the Covid crash was the Lehmann brothers collapse more commonly known as the global financial crash of 2008 where the house prices in London went down by 12.6%. Everyone expected the trend to follow suit in 2020 but what has played out has shocked many of the busiest brains in the industry. The concept of buying a house at its fair price has become a Herculean task nowadays and this project aims at analyzing the process to make the buyer get their home at their preferred price. Buying a house is one the most important decision that an individual or a family could make in their lifetimes.

The need for knowing the exact fair price of a dream home is a basic right of every buyer and that is the major motivation for his project. There have been many instances while our House search when we really felt the need for a reliable source to tell us the exact price for our need. There are many websites in one's fingertips which tells how much is a current property worth but the very property selling websites only showcases the prices which the seller has inputted in. An new buyer wanting property at a said location with an idea of the area required and other required amenities should be able to find a perfect price for the desired dream property.

Our project aims at analyzing the current market trend in the city of London to predict the house prices based on location, total area and number of bathroom and bedrooms respectively to predict a price for the desired house.

II. ANALYTICAL QUESTIONS AND DATA

A. Analytical question

Based on our abstract and introduction we are mainly trying to answer a very simple question.

1. What is the price of a property in London?

   We are trying to approach this question by the needs of the buyer for example the price of a property based on the locality, Size, bathroom and bedroom

During the process of identifying the price we try to answer the set of other sub questions during the analytical process

2. What aspect contributes heavily to the price of a House?
3. What is the greatest outlier known to human knowledge which contributes most to the price.

B. Data

For the said series of questions, we require a data which should be recent and up to date to attempt at finding the recent prices and also should be having various aspects required for a property prediction including the Locality, Area (size of the property), Bathrooms, bedrooms, inside and outside London etc and the data – Housing Prices in London by Arnav Kulkarni recorded from Kaggle.com is perfect for the prediction.

III. DATA (MATERIALS)

A. Source and Characteristics

The data collected for this analysis is a special one. We have used the Housing Prices in London data set from the Kaggle free dataset as our primary data set which has been updated till recently to fit all the recent fluctuations in the price of the data in London. The entire dataset consists of 3480 rows 9 features and 1 target variable.

Totally up to 3480 instances of detailed transaction belonging to property description with various feature and target variable.

For each of the transaction the following various information are provided which we will consider as our independent variables.

1. Unnamed serial number.
2. Property Name (Street name)
3. HouseType (Flat/House/apartment/bungalow… etc)
4. AreainSQFT
5. Bedroom (No of bedrooms)
6. Bathroom (No of bedrooms)
7. Location (Area in London)
8. City/Country (London /Essex etc)
9. Postcode

And the dependent target variable

10.Price in GBP

The data as discussed in section II consists of all the key characteristics required for answering our questions for the project motivation.

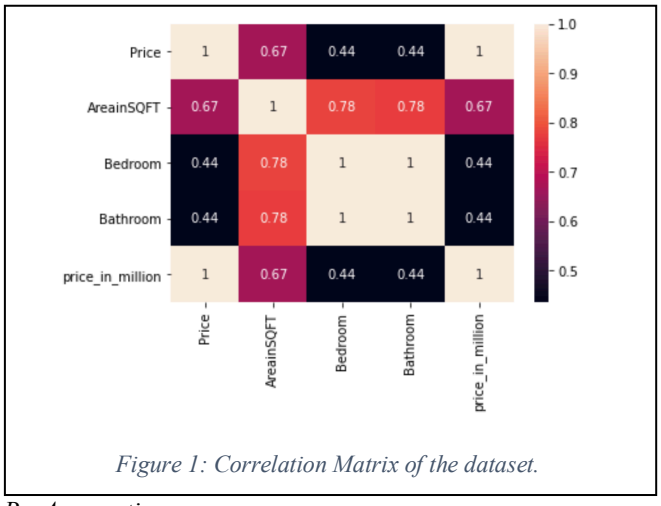

Figure 1: Correlation Matrix of the dataset.

B. Assumptions:

The data obtained is from Kaggle.com scrapped by the Contributor Arnav Kulkarni as mentioned is recent and has all the values listed as per the website scrapped for, We make the assumption that the above statement is true.

Other than the above mentioned assumptions we consider our data to be perfect for our analysis. If requires a lot of data cleaning and feature engineering for the data modelling to be performed but the raw data obtained is suitable for our analysis.

IV. Analysis

A. Data Preperation

Data preparation is the first and most important process of a data business analysis project. During this phase the raw unclean data is transformed into useful data and the information is sent to the data derivation stage for useful decision making.

In our project the data derived is found to be unclean and not so fit for our analysis hence we decide to clean up the data as per the following procedures

i) Heat-Map is plotted to identify the missing values in the dataset.

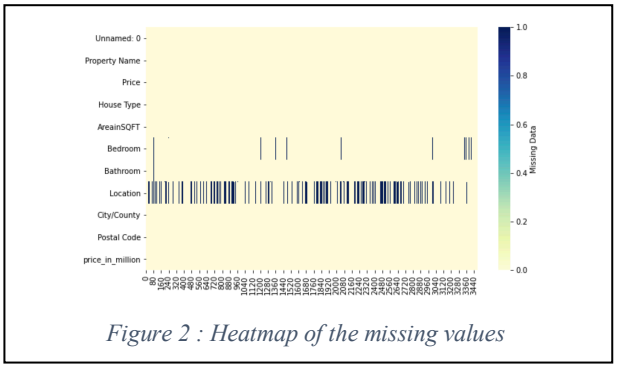

Figure 2 : Heatmap of the missing values

ii) Dropping few unnamed columns and the house types column from the data set as we are having the basic aim to find the price of a general house and not based on different house types.

iii) Comparing the Price feature to other features in our data set gives very interesting result. Even though the pattern mostly follows that if the area goes up price goes up rule but the significant outliers reveal that this is not the case always.Hence we shall discuss this in our findings.

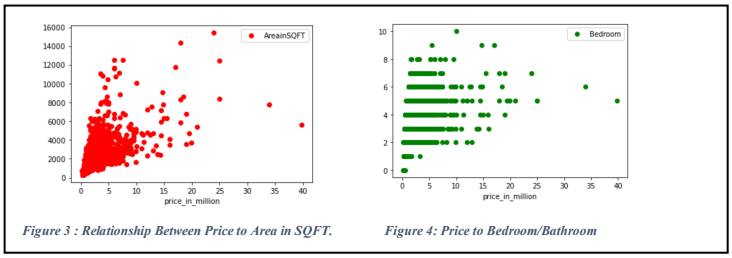

Figure 3 : Relationship Between Price to Area in SQFT.

Figure 4: Price to Bedroom/Bathroom

iv) We open a new column named price to millions to view the price of a property in its millionth currency value for better visualisation.

v) As per the Missing plot graph Figure 5 we identify that there are significant missing values in the dataset especially in the location features and few in the bedroom and bathroom features. As replacing the NaN values with the mean, median or mode wouldn't do justice to the project since we are building a prediction to identify the exact property prices we are decide to drop the same also because its total percentage is very less compared to the entire dataset.

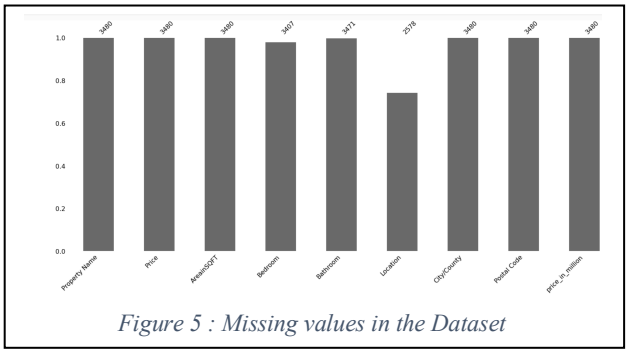

Figure 5 : Missing values in the Dataset

B. Data Derivation

After the unclean data is cleaned and prepared we move on to the data derivation part when further feature engineering an expletory analysis of the prepared data is done to make data more presentable.

i) Since we are creating a model for the generalised house across all verticals, more than 7 bedrooms for a house is present only in penthouses and bungalows Hence we remove the houses having more than 7 bathroom, As per the correlation matrix since the No of Bedrooms = No of bathrooms In our dataset, by removing houses with 7 bathrooms we automatically remove the houses with more than 7 bathrooms as well.

ii) Rate of the house per Sq/Ft is converted for easy interpretation of the result and for making further progress in the engineering side of things

iii) Some of the location in our dataset has houses which occur less than 5 times in the dataset. Since they are very sparsely located areas, and to prevent in distortion in our finding we name them as 'Others' (figure 6) for further processing.

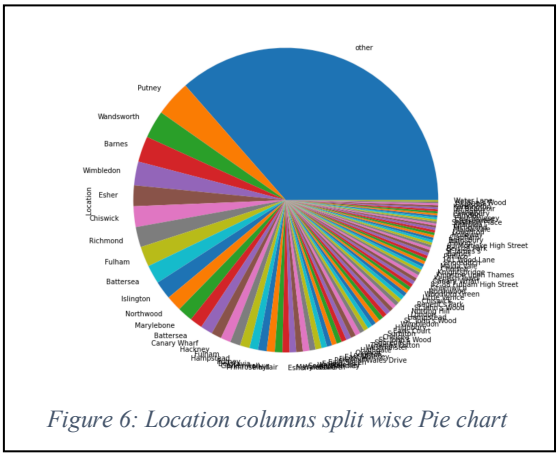

Figure 6: Location columns split wise Pie chart

iv) An ideal home as per the Buildings regulation act of 2010 in document M states that there should be a bedroom for every 250 Sq. Ft otherwise it would be very difficult for the people residing to maintain proper personal space and hygiene. Hence to maintain the standards we negate the houses which are don't have a sperate bedroom for every 250 sq. Ft.

v) To remove the outliers in the dataset for the ambition to obtain a normal distribution we remove the said such outliers in the dataset which are located one standard deviation away from the mean.

vi) As per figure 7 Few Houses tends to have smaller overall area than houses which have a higher number of bedrooms than theirs. This condition will affect the dataset as we are looking to predict an average house. Hence, we remove such houses for the dataset. Example 2 bedroom houses have more area than three bedroom houses. We remove such outliers.

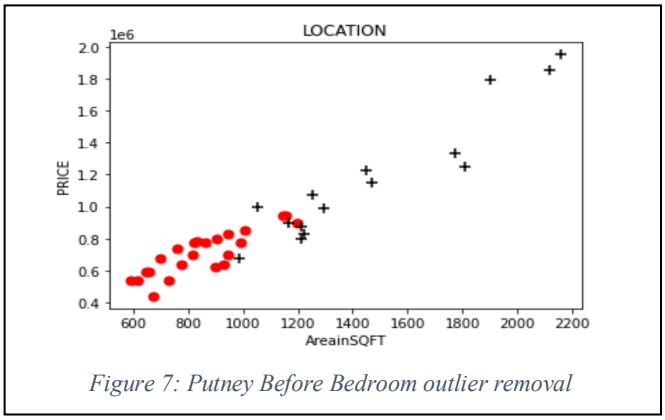

Figure 7: Putney Before Bedroom outlier removal

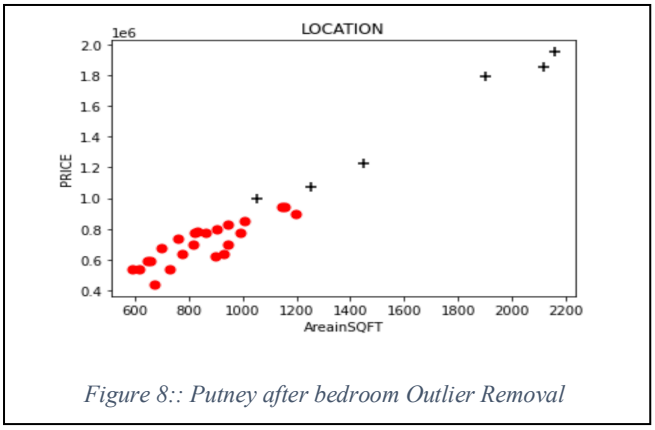

Figure 8:: Putney after bedroom Outlier Removal

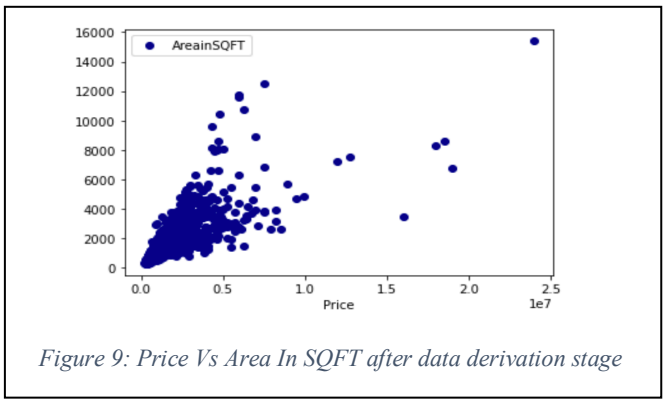

Figure 9: Price Vs Area In SQFT after data derivation stage

C. Construction of Models:

Since we are predicting price of an individual house based on the customer inputs we are building a regression model

There are pre-requisites before constructing the model. They are:

i) Creating dummies for the location ares. We create dummies for the location to make the dataset to be model as interpretable.
ii) Splitting the datasets: In order to maintain balance we split the dataset into 75% for training and 25% for testing.
iii) We build the models and calculate the R- Squared for the training and test set
The R squared value is the measure in statistic of how close the fitted regression line of our developed model is to our original data.

For our model we attempt to model the data using three regression techniques:

a). Linear Regression :

Linear regression (LR) modelling on this dataset is preferred to this dataset mainly due to the simplicity it processes which takes in series of continuous dependant variable to predict or estimate a continuous independent variable.

b). Decision tree Regression :

Decision tree is a special case model which can be used for both classification and regression tasks. Each tree is very simple model with sperate branches , nodes and leaves.

c). LASSO Regression :

The term LASSO stands for Lean absolute shrinkage and selection operator. The main advantage of lasso is it uses shrinkage in its modelling. The models considers a mean and shrinks all the point to the central point as its mean.

	Linear Regression	*Decision Tree Regression*	*LASSO Regression*
R Squared of Training set	82.68	99.94	82.68
R Squared of Test Set	84.70	79.01	84.70

Table 1: R squared for the regression models

From the arrived results we identify that the R-squared value is similar to Liner and Lasso and they provide the best accuracy.

From the results to further elevate the model we shuffle the data and perform hyperparameter tuning on the modelling to fine tune the model for sending it to deployment.

D. Validation of results:

a). K Fold Cross validation :

K fold cross validation is a very efficient shuffling technique used in machine learning where the training and test day is split into a number of K folds. In this manner every fold becomes the test set for a fully even shuffling to be undergone.

b). Hyperparameter tuning using GridSearchCV :

Grid search is the simplest form of hyperparameter tuning where the domain of the hyperparameters is divided into discrete girds Then every combination within the grid is tried while calculating the performance metrics of the model using the cross val function. The point of the process where the grid is maximum in the cross valuation average, is the exact optimum hyperparameter combination values.

The parameters in the said step is different for each model as it all involves various stages of modelling on its own.

Hyper parameter results:

	model	best_score	best_parms
0	linear_regressor	0.798661	{'normalize': True}
1	lasso	0.798668	{'alpha': 2, 'selection': 'cyclic'}
2	decision_tree	0.678091	{'criterion': 'mse', 'splitter': 'best'}

Table 2 Best model with Parms and accuracy score.

V. Findings, Reflections and Further work.

a). Findings and Reflections :

The predicted and the original set comparison as per figure 10 proves that the findings are totally valid as there is very little discrepancy in the prediction. Thereby answering our first question that we are able to make an accurate prediction based on the customer's needs.

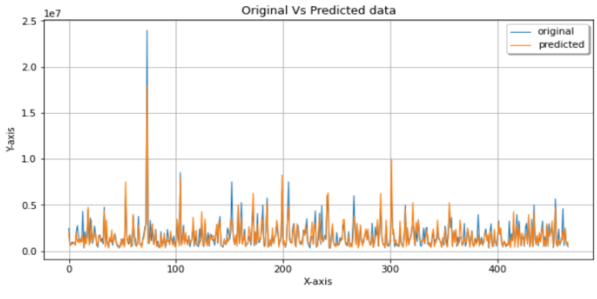

Figure 10: Original vs predicted data comparison

To identify the basic trend for a price of a property, we try to understand if there is any trend suited for the way a price of property is determined.

Figure 3 and 4 Shows the relationship between the Area in Sqft and Price of a property which shows that even though the prices increase with the size of the property but the trend is not quite predominant as we can find highly expensive properties with less size and cheap properties which are large in size. The same trend is prevalent in no of bathrooms/bedrooms to price.
The location of the property also follows a trend but there are some properties in the same location whose price fluctuate.

Hence, we conclude that there is a huge multicollinearity within the data and it takes various attributes grouped together to predict the price of a house Which answers our second question.

Based on our findings, the price of a property is decided by various features and to identify if any said outlier contributes heavily to a price like the construction of bathroom, availability of a garden etc, further research and data should be collected to answer the question no 3 more efficiently.

We are able to answer our main analytical question of what is the price of a house in London based on the buyers requirements with an accuracy of 84.7% using Linear regression and we are also able to answer the other analytical sub question too.

b). Critical Analysis and future work

1) Our raw dataset contains 3840 instances of the data and for modelling it was cleaned and filtered to be 1980 instances, which is not the real refection of the exact societal trends. In future we attempt to replicate this model with a much higher collection of data and thereby trying to reflect on the result obtained.

2) A discussed in section 3 we chose to ignore the Housetypes in the data set. For example a bungalow with a small area located in the outskirts would be very expensive compared to the other houses of the same area because it may have various inhouse designer interiors, garages, swimming pools etc. These contribute heavily to the price of the house. In the future work we try to do separate modelling for the elite houses and average house to get a balanced model for both sets of housing classes.

3) Since we are using a fairly small dataset, the Lasso regression is giving a almost similar modelling as that of linear regression. In other words, the modelling doesn't have any problems for the shrinkage ability of LASSO to solve hence it gives a similar score to that of linear regression. During our future work of using higher sets of data, LASSO would be very useful tool to model the high multicollinearity property data but not so useful in this case.

4) The property prices worldwide as discussed in section 1 is subject to change with the economic changes in the real world. The Lehmann collapse plummeted the reality price while the covid crises shot up the prices, hence with the uncertain world the mentioned model is a working format only for the current normal scenarios to be prevalent throughout the years but on the occasion of any unforeseen event, the Prices might fluctuate and in future a suitable modelling which can taken into account every possible outcomes should be designed.

Word Counts

- Abstract : 87 words
- Introduction : 399 words
- Analytical questions and data: 177 words
- Data(Materials) : 251 words
- Analysis : 1102 words
- Findings, reflection and further work : 625 words

Total words : 2641 out of 2650 Allowed

REFERENCES

[1] Balakumar, B., Raviraj, P. and Essakkiammal, S. (n.d.). *Predicting Housing Prices using Machine Learning Techniques*. [online] Available at: http://sajrest.com/Archives/vol4issue4_2019/v4i4p1.pdf.

[2] Hamizah Zulkifley, N., Abdul Rahman, S., Ubaidullah, N.H. and Ibrahim, I. (2020). House Price Prediction using a Machine Learning Model: A Survey of Literature. *International Journal of Modern Education and Computer Science*, 12(6), pp.46–54.

[3] housemetric.co.uk. (n.d.). *Analysis of house prices in London (NW10 4)*. [online] Available at: https://housemetric.co.uk/house-price-analysis/NW10-4/London [Accessed 20 Dec. 2021].

[4] Kirenz, J. (2021). *Lasso Regression with Python*. [online] Jan Kirenz. Available at: https://www.kirenz.com/post/2019-08-12-python-lasso-regression-auto/ [Accessed 20 Dec. 2021].

[5] Lowrance, R., Lecun, Y. and Shasha, D. (2015). *Predicting the Market Value of Single-Family Residential Real Estate*.

[6] Marsden, J. (2015). House prices in London – an economic analysis of London's housing market. *House prices in London – an economic analysis of London's housing market*, Working paper 72(1), pp.11–19.

[7] Mujtaba, H. (2020). *An Introduction to Grid Search CV | What is Grid Search*. [online] GreatLearning. Available at: https://www.mygreatlearning.com/blog/gridsearchcv/.

[8] Ng, A. (2015). *Machine Learning for a London Housing Price Prediction Mobile Application*. [Individual report : Final project] pp.18–35. Available at: https://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf.

[9] kaggle.com. (n.d.). *Housing Prices in London*. [online] Available at: https://www.kaggle.com/arnavkulkarni/housing-prices-in-london.

[10] Zaman, U., Waqar, M. and Zaman, A. (2021) . Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. *Sof t Computing and Machine Intelligence Journal*, (1), p.2021

[11] Pow, N., Janulewicz, E. and Liu, L. (Dave) (n.d.). *Applied Machine Learning Project 4 Prediction of real estate property prices in Montr eal*. [online] www.readkong.com. Available at: https://www.readkong.com/page/applied-machine-learning-project-4-prediction-of-real-4501082.

[12] Refaeilzadeh, P., Tang, L. and Liu, H. (n.d.). *C Cross-Validation*. [online] Available at: http://leitang.net/papers/ency-cross-validation.pdf.