ECS784P Data Analytics Project
Machine Learning for Housing Price Prediction

Alberto RAMIREZ MUNGUIA
ID: 180796108
a.ramirezmunguia@se18.qmul.ac.uk

March 30, 2019

Group size: 4

Table of individual contribution by each member of my group, based on my subjective opinion:

Student	**Effort**
Juan Carlos FERNANDEZ GOMEZ	25%
Nicholas Zhi LI	25%
Alba Evelyn MARQUEZ MARQUEZ	25%
Alberto RAMIREZ MUNGUIA	25%

Abstract

We present multiple linear regression (MLR), polynomial regression (PR), artificial neural network (ANN) and long short-term memory (LSTM) models for daily housing price analysis and prediction in three London boroughs for the time period between 1995 and 2018. MLR was chosen as the baseline model, while PR, ANN and LSTM were chosen as the more advanced data analytics and machine learning models. RMSE was used as the performance metrics. The results showed that ANN was the the best performing model, while LSTM was outperformed by the baseline MLR model. This suggested the ANN model was the optimal model to be used with the features studied in this work, namely: location, type of property, type of tenure (or duration) and whether the property was newly built.

1 Introduction

Real estate is an asset class crucial to a household, which forms one of the largest proportion of many household wealth and savings [1]. The UK housing market has become a contentious issue in recent years, with politicians from all parts of the political spectrum reaching consensus [2]. The consensus being that the housing prices, particularly in London and the South-east of England, are rising due to: (a) supply and demand driven chronic shortage of affordable housing [2], and (b) activity by speculative investors. Indeed, while the number of households increased by over 30% from 1981 to 2014, the population increased by just under 15% over the same period of time [1].

The objectives of this work is to illustrate the change in daily housing prices in three London boroughs, between the years of 1995 and 2018. A multiple linear regression model was used as the baseline to establish the relationship between the prices and various features extracted from the data. This is then used to make a baseline prediction for future house prices. More sophisticated polynomial regression and machine learning models such as artificial neural network (ANN) and long short-term memory (LSTM) model were then used to further learn the patterns within the data. These were then used for prediction and tested for their accuracy.

2 Background and Related Work

2.1 Factors Affecting Housing Price

Although housing prices are primarily a function of supply and demand, the underlying factors that correlates with housing prices may also include the features of the property e.g. living area, number of bedrooms, number of bathrooms, lot size, age of house [3], tenure (i.e. freehold or leasehold) and type of property (i.e semi-detached house, detached house or flat). This work estimates the relative change in housing prices over the specified time period and attempt to attribute the change to various contributing factors, thus attempt to make future housing price predictions with said contributing factors.

Research in housing price analysis and prediction is abundant. Bhalla (2008) discussed the housing market development and performance as a financial industry segment, and concluded housing finances grow at the rate of 36% [4]. Bhalla, Arora and Gill (2009) further examined the performance of housing sector and showed that due to continuous change in the *global macroeconomic environment* resulted in a slow and gradual growth in the housing sector. They observed reduced *financing costs* which contributed to global real estate price increase. In addition to *location*, *local services* was also a contributing factor in determining property prices. For instance, the presence of local services such as parks provide additional environmental quality. Housing prices may be influence by these additional qualities [5]. Other macro-economic factors that influence first time buyers the most were: *house prices increase*, the amount of deposit required and the level of *personal income* [6].

2.2 Models for Housing Price Analysis and Prediction

Slone *et. al.* (2014) analysed the relationship between the asking prices and various characteristics of residential properties using both simple linear regression and multiple linear regression. They used the square footage as the only explanatory variable in the simple linear regression, while the MLR contained addition variables such as the land parcel size, the number of bedrooms and the year of construction. The MLR results proved the bias due to the omission of crucial factors in the simple linear regression. They found that the square footage was the most important factor in determining residential property prices [7]. A number of data analytics methods have been employed in housing price research. These include econometrics techniques (e.g. ARIMA, multiple linear regression) to machine learning techniques (e.g. artificial neural networks, fuzzy logic) [8][9]. Comparison was drawn between a hedonic pricing model and artificial neural networks and while the results from hedonic pricing models supported previous findings, it was found that the artificial neural networks offered better performance in out-of-sample forecast, based on a trial-and-error strategy [10].

When comparing the performance between MLR and ANN, one study found that the value of R^2 in neural network model is higher than MLR model by 26.475%. The value of Mean Squared Error (MSE) in neural network model also lower compared to MLR model [3]. The ANN [11] and LSTM [12] models have gain popularity in housing price analysis and prediction.

3 Problem Statement and Hypothesis

The market value of real estate has long been dis-associated from their intrinsic, or brick and mortar, value. The value of a house should come close to the present value of the cash inflow generated by the property minus any services or maintenance over its lifetime. But as we see later, this is not the case. Indeed, people do not value houses based only on the perpetual flow of accommodation and amenities provided by the property. A time-series analysis has thus been conducted to illustrate the disjoint between the intrinsic and real prices. [1]

If a house is freehold, the purchaser would also buy the land on which the house is built and will hold the property indefinitely. If a house is lease-hold, the purchaser is essentially pre-paying the rent for the lease period (be it 99 years or 999 years) upfront. Type of tenure (or duration) is investigated as a feature in this report.

The aim of this word is testing various models specified in Section 4 to predict future property prices. The underlying hypothesis for this work is that the real house prices would invariably increase with time, as this is driven by population increase given a finite living spaces in a city. However, we hypothesise that external political events would cause greater fluctuation to properties in a luxury area than properties in relatively deprived area. Furthermore, we also hypothesise that properties in luxury area would not follow the natural evolution of price, but would be much more susceptible to supply and demand caused by the geopolitical climate.

4 Methodology

4.1 Multiple Linear Regression

Multiple linear regression (MLR) method is a generalisation of the simple linear regression method, MLR considers more than one explanatory variables (or independent variables) to calculate the response (or the dependent variable). In contrast, simple linear regression method uses only one independent variable [13]. In this study, the dependent variable to be predicted is the housing price itself, the independent variables are the features (or parameters) related to each house, e.g. *"date of transfer"*, *"property type"*, *"old or new build"* and *"district"*. The linear regression model can be represented as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_x + \epsilon \tag{1}$$

Equation (1) is the MLR model for the k regressors. The parameters β, j = 0,1,...,k are the regression coefficients, and β_0 is the intercept. sample unit form, where ϵ is a random error. [13]

4.2 Polynomial Regression

Polynomial models follow similar principles to linear models, where one (or several) explanatory variables are used to estimate and predict the response (or dependent variable) over a specific range of inputs. However, these models are used when the relationship between the explanatory variables and the response is curvilinear. In this case, the model has the form of the equation below:

$$y_i = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_{11} z_1^2 + \beta_{22} z_2^2 + \beta_{12} z_1 z_2 + \epsilon_j \tag{2}$$

Equation (2) is the quadratic (second-order) polynomial model for two explanatory variables, where the single x-terms are called the main effects and the second-order terms are called the quadratic effects. These are the elements that allow the model to follow the curve in the response. The cross-product terms are used to model interactions between the explanatory variables [14].

A higher order can be introduced to the model by following the same relationships between main effects, exponential effects and cross-products. However, in this document, only a quadratic model will be considered.

4.3 Artificial Neural Network

Artificial neural network (ANN) refers to a computational model similar to the human brain and individual neurons behaviour. An ANN is composed of an input layer, an output layer and hidden layers (intermediate layers), layers are composed by neurons (nodes). The neural network architecture defines how the neurons are interconnected, each connection for each neuron posses an individual weight. The weight represents the intensity of the signal that is being passed to the next neuron.

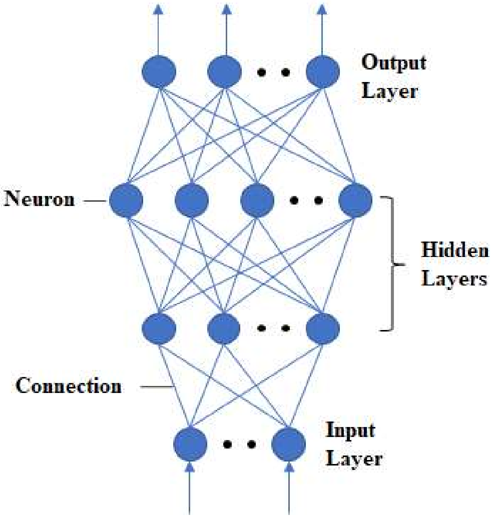


Figure 1: Feed-forward neural network structure with two hidden layers. [15]

The neurons or nodes are the smallest processing unit by a Neural Network and it can receive inputs from other neurons and a bias adjustment, then using an activation function it produces an output. The output of each neuron $node_j$ is the computation result of applying a transfer function ϕ to the summation of all signals from each connection A_i times the value of the connection weight between node j and connection i (W_{ji}) (refer to equations 3 and 4).

$$Sum_j = \sum_j (W_j i A_i) \tag{3}$$

$$O_j = \phi(Sum_j) \tag{4}$$

Where O_j is output for node j and ϕ is transfer function which can take many different forms: linear functions, linear threshold functions, step linear functions, sigmoid function or Gaussian functions [15]. The training of the NN consists of using labelled data, in the case of house price prediction is the price and the attributes of features of the house. The information of the features is forward propagated trough the Neural Network and the result is compared to the actual price of the house, the error obtained is the input to the loss function and back propagates to update the weights in the network and minimise the error. The objective is to reduce the error every time that the house price is calculated. Therefore, a number of iteration times called epochs are defined for the training process, in conjunction with a loss function.

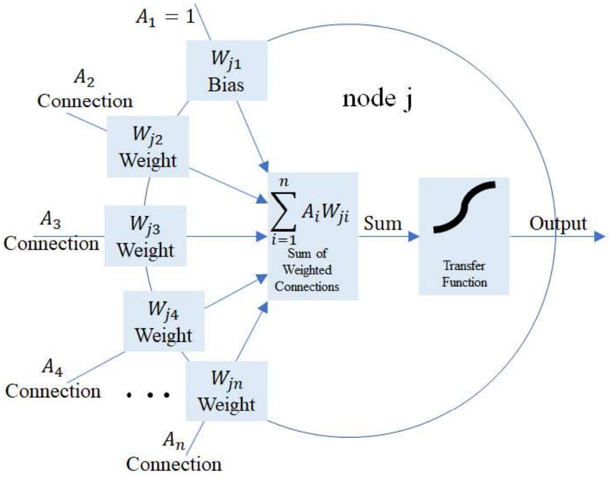


Figure 2: Structure of a computational unit (node j). [15]

4.4 LSTM

Long short-term memory (LSTM) networks are a type of recurrent neural networks (RNN). Introduced by Hochreiter & Schmidhuber in 1997 [16], LSTM are able to learn long-term dependencies, work very well on a large variety of problems and have gained popularity. The fact that LSTM models are able to memorise long-term information meant that they are able to avoid the long-term dependency problems other models succumb. The difference between a standard RNN and LSTM is that in standard RNNs, the repeating module will have a very simple structure, such as a single *tanh* layer; whereas LSTMs have a chain-like structure and the repeating modules have different structures.

Figure 3 illustrates a typical LSTM module, where X_t is the input vector, c_{t-1} is the memory from previous state, h_{t-1} is the output from previous state, C_t is the memory of current state, h_t is the output of current state and W is the weights. The intermediary parameters f_t, i_t, $\tilde{C}_t$ and o_t are defined in equations (5), (6), (7) and (9) respectively. The key to LSTMs is the cell state c_{t-1} and C_t which runs through the LSTM modules like a conveyor belt. The modules have the ability to alter information to the LSTM cell state which are carefully regulated.

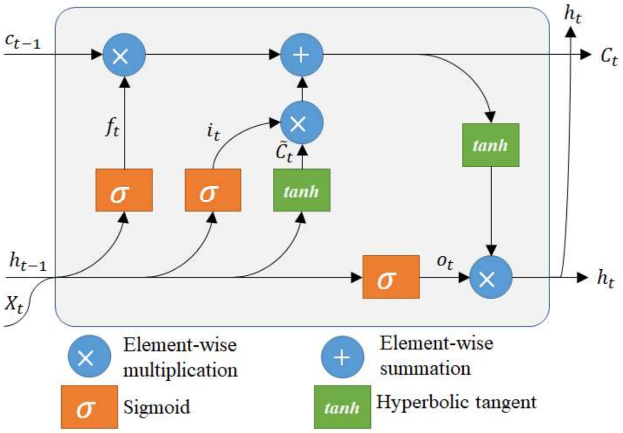


Figure 3: A typical LSTM module [16]

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{6}$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t + b_C) \tag{7}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{8}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{9}$$

$$h_t = o_t * tanh(C_t) \tag{10}$$

5 Preliminary Analyses

5.1 Dataset

For model analysis purposes, the ”price paid data” which tracks property sales in England and Wales submitted to HM Land Registry for registration was selected. Price paid data is based on the raw data released each month. The file selected include standard and additional price paid data transactions received at HM Land Registry from 1 January 1995 to the most current monthly data.

The amount of time between the sale of a property and the registration of this information with HM Land Registry varies. It typically ranges between 2 weeks and 2 months. Data for the two most recent months is therefore incomplete and does not give an indication of final monthly volumes. Occasionally the interval between sale and registration is longer than two months. The small number of sales affected cannot be updated for publication until the sales are lodged for registration [17]. For the scope of this document, two different sets of observations will be considered. We will be talking about Nominal prices, for all values registered in their actual price at the moment of sale, and Real prices, normalised to 1995 in consideration to inflation, so we can prevent this economical factor from altering the real behaviour of the models and statistic analysis.

5.1.1 Pre-processing the Dataset

After initial exploration of data structure, a processing phase was carried over. This phase included:

- Dataset size exploration with an initial file containing 16 columns (features) and 20,020,162 records.
- A first dataset size reduction in where a filter was applied to obtain only records related to 'Greater London' county. Sub set obtained was 3,154,384 records.
- Second dataset reduction to obtain a subset of records that included boroughs of : City of Westminster, Lambeth and Croydon. Motivation for this borough selection can be read in next section. This last reduction resulted in a dataset containing 6 columns and 377,829 records.
- Having obtained a smaller version of the initial dataset, checks for consistency was carried over(nulls, counts, data types) as well as an statistics exploration described in next sections.
- This final smaller version of dataset was used as a base across analysis and model implementation.
- Using the base dataset, further processing was made to obtain additional subsets such as: removing property type "Others", remove property type "Others" and outliers and also a subset to consider real house price by calculating it using Consumer Price Index for Housing (CPIH) data. Base year used for price standardisation was 1995.

5.1.2 Features included and brief description

Feature	**Explanation**
Transaction unique identifier	A reference number which is generated automatically recording each published sale.
Price	Sale price stated on the transfer deed.
Date of Transfer	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. Note that postcodes can be reallocated and these changes are not reflected in the Price Paid Dataset.
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.

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
Duration	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.
PAON	Primary Addressable Object Name. Typically the house number or name.
SAON	Secondary Addressable Object Name. Where a property has been divided into separate units (for example, flats), the PAON (above) will identify the building and a SAON will be specified that identifies the separate unit/flat.
Street	Street name of the property address
Locality	Locality name of the property address
Town/City	Town or city name of the property address
District	District name of the property address
County	Country name of the property address
PPD Category Type	Indicates the type of Price Paid transaction. A = Standard Price Paid entry, includes single residential property sold for full market value. B = Additional Price Paid entry including transfers under a power of sale/repossessions, buy-to-lets (where they can be identified by a Mortgage) and transfers to non-private individuals. Note that category B does not separately identify the transaction types stated. HM Land Registry has been collecting information on Category A transactions from January 1995. Category B transactions were identified from October 2013
Record Status - monthly file only	Indicates additions, changes and deletions to the records.(see guide below). A = Addition C = Change D = Delete. Note that where a transaction changes category type due to misallocation (as above) it will be deleted from the original category type and added to the correct category with a new transaction unique identifier.

5.1.3 Selected features and motivation for this decision

Selected Attribute	Explanation
Price	This is our dependent variable and the central feature of our dataset. We will be running the models around it.
Date of transfer	The trend shows a consistent increase in our dependent variable related to time, even after inflation.
Property type	We observe a different behaviour for each of the property types, inferring the requirement of a special analysis for different subgroups.
Old/New	The distribution of old/new for each borough shows a considerable variability, so we will define if it influences the overall price.
Duration	The distribution of duration, similar to the previous feature, shows a considerable variability for each borough, so we will also define if it influences the overall price.
District	Consistent with scope of the project, we will be analysing three different districts to compare the price evolution among them and the accuracy obtained by the models.

5.2 Selection of Boroughs

The chosen dataset included the 32 boroughs that make up the Greater London county. For this analysis, it was decided that the selection of three representative areas in the city would serve to compare both the evolution of the market and the performance of the predictive models for each one of them.The selection of boroughs was based on three critical elements:

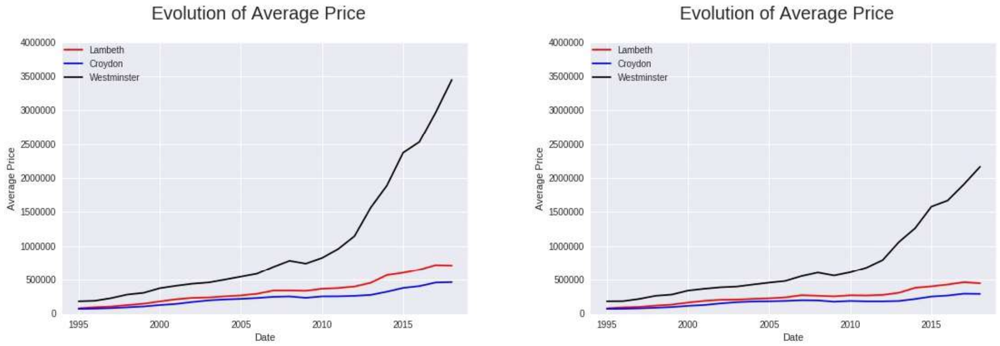
- The amount of transactions available for the 23-year interval.
- The variability of the average price during the sampling period.
- The diversity of tier representation according to price, which means that, after sorting from the highest to the lowest average price, a borough from each one of the tiers should be selected to avoid a repetitive or biased behaviour.

In table 2 we can observe the top ten boroughs according to the transaction count. Out of these, we selected three that represented each of the tiers sorted by price, and finally, we compared the trends to identify those that showed a significant variability in price evolution during the period.

Table 2: Top ten boroughs per transaction count

District	Average Price	Count
Wandsworth	176,506	55,008
Croydon	101,905	48,088
Bromley	131,496	46,754
Barnet	160,728	45,946
Enfield	111,562	42,662
City of Westminster	291,274	40,473
Ealing	140,137	39,009
Lambeth	141,387	38,243
Redbridge	107,731	35,232
Lewisham	93,936	34,737

As shown in Figure 4, the three candidate boroughs show a very distinctive behaviour one from the others, complying with the required variability, both for nominal prices and for real prices. For this reason, after this analysis, *City of Westminster* was selected as the top-tier, *Lambeth* as the mid-tier and *Croydon* as the lower-tier representation respectively.



(a) Nominal price

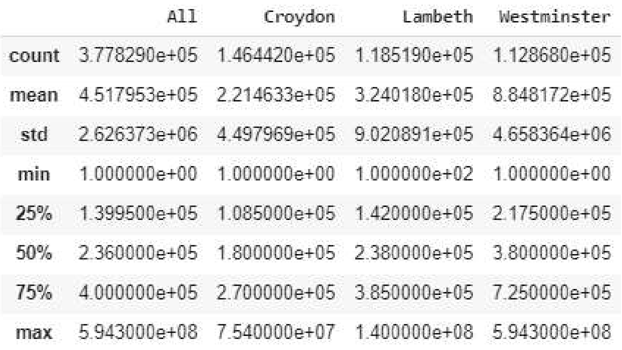
(b) Real price

Figure 4: Evolution of average price on selected boroughs

5.3 Centrality and Dispersion

One of the first analysis of this dataset involves the comparison between centrality and dispersion metrics among the three selected boroughs, both in nominal and real prices. This will allow us to understand the general characteristics of the distributions. In figure 5, the difference between all boroughs is shown. As expected, Westminster is in average around 4 times the price of Croydon. However, the standard deviation shows a difference of 10x, showing a higher variability that could be prejudicial to our prediction model.

If we consider the range of the first three quartiles and compare it against the fourth one (between 75% and max), we can observe a considerably larger range in all boroughs, suggesting a significant number of outliers on our dataset. We will have to take this into consideration for further analysis.



	All	Croydon	Lambeth	Westminster
count	3.778290e+05	1.464420e+05	1.185190e+05	1.128680e+05
mean	4.517953e+05	2.214633e+05	3.240180e+05	8.848172e+05
std	2.626373e+06	4.497969e+05	9.020891e+05	4.658364e+06
min	1.000000e+00	1.000000e+00	1.000000e+02	1.000000e+00
25%	1.399500e+05	1.085000e+05	1.420000e+05	2.175000e+05
50%	2.360000e+05	1.800000e+05	2.380000e+05	3.800000e+05
75%	4.000000e+05	2.700000e+05	3.850000e+05	7.250000e+05
max	5.943000e+08	7.540000e+07	1.400000e+08	5.943000e+08

Figure 5: Metrics for all boroughs in Nominal Prices

5.4 Statistical Analysis

A plot was constructed to visualise the general distribution of the transactions. In Figure 6, the comparison of the three sets of data is shown for nominal prices. The outliers observable in all the graphs are relevant, since it is apparent that we are dealing with extreme-value distributions. This characteristic could affect our predictive models, so a deeper understanding is required.

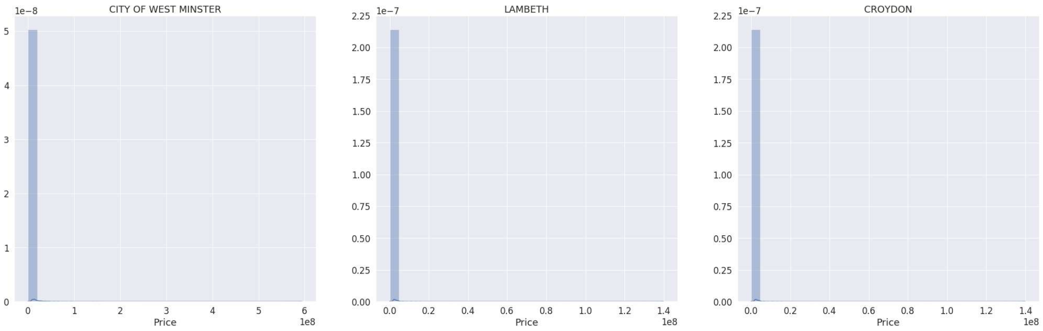
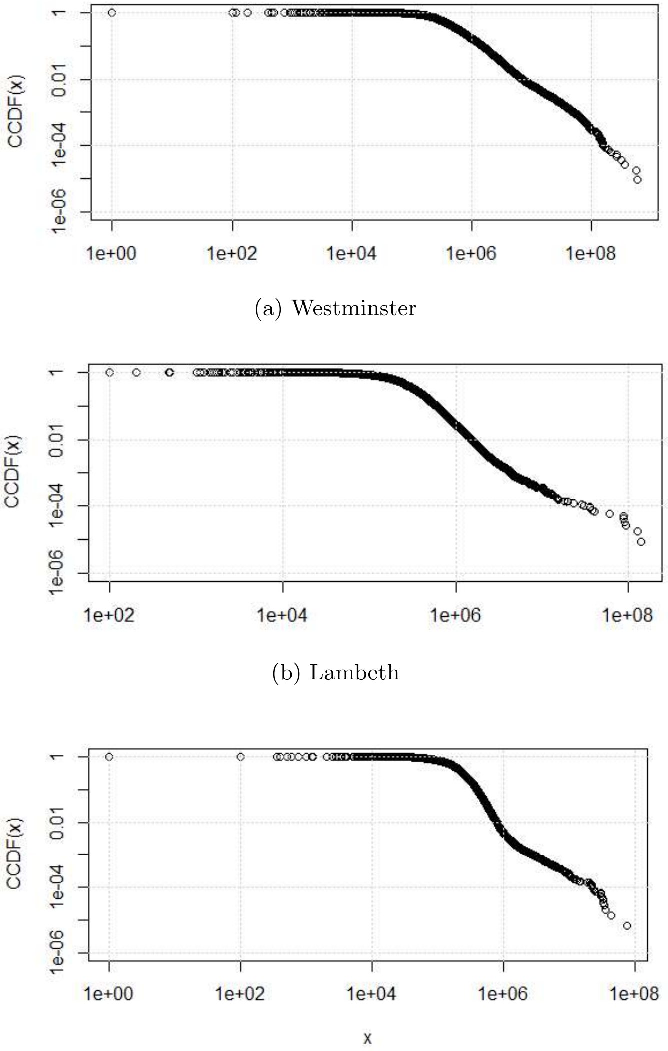


Figure 6: Distribution per borough

By looking at the tail of the complementary cumulative distributions shown in Figure 7, we can confirm our previous supposition, where the three datasets present extreme values that must be considered in the development of the models to guarantee an increased performance. Croydon and Lambeth, in particular, show in the cadence of the tail that a very small amount of data represents the most extreme values, while Westminster shows more consistency.



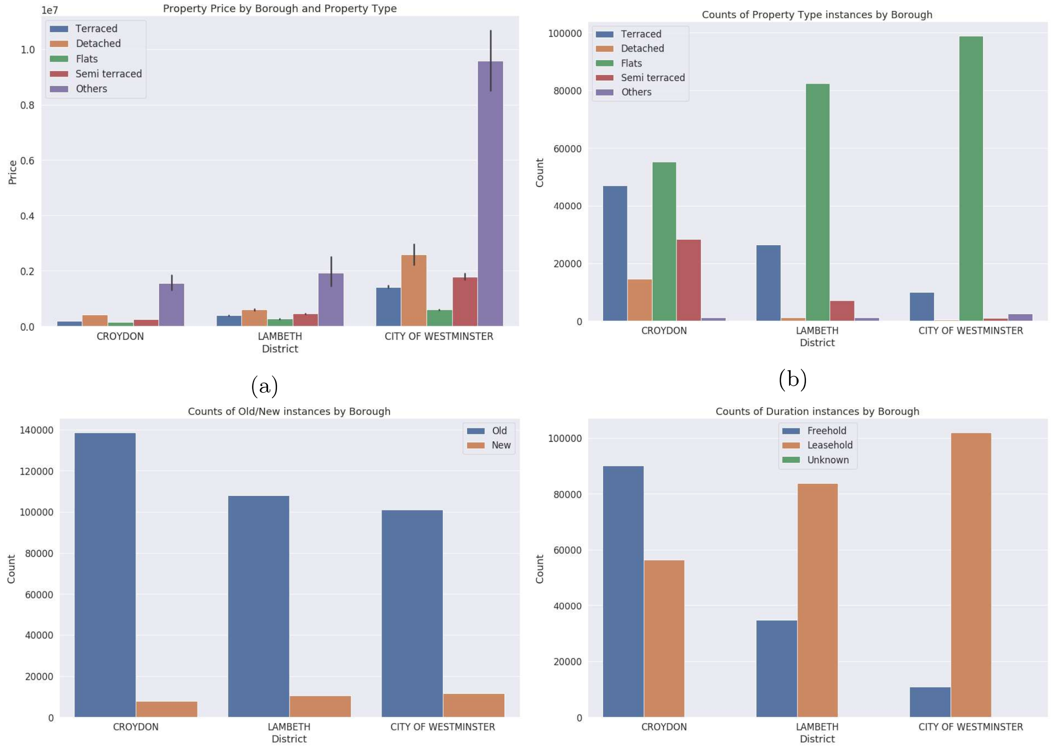
(a) Westminster

(b) Lambeth

(c) Croydon

Figure 7: Complementary cumulative distribution function (CCDF) of all boroughs

After breaking down the datasets by considering each of the selected features, like duration or property type, we can observe several contrasting trends. For example, while Westminster, which represents the top-tier of the properties by average price, shows that 90 % of the properties are leasehold, less than 40% of the transactions in Croydon have this type. This suggests that the real perceived value in Croydon is much lower than expected, since even for a "better" deal like a freehold, people are not willing to invest larger amounts of money in that neighbourhood.



(a) (b)

(c) (d)

Figure 8: Decomposition of data per attributes

It is observable in Figure 9 that the property type ”Others”, while representing less than 1% of all the data, incurs in most of the outliers for all three boroughs, specially Westminster. This is understandable since this category represents unique transactions for properties such as residential buildings or commercial compounds that remain out of the standard categorisation.

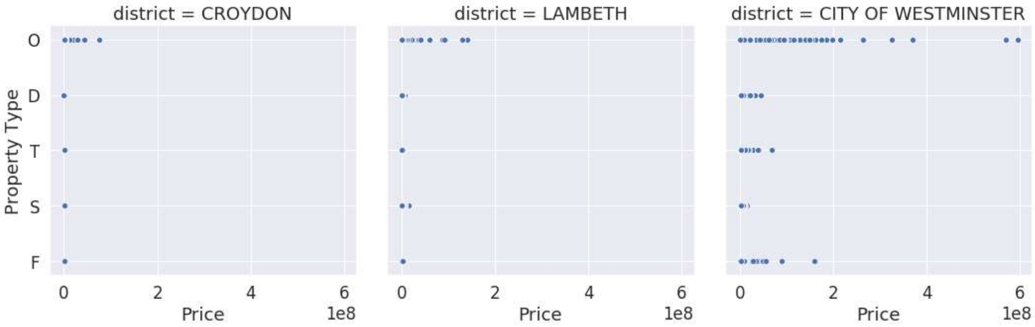
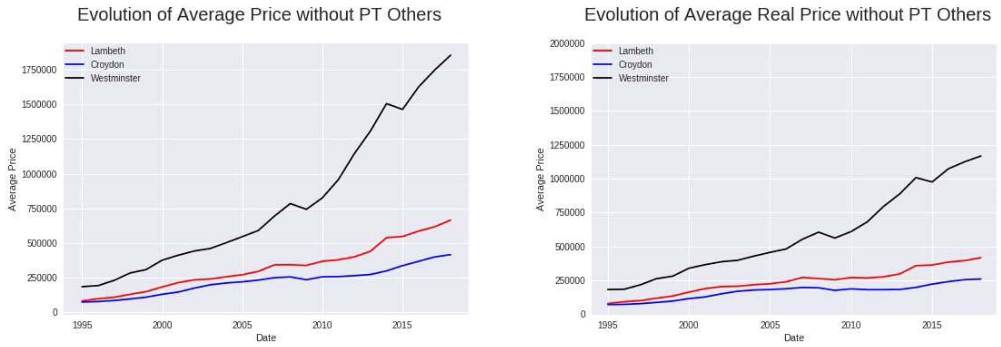


Figure 9: Price distribution per property type

For this reason, property type ”Others” will be excluded from our analysis and prediction models, since it shows a stochastic trend with larger variability. This is shown by plotting the price evolution for all property types besides ”Others” in Figure 10, where we can observe that, by separating this type, a much more linear trend than the previous graph in Figure 4 is obtained. This proves that the behaviour in both groups follows a very contrasting trend. For this reason, a separate analysis for all types of property besides ”Others” will be developed. The models will also be constructed by taking into consideration this constraint.



(a) Nominal

(b) Real

Figure 10: Price evolution of selected boroughs without property type ”Others”

By excluding property type ”Others” from the analysis, it is observable in Lambeth and Croydon that the distribution of transactions per borough shows now a more defined shape, where the curve represents how the price is distributed in the different boroughs. City of Westminster, on the contrary, still shows an extreme-value behaviour. This will be further analysed in subsequent sections.

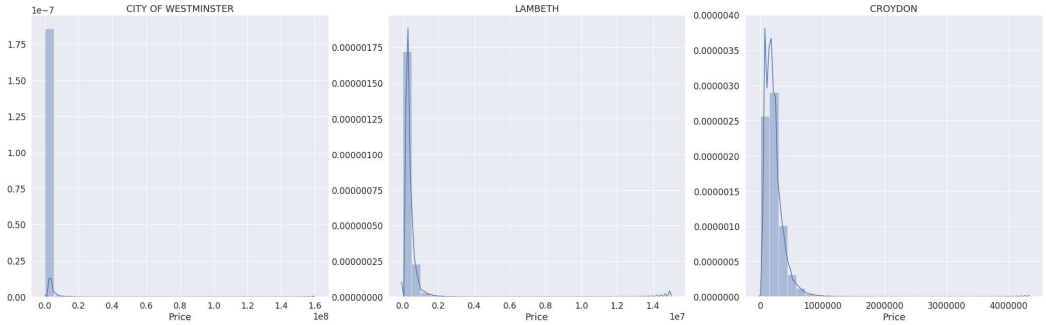
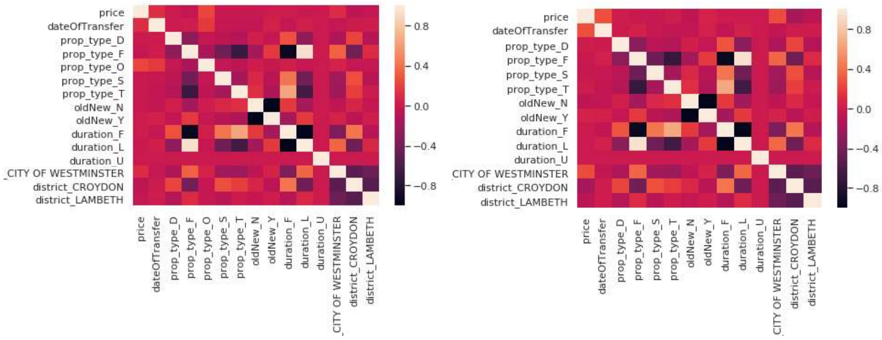


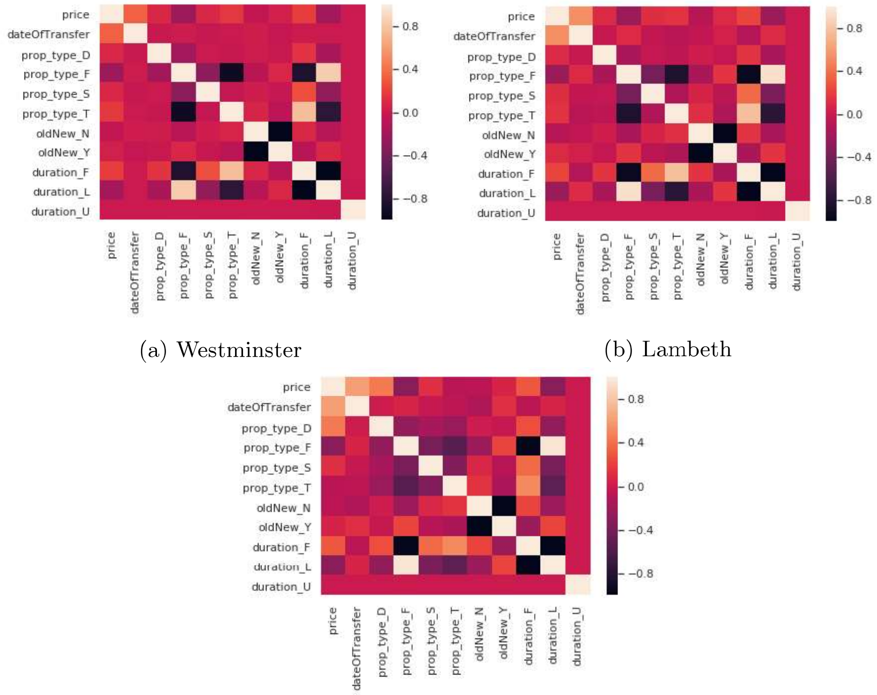
Figure 11: Distribution per borough excluding property type ”Others”

A similar analysis of Pearson correlations and trends between features was done for specified data subsets: all property types, excluding property type ”Others” and the latter split by borough. In Figure 12, is observed that from price correlations, date of transfer happens to be the most correlated regardless the exclusion of property type ”others” and no significant correlation is found with other features. However in Figure 13, once the subset of data split by boroughs is observed that correlations get stronger, specially for Croydon and Lambeth. These observations contributed to decision to carry over prediction exercises with different subsets of the data.



(a) All property types (b) Excluding property type ”Others”

Figure 12: Correlation between the different variables for the three boroughs



(a) Westminster (b) Lambeth

(c) Croydon

Figure 13: Correlation between the different variables per three boroughs and excluding property type ”Others”

6 Model Architecture

6.1 Fundamentals

The proposed analysis includes some of the widely used methodologies for house price prediction presented in section 2 (MLR, PR, ANN, LSTM), and a comparison of the obtained results. The performance measure selected for the comparison between the models is the Root Mean Squared Error. RMSE is a typical performace measure for regression problems [18]. The following is the mathematical formula to compute the RMSE.

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)} - y^{(i)})^2} \tag{11}$$

Equation (11) calculates the RMSE, where m is the number of instances in the dataset, $x^{(i)}$ is a vector of all the feature values (excluding the label) of the i^{th} instance in the dataset, and $y^{(i)}$ is its label (the desired output value for that instance). MLR was chosen to introduce the most basic model to predict the house price taking into considerations the variables available for the prediction. Polynomial Regression was chosen to improve the accuracy obtained by the regression since it introduces a higher degree of complexity that better adjusts to the data.

Artificial Neural Network model was considered because it is one of the methods that according to section 2, performs better than previous methods used for house price prediction [3]. LSTM method was chosen since it is one of the latest methods introduced for prediction problems. More over, it has proven to outperform other methods in house price prediction [12].

The Google Colaboratory platform was selected to use for this project [19]. This was chosen because of the advantages it provides such free to use, cloud platform, ability of share work and dedicated GPU. The Colab platform is similar to a Jupyter Notebook and it is been widely adopted given that its purpose is education and research.

Code was written in Python considering the straight forward implementation of most machine learning techniques and that is fully supported by Google Colab. Scikit-learn [20] tools were used for the construction of MLR and Polynomial regression models. The ANN and LSTM models were constructed using Tensorflow and Keras [21]. Plotting was made using Matplotlib and Seaborn [22].

6.2 Data

As we could observe in section 5.4, the different behaviour between the various property types for all boroughs, suggests that a lower error will be achieved in the evaluation if we removed property type "Others" from the data. This analysis allows us to construct the models in consideration of this constraint. Furthermore, as even without "Others" the datasets show a large number of outliers, we will be excluding some of the extreme values below £20,000 and above £30,000,000, which represent around 0.5% of all the instances, to evaluate the improvement in performance.

The correlation tables in Figure 13 show that price is mostly influenced by the date of transfer, so we are assuming a time-series behaviour overall. All the models will be taking this element into account. To safeguard the integrity of the models, we will be constructing the models around both Nominal and Real prices to make sure that inflation is not affecting the performance and results of the analysis.

Finally, to compare the accuracy of all the models, we selected the Root Mean Squared Error as our loss function, since it expresses the average prediction error in units of the variable of interest with negatively oriented scores and is useful when large errors are undesirable. [23]

6.3 MLR Architecture

MLR model is considering the features specified in 5.4, where the independent variables for our model are date of transfer, duration, property type and old/new. The first step is to convert all the categorical variables into dummy values to make sure that they are properly considered into the prediction model. Date of transfer will be converted into ordinal value for this same reason. As mentioned in 5, the models will be constructed for each of the boroughs independently, to analyse each one of them separately.

We separated the general dataset into two different groups, 80 % of the instances randomly selected for training and 20% for testing, repeating the same process for every borough and subsequently, for the different datasets in scope (property types excluding "Others" and excluding outliers).

6.4 PR Architecture

The polynomial regression follows a similar construction as the one specified in the previous section, where the model construction was separated by borough and then, separated into training and testing sets, 80 % and 20% of the instances respectively. A second order (quadratic) function was selected as the best, since it showed more stability during the multiple evaluations for all boroughs and all datasets.

6.5 ANN Architecture

The model architecture consists of one input layer, two hidden layers and the output layer that deliver the prediction of the house price. The input layer dimension is 11 for all data (10 for data without property type "Others"), since the model considers 11 features (10 without property type "Others") as input to predict the house price. The first hidden layer dimension is 20 and the second layer 10. The output layer dimension is 1, the result for a single row with house features is the price prediction. Parameters such as learning rate, optimiser and number of nodes that provide the optimal result for housing prediction was achieved by trial and error. According to previous analysis the theory for choosing the optimal parameters is not specifically defined.[10]

The optimiser chosen was Adam with the following parameters: epochs *epochs = 20*, learning rate *lr = 0.001* decay*decay= lr/epochs = 0.00005*. The Adam optimiser has proven to be robust in regression problems, since it require less tuning of the hyper-parameter n (learning rate) and provide faster convergence. [18]

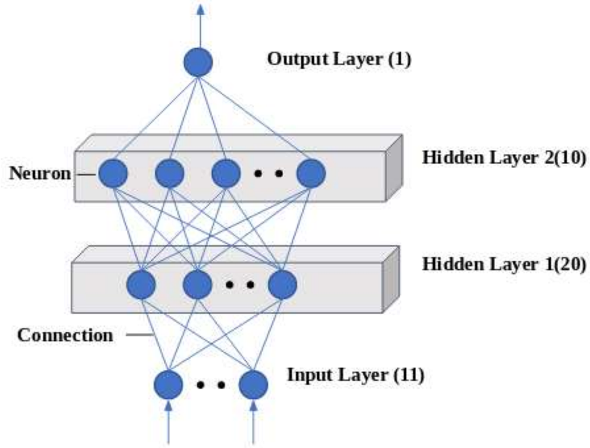
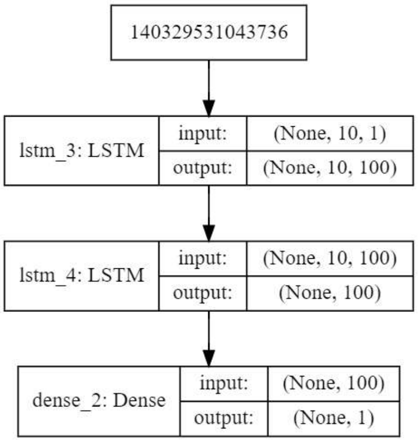


Figure 14: ANN model architecture

The loss function used in the model was the *mean squared error* (MSE), and the optimiser Adam, that presents some benefits over the Stochastic Gradient Descent (SGD) for generalisation such as faster convergence.[18]

6.6 LSTM Architecture

The LSTM architecture contains one input layer, two LSTM layers and the output (dense) layer for the model output. Similar to ANN, the input layer dimension contains 11 features (10 without property type "Others"). There are two LSTM layers in this model and both layers contain 100 hidden neurons. The output layer dimension is 1, containing in a single row of house prices. The loss function chosen was MSE, and the optimiser chosen was Adam. The other parameters are: number of epochs = 20, learning rate = 0.001 (default in Adam) and batch size = 64.



140329531043736

lstm_3: LSTM | input: (None, 10, 1) | output: (None, 10, 100)

lstm_4: LSTM | input: (None, 10, 100) | output: (None, 100)

dense_2: Dense | input: (None, 100) | output: (None, 1)

Figure 15: LSTM model architecture.

7 Results and Evaluation

The results are tabulated in Sections 7.1 and 7.2. Overall, the general trend observed while developing the models is that the larger the standard deviation is for a specific borough, the harder it is to predict the price. This is consistent for all our tests and all our models, where the construct for Westminster was always the worst performer, with Croydon being consistently the best. For both nominal and real prices, excluding property type "Others", and subsequently, excluding the outliers, turned out to be an effective strategy to reduce the RMSE for all three boroughs. We can observe this in the following tables, where the RMSE decreases by more than 50% in all cases after we dropped property type "Others" and more than 60% after we dropped the outliers.

Interestingly enough, we observed in Figure 10 at Section 5.4 that Westminster tends to react more to external changes, such as the price drop in 2009 (global financial crisis) and 2015 (Brexit referendum) while Lambeth and Croydon tend to remain more stable. This factor appears to be affecting the performance of the models as well.

For Westminster, ANN performed better in all cases compared to the rest of the models, showing that this approach has a better performance in cases where the variation of the data is larger. However, while we reduced this variation by excluding the outliers and linearised the trend of the dependent variable (Price), Multiple Linear and Polynomial regression scored better in Croydon, where the trend was relatively linear.

Furthermore, the non-linear relationship between house attributes and house prices, the lack of some environmental attributes of the properties, and inadequate number of sample sizes could be contributing to the poor performance of the linear regression models. However, it should be noted that the optimal artificial neural network model was created by a trial-and error strategy. Without this strategy, results may not outperform that of the neural network model, such as seen in the results of the LSTM model.

7.1 Results with Nominal Prices

Table 3: RMSE for Nominal Prices with All values

Borough	**MLR**	**PR**	**ANN**	**LSTM**
Westminster	£3,902,374	£3,830,236	**£1,456,152**	£4,717,127
Lambeth	£658,481	£654,648	**£322,468**	£799,877
Croydon	£569,279	£565,357	**£209,221**	£558,271

Table 4: RMSE for Nominal Prices without Property Type Others

Borough	**MLR**	**PR**	**ANN**	**LSTM**
Westminster	£1,063,119	£1,040,744	**£964,826**	£2,462,343
Lambeth	£235,969	£229,004	**£209,979**	£239,044
Croydon	£97,919	**£93,946**	£156,798	£180,815

Table 5: RMSE for Nominal Prices without Property Type Others and Outliers

Borough	**MLR**	**PR**	**ANN**	**LSTM**
Westminster	£1,037,774	£1,019,087	£**947,444**	£1,491,830
Lambeth	£214,314	£**207,891**	£215,022	£272,415
Croydon	£94,300	£**89,750**	£158,259	£159,048

7.2 Results with Real Prices

Table 6: RMSE for Real Prices with All values

Borough	**MLR**	**PR**	**ANN**	**LSTM**
Westminster	£2,556,258	£2,509,908	£**1,009,118**	£3,120,706
Lambeth	£440,697	£438,298	£**234,380**	£509,403
Croydon	£379,129	£376,289	£**152,090**	£348,405

Table 7: RMSE for Real Prices without Property Type Others

Borough	**MLR**	**PR**	**ANN**	**LSTM**
Westminster	£742,976	£730,978	£**648,978**	£1,544,915
Lambeth	£166,541	£162,652	£**142,288**	£135,209
Croydon	£72,724	£71,428	£**63,116**	£88,949

Table 8: RMSE for Real Prices without Property Type Others and Outliers

Borough	**MLR**	**PR**	**ANN**	**LSTM**
Westminster	£766,522	£751,250	£**444,339**	£836,363
Lambeth	£168,762	£166,105	£**143,818**	£146,614
Croydon	£69,815	£**68,195**	£93,742	£89,660

8 Conclusions and Discussion

MLR results provided a baseline for house price prediction and RMSE value for each model. The model performance served as a baseline reference for the other models. Performance was marginally improved by applying a PR model with a degree of two. Furthermore, to increase the polynomial order of the model deteriorated the performance resulting in higher RMSE values than MLR.

The ANN model performance was significantly better than MLR and PR models. The performance improvement confirmed that NN models can perform better for regression problems as considered in 2.2. However, the selection of the best hyper-parameters for the model (e.g. *learning rate, epochs*) required additional time to consume during the training. This tendency was early noticed when applying the LSTM model for the prediction. LSTM provided a marginal increase of performance for the price prediction in some cases according to Table 8. However, in most cases the performance of the LSTM was inferior than ANN. The drop of the performance could be explained by the reduced training time used for this model.

In conclusion, different models for house price prediction were successfully implemented. The performance obtained suggests that the ANN models could be applied for house prices prediction with a fair generalisation. One of the limitations encountered in the project was the data availability and relevant features in the data-set. The model construction was limited to a couple of numerical features and thus affecting the accuracy of the regressions. House prediction is affected by many other factors that should be taken into consideration for the price prediction. External factors like the financial crisis and Brexit news make the prediction harder to model. Another limitation was the training time of the model, since the data-set was a significant sample of the complete data, the time required for training and the number of possible combinations of hyper-parameters made the model training more time consuming. In comparison, some papers focus only on a small sample of a data-set, this simplification could improve the model performance.

9 Recommended Future Work

The main limitation with this work was that due to the nature of the dataset, property-specific parameters could not be obtained (e.g. square footage, number of bedrooms). As suggested by prior work denoted in Section 2, these parameters could significantly improve the model performance. For future work, it is suggested that more complex models are to be explored, with more suitably adjusted hyper-parameters (e.g. loss function, activation function, learning rates for ANN and LSTM). More features can also be used for the MLR and PR models to achieve better results. As alluded to in Section 7, trial-and-error with more exhaustive parameter search could be conducted with the LSTM model in order to achieve better performance.

References

[1] Jagjit S. Chadha. Commentary: The housing market and the macroeconomy. *National Institute Economic Review*, 243(1):F4–F9, 2018.

[2] Ministry of Housing Communities & Local Government. *Analysis of the determinants of house price changes.* 13 Apr 2018.

[3] Nur Khalidah Khalilah Binti Kamarudin Azme Bin Khamis. Comparative study on estimate house price using statistical and neural network model. *INTERNATIONAL JOURNAL OF SCIENTIFIC TECHNOLOGY RESEARCH*, 3, 12 2014.

[4] A. K. Bhalla. Housing finance in india: Development, growth and policy implications. *PCMA Journal of Business*, 1(1):51–63, 2008.

[5] Kanojia Anita. Valuation of residential properties by hedonic pricing method-a state of art. *International Journal of Recent Advances in Engineering Technology*, 2016.

[6] Jordan De Silva. An analysis of economic factors affecting first-time buyers; a case of the uk housing market. *Cardiff Metropolitan University*, BA Dissertation, 2017.

[7] Eric Slone, Haitian Sun, and Po-Hsiang Wang. Market prices of houses in atlanta. 2014.

[8] Elli Pagourtzi, Vassilis Assimakopoulos, Thomas Hatzichristos, and Nick French. Real estate appraisal: a review of valuation methods. *Journal of Property Investment & Finance*, 21(4):383–401, 2003.

[9] Chris Brooks and Sotiris Tsolacos. Real estate modelling and forecasting. 2010.

[10] Visit Limsombunchai, Christopher Gan, and Minsoo Lee. House price prediction: Hedonic price model vs. artificial neural network. *American Journal of Applied Sciences*, 1, 03 2004.

[11] Julio Gallego Mora-Esperanza. La inteligencia artificial aplicada a la valoracin de inmuebles un ejemplo para valorar madrid. *JSSN 1138-3488*, 50:51–68, 2004.

[12] Xiaochen Chen, Lai Wei, and Jiaxin Xu. House price prediction using lstm. *CoRR*, abs/1709.08432, 2017.

[13] Douglas C. Montgomery. *Introduction to Linear Regression Analysis.* John Wiley Sons, Incorporated, 2013.

[14] Carroll Croarkin and Paul Tobias. *Engineering Statistics Handbook.* Apr 2013.

[15] James Coakley and C.E. Brown. Artificial neural networks in accounting and finance: Modeling issues. *International Journal of Intelligent Systems in Accounting*, 9, 06 2000.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[17] HM Land Registry. Price Paid Data. `https://www.gov.uk/guidance/about-the-price-paid-data`, 14 Aug 2014. [Online; accessed Feb-2019].

[18] Aurlien Gron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow.* O'Reilly, 2017.

[19] Google. Colaboratory. `https://colab.research.google.com`, 2018. [Online; accessed Mar-2019].

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] François Chollet et al. Keras. `https://keras.io`, 2015.

[22] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

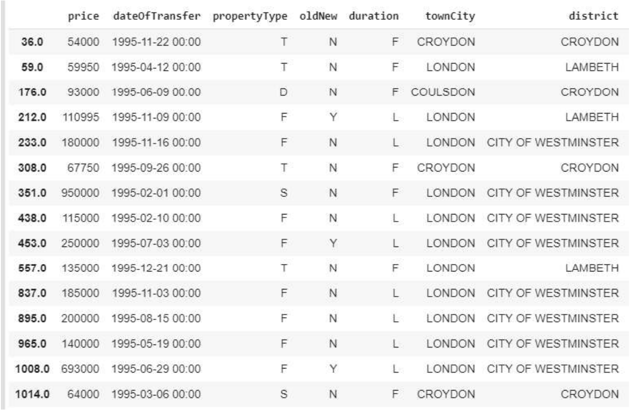
[23] JJ. MAE and RMSE. `https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d`, 23 March 2016. [Online; accessed Mar-2019].

Appendix

A Dataset

A.1 Data Sample

In Figure 16 a sample of data used across the analysis is observed.



	price	dateOfTransfer	propertyType	oldNew	duration	townCity	district
36.0	54000	1995-11-22 00:00	T	N	F	CROYDON	CROYDON
59.0	59950	1995-04-12 00:00	T	N	F	LONDON	LAMBETH
176.0	93000	1995-06-09 00:00	D	N	F	COULSDON	CROYDON
212.0	110995	1995-11-09 00:00	F	Y	L	LONDON	LAMBETH
233.0	180000	1995-11-16 00:00	F	N	L	LONDON	CITY OF WESTMINSTER
308.0	67750	1995-09-26 00:00	T	N	F	CROYDON	CROYDON
351.0	950000	1995-02-01 00:00	S	N	F	LONDON	CITY OF WESTMINSTER
438.0	115000	1995-02-10 00:00	F	N	L	LONDON	CITY OF WESTMINSTER
453.0	250000	1995-07-03 00:00	F	Y	L	LONDON	CITY OF WESTMINSTER
557.0	135000	1995-12-21 00:00	T	N	F	LONDON	LAMBETH
837.0	185000	1995-11-03 00:00	F	N	L	LONDON	CITY OF WESTMINSTER
895.0	200000	1995-08-15 00:00	F	N	L	LONDON	CITY OF WESTMINSTER
965.0	140000	1995-05-19 00:00	F	N	L	LONDON	CITY OF WESTMINSTER
1008.0	693000	1995-06-29 00:00	F	Y	L	LONDON	CITY OF WESTMINSTER
1014.0	64000	1995-03-06 00:00	S	N	F	CROYDON	CROYDON

Figure 16: Data sample.

A.2 Pointers to Data

HM Land Registry URLs to download price paid data single file in CSV format, which includes price paid data transactions received at HM Land Registry from 1 January 1995 to the most current monthly data.

Reference: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloadssingle-file

Actual CSV file (3.7GB): http://prod.publicdata.landregistry.gov.uk.s3-website-eu-west-1.amazonaws.com/pp-complete.csv

B Programming Code

B.1 Preliminary analysis and plotting

```
import pandas as pd
from datetime import datetime
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats

import google
google.colab.drive.mount('/content/drive')
```