ESTIMATING UK HOUSE PRICES USING MACHINE LEARNING

ADEBAYOSOYE ABDULFATAI AWONAIKE

A thesis submitted in partial fulfilment of the requirements of the University of East London for the degree of Professional Doctorate in Data Science

January 2022

Abstract

House price estimation is an important subject for property owners, property developers, investors and buyers. It has featured in many academic research papers and some government and commercial reports. The price of a house may vary depending on several features including geographic location, tenure, age, type, size, market, etc. Existing studies have largely focused on applying single or multiple machine learning techniques to single or groups of datasets to identify the best performing algorithms, models and/or most important predictors, but this paper proposes a cumulative layering approach to what it describes as a Multi-feature House Price Estimation (MfHPE) framework. The MfHPE is a process-oriented, data-driven and machine learning based framework that does not just identify the best performing algorithms or features that drive the accuracy of models but also exploits a cumulative multi-feature layering approach to creating machine learning models, optimising and evaluating them so as to produce tangible insights that enable the decision-making process for stakeholders within the housing ecosystem for a more realistic estimation of house prices. Fundamentally, the MfHPE framework development leverages the Design Science Research Methodology (DSRM) and HM Land Registry's Price Paid Data is ingested as the base transactions data. 1.1 million London-based transaction records between January 2011 and December 2020 have been exploited for model design, optimisation and evaluation, while 84,051 2021 transactions have been used for model validation. With the capacity for updates to existing datasets and the introduction of new datasets and algorithms, the proposed framework has also leveraged a range of neighbourhood and macroeconomic features including the location of rail stations, supermarkets, bus stops, inflation rate, GDP, employment rate, Consumer Price Index (CPIH) and unemployment rate to explore their impact on the estimation of house prices and their influence on the behaviours of machine learning algorithms. Five machine learning algorithms have been exploited and three evaluation metrics have been used. Results show that the layered introduction of new variety of features in multiple tiers led to improved performance in 50% of models, a change in the best

performing models as new variety of features are introduced, and that the choice of evaluation metrics should not just be based on technical problem types but on three components: (i) critical business objectives or project goals; (ii) variety of features; and (iii) machine learning algorithms.

Declaration

I declare that this written submission represents my ideas in my own words, and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented, fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the university and can also evoke penal action from any sources which have not been properly cited or from whom proper permission has not been sought where needed.

Adebayosoye Awonaike
U1440387
Date: 21 January 2022

Table of Contents

List of Figures

List of Tables

Acknowledgements

Isaac Newton said, *"If I have seen further than others, it is by standing on the shoulders of giants".* Therefore, I am immensely grateful to the Almighty God for inspiring and enabling me to start and complete this programme. Furthermore, my deep appreciation goes to Funke (my wife), Kiitan (my son) and Kiishi (my daughter) for the individual and collective roles they played in ensuring I got over the finish line. I also couldn't under-estimate the contributions of the stewards at The Hope Church East Grinstead for how they covered for me individually and collectively during the final stages of my research. Finally, my appreciation goes to my supervisors, friends and mentors that God placed in my life for a time like this; without you this achievement would have been impossible.

Abbreviations

ARIMA – AutoRegressive Integrated Moving Average

CAMA – Computer Assisted Mass Appraisal

DSL – Domain Specific Language

FDI – Foreign Direct Investment

GDP – Gross Domestic Product

GMM – Generalised Method of Moments

MAE – Mean Absolute Error

MfHPE – Multi-feature House Price Estimation

MLR – Multiple Linear regression

MRR – Material Removal Rate

MSE – Mean Square Error

NARX - Nonlinear Autoregressive with Exogenous Variables model

NDCG – Normalised Discounted Cumulative Gain

NSPL – National Statistics Postcode Lookup

OLS – Ordinary Least Squares regression

ONS – Office of National Statistics

PCA – Principal Component Analysis

RF – Random Forest

RMSE – Root Mean Square Error

RNN – Recurrent Neural Network

ROC – Receiver Operating Characteristic

ROI – Return on Investment

SVR - Support Vector Regression

TB – Tuberculosis

WEO – World Economic Outlook

Chapter 1: Introduction

1.1 Background and motivation

Residential housing is a significant part of human existence, as everyone would want to be able to go to a place they can call 'home'. This can be after work, travel or just some time being away from the place called 'home'. For most individuals and/or families, their house is also their home, and this place called home could be in the city (urban) or the country (suburban), depending on the preferences of the individual or family. Furthermore, these places range in size, from a room in a shared house to a mansion with multiple rooms, a garden, a swimming pool and all sorts of features. Depending on the location, size, age and a range of factors, including economic factors, the value of houses vary and also change over time.

As a home owner myself, and as with most of my peers, the reasons most people buy a house are usually a combination of two factors, being (i) their lifestyle, as well as (ii) the fact that it is seen as an investment. However, the investment angle seems to be driven by certain credence associated with house prices. According to Monnery (2011) these can be summarised as follows:

- Medium to long term, house prices increase
- A fall in house prices is an opportunity to get on the ladder, as they are brief
- Houses are good investments because house prices grow, well ahead of inflation rates

More so, in a rapidly changing world and with the measure of uncertainty on what the new world will look like after a global pandemic like Covid-19, it has become relatively safe to wonder if the credence associate with house prices as stated above will stay true in the future as it has over the past decades.

With the average income in the United Kingdom currently around £38,600 for people in full-time jobs and £13,803 for people in part-time jobs (Office of National Statistics, 2021a), the estimated earnings after tax based on the gov.uk tax service calculator is £29,889 and £13,027 respectively. Based on these figures,

a 40-year working career for either of these working groups will generate an estimated £1,195,592 and £521,113 respectively of lifetime after-tax income in today's money.

However, despite the fall in house prices since the first quarter of 2020 by 0.2%, the housing market has experienced an annual price rise of 2.1%. As a result, the average residential property in the United Kingdom is valued at £231,885 (HM Land Registry (2020). This shows that a significant part of our working life is committed to creating the wealth required to own a home. Owning a home will probably be more realistic for an individual earning around the average full-time income stated above compared with a part-time income earner. It follows that individuals with an annual income around the average part-time income are unlikely to be home owners, considering the current value of the average property in the UK. These individuals are therefore most likely to be renters of homes owned by private landlords and other investors in the property market.

Table 0.1: Average price by property type for England Source: HM Land Registry (2020)

Property Type	March 2020	March 2019	Difference %
Detached	£379,050	£369,683	2.5
Semi-detached	£232,901	£228,288	2.0
Terraced	£199,959	£195,955	2.0
Flat/Maisonette	£226,383	£221,555	2.2
All	£248,271	£242,982	2.2

Based on these facts, this research will explore the estimation of prices and how this affects or influences the behaviour of a range of stakeholders in the property market including investors, developers, landlords and tenants.

1.2 The research problem

Across the cities of the UK, the determination of house price is as much an art as it is a science. This is a result of the combination of factors considered for house price estimation. These include tangibles, such as floor area, number of bedrooms, year built or air quality, and intangibles, such as greenery, scenic views or prestige. Also, the price of a house is influenced by other factors such as roads, access to social amenities and its neighbourhood. Furthermore, the estimation of house price is dependent on broader macroeconomics that are prevailing in the economy, such as liquidity, level of income, cost of capital,

inflation and GDP. Most of these factors have been discussed in existing literature (see Chapter 2), with a focus on how they influence the final price of a house using machine learning. The authors who have written on the subject of this research have exploited multiple approaches including (i) a focus on a few factors and a single ML algorithm, thereby creating a single model, or (ii) a few factors and multiple ML algorithms, thereby creating multiple models with different combination of factors.

These approaches have provided insights on how different ML algorithms perform, how algorithms compare with one another, and what factors may have a positive, negative or neutral impact on the estimation of house prices. The challenge is that in a real-life scenario and based on geographic location, all or some of the various parameters used in machine learning models exist in a static or changing continuum. Parameters like house physical properties or features and neighbourhood amenities mostly exist in a static continuum, while features like economic metrics vary every month or quarter in a changing continuum. Therefore, it becomes imperative that all known factors or parameters are considered through the creation of a framework in which multiple new or existing machine algorithms can be exploited on the various parameters that actually co-exist in real life for the estimation of house prices.

In this thesis, the terms 'factors' and 'parameters' have been used interchangeably to represent the variables of the datasets exploited. Sections 2.2 and 2.3 provide a broad overview of the variables/attributes of 'factors' and machine learning respectively, so as to provide some context to anyone new to data and machine learning.

1.3 The research hypothesis and questions

The research hypothesis for this thesis states:

*'Deploying **standalone** and **ensemble** Machine Learning (ML) algorithms on publicly available data can create a deeper understanding of how different algorithms perform based on variation in datasets and also influence the behaviour of a range of **stakeholders** through the estimation of house prices.'*

To give some clarity on some of the different components of the research hypothesis stated above based on the context of this research; *First*, standalone machine learning algorithms are simply those that are not in a group, while ensemble machine learning algorithms means a group of machine learning algorithms. This is being considered because ensembles may give a boost in accuracy on the range of datasets this thesis will seek to harness. In this thesis, the standalone algorithms exploited are **Random Forest** – for (i) strong performance (ii) convenience in the handling of categorical data with many levels (iii) adequately works with missing data (iv) allows for nonlinear and unsteadiness of variables (v) doesn't require detailed model specification, **Light Gradient Boosting Machine** - for accuracy, efficiency and cost, and **Extreme Gradient Boost** - for scalability, while the ensemble machine learning algorithms used are **Hybrid Regression** and **Stacked Generalisation**. *Second*, the stakeholders mean individuals or groups who are relevant to the housing market. In the context of this thesis, the stakeholders are investors, developers, landlords and tenants. The investors can be individuals or corporate entities who invest their capital in residential properties with a focus on ROI. The developers can also be individuals or corporate entities who build residential housing. The landlords are individuals or corporate entities who either buy or build houses with the purpose of renting to tenants. The tenants are strictly individuals who rent the house they reside in from individual or corporate landlords, but not from local authorities or housing associations.

A detailed review of existing literature has been documented in Chapter 3 for a thorough understanding of the benefits and limitations of existing approaches to the estimation of house prices. Ultimately, the review backed up by the framework design documented in Chapter 4 is expected to showcase a robust, data-driven, machine learning enabled framework that produces relevant insights that should inform the behaviour of a range of stakeholders. However, owing to the complexities around house prices, it is impossible to take an affirmative position that the output of this thesis will be applicable for all possible scenarios. As a result, below are the research questions in view:

Research Question 1: What data-led methods have been used for estimating house prices?

Research Question 2: What are the house characteristics, neighbourhood factors, macroeconomic indicators and other factors that influence the value of house prices?

Research Question 3: Can machine learning be used to understand the influence different groups of factors have on the estimation of house prices?

Research Question 4: What evaluation approaches exist in this industry and how will this research work be evaluated?

Research Question 5: Can multiple datapoints be integrated so as to improve the accuracy of house price estimation?

Research Question 6: What is the impact data volume and variety on the accuracy of house price estimation?

Research Question 7: How do different machine learning algorithms respond to changes due to data variety?

1.4 Research contributions

The primary contribution of this research is the minimisation of the research gap identified in the review of existing literature in Chapter 3 by creating a near real-life scenario where all the features possibly relevant to house price estimation can be introduced into a machine learning enabled framework. This will be done by taking a cumulative multi-feature layering approach to (i) improving the accuracy of house price estimation and the machine learning algorithms, (ii) understanding the impact of the introduction of a variety of features have on the behaviour of algorithms, and (iii) examining the overall effect of insights produced on stakeholders in the housing market. The research as documented in this thesis has explored a process-oriented, data-driven and machine learning enabled approach to the development of a framework. A summary of the main research contributions is as follows:

I. The modularised **Multi-feature House Prices Estimation (MfHPE) Framework**: This is the main design output of this research based on design science methodology. It is described as '**modularised**' because it is made up of nine different modules, explained in Section 4.6, and '**multi-feature**' because it leverages over 25 features from ten datasets from multiple sources. These features are then grouped into three different tiers, as discussed in Chapter 4, and these tiers form the basis for the novel layering approach taken in the development of the modules. The modularity of the MfHPE framework makes it robust and enables (i) updates to existing datasets, (ii) introduction of new datasets, and (iii) exploitation of other interesting machine learning algorithms.

II. The **Cumulative Multi-feature Layering** of groups of multiple parameters throughout the model development: An existing study used geo-data from multiple sources and machine learning to understand the appreciation of house prices. Rather than building a machine learning model with all the datasets, this study created six different models which included the baseline data, then a combination of baseline and house photos, baseline and street view, baseline and mobility data, baseline and socioeconomic data and then all data sources. However, the research documented in this thesis has explored a **cumulative multi-feature layering** approach comprising three groups of parameters: the baseline transaction data, followed by a layer of neighbourhood data, and a layer of macroeconomic datasets. This is unlike the approach prior studies on this subject have followed, hence the novelty of the modularised MfHPE framework. In this thesis, the MfHPE framework creates a total of 48 models that exploit five different machine learning algorithms by introducing layers of new groups of data, as discussed in Chapter 4. Firstly, it is a 'layering' framework because groups of features are introduced as new layers into the framework. Secondly, it is described as a 'multi-feature' approach because the framework has leveraged ten datasets (parameters) from multiple sources and has the capacity to have more introduced by design. Thirdly,

it is described as ‘cumulative’ because the layering approach allows the introduction of new layers without the removal of existing layers in each model, thereby creating a near real-life scenario where all relevant features co-exist.

III. The **Research Dataset:** A UCL working paper series titled ‘Creating a new dataset to analyse house prices in England’ described the HM Land Registry Price Paid Data as ‘the official house price dataset in England’. They created a geocoded version of the Price Paid Data by linking it with OS MasterMap and OS AddressBase Plus. However, in this thesis, the Price Paid Data is geo-coded by blending it with the ONS NSPL product. The variables of the geo-coded HM Land Registry Price Paid Data are used to create new variables that enable a further data-blend with neighbourhood and macroeconomic datasets to create the complete research dataset. The process for the creation is discussed extensively in Chapter 4. This research dataset created is therefore an addition to knowledge as it is comprised of price paid transaction data for London boroughs published by HM Land Registry blended with ONS NSPL product being Tier 1 then with bus stops, retail locations, national rail and underground stations being Tier 2 features, and then macroeconomic indicators including GDP, inflation rate, employment rate, unemployment rate and consumer price index being Tier 3 features, as shown in Figure 4.2.

IV. **Response** of **Machine Learning Algorithms** to changes in **Data Variety**: Within the context of predicting or estimating house prices, multiple research papers have exploited multiple machine learning algorithms on a variety of datasets. Their focus has ranged from comparison between algorithms to model accuracy, algorithm performance, predictors and model explainability. However, the cumulative multi-feature layering approach explored in this research unveiled how machine learning algorithms respond to a changing landscape of features as multiple tiers

of features were introduced into the framework. The detail of this contribution is further discussed in Chapters 6 and 7.

V. **Evaluation Metrics respond differently to Features and Machine Learning Algorithms:** As discussed in Section 3.9 of the literature review, the choice of evaluation metrics is usually driven by the type of machine learning problem. For example, Classification Problem (F1-Score, ROC, Precision), Regression Problem (RMSE, MAE, MSE), Ranking Problem (NDCG, MRR), and Statistical Problem (Correlation). However, beyond these, the choice of evaluation could be influenced by a blend of business problems or project goals, and the variety of features and machine learning algorithms. The analysis and evaluation of the validation data in this thesis, as detailed in Section 6.4, proposes that this tripartite view to choosing an evaluation metric is more likely to provide the best-fit insights that enable decision making.

1.5 Research publications

The papers shown below are a product of the work done within the context of this research:

AWONAIKE, A., GHORASHI, S. A. & HAMMAAD, R. 2021. A Machine Learning Framework for House Price Estimation. International Conference on Intelligent Systems Design and Application. Online.

AWONAIKE, A., GHORASHI, S. A. & HAMMAD, R. 2022. Machine Leaning based Cumulative Multi-feature Layering Framework for House Price Estimation. Expert Systems with Applications. [Status: Submitted]

1.6 Upcoming publications

AWONAIKE, A., GHORASHI, S. A. & HAMMAD, R. 2022. Creating an Enriched Dataset for House Price Estimation in England and Wales. Expert Systems with Applications. [Status: Manuscript]

AWONAIKE, A., GHORASHI, S. A. & HAMMAD, R. 2022. Impact of Data Variety on Machine Learning Algorithms. Expert Systems with Applications. [Status: Manuscript]

1.7 Presentations

The research covered by this thesis has been presented and discussed extensively with peers in the industry and academia through both organised conferences and workshops, as follows:
AWONAIKE, A. 2021. A Machine Learning approach to Estimating UK House Prices being a Macroeconomic Indicator. School of Architecture, Computing & Engineering Postgraduate Research Conference, University of East London.
AWONAIKE, A. 2021. Using Machine Learning for the Estimation of House Prices. Get Curious about Machine Learning, Legal and General, London.

1.8 Thesis outline

Chapter 2 captures a range of background information that is expected to provide some additional information on the concepts exploited in the thesis. In Chapter 3, the rationale and motivation of this thesis is supported with a review of existing literature on (i) London's housing market, (ii) the factors that influence house prices, (iii) existing forecasting models, (iv) the use of machine learning for the estimation of house prices, (v) the value of generating insights from multiple data points, (vi) a review of the evaluation approaches that have been explored for similar researches so as to define what evaluation method to be used for this research, (vii) a possible correlation between house prices and rental cost, and (viii) some clarity on the research gap this thesis fills.

Next, Chapter 4 captures the fundamentals of the research framework design. This is a **process-based**, **modularised**, **data-driven** and **machine learning enabled framework** designed to estimate house prices by layering known

factors that influence house prices, and the research focus is to estimate the price for existing houses in London. The chapter provides (i) an overview of the research framework design approach used, (ii) a detailed overview of the profile of each dataset exploited, (iii) an explanation of the concept of data modelling and how the relationship between the range of datasets is exploited, (iv) a showcase of the significance of creating pipelines for the development of a robust solution, (v) an overview of modular programming and a detailed view of the modules that make up the MfHPE framework, (vi) an in-depth view of the initial exploration data analysis which gives an understanding of the trends, patterns and quality of the research data, (vii) details on the engineering of features or data attributes in preparation for modelling, and (viii) the actual modelling simulations exploring an ensemble of machine learning algorithms. The results, from modelling performed, as well as evaluation and optimisation, are presented in Chapter 5 and discussed in detail in Chapter 6, while Chapter 7 focuses on the conclusion to the research topic, with recommendations for further research.

1.9 Conclusion

This chapter has given insights into the motivation behind this research and, consequently, the presentation of the research problem. Seven research questions were raised in the quest to prove or disprove the research hypothesis. Five proposed contributions to knowledge were presented as well as reference to presentations, published papers and unpublished papers that have been inspired by this research. In conclusion, the chapter was wrapped up with an outline of what to expect in the remaining chapters of the thesis.

Chapter 2: Background

2.1 Introduction

This chapter, an extension to chapter 1 outlines and discusses fundamental concepts that are potentially relevant for the understanding of non-technical readers of this thesis focused on the development of a framework for the estimation of house prices by applying machine learning techniques to a range of data from multiple sources. These data are made up of multiple variables (later referred to in this thesis as features). Therefore, the chapter highlights the different types of variables that have been exploited in this and existing research, what machine learning is and the various types of machine techniques that exist.

2.2 Understanding variables

The variables used in machine learning frameworks differ depending on the demands of the algorithm. Data variables are either independent (inputs) or dependent (outputs). Unsupervised machine learning models take into account the inputs (independent variables) of data to establish the patterns and develop a forecast, while supervised machine learning models apply both inputs and outputs to classify data (Mali et al., 2021). House prices are affected by a lot of factors, and the effective development of a good machine learning system is dependent on the variables used. Previous research has identified the main variables used by the majority of house estimation machine learning models as follows:

2.2.1 Geographical and non-geographical variables

Geographical variables include neighbourhood features such as distance from the nearest school, the quality of the schools nearby and the distance to the nearest city centre. Non-geographical variables take into account the intrinsically descriptive house characteristics which are unique to every house, such as the number of rooms, total floor area, number of living rooms, number of bathrooms,

outdoor spaces and lots more (Gao et al., 2019). Machine learning models that exploit geographical variables show that the value of a house is dependent on the amenities within its neighbourhood; therefore, the location and neighbourhood of a house have an effect on its expected price. Kuvalekar et al. (2020) stated that systems that use real-time neighbourhood data can be efficient in developing precise real-world valuations of houses. Koktashev et al. (2019) combined both geographical and non-geographical features in their machine learning system designed to estimate the value of 1,970 houses in Krasnoyarsk.

2.2.2 Visual and non-visual variables

Studies have identified potential relationships between the non-visual attributes of a house with its visual appearance. Arietta et al. (2014) designed a non-linear Support Vector Regression (SVR) system for the identification and validation of predictive relationships between the visual appearance of the geographic area where houses are located and the corresponding non-visual variables such as crime statistics, danger perception and population density. They found the existence of a predictive relationship which ultimately can influence the value of houses in a location-specific way.

2.2.3 Comparing sales variables

These models predict the price of a house depending on sales variables such as previous sale price, the prices of similar houses (either in the same or similar geographic domains), budgets and priorities of prospective buyers, any prevalent economic issues and lots more.
(Kim et al., 2020) designed a procedure based on Comparable Sales Method (CSM) where the price of a house is assessed on sales of houses that are comparable to it. The criteria for the assumption of comparability are twofold: firstly, houses that are located near each other have similar price volatility; secondly, houses that have shown similar prices in the past have the same price volatility.

2.3 What is machine learning?

The term machine learning means the ability given to machines to learn without having to be programmed explicitly. (Mohammed et al., 2020) describe machine learning as a natural outgrowth originating from an intersection of statistics and computer science with the intention of getting computers to program themselves from experience and to make conclusions inferred from datasets presented to them. As the models are exposed to more new data, they can adapt independently and reproduce reliable and repeatable results, thus aiding in decision making.

The objective of machine learning is for machines to perform clustering, make predictions, derive associations and make decisions from a given dataset. As a subfield of artificial intelligence, machine learning is closely related to data mining in the sense that they deal with the discovery of new interesting patterns from large data sets. The key difference between the two, however, is that data mining focuses on the discovery of implicit knowledge and regularities in data, whereas machine learning concentrates more on operational use and adaptive behaviour (Taranto-Vera et al., 2021). The development of machine learning has seen its application in many fields such as cognitive computing, image processing, knowledge representation, pattern recognition, gene function prediction, house price prediction and so forth.

There are four general machine learning methods: unsupervised machine learning, supervised machine learning, semi-supervised machine learning and reinforcement machine learning. These will further be used in explaining the profile of the machine learning powered framework which is the main output of this thesis.

2.3.1 Supervised machine learning

Supervised learning gives the machine the ability to learn human or object behaviour and to use the new knowledge to perform similar tasks. Supervised algorithms are predetermined by human action whereby classes are created from a finite set. In this method, the model is presented with sample inputs and some

of the desired outputs with the goal of learning the general rules that map inputs into outputs. If the output has a finite set of values, the inputs are mapped using classification; when the outputs take continuous values, regression is applied (Liu and Wu, 2012).

2.3.2 Unsupervised machine learning

In this method, the model is not presented with sample inputs; no labels are associated with the algorithm. The goal of unsupervised learning is for the model to establish structures from the input. This approach is therefore applied to discover some of the hidden patterns in the data. Cluster analysis is used in unsupervised learning to segment datasets that share attributes and examine the algorithmic relationships originating from them. Standard k-means is the most used partitional clustering algorithm, which employs an iterative relocation scheme to produce k-way clustering and reduce distortion between data objects (Greene et al., 2008).

2.3.3 Semi-supervised machine learning

Semi-supervised machine learning models combine both the labelled and unlabelled examples to come up with appropriate classifiers and functions (Nasteski 2017). The major objective of semi-supervised machine learning is overcoming the drawbacks that exist between unsupervised and supervised machine learning. The machine is provided with supervision information that does not fit all the examples. An imperative extension of transduction where the entire set of problem instances are known at the time of learning only reveals that the targets are missing; there is a labelled training set and an unlabelled test set (Reddy et al., 2018). Transduction tries to predict newer outputs on the basis of training outputs, training inputs and new inputs.

2.3.4 Reinforcement machine learning

In reinforcement learning, the inputs and outputs are not introduced to the learning system; however, the system receives a reward for every action it performs with the goal to maximise the cumulative reward from all the processes (Lawrynowicz and Tresp, 2014). To this end, the computer interacts with a dynamic environment whereby it has to perform a certain goal without knowing the specifics of the goal. The algorithm learns the policies by observing the impact of every action on the environment and the environment relays feedback which is critical in guiding the algorithm further (Nasteski 2017).

2.3.5 Batch and online learning

Further categorisation of machine learning can be done by observing the way the model accesses its databases over time. Access to data is a critical aspect of machine learning. There are two main design choices available for selection in creating a modelling pipeline: batch learning and online learning. In batch learning, the system cannot learn incrementally, it must be trained using all the available data. As such, the model is built while the model is at rest. A batch learning algorithm builds a statistical assumption over the product space of X*Y whereby the batch learning algorithm is expected to generalise that the output hypothesis predicts labels 'Y' on unseen examples of 'X' from the distribution (Burlutskiy et al., 2016). On the other hand, online learning entails training a system incrementally through feeding data sequentially. In the online prediction model, the learner operates on a sequence of data entries where the learner receives an example in d-dimensional feature space (Burlutskiy et al., 2016). This can be done either in small groups (mini-batches) or individually, therefore enabling the system to receive data as a continuous flow. Online learning is efficient for predicting big data, since once the data is consumed there is no need to store the data. Additionally, the system does not make assumptions through identification of the data distribution; rather, as the data changes, the model adapts to keep with the trends.

2.3.6 Instance-based and model-based learning

Extensive machine learning categorisation can be done by grouping machine learning models as either instance-based or model-based. Instance-based learning extends the techniques of classification and regression, which produce a prediction on the basis of the similarity of the query to its nearest neighbours in the training set. This model stores data and derives answers to queries from the examination of the nearest neighbours of the model. The models do not perform explicit generalisation, but rather compare queries with instances seen in training (Shaier, 2019). Some of the approaches of instance-based learning include: K-nearest neighbour (KNN), Learning Vector Quantisation (LVQ), Self-Organising Map (SOM) and Locally Weighted Learning (LWL). On the other hand, model-based learning is an approach based on the combination of a model with an inference model. All the assumptions about a problem, and the specific queries, are made explicit through the creation of a set of assumptions in a precise mathematical form. The idea of this approach is to create a machine learning model tailored for each new application (Bishop, 2013).

2.3.7 Typical machine learning workflow

A ML workflow is the collection of different phases that are implemented when undertaking a machine learning project. Figure 2.1 presents a simple machine learning process, which will be explained further in Chapter 3.

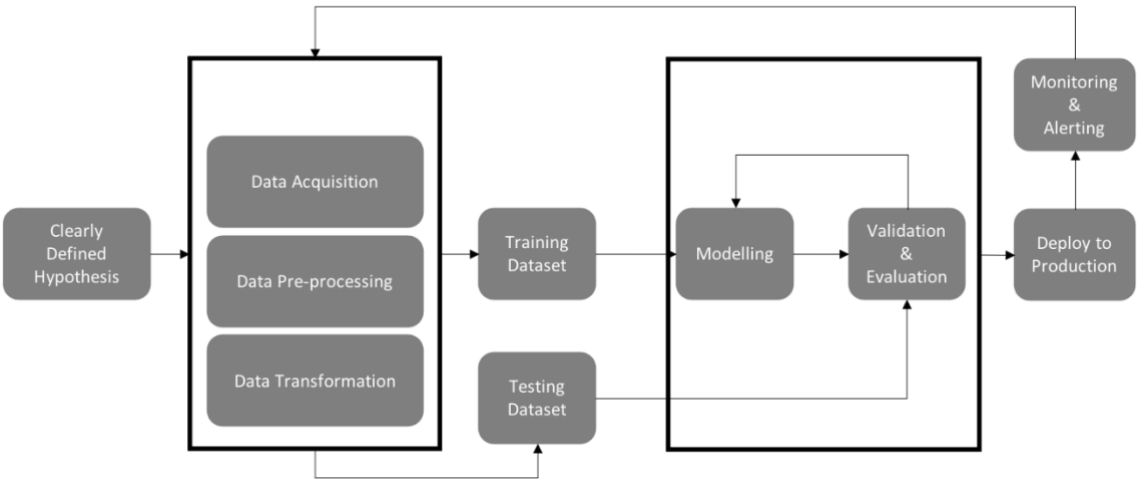


Figure 2.1: High-level machine learning workflow

2.4 Conclusion

Since machine learning is about algorithms learning from historic data to predict or estimate present and future occurrences or values, chapter 4 of this thesis provides details of the data, features and machine learning algorithms used. Based on the overview of the four machine learning methods described above and the categorisation of models based on access to data, the framework produced by this thesis can be described as a ***batch***, ***model-based*****,** ***unsupervised learning*** **system** for estimating house prices and with a focus on London, United Kingdom.

Chapter 3: Literature Review

3.1 Introduction

Further to the introductory chapter, this thesis focuses on estimating the performance of a macroeconomic indicator – house prices – by using an ensemble of machine learning algorithms and data. Park and Bae (2014) stated that developing predictive models for house price estimation could also assist in the establishment of relevant policies beyond regular expectations of predicting future house prices. The outputs of this thesis are expected to guide the behaviour of a range of relevant stakeholders.

This review of literature presents the existing research on (i) London's housing market, (ii) data-led methods/approaches that have been exploited for the estimation of a macroeconomic indicator, house prices, and (iii) the factors that influence the value of house prices. It will then take a dive into exploring (iv) machine learning techniques that have been exploited to estimate the value of house prices, (v) layering of multiple parameters for the creation of relevant insights, (vi) possible correlation between estimated house prices and cost of rental, and whether the cost of rental can improve the estimation of house prices, (vii) the possible impact of data volume and variety on the accuracy of house price estimation, and finally (viii) an identification of evaluation approaches that exist in this industry and how this research work will be evaluated.

3.2 Literature review approach

The literature review is approached as presented in Figure 3.1, as an in-depth look into the existing literature on various factors that are known to influence the estimation or prediction of house prices. This includes property characteristics, neighbourhood amenities and macroeconomic indicators. These factors are presented in the literature with the aim of highlighting their impact and/or performance trends and what they mean for policy makers and professionals in the housing market. For macroeconomic indicators, the methodologies used for

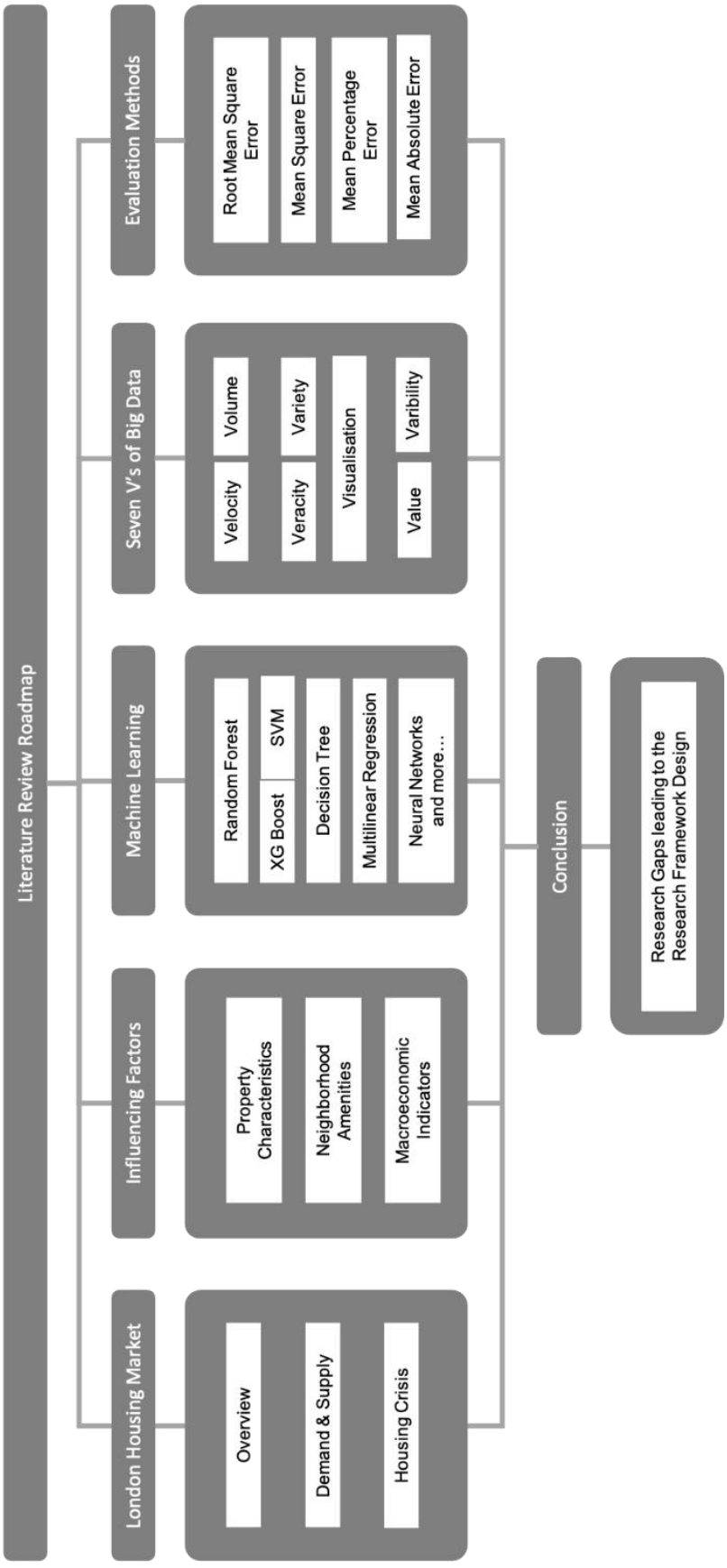


Figure 3.1: Literature review approach

forecasting the performance of the different indicators is presented, including the estimation of house prices. Then, a background on machine learning algorithms and their use in estimating housing prices is presented as a major part of the literature review, with a focus on what they have explored and their rationale. By assessing the literature on housing prices in the UK and their estimation using traditional forecasting approaches and machine learning, the literature review provides a basis for identifying research gaps to be addressed in this thesis.

3.3 London housing market – an overview

London is a Metropolis in the United Kingdom (UK) of approximately 9 million people (Greater London Authority, 2020). Its people need a place to retire to after work – a house or home – and house prices have steadily increased over the years from £250,000 in the aftermath of the 2008 financial crisis to reach an average price of £500,000 in April 2021. This price is out of reach for most Londoners, especially the low-income earners. Figure 3.2 shows the trend of average house prices across the UK in comparison to London.

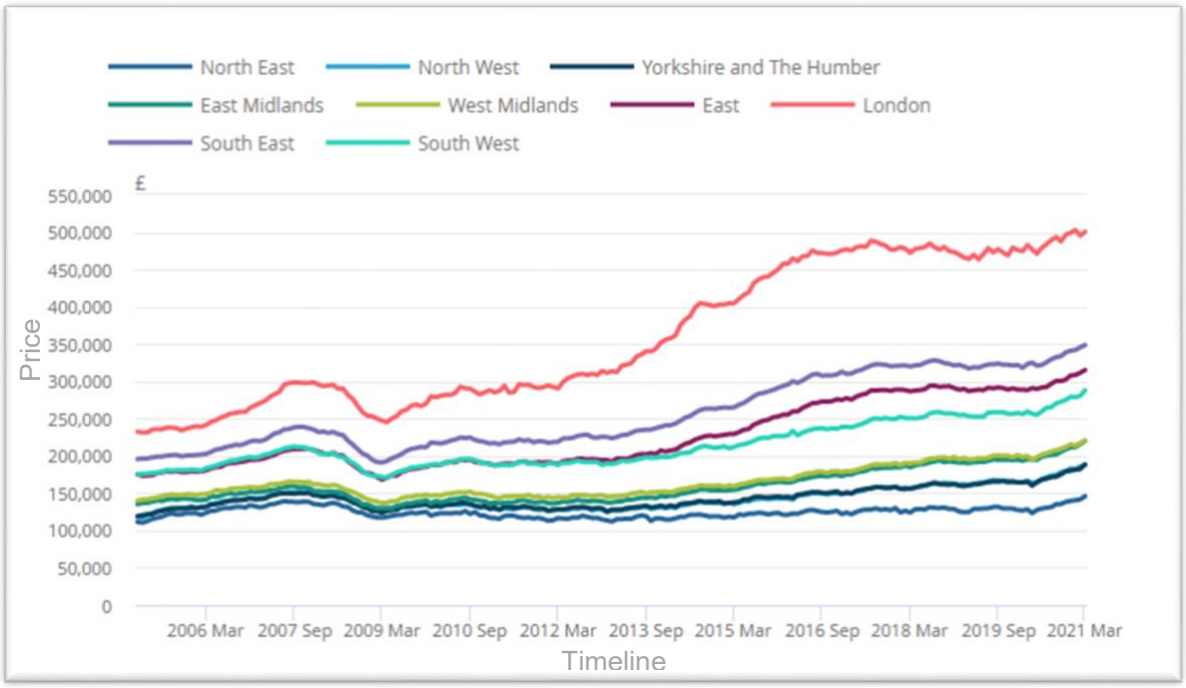


Figure 3.2: Average house prices in UK and London, 2005 to March 2021. Source: UNO, 2021

House price growth had maintained an average of 9.2% in London from 2009 to 2015, while the rest of the UK had managed 5.2% over the same period (Marsden, 2015). This trend is expected to continue since the market did not correct during the lockdown periods of Covid-19. House prices in London are influenced by the interplay of demand and supply, new houses, overseas investment, government deregulation and financialisation of homes.

3.3.1 Demand and supply of housing in London

There is high demand for housing in London both from Londoners, the UK populace and overseas investors (Wallace et al., 2017). On the other hand, there is a limited supply of housing, partly due to inadequate investment, unavailability of construction land, and obstacles in councils approving plans and construction permits. Presently, Londoners have decried the type and standards of housing constructed as it does not meet either UK or European standards, hence exposing people to overcrowding, low-standard houses and displacement and disintegration of communities (Marsden, 2015). As such, councils are demolishing old houses to pave way for modern flats that will house more wealthy people (Gallent et al., 2017). Due to the low supply of houses in London (less than 1% of existing houses), buyers and investors are ready to take whatever is available and thereby push the prices up.

It is no secret that overseas first-time buyers are an integral part of the London housing market. Wallace et al. (2017) sought to find out the proportion of new homes bought by overseas investors and the use of these houses. They found that 13% of new houses were owned by overseas owners in 2013, and this rate increased by 2% annually. Southeast Asian countries China, Singapore, Hong Kong and Malaysia accounted for the highest proportion of overseas buyers. Regarding areas, overseas investors concentrated more on inner city boroughs as opposed to the periphery of the city. The overseas buyers bought properties across the price spectrum, from as low as 0.2 million up to 5 million, while there was a split between mortgaged and non-mortgaged purchases.

The propensity to leave a property empty is higher in the inner city and more valuable. However, only a small portion (less than 1% of houses) are left un-

occupied by foreign investors for speculation. The rest of the properties are dedicated to renting to Londoners (occupancy by students and families).

There are several repercussions of overseas purchase of new houses. First, the pre-sale of houses enables the developer to build quickly hence availing more houses earlier. And finally, house prices are inflated to match the demand (Marsden, 2015). Despite these repercussions, the housing market crisis in London would persist in the absence of overseas investors.

3.3.2 Housing crises

London's housing crisis was not instigated by overseas investors solely, but by undersupply, rising cost of living, house and rent inflation, and under-usage of houses. Also, there is a failure to identify the land, approve planning and construct new buildings (Snelling et al., 2016).

The housing crisis in London is bleak, and characterised by homelessness, poor quality accommodation and displacements. The situation is pervasive and damaging with little help from the Government or councils. Although the housing crisis in London is not new, it has been intensified by neoliberal policy that facilitates Government initiatives such as the financialisation of the London housing market, privatisation, deregulation and gentrification. Gentrification is the process of converting boroughs that catered to the housing of the lower class to the tastes of the middle class, leading to low-income earners being driven further out of the city centre (Hamnett, 2003). For instance, the Heygate estate in London was put up for regeneration. The repercussions were replacement of old houses with new flats, disintegration of communities and increased rents. Gentrification benefits the investors and owners of real estate at the expense of working-class people.

Leccis (2019) investigated the effect of housing regeneration in Bankside, London, and asserted that even the most innocent regeneration programmes turn out to be gentrification. These regeneration programmes result in a loss of social cohesion and diversity as former residents move to cheaper places. Therefore, involvement of local communities is essential to successful regeneration

programmes. When a regeneration scheme turns out to be gentrification it denies humans, especially previous occupants, the right to adequate housing.

3.4 House price estimation: influencing factors

There are many factors that influence house price, value and its estimation. These factors can be classified broadly as environmental, macroeconomic, locational, structural and neighbourhood. A house is a heterogeneous good, with the price determined by several factors, such as motorway proximity, garages, pools and lawns, whereby each of these factors have no market value individually.

The reasons for appraising, valuing and estimating house prices depend on housing stakeholders. Bankers appraise the houses to conform with Basel II Accord, issued in Basel in 2008, that states '*the bank is expected to monitor the value of the collateral on a frequent basis and at a minimum once every year',* (Hong et al., 2020). In this light, frequent monitoring and appraisal is required as the market is subject to periodic fluctuations and significant changes. For the banks, valuing house prices is an ongoing process, and hence reliable methods must be employed. Local authorities' intent is to value properties for tax purposes, for instance in the UK Computer Assisted Mass Appraisal (CAMA) is an accepted tool for mass appraisal.

3.4.1 House characteristics

Structural house characteristics are the physical factors that describe a house, including the type of the house, age of the house or construction date, floor space, number of bedrooms as well as number of bathrooms, among other things. Innumerable studies have been carried out to relate the price of a house to these factors. Hong et al. (2020) used elapsed year, floor area, floor level of the property and heating system as structural factors to determine their impact on price of a house. The outcome showed that elapsed year has a negative correlation with price, while floor area has a positive impact on the price of house. Buyers were indifferent on the presence of a heating system or not. Regarding floor level of

the apartment, the lowest floor of an apartment influences prices positively, while higher apartment floors affect price negatively.

Similarly, (Ferlan et al., 2017) had similar findings on structural factors of houses in Slovenia. They asserted that impact of the floor level of an apartment depends on context. A floor level is dis(amenity) if the apartment block has no elevator, but incentive if there is an elevator. The age of a house is negatively correlated with the price, a major driving factor for this outcome is depreciation attached to the house. Orientation of the house affects price positively if the right orientation, southwest – otherwise the house price deteriorates, as there will be insufficient sunlight hours per day. Layout of the house and presence of parking lot also affect the price of a house positively.

Nguyen (2020) conducted a study to determine the hedonic determinants of house prices in two cities in Vietnam. The country has an emerging housing market, where most of the people prefer to own homes over renting. To establish the house price determinants, (Nguyen, 2020) applied Ordinary Least Squares (OLS) regression in conjunction with statistics on house data collected on Ho Chi Minh and Ha Noi cities. The findings indicated that the number of bedrooms, size of house, type of house and structure influence the house price positively. Two house determinants were found not to have a statistical significance on price: number of bathrooms and availability of pool.

Ndegwa (2018) carried out a study in Nairobi metropolitan area to establish the houses' structural factors that significantly influence the price. The study employed both primary and secondary data. Upon the analysis of the data using Multiple Linear Regression (MLR), the results showed that land price and size of the apartment are the most significant in influencing price of the house positively. Factors that had no impact on price include nearness to informal settlement and availability of balcony.

3.4.2 Macroeconomic indicators

Macroeconomic factors such as labour participation, interest rate and population influence the price of a house. Relevant macroeconomic factors included in (Hong et al., 2020) are transaction period (year), gross domestic product (GDP),

growth in real GDP, land price fluctuation and interest rate offered on mortgage. Transaction year had the most positive impact on house price. GDP and mortgage interest rate also affect the price – in this case, a decrease in mortgage rate influenced price positively, while an increase in GDP also affected the price positively. Growth in GDP rate had no appreciable influence on price of houses.

Trinh et al. (2021) examined the short- and long-run impact of Foreign Direct Investment (FDI) inflows, energy intensity and financial development on house prices. The study employed data from 35 countries during the 1980–2018 period. The study applied multiple data analysis method such as OLS, GMM and Multiple Linear regression (MLR). Results from GMM revealed strong cointegration between house price, financial development, energy intensity and economic growth. The regression methods produce consistent results with GMM.

In developed financial markets, housing is easy to buy as an asset, hence facilitating housing investments. A stable financial development improves demand of housing as well as stabilising prices (Yildirim et al., 2021). The house values and ultimately the prices are likely to rise with financial progress. However, the housing price rates increases during a housing boom, regardless of economic growth level. Trinh et al. (2021) note that the impact of a housing boom is lesser in economies with good financial development than in undeveloped financial markets. In countries with underdeveloped financial markets, high capital requirements and loan limitations depress home prices by lowering demand. Therefore, once finance growth is realised, housing demand and boom result. Trinh et al. (2021) included various variables in their research such as housing price index (HPI), FDI, Energy intensity (EI), financial development index, GDP, labour participation, urbanisation, inflation and trade.

Shimizu (2014) observed that office worker ratio in a neighbourhood leads to increased house prices. Since the income and academic level of these people is expected to be high. This observation is consistent with (Rosen, 1974), who found that house prices in a neighbourhood with high income as well as substantial wealth tend to be higher than other places. Therefore, house prices are not affected purely by structural and physical factors in high-income communities.

Green (2018) carried out research on the relationship between immigration and house prices in the UK. A total of 80 local authorities' data across the UK from

2010 to 2016 was used in the analysis, using GMM regression. The research unravelled that immigration correlated negatively with house prices. But it was also uncovered that immigration lowers the level of crime in a community. The reason for house price decline upon inflow of migrants in each neighbourhood is due to native out-migration in response to the inflow of migrants. With Brexit in place, the net migration is expected to reverse, and the effect could bring about a positive wealth effect in the coming years. In predicting house prices in a given local authority, high inflows of migrants will result in reduced house prices. Therefore, the immigration factor should be taken into consideration when predicting house prices.

Karagöz and Özkubat (2019) conducted a study in the Aegean region to investigate the impact of macroeconomic factors on house prices. The duo decided to carry out their study on this region due to intensive industrialisation and intensive immigration witnessed in recent years. The data obtained from varied sources was analysed using the regression method. The results showed three varied factors were at play: in Izmir, house prices were affected by general price and gold price; in the Aydin sub-region, population, interest rate, general price level, gold price and exchange rate influenced house prices; in the Manisa sub-region, exchange rate, interest rate and general price dictated the price of the houses.

The performance of economies is central to policy decisions because it provides opportunities for policy makers to understand and get insights into the future. They can more effectively understand what to expect in the future and develop frameworks for mitigating possible adverse effects. It can also set the stage for development of policies that take advantage of positive turns in the economy. Equilibrium macroeconomic variables including supply side performance can present significant opportunities for policy makers to see areas of weaknesses or challenges. Differences in the performance of economies across the globe also play a major role in determining the exchange rate differentials and trade related variables (Ellison and Scott, 2000). Different modelling approaches have been adopted in assessing economic performance and other relevant elements and deviation from the expectations. To ensure their effectiveness, these models are

calibrated to the critical aspects of economies such as the Eurozone and North America (Cicceri et al., 2020).

A major issue cited in (Hume and Sentance, 2009) is that adverse economic events that are unpredictable and nonlinear are difficult to predict and present significant challenges for policy makers. The focus of the policy makers is to have effective ways of predicting the occurrence of such events, and effectively understanding their nature before they occur. A good example is the credit crisis of 2008 that affected the global economy resulting in massive losses. Economic forecasting can have significant benefits for policy makers as well as the general population by alerting them of potential economic crises before they occur. The available econometric models have had significant gaps in offering the required predictive capabilities due to the insular nature of the discipline and its reluctance to adopt knowledge and advancements from other fields. The following section discusses the traditional econometric models and forecasting approaches as they are applied in the sector. It also presents the top macroeconomic indicators used for forecasting in the UK.

A major aspect of housing prices as a factor influencing the economy as a whole is that they are the outcome of a wide range of other factors. Housing prices affect the economy through different elements such as **credit**, **disposable income** and **interest rates**. One of the main aspects of the economic determinants of housing prices is the nominal interest rate, which is an indicator of the investment environment and economic conditions in a country in the future. The nominal interest rate is a predictor of the appetite people have for investing in different types of securities and investment opportunities (Xu and Tang, 2014).

A study by (Adams and Füss, 2010) posits that economic variables such as **credit availability** and **money supply** have multidirectional linkages to housing prices. Purchasing a house has a significant impact on the real income in a family. In the same way, the **disposable income** in a family has a significant impact on consumer spending in different activities and consumer confidence (Tajik et al., 2015). This means changes in housing prices will have a significant impact on the level of consumption in the economy. They also argue that economic activities are the outcome of different elements of real **GDP** such as **employment rates**,

industrial production and **consumption**. In the long run, housing prices are a central element of economic growth and development (Chen et al., 2014).

Disposable income is considered as a proxy for affordability in the housing market, with real housing prices being positively associated with disposable income. Higher income leads to an increase in demand for housing, and reduction in stock leads to increased prices. **Consumption** is perceived to be strongly dependent on housing and stock market income (Case et al., 2013). A major factor that has been studied as having a major impact on the housing market is interest rate. It has been evaluated as a factor in the housing market as influencing demand and prices. Different types of interest rates have been explored in the research including the **long-term interest rate**, **treasury bill rate**, **mortgage rate** and **real interest rate**. In studies such as (Barksenius and Rundell, 2013), the three-month treasury bill rate is used as the nominal interest rate. The findings showed a strong negative relationship between term spread and the nominal interest rate. Real interest rate, which is the cost of financing, was shown to have a significant and negative impact on the real cost of housing. Additionally, in a study by (Hilbers et al., 2008), the dual role of interest in affecting the housing market in Europe was reviewed. They concluded that it affects the mortgage rate, which shows in the housing costs, and the risk-free rate, showing the opportunity cost of investing in a house as opposed to another venture. They also found that **real interest rate** had a negative and significant effect on housing prices, with the market in Sweden being more sensitive than in the UK.

Unemployment rate is a significant economic variable that strongly influences the housing market. It has been studied in different research articles as an indicator of economic conditions as a whole. Studies such as (Barksenius and Rundell, 2013) show a strong association between housing price return and unemployment rate. They found that unemployment declines in the boom period, but the bursting of the bubble leads to a major increase in the unemployment rate. An important aspect of unemployment as a factor in the housing market is that it increases uncertainty. It raises concerns among financiers about the capacity of borrowers to keep up with their repayment of mortgages and other credit. Unemployment may deter first time buyers from mortgages and cause

them to prefer alternatives such as unsecured loans. It causes lower growth in the wage rates and increases uncertainty of future income levels.

Construction costs play a major role in determining the prices of new dwellings and the overall cost of housing in the economy. This incorporates the cost of labour and construction materials, with the research indicating that the higher costs of construction result in reduced supply of housing units and stock in the market. The reduction in housing space generates an increase in the prices of houses as well as rent. This means construction costs are negatively associated with the supply of housing, and so have a positive impact on the overall prices of housing. The relevance of this relationship was based on the power of property developers to transfer the costs of construction to the buyers. Additionally, changes in construction costs affect the construction activities in general, resulting in a highly dynamic impact on the overall housing market (Barksenius and Rundell, 2013).

In a review by Cohen and Karpaviciute (2017), the determinants of housing prices such as economic, financial and social considerations were included. One of the key factors identified was the impact of growth in real per capita **GDP**. The main reason for this is that it leads to the perception of higher income over the lifetime of an individual and the willingness of individuals to spend a larger share of their income on housing. This means higher growth in personal income is positively associated with stronger growth in the demand for housing. Other factors identified in the study were credit conditions such as loan to value, debt to income and down-payment requirement. Chu (2014) identified that housing prices respond heavily to changes in the requirements for down-payments. This impact was very strong for housing markets where the owner-occupied and rental housing were inelastic in supply. The DTI ratio lowers the purchasing power of individuals who would be willing to purchase homes, while the LTV ratio lowers the pool of borrowers who can access the financing necessary to purchase a home, hence reducing the demand pressure in the market. Durganjali and Pujitha (2019) identified physical attributes, location and several economic factors to be responsible for influencing house resale price. In Cohen and Karpaviciute (2017), who focused on the housing market in Lithuania, the impact of economic factors such as **GDP**, unemployment and inflation were incorporated. The results

indicated that the unemployment rate and **GDP** were significant causal factors of housing prices. The study also showed a causal association between inflation and housing prices. However, the causation was from housing to inflation rates, indicating that the rate of inflation does not affect housing prices, but instead the opposite relationship is evident (Cohen and Karpaviciute, 2017).

3.4.3 UK macroeconomic indicators

Some of the key issues in the existing models include the need to compare feature packed models against more abstract ones that are easier to digest. (Gualdi et al., 2015) argue that larger models have to be developed before researchers can unpick them to identify the underlying effects and relationships that are relevant in explaining the phenomenon at hand. A key conclusion from the literature on agency-based models is that their effectiveness and applicability depends on their specific assumptions and how they have been used. Lack of restrictions in the models can be a major challenge. This flexibility of the models is what give rise to the risks identified, but also contributes to certain benefits that can make them effective for economic modelling.

As cited in (Haldane and Turrell, 2017), the models are effective in areas where heuristics dominate and there is plenty of granular data that results in the failure of analytical models. Applicability of these models includes areas such as the relationships among financial institutions where the dependencies increase system-wide risk and stress.

In testing the effectiveness of different models for predicting economic outcomes, Balcilar et al. (2015) used a small set of variables including inflation, real GDP, and short-term interest rates. They applied a wide range of models including linear and nonlinear, time series and classical ones. They tested the different models with the aim of assessing the US economy *ex-ante* to identify where the global financial crisis could be effectively captured. The nonlinear DGSE model was shown to be the most effective in providing the necessary predictions regarding the recession of 2007/2008. However, the outcome still showed major limitations that could be addressed using artificial intelligence.

The applied approaches involved stochastic data models that sought to identify what underlies the data generation process. A more effective approach would have been one that seeks to find a function that can best predict the output and the relationships given the identified inputs. Some of the key macroeconomic indicators used in the UK include the level of household growth and population. Real income growth in the country and interest rates are also considered as key macroeconomic indicators used for predicting other elements of economic performance in the UK. Housing affordability in terms of the mortgage rates offered to home owners is a key macroeconomic indicator used in the UK for forecasting purposes. The number of dwellings completed and the overall supply level in the market is another important macroeconomic indicator applied in the forecasting of the UK economy (Cohen and Karpaviciute, 2017).

The following sections in this chapter will now focus on the review of existing literature, with emphasis on the seven research questions stated in Section 1.3.

3.4.4 Neighbourhood amenities

Ferlan et al. (2017) revealed that absence of industrial facilities, quietness (absence of noise), tidiness of neighbourhood and open view increase the price of a house in Slovenia. When there was unacceptable noise level in a neighbourhood the price of a house decreased 12% compared to a neighbourhood with acceptable noise levels. Open view contributes an increase of 12% on a house price, and open view to the sea increased the price by up to 24%.

Shimizu (2014) incorporated neighbourhood variables such as floor area ratio, zoning, average building area, building density, standard deviation of building area, rate of office worker and road traffic noise. The study applied Ordinary Least Square (OLS) regression and Multiple Linear regression to analyse the data. The data was obtained from various sources, i.e., websites, government archives and GIS. The land-use conditions had a positive correlation with the price of the house; an increase in average building area positively affected the single-family house price. The standard deviation of the building area had an inverse correlation with price, indicating that house prices increase when there is

uniformity of houses of the same size, and the prices decreases where there is non-uniformity of the houses. The density of the houses had a positive correlation until the 95% threshold, beyond which the price of the houses decreases. The high prices in regions with large houses and uniform distribution of houses is explained by pleasant local environments and orderliness of towns. On the other hand, variation of building areas causes the local environment to deteriorate, hence decreasing house prices.

Hong et al. (2020) investigated the effects of neighbourhood factors on the price of houses. The neighbourhood factors they included in their model were: apartment brand, number of units in an apartment complex, parking lot, floor area ratio, number of buildings in the apartment complex, building coverage ratio and top and lowest floors of an apartment. Number of buildings in the apartment complex was the most critical factor in neighbourhood category in influencing the price positively. Other factors that influenced the price positively are; the lowest floor of an apartment, units (rooms) available in apartment and presence of a parking lot.

An appropriate amount of building concentration has a positive impact on house prices. Therefore, the right density of houses in a neighbourhood should be maintained to attract beauty, orderliness and reduce traffic. Neighbourhoods that have high density of houses can be marred by social ills such as crime, high traffic jams, unpleasant environment and waste disposal problems. Neighbourhood variables have a significantly impact house prices so, a model that doesn't incorporate these variables stands to be declared incomplete. Omitting these variables introduce bias in a model that is purely hedonic (Shimizu, 2014).

3.4.5 Environment

Chen and Jin (2019) investigated the impact of environmental pollution on house prices in 286 cities in China from 2005 to 2013. The study applied Ordinary Linear Regression (OLS) to analyse the data. They established that air pollution imposed by fossil fuel burning does indeed have a negative impact on house prices. Specifically, the results suggest that a 10% increase in $PM_{2.5}$ concentration brought about a 2.4% reduction in house prices locally. These

researchers also noted that air pollution also impedes urbanisation, menaces cities' human capital formation and alters expectations of people's housing prices. Therefore, the findings by Chen and Jin (2019) could be generalised to the cities around the world with a similar profile on matters air cleanliness. In this regard, in estimating the house prices, air cleanliness is a factor that surely influences house prices. Their findings align with those of Deng et al. (2012), who observed that after decommissioning of a power plant in Chengdu, China, the occupancy of houses rose by 54.6% and prices in excess of 6.8%. The reduced housing activity in air polluted areas is premised on reduced investment in housing by investors, and the willingness to sell a house as soon as possible even at a discount, hence depressing the house prices.

Belcher and Chisholm (2018) investigated the effect of vegetation on residential property value in Singapore. The duo observed that vegetation increases the selling price of a property by 3%. Managed vegetation accounted for the biggest effect on price, followed by high conservation value vegetation, and lastly spontaneous vegetation. Therefore, to attract high prices houses should have managed vegetation nearby. Vegetation provides non-quantifiable service to society and is mostly undervalued in land-use decision making. Some of the services provided by vegetation are shielding of urban heat, aesthetic value, improved air quality and recreation. Similar studies have shown that presence of recreational parks in a neighbourhood explain increments of as much as 10% on property prices. The presence of vegetation dictates prices differently from region to region, hence local preferences should be determined as estimated values cannot be applied in other regions (Belcher and Chisholm, 2018).

Trojanek et al. (2018) analysed the effect of proximity of greenness to urban areas on apartment price in Warsaw. The duo obtained over 43,075 geocoded data for transactions from 2010 to 2015. A number of analysis methods were employed in the research including Ordinary Least Squares (OLS), Median Quantile Regression (QR) and Weighted Least Squares (WLS). The results showed that the apartments within 100 metres of a green area increases the dwelling price by 2.8–3.1%. For houses built after 1989 the presence of a park/forest nearby yields higher change, while in houses built before 1989 the presence of urban green produced a higher implicit price. Therefore, the

greenness of an urban area, and parks or forests in the vicinity, increase the price of the houses (Donovan et al., 2019). Hence it should be considered in house development and prediction.

Trojanek et al. (2018) singled out the benefits of parks, greenness and forests nearby houses from a vast amount of literature. First, environmental benefits include ecological, cooling urban places, pollution reduction through carbon sequestration and increased biodiversity and wildlife. Second, economic benefits include energy saving, good water balance, attractiveness to tourists as well as increased property values. Third, social and psychological benefits such as entertainment and recreation, crime reduction, strengthening social bonds and improved overall health. Lastly, planning and designing benefits are derived as well, which include aesthetic values, perception of green areas and planning and designing green areas.

Iqbal and Wilhelmsson (2018) observed that parks and open spaces are a desirable part of city scenery. However, property buyers prefer open spaces to parks and forests as they can be associated with crime or dangerous wildlife, such as snakes, monkeys and mosquitoes. When a buyer's perception of a nearby park is negative, the properties fetch lower prices in the market. So there are some dis(amenities) associated with parks and forests, however the parks' benefits far outweigh these dis(amenities), as shown by Trojanek et al. (2018). Iqbal and Wilhelmsson (2018), in their research on desirability of parks and green spaces, found that grass parks and park blocks are more desirable than landscaped parks. Also, parks in a city centre have greater impact than parks on the periphery of Stockholm on house prices. The level of crime in a park affects apartment prices; low crime brings about positive prices and vice-versa.

3.4.6 Locational

Nguyen (2020) observed that a house price is negatively affected by its proximity to the city centre, while a park is positively correlated with price. Ndegwa (2018) observed neighbourhood determinants that influence the price of a house positively include nearness to schools, malls and a city centre, while proximity to slums had no effect on price. Cordera et al. (2019) inferred that accessibility of a

house by public transport can have a positive impact on the real estate price. However, the benefits derived from accessibility vary from one area to another. For instance, in a city with mobility problems, such as Rome, Italy, the accessibility of public transport impacts positively on the price of a house. Whereas in a city without mobility problem, such as Santander, Spain, the availability of public transport does not influence the price of a house as much. Hong et al. (2020) observed no significant effect of locational attributes on house price. However, Huang and Hess (2018) established that there is a positive, statistical significance to house price in relation to proximity of schools. In regard to nearness to a city centre, it is the most crucial factor in influencing house price positively in Slovenia (Ferlan et al., 2017).

3.5 Forecasting models

A wide range of forecasting models have been applied in assessing the changes in housing prices in the UK and elsewhere while at the same time identifying the key determinants. In 1993, (Drake) used a very simple model for forecasting housing prices where he included disposable income, number of houses being constructed and mortgage interest. The number of houses started in the period was found to be a less effective measure because it has a lagged effect on housing stock, unlike the number of houses completed which directly feeds into the supply side of the market (Cohen and Karpaviciute, 2017). This model is presented in Equation 2.1.

$$In(P) = \beta 1 + \beta 2\, In(Y) + \beta 3R + \beta 4\, In(B)$$

Where:

P = housing price index

Y = disposable income

R = mortgage interest

B = new dwellings

Equation 0-1: Forecasting Model by Cohen and Karpaviciute

A major element of this model is that it showed a very bad fit for the regression equation – a coefficient of determination, R-Squared, of only 7.3%. The mean

error for the regression was also shown to be very low, indicating the inaccuracy of the model.

The lifecycle approach is also adopted in these predictions as a way of improving the accuracy of the forecasts by accommodating the multipurpose model, where households are assumed to be focused on maximising their lifetime consumption of housing and other commodities that are essential to them. Households operate under a lifetime budget constraint, meaning that current income is not the only limitation when deciding about housing. Instead, households also consider their expected future income in determining the type of house to purchase or rent. This also means that responsiveness to exogenous shocks is less direct since it can be spread to other periods or years in the future, as presented in Equation 2.2 (Meen, 1999).

$$R(t)/g(t) + g(t)/g(t) + \pi(t) - \delta(t) = \left(1 - \theta(t)\right) * i(t)$$

Where:

$R(t)$ = Marginal rate of substitution between housing and consumption

$g(t)$ = Purchase price of dwellings

$\theta(t)$ = Household marginal tax

$i(t)$ = Market interest rate

$\delta(t)$ = Depreciation rate on housing

$\pi(t)$ = General inflation rate (depreciation of financial assets)

(*) = Time derivative

Equation 0-2: House price forecasting model by Meen (1999)

This approach is preferred due to its versatility in modelling risk in the market and credit market constraints. In spite of its effectiveness and advantages, the model was found to have flaws in relation to its assumption that the housing market is economically efficient in terms of capturing all information.

More recent forecasting models have been developed to address the challenges identified in earlier approaches. One of the main issues identified in the earlier models was the assumption that prices should react instantly to the explanatory

or exogenous variables. For example, major changes in the exogenous variables are overestimated in their impact when using Ordinary Least Squares (OLS) regression.

To counter this problem, more accurate predictions can be made using the Autoregressive Moving Average Model (ARMA). This approach is effective because it introduces two new elements to the model: the autoregressive element, which incorporates the past values of the dependent variable alongside the explanatory variables; the use of the moving averages of all variables, hence eliminating the adverse effect of one-off deviations in the variables (Balcilar et al., 2015, Wilhelmsson, 2009).

The hedonic regression model refers to the weighting of the relevance of different components in constructing an index of usefulness. This refers to the intrinsic pricing of the attributes revealed through observed prices of the products and the specific levels of their associated characteristics. The model has been applied in a wide range of studies such as (Balcilar et al., 2015) focused on nonlinear pricing models for incorporating the non-observable values of the housing market attributed to neighbourhoods, such as access to hospitals, traffic and noise pollution. Wilhelmsson (2009) expanded the model to incorporate physical attributes of the houses, such as number of rooms, living areas and number of bathrooms.

A key aspect of the housing market is the heterogeneity of houses, since they differ based on location, construction details, size and services accessible. Since these features are not directly traded, their impact on the prices is not explicit and the hedonic pricing model is applied as an effective framework for ensuring that their marginal contribution can be effectively captured in the forecasting models. The transaction price for the house is used as the proxy for its value and it can be compared to other measures. However, the model does not indicate the independent variables to be incorporated, hence Abdulai and Owusu-Ansah (2011) recommend that selection of variables should be guided by the problem of multicollinearity, where they select only the variables that are not highly correlated. The form of the equation in the model could be linear, quadratic or semi-logarithmic, as indicated in Yusof and Ismail (2012).

A housing price model was developed by Gu (2018) that incorporates different variables for the demand and supply side. Seven factors affecting house prices were identified in the research. These included land costs, loan interest rates, real estate investments and the number of completed residential units as the supply variables. Gross Value Added, real income and population growth were used as the demand variables. They were applied in line with the ideas presented by XU et al. (2016) regarding the relevance of the demand and supply side variables in the prediction models. Principle component analysis (PCA) and multicollinearity avoidance were used to determine the variables to be extracted and those eliminated from the analysis. PCA is a dimensionality-reduction method used to reduce the dimensionality of large data sets, initiated by transforming a large set of variables into a smaller set without a significant loss of information in comparison with the large set. In the model, PCA1 (the first PCA iteration) and PCA2 (the second PCA iteration) were included in the regression analysis, with the model indicating capability of explaining almost 100% of the variations in housing prices. While the model was shown to be effective, the individual variables were shown not to be significant since they had relatively high p-values. After accounting for the multicollinearity, the model was found be effective in explaining 96.3% of the variations in housing prices (Gu, 2018).

Housing prices have been evaluated using different regression methods applying economic, demographic and spatio-temporal data. Hedonic regression has been cited as an effective approach for assessing price variations in the housing sector because it captures non-monetary factors that may not be clearly observable. Additionally, it captures the factors that do not directly or explicitly influence the prices of housing in the economy. Another approach is to use spatio-temporal data to estimate house price variations (Chica-Olmo et al., 2019). This approach was adopted in a study on the housing market in Granada Spain. The analysis was undertaken using the regression-cokriging (RCK) method and the outcomes compared to the universal cokriging method (UCK). Kriging is a multistep process initiated by an exploratory statistical data analysis, surface creation using variogram modelling, and the exploration of a variance surface (which is optional). This process is explored when there is a known directional bias or spatially correlated distance in the data. The kriging statistical method improves the

performance and effectiveness of predictions using spatial hedonic models (Kuntz and Helbich, 2014).

These models are applied due to their higher accuracy compared to OLS and spatial error models. Another important element of these models is that they effectively incorporate temporal and spatial components in the analysis. Spatial dependence is common in the housing market since prices are more alike within the same area. However, the presence of correlations in the disturbance term of the model means that the OLS estimator is inefficient. As an alternative, the General Least Squares method is used to estimate the relevant parameters and predictions. The model is represented as shown in Equation 2.3

$$Z = X\,\beta + u$$

Where each element is a matrix as presented below:

$$z = \begin{bmatrix} z1 \\ z2 \\ . \\ . \\ . \\ zq \end{bmatrix} \quad X = \begin{bmatrix} x1 & 0 \dots 0 \\ 0 & x2 \dots 0 \\ . & . \quad . \\ . & . \quad . \\ . & . \quad . \\ 0 & 0 \dots Xq \end{bmatrix} \quad \beta = \begin{bmatrix} \beta 1 \\ \beta 2 \\ . \\ . \\ . \\ \beta q \end{bmatrix} \quad u = \begin{bmatrix} u1 \\ u2 \\ . \\ . \\ . \\ uq \end{bmatrix}$$

Equation 0-3: General Least Squares

The GLS estimator of β in the model is the Best Linear Unbiased Estimator (BLUE). The key advantage of the RCK method is that it can be used in combination with other models to give a hybrid approach that would be more efficient in supporting the predictions. In this model, the X matrix incorporates other explanatory variables, as in the case of a single equation model. It explains any type of explanatory variable, which is obtained by incorporating the drift and ordinary kriging of residuals. The main advantage of this approach is that it is flexible for modelling and mapping since it may be applied with other methods (Montero et al., 2015).

The problem of economic modelling has been considered in many settings, with scholars applying different strategies to predict or identify changes in housing prices. The performance of the housing market is a key indicator of the economy of a country. This is because a wide range of factors have been identified as causing changes in the aggregate prices of housing (Cohen and Karpaviciute, 2017). Housing prices are a function of the demand and supply in the market,

with demand being positively correlated with the number of households and their real income level. It is negatively correlated with interest rates, and the impact of demand on housing prices is influenced by the relative effect of supply and demand. The change in housing prices over time is a factor of a wide range of determinants that indicate the changes in the state of an economy (UK Government, 2018).

The UK government has been interested in efforts to identify the factors influencing housing prices to determine affordability. The National Housing and Planning Advice Unit (NHPAU) have been conducting analysis to identify the factors that influence housing prices in the UK using the affordability model (UK Government, 2018, Cohen and Karpaviciute, 2017, Xu and Tang, 2014). Considering the complexity of the housing market in the UK and globally, the analysis is not designed to be exhaustive, but instead seeks to provide an accurate and close estimation of the relationship and apparent impact of the different factors on housing prices. The NHPAU published the report, dubbed 'Affordability Still Matters', where the key drivers of affordability are estimated alongside their relationship with the affordability of the housing market (UK Government, 2018). Some of the main findings from the report include:

- A 1 percent increase in the number of households would cause house prices to rise by 2 percent.
- 1 percent increase in real income causes house prices to increase by about 2 percent.
- A rise in interest rates by 1 percentage point can cause housing prices to fall by about 3 percent.
- An increase in the stock of housing units by 1 percent causes a fall in the prices of housing by about 2 percent.

The analysis was cited as being reliant on the view that other factors are held constant. While this may be helpful in showing how the factors have changed over time, it is highly limited and should only be used to generate stylised inferences about the changes and impact of the identified variables on the housing market. A major weakness of the model is that shifts in one of the parameters do not happen in isolation – changes in housing prices are due to an interaction of different demographic, societal and economic factors (Reed, 2016).

Housing prices in the UK have been on a general upward trend since 1970, with a few short-term contractions in the early 1990s and 2007, as indicated in Figure 3.3.

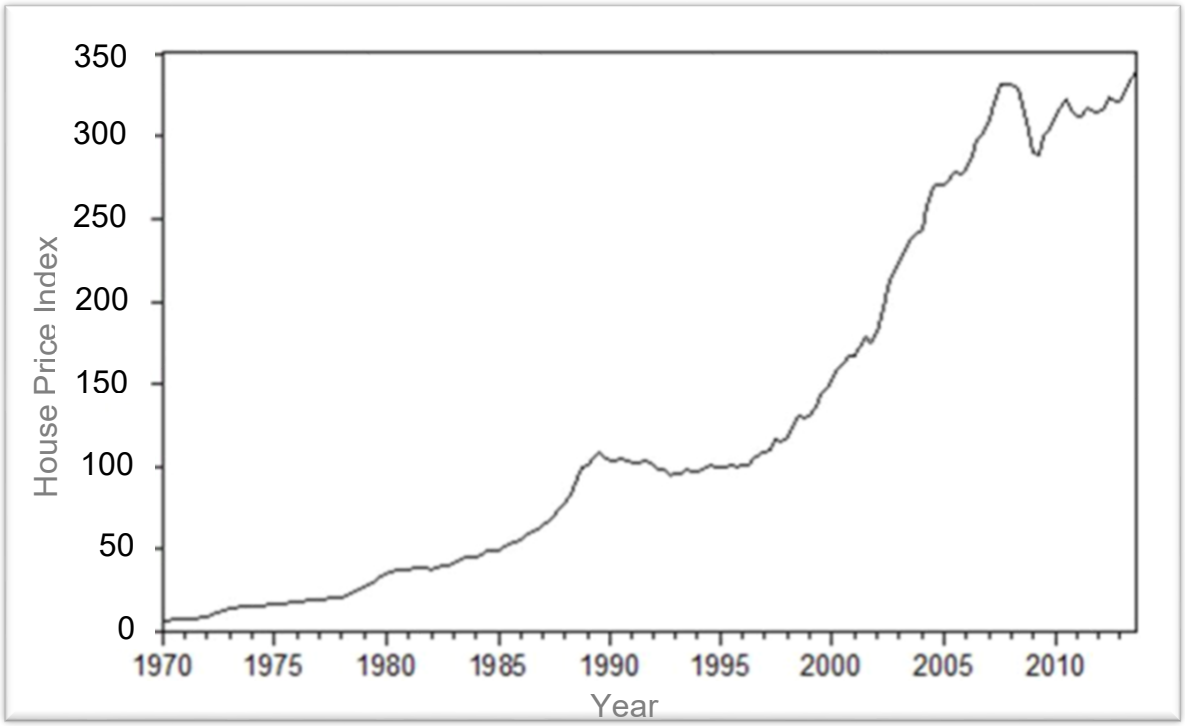


Figure 3.3: UK House Price Index

Most international organisations such as the World Bank and IMF engage in economic forecasting with the aim of showing their prospects. Most of these organisations rely on traditional models of economic forecasting where the macroeconomic variables are fitted into pre-specified relationships between the inputs and outputs. The models assume a stochastic process of the relationships between the variables. This means the models are relatively static and are only as good as their specifications, and the models rely on these specifications irrespective of what the available data may suggest. A lot of economic modelling has been undertaken using different approaches such as time series analysis and the general equilibrium model. The focus of this analysis, such as the one undertaken by (Benigno and Thoenissen, 2002), has been to develop dynamic models. To enhance the effectiveness of such models, they incorporate different dimensions of the economy to offer a comprehensive framework.

In Benigno and Thoenissen (2002), the forecasting model focusing on supply side elements was modified to understand the determinants of the real exchange rate and dynamic adjustment paths. Other elements in the traditional models include the introduction of imperfections in the goods and labour markets. The specification of nominal rigidities (a situation where a nominal price is resistant to change) is a major element of the model, which was adopted to enhance its predictive capabilities. Using the panel cointegration techniques, the model showed a long run relationship between real interest rate differentials and real exchange rates in the economy. An interesting finding is that the model was found to have a positive impact for the small open economy, although it would have been rejected using data from larger economies.

Time series analysis refers to the consideration of ordered sequences of values at equally spaced time intervals for each variable. The basic assumptions of time series analysis are that data points taken over time have internal structures, such as trends, autocorrelation and seasonal variations, that can be evaluated. They often utilise multiple linear regression models. The relationship between the observed response and the contemporaneous variables is used in the models for predictions. The main issue with these models is that they are limited due to the nonlinearities that exist in the relationships, and these make it difficult for the relationship between the observed variable and the predictors to be accurately predicted (Vrbka, 2016).

Time series analysis and its application in economic forecasting is also associated with the impact of disaggregation. As cited in Poncela and García-Ferrer (2014), forecasts derived from the aggregated time series based on univariate models were compared to others derived from each component of the aggregate. The outcome was applied in assessing the economic performance of different EU economies. The results of the modelling were compared to the modelling and forecasting predictions. The outcomes indicated that the factor models were more accurate due to their greater effectiveness in explaining behaviour at the turning points.

In Haldane and Turrell (2017) an agent-based interdisciplinary model of macroeconomics was applied. The model was developed with the aim of addressing the shortcomings identified in the forecasting approaches prior to the

global financial crisis. Some of the key issues with the models used before the crisis was their restrictive nature and the fact that they were not effectively supported by empirical evidence. Haldane and Turrell (2017) explain the relevance of different types of economic models, including their application in showing different relationships. They argue that statistical models do not say anything about heterogeneous models. Agency-based models and Dynamic Stochastic General Equilibrium (DSGE) models are more effective in accomplishing this task, although they are not applicable for every problem. Agency-based models are more effective for conditional forecasts used in assessing the impact of particular policies. In this way, the models are effective in forecasting complex behaviour among heterogeneous agents, mainly due to their flexibility.

3.6 Machine learning for house price estimation

Machine learning is a very important and valuable field, which is used in several fields of life, and it effectively provides valuable services in several industries. Furthermore, the machine learning technology became famous throughout the world due to face or image recognition systems, natural speech comprehension, spam detection and other technologies. Machine learning technology is also facilitating developments in the medical industry in which paramedical staff can easily diagnose patients quickly and ensure immediate medication for each condition. On the other side, it is also used in the e-commerce field to recommend the latest, most popular and appropriate products to meet customer needs. It is now determined that the machine learning technology is very effective as well because it has the capability to facilitate many different business areas. Machine learning is a technology that can be used in every field of life, including in property industry where it can be used to effectively estimate the prices of houses according to conditions provided to the system.

There are several types of studies available, which describe how machine learning is helpful in the prediction of house prices. Many scientists and machine learning developers have used different types of algorithms to develop an

effective system for the property industry to facilitate the property agents as well as the customers. The machine learning systems for estimating house prices make the process easier, and can also reduce fraudulence in house pricing.

As described by Ge et al. (2019) urban housing price prediction or estimation is a subject of interest to both academic researchers and business leaders. The researchers proposed a fine-grained model for price predictions. In their study, they described how this machine learning model can help to handle the problem of property pricing. To develop this model, the researchers used **FTD DenseNet** as well, because it incorporates more economic and social features and it also makes complete use of spatio-temporal features at all levels.

Masrom et al. (2019) described that designing the machine learning model for the classification or prediction problem is a very complicated and difficult task because it requires a lot of programming and computing knowledge and skills to develop. Furthermore, the most approporiate method of reducing this complex design is through using AML (automated machine learning), which can optimise intelligently the best suitable pipeline for the dataset. For developing the house pricing prediction model, the **Genetic Programming (GP)** algorithm is used with automated machine learning. This algorithm is a meta-heuristic algorithm, and generated the best pipeline of machine learning with reduced error and high accuracy (Masrom et al., 2019).

Sawant et al. (2018) have also worked on machine learning and developed a system to predict house prices using **Random Forest Algorithm for price prediction**. There is a great need for a machine learning based system in India because property in the coutry has been predicted to grow in the future. Pune is considered a metropolitan city, and many major companies also exist in this region, so it is an ideal place for constructing and buying a house. To satisfy the interests of buyers and sellers, the machine learning algorithm will help because they do not underestimate or overestimate the price of property. The proposed model of house price prediction helps buyers, sellers and property agents to make effective decisions by providing information based on the model (Sawant et al., 2018).

Zhao et al. (2019) have worked on a machine learning model for price prediction of properties. The name of this model is Property Appraisal, which is a very

important tool for evaluating prices and values at the time of the selling, insuring and purchasing. In the devleopment and desiging of the system, the developers used deep learning combined with **eXtreme Gradient Boosting (XGBoost)** which is effective in analysing historical sale records along with image of the houses or property sites, as well as generating the most effective and valuable results on price prediction. Furthermore, the system provides accurate results related to house prices (Zhao et al., 2019).

As described by (Zheng and Hao, 2018), housing price prediction is a very valuable task, but many buyers and sellers have to face problems due to inefficient prices given by property agents and developers. This research describes house pricing prediciton on the basis of a deep learning or AI based dynamic model and averaging model combined with a web search index. Furthermore, the combination of two approaches **DMA** and **DMS** are used for forcasting the price of houses (Zheng and Hao, 2018).

Wang et al. (2019) have also worked on machine learning for house price prediction. Their model is designed for facilitating the property industry because buyers and sellers have to face many problems at the time of selling or buying a house. The researchers used the **ARIMA** model and **SVR** method for developing the system. The use of these methods helped in predicting accurate prices for houses.

Varma et al. (2018) also described that the least transparent industry is the property industry, in our ecosystem. The prices of property are changing day to day, based on the changing valuation of properties. In this case, price determination can be difficult. To develop a model for dealing with this price problem, the researchers used **Linear Regression**, **Forest regression**, **boosted regression** and **neural network** algorithms in this research (Varma et al., 2018).

As described by (Madhuri et al., 2019), people are now more careful in purchasing properties such as land, plots or houses because sellers and buyers have limited knowledge about property prices. Due to this problem, both buyers and sellers may overestimate or underestimate the price of land. A valuable and efficient machine learning system is required which can predict the prices effectively and will be developed using different models. In this journal, the

researchers used different types of regression techniques for price prediction such as **multiple linear**, **Lasso**, **Elastic Net**, **Gradient boosting**, **linear** as well as **Adat boost regression** (Madhuri et al., 2019).

Phan (2018) has also worked on house price prediction using machine learning algorithms. In this work, Phan described different aspects related to machine learning, the price prediction problem as well as the algorithms. The price prediction for houses was a very complicated and difficult task, but the developed model can facilitate users to predict prices. As compared to others researchers, Phan used a **Support vector machine (SVM)** algorithm to solve this problem (Phan, 2018).

Manasa et al. (2020) also worked on machine learning for the predicting of house prices. Manasa et al. described that predicting models for determinination of house sale prices is a very challenging task becasuse it depends on a number of interdependent factors. In this journal, the predictive model for price evaluation on the basis of these factors was developed using regression techniques. In this research, **Ridge**, **Lasso** and **linear regression** algorithms have been used. Furthermore, the **XGBoost algorithm** is also used to enhance processing speed (Manasa et al., 2020).

(Peng et al., 2019) conducted a study analysing several types of data for enhancing price prediciton of houses in the property sector. To make their system better for price prediction, they described using multiple regression analysis algorithms – **linear regression**, **decision tree** as well as **XGboost** – to increase the accuracy of pricing for houses. These authors further described the results showing that XGboost provided the most accuracy in predicting prices compared to other algorithms. Furthermore, it also increased the speed of the process (Peng et al., 2019).

Wang and Wu (2018) have also researched price prediction strategies for properties and houses. They found many effective and accurate results by using an artificial neural network (ANN), but practical implementations have rarely been documented in the real world of such memory-based networks. A multivariable regression model with a back-propagating algorithm was designed to train an artificial neural network based on memorisers during this work. The ANN has the

potential to understand and make predictions after online testing from samples (Wang and Wu, 2018).

Jiang and Shen (2019) also provide information on the prediction of house prices using machine learning methods. In this study they were able to estimate the price of second-hand housing in Shanghai using housing data for the home network. Then, the URL text information from the json request address and the BeautifulSoup parser was parsed using crawler technology. The deeper learning library Keras was then used to construct a multi-layer feedforward neural network model trained by an error inversely propagated algorithm. The experimental findings revealed that the relative error between the prediction and the true value of the Gaussian noise models is 95.59 percent. In house price prediction, this model had a positive impact (Jiang and Shen, 2019).

3.6.1 Theory of ML in economic forecasting

Bernard Marr, an enterprise tech contributor to Forbes, describes artificial intelligence as the broader concept of machines being able to carry out tasks in a way that we would consider 'smart', and machine learning (ML) as a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves. There is also another member of this family known as 'deep learning' (DL). So, in lay terms, ML is a subset of AI, while DL is a subset of ML that enables computers to solve more complex problems (Jung et al., 2018, Tiffin, 2016).

The learning method in machine learning refers to the strategies used by the machine to determine the best fit between the variables. The algorithm applied models the relationship between the inputs and outputs within the model. The learning methods in machine learning can be categorised into supervised and unsupervised techniques. In supervised techniques, the output or dependent variables from the model are clearly known, although the specific relationship is not clear. In the supervised models, the machine tries to quantify the impact of a series of independent explanatory variables on the dependent, and these include the traditional econometric models. Unsupervised learning, on the other hand, has no specific output that is defined beforehand. The machine seeks to detect

latent or underlying patterns in the input variables. The machines learn from the datasets given to them with the aim of recognising patterns in the data and determining the output classification (Vrbka, 2016).

The machine learning algorithm is tested and validated using a training dataset and a test data set that aid in in fine-tuning the generalisations made. The generalisations and predictive power of the different machine learning algorithms are evaluated to ensure that they have low error rates and do not show 'over fitting'. Validation plays a central role in calibrating the model based on the testing applied in making it more accurate and effective. Different parameters can be tuned in the model, such as the number of trees grown in the decision tree algorithm. Different machine learning algorithms have been developed including elastic net, super learner and neural networks. These algorithms have different applications in economic forecasting and other areas of data analytics. They involve varied considerations and methods used in running the predictive analysis (Jung et al., 2018).

Economic forecasting using machine learning can be undertaken using the elastic net algorithm, which is a combination of two regression methods: least absolute shrinkage and selection operator (LASSO) and ridge regressions. They involve improvement of ordinary least squares regression through variable election and/or dimension reduction. Ridge regression involves reduction of the residual sum of squares and the shrinkage penalty. The optimal result is achieved when correlated regressors are shrunk. The minimisation problem is presented in Equation 2.4, where p is the number of explanatory variables and n is the number of observations (Tiffin, 2016).

$$\hat{\beta} = arg \min_{\hat{\beta}_j} \left[\underbrace{\sum_{i=1}^{n} (Y - X\hat{\beta})^2}_{\text{RSS}} + \underbrace{\lambda \sum_{j=1}^{p} (\hat{\beta}_j)^2}_{\text{ridge penalty}} \right]$$

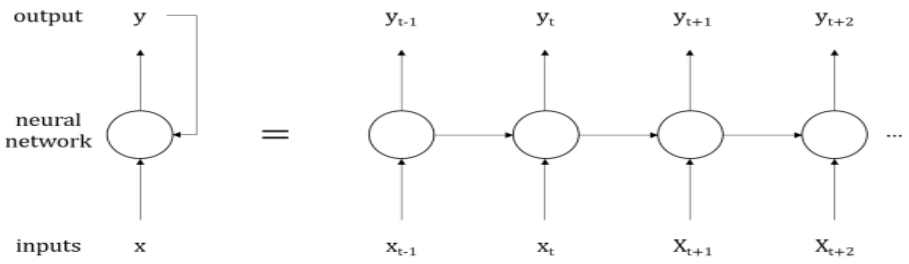
Equation 0-4: Elastic Net Algorithm

The LASSO regression uses a different shrinkage term and coefficient values of 0 are possible if the λ parameter is large enough. Additionally, the combination of the two means that the LASSO being capable of variable selection and the ridge regression lowering the coefficients close to zero means that the elastic net

algorithm combines the penalty elements to regulate its size through the previously known λ parameters. Advantages of the elastic net algorithm include its high computational efficiency and its intuitiveness. The approach is also effective in producing an output that is resilient against multicollinearity among the regressors. In a simulation study by Smeekes and Wijler (2018), elastic net was shown to be more vigorous in preventing mis-specification.

Neural networks are machine learning algorithms developed to mimic the human brain by running the inputs through learning nodes. The commonly used sigmoid neuron can process discrete and continuous inputs as well as outputs. The inputs are run through a linear or nonlinear model to produce the desired output variable. Weights are introduced for the input variables to indicate the relevance in determining the output. A large network of perceptrons is used to influence decision making involving inputs and outputs. A network of perceptrons in different layers is linked to each other in a whole system of neurons. If the information is passed from one neuron to the other in one direction, the outcome is a feed-forward neural network.

Specification of the neural network, including the number of layers and neurons, can be determined in an arbitrary manner. The academic literature suggests that the number of nodes may be located between the input and output layer sizes, with others suggesting that the neural networks should have as many hidden nodes as the dimensions necessary to capture at least 70% of the variance in the input data. One drawback of this approach is that it may contribute to the quasi-treatment of input data as cross-sectional. This feature of feed-forward neural networks makes it unsuitable for time series data since it omits the temporal component. One solution to this problem is the recurrent neural network (RNN). This is like a plain feed-forward neural network with the feature of the estimated output value being passed on to the next output value being estimated. The output of the t+1 observation depends on the output computed in observation t. The functioning of an RNN is presented in Equation 2.5.



Equation 0-5: Recurrent neural network (RNN)

RNNs may be perceived as copies of the same network ordered in a successive manner with each passing a message to its successor. They are better performing in time series problems by incorporating more layers of neurons in their predictions. Different extensions and models based on RNN have been developed, such as long short-term memory and the gated recurrent units (Cho et al., 2014). RNNs for economic applications include the Elman network, which has an additional input layer besides the multilayer perceptron. This incorporates a state space model for time series analysis. Jung et al. (2018)) used the Elman model specified with two layers of nodes. This model was selected due to its high level of forecast performance, and they developed a multivariate recurrent neural network for macroeconomic forecasting.

An ensemble of different machine learning algorithms can be developed to classify or predict new data points through a weighted vote of the learners to develop a super learner algorithm. Different approaches have been developed for the super learner including the original Bayesian averaging. The algorithm utilises cross validation to identify the combination of learners that perform best on a specific problem. Weights are assigned to the different learners and adjusted iteratively to minimise the Root Mean Square Error (RMSE). The context of the prediction problem under consideration is a key aspect of the decision making involved in choosing the ensemble of learners from different libraries.

In Jung et al. (2018), the elastic net, RNN and super learner algorithms were used in a study to predict GDP growth rates using AI. The aim was to utilise the method to evaluate the performance of the GDP forecasts in predicting adverse economic events such as the global financial crisis. The sample included the countries of Mexico, Germany, United Kingdom, Philippines, United States, Spain and

Vietnam. Data from the World Economic Outlook (WEO) from April 2017 was used in the analysis. The predictions focused on one year ahead growth forecasts of annual real GDP. The accuracy of the machine-learning-based forecasts was assessed against IMF forecasts and the actual performance of the economies. The results indicated that the use of machine learning algorithms produced more accurate GDP forecasts compared to the WEO. Additionally, the super learner and elastic net models consistently outperformed the benchmarks. The RNNs outperformed WEO forecasts only once, in the case of Philippines. The super learner algorithm was the best performer, with prediction accuracy increasing by an average of 61% for all datasets.

In Liao (2017), the applicability of artificial neural networks in time series macroeconomic forecasting was undertaken. The focus of the experiment was to assess the application of artificial neural networks in inflation rate forecasting. The study used 11 time series data sets as its baseline inputs for the analysis, including real GDP growth, stock index returns and bond spread among others, for the period 1968 Q4 to 2015 Q4. The study utilised a Markov Regime Switching Model, which proposes that the parameters of the auto-regression can be perceived as the outcome of discrete state Markov processes. The k-means method divides the set of samples into disjoint clusters separated by the mean of each sample. The performance of the k-means Markov model was found to be much better than standard time series forecasting models.

In Vrbka (2016), the application of AI through neural networks is evaluated for countries in the Eurozone. The GDP growth rate for Eurozone countries to the year 2025 is highlighted. The study used data for the period 1960 to 2015 from the World Bank. The authors generated 1000 artificial neural structures from which the 5 most appropriate were selected. The study used a sample of 11 neurons in the hidden RBF layer and 20 in the 3-layer perceptron neural network. The neural networks used were based on the radial basic function, with all showing positive and excellent characteristics regarding performance of the predictions and error. Residual analysis of the models and the resulting predictions indicated that neural networks are the most useful and effective tools for predicting GDP performance. As a result, the models were applied in predicting GDP growth rates and levels to 2025.

In Cicceri et al. (2020), the application of machine learning as a tool for evaluating economic performance and the occurrence of recessions was evaluated. The occurrence of recessions as GDP failures that are episodic and nonlinear makes them difficult to predict using ordinary stochastic models. The authors were interested in showing the relevance of machine learning as an approach for predicting economic performance and short-term forecasting accuracy. The article is a case study using data from the Italian economy for the period between 1995 and 2019. The study was comparative in nature, where it compared the GDP predictions from ML and classic linear regression models. Autoregressive models were used as benchmarks for stochastic processes varying over time. The prediction of economic performance in line with other variables was based on the average of the values of the neighbourhood of the query point.

The article also uses adaptive boost, which is an ensemble ML technique used with other algorithms to improve the final model in classification problems involving discrete data. The strength of this approach is that it builds the predictive model from the residuals of other weak predictive models. The applied model consists of decision trees that are applied to ensure that the final output corresponds with the weighted sum of the outputs from other algorithms. This approach is what is referred to as a super learner algorithm in machine learning. The training instances involve splitting of the space of the predictors to develop a set of training instances that are applied in an iterative process. It begins with the forecast of the original dataset where equal relevance is given to each observation (Cicceri et al., 2020). They also made use of nonlinear SVR, which are non-parametric ML methods applied heavily in regression analysis based on kernel functions. Due to their high applicability in classification problems, the authors had to set a margin of tolerance for the regression problem in line with the working of support vector machines. The results of the ML experiment indicated that most of the models used were effective in accurately predicting the economic crisis hitting Italy. It was clear that ML models are better suited to give recession predictions than the stochastic models. They had higher predictive power of the crises and lower error rates. Despite this effectiveness in predicting the crisis, all models missed a crucial turning point.

For Cicceri et al. (2020), the most powerful model was the Nonlinear Autoregressive with Exogenous Variables model (NARX). The NARX model had the lowest error rate and was able to accurately forecast the recession period and the two crises that hit the economy. According to the authors, the NARX model was able to predict the 2008 and 2011 economic crisis in Italy for two and one quarters, respectively, before they occurred. The main advantage of the NARX model applied in the forecasting was its effectiveness in avoiding false positives. A key conclusion was that the AR model was found to be ineffective as a predictor of economic trends in evaluating trend and variation.

The key aspects of the studies such as (Cicceri et al., 2020) and (Benigno and Thoenissen, 2002) are their use of multiple machine learning models. They are also effective in presenting the associations or differences in predictive capacity and performance of the models. Additionally, this means the studies have highlighted the variations in the performance levels of the different models for macroeconomic data. By having a wide range of models and many variables such as stock market indices, bond performance, inflation, GDP and unemployment, the models are effective in highlighting the deep relationships among the variables. The success of ML models in predicting economic performance is also influenced by the datasets used in their training as well as the testing and validation approaches.

3.7 Generating insights from multiple data points

Wang and Wu (2018) benchmarked the Random Forest machine learning algorithm with linear regression model for estimating house prices in Arlington, North Virginia. The duo observed that the Random Forest algorithm is able to capture hidden non-linear relationships among various features of a house and ultimately give a better house price estimation. Therefore, the resultant model can be used to predict future real estate prices. In their model, they included influencing factors such as zip code, location of the house, year the house was built, house price and lot size. A total of 27,649 data points were collected from Arlington County, Virginia, USA in 2015. All the data were for single-family houses. They randomly selected 30% of their dataset as test data while the rest

was used as training data. The Random Forest algorithm performed better in terms of R2 and RMSE.

Law et al. (2019) used a deep neural network model to show that street and satellite images capture elements of urban quality such as scenic features, prestige and convenience to improve house price estimation. Two types of data sets were employed: traditional housing features such as size of the house, type, size and accessibility; and images from satellite and Google Street View. Data was collected from the UK Land Registry Price Paid dataset and Nationwide Housing Society. A total of 130,557 convectional data points were collected which were matched to 40,470 street views. This study established that traditional house features account for the majority of the variance in house price, and a combined use of traditional house attributes and images depicting these features have improved price estimations. One example of an image attribute that has an impact on price is a visually desirable neighbourhood.

Ho et al. (2021) investigated the performance of three machine learning algorithms in appraising house prices in Hong Kong for a period of 18 years, to 2020. These three algorithms are Random Forest (RF), Gradient Boosting Machine (GBM) and Support Vector Machine (SVM). The performance was compared on Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) metrics. Historically, RF and GBM are known to produce better performance due to their predictive power, but Ho et al. (2021) observed that the SVM algorithm is still a useful algorithm, as it is capable of producing near accurate forecasts in a reasonable amount of time. This study used conventional house attributes that influence price such as floor area, age of the house, floor level and proximity to social amenities and CBD, as well as orientation (i.e. east, west, etc.). Approximately 40,000 data points were used in the study.

Pai and Wang (2020) undertook a study to predict real estate prices in Taiwan. The duo decided to apply advanced ML methods such as Least Squares Support Vector Regression (LSSVR), General Regression Neural Networks (GRNN), Classification and Regression Tree (CART) and Backpropagation Neural Networks (BPNN). The study employed cleansed data of 32,215 data observations and 27 attributes. Two performance metrics MAPE and NMAE were

used to compare the results of these advanced ML methods. The results showed that LSSVR is a better forecasting ML method compared to the other three.

Baldauf et al. (2020) studied whether house prices portray believed differences about climate change. They used a wide array of data sources to relate prices of individual homes with beliefs about climate changes. The outcome of the research revealed that houses in neighbourhoods believed will be underwater in the future sell at a discount. The analysis data was aggregated from six sources: Howe et al. (2015); Internal Revenue Service; North America Land Data Assimilation System; Climate Central; American Community Surveys; and Harvard Election Data Archive. Real estate is arguably the right asset class to establish whether beliefs about climate change affect the price of a houses since it is a long-term investment and important to most households.

Dohaiman (2017) compiled and analysed monthly data from S&P Saudi Arabia real estate index, Tadawul All Share index, stock returns volatility, short term interest rates from Saudi Interbank offering rate and CPI, money supply (M3 growth) and OPEC spot oil price to determine the impact of the stock market and other macroeconomic variables on real estate prices in Saudi Arabia. The paper also aims to identify the variables that led to real estate price dynamics. In Saudi Arabia, speculation in the housing market produces high return due to high income and high population density. However, after the crash of 2006 most investors are risk averse and would prefer to put their money in areas with little risk. Dohaiman (2017) analyses data using linear quantile regression. Empirical studies have shown that relationships exist between macroeconomic variables and real estate price dynamics. However, in Saudi Arabia the variables that influence real estate prices are not apparently known. This study can be used to assist investors in forecasting prices in real estate, and its results reveal that in a bullish market real estate price follows that of stock return volatility, while in a bearish market it moves alongside the price of oil.

Albuquerque et al. (2021) used GAAP accounting rules to find out whether real estate CEOs are paid for luck or reacting to luck. CEOs being paid for lucky events out of their control is seen as inefficient contracting; however, CEOs being incentivised to react to lucky events is seen as efficient contracting. Their findings revealed that compensation was for real estate CEOs' response to luck, but

challenged the idea of being paid for luck. This study used data from Execucomp database for CEO-firm-years, CRSP and Compustat databases for stock returns as well as accounting data and house price data from the Federal Housing Finance Association (FHFA).

Herland et al. (2018) focused on detecting Medicare fraud using a number of datasets from CMS (Centres for Medicare and Medicaid Services). The first data set was physicians and suppliers, prescribers, equipment and hardware. The data processing steps include data imputation, variable selection, transforming data from procedure-level to provider-level to match LEIE datasets and combining the various datasets into one database. The combined data was then held in Hadoop, as it effectively handles large unstructured data. In the analysis stage, machine learning libraries such as Random Forest (RF), Gradient Tree Boosting (GTB) and Logistic Regression (LR) were used to gauge fraud detection; Logistic Regression produced overall better performance in detecting fraud.

Guzman and Juan Silva (2018) carried out a study to determine the factors that influence the international price of copper. In mineral economics, market fundamentals (physical demand and supply) are considered to largely explain the commodity price fluctuations. However, in recent times there has been a price boom without accompanying market fundamentals. Therefore, there is some role played by non-fundamental indicators in influencing the price of copper. These non-fundamental indicators include money supply in key countries, financial speculation and financialisaton of commodities. They analysed a sixteen-variable data set of both non-fundamental and fundamental indicators using Vector Autoregression (VAR). The results showed that fundamentals are not solely the indictor of the final price of copper, but also liquidity, financial speculation and the Dollar index. This study concluded that liquidity not only affects the general level of prices of goods but also boosts demand by expecting growth in physical demand from stimulation of industries.

Weng et al. (2018) set out to develop a financial expert system that incorporates real-time or near-real-time data from the internet. In this endeavour, they acquired data from various sources such as time series stock market data, finance news and sentiments, Google search trends as well as technical indicators and

Wikipedia hits. Once the data was collected it was pre-processed by removing outliers and missing data. The cleansed data was analysed using various machine learning algorithms to produce and predict stock prices. Some of the machine learning algorithms included Boosted Regression Tree (BRT), Support Vector Regression Ensemble (SVRE) and neural network regressions. The relevance of the output was appraised using ML's MSE and MAPE performance metrics. The system produced superior performance.

Table 3.1 presents an overview of existing research that has exploited data from multiple sources for machine learning based models for house price estimation.

Table 0.1: Use of multiple datasets for insight generation

Author	Research Focus	Number of data sources	Type of insight
Wang and Wu (2018)	House price estimation using machine learning.	6	Unlike linear regression algorithm, random forest ML algorithm is able detect inherent non-linear relationship in the data set hence producing better estimations.
Law, Paige and Russell (2019)	Combined house feature with exterior images of google to estimate price of a house using deep neural network.	9	Images improve the house price estimation.
Ho, Tang and Wong (2021)	Comparison of effectiveness of various machine learning algorithms.	13	Support Vector Machine (SVM) is still relevant algorithm to predict house prices.
Pai and Wang (2020)	House price prediction using advanced machine learning algorithms such as LSSVR, CART, GRNN and BPNN.	24	LSSVR outperforms the other advanced ML methods in forecasting real estate prices.
Baldauf, Garlappi and Yannelis (2020)	Climate change effect on real estate price.	37	Real estate which are projected to be underwater sell at a discount.
Dohaiman (2017)	Real estate correlation with macroeconomics variables.	7	Stock return volatility and oil price are key determinants of real estate price in Saudi Arabia.
Albuquerque et al. (2021)	Real estate CEO compensation. Is it efficient or inefficient contracting?	3	Real estate CEOs are paid for acting on luck rather than paid for luck. Hence, efficient contracting.
Herland, Khoshgoftaar and Bauder (2018)	Fraud detection using big data and machine learning algorithms on multiple Medicare data.	3	Logistic regression is better in fraud detection in big data.

Guzman and Silva (2018)	Non-fundamentals also affect the price of copper in the world market as opposed to fundamentals only.	16	Non-fundamental macroeconomics such as liquidity and volatility index also influence the price of copper.
Weng et al. (2018)	Predicting the future of stock price	13	The model has good predictive performance.

3.8 House prices versus rental cost

The question here is: is there possibly a correlation between house prices and rental cost, and what is the impact of this on house prices? Economic theory indicates that the value of an asset is equivalent to the present value of the future income generated from it. Income earned today is worth more in present value than income that is to be earned in the future. Future income has a cost because it involves foregoing the opportunity to earn interest in the present. It is discounted at a rate reflecting the opportunity cost of the investment income. The value of the asset can be determined using the formula presented in Equation 2.6.

$$V - \frac{R - C}{r}$$

Where:

V = value of the asset

r = capitalisation rate

C= annual cost

R = rent income

Equation 0-6: Calculating the value of an asset

In the rental housing market, it can be assumed that the buyers and sellers are unsophisticated investors who use the gross income in their calculations instead of the net income (Das and Gupta, 2012). The simplified capitalisation method resulting from this view is as presented in Equation 2.7.

$$V = {}^{R}/_{r}$$

Equation 0-7: Value of an asset – simplified classification

This equation does not explicitly calculate the income stream arising from the residual value of the property at the end of its useful lifetime. However, it is

effective in reflecting investor expectations of the future changes in property values. The capitalisation rates are lower than the mortgage interest rates, which indicates a significant expectation of future growth in property value, since real estate is often used as a medium-to-long-term investment, which means that buyers should be indifferent between owning and renting a house. In this case, the value or price of a house should be strongly associated with the discounted value of its future rent income (Hargreaves, 2008).

In the study by (Hargreaves, 2008) it was found that the rent and prices of houses were strongly correlated. The correlation coefficients were found where the rental data was lagged 6 months. The key issue from the data was that the changes in rent levels are not immediately reflected in the house prices. It was found that rent moves before house prices, since rent is more closely associated with wages and salaries than house prices. Landlords increase rent as wages increase in the economy, and these increases are not necessary in line with the value or prices of houses (Shiller, 2015, Chen et al., 2014).

The size of the rental market has been found to vary significantly, especially for peripheral countries such as Spain, Greece and Ireland. According to Eurostat data, Germany was found to have the largest rental market, accounting for 47% of all households. The data also shows that the rental market only changes gradually. In a review by (Gallin, 2008) it was found that when house prices are significantly high relative to rent, the changes in real rent are larger than usual, while house prices tend to change at rates that are significantly lower than usual. In the study, a standard error correction model was used to assess whether the rent-to-price ratio had a significant predictive power (Albouy, 2016). The error correction models adopted produced inconclusive results about the predictive power of the rent-to-price ratio at a quarterly frequency The result was consistent even when the authors included a measure of the use cost of capital.

A long horizon regression approach was found to produce biased estimates of the degree of error correction if prices have a unit root and do not follow a random walk. In the presence of the bias, it was found that the rent-to-price ratio was a significant indicator of how the housing market is valued (Gallin, 2008). In the study, the Conventional Mortgage House Price Index (CMHPI) was used because it is based on the value of houses that are resold, hence it is not affected

by the composition of the homes and excludes those with mortgages. The application of the repeat sales methodology in the CMHPI was found to be upward biased since homes that are more likely to change hands tend to have stronger price appreciations.

3.9 Conclusion and research gap

The literature review indicates that there are different strategies adopted for predicting changes in house prices in the UK and other countries. Each model incorporates different features and strategies. The demand and supply elements of the prediction models have been shown to be significant in determining how house prices change in line with the economic, social and spatial variables. Quite apparent in the literature is the complexity on the subject of house price estimation and the wide range of variables that have different levels of correlation and work together to influence prices and outcomes.

Lifecycle models on price estimation for the housing sector are relevant in incorporating present and future changes in income and other socioeconomic features such as population growth, construction costs, interest rates, inflation, unemployment and spatial location. The different models adopted involve varied considerations applied to addressing the effectiveness of the estimations. These reviews indicate the need for more **complex**, **iterative** and **effective** approaches for predictions or estimations of housing prices that will take into account the wide range of features and their associations. Though some of the existing models are able to process the complexities of house pricing systems, research shows that models that are stochastic are relatively less effective than machine learning enabled models in providing accurate predictions.

In response to these weaknesses of the stochastic models, machine learning models have been developed. These include the application of neural networks, super learners and elastic nets that can be adopted in varied settings to produce robust predictions on economic performance. This review of existing literature indicates that the use of machine learning algorithms produces more accurate forecasts compared to the World Economic Outlook (WEO); the super learner algorithms and elastic net models consistently outperformed the benchmarks.

Recurrent neural networks (RNN) were also shown to outperform WEO forecasts in some cases. The super learner algorithms applied in economic predictions increase prediction accuracy by an average of 61% for all datasets. Ensemble ML techniques were used with other algorithms to improve the final model in classification problems involving discrete data. The strength of the super learner algorithms is that they build the predictive model from the residuals of other weak predictive models. The results of the ML experiments cited in (Cicceri et al., 2020) indicate that most of the models used were effective in accurately predicting the economic crisis hitting Italy.

The reviewed literature shows there have been a lot of advances in house price forecasting, and predictive models have shown some level of effectiveness in providing accurate predictions of housing prices. The trend in literature indicates that machine learning should be deployed as an alternative in order to develop more accurate predictions for the housing market that could effectively show the changes in the housing market, including issues such as affordability and identification of ideal locations to enhance return on investment. However, they present some limitations in their capacity to take an iterative approach that explores the unique impact of the varying factors that influence house price estimation in the UK. The rationale for the use of machine learning in the estimation of housing prices with due consideration for the possible impact of every factor is that it may provide better solutions for making predictions more accurate and for helping investors identify ideal locations, prices and factors that influence house prices.

In this thesis, the proposed cumulative Multi-feature House Price Estimation framework iteratively creates a total of 48 models that exploit five different machine learning algorithms by introducing layers of new groups of data, as discussed in Chapter 4. First, it is described as ‘cumulative’ because the layering approach allows the introduction of new layers without the removal of existing layers in each model. This is an essential part of this thesis, as cumulative layering tries to build towards a real-life scenario whilst assessing the impact of each layer introduced on house prices estimation and the performance of algorithms. Second, it is described as ‘multi-feature’ because the framework has leveraged ten datasets from multiple sources and has the capacity to have more

introduced by design. Third, it is a ‘layering’ framework because groups of datasets (also described as parameters in the context of this paper) are introduced as new layers into the framework.

Chapter 4: Methodology

4.1 Introduction

Further to the review of multiple research materials including conference and journal papers with a focus on how machine leaning algorithms can be used to estimate house prices, a few research gaps have been identified as discussed in Section 3.9. Therefore, this chapter, which may also be described as 'The Research Framework Design', is expected to detail the systematic approach taken in this thesis to respond to the identified gaps. The research framework design in Section 4.2 captures the overarching methodology and all embedded stages. The profiles of each of the ten datasets exploited in this thesis are fully presented in Section 4.3. Section 4.4 provides an overview of the conceptual, logical and physical modelling required to establish the relationship between the multiple datasets exploited in this thesis. To guarantee the robustness of the research framework, Section 4.5 provides a detailed explanation of the concept and importance of pipelines and Feature Union. Section 4.6 provides an overview of what modular programming is and the different modules that make up the MfHPE framework. These modules include the data ingestion module in Section 4.7 and data pre-processing module in Section 4.8. After ingestion and a bit of pre-processing, Section 4.9 then captures the initial exploratory data analysis, which provides insights into patterns and trends in the data as well as data quality issues. The engineering of the features of the research data in preparation for the development of machine learning models is detailed in Section 4.10, while Section 4.11 captures the detail of the development of the machine learning model and the algorithms explored.

Overall, this chapter provides a detailed description of the proposed multi-feature house prices estimation framework that is modularised, process-based, data driven and machine learning enabled.

4.2 Research framework design

Another look at the series of past research on the subject of house price estimation as shown in Chapter 3 reveals a number of research methods that

have been explored and similar studies on the subject of house price estimation or prediction. These include experimental (Park and Bae, 2014), comparative study (Madhuri et al., 2019) and systematic sampling (Rico-Juan and Taltavull de La Paz, 2021). Since this thesis aims to explore a cumulative layering approach for the design of a multi-feature house prices estimation framework, the *design science research methodology* will be explored. Peffers et al. (2007), cited in (Hammad, 2018) presented the Design Science Research Methodology (DSRM) with stages including: (i) problem identification and motivation; (ii) definition of the objectives for a solution; (iii) design and development; (iv) demonstration; (v) evaluation; and (vi) communication. Furthermore, (Hammad, 2018) reiterated the need for the DSRM to be implemented in an agile or iterative way, and this aligns with the approach this research is exploring.

The *first* step of the process for implementing the design science research methodology, **problem identification and motivation**, is discussed in Section 1.1 of this thesis. The *second* step, **definition of objectives** for a solution, is discussed in Section 1.5, and even more extensively in the review of existing literature in Chapter 3. The *third* step of the process, **design and development**, is the main focus of this chapter, while the *fourth* step, focused on the **demonstration** of the process-based, data driven and machine learning enabled multi-feature house prices estimation framework, is also captured in this chapter. Chapters 5 and 6 will focus on the *fifth* step, **evaluation**, and *sixth* step, **communication**, respectively.

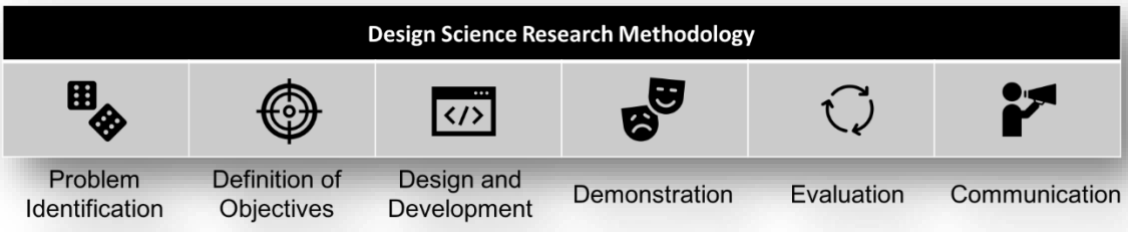


Figure 4.1: Design Science Research Methodology for MfHPE framework

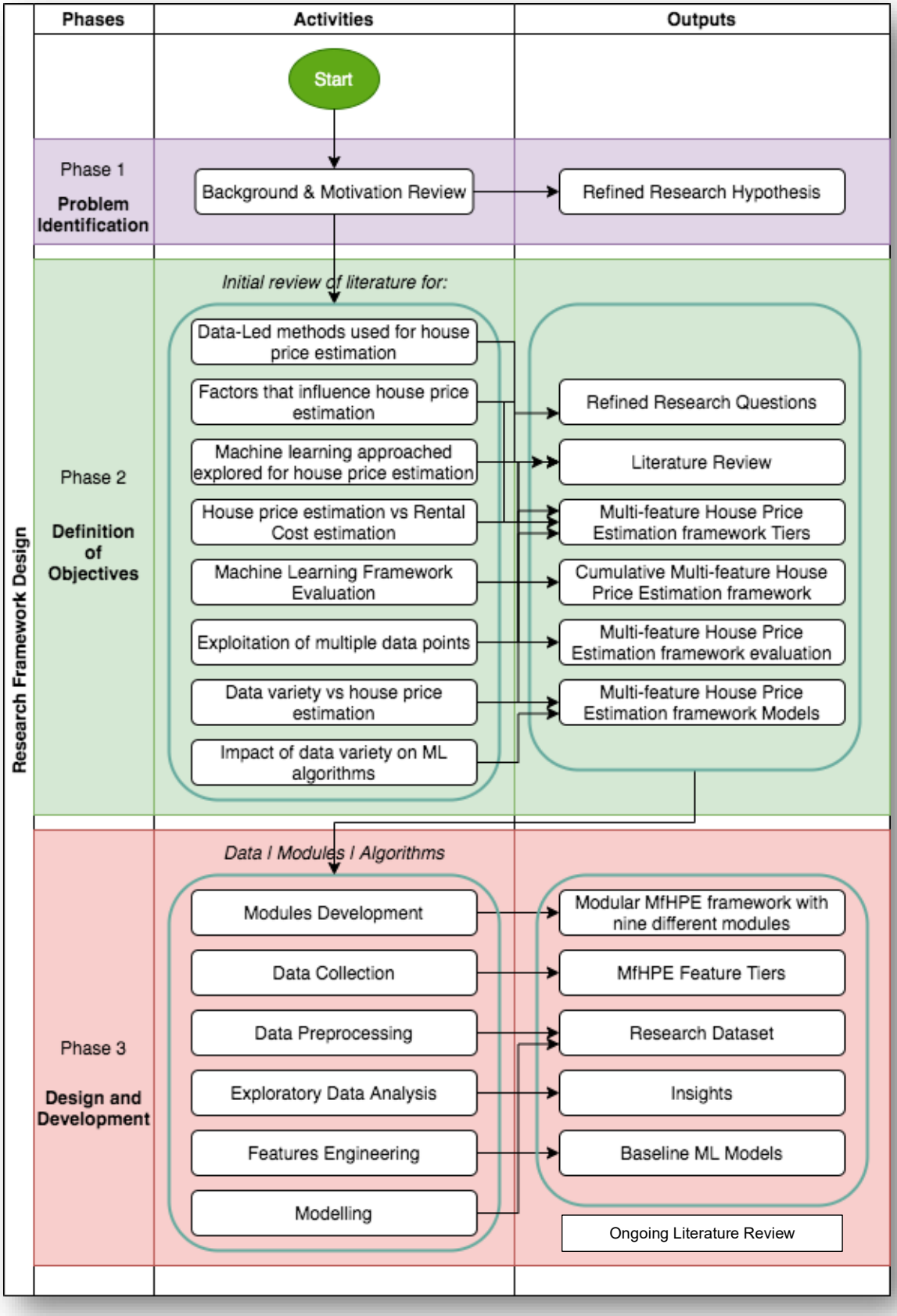


Figure 4.2: MfHPE framework (Phases 1–3)

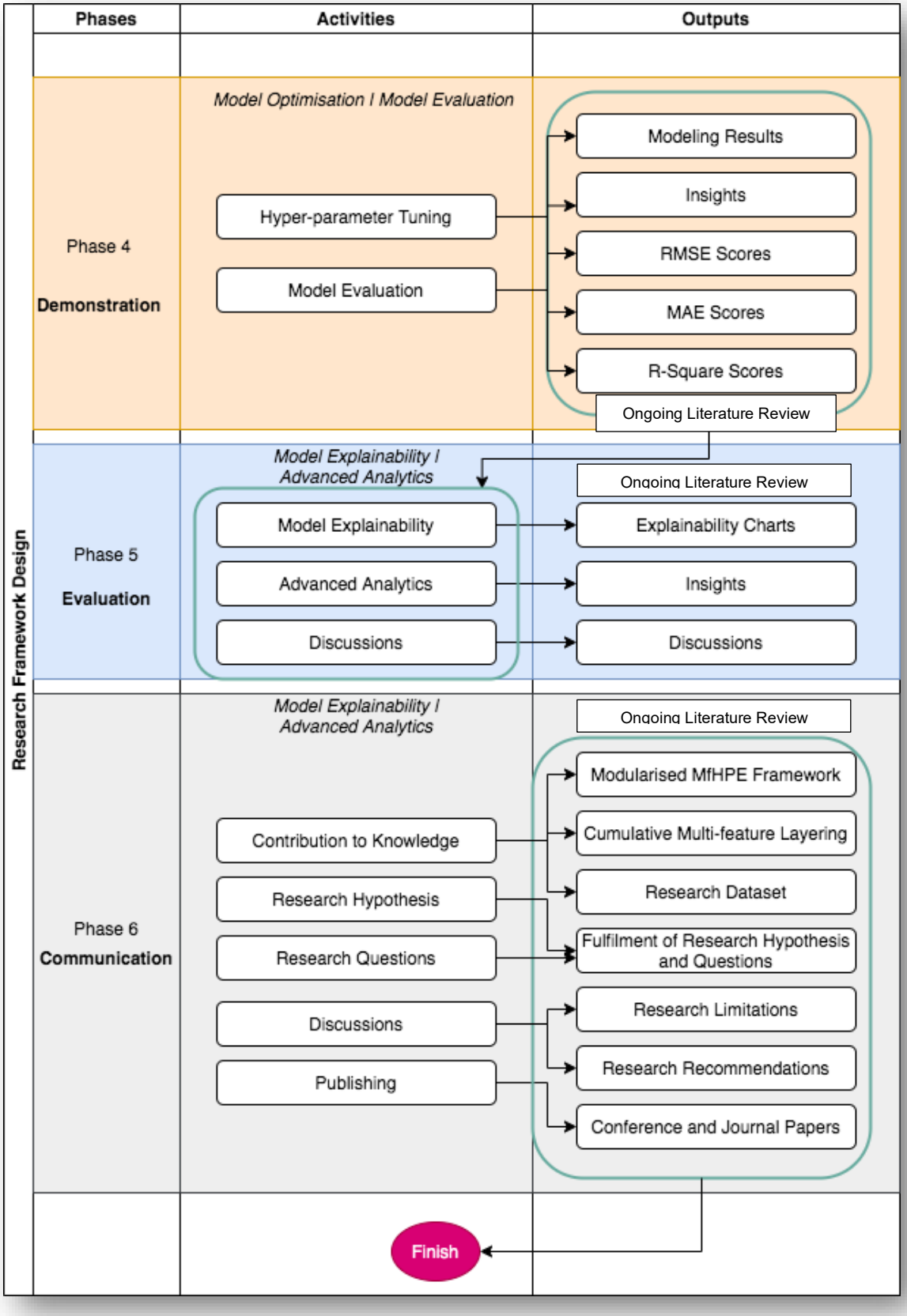


Figure 4.3: MfHPE framework (Phases 4–6)

4.3 Datasets and sources

All the datasets exploited in this research are publicly available, and a detailed profile of each has been captured below as follows: (i) Price Paid Data, Table 4.1; (ii) rail stations, Table 4.2; (iii) supermarkets, Table 4.3; (iv) bus stops, Table 4.4; (v) GDP, Table 4.5; (vi) unemployment rate, Table 4.6; (vii) employment rate, Table 4.7; (viii) inflation rate, Table 4.8; (ix) Consumer Price Index (CPIH), Table 4.9; and (x) Office of National Statistics (ONS) National Statistics Postcode Lookup (NSPL) data, Table 4.10. These datasets are then further split into tiers owing to the layering approach explored by the **Multi-feature House Prices Estimation** (MfHPE) Framework. The tiers are shown below in Figure 4.4.

Firstly, Tier 1 is comprised of the geo-coded Price Paid Data which holds locational, transactional and descriptor information on house sales. This is expected to inform the baseline estimation models in the framework, as seen in Section 4.11. *Secondly*, Tier 2 is comprised of 'neighbourhood' datasets. These include rail stations, supermarkets and bus stops. *Finally*, Tier 3 is comprised of macroeconomic indicators and will see five additional layers being considered – these include GDP, employment rate, unemployment rate, inflation rate and Consumer Price Index. Therefore, there are a total of eight layers of data, with multiple features explored beyond the location-based layer and the regular descriptors of a house being the base layers.

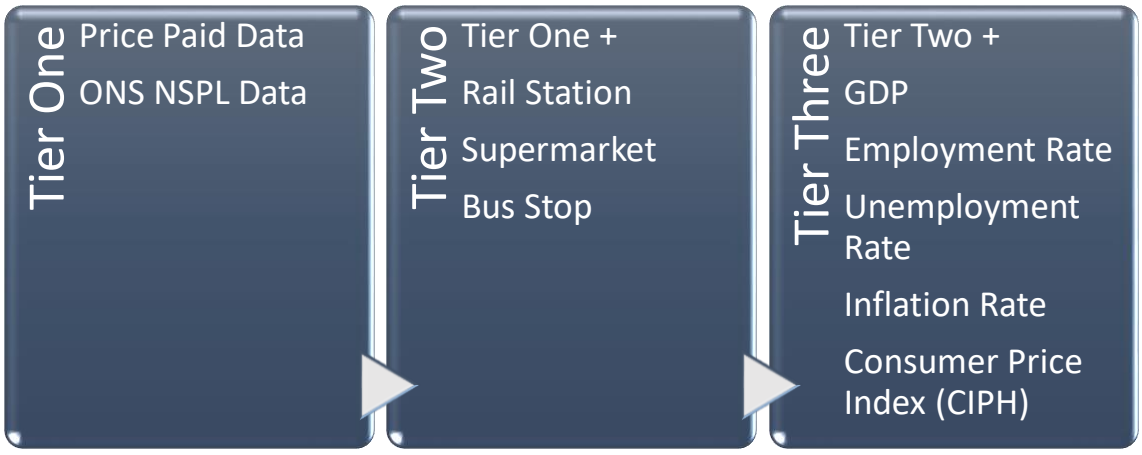


Figure 4.4: Tiers – multi-feature data enabled framework

4.3.1 Price Paid Data

The 'Price Paid Data' is owned by HM Land Registry and available for download from the central UK Government website, gov.uk. The single download file for this dataset contains over 25 million records, with 16 variables representing information on all property sales transactions in England and Wales from 1 January 1995 to date (HM Land Registry, 2021). Rico-Juan and Taltavull de La Paz (2021) in their experiment utilised a randomly selected 30% of available property data in the choice f database for pre-processing and machine learning, primarily due to volume of records. Table 4.1 shows the profile of the HM Land Registry Prices Paid Data. (Chi et al., 2019) described the HM Land Registry Price Paid Data as 'the official house price dataset in England'. This research has only used about 1.1 million records of the Price Paid Data, representing all transactions in a London borough through a defined selection process detailed in Section 4.7.

Table 0.1: Profile of complete 'Price Paid Data'

Price Paid Data			
About			
The Price Paid Data is a compilation of the submitted sale price of properties sold in England and Wales to HM Land Registry for registration. There are three options for the download of this data. These are: A single file: This contains records from 1995 to date. Monthly file: This contains records for transactions from the first to the last day of a single month. Yearly file: This contains annual data from 1995 to date.			
Cost	Free of Charge	File Size	2.4 GB
Updated	Monthly	No. of Records	25,914,816
Licence Type	OGL	No. of Variables	16
Source	https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads	File Name	pp-complete
		File Format	.csv
Variable	Description		
Transaction ID	Automatically generated reference number for each published sale		
Price	Price of sale as captured on the transfer deed		
Date of Transfer	Completion date as captured on transfer deed		
Post Code	Postcode during transaction		
Property Type	Detached (D), Semi-Detached (S), Terraced (T), Flats/Maisonettes (F), Other (O)		
Old/New	This refers to the age of the property. 'Y' for new builds and 'N' for existing properties.		
Duration	This refers to the Tenure. It can either be Freehold (F) or Leasehold (L).		
PAON	Primary Addressable Object Name. This is usually the house number or name.		
SAON	Secondary Addressable Object Name. This is applicable where buildings are subdivided. The SAON identifies the subdivided unit/flat.		
Street	*No unique description stated*		
Locality	*No unique description stated*		

Town/City	*No unique description stated*
District	*No unique description stated*
County	*No unique description stated*
PPD Category Type	Price Paid transaction type A = Standard Price Paid. This includes single residential property sold for value. B = Additional Price Paid. This includes repossessions, buy-to-lets and transfers to non-private individuals.
Record Status	This is only applicable to the monthly file. It flags additions (A), changes (C) and deletions (D) where applicable.
Attribution Statement	Contains HM Land Registry data © Crown copyright and database right 2021. This data is licensed under the Open Government Licence v3.0.

The single .csv file of the complete Price Paid Transaction Data was downloaded and ingested into through the data ingestion module of the MfHPE framework. The ingestion module is designed to then remove all records with NULL values for postcode, and then identify records with the 'District' value being a London borough AND 'Date of Transfer' between 1st January 2011 and 31st December 2020. This translates to 1,097,302 records of individual residential property sale transactions across London being used for the model design.

4.3.2 GB Rail Stations

The 'GB Rail Stations' dataset is a list of all the train stations in the United Kingdom. For each train station the attributes held in the dataset are as shown in Table 4.2. The volume of passengers travelling through each station either as the start or end of their journeys, or even as an interchange, is captured. Therefore, this research will be exploring the possible impact of the location of train stations (with a focus on the distance to nearest station and the number of train stations within a two-mile radius) on house prices.

Table 0.2: Profile for rail stations, including underground

GB Rail Stations			
About			
This dataset is a list of all Great Britain stations. From source, this data can be downloaded in two different formats; .csv or .kml			
Cost	Free of Charge	File Size	603 KB
Licence Type	Unknown	No. of Records	2,569
Source	https://www.doogal.co.uk/UkStations.php	No. of Variables	39
File Format	.csv	File Name	GB stations
Variable	Description		
Station	This is the station name		
Postcode	Postcode allocated to station		
Latitude	Latitude projection based on WGS84 format		

Longitude	Longitude projection based on WGS84 format
TLC	Three-letter abbreviation for identifying station
NLC	No definition found for this variable
Owner	Owner and operator of the station
Entries and Exits	Number of entries and exist from station per year. There are 16 variables for 2005–2020.
Interchanges	Number of interchanges at station per year. There are 16 variables for 2005–2020.

4.3.3 Supermarket location

Retailers are opening up in multiple locations and in different formats, thereby providing customers with choice. With so many new stores, it becomes relatively easier to know where the competition is, and consequently the new markets being targeted by retailers. This research will be exploring the impact of the location and additional data of these supermarket on house prices.

Though the required information is not widely available in a single web-based source, Geolytix has created this single source of data for supermarkets, as it then serves as the baseline dataset for other projects Geolytix delivered being United Kingdom Supermarket 'Retail Points' is a database of supermarkets with store names and addresses with postcodes. It also has a range of other attributes including opening date (where known) and a four-way classification for size band, as shown in Table 4.3.

Table 0.3: Profile for supermarket locations

UK Supermarket Retail Points			
About			
This dataset is produced by GEOLYTIX. It contains location data for over 10,000 supermarket retailers like Tesco, Marks and Spencer, Waitrose, Asda, Wholefoods, Aldi, Lidl, and more. Contains public sector information licensed under the Open Government License v3.0 Contains Ordnance Survey data © Crown copyright and database right 2020. Contains Royal Mail data © Royal Mail copyright and database right 2020. Contains National Statistics data © Crown copyright and database right 2020. From source, this data can be downloaded in three different formats: .txt, .csv or .xls			
Cost	Free of Charge	File Size	3.9 MB
Licence Type	GEOLYTIX Retail Points ODL	No. of Records	16,991
Source	https://www.geolytics.com	No. of Variables	17
File Format	.csv	File Name	Geolitixs_retailpoints_v19_202102
Variable	Description		
Id	10 digit Geolytix Unique Identifier. This is prefixed by 101		
Retailer	Retailer name		
Facia	Name 'above the door'		
Store_name	Store name as shown on website		
Add_one	First line of address		
Add_two	Second line of address		

Town	Town store is located. This is based on a Geolytix definition.
Suburb	Suburb store is located. This is based on a Geolytix definition.
Postcode	Royal Mail format full postcode
Long_wgs	Longitude based on WGS84 format
Lat_wgs	Latitude based on WGS84 format
Bng_e	Eastings in British National Grid format
Bng_n	Northings in British National Grid Format
Pqi	Postcode Quality Indicator: PQI 1 – Rooftop geo-coded by Geolytix PQI 2 – Rooftop geo-coded by third party PQI 3 – Postcode geo-coded
Open_date	Date store opened (YYYYMMDD)
Size_band	A – Less then 3,013sqft (280msq) B – 3,013 – 15,069sqft (280msq – 1,400msq) C – 15,069 – 30,138sqft (1,400msq – 2,800msq) D – 30,138sqft + (2,800msq +)
County	County store is located. This is based on a Geolytix definition.

4.3.4 Bus stop

This dataset is published to the National Public Transport Data Repository by the Department for Transport. The data in this repository is available for the period October 2004 to October 2011. However, it is now static and superseded by the Traveline National Dataset. Table 4.4 provides a detailed profile for the bus stop data used for this research.

Table 0.4: Profile for bus stop data

Bus Stops			
About			
It is a compilation of: local public transport information from each of the traveline regions; coach services from the national coach services database; rail information from the Association of Train Operating Companies (ATOC). From source, this data can be downloaded as .csv			
Cost	Free of Charge	File Size	85.3 MB
Licence Type	Open Government License	No. of Records	406,873
Source	https://data.gov.uk	No. of Variables	25
File Format	.csv	File Name	Bus Stop
Variable	Description		
ATCOCode	Unique ATOC code for bus stops		
GridType	No description available		
Easting	Eastings in British National Grid format		
Northing	Northings in British National Grid Format		
Lon	Latitude projection based on WGS84 format		
Lat	Longitude projection based on WGS84 format		
CommonName	No description available		
Identifier	Identifier for bus stop. Usual a sign on a post		
Direction	Direction of travel		
Street	Street Name		
Landmark	Landmark identifiable with bus stop		
NatGazID	*No description available*		
NatGazLocality	*No description available*		
ParentLocality	*No description available*		

GrandParentLocality	*No description available*
Town	Town Name
Suburb	Suburb name where applicable
StopType	*No description available*
BusStopType	*No description available*
BusRegistrationStatus	*No description available*
RecordStatus	*No description available*
Notes	*No description available*
LocalityCentre	*No description available*
SMSNumber	*No description available*
LastChanged	*No description available*
ATCOCode	*No description available*
GridType	*No description available*

4.3.5 GDP

GDP is an acronym for Gross Domestic Product and is described by (Investopedia, 2021a) as the monetary value of all finished goods (manufactured) and services (provided) within the geographic boundaries of a country over a specific timeframe. There are three known approaches to calculating this macroeconomic indicator – these use (i) expenditures, (ii) production or (iii) incomes.

GDP is one of the macroeconomic indicators being explored in this research for the estimation of house prices because it is a leading tool used to guide strategic decision making for business leaders, investors and policy makers, despite its limitations.

Table 4.5 provides a detailed profile for the ONS GDP data downloaded for this research.

Table 0.5: Profile for GDP data

GDP			
About			
The Gross Domestic Product (GDP) is a measure of the size and health of a country's economy over a period of time (usually one quarter or one year) (Bank of England, 2021). The size of economies can also be compared based on this macroeconomic indicator. This data is produced by the Office of National Statistics (ONS) quarterly. Each file contains records from quarter 2 (Q2) of 1955 to date. From source, this data can be downloaded in three different formats; image, .csv or .xls			
Cost	Free of Charge	File Size	4 KB
Updated	Quarterly	No. of Records	272
Licence Type	OGL	No. of Variables	2
Source	www.ons.gov.uk	File Name	Series-050321
Unit	%	File Format	.csv
Variable	Description		
Period	Quarter of the year		
Value	GDP value in percentage		

4.3.6 Unemployment rate

The UK unemployment rate is described as the rate of unemployment measured by the Labour Force Survey (LFS), and based on the International Labour Organisation's definition of unemployment (Office of National Statistics, 2021c). The profile of the data is as detailed in Table 4.6.

Table 0.6: Profile for unemployment rate data

Unemployment Rate (aged 16 and over, seasonally adjusted)			
About			
The Office of National Statistics (ONS) is responsible for the measurement of the rate of unemployment in the United Kingdom. The data is produced monthly and with average quarterly and yearly values from 1971 to date. From source, this data can be downloaded in three different formats; image, .csv or .xls			
Cost	Free of Charge	File Size	12 KB
Updated	Monthly	No. of Records	848
Licence Type	OGL	No. of Variables	2
Source	www.ons.gov.uk	File Name	Series-050321
Unit	%	File Format	.csv
Release Date	23/02/2021	Source Dataset ID	LMS
Variable	Description		
Period	Month, Quarter or Year		
Value	Unemployment rate value in percentage		

4.3.7 Employment rate

According to the (OECD Employment Outlook, 2020), employment rate is described as a measure of the extent to which people available to work are being engaged. It is calculated as the ratio of 'the employed' to 'the working age population'. The profile of the data is as detailed in Table 4.7.

Table 0.7: Profile for employment rate data

Employment Rate (aged 16 and over, seasonally adjusted)			
About			
The Office of National Statistics (ONS) is responsible for the measurement of the rate of employment in the United Kingdom. The data is produced monthly and with average quarterly and yearly values from 1971 to date. From source, this data can be downloaded in three different formats; image, .csv or .xls			
Cost	Free of Charge	File Size	12 KB
Updated	Monthly	No. of Records	848
Licence Type	OGL	No. of Variables	2
Source	www.ons.gov.uk	File Name	Series-050321
Unit	%	File Format	.csv
Release Date	23/02/2021	Source Dataset ID	LMS
Variable	Description		
Period	Month, Quarter or Year		
Value	Unemployment rate value in percentage		

4.3.8 Inflation rate

The Office of National Statistics (ONS) describes inflation rate as the change in prices for goods and services over a period of time (Office of National Statistics, 2021b). It also states that the measures of inflation include the House Price Index (HPI), producer price inflation and consumer price inflation. However, (Investopedia, 2021b) takes the description of the inflation rate one step further, beyond just the change in price of goods and services, to include the rate the value of currencies falls.

The rate of inflation can be relative positive or negative, depending on the perspective of the individual and the rate of change. Investors in tangible assets, like houses (residential and commercial) or stocked commodities, may favour some inflation, as that consequently raises the value of their investments/assets. However, individuals or groups holding cash are unlikely to favour inflation, as it weakens the value of their cash. Overall, an optimum level of inflation is essential as it promotes spending rather than saving, thereby fostering economic growth.

Table 4.8 shows the profile of the inflation rate data downloaded from the ONS website.

Table 0.8:Profile for inflation rate data

Inflation Rate			
About			
The Office of National Statistics (ONS) is responsible for the measurement of the inflation rate in the United Kingdom. It is also identified as CPIH Annual Rate. The data is produced monthly and with average quarterly and yearly values from 1989 to date. From source, this data can be downloaded in three different formats; image, .csv or .xls			
Cost	Free of Charge	File Size	7 KB
Updated	Monthly	No. of Records	545
Licence Type	OGL	No. of Variables	2
Source	www.ons.gov.uk	File Name	Series-050321
Unit	%	File Format	.csv
Release Date	17/02/2021	Source Dataset ID	MM23
Variable	Description		
Period	Month, Quarter or Year		
Value	Inflation rate value in percentage		

4.3.9 Consumer Price Index (CPIH)

The known most comprehensive measure of inflation is the CPI including owner occupiers' housing costs (CPIH). The CPIH extends the Consumer Price Index (CPI) to include a measure of the costs associated with owning, maintaining and

living in one's own home. This cost is known as owner occupiers' housing costs (OOH), which includes Council Tax. These two costs are assessed as significant expenses for many households but are not included in the CPI. Therefore, in the bid to have as much relevant household expenses considered as possible, the CPIH is considered for this research rather than CPI. Table 4.9 details the profile of the CPIH data.

Table 0.9: Profile for Consumer Price Index – CPIH

Consumer Price Index – CIPH			
About			
The Office of National Statistics (ONS) is responsible for the measurement of the consumer price index in the United Kingdom. It is described as the measure of changes in price levels of a weighted average market basket of consumer goods and services bought by households. This data is produced by the Office of National Statistics (ONS) monthly. Each file contains records from 1989 to date. From source, this data can be downloaded in three different formats; image, .csv or .xls			
Cost	Free of Charge	File Size	7 KB
Updated	Monthly	No. of Records	548
Licence Type	OGL	No. of Variables	2
Source	www.ons.gov.uk	File Name	Series-140521
Unit	%	File Format	.csv
Release Date	21/04/2021	Source Dataset ID	MM23
Variable	Description		
Period	Month, Quarter or Year		
Value	CPI value in percentage		

4.3.10 ONS postcode data

To combat the shortcomings of the HM Land Registry Price Paid Data raised by (Chi et al., 2019), this research has blended the Land Registry data with the ONS NSPL product to create a latitude and longitude variable. The result of this is a fully geo-referenced version of the HM Land Registry Price Paid data. The profile of the data is detailed in Table 4.10.

Table 0.10: Profile for National Statistics Postcode Lookup

ONS Postcode Data			
About			
The Office of National Statistics (ONS) produce two main postcode products. These are (i) ONS Postcode Directory (ONSPD) and (ii) National Statistics Postcode Lookup (NSPL). These products are widely used by a range of customers including central and local government, commercial organisations and academia (ONS, 2021).			
Cost	Free of Charge	File Size	1.13GB
Updated	Monthly	No. of Records	2,661,131

Licence Type	OGL	No. of Variables	41
Source	www.ons.gov.uk	File Name	NSPL_MAY_2021_UK
Resolution	1 meter	File Format	.csv
Variable	Description		
Pcd			
Pcd2			
Pcds			
Lat	Latitude based on WGS84 format		
Long	Longitude based on WGS84 format		

4.4 Data modelling

Data modelling aims at communicating the connections between data structures and data points through a process that creates visual representations of parts of an information system. Through this process, the relationships between the data used and stored can be organised in accordance to their attributes and format (Ribeiro et al., 2015). Data modelling employs the use of symbols, diagrams and text to represent the way which data interrelates.

Since the structure of data modelling imposes itself on a data system, the process of data modelling improves the consistency of naming in data, the rules and data security, while optimising data analytics. The model also highlights what further data is needed and how it should be organised. However, data modelling does not dictate the actions that a data architect can perform on the data (Cariou, 2020).

In this thesis, the goals of this section can be defined as: ensuring that all data objects exploited are accurately represented; defining the relationship between data tables; identification of primary and foreign keys; giving a visual representation of base data that can subsequently be used to create a physical database; identifying redundant or missing data; making subsequent data infrastructure maintenance and upgrade faster and cheaper.

There are three main types of data modelling, which serve different purposes and bear their own advantages: conceptual data modelling; logical data modelling; and physical data modelling (Gaur, 2020). Sections 4.4.1 to 4.4.3 capture a generic overview of these types and then then their context in line with this thesis.

4.4.1 Conceptual data model

Conceptual data modelling is also referred to as domain modelling. It gives a high-level view of what a data system will contain, its organisation and the rules used in the data structure (Ribeiro et al., 2015). The rules and organisation in a data structure include entity classes, their constraints and characteristics, the relevant security and integrity requirements as well as the relationships that are forged between the entity classes. Figure 4.5 shows how all ten datasets exploited in this thesis are connected conceptually.

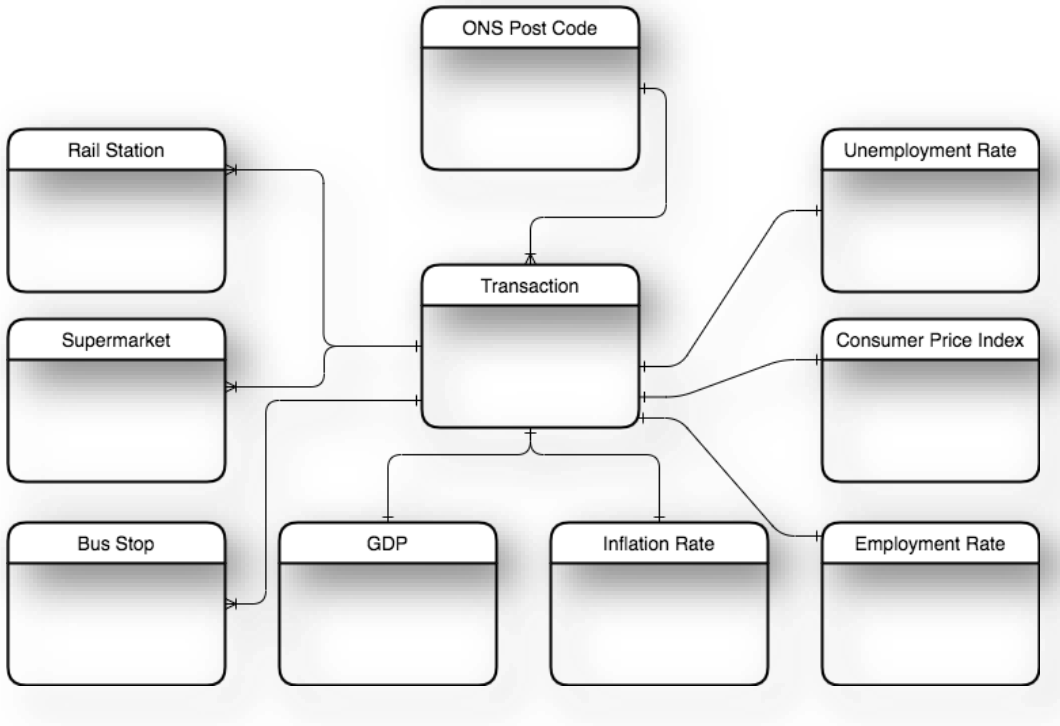


Figure 4.5: MfHPE framework conceptual data model

4.4.2 Logical data model

Logical data models are a refinement of conceptual data models. They detail domain relationships and entities while acting as stand-alone and platform-independent models. Logical data models are less abstract compared to conceptual data models, and provide in-depth detail of the concepts and

relationships of a domain (Gaur, 2020). Figure 4.6 shows the developed logical data model for the cumulative Multi-feature House Price Estimation framework.

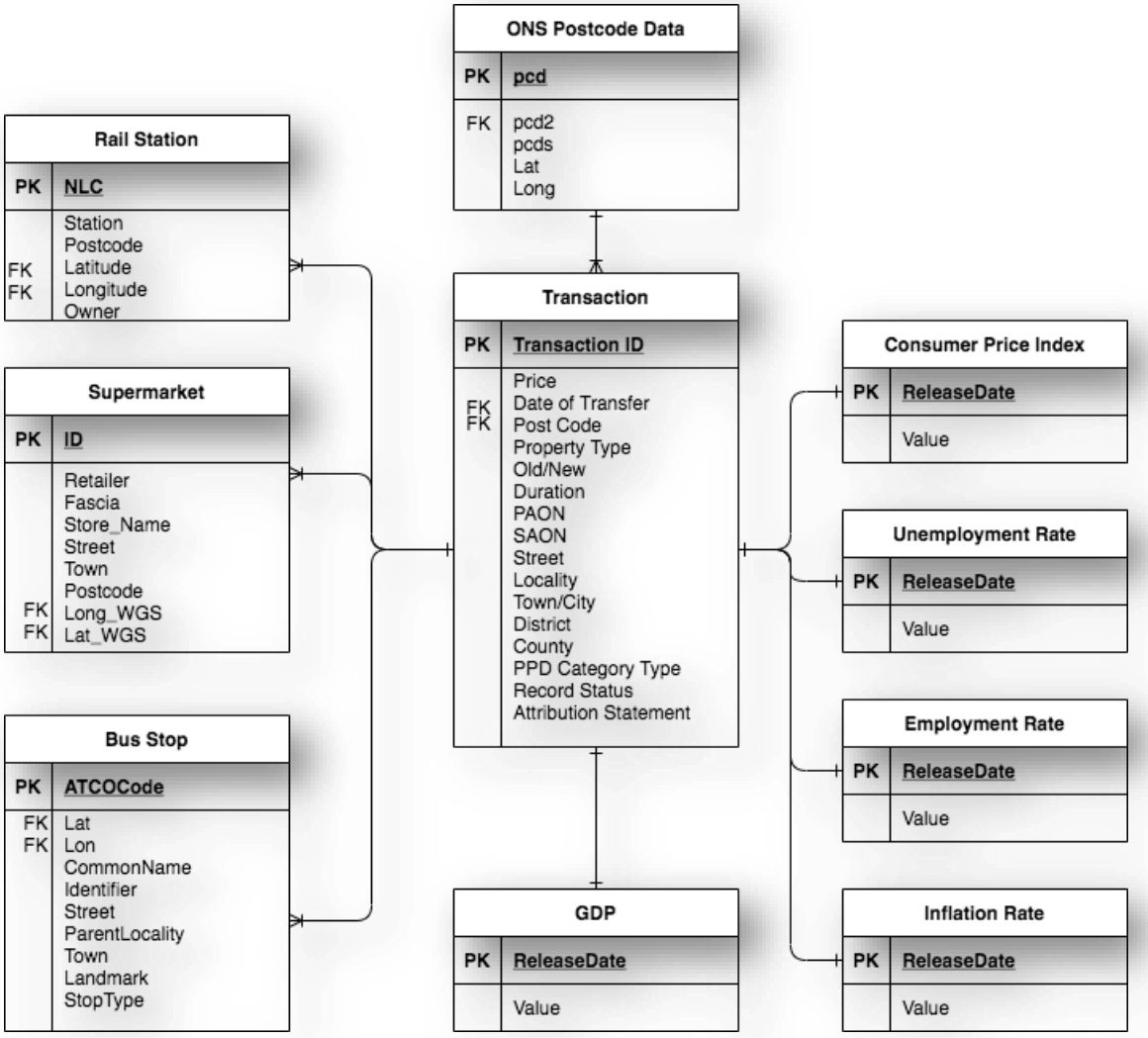


Figure 4.6: MfHPE framework logical data model

The logical data model does not specify the technical requirements that the build of the research data requires, but depicts the attributes of every entity – such as each entity's unique identifier and primary key – and the relationship between the said entities – the keys identifying the relationship between entities and the foreign keys (Ribeiro et al., 2015). They are more useful in environments that have highly procedural implementation requirements and projects that are highly data oriented.

4.4.3 Physical data model

Physical data modelling is an approach that applies data specific modelling. It is ideally used for a specific project, but it can be integrated into other physical models for a comprehensive view. Physical data models give more details on the column constraints, column keys and the primary and foreign keys (Cariou, 2020). The physical data model is therefore critical in the design and development of the schema of the data model for the research data, and ultimately for the MfHPE. Figure 4.7 shows the developed logical data model for the cumulative Multi-feature House Price Estimation framework.

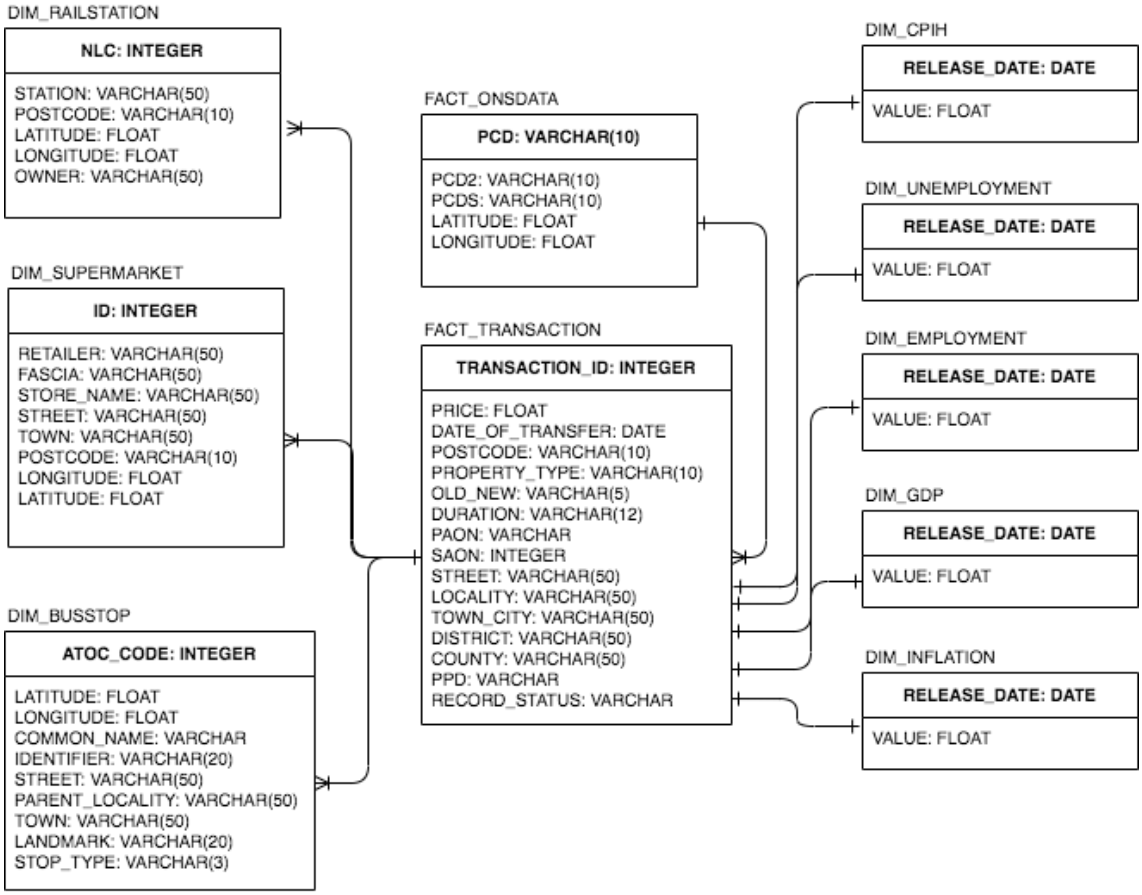


Figure 4.7: MfHPE framework physical data model

As mentioned above, Gaur (2020) stated that physical data modelling differs from both logical and conceptual data modelling because it is platform specific; it reflects the database schema of a singular data system, thus giving the

advantage of showing platform-specific attributes such as query language extensions and database-specific data types.

4.5 Pipeline and Feature Union for MfHPE

There are many data transformation steps required for the effective running of the MfHPE framework. Even more important is the fact that these steps will be required to be executed in a defined order. Therefore, the development of the MfHPE framework has leveraged the Scikit-Learn *Pipeline class*.
Scikit-Learn is free and open-source library for machine learning based on Python programming language, although it is written in a combination of languages (Python, Cython, C and C++). The library is built upon NumPy, SciPy and MatplotLib Python based libraries, and it also integrates well with these supporting libraries. Therefore, it is prudent to be familiar with these libraries before delving into Scikit-Learn. This library features various regression, clustering, classification, dimensionality reduction, model selection and pre-processing algorithms such as K-Means, feature selection, non-negative matrix factorisation, metrics, grid search, cross-validation, pre-processing, spectral clustering, mean-shift, SVR, SVM, random forest and nearest neighbours, among others (Scikit-learn, 2021a).

4.5.1 Pipeline

A pipeline is a collection of transformers followed by an estimator (an estimator is an object that carry out a fit and transform methods). Essentially, pipelines encapsulate sequences of estimators into one for convenience purposes. Figure 4.8 provides a high-level view of the implementation pipeline for the cumulative Multi-feature House Price Estimation framework. Pipelines are useful for the following reasons: compactness – writing few lines of codes; clarity – easy to write and visualise; ease of handling; and joint parameter selection.

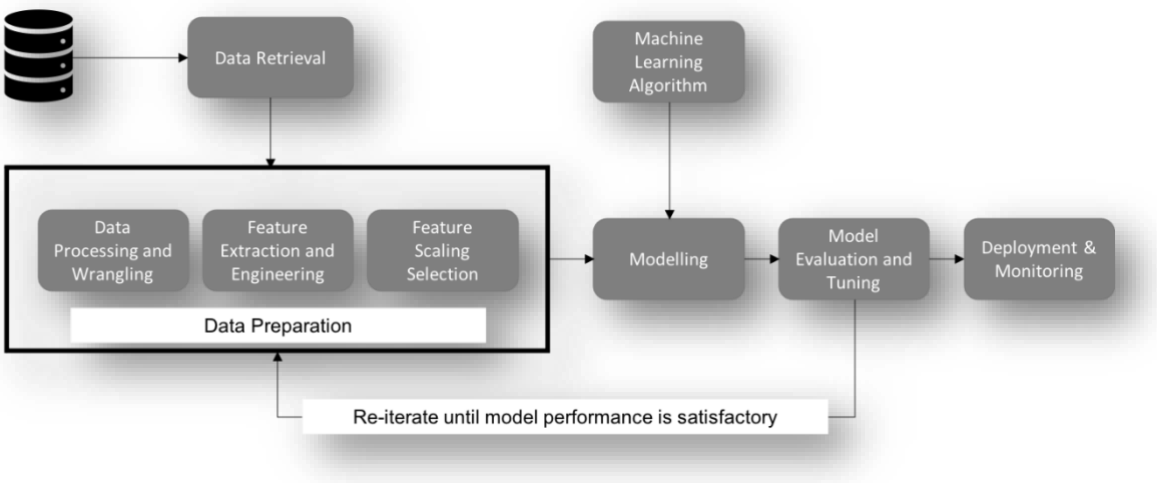


Figure 4.8: A high-level presentation of MfHPE implementation pipeline

4.5.2 Feature Union

Feature Union concatenates (joins) the outputs of multiple transformers. For instance, column transformers are used to transform data in distinct columns. Hence, Feature Union will come in handy in concatenating the outputs of each column. Essentially, it allows for combining different feature extraction transformers into one transformer (Scikit-learn, 2021b). Scikit-Learn is a Python library which has received widespread acceptance in ML. This library contains pipeline and Feature Union methods for encapsulating and concatenating the outputs of different frameworks, respectively. Scikit-Learn is widely used in various areas such as medicine, weather, population and housing, and has been used for the implementation of the cumulative Multi-feature House Price Estimation (MfHPE) framework.

Several pieces of research have focused on house price estimation that used pipelines and features union. Hao and Ho (2019) identified the features that make the Scikit-Learn library stand out among many pieces of machine learning software. First, Scikit-Learn has a wide coverage of ML methods. This coverage is informed by a community review procedure which helps to identify which methods to include and which to discard or leave out. Therefore, a balance is struck between extensive coverage and selectivity of ML. Second, the algorithm implementation is optimised for computational efficiency. Third, Scikit-Learn is backed by strong community support for quality assurance, bug tracking and documentation. Finally, it maintains a uniform data input/output convention,

hence switching from one method to another is effortless. In relation to machine learning, Scikit-Learn covers data transformation by NumPy data structures, supervised learning, unsupervised learning and model evaluation and selection.

Thamarai and Malarvizhi (2020) carried out research on house price prediction using Scikit-Learn's decision tree classification, decision tree regression and multiple linear regression machine learning algorithms. The attributes of the house considered in the study were number of bedrooms, age of house, travelling amenities, availability of school facilities and availability of shopping mall nearby. The sample dataset was split into training and testing datasets in the ratio of 80:20 using the Scikit-Learn tool. The performance evaluations were measured with Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results showed that Multiple Linear Regression (MLR) has a better performance compared to decision tree regression in predicting the house prices. The framework proposed by this research also leverages similar features, machine learning algorithms and evaluation metrics but the unique difference is that whilst their approach lumped all features into single models based on the machine learning technique, this research will create models by cumulatively introducing groups of features.

Truong et al. (2020) explored the various house features in predicting house prices using both traditional approaches and advanced machine learning algorithms. They also validated multiple techniques in model implementation on regression and hence gave a more optimistic house price prediction. The Scikit-Learn library was used to pre-process data, split between training and testing data and also in classifying and clustering using RandomForestClassifier class by Scikit-Learn. The classifier was specified to as many as 900 trees, with a maximum tree depth of 20 and at most 10 rows before a tree could be split. Although XGBoost and LightGBM exist as separate packages from Scikit-Learn, a function from Scikit-Learn was used extensively to fine-tune the results of various algorithms. The results showed that Scikit-Learn RandomForestClassifier outperformed the other algorithms in accuracy of the results.

Biswas and Rajan (2021) observed that machine learning models exhibit discrimination towards people based on age, sex, race and locality, among other factors. Although research has been conducted to enumerate and mitigate

unfairness in machine learning, most of the research has considered a single classifier solution. Some studies have shown that the root cause of unfairness is ingrained in the data itself instead of the model. Therefore, Biswas and Rajan (2021) pursue fairness in data with a ML pipeline. Their ML pipeline was adopted from the canonical definition of a pipeline from Scikit-Learn specifications: a ML pipeline is an ordered set of m stages with a set of pre-processing stages (S_1, S_2, … S_{m-1}) and a final classifier (S_m). Each data pre-processing stage acts on the output data of the preceding stage. A pre-processing stage can be a data transformer or custom operations. A data transformer is a method or an algorithm that performs a specific operation on a set of data, such as variable encoding, dimensionality reduction, feature extraction and selection. The results show that many stages in data pre-processing induce unfairness/bias in the prediction model. Therefore, fairer or custom pipelines need to be constructed.

Sugimura and Hartl (2018) found that one of the weakness of Scikit-Learn is irreproducibility, meaning that the scope and the design are restricted to a single model. Therefore, developers who need their work to be reproducible are greatly restrained by Scikit-Learn.

4.6 Modular programming in the MfHPE framework

Modular programming is a concept that originated and is widely applied in computer science. It refers to a software design technique that entails separating functions of a program into independent, interchangeable modules whereby each module executes only one aspect of the desired functionality (Lavy and Rami, 2018). Other terms that refer to modular programming are functions, modules, procedures or sub-routines. The essence of modules is to allow for the breakdown of complicated programs into smaller and more manageable programs. Also modules are reusable by other programs. Modular programming has been used in other industries besides computer science. Some of the recent applications include: enabling communication and synchronisation in parallel programs (Veen and Jongmans, 2018); spam detector (Subhan et al., 2021); tuberculosis control modelling, given complexity and heterogeneity of cases

(Trauer et al., 2017); and creating hybrid composites in polymers (Feng et al., 2019).

As shown in Figure 4.9, the modules of the MfHPE framework include: (i) **assets** – holding all raw data to be ingested, as detailed in Figure 4.4; (ii) **data ingestion** – comprising functions designed to ingest all research data from the asset module and also extract the specific records required as baseline; (iii) **data processing** – this is made up of the functions designed for data cleansing and initial exploratory data analysis; (iv) **features engineering** – this caters for the transformation of data across all three tiers; (v) **model building** – this is where the machine learning algorithms are exploited on the training data, test data and validation data; (vi) **params** – this module serves as the store for multiple dictionaries created to hold groups of data that belong to the same tier; (vii) **utils** – this comprises classes and functions designed to handle processing tasks like the *label encoding* of categorical features and *Feature Union* of numerical features; and (viii) **Main.py** is the primary environment where the functions and classes across other modules are called into action.

Lavy and Rami (2018) assert that to be able to develop modular program, one should be able to process abstract thinking abilities in deconstructing the solution into logical parts, and eventually create a complete solution by integrating the individual parts. The quality of code is determined by the level of modularity employed. Minimal modularity could reflect laxity, lack of coding skills to design modularity from the start, or unawareness of the principle. (Lavy and Rami, 2018) aimed to identify the circumstances under which beginner programmers will use modularity in their code. The results were explained by two psychological theories: dual process theory and cognitive dissonance theory. One of two groups of students was asked to handle an incremental problem task, while the other group had full scope of the problem from the start. Most of the participants in the second group realised the necessity of modularity, while only 15% of the first group found modularity necessary. The actions of the first group, of minimal use of modularity in their code, is explained by dual-process theory that asserts that people will tend to apply minimal effort on the problem they confront despite compromising the quality of the final results. Even after the study group had learned and understood the importance of modular programming, they refused to

revise their work to maintain quality, evoking cognitive dissonance. Therefore, a programming team should be provided with the full scope of their work from the beginning to avoid mental discomfort that comes with reworking.

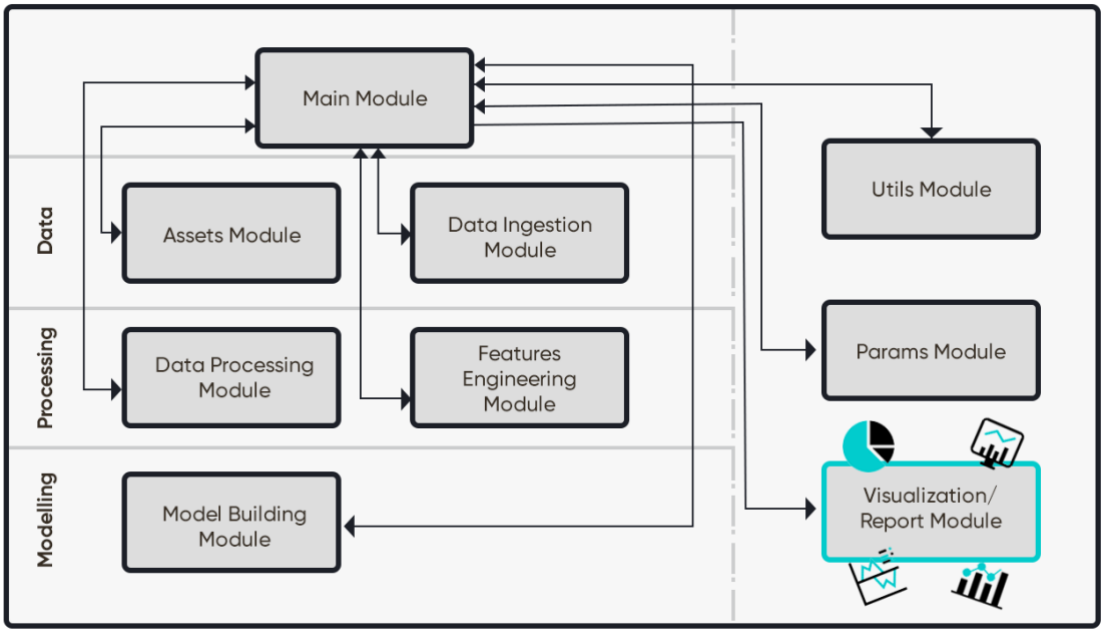


Figure 4.9: MfHPE framework modules

Veen and Jongmans (2018) used modular programming to implement communication and synchronisation among tasks in parallel. Their approach used high-level DSL by generalising the existing protocol language Reo, from supporting only a set number of tasks to a dynamic number of tasks. The new modular programming using Reo language outperforms the existing approach, although the new approach requires more work to be done at run-time. This study used experimentation with Reo backed program packages and the existing program.

Modular programming principles have also inspired development of non-IT products. Dennis (2003) proposed a multiprocessor chip guided by principles of modular programming. Multiprocessors are programmed differently than modular-based programs. To improve the efficacy of multiprocessor chips it is important to incorporate principles of modular software. He envisioned a multiprocessor architecture incorporating simultaneous multithreading, the use of shared addresses and no update of the memory. The six principles that are widely used in modular software and neglected in multiprocessor chips are: hiding

information; invariant behaviour; data generality; secure arguments; recursive construction; and system resource management. Finally, Dennis (2003) illustrated the use of array data structures, linear system solvers and job collection functions to increase the performance of multiprocessor chips.

Trauer et al. (2017) developed a modular-based application for controlling, simulating and managing tuberculosis (TB). They agreed that existing mathematical models are not powerful enough for control of TB, hence modular-based programs being flexible could capture the heterogeneity and complexity of TB. The methodology entails development of application named AuTuMN that adopted the basic principles of software engineering for simulation. Table 4.11 provides a summary of some existing research that has exploited the concept of modularised programming.

Table 0.11: A summary of modularised programming in existing research

Author	Objective	Methodology and data	Results and finding
Lavy and Rami (2018)	Propensity of beginner programmers to use modular programming.	Experimentation with two study groups.	There is application of minimal efforts to achieve solutions to problem even at the expense of quality. Modularity was not used by most of the student from lack of skills, laxity and afraid to start-all over again.
Van Veen et al. (2018)	Reo modular programming is a better alternative to communication and synchronising for parallel programs.	Experimentation	Reo can be generalised to support parallel programming and the new approach outperforms the existing approach.
Subhan et al. (2021)	Spam detection using modular programming in websites.	Development and experimentation	The application detected and removed spams from websites.
Dennis (2003)	Development of multiprocessors chips inspired by modular programming principles.	Prototyping	High degree of security will be realised, high performance and concurrency computation.
Trauer et al. (2017)	Development of modular based application for control and simulation of TB.	Development and prototyping	Development of AuTuMN platform which will enable quick simulation and control, minimise errors and enable collaboration.

To enable modularity in the MfHPE framework, I have written *functions* that power each module rather than manually creating every step. This will allow: (i) transformations to be easily reproduced on datasets, especially when new or

updated data is ingested by the framework; and (ii) a progress build of a *library* of functions, reusable at multiple stages of development.

4.7 Data ingestion module

The data ingestion module is designed to call four different functions that initiate the ingestion of all the datasets described in Section 4.3 into the framework. These functions are detailed as follows:

(i) **def london_borough_transactions**

This function gets the original Price Paid Data from assets, extracts transactions set in only London boroughs between 01/01/2011 and 31/12/2020, and then remove records without postcode values. This results in the ingestion of 1,097,302 rows of data and 14 columns.

(ii) **def neighbourhood_data**

This function gets the rail station data, supermarket data and bus stop data from the neighbourhood *dictionary* in the asset. Since there are multiple sets of data in this category, a dictionary is created to hold them so they are ingested as a batch. This dictionary is stored in the params or parameters module. The neighbourhood data is also referred to as the Tier 2 data sets for the MfHPE framework.

(iii) **def ons_data**

This function gets the ONS data from the assets module and extracts three attributes – pcds, lat and long – and renames them as ‘postcode’, ‘latitude’ and ‘longitude’ respectively.

(iv) **def get_macroeconomic_data**

This function gets all macroeconomic data including Consumer Price Index, GDP, employment rate, unemployment rate and inflation rate.

4.8 Data pre-processing module

As a part of the data ingestion module, some pre-processing had been initiated primarily to reduce the volume of data, and as a consequence compute cost for this stage. For absolute clarity, the full scale of data pre-processing initiated in the data ingestion module (Section 4.7) and this module, as shown in Figure 4.10,

include: (i) extracting records for London-based transactions from the Price Paid Data; (ii) removing records with no postcode value from the Price Paid Data, as this is a must for upcoming data enriching steps; (iii) extracting only transactions that occurred between 01/01/2011 and 31/12/2020, being the focus time range for this research; (iv) removing duplicate records from the ONS postcode data; (v) blending the ONS data and price paid data to enrich or geo-code the price paid data with latitude and longitude values for each record; (vi) extraction of date entities, as this may be useful later; (vii) replacing categorical values for some *features* on the Price Paid Data with more appropriate values; (viii) scaling the price feature by presenting the values in millions; and (ix) binning the latitude and longitude features.

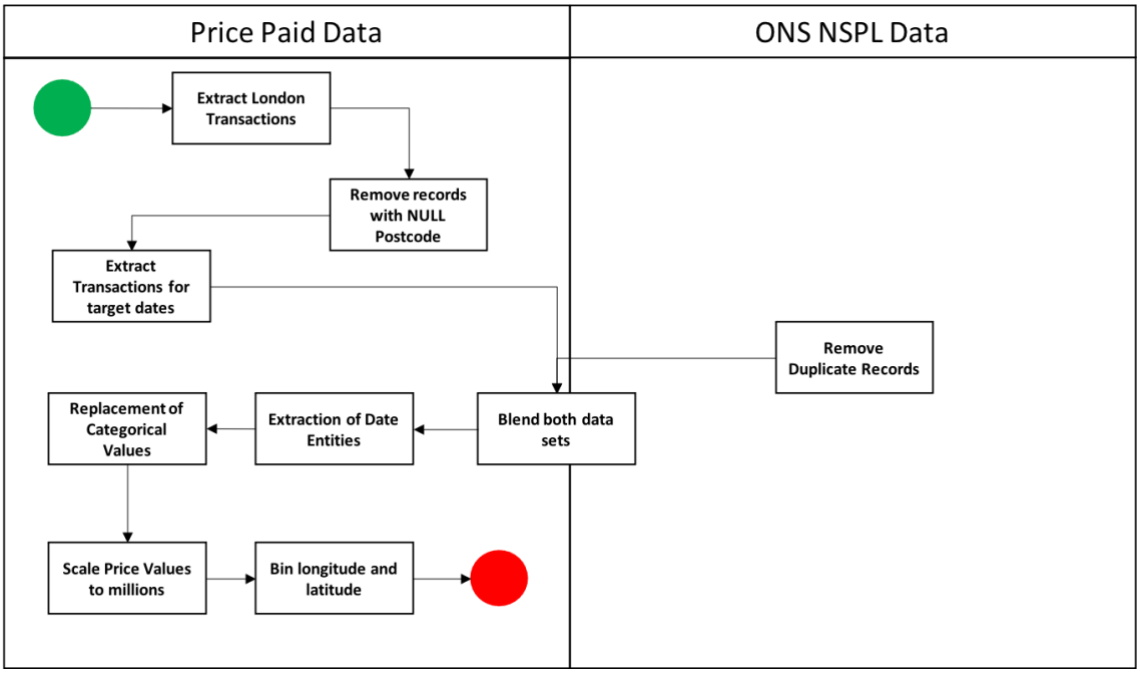


Figure 4.10: Geo-coding the Price Paid Data

4.8.1 Handling text and categorical variables

Machine learning is acknowledged as a strong approach with the ability to solve a range of problems based on the analysis of datasets by different algorithms. The growth of information technologies and services has made data storage, transmission and processing easier, thus leading to an influx in the sources of data. The large amount of data collected and stored makes it possible for

machine learning models to evaluate patterns and consequently describe the behaviour of different phenomena. The extensive data-mining resources have enabled the collection of large volumes of data, which are both non-numerical data (categorical data) and numerical data, and which are equally critical in the evaluation of a dataset and identification of patterns within the dataset. However, computers work on the numerical representation of data, meaning that datasets must be encoded in numerical values in order to compute arithmetic operations such as distance measures, dispersion and central tendency (Pargent, 2019). Categorical data is non-numerical, making it hard to satisfy some of these arithmetic operations. This, therefore, introduces the need for a critical pre-process in the creation of machine learning models, where categorical features are transformed into more operational numerical features which can be processed by a computer.

Understanding categorical data

Categorical data can be defined as variables that possess label values instead of numeric values. The possible number of label values are limited to a fixed set (Brownlee, 2020). Potdar et al. (2017) present that there are two broad groups of categorical data: nominal scale data and ordinal scale data. Nominal data means that the values in the data do not have any qualitative value – they are purely quantitative, for example gender (female/male) or marital status (single/married/separated/divorced). The variables in this category contain a finite set of discrete values which have no relationship within themselves. Ordinal data is a sub-group of categorical data which proposes that there is an order associated with the category. In essence, the variables contain a finite set of values that are presented in a ranked structure. Features like economic status, which possess categories such as high, medium and low, can be grouped as ordinal data.

Categorical data encoding

It is therefore necessary to encode categorical data to make it more amenable to the different mathematical calculations and operations necessary in machine learning. Encoding is a critical pre-process which attempts to transform the categorical features in a dataset and make an attempt to extend the numerical notions (such as centrality, dispersion and distance) of data which is otherwise non-numerical (memory efficient). The pre-processing approach entails creating binary values that indicate either the presence or absence of different numerical variables. In Table 4.12 below, some of the applicable approaches of categorical data encoding are presented along with a short sample of code that can be run in Python.

Table 0.12: Data encoding options

Approach	Description
Ordinal (integer) encoding	In this approach, every unique category within the dataset is given an integer value provided that the number of categories in the dataset are known. The integer values herein have an ordered relationship with one another and models can decipher the nature of the relationship.
One hot encoding	One hot encoding compares every level in the categorical variable with a fixed reference level by transforming single variables (with *n* observations and *d* distinct values to *d* binary values with *n* observations. The observations can therefore indicate either the absence (0) or the presence (1) of binary values. One hot encoding creates vectors that are in line with orthogonal and equidistant nominal categories (Cerda and Varoquaux, 2020).
Dummy variable encoding	A major problem that is associated with one hot encoding is that it creates one binary variable in each category, which may lead to redundancy. Additionally, regression models which have bias terms such as linear regression necessitate inverting the input data which is prevented by one-hot encoding. Dummy variable encoding is perceived to be less redundant since it allows many categories to be encoded and different columns to be selected using a prefix. It is also more flexible since it allows proper naming which makes analysis easier.
Feature hashing	This approach is proposed as an alternative to one hot encoding, more so when dealing with large scale datasets. It is a simple process that gives the user the ability to pick the output dimensionality by hashing the input value and subsequently dividing it with the output dimensionality. (McGinnis, 2016). In application, the hash function is a typical integer number which indexes a feature vector (Seger, 2018).
Target encoding	This approach is applied to reduce the number of levels in hierarchical clustering. The main concept of this approach is using a training set to make a prediction of the target in each level and using the predicted value as the numerical feature for the level. It is a quick and simple approach that limits the dimensionality of a dataset (Svideloc, 2020).

Handling of categorical variables in the MfHPE framework was done using ‘one hot encoding’.

4.9 Exploratory data analysis

Exploratory data analysis (EDA) is a pivotal process of any research analysis. It is a model of looking at data which does not follow the formal statistical inferences and modelling, therefore making it an imperative part of data analysis. In its natural form, exploratory data analysis does not adhere to strict rules of analysis – the analyst is free to explore any idea originating from the data, whereby some ideas may work out and be productive, while others may only end up being dead ends (Patil, 2018). To this end, EDA is in theory a creative process, whereby the analyst aims at asking qualitative questions that generate further quantitative questions. To get the best out of EDA, the analyst may need to raise the following questions: what types of variation occur within the variables in the data, and what types of co-variation occur with the data.

Some of the main reasons why exploratory data analysis is essential include: (i) maximising insight gained from a data set; (ii) detecting outliers; (iii) developing parsimonious models; (iv) detection of mistakes; (v) checking for the presence of assumptions; (vi) determining optimal factor settings; (vii) establishing whether there are relationships amongst the exploratory variables; and (viii) further assessing the relationship between dependent and independent variables (Grolemund, 2021); (Patil, 2018).

EDA can be cross-classified in two ways. First, the EDA approach can either be graphical or non-graphical. Graphical EDA methods usually summarise data using a pictorial or diagrammatic way, while non-graphical methods involve a summarised calculation of statistics. The graphical techniques used in EDA are more often than not simple approaches which include: plotting simple statistics like mean plots, box plots and standard deviation; plotting raw data such as lag plots, Youden plots and probability plots; and using multiple plots to maximise pattern recognition (Grolemund, 2021). Secondly, the approach can either be variate or bivariate. Variate EDA looks at one data column (variable) at any given time, while multivariate (or bivariate) data explores two or more variables at a given time.

4.9.1 EDA techniques

Based on the definition of exploratory data analysis given above, there are various techniques which apply EDA in data analysis. Komorowski et al. (2016) present the following techniques by grouping them in accordance to graphical and non-graphical EDA as well as variate or non-variate EDA.

Variate non-graphical EDA

Tabulation – Simple table containing the count of data and its frequency for each category.

Quantitative data characteristics – Expressions of the characteristics of a data sample using limited parameters. These characteristics may express: the central tendency of the data (mode, median, mean); its spread (variance, interquartile range, standard deviation); and its distribution (skewness, kurtosis).

Multivariate non-graphical EDA

Cross tabulation – This is an extension of tabulation which works with quantitative and categorical data with few variables.

Covariance and correlation – The measure of the relationship between random variables and an expression of how the two variable change together.

Variate graphical EDA

Histograms – Expresses distribution, central tendency, modality and outliers.

Stem plots – Substitutions for histograms which show distribution shape and all data values.

Other techniques include boxplots and 2D line plots.

Multivariate graphical EDA

Side-by-side boxplots – Presents several boxplots together for easier comparison.

Scatterplots – Built using two continuous, discrete or ordinal variables.

Curve fitting – Quantifies relationship between change of variables over time.

There are more complicated EDA techniques which handle more complex relationships. Such approaches include heat maps, 3D surface plots and support vector machines.

The EDA follows closely after the range of datasets considered for the design of the MfHPE framework have been ingested. Further to the tiers defined in Figure 4.4, the EDA will seek to understand and present as many insights as possible from each dataset.

The Tier 1 data is the baseline data for the framework, being the transaction data for houses sold in England and Wales from 1st January 1995 to date. Table 4.1 gives a snapshot of the profile of the data, while this sub-section presents further insights.

The distribution of the sum of daily transactions between 2011 and 2020 is as shown in Figure 4.11.

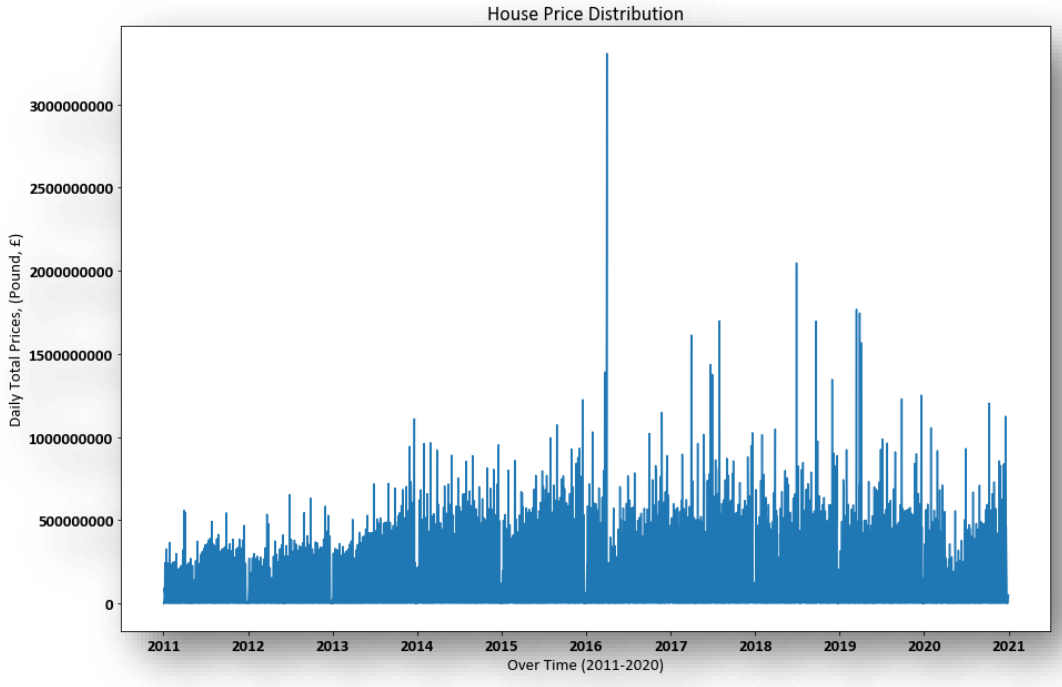


Figure 4.11: Distribution of transactions, daily

These are 1,097,302 transactions over the period, with a significant spike on 31st March 2016 worth a total of three billion, three hundred and six million, four hundred and ninety-two thousand, eight hundred and sixty-seven pounds (£3,306,492,867). Could the Five Point Plan for housing passed by the UK

parliament on 16th March 2016 have been the trigger? This plan focused on low-cost home ownership for first-time home buyers, and included commitments such as (Select Committee on Economic Affairs, 2016):

(i) Extension of Right to Buy to Housing Association tenants;
(ii) Delivery of 400,000 affordable housing starts by 2020-21;
(iii) Acceleration of housing supply by getting more homes built;
(iv) Availability of Help to Buy: such schemes include Equity Loan scheme to 2021, London Help to Buy, and offering a 40% equity loan;
(v) Increase Stamp Duty Land Tax (SDLT) on purchases of additional residential properties by 3% in each tier.

For the research time frame, being 01/01/2011 to 31/12/2020, additional insight from the data showed there was a total of 240 days within the time frame in view when there were no transactions. Figure 4.12 shows the distribution of daily averages with an upward trend observed until 2019 and then a drop in 2020, which can be explained as the impact of Covid-19.

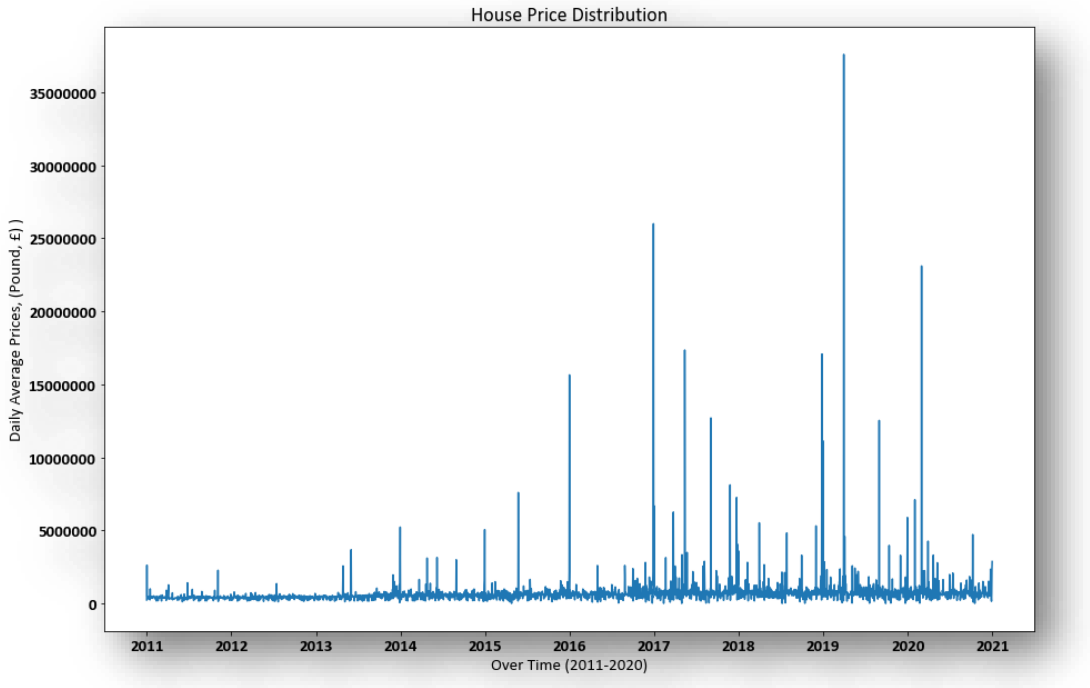


Figure 4.12: Distribution of transactions, daily averages

Beyond the spike in March of 2016, and with a view of the distribution of transactions weekly patterns as shown in Figure 4.13, the line chart shows some trend and seasonality in the weekly distribution that is worth further investigation,

beyond the scope of this thesis. Furthermore, the plot of the monthly distribution of transactions, in Figure 4.14, then shows a significant dip in 2020 which aligns with the reduced volume of transactions as a result of the impact of Covid-19. The spike in 2016 and the dip in 2020 are two events that would have to be given some extra consideration during the model design. March 2016 is also observed to have the highest total value of transactions in the ten-year period, with total transactions worth thirteen billion, six hundred and forty-six million, three hundred and eighty-three thousand four hundred and eight pounds (£13,646,383,408), while February 2011 has the lowest transaction value at two billion, three hundred and ninety-two million, seven hundred and sixty-six thousand, eight hundred and eighty-six pounds (£2,392,766,886).

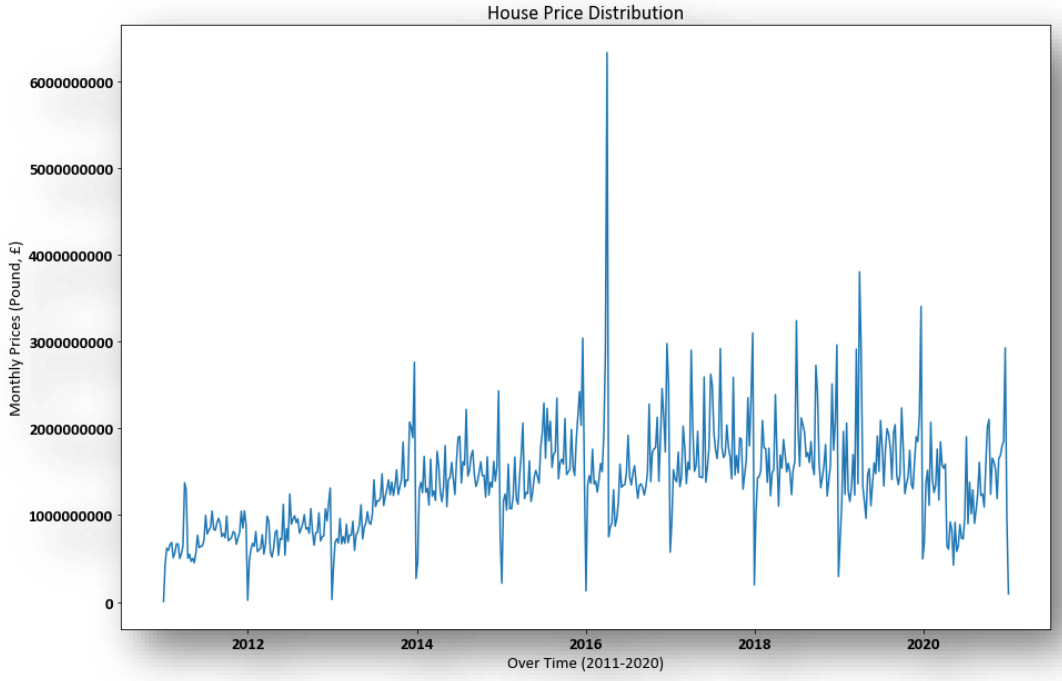


Figure 4.13: Distribution of transactions, weekly

The possible effect of Covid-19, as highlighted in Figure 4.14, is further amplified in Figure 4.15. However, it is worth noting that the spike or anomaly of 2016 and the dip of 2020 may have a significant impact on the house price estimation modelling, being unusual circumstances. Therefore, modelling will be implemented with outliers considered. Although the trend in the volume of transactions was already downward since 2018, Covid-19 lockdowns with the

unexpected reduction in the velocity and volume of economic activities created an even steeper downward trend, as seen in Figure 4.15.

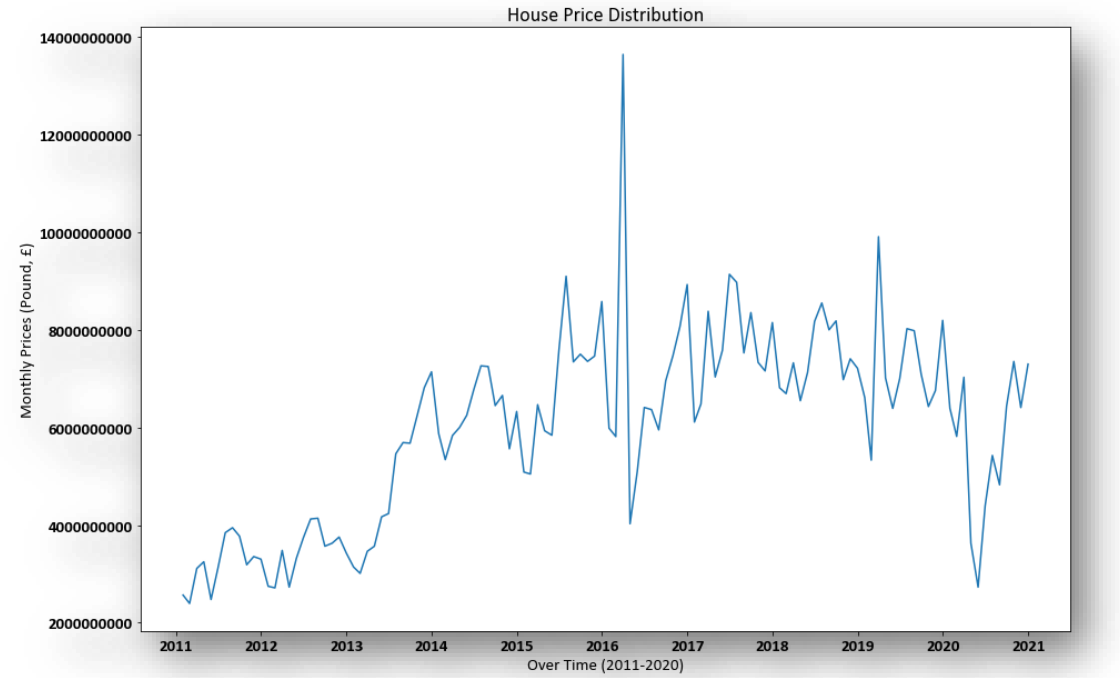


Figure 4.14: Distribution of transactions, monthly

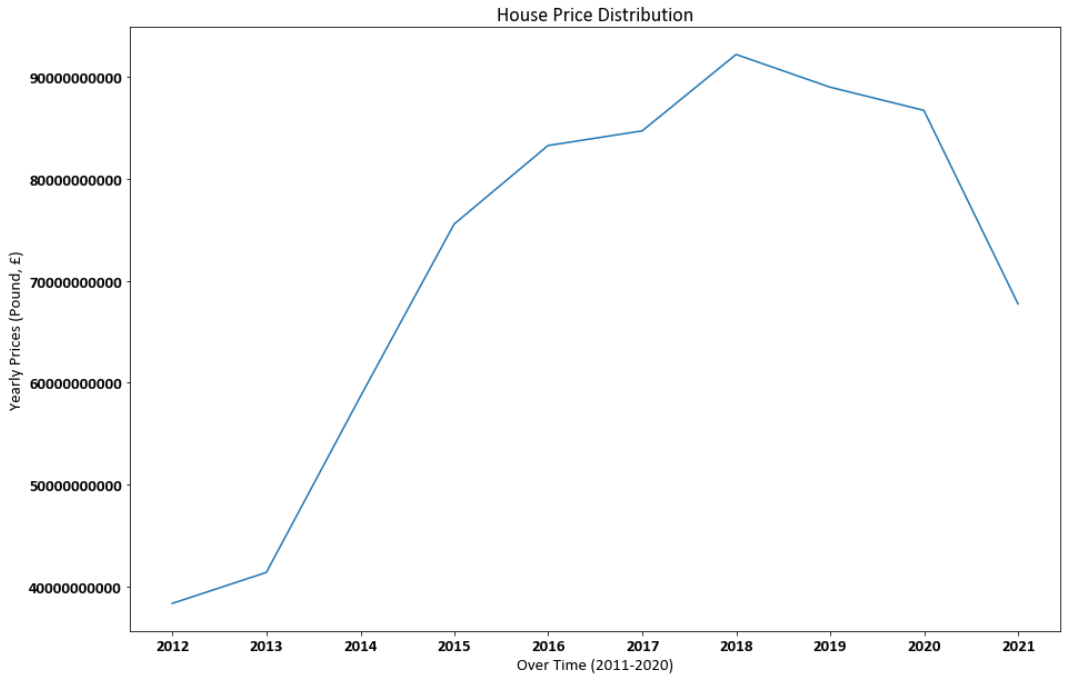


Figure 4.15: Distribution of transactions, yearly

Simple data analysis gives a basis for measuring the central tendencies of data, the spread of data and outliers. This is a critical basic step that entails measuring

the location and variability of the data set. Beyond the measures of dispersion and central tendency, further analysis of the data is critical to explain more in-depth characterisation insight, more so in terms of the characteristics of the distribution. Skewness and kurtosis are examples of such techniques that offer in-depth insight into datasets. These techniques ensure that the normality in a data set is analysed, since they are measures of shape. They present critical information about the distribution of data in instances where graphical methods of data analysis do not present effective results or cannot be used (Komorowski et al., 2016). The understanding of the shape of data is crucial because it helps identify where most of the information is lying, and consequently forms the basis of the analysis of outliers in the dataset. This section also presents skewness and kurtosis; their application and why they are imperative in statistical data analysis. The original distribution of the target variable in the 'Price Paid Data' is shown in Figure 4.16. This distribution is significantly skewed to the left and the kurtosis is high, i.e. it shows a significant number of outliers.

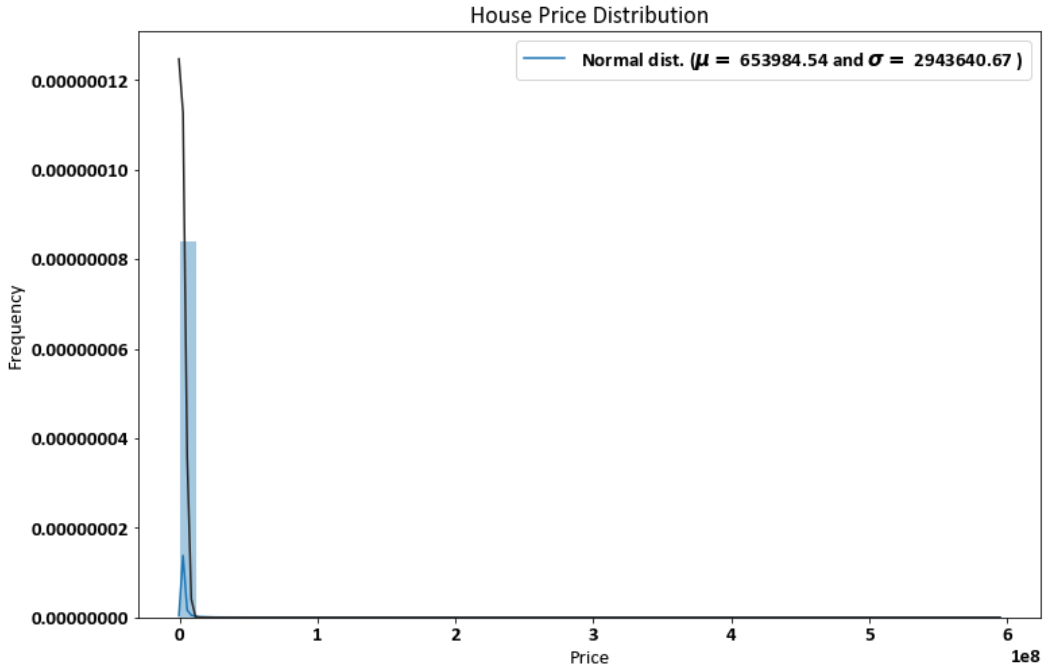


Figure 4.16: Original price distribution of Price Paid Data

4.9.2 Skewness

The term 'skewness' means a departure from symmetry or the lack of distribution. To this end, when the distribution in a dataset is not symmetrical, it is referred to

as a ‘skewed distribution’. Inversely, when the distribution is symmetrical, it is referred to as a ‘normal distribution’. The measurements gained from skewness indicate the differences in the manner in which the observations are distributed in comparison with a symmetrical distribution. Therefore, skewness is described as the degree of distortion from the curve of a normal distribution, or from the symmetrical bell curve. Skewness differentiates the extreme values in one tail versus the other tail in the curve.

There are two types of skewness: positive skewness and negative skewness. Positive skewness essentially means that the tail is longer and/or fatter to the right end of the distribution, with the median and mean higher than the mode of the data. A positively skewed distribution means that the dataset contains some values which are much larger than most values in the observation. The mean is therefore pulled towards the high-valued item, hence a bend towards the right. On the other hand, negative skewness is a phenomenon where the left tail side in the distribution is fatter and/or longer compared to the tail to the right of the curve. In a negatively skewed dataset, the median and mean of the data are less than the mode. The dataset contains some values which are much less than the majority of the observations, hence the curve is pulled towards the low-valued item which is on the left.

There are different approaches used to define skewness. This thesis presents three approaches that are applied in defining skewness: (i) Pearson’s first coefficient of skewness; (ii) Pearson’s second coefficient of skewness; and (iii) Galton’s skewness (which is also called Bowley’s skewness).

Pearson’s first coefficient

$Pearsons\ first\ coefficient\ =\ (mean - mode) \div\ standard\ deviation$

(Gawali, 2021).

This approach is helpful in instances where the data presents a high mode. However, when the data has various modes or low modes, then it is preferable to apply Pearson’s second coefficient which does not rely on mode.

Pearson’s second coefficient of skewness

$Pearson's\ second\ coefficient\ =\ 3\ (mean - median)\ \div\ standard\ deviation$

(Gawali, 2021)

Galton skewness

$Galton\ skewness\ =\ (Q1 + Q3\ -2Q2) \div\ (Q3\ -Q1)$

(Nist.gov, 2021)

When either of these approaches is applied, the results indicate the following:

- If the skewness is between 0.5 and −0.5, the data is close to symmetrical.
- When the skewness is between −1 and −0.5 (for negatively skewed data) or between 0.5 and 1 (for positively skewed data) the data is only slightly skewed.
- When the skewness is more than 1 (for positively skewed) or lower than −1 (for negatively skewed) the data is greatly skewed.

(Dugar, 2018)

4.9.3 Kurtosis

Kurtosis can be defined as the measure of the extreme values in a curve in relation to the other extreme value. This measure analyses the number of outliers that are available in a distribution.

Kurtosis can also be defined as a measure of the degree of 'peakedness' in a frequency curve, because it can also be used to show how tall and/or sharp the central peak of a curve is in relation to a standard bell curve of distribution.

This is the formula used in the calculation of kurtosis (McNeese, 2008):

$$Kurtosis = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\frac{(X_i-\bar{X})^4}{s^4}\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Types of kurtosis

(i) High kurtosis – This is an indicator that the data has heavy outliers or heavy tails.

(ii) Low kurtosis – This indicates that the data has light tails and lacks outliers. In instances where the Kurtosis is too low or too high, it is critical to check the viability of the dataset (Dugar, 2018).

(iii) Mesokurtic – This is when the kurtosis is equal to zero, making the curve appear normal in its shape. The distribution static here is similar to that of normal distribution, meaning the extreme values in the curve

will be identical to the extreme values of a normal distribution. The standard normal distribution has a kurtosis of 3, indicating normality (Dugar, 2018).

(iv) Platykurtic – This curve is flatter and spread out and wide. The kurtosis is less than 3 and the frequencies all over the curve are closer to being equal. The peak is, therefore, lower and much broader compared to the mesokurtic curve due to the lack of outliers in the data (Dugar, 2018).

(v) Leptokurtic – This is where the kurtosis is more than 3. There are higher frequencies in the small part of the curve making the curve more peaked. The shape of the curve is high and thin because the outliers stretch the horizontal part of the curve, making the bulk of the data appear in a narrow vertical range (Dugar, 2018).

4.9.4 Q-Q plot

Also known as quantile-quantile plots, Q-Q plots are graphical techniques that are used to assess whether datasets come from populations with normal distributions. Quantiles can be described as the fraction or percentage that the values in that particular quantile fall below. To put this into perspective, the 0.4 (or 40%) quantile is the point where 40% of values in the dataset fall below and 60% of the values fall above (Ford, 2015). Quantiles are also known as percentiles.

A normal Q-Q plot is an important diagnostic tool that is used to assess the assumption of normality in a dataset. Therefore, Q-Q plots are plots of two quantiles that are placed against each other to purposely establish whether the two sets of data originate from the same distribution (Loy et al., 2016). A Q-Q plot takes the values from the dataset and arranges them in ascending order. If the values originate from a normal distribution, they should form a line that is somewhat straight.

Plotting Q-Q plots

A quantile-quantile plot is created from a sample by plotting the assumed

(theoretical) quantiles against sample quantiles. When the theoretical quantiles are consistent with the sample quantile, the points of the data set in the Q-Q plot fall in a line of identity. The theoretical quantiles are plotted along the x-axis while the sample quantiles are plotted along the y-axis (Loy et al., 2016), as shown in Figure 4.17. A utopian consistent Q-Q plot would be a line ascending at a slope of 45 degrees. Therefore, Q-Q plots contain a 45-degree line which is used to assess the normality of the plot. In any sample distribution (whether a normal distribution, a log-normal distribution or an exponential distribution) there will still be some form of association to this line of identity. The appearance of distribution on Q-Q plots can reveal: (i) normally distributed data; (ii) right-skewed data; (iii) left-skewed data; (iv) under-dispersed data; and (v) over-dispersed data.

Normally distributed data will appear like a 45-degree line. *Right-skewed data* is a plotted positive skew, which will appear to be curved towards the y-axis. *Left-skewed data* is a plotted negative skew, which will appear to be curved towards the x-axis. The negative exponential distribution is the opposite of right-skewed data. *Under-dispersed data* has a negative excess kurtosis. A dataset with a negative excess kurtosis will ideally have a reduced number of outliers, meaning the distribution will have thinner tails. On the plot, the data will appear to be S-shaped. *Over-dispersed data* has a positive excess kurtosis, meaning that it would have a higher number of outliers compared to normally distributed data and fatter curve tails. When plotted on a Q-Q plot, over-dispersed data will appear to have a flipped S shape (Yearsley, 2015). Overall, Q-Q plots are virtual subjective visual checks which give an insight into the normality of a data set, whether normality in the dataset has been violated, and which values in the dataset contribute to the violation.

Since the target variable in the Price Paid Data is left-skewed with high kurtosis and over-dispersed, it will be essential to have the data normalised.

In statistical databases, normalisation is described as the process that ensures data is structured in a more robust and logical format. Normalisation does not change the values associated with attributes of entities, but rather develops structures based on the logical linkages and connections in the data. To this end, normalisation can be described as a technique that is used to produce sets of relations that have desirable properties given the data requirements (Eessaar,

2016). Normalisation is applicable in the following scenarios, for example: when the distribution of data is not Gaussian; when data has varying scales; and when the algorithm being used does not make assumptions based on data distribution, such as K-nearest neighbours and artificial neural networks. (For data that is not Gaussian, standardisation substitutes normalisation as will be seen in a subsequent section.)

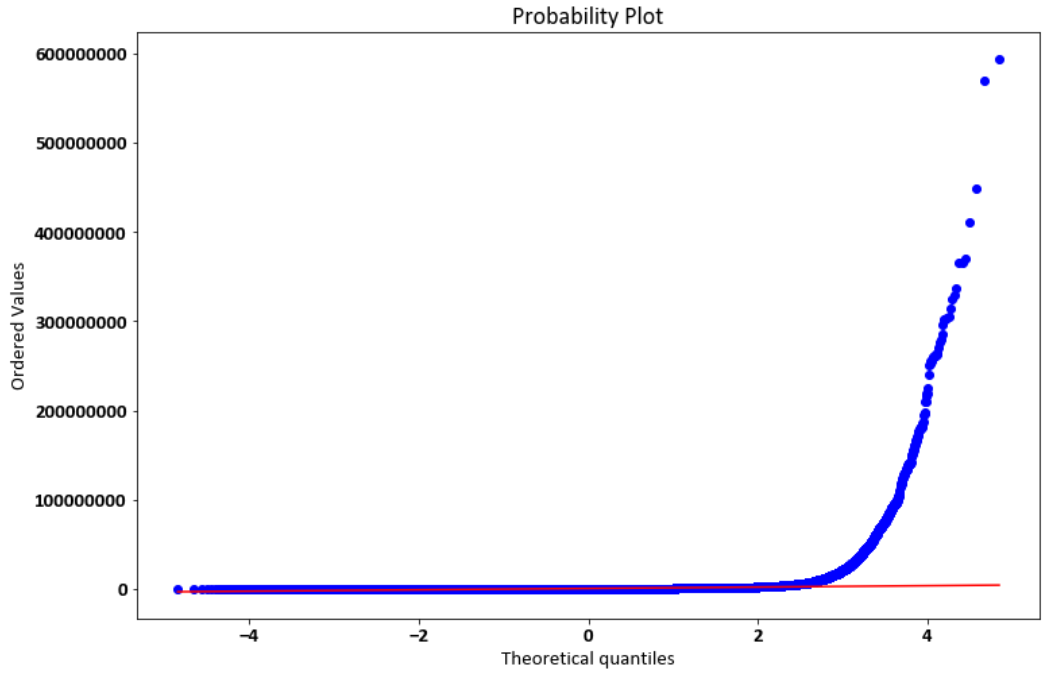


Figure 4.17: Original price Q-Q plot

The formula for calculating the normalised score using min-max scaling is:
$X\ new\ = (X - Xmin)/(Xmax\ -\ Xmin)$
Where *Xmin* and *Xmax* are minimum and maximum values of the feature respectively
The normalisation of the price distribution using **log transformation** and corresponding Q-Q plot are as shown in Figure 4.18 and 4.19 below. The Q-Q plot in Figure 4.19 now shows a better distribution of the data post-normalisation.

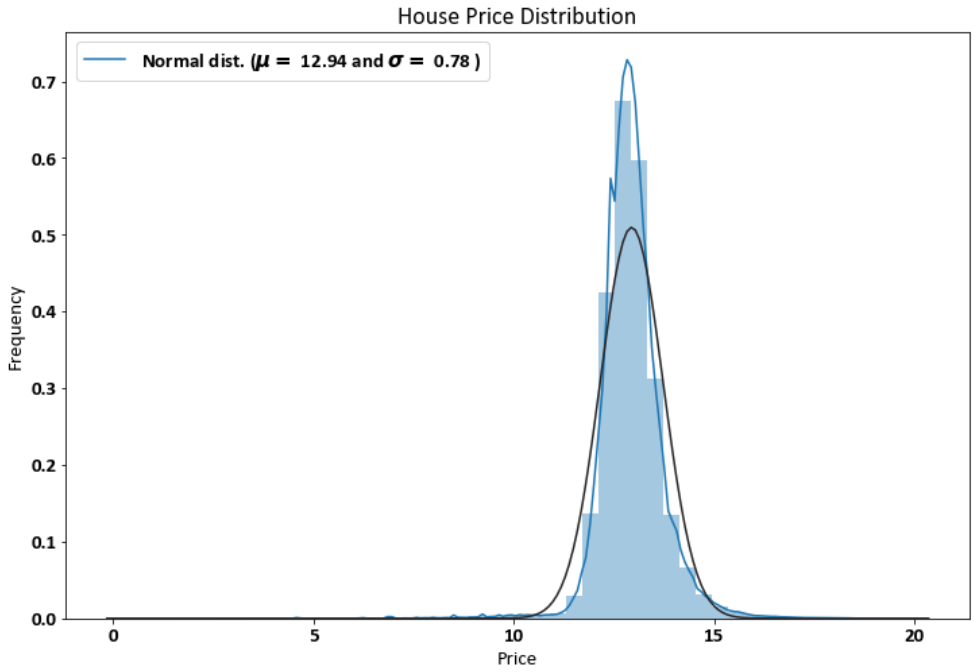


Figure 4.18: Log transformed price distribution of Price Paid Data

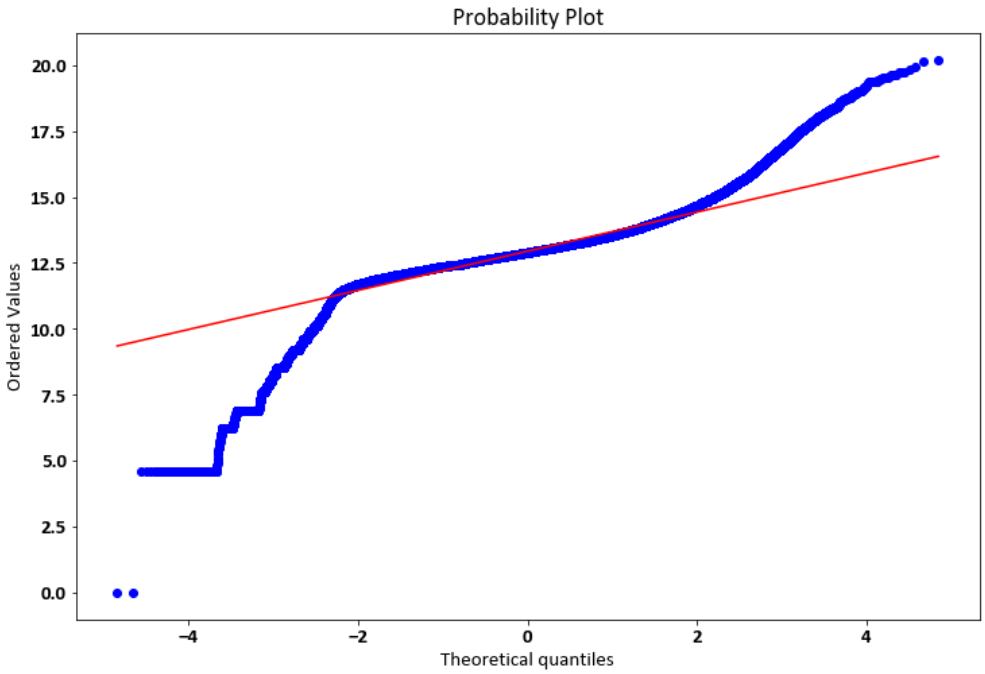


Figure 4.19: QQ plot for log transformed price distribution of Price Paid Data

However, with further exploration, a **Box-Cox transformation** is applied just to assess its impact on normalising the data. The distribution curve and Q-Q plot are as shown in Figures 4.20 and 4.21 respectively, with Lambda for the Box-Cox transformation set to 0.15.

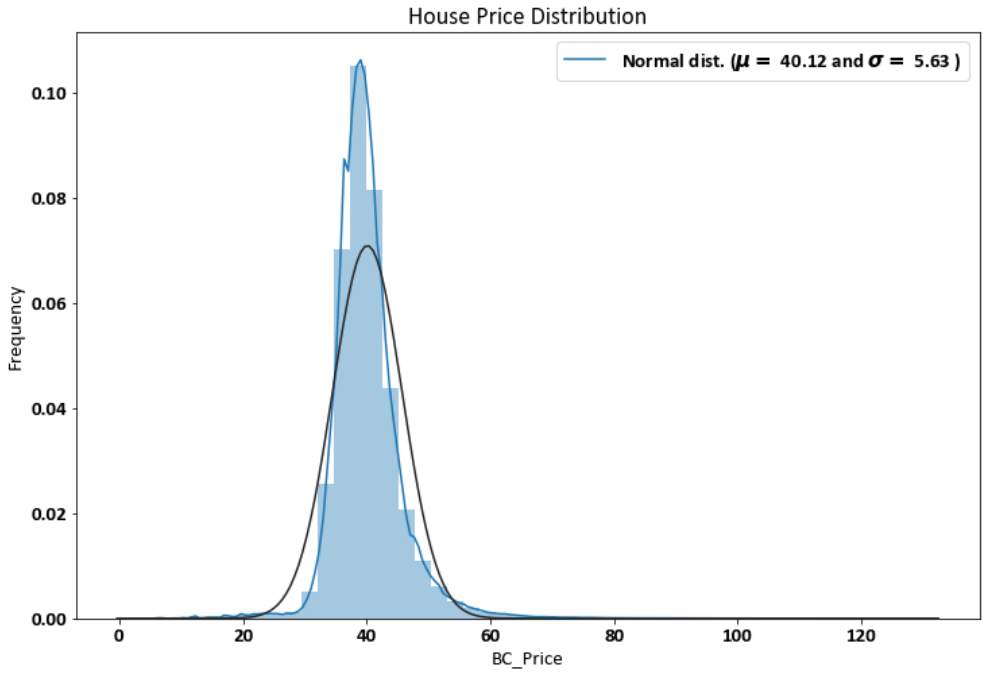


Figure 4.20:Box-Cox transformed price distribution of Price Paid Data

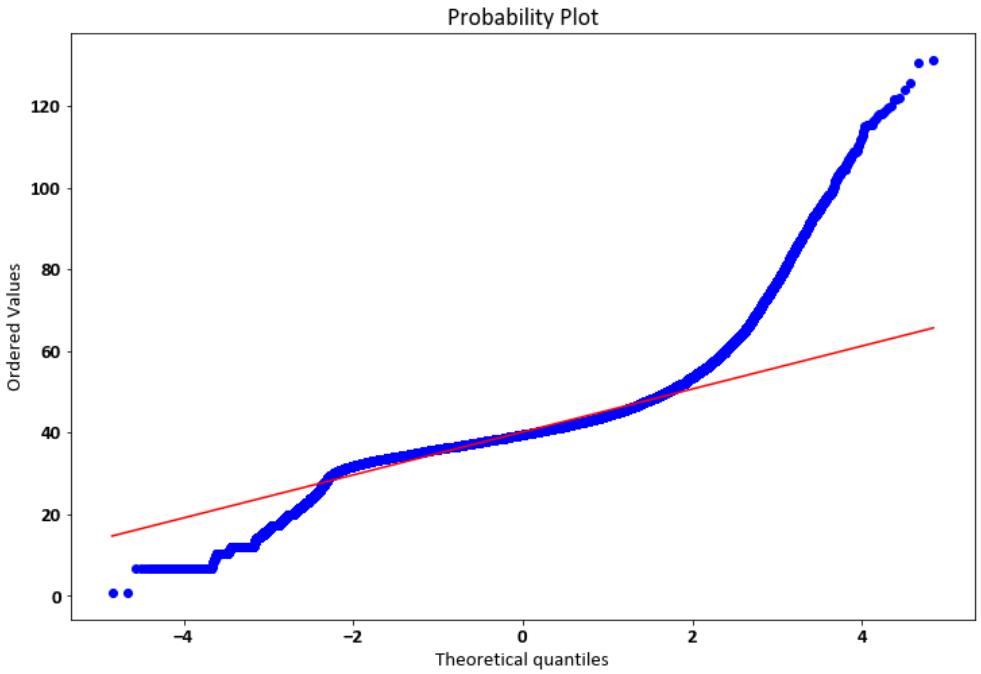


Figure 4.21: QQ plot for Box-Cox transformed price distribution of Price Paid Data

Table 4.13 shows a comparison between the original and transformed statistical metrics for the target variable, price.

Table 0.13: Original vs transformed statistical metrics for target variable

Distribution	**Mean / mu**	**Standard Deviation / Sigma**
Original distribution	653984.54	2943640.67
Log transformation	12.94	0.78
Box-Cox transformation	40.12	5.63

Outlier effect

To assess the possible impact of outliers evidently present in the data, an outlier detector was created based on the concept of interquartile range. The lower, median and upper quartile stats before the removal of outliers were calculated using:

```
Iqr_dict = {'q1':tier_three_df['Price'].quantile(.25), 'median':
tier_three_df['Price'].median(), 'q3': tier_three_df['Price'].quantile(.75)}
statBefore = pd.DataFrame().append(iqr_dict, ignore_index=True)
```

Result: '**q1'**: 275000.0, '**median'**: 400000.0, '**q3'**: 600000.0

The interquartile range is then calculated using:

```
iq_range = statBefore['q3'][0] – statBefore['q1'][0]
```

Result: iq_range = 325000.0

Further to the calculation of the interquartile range, outliers in the data are then filtered out by the creation of a function as shown below:

```
Def outlier_(x):
    If x > (median + (1.5 * iq_range)) or x < (median – (1.5 * iq_range)):
        return True
    else:
        return False
```

This function is then applied to the data and a new variable, 'outlier', is created which helps to identify all records that are classified as outliers. Table 4.14 shows there are 125,801 records classified as outliers.

Table 0.14: Full data vs data without outliers

	No. of records	No. of variables
Full data	1,093,302	73
Data without outliers	967,501	73

The stats for over 960,000 records classified as non-outliers are shown in Table 4.15. Figure 4.22 shows that the removal of outliers in the data does not seem to have made a significant difference, although it shows a better distribution, and

the Q-Q plot shows more records aligning with the line of best fit, as shown in Figure 4.23.

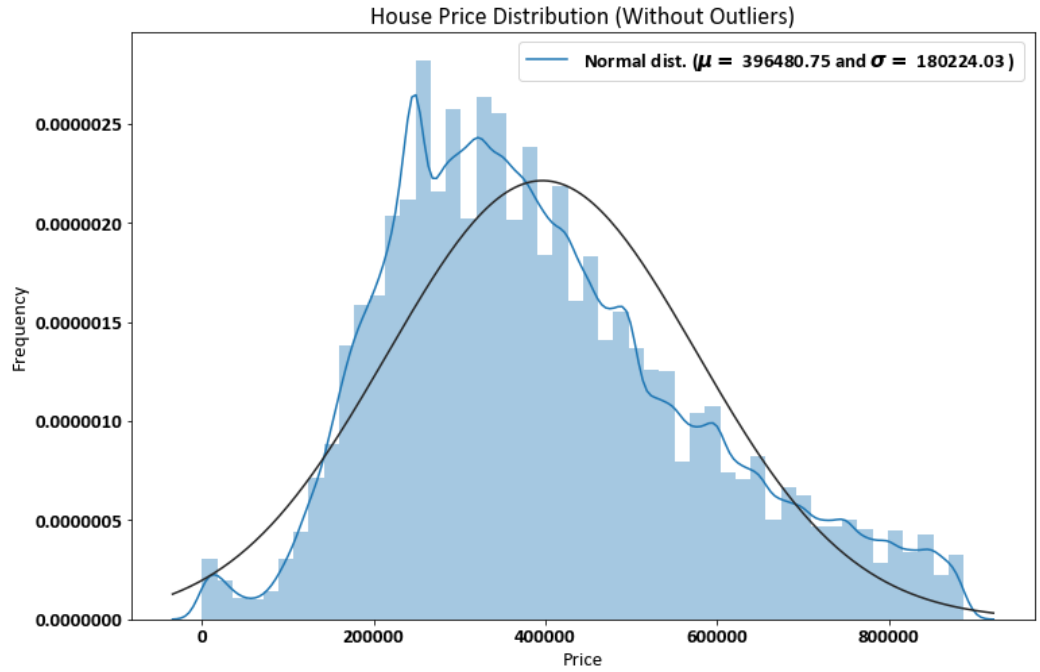


Figure 4.22: Price Paid Data distribution without outliers

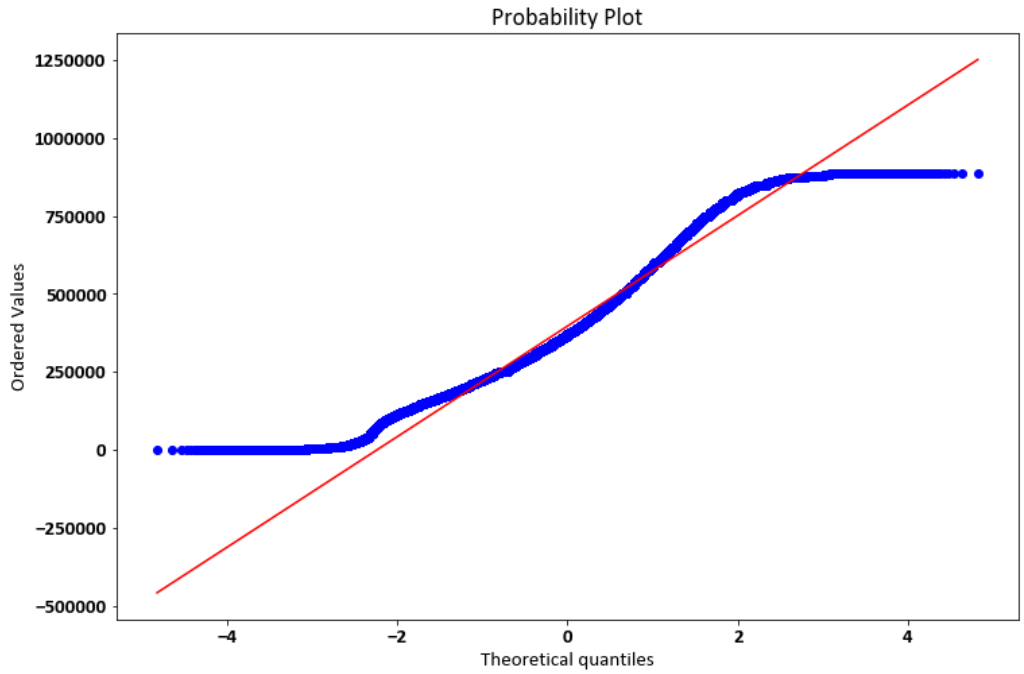


Figure 4.23: QQ plot for Price Paid Data distribution without outliers

Table 0.15: Stats for full data and data without outliers

	Full data with outliers	Full data without outliers
Mean	653984.54	396480.75
Median	400000.0	369000.0
1st Quantile	275000.0	260000.0
3rd Quantile	600000.0	500000.0
Standard Deviation	2943640.67	180224.03

Figure 4.24 shows an overlay of the full data with outliers and full data without outliers, and without any form of transformation reveals that there is no significant difference in the state of the research data with or without outliers.

Therefore, the design of the machine learning models later in this chapter will be completed with outliers because of domain knowledge.

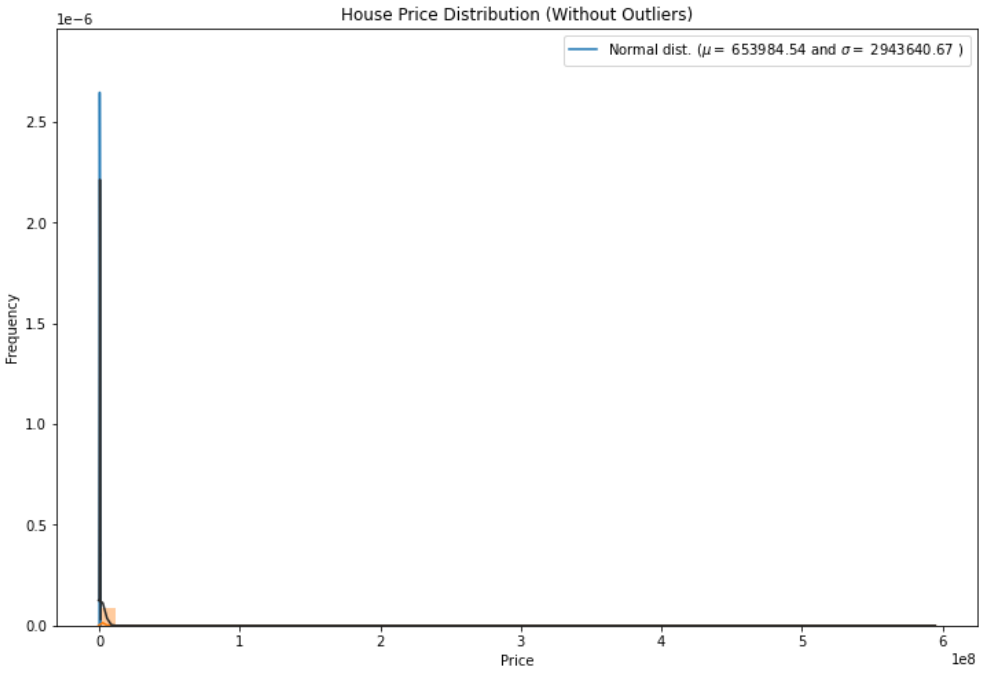


Figure 4.24: Overlay of the full data with outliers and full data without outliers

With a further review of the outliers the mean, min and max values do not necessarily depict anomalies, considering the research domain is London. This then substantiates the need to initiate ML modelling with and without the outliers and then explore a comparison.

The log transformation of the research data without outliers, as shown in Figures 4.25 and 4.26, also validates the fact the outliers do not have a significant impact on the usability of the data.

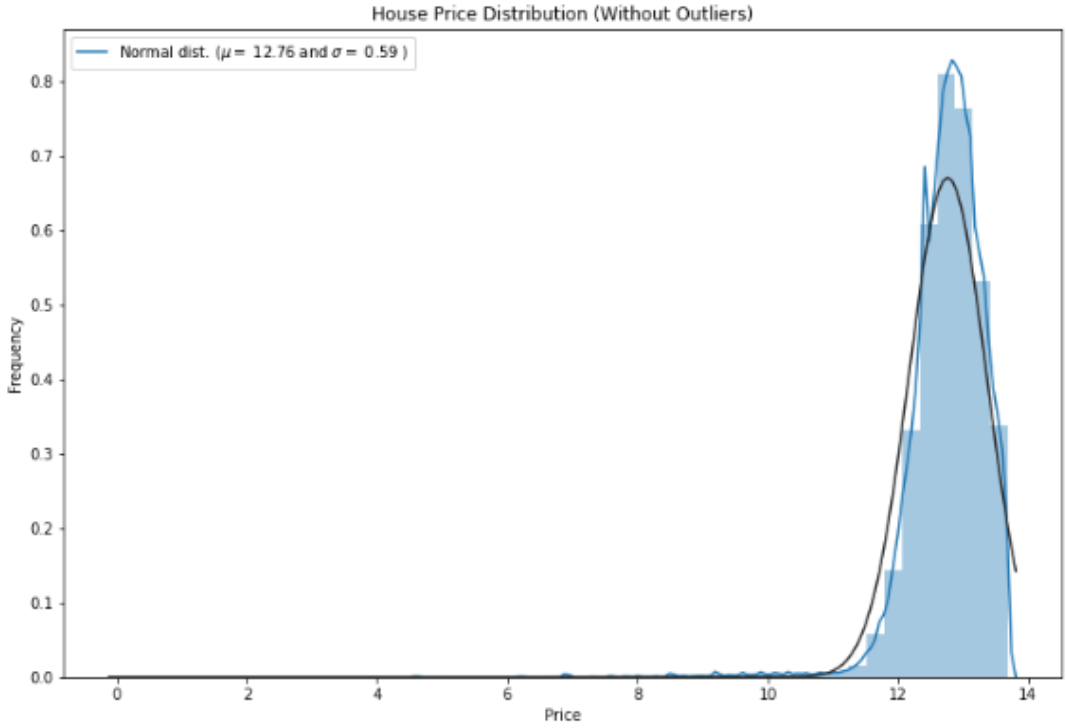


Figure 4.25: Log transformation of the research data without outliers

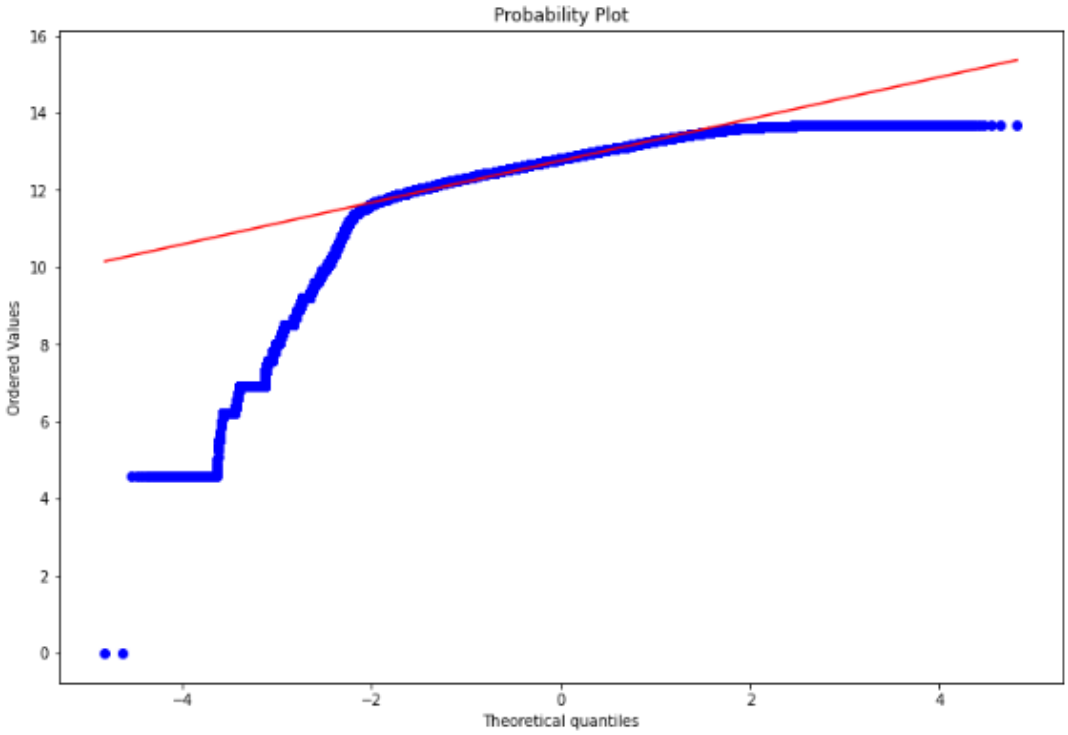


Figure 4.26: Q-Q plot for log transformation of the research data without outliers

Functional dependencies

A functional dependency can be described as a constraint that exists between two sets in a database. It therefore describes the relationship between attributes in a relation. To put this into context, if A and are attributes relative to X, B will be functionally dependant on A if every value of A in X is associated with one value of B in X (Ebrahim and Mahmoud, 2014).

A candidate key is an attribute that uniquely identifies a row in a relation. In this sense, a candidate key can be a combination of attributes that are non-redundant. Subsequently, every non-key field is functionally dependant on a candidate key. As a scaling technique, normalisation applies a method through which data points are rescaled and shifted to be within the range of 0 to 1 through min-max scaling. Several reasons make the performance of normalisation of data necessary; a majority of which aim at preventing the corruption of the database based on anomalies as expounded below:

- Insertion anomalies: This is the inability to include more data into a database because of the lack of some other data, thus leading to inconsistencies in the data because of some omission.
- Deletion anomalies: In this anomaly, there is an unintentional loss of data because some other data has been deleted. This, therefore, results in inconsistencies.
- Modification anomalies: This is an inconsistency generated by the addition, change or deletion of data from a database (Ebrahim and Mahmoud, 2014).

The existence of these anomalies leads to either duplication within the dataset and/or unnecessary dependencies within entities, thus necessitating normalisation. The normal form of a relational database is a criterion that is used to measure and determine the degree of a table's immunity against anomalies. In theory, the higher the normal form of a table is, the less vulnerable it is to anomalies and logical inconsistencies (Ebrahim and Mahmoud, 2014). To this end, a table within a dataset will meet its highest normal form and additionally meet the requirements for all the normal forms lower than its HNF. To put this into context, a table meeting the requirements for 3NF will have to meet the

requirements for 1NF and 2NF. Table 4.16 shows the different levels of normal and the corresponding process for implementation.

Table 0.16: The levels of normal forms

Step	Approach/Process	Description
First normal form; 1NF	Remove multivalued attributes	This is the initial normalisation process that converts unnormalised data in regards to the first normal form in the sense that repeating groups of data are extracted and replaced with values that have at most one value associated with them. The first normal form stipulates that; no rows and columns may be duplicated, no columns/rows intersections should contain multivalued fields and/or null values
Second normal form; 2NF	Remove partial dependencies	This is the second step in normalising a database. It builds on the first normal form. Its basis is full functional dependency dependent on X, but not on any subset of X. Every non-key attribute becomes fully functionally dependant on the primary candidate key, not as a part of the key and thus eliminating functional dependencies.
Third normal form; 3NF	Remove transitive dependencies	Transitive dependencies occur when a non-key attribute depends on a duplicate non-key attribute. Normalisation decomposes the original relations using algebra projections thus removing the transitive dependencies and placing them in new relations with copies of their determinant.
Boyce-Codd normal form; BC-NF	Removing anomalies from multiple candidate keys	Violation of BC-NF occurs when a specified relation has multiple composite candidate keys that overlap and share one or more attributes. Transformation to BC-NF is successful when violating functional dependencies are removed and placed in a new relation.
Fourth normal form; 4NF	Removing multivalued dependencies	The fourth normal form builds from the third normal form by removing a multi-valued dependency. Multi-valued dependency occurs when there are no fewer than 3 attributes in a relation (A, B, C), whereby for every value of A there is a defined set of values for B and a defined value for C, but the values for B are independent of those of C.
Fifth normal form; 5NF	Eliminating joint-dependency	The fifth normal form is an extension of the fourth normal with eliminated joint dependency. It is satisfying when all the tables within the dataset are broken into many tables in order to avoid redundancy.

Standardisation

Standardisation is a scaling method whereby values have a centre on the mean with a unit standard deviation. The formula for standardisation is:

$$Z = (X - \mu)/\sigma$$

Where:

μ = mean of the distribution

σ = standard deviation of the distribution

Z = standard score – number of standard deviations above and below the mean (Rizwan, 2020).

Further EDA shows the distribution of the *total number of house sale transactions* per year between 2011 and 2020, as seen in Figure 4.27. Figure 4.28 then shows a *duration split* for the total number of properties sold year-on-year. The prevailing trend is that more houses are sold on leasehold than freehold every year, with a slight reversal in 2020. The boxplots in Figures 4.29 and 4.30 show the distribution of the duration/tenure of house sale transactions with and without outliers respectively.

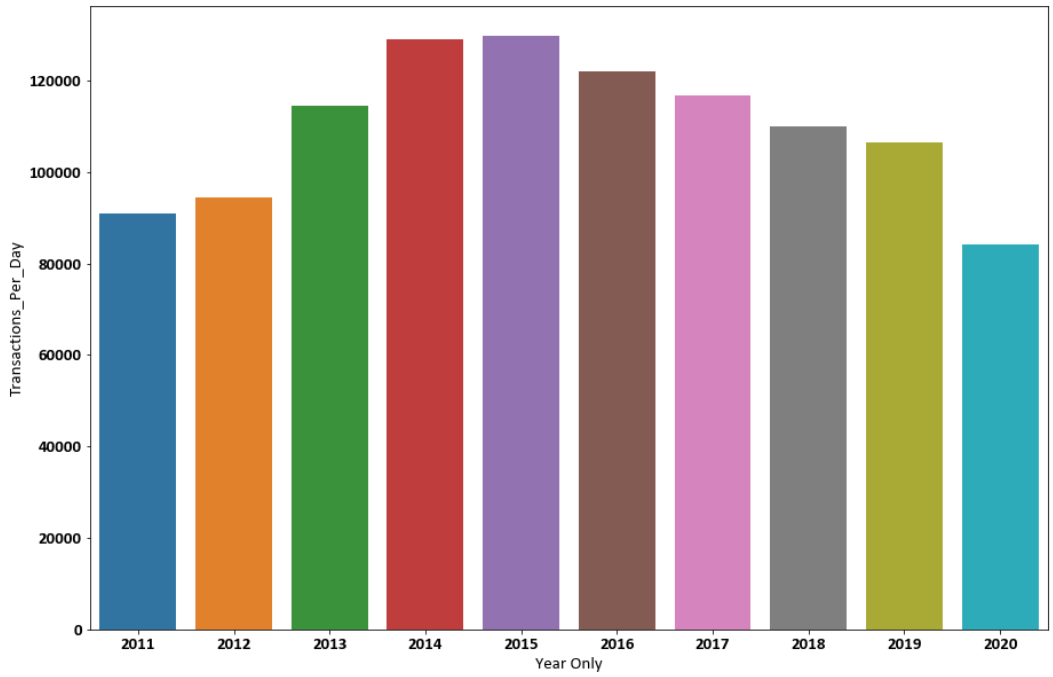


Figure 4.27: Distribution of the total number of house sale transactions per year between 2011 and 2020

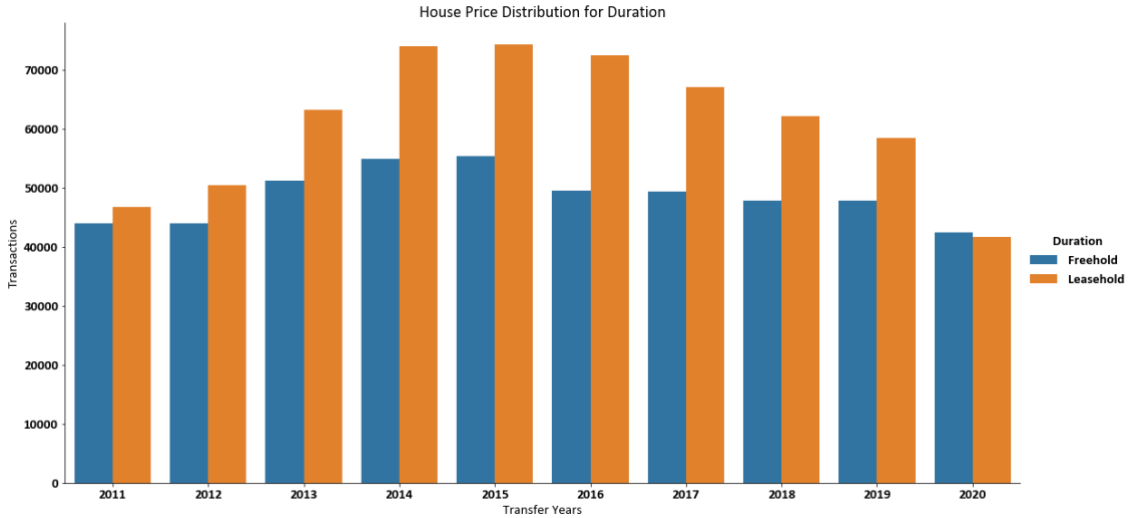


Figure 4.28: Distribution based on tenure

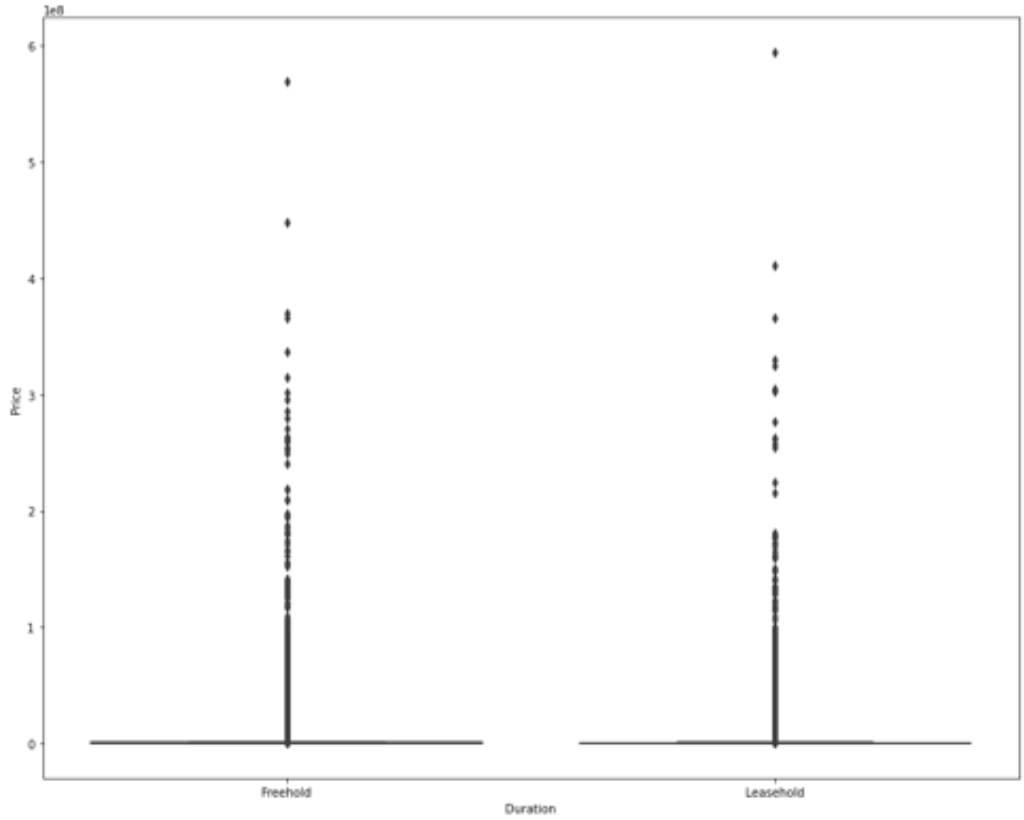


Figure 4.29: Boxplot for tenure with outliers

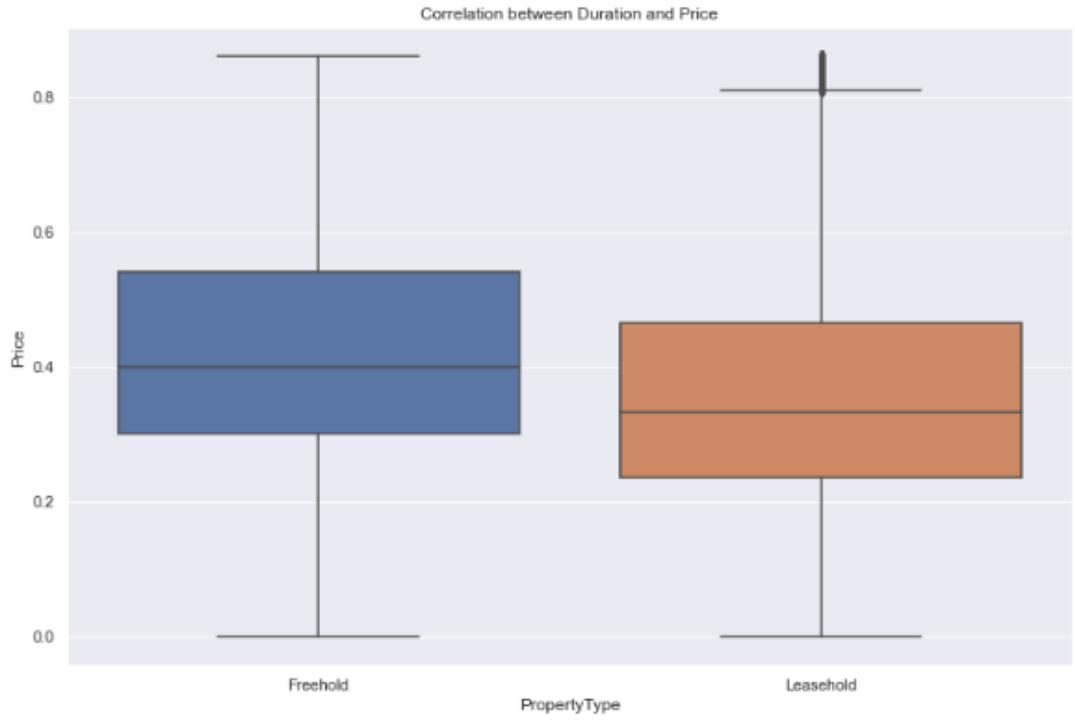


Figure 4.30: Boxplot for tenure without outliers

The research data contains different types of houses. These include: (i) D = detached; (ii) S = semi-detached; (iii) T = terraced; (iv) F = flats/maisonettes; (v) O = others. As shown in Figure 4.31, there are more flats or maisonettes and terraced houses sold in London every year than any other house type. While 2014 and 2015 showed the highest number transactions for terraced and flats/maisonettes, the number of transactions on detached or semi-detached houses are relatively similar year-on-year. The boxplots in Figures 4.32 and 4.33 show the distribution of the house sale transactions for different house types with and without outliers respectively.

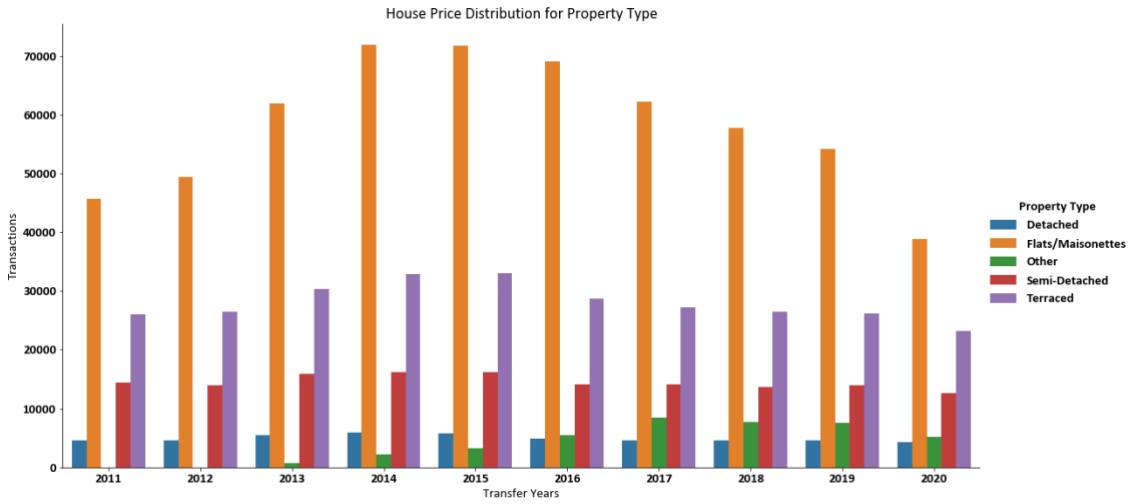


Figure 4.31: Distribution based on house type

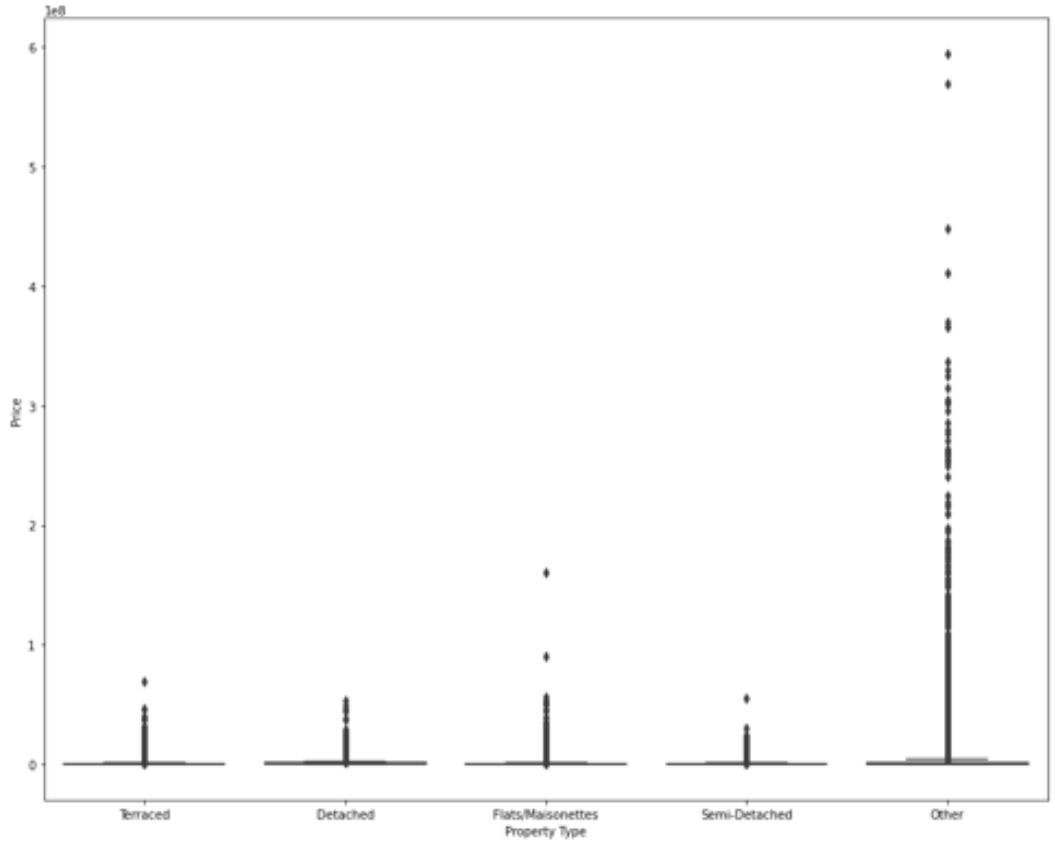


Figure 4.32: Boxplot for house type with outliers

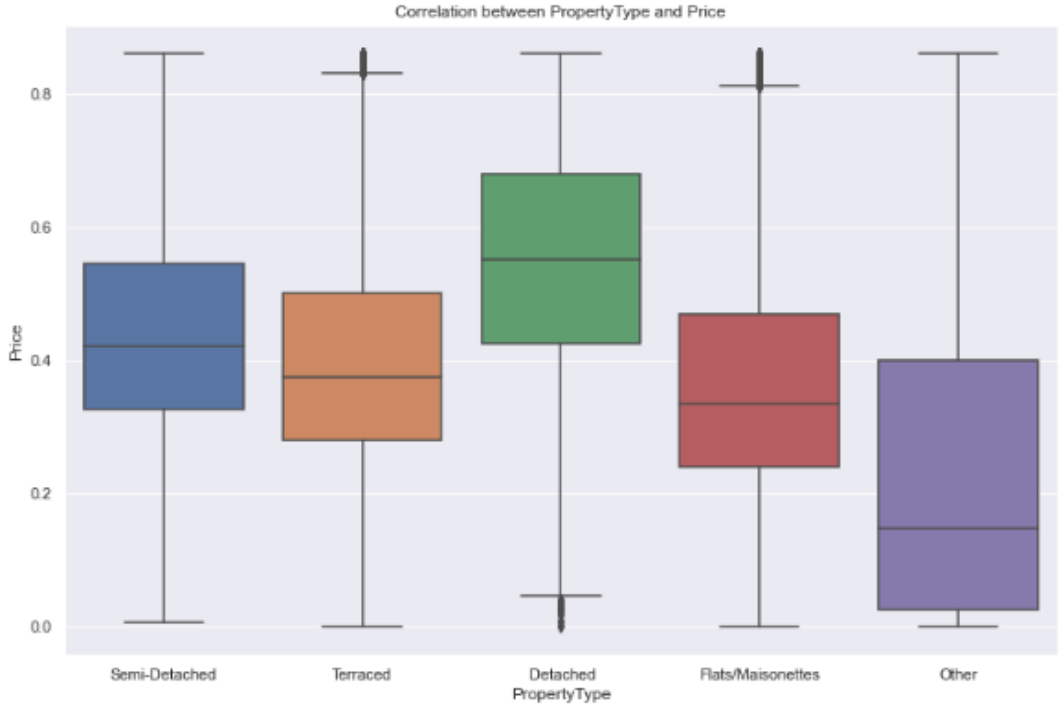


Figure 4.33: Boxplot for house type without outliers

Although there has been an upward trend in the number of house sale transactions for ‘new build’ houses between 2011 and 2016, there has been a decline in the trend since 2017, and with a significant dip in 2020, which is traceable to the impact of Covid-19 lockdowns. New build sales in 2020 were at their lowest since 2011.

Figures 4.34 and 4.35 show the distribution and boxplot with outliers respectively.

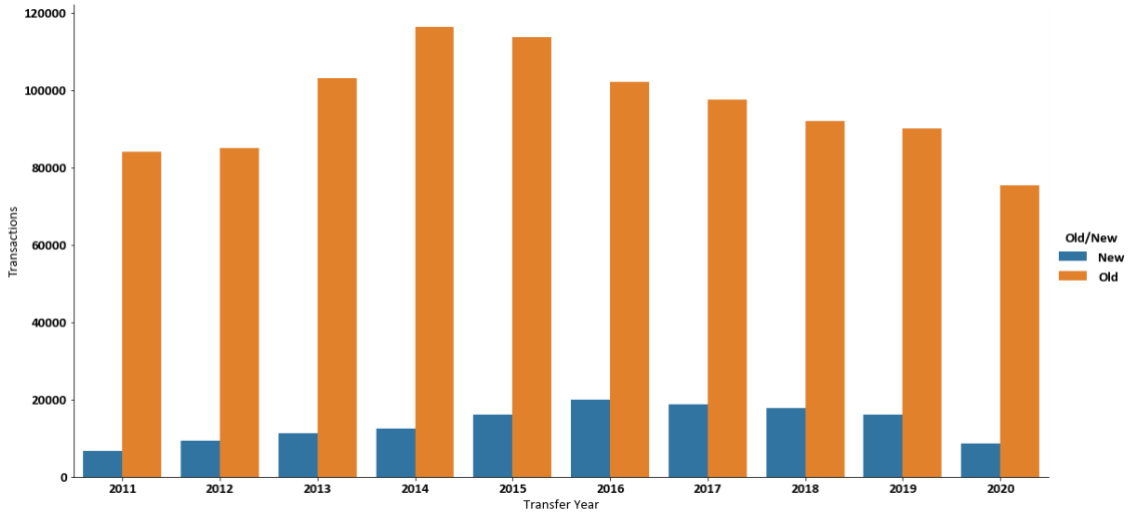


Figure 4.34: Distribution based on new/old

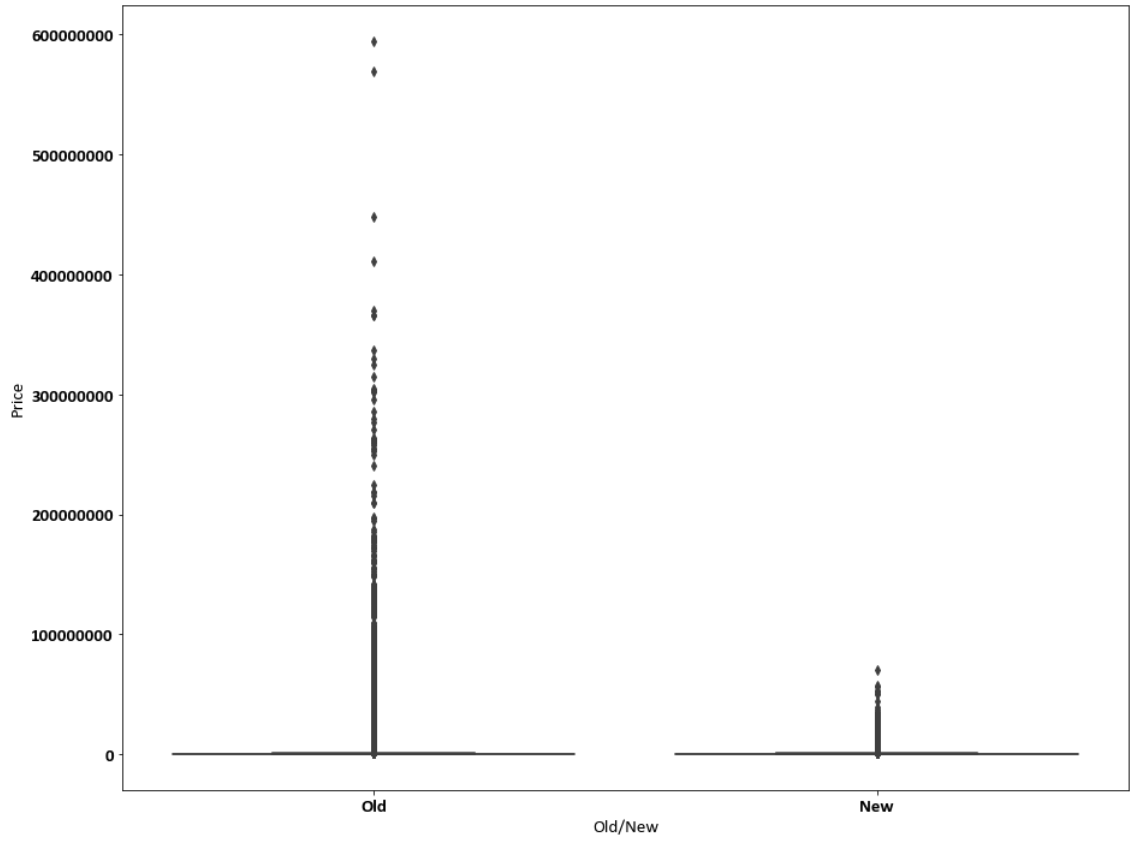


Figure 4.35: Boxplot for new/old with outliers

4.9.5 Further data insights

The extraction of London based transactions from the HM Land Registry price paid data leaves a total of 1,097,302 transaction in the research dataset. Across the London boroughs, Appendix 1 shows that 'City of London' had the least volume of transactions every year between 2011 and 2020. Other boroughs that featured in the bottom three included Barking and Dagenham, Newham, Harrow, Islington, and Kensington and Chelsea. The London Borough of Wandsworth consistently recorded the highest volume of transactions for nine years from 2011 to 2019, but lost the position to the London Borough of Bromley in 2020. The margin that the London Borough of Bromley has in 2020 was enough to be the borough with the highest volume of transactions over the ten-year period because it has consistently been in the top three. Other boroughs that featured in the top three include Barnet, Lambert and Croydon. This shows the areas where more stakeholders in the housing market are either moving to or investing in.

As a consequence of the low volume of transactions, the total value of transactions for City of London was lower than the top performers, but it performed better than a few other boroughs that never featured in the bottom three for volume of transactions. These include the London boroughs of Bexley, Sutton, Redbridge, Waltham Forest, Havering, Kingston upon Thames and Enfield. The London Borough of Barking and Dagenham performed the least, with a total value of transactions being just under six billion pounds (see Appendix 4). The City of Westminster consistently featured in the top three for total value of transactions, as shown in Appendix 3, and consequently the overall highest with over eighty billion pounds (Appendix 4).

4.10 Features engineering

The features engineering module is a module of modules, as it is made up of three modules each managing the engineering of the features in the datasets for

each tier, as shown in Figure 4.4. These sub modules, as shown in Figure 4.36, are: (i) Tier 1; (ii) Tier 2; and (iii) Tier 3.

Tier 1 is a prime sub module because it facilitates the foundational step to one of the contributions to knowledge of this research as stated in Section 1.6, the research dataset. In this sub module, the ONS postcode data with longitude and latitude features alongside all existing postcodes is blended with the Land Registry Price Paid Data which also has the postcode of houses as a feature. The creates a geo-coded version of the Land Registry's Price Paid Data.

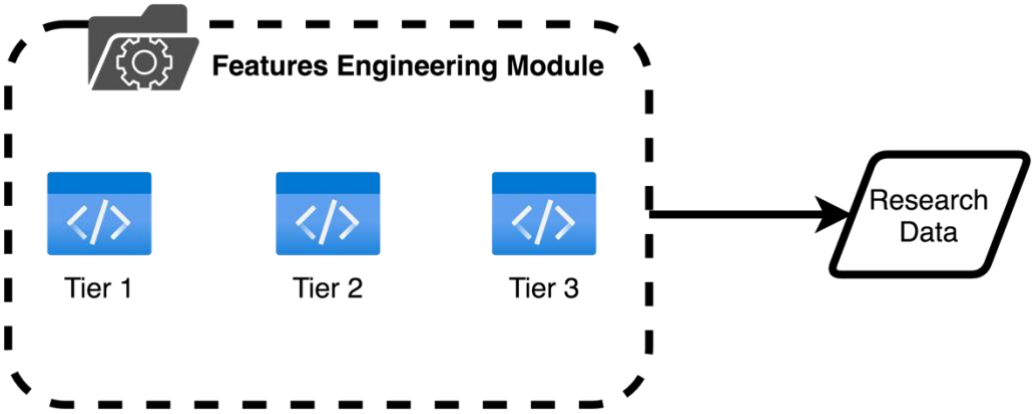


Figure 4.36: MfHPE features engineering module

Tier 2, the second sub module, facilitates engineering of Tier 2 features through the second layer of data enrichment to the Price Paid Data. This sub module blends the geo-coded base data, the output of the first sub module with three location enabled datasets being: (i) rail stations (including underground stations); (ii) supermarkets; and (iii) bus stops.

The nearest neighbour algorithm was then used to calculate the distance between the nearest rail station, supermarket and/or bus stop and each house with a sale record in the price paid data. For the purpose of this research, binary features were also created to flag if each of the neighbourhood features stated above is within a 2km radius of a sold house. Although 2km has been used, this could be changed to match the particular requirements of any user of the MfHPE framework.

Finally, Tier 3 is the final value-add sub module as it facilitates the enrichment of the already geo-coded and neighbourhood feature enabled base transactions data with macroeconomic indicatords including: (i) GDP; (ii) unemployment rate;

(iii) employment rate; (iv) Consumer Price Index; and (v) inflation rate. For the timeline in focus being 01/01/2011 to 31/12/2020, the monthly or quarterly value for each macroeconomic indicator is assigned to every house sold in the same month or quarter. In essence, this creates a richer data set, the **research data**.

Table 0.17: Modelling-ready research data including standard and engineered features

Attribute	Data Type	Dataset
Transaction-id	Object	Price Paid Data
Property Type	Object	Price Paid Data
Old/New	Object	Price Paid Data
Duration	Object	Price Paid Data
Station_less_2km_count	int64	GB Rail Stations
Station_less_5km_count	int64	GB Rail Stations
Station_shortest_distance	float64	GB Rail Stations
Station_within_2km	int64	GB Rail Stations
store_less_2km_count	int64	UK Supermarket
store_less_5km_count	int64	UK Supermarket
store_shortest_distance	float64	UK Supermarket
store_within_2km	Int64	UK Supermarket
busstop_less_2km_count	int64	Bus Stop
busstop_less_5km_count	int64	Bus Stop
busstop_shortest_distance	float64	Bus Stop
busstop_within_2km	int64	Bus Stop
Transfer_Year	int64	Price Paid Data
Transfer_Month	Int64	Price Paid Data
Transfer_Day	Int64	Price Paid Data
Quarters	object	Price Paid Data
GDP	float64	ONS GDP
Employment_Rate	float64	ONS Employment Rate
Unemployment_Rate	float64	ONS Unemployment Rate
Inflation_Rate	float64	ONS Inflation Rate
CPIndex	float64	ONS CPIH
Latitude	Float64	ONS NSPL Product
Longitude	Float64	ONS NSPL Product
Price	int64	Price Paid Data

4.11 Baseline model building

The modelling-ready data was split into training and test sets, while data for 2021 was kept as unseen data for the validation of modelling results. Thirty baseline models were created for Tier 1, Tier 2 and Tier 3 data using five modelling techniques but using default parameters (i.e. no tuning at this point). These models have been described as ‘baseline’ because they have used the default parameters of each algorithm (LightGBM, 2021, XGBoost, 2021, Scikit Learn, 2021). These are: (i) LightGBM; (ii) XGBoost; (iii) Random Forest; (iv) Hybrid Regression; and (v) Stacked Generalisation.

4.11.1 LightGBM

LightGBM (Light Gradient Boosting Machine) is a machine learning algorithm based on a decision tree algorithm. LightGBM is free and open-source, and based on a gradient boosting framework for machine learning, developed by Microsoft. It has wide application in real-life such as **ranking**, **classification** and tasks based on machine learning. Its development focuses on performance and scalability. LightGBM has the following advantages: sparse optimisation; early stopping; parallel training; bagging; regularisation; multiple loss functions. One major difference between XGBoost and LightGBM is in the construction of trees. XGBoost grows trees level-wise, while LightGBM grows tree leaf-wise, as shown in Figure 4.37. LightGBM selects a tree that produces the largest decrease in loss through optimisation. **Exclusive Feature Bundling** (EFB) and **Gradient-based one-side sampling** (GOSS) are the two powerful techniques used by LightGBM to improve accuracy, efficiency and memory consumption as well as speed (Ke et al., 2017).

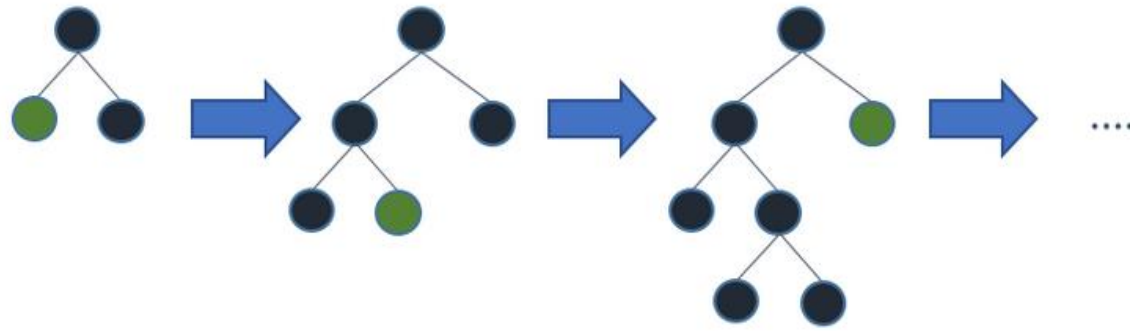


Figure 4.37: Leaf-wise tree growth for LightGBM

Truong et al. (2020) compared the performance of various machine learning algorithms, i.e. Random Forest, LightGBM, XGBoost, Hybrid Regression and Stacked Generalisation, with the observation that LightGBM is the most useful algorithm with regard to time and accuracy dimensions. Hybrid Regression and Stacked Generalisation are the worst for time utilisation, since they apply K-fold cross-validation in their mechanism. However, their results are as accurate as other methods. The default parameters used for the baseline modelling are as shown in Table 4.18.

Table 0.18: Baseline LightGBM parameters

Parameter	Description	Value
`objective`	This can be '`regression`' for LightGBM Regressor, '`binary`' for LightGBM Classifier and '`lambdarank`' for LightGBM Ranker	`regression`
`num_leaves`	This is the maximum number of tree leaves for base learners	36
`learning_rate`		0.15
`n_estimator`	This is the number of boosted trees to fit	64
`min_child_weight`	This is the minimum sum of instance weight required in a child	2
`colsample_bytree`	The subsample ratio of columns during the construction of each tree	0.8
`reg_lambda`	This is the L2 regularisation term on weights	0.45

Source: (LightGBM, 2021)

4.11.2 XGBoost

XGBoost exists as an open-source package for tree boosting. It is known to be a scalable machine learning system. Its scalability feature allows users to quickly define their objectives. Other features that are inherent in this algorithm are: very fast as it performs parallel computation; accepts a wide array of inputs; sparsity; customisation; and better performance (Chen and Guestrin, 2016).

XGBoost mostly attributes it success to scalability, where it scales to billions. This scalability is enabled by algorithmic optimisations such as the novel tree learning algorithm for handling sparse data (Chen and Guestrin, 2016). XGBoost is capable and able to conveniently solve real-world scale problems with minimal resources. Avanijaa (2021) also explored XGBoost in predicting house prices and considered a range of steps. The value and description for the default XGBoost parameters used are as shown in Table 4.19.

Table 0.19: Baseline XGBoost parameters

Parameter	Description	Value
booster	There are a range of boosters to explore. These are gbtree, gblinear or dart.	gbtree
verbosity	There are four valid values including 0 – silent, 1 – warning, 2 – info and 3 – debug.	1
validate_parameters	This is a parameter that is still being experimented. It performs a validation of the input parameters as assess whether or not the parameter is used.	False
nthread	The nthread is the number of parallel threads used to run XGBoost	
disable_default_eval_metric	This is a flag used to disable default metric	False
num_pbuffer	This is the size of the prediction buffer. It is normally set to the number of training sets	set automatically by XGBoost
num_feature	This is the feature dimension used in boosting. It is set to the maximum dimension of the feature	set automatically by XGBoost

Source: (XGBoost, 2021)

4.11.3 Random Forest

Random Forest is a machine learning (ML) algorithm based on weak learner decision trees (Wang and Wu, 2018). The ensembling manner of the decision trees eliminates instability problems as well as overcoming the high variance of decision trees, since decision trees are generated by a random sampling method, hence the name Random Forest. The algorithm randomly selects a subset of explanatory variables which are trained with weak learners separately. Hence, it builds the prediction model by independently training a weak prediction model: decision trees. Finally, the prediction of each tree is averaged.

Some of the advantages of Random Forest over other machine learning algorithms, as noted by (Antipov and Pokryshevskaya, 2012), are: first, it has strong performance; second, it conveniently handles categorical data with many levels; third, it adequately works with missing data; fourth, it allows for nonlinearity and unsteadiness of variables lastly, it does not require detailed model specification. Random Forest has a shortcoming though (Oshiro et al., 2021) – the large number of trees in the forest increases computational costs without significant performance gain. Therefore, (Oshiro et al., 2021) observed that RM

with trees between 64 and 128 strikes a good balance between processing time and memory usage. (Probst and Boulesteix, 2017) established that the biggest performance gain is realised in the first 100 trees.

Wang and Wu (2018) used Random Forest to estimate the house prices in Arlington, Virginia. The results were compared to linear regression model results and they established that the Random Forest algorithm is a superior estimation method since is captures the non-linear relations between price and given features of the house. Similarly, (Hong et al., 2020) benchmarked the performance of Random Forest with Ordinary Least Square (OLS) linear regression. OLS is hedonic based and has been criticised for imposing strong assumptions on the model. Hong et al. (2020), (Čeh et al., 2018) established that random forest is by far a superior algorithm. The value and description for the default Random Forest parameters used are as shown in Table 4.20.

Table 0.20: Baseline Random Forest parameters

Parameter	Description	Value
n_estimator	This is the number of trees in the forest	100
criterion	A function that measures the quality of a split	mse
max_depth	This is the maximum depth of the tree	None
min_samples_split	This is the minimum number of samples needed to split an internal node	2
min_samples_leaf	This is the minimum number of samples needed to be at a leaf node	1
min_weight_fraction_leaf	Minimum weighted fraction of the sum of total weights needed to be at a leaf node.	0.0
max_features	No of features to consider for the best fit	Auto
max_leaf_nodes		None
min_impurity_decrease		0.0
min_impurity_split		None
bootstrap		True
oob_score		False
n_jobs		None
random_state		None
verbose		0
warm_start		False
ccp_alpha		0.0
max_samples		None

Source: (Scikit-Learn, 2021)

4.11.4 Hybrid Regression

Hybrid Regression is and ad-hoc and user-defined ML method. In this regard, Hybrid Regression entails combining two or more ML methods to develop a unique ML method. The combined outcome of these ML methods is far better than the results of each ML method by itself. For instance, (Truong et al., 2020) tested houses Prices in Beijing with a model consisting of 33.33% Random Forest, 33.3% LightGBM and 33.3% XGBoost. The hybrid model achieved a better result with a RMSE of 0.14969, far better than each algorithm run separately. Similarly, (Lu et al., 2017) realised a better overall result using a hybrid model consisting of 65% Lasso and 35% Gradient Boosting.

4.11.5 Stacked Generalisation

This method was introduced by Wolpert (1992). It is a Python-based package, and the main idea behind this method is to use predictions of the previous models as features for the present model. Stacked Generalisation uses K-fold cross-validation to avoid overfitting. (Truong et al., 2020) used 2-level stacking architecture, the first stack comprised of Random Forest and LightGBM while the second stack was comprised of XGBoost to predict the house prices. They found that the combined results were not as impressive as the Hybrid Regression.

4.12 Model optimisation

Model optimisation in machine learning is, without doubt, one of the most challenging aspects of the implementation of ML solutions. There is immense attention given to deep learning theories and machine learning to achieve the optimisation of models. There are two types of algorithm parameters usually considered when building machine learning systems: (i) model or default parameters, which possess the ability to be initiated and consequently updated through data learning; and (ii) hyperparameters, the parameters which are used to configure a machine learning model and to specify the algorithm which is used in minimising the loss function (Zhang, 2012). Section 4.2 of the development of

this MfHPE framework focuses on the exploitation of the default or model parameters of each algorithm used, while this section will focus on the use of hyperparameters to tune and further improve the accuracy of the machine learning models. However, because of (i) computational cost, (ii) the performance of the baseline models, and (iii) the fact that the two ensemble techniques are composed of the standalone techniques, only the LightGBM, Random Forest and XGBoost models will be optimised.

4.12.1 Hyperparameter tuning

Hyperparameters differ from first-level model parameters because they are second-level tuning parameters that achieve maximum performance once they are carefully optimised. Further to the development of thirty baseline models using default parameters, optimal hyperparameters were selected for the configuration of the baseline models in a bid to improve their performance and accuracy. Table 4.21 provides a generic view of various approaches to model optimisation to achieve a balance between generalisability and accuracy.

Table 0.21: Approaches to hyperparameter tuning

Approach	Methodology
Manual hyperparameter tuning	This is a traditional approach that makes use of trial and error. In this case, the model user makes a 'guess' on which parameter values would have the highest accuracy in the model. This process, however, is relatively slow and more prone to human error, therefore necessitating the search for a more sophisticated and automated approach for faster results (Agrawal, 2020).
Grid search	This is a combination of all the hyperparameter values which can possibly be used to create a large hyperparameter set that can be reproduced and has the ability to be automated (Panda, 2019). It is used extensively because it is a relatively easy model to implement. The grid search model is not limited in its applicability in machine learning models – it can be used across all model types to evaluate the best parameters to use (Chauhan, 2020a). There are a couple of problems that make grid search less optimal in hyperparameter tuning: the function evaluations grow exponentially, therefore introducing dimensionality within the configuration space; and the increase in resolution of discretisation increases the number of function

	evaluations, thus escalating dimensionality further. This means that a grid search will guarantee the production of the perfect solution.
Random search	Random search chooses the values in random sample points in a grid. The algorithm sets up a grid of hyperparameter values and selects different random combinations which it trains thus giving the user more control over the number of parameter combinations that will be attempted. Random search performs better because it does not make assumptions about the machine learning model. It is considered to be the best search approach when there are fewer dimensions, given the fact that less time is taken to find the right set with fewer iterations (nuggets). One of the major challenges of the random search approach is that it is relatively non-adaptive in nature, given the fact that the hyperparameter sets which are selected are not easily reproducible.
Smart hyperparameter tuning	Smart hyperparameter tuning selects a few hyperparameter settings, evaluates the quality of their validation matrices, adjusts the hyperparameters and consequently re-evaluates the validation matrices (Chauhan, 2020a). This process is not parallelisable given that the process is sequential and inherently iterative. The goal of smart hyperparameter tuning is to make fewer overall evaluations while saving on the overall computational time. The following are some examples of smart hyperparameter tuning approaches: Hyperopt, which tunes hyperparameters using tree-based estimators; and Spearmint, which optimises hyperparameters using Gaussian processes.
Bayesian Optimisation	Bayesian Optimisation is one of the sequential model-based optimisation algorithms which allow results from previous iterations to improve the sampling method in the next experiment. The model has two main components: a surrogate model which applies probability and an acquisition function that decides which point to evaluate next (Quitadadmo et al., 2017). As the observation points increase, the posterior distribution improves, thus giving the algorithm more certainty on which parts of the model are explorable and which are not. Bayesian hyperparameter tuning applies Gaussian processes to model the target functions (Feurer and Hutter, 2019).
Gradient-based optimisation	This is used specifically for neural networks to compute gradient with consideration to hyperparameters and optimises them by applying gradient descent algorithm. The approach is applicable when a range of continuity conditions and differentiability conditions are adequately satisfied (Choudhury, 2021).
Multi-fidelity optimisation	With increased dataset sizes and more sophisticated models, training a single hyperparameter configuration on large datasets is a long and expensive process. Multi-fidelity methods are used to speed up the process by introducing an algorithm configuration within a small

	dataset and optimising these via cross-validation. The model uses low-fidelity approximations to minimise the loss function. This therefore increases speedups, which outweigh the approximation error while introducing a trade-off between performance and runtime (Feurer and Hutter, 2019).

4.13 Evaluation of machine learning models

Machine learning is the ability to improve performance through experience gained by analysing information and generalising it for the extraction of new knowledge through automation. Machine learning offers the unequivocal ability for computers to gain insight or experience from data and consequently help in making better predictions for future scenarios. The process of machine learning involves the construction of different forms of mathematical models to understand new data by fitting it to previously seen data and predicting the newly observed data. The goal of ML is not to create new models but rather to ensure that ML models' predictive power is high – ML models are only useful if the quality of their predictions is high. Therefore, after the creation of a machine learning model and subsequently training it with some data, it is pivotal to conduct some evaluation of the predictive power of the model. Building a fit machine learning model entails a critical process of checking for errors and ensuring that all the gaps in the system are removed and the relevant improvements made, which is defined as machine learning model evaluation.

Model evaluation is the process of evaluating the correctness and accuracy of machine learning models on test data. Evaluating machine learning models has been a critical process in the development and application of machine learning systems. Every machine learning model has several limitations; no model is perfectly accurate because they depend on estimations which have limitations relative to data distribution. Jordan (2017) proposes that the following three questions should guide the evaluation of a machine learning model. Is the model useful? What more features are needed to improve it? Will more training of the model with data improve its performance? Using different metrics to evaluate the performance of a machine learning model increases the predictive power of the model.

As mentioned above, machine learning models are not perfect – their accuracy needs to be tested using different metrics. The tables below present a couple of machine learning evaluation metrics that are used in machine learning evaluation, their approach and how they are calculated (formula).

4.13.1 Classification metrics

The discussion of values in this section will be based on common binary classification approaches as shown below – assuming a person has a computer tablet (positive) or does not have a computer tablet (negative) (Nighania, 2018).

True Positives (TP): Predicted to have a computer tablet and has it

False Positives (FP): Predicted to have a tablet and does not have a computer tablet (also referred to as a Type 1 error).

True Negatives (TN): Predicted not to have the tablet and does not have it.

False Negatives (FN): Predicted not to have a tablet but has it (also referred to as a Type 2 error).

Table 4.22 provides an overview of machine learning model evaluation metrics exploited for classification problems.

Table 0.22: Metrics used for the evaluation of classification problems

Metric name	Description	Approach
Confusion Matrix	After a machine learning model has been created, it is imperative to evaluate the performance of the model. A confusion matrix is the representation of the binary parameters presented above in a matrix form (Kumar et al., 2021). It gives the results of any binary testing in a matrix representation and forms the basis of other classification machine learning evaluations metrics such as accuracy score, error rate, precision, specificity and recall. It is also a standard model used in evaluating statistical models, and forms a basis for the creation of ROC graphs (Bhattacharya, 2019).	<table><tr><td>True Positives (TP)</td><td>False Negatives (FN)</td></tr><tr><td>False Positives (FP)</td><td>True Negatives (TN)</td></tr></table>
Accuracy Score	An accuracy score is a simple evaluation metric to use because it is simply the proportion of the observations which have been made correctly in relation to the whole dataset. Accuracy score is a common evaluation metric for classification problems as it gives the number of correct/accurate predictions made in relation to all the predictions made (Novaković et al., 2017). In binary form, accuracy is measured depending on the number of positives in the whole	$(TP + TN) \div (TP + FP + TN + FN)$

	data set. It can therefore be calculated with ease by dividing the total number of correct predictions by the total number of predictions in the dataset (TP + TN). Accuracy score is simple in approach but suffers from a paradox of applicability, more so in imbalanced classes where the accuracy level is high, but has loopholes in predictive power.	
Error rate	This is ideally an extension of the accuracy score; which entails classification of the error rate rather than success rate (Chauhan, 2020b).	$(FP + FN) \div (TP + FP + TN + FN)$
Recall	A recall evaluation is a proportion of observations that are predicted to belong in the positive class and truly belong in the positive class (Analytics Vidhya, 2021). It gives an idea of the ability of a machine learning model to observe which observation truly belongs in the positive class. Ideally, this model answers the question: if a prediction value is positive, how often does the machine learning model predict that the value is positive? The fraction gained (true positive rate) gives a better way of evaluating the performance of a machine learning model when there is a class imbalance (Novaković et al., 2017).	$TP \div (TP + FN)$
Specifity	This is the ratio of negative instances from the total actual negatives. This approach is similar to Recall, but it shifts the attention to the negative instances (Nighania, 2018).	$TN \div (TN + FP)$
Precision	Precision is a classification approach that is used to identify correctness in the true positives observed. The equation gives a ratio of the positive predictions (true positives) to the total number of predictions that were made as positives (true positives and false positives) (de Melo Junior et al., 2017). The higher the ratio, the higher the precision, and as such the greater the ability of a model to classify the positive classes correctly. This model is mainly applicable when there is a need to identify the correct positive classes and subsequently reduce the number of false positives. It can therefore be simply defined as the fraction of all the true positives with the summation of true negatives and true positives.	$TP \div (TP + TN)$
F1 Score	Recall and precision are useful in cases where classes are not evenly distributed. A combination of both values can be used to increase the accuracy of a prediction by measuring the overall accuracy of the model within a positive prediction environment. F1 score can therefore be ideally defined as the harmonic mean between recall and precision, thus the name 'harmonic mean of the precision and recall evaluation metrics' (de Melo Junior et al., 2017). Since it is a combination of recall and precision, a high F1 score means a higher accuracy of the model. A model which does well in the F1 score predicts the actual positives – precision; does not	$(2 * Precision * Recall) \div (Precision + Recall)$

	miss any of the positives and predicts the negatives correctly – recall. Further accuracy can be achieved by the application of the PR curve, which is a curve between recall and precision for values in various thresholds (Nighania, 2018).	
ROC Curve	The Receiver Operating Characteristics Chart is an evaluation metric that is used to check the performance of classification models. ROC curves are used to evaluate classifiers that contain only two target classes. It is plotted on a two-dimensional plot called the ROC space. The x-axis in this chart is computed as the false approved rate (FDR), while the y-axis is computed as the true approved rate (TDR) (de Melo Junior et al., 2017). The area under the ROC curve is used to assess the quality of the classification model. The receiver operating characteristics graph, therefore, helps separate a signal from the noise by visualising how well a machine learning classifier performs (Kumar et al., 2021).	TPR = $TP \div (TP + FN)$ FPR = $FP \div (FP + TN)$
Gini Coefficient	The Gini Index is mainly applied for class values that are imbalanced. It was developed by Corrad Gini as a statistical measure of distribution with coefficients that range from 0 to 1, with 0 being perfect equality and 1 being perfect inequality (Bhattacharya, 2019). If the value of the coefficient is high, the data will consequently be more dispersed and vice versa. The Gini Coefficient can be calculated from the ROC curve by mathematical evaluation of the area under the curve, as shown in the formula.	Gini Coefficient $=(2 * ROC\ Curve) - 1$
Gain and Lift Chart	This is a measure of the effectiveness of classification models through the use of a graph as a visual aid to help in the evaluation of the performance of classification models (Choudhury, 2019). While the confusion matrix assesses a machine learning model on a whole dataset, the gain and lift chart assess the model on a portion of a population. As such, when a lift is higher from the baseline, the better the ML model.	This is calculated as a ratio between the results observed when the model is used and the results observed when the model is not used.

4.13.2 Regression metrics

As seen in existing literature, a range of evaluation metrics have been explored to measure the performance of regression-based ML algorithms. These include: (i) Root Mean Squared Error (RMSE); (ii) Mean Absolute Error (MAE); (iii) Mean Squared Error (MSE); (iv) Mean Percentage Error (MPE); (v) Explained Variance; (vi) R-squared; and (vii) Adjusted R-squared. Shinde and Gawande (2018) compared the accuracy of a range of regression models based on a range of error metrics including MAE, MSE, R-squared and RMSE. Table 4.23 provides

an overview of machine learning model evaluation metrics exploited for regression problems.

Table 0.23: Metrics used for evaluating regression problems

Metric	Description	Approach
Explained Variance	This approach compares the variance in a data set with the predicted variance. This gives an idea of the amount of variation in the actual data set that the model was able to explain (Jordan, 2017).	EV = $1 - ((Var(ytrue - ypred) \div ytrue)$
Mean Squared Error	This is the average of the squared errors which are used as the loss function. The mean square error is the sum of the difference between the predicted variables and the actual variables (Jordan, 2017). The MSE, therefore, helps in defining how close a set of points is to a regression line.	Find the regression line. Insert the values of Y in the regression equation and use these to find Y values. Subtract the new values of Y from the original values. Square the errors, add them up and find their mean.
Root Mean Squared Error	The Root Mean Squared Error is defined as the difference between the values predicted by a model and the values which are observed (Kumar et al., 2021). It is the square root of the mean squared error, which is explained above.	$\sqrt{Mean\ Squared\ Error}$
R-squared	This is a statistical measure that is applied to assess the goodness of fit in a regression model. In this approach, the ideal measure would be 1. Therefore, the closer the value is to 1, the better it can be fitted into the model. The R-squared value is a comparison of the sum of squares with the total number of squares. The R-squared tends to have an inability to decrease when new parameters are added, thus limiting it to assessing whether the model does better with fewer parameters and/or worse with more parameters (Nighania, 2018). This necessitates the optimisation of the R-squared model, thus the creation of Adjusted R-squared (see below).	R Squared =$1 - (Variance(model)/ Variance(average)$
Adjusted R-squared	The Adjusted R-squared removes the inability of the R-squared to reduce in value when new parameters are added by essentially punishing/penalising the value as more features and parameters are added into the R-squared (Chauhan, 2020b).	$adjR_2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$

However, in this research the performance of the machine leaning models was measured using regression metrics rather than classification metrics, as the problem this research seeks to provide answers to is a regression problem. The three metrics exploited in this research are further discussed below:

RMSE is a good estimator for the standard deviation σ of the distribution of our errors! In data science, RMSE has a double purpose: (i) To serve as a heuristic for training models and (ii) To evaluate trained models for usefulness / accuracy This is a standard way to measure the error of a model in predicting quantitative data. (Moody, J., 2019)

MAE tells us how big of an error we can expect from the forecast on average.
As the name suggests, the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. MAE is robust to data with outliers

R-Square

The most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 60% reveals that 60% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

However, it is not always the case that a high r-squared is good for the regression model. The quality of the statistical measure depends on many factors, such as the nature of the variables employed in the model, the units of measure of the variables, and the applied data transformation. Thus, sometimes, a high r-squared can indicate the problems with the regression model.

A low r-squared figure is generally a bad sign for predictive models. However, in some cases, a good model may show a small value.

There is no universal rule on how to incorporate the statistical measure in assessing a model. The context of the experiment or forecast is extremely important and, in different scenarios, the insights from the metric can vary.

4.14 Model explainability

Model Explainability is a wide subsection that seeks to analyse and understand the results that machine learning systems yield. Explainability is applied in the case of black box models whose results are hard to elucidate. Black box models include deep neural networks, vector machines and decision trees with many nodes (Roscher et al., 2020). In some cases, developers of the algorithms also have a hard time understanding and spelling out how the models make the decisions. Preventing the exploitation of the model is harder when the designers do not understand the features which are most salient in how a model arrives at its predictions. Human beings need an explanation of certain decisions which are made by machine learning models. Burkart and Huber (2021) highlight the following as the motivating aspects of explainability:

Trust: The strengths and weaknesses that a model must identify in order to make a judgement about the success of the system.

Transferability: Models need to showcase their ability to make decisions, similar to those gained from training data, before they can be deployed into post-development use.

Informativeness: It is critical to identify whether the system actually serves the purpose it was intended for in the real world. This goes beyond the purpose which the system was trained to achieve.

Adjustments: Developers should have the ability to make changes to the prediction models by changing the code or changing the parameters. Explainability is therefore core in identifying bugs and failure modes.

Accountability: Where prediction problems and data shift challenges a system, the system should have the ability justify and explain its decisions.

Ethical Decision Making: there is a variety of laws which have been put in place by bodies such as the EU to ensure fairness and conformity of systems to social and legal standards. Explainability offers a robust way of analysing whether a model conforms to the stipulated ethical standards.

4.14.1 Interpretability and explainability

Explainability and interpretability have often been used interchangeably as a definition of the level to which a ML model is understandable. However, there are subtle differences between the two terms. On one hand, interpretability refers to the ability of a model to be understood by its users who apply the model. An interpretable model makes it easier for a user to verify whether the algorithm satisfies the purpose it is intended to serve (Bibal et al., 2021). It is imperative for the end user to understand why a model behaves the way it does. A transparent and interpretable model builds trust the consumer has in a model and shows that they can rely on the model for future applications.

Explainability gives data scientists and developers insight into the behaviors of a model and helps in addressing the challenges that are experienced in building and deploying machine learning models such as model training, debugging, monitoring, transparency and audit (Bhatt et al., 2020). Machine learning stakeholders use interpretability and explainability interchangeably because their purpose is making machine learning models understandable to a human observer. Explainability is not demanded in every domain that applies machine learning. Model explainability is specifically demanded in fields where critical decisions have to be made: those involving human lives, such as medicine, judicial systems and information; and those that involve huge amounts of money, such as finance and banking (Burkart and Huber, 2021).

4.14.2 Scope of explainability

There are different approaches used to explain a machine learning model. They are widely grouped into two: global explainability and local explainability. Local explainability focuses on a single prediction from the model and maps it out in order to highlight its important features and contrast it with other predictions made by the model (Burkart and Huber, 2021). Global explainability is an approach that attempts to explain the whole model at once by characterising all the predictions made by the model (Bhatt et al., 2020). Table 4.24 shows an overview of some of the commonly used model explainability methods.

Table 0.24: An overview of some of the commonly used model explainability methods.

Method	Application
Feature Importance	Feature importance describes how important a feature is in the performance of a model. It is referred to as an individual measure of the contribution of a feature in the classifier regardless of its shape or direction. One of the most applied models for measuring feature importance is based on Shapley values, which define the cooperation of the features in a dataset with the prediction from the model (Bhatt et al., 2020).
Local Interpretable Model-agnostic Explanations (LIME)	LIME generates random points within a sample, computes their output from the model and subsequently trains a model which is embedded on top of the output. The surrogates have to be simple and explainable so as to approximate and analyse the predictions in the underlying model. LIME is applicable for tabular data as well as image- and text-based data (Garbacz, 2021).
Individual Conditional Explanation (ICE)	ICEs are used to display a line on a data point to plot a graph that gives insight into how the data point varies as a feature undergoes change across datapoints. ICE plots show interactions and individual differences of datapoints by disintegrating partial dependence of the points. The plot then visualises one variable at a time in order to reduce chances of a visualisation overload (Wright, 2018).
Partial Dependence Plots (PDPs)	A PDP shows the effect which a single feature has on the overall outcome of a machine learning model. It works by segregating the feature of interest from another feature. Its equation for regression is: $\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$ The features of less importance are denoted by *c* while the feature of interest are denoted by *S*. PDPs marginalise these features and arrive at a function that is dependent on *S*, thus making it easier to understand how they influence the prediction made by the model (Aguiar, 2019).
Accumulated Local Effects (ALE) plot	This is an unbiased and faster alternative to partial dependence plots. It is more effective when applied to correlated features. PDPs are biased in correlated features because they give samples which vary highly. ALEs address this by establishing the changes in predictions rather than the averages (Aguiar, 2019).

4.15 Conclusion

This chapter has focused on the approach taken for the delivery of the cumulative Multi-feature House Price Estimation framework. The approach leverages the Design Science Research Methodology which is made up of six unique stages that underpin the framework design. Ten datasets from multiple sources were exploited in a modularised framework comprising nine modules. Exploratory Data Analysis show that price distribution in the HM Land Registry Price Paid Data for London was skewed as a result of residential houses with significantly high prices. Two transformation techniques, log transformation and Box-Cox transformation, were used to normalise the price distribution curve, however

there was no significant difference between the results of both. The outlier effect was calculated by creating an outlier detector based on the concept of interquartile range and results did not provide enough justification for the removal of outliers because of domain knowledge. Multiple features were engineered in preparation for machine learning models which leveraged both baseline and optimised parameters of a total of five algorithms. Three evaluation metrics were exploited and both local and global explainability were explored to understand the features that drove the predicted house price for specific houses and the entire dataset respectively.

Chapter 5: Results, Evaluation and Optimisation

5.1 Introduction

Further to the details provided in Chapter 4 on the research data, Design Science Research Methodology and the creation of a robust pipeline which enabled a seamless development experience cutting across the implementation of an exploratory data analysis, data pre-processing, features engineering, modelling, model evaluation, model optimisation and explainability, this chapter can be described as the '**Demonstration**' stage and will focus on a review of baseline modelling results for all the algorithms explored, using default parameters and optimisation of each model produced. In the next chapter, the evaluation of each model will be presented using evaluation metrics discussed in Section 4.13.2, selection of the best performing model, the explainability of the best performing model and testing.

5.2 Baseline model results and evaluation

As discussed extensively in Section 4.11, thirty baseline models have been created and three metrics – Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-squared – have been used to measure the accuracy of their results. The figures that follow are charts that have been created with only the first 200 records from the modelling results for each tier and algorithm. This selection of transactions has house prices ranging from below £250,000 to over £1,000,000 with multiple characteristics and locations.

Figure 5.1 shows a plot of Tier 1-based LightGBM predicted prices in comparison with the actual prices. This shows that in some cases the predicted prices are higher than the actual prices, in some cases lower, whilst in some cases a very near or perfect match. Figures 5.2 and 5.3 show the results for LightGBM models using the default parameters after a layer of neighbourhood and macroeconomic features have been introduced, respectively. The introduction of Tier 2 and Tier 3 features has been described in this thesis as **cumulative multi-feature**

layering. Figures 5.2 and 5.3 show an improvement in the accuracy of the predicted prices for some of the transactions due to the introduction of neighbourhood and macroeconomic features respectively.

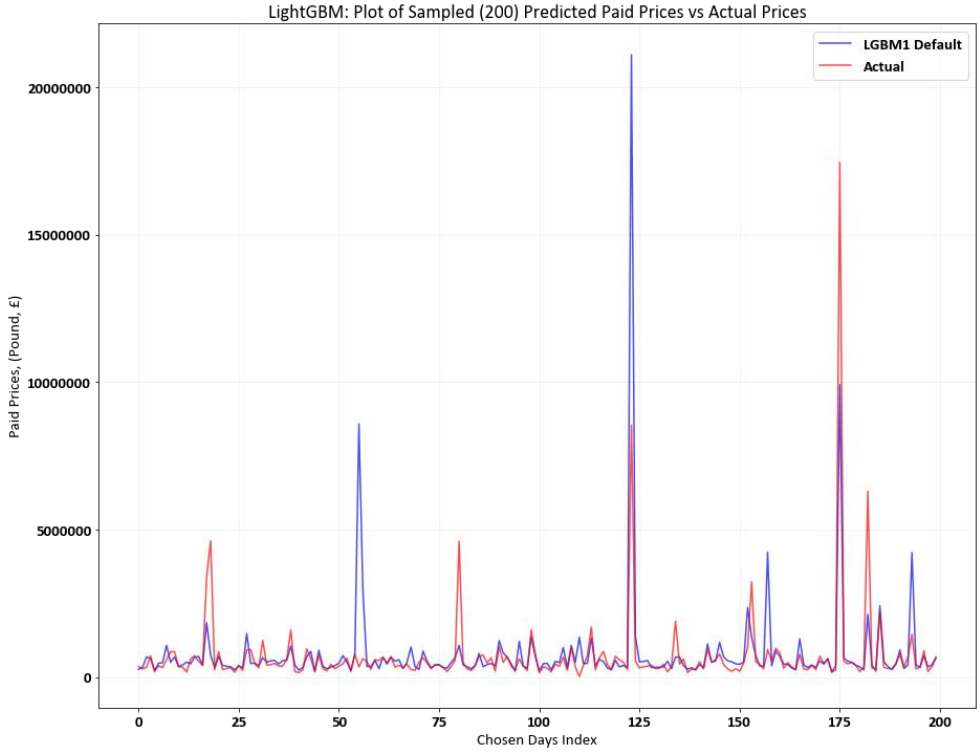


Figure 5.1: LightGBM (default) Tier 1 predicted paid prices versus actual prices

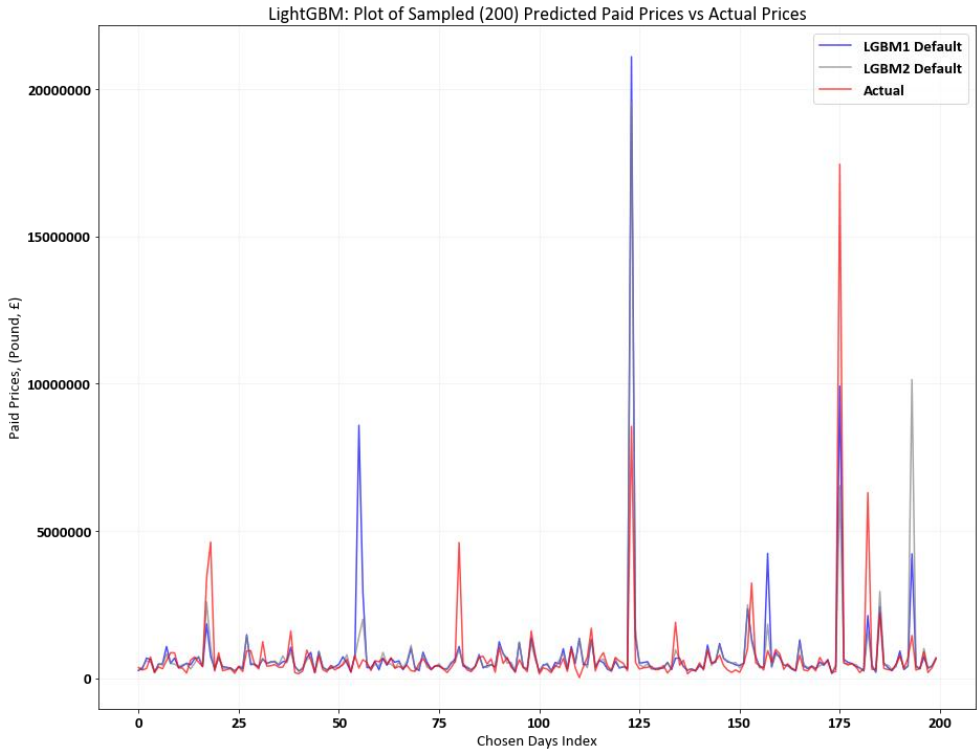


Figure 5.2: LightGBM (default) Tier 2 predicted paid prices versus actual prices

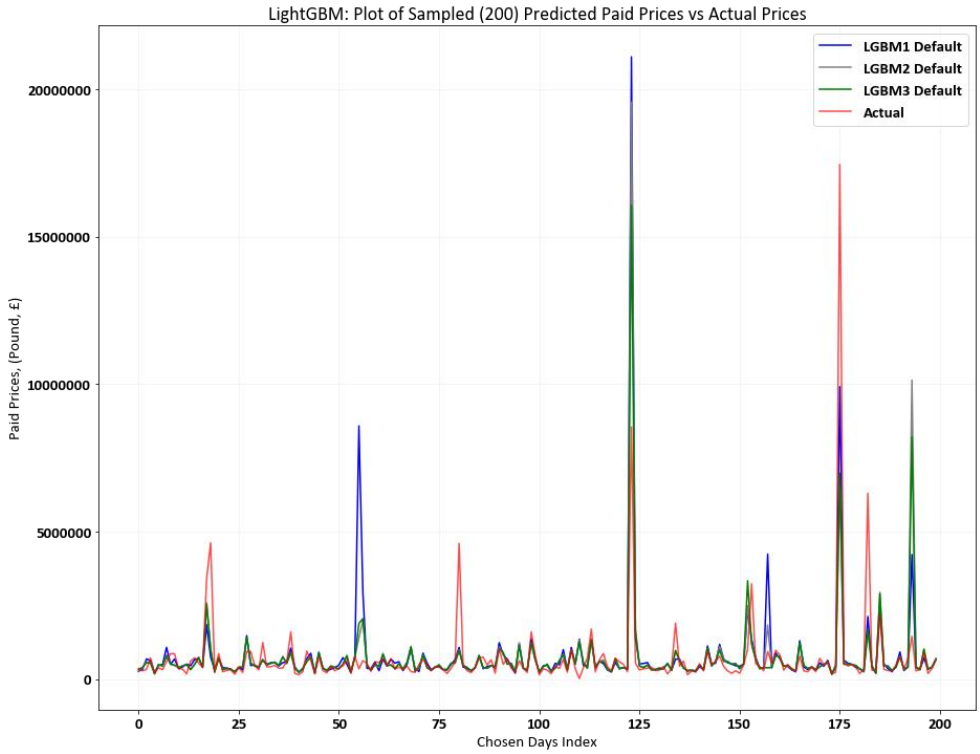


Figure 5.3: LightGBM (default) Tier 3 predicted paid prices versus actual prices

Figures 5.4 to 5.15 now show the plot of the results for four other models and the three cumulative tiers.

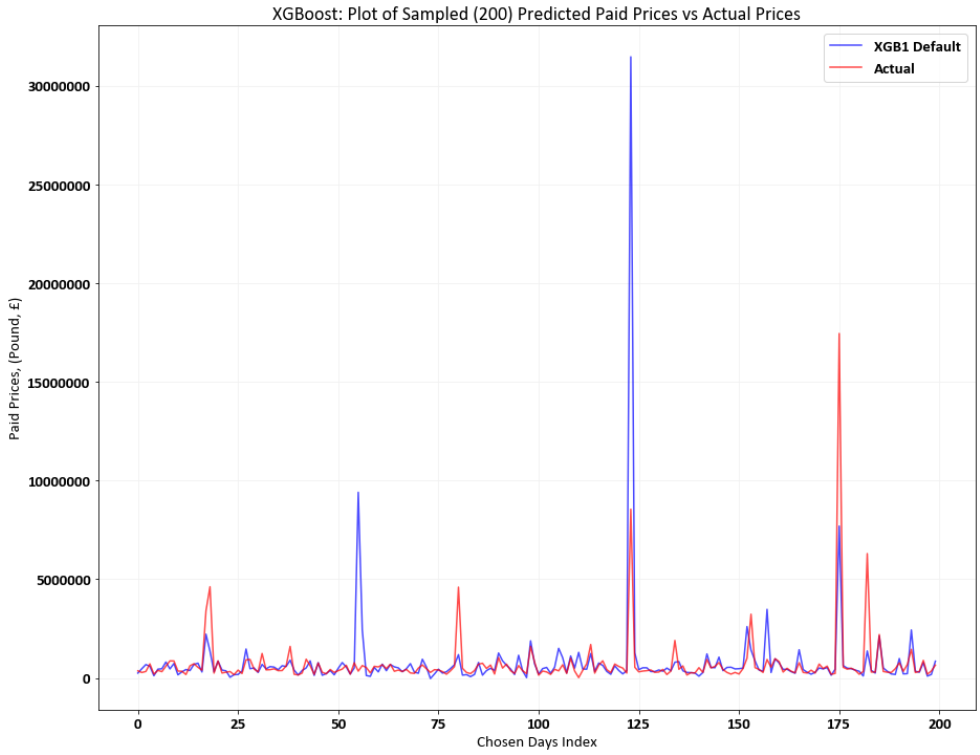


Figure 5.4: XGBoost (default) Tier 1 predicted paid prices versus actual prices

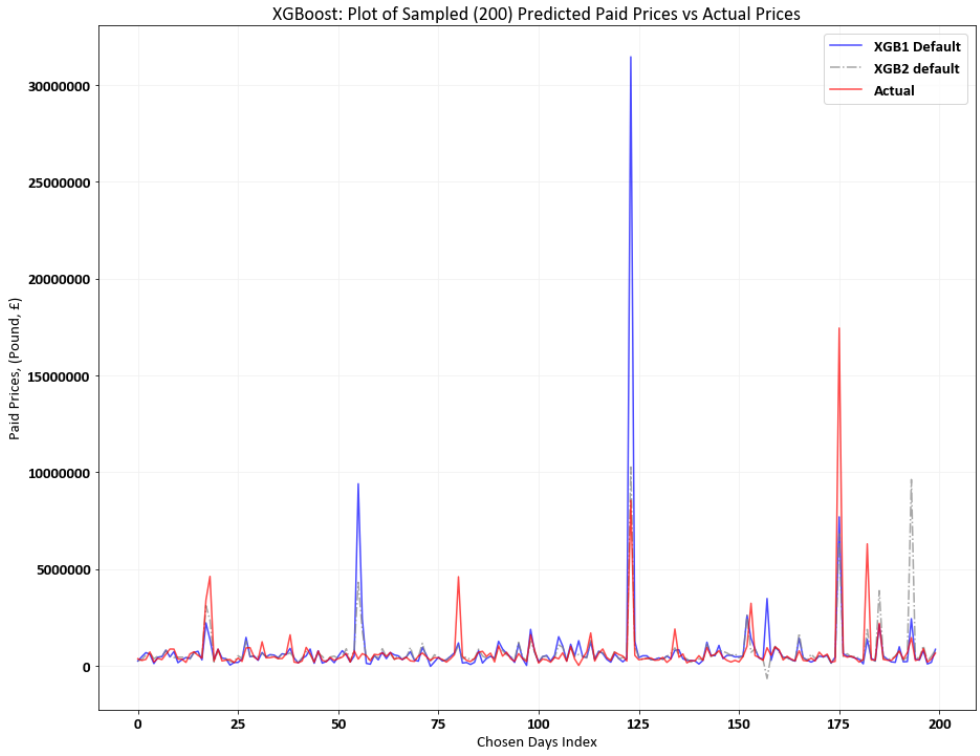


Figure 5.5: XGBoost (default) Tier 2 predicted paid prices versus actual prices

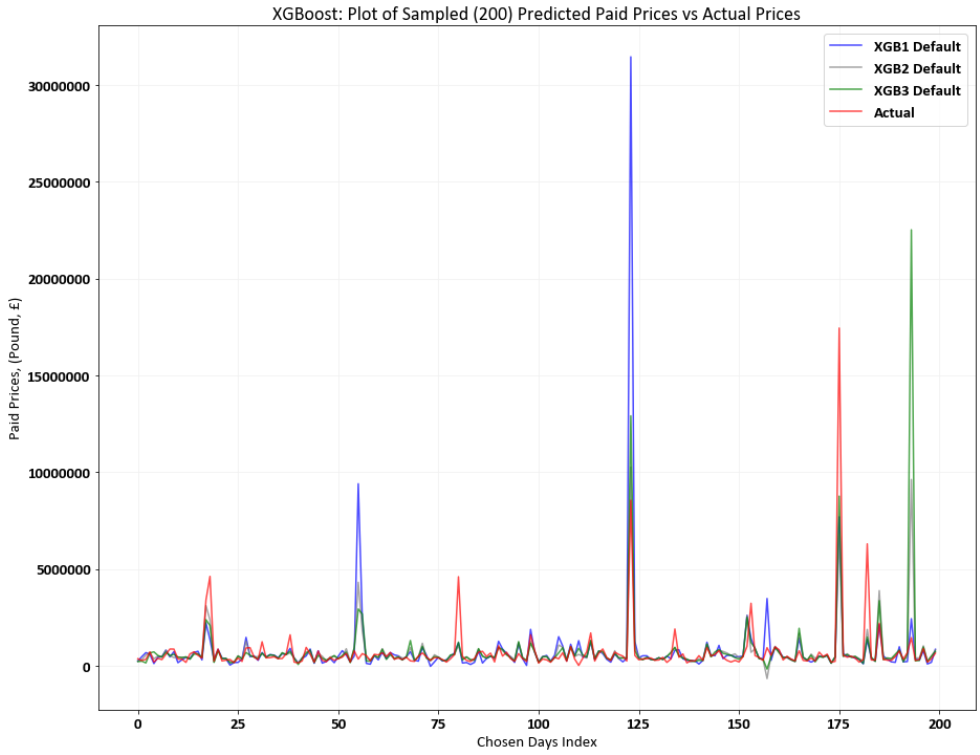


Figure 5.6: XGBoost (default) Tier 3 predicted paid prices versus actual prices

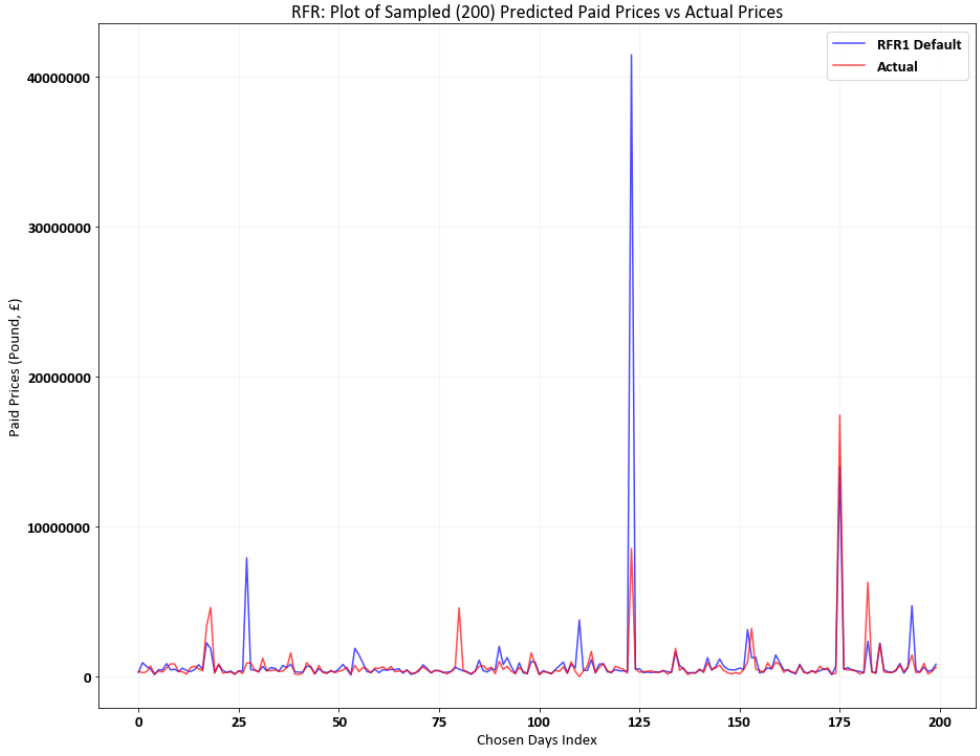


Figure 5.7: Random Forest (default) Tier 1 predicted paid prices versus actual prices

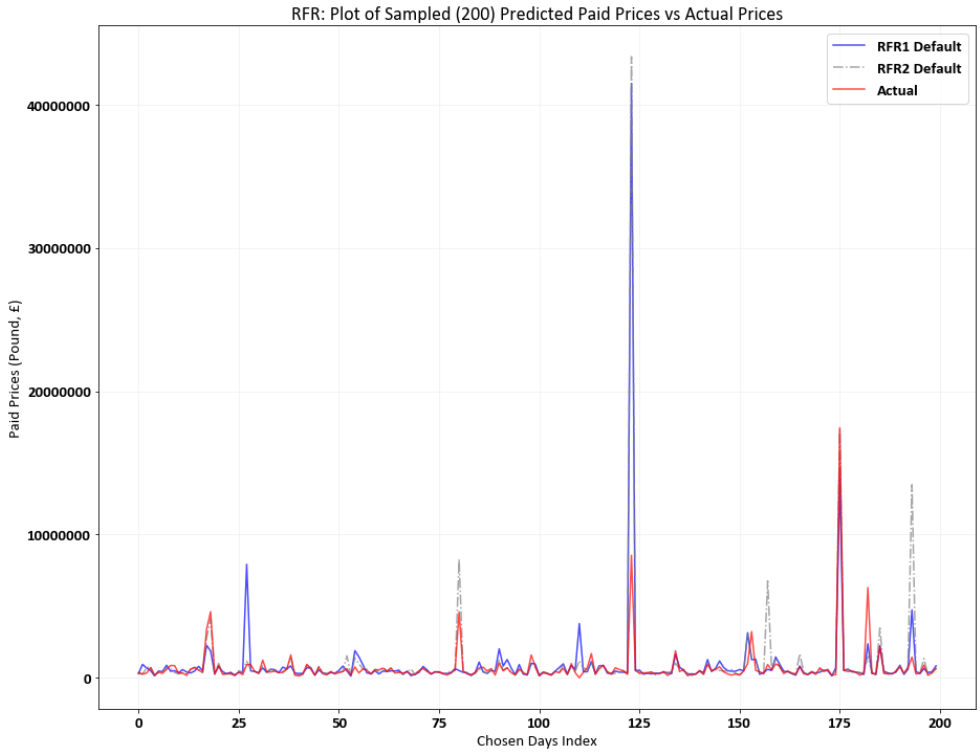


Figure 5.8: Random Forest (default) Tier 2 predicted paid prices versus actual prices

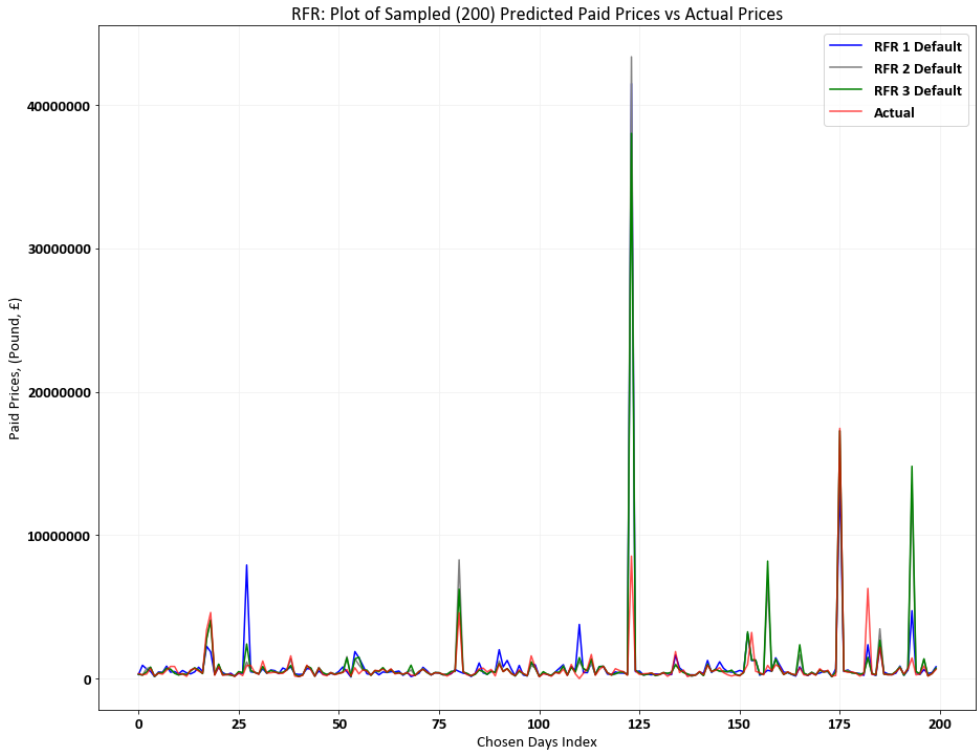


Figure 5.9: Random Forest (default) Tier 3 predicted paid prices versus actual prices

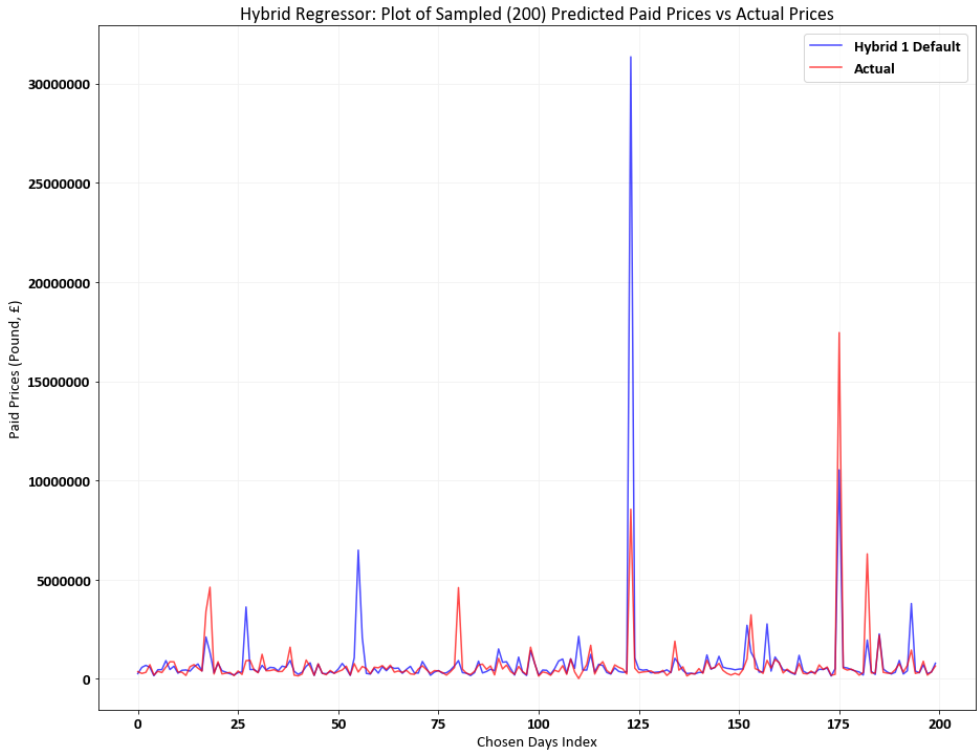


Figure 5.10: Hybrid Regression (default) Tier 1 predicted paid prices versus actual prices

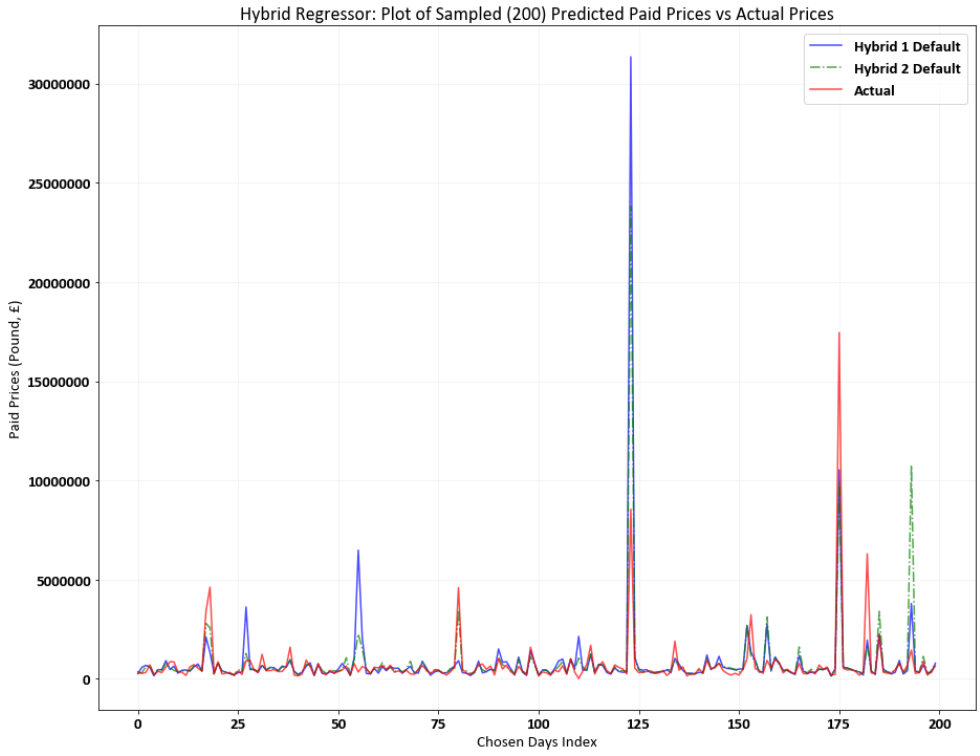


Figure 5.11: Hybrid Regression (default) Tier 2 predicted paid prices versus actual prices

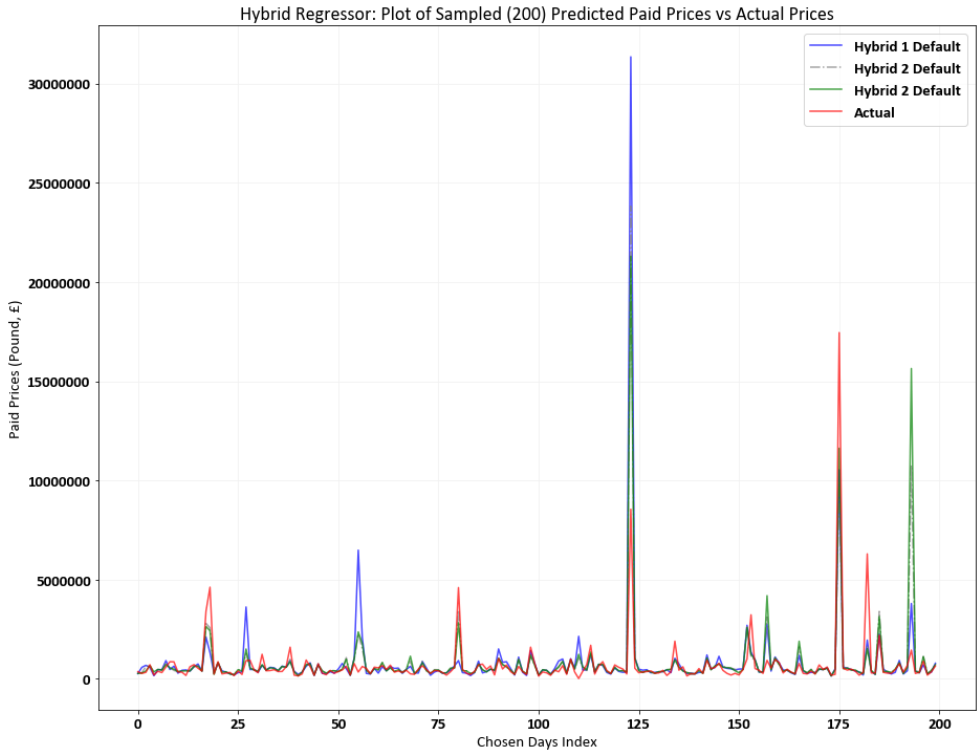


Figure 5.12: Hybrid Regression (default) Tier 3 predicted paid prices versus actual prices

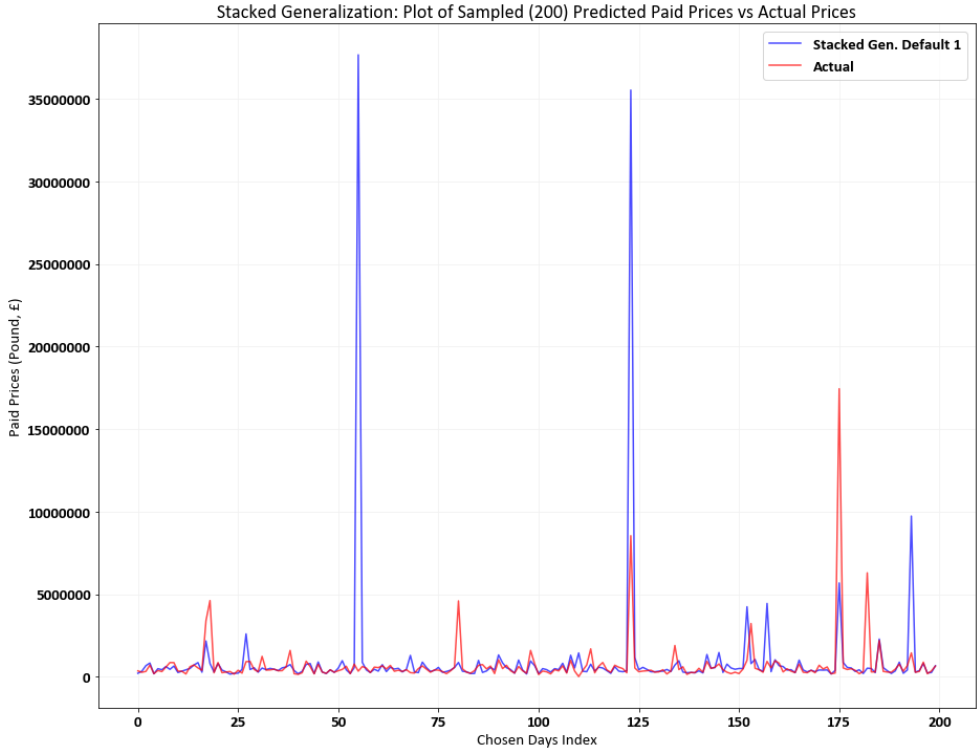


Figure 5.13: Stacked Generalisation (default) Tier 1 predicted paid prices versus actual prices

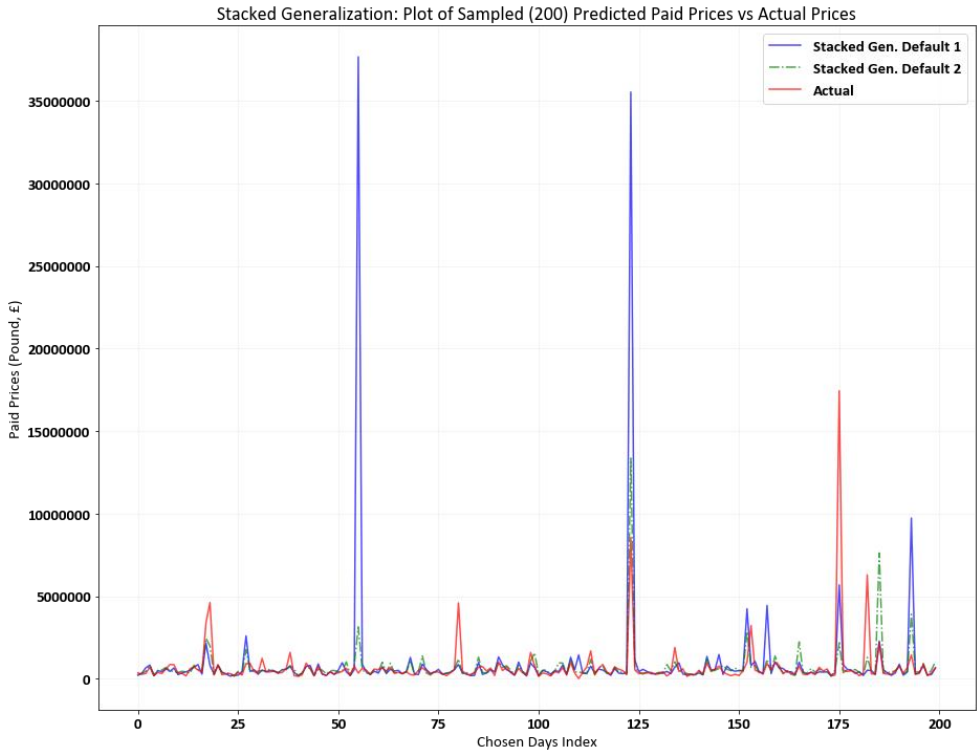


Figure 5.14: Stacked Generalisation (default) Tier 2 predicted paid prices versus actual prices

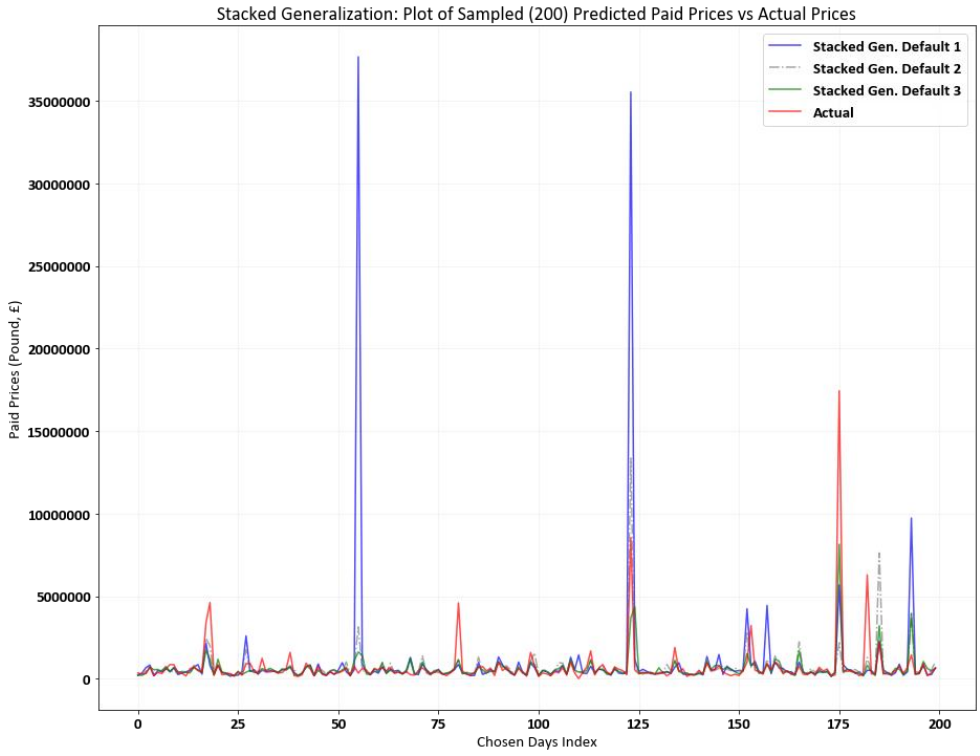


Figure 5.15: Stacked Generalisation (default) Tier 1 predicted paid prices versus actual prices

Root Mean Square Error (RMSE) is a common metric used to measure the error of a model predicting quantitative data (Moody, 2019). It estimates the standard

deviation of an observed value from the model prediction. According to (Moody, 2019), the observed value is equal to the sum of the predicted value and predictably distributed random noise with mean zero. If the noise is negligible as estimated by RMSE, the model is assessed as good at predicting the observed data. This means that in the cumulative Multi-feature House Price Estimation Framework, if RMSE is low for models in a tier or decreases when comparing model results across tiers, the model is assessed as good at predicting the observed data. However, when RMSE is large for models in a specific tier, or increases when comparing similar model results between tiers, it means that the model is not accounting for important features in the data.

Mean Absolute Error (MAE) is described as the mean of the absolute errors, as the name suggests. Therefore, it is the average of the absolute value of the difference between the forecasted value and the actual value. So, the question is what does MAE tell us in the context of machine learning models? It simply shows the size of the expected error from the predicted results (Watson, 2012). For a system like the cumulative MfHPE framework, with multiple levels of modelling, a low or reducing MAE shows that the predicted house prices are close to the actual house prices captured by HM Land Registry price paid data (Micalizzi, 2020).

The third metric explored for the evaluation of the thirty baseline models is R-squared. It is described by (Fernando et al., 2021) as a statistical metric that shows the proportion of the variance for a target variable that is explained by a range of independent variables in a regression model. A high R-squared value shows that predicted values are close to the actual values.

Tables 5.1, 5.2 and 5.3 show the results for all models using the default parameters for each algorithm stated in Section 4.11 on only Tier 1, Tier 2 and Tier 3 features. For Tier 1, LightGBM models were the best performing models when assessed using RMSE or R-squared, while Hybrid Regression models were the best performing model based on MAE scores for Tier 1 features. Furthermore, for Tier 2 Hybrid Regression models were best performing based on RMSE and R-squared scores, while Random Forest model was best performing for MAE. In conclusion, for Tier 3 Hybrid Regression models were the best performing based

on RMSE and R-squared scores, while Random Forest models were the best performing based on MAE score.

Table 0.1: Modelling results using default parameters (Tier 1)

	Model	RMSE	MAE	R-squared
Tier_1_Default	**LightGBM**	**3520071.29**	359972.67	**0.0878**
	Random Forest	3702246.49	370818.23	-0.0090
	XGBoost	3705299.63	389103.92	-0.0107
	Hybrid Regression	3552304.53	**347270.39**	0.0711
	Stacked Generalisation	3705613.80	401373.52	-0.0108

Table 0.2: Modelling results using default parameters (Tier 2)

	Model	RMSE	MAE	R-squared
Tier_2_Default	**LightGBM**	3477427.88	348603.42	0.1098
	Random Forest	3492294.16	**289294.10**	0.1022
	XGBoost	3705299.63	389103.92	-0.0107
	Hybrid Regression	**3416892.70**	309849.29	**0.1405**
	Stacked Generalisation	3631184.76	375124.47	0.0294

Table 0.3: Modelling results using default parameters (Tier 3)

	Model	RMSE	MAE	R-squared
Tier_3_Default	**LightGBM**	3479711.34	348552.42	0.1086
	Random Forest	3487379.03	**293799.64**	0.1047
	XGBoost	3507266.28	348531.83	0.0945
	Hybrid Regression	**3426453.28**	313405.37	**0.1357**
	Stacked Generalisation	3535413.97	373892.93	0.0799

Despite the performance of the models within the isolated tiers, the summary of results shown in Table 5.4 reveals the impact of the concept of cumulative multi-feature layering, which this thesis has introduced, as it highlights the improvement in model performance across tiers based on RMSE scores. A review of Figure 5.16 and Table 5.4 concurrently shows a reduction in RMSE scores for 80% of models in Tier 2 compared with Tier 1, with an overall improvement in the accuracy of the models due to the introduction of Tier 2 features. Furthermore, the introduction of Tier 3 features then shows a further reduction in RMSE scores for 60% of the modules compared with the improvements from Tier 2. This means a further improvement in the accuracy of 60% of the models because of the introduction of Tier 3 features. Therefore, from the perspective of cumulative

multi-feature layering, the Random Forest model is the overall best performing model for Tier 2, while XGBoost is the overall best performing model at Tier 3, based on RMSE scores.

Table 0.4: Impact of cumulative multi-feature layer – baseline models (RMSE)

	RMSE				
Model	**Tier_1_Default**	**Tier_2_Default**	**% Improvement**	**Tier_3_Default**	**% Improvement**
LightGBM	3520071.29	3477427.88	1.2%	3479711.34	-0.1%
Random Forest	3702246.49	3492294.16	**5.7%**	3487379.03	0.1%
XGBoost	3705299.63	3705299.63	0.0%	3507266.28	**5.3%**
Hybrid Regression	3552304.53	3416892.70	3.8%	3426453.28	-0.3%
Stacked Generalisation	3705613.80	3631184.76	2.0%	3535413.97	2.6%

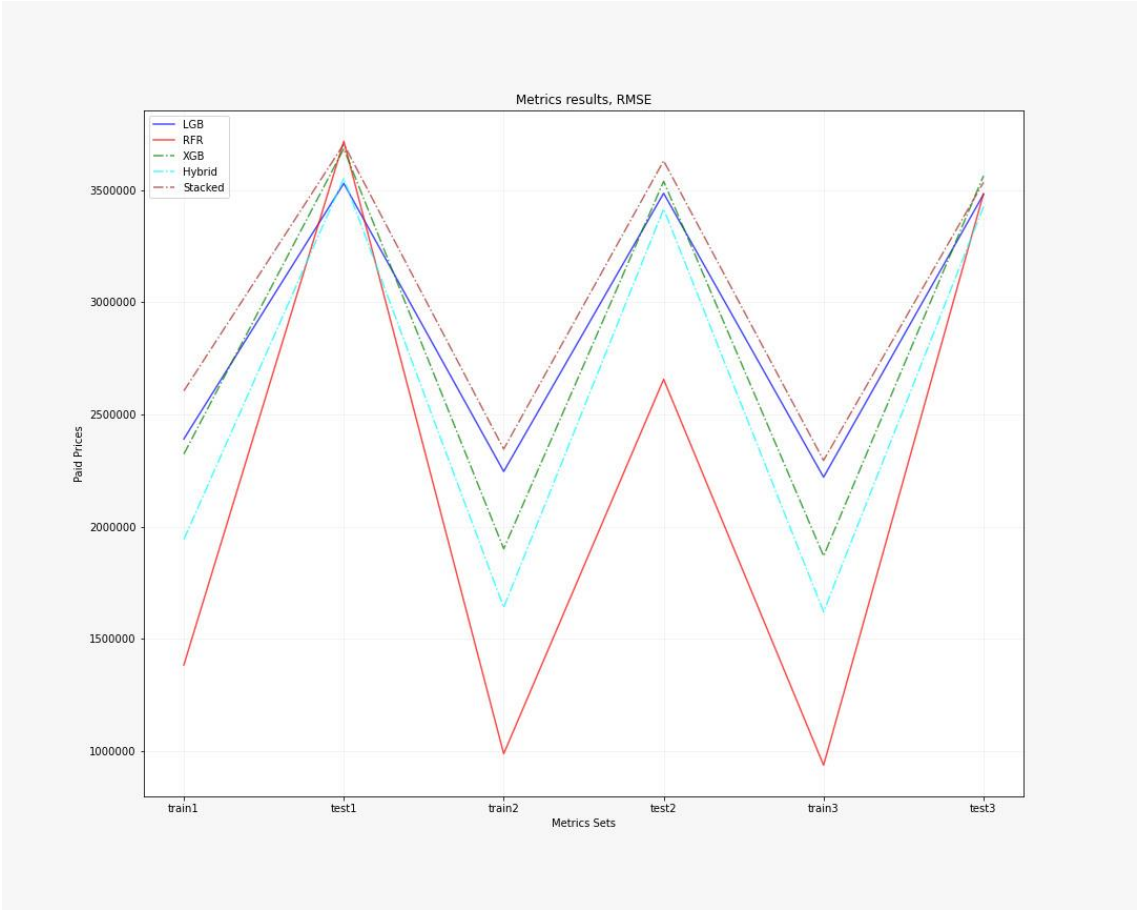


Figure 5.16: Impact of cumulative multi-feature layering – baseline models (RMSE)

As shown in Table 5.5 and Figure 5.17 concurrently, a reduction in MAE values was also recorded for 80% of the models in Tier 2 in comparison to Tier 1, showing improvements in the accuracy of the modules due to the cumulative layering of Tier 2 features. The cumulative layering of Tier 3 features resulted in

improved accuracy of 40% of the models, like the RMSE scores. Therefore, from the perspective of cumulative multi-feature layering, the Random Forest model is the overall best performing model for Tier 2, while XGBoost is the overall best performing model at Tier 3, based on RMSE scores.

Table 0.5: Impact of cumulative multi-feature layer – baseline models (MAE)

	MAE				
Model	**Tier_1_Default**	**Tier_2_Default**	**% Improvement**	**Tier_3_Default**	**% Improvement**
LightGBM	359972.67	348603.42	3.2%	348552.42	0.0%
Random Forest	370818.23	289294.10	**22.0%**	293799.64	-1.6%
XGBoost	389103.92	389103.92	0.0%	348531.83	**10.4%**
Hybrid Regression	347270.39	309849.29	10.8%	313405.37	-1.1%
Stacked Generalisation	401373.52	375124.47	6.5%	373892.93	0.3%

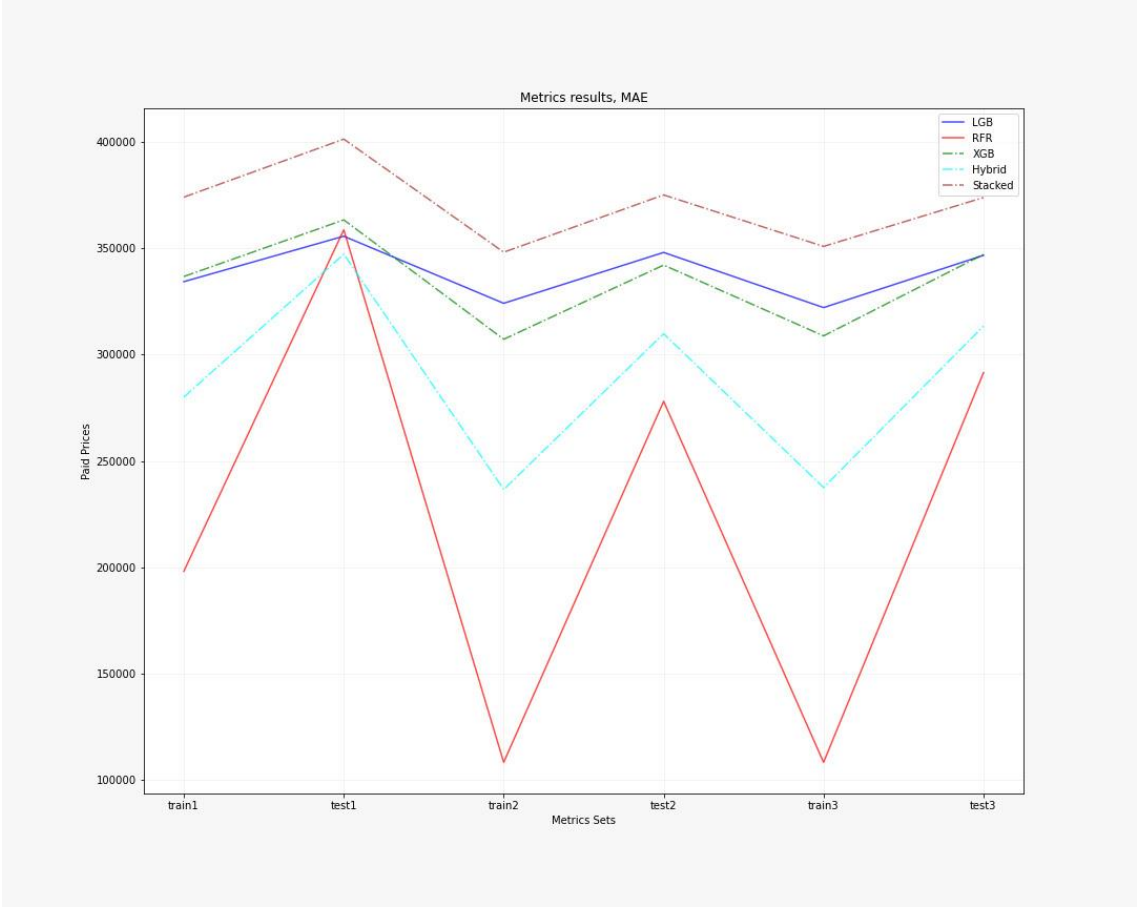


Figure 5.17: Impact of cumulative multi-feature layering – baseline models (MAE)

In conclusion, a concurrent view of Table 5.6 and Figure 5.18, being an output of this process-based, data driven and machine learning based house price estimation framework like the MfHPE, shows an improved performance for 40% of the models after the cumulative layering introducing Tier 2 features into the framework.

Table 0.6: Impact of cumulative multi-feature layer – baseline models (R-squared)

	R-squared				
Model	**Tier_1_Default**	**Tier_2_Default**	**% Improvement**	**Tier_3_Default**	**% Improvement**
LightGBM	0.09	0.11	-25.0%	0.11	1.1%
Random Forest	-0.01	0.10	**1234.4%**	0.10	-2.5%
XGBoost	-0.01	-0.01	0.0%	0.09	**985.2%**
Hybrid Regression	0.07	0.14	-97.8%	0.14	3.4%
Stacked Generalisation	-0.01	0.03	370.7%	0.08	-172.1%

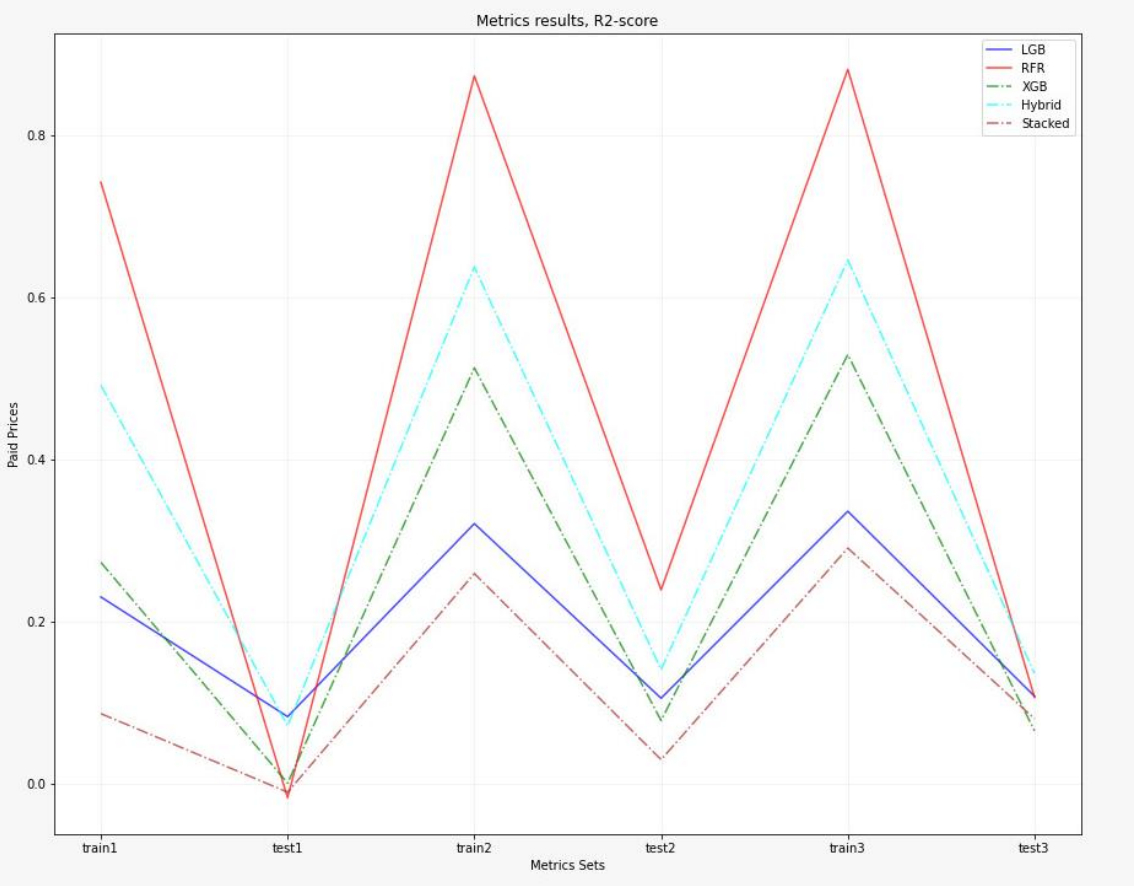


Figure 5.18: Impact of cumulative multi-feature layering – baseline models (R-squared)

The cumulative introduction of Tier 3 features produced an improved performance on 60% of the models. Overall, the baseline model results established that the cumulative introduction of features through the tiers of the MfHPE framework leads to improved model accuracy. However, it also shows a change in the best performing algorithm as the multiple features are layered through the framework. Random Forest was the best performing algorithm for both Tier 1 and Tier 2, while XGBoost was the best performing at Tier 3. This also establishes the fact the performance of machine learning algorithms has a dependence on the data ingested into the ML system.

5.3 Model optimisation

Model optimisation in machine learning is, without doubt, one of the most challenging aspects of the implementation of ML solutions. There is immense attention given to deep learning theories and machine learning to achieve the optimisation of models. There are two types of algorithm parameters usually considered when building machine learning systems: (i) model or default parameters, which possess the ability to be initiated and consequently updated through data learning; and (ii) hyperparameters, the parameters which are used to configure a machine learning model and to specify the algorithm which is used in minimising the loss function.

The MfHPE has leveraged Bayesian Optimisation as the choice approach for the optimisation of models. As stated above, and like the discussions on the baseline modelling in Section 5.2, the results for the eighteen optimisation models created are presented in this section in Tables 5.8 to 5.13. Furthermore, Figures 5.19 to 5.24 show how the optimised models for Tier 1 to Tier 3 for both LightGBM and XGBoost compared against the corresponding default models. The accuracy of the predicted house prices of the optimised models are observed to be an improvement over most of the predicted house prices of the default models. The predicted house prices for each tier of the Random Forest models are also compared in Figures 5.25 to 5.27, and it was observed that the overall accuracy of the predicted house prices improved as new tiers of features were introduced.

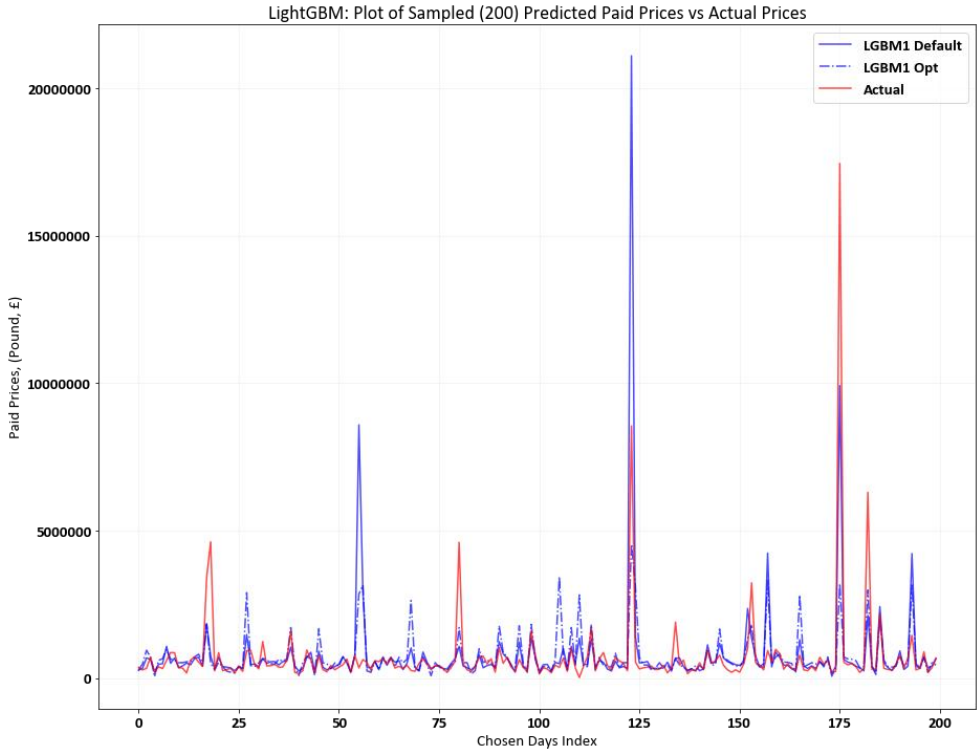


Figure 5.19: LightGBM Tier 1 – actual prices versus default prediction versus optimised prediction

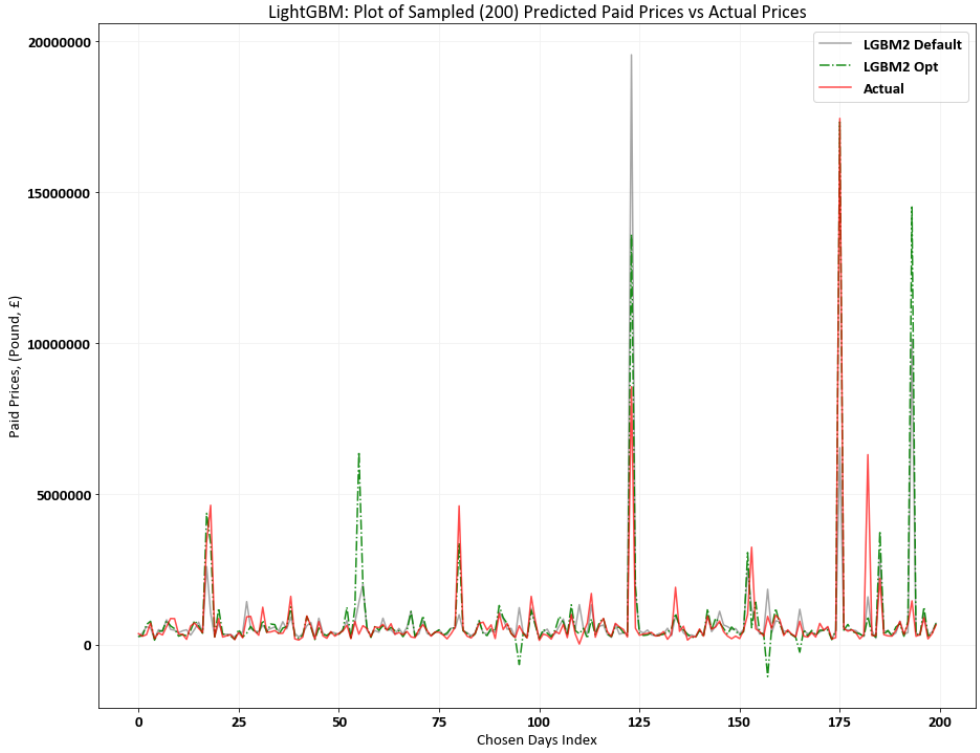


Figure 5.20: LightGBM Tier 2 – actual prices versus default prediction versus optimised prediction

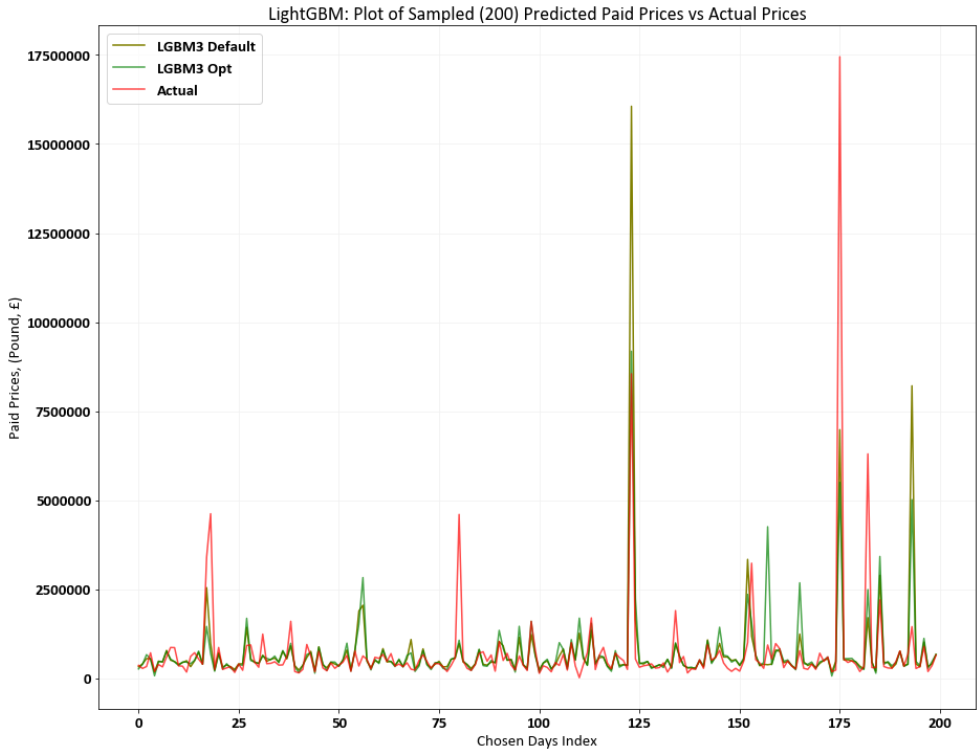


Figure 5.21: LightGBM Tier 3 – actual prices versus default prediction versus optimised prediction

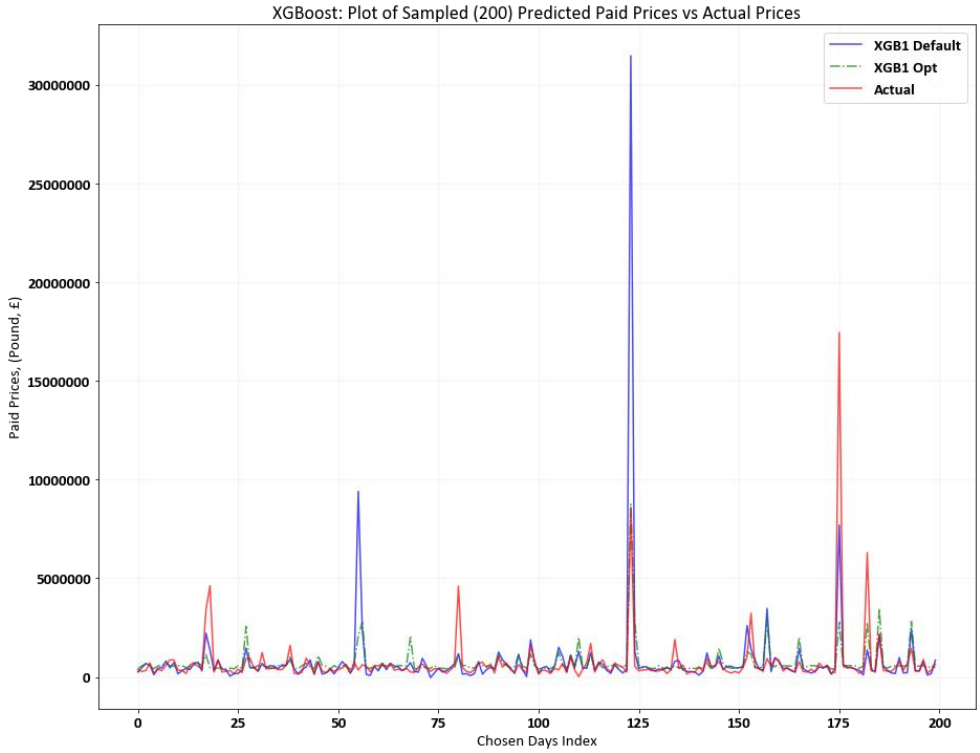


Figure 5.22: XGBoost Tier 1 – actual prices versus default prediction versus optimised prediction

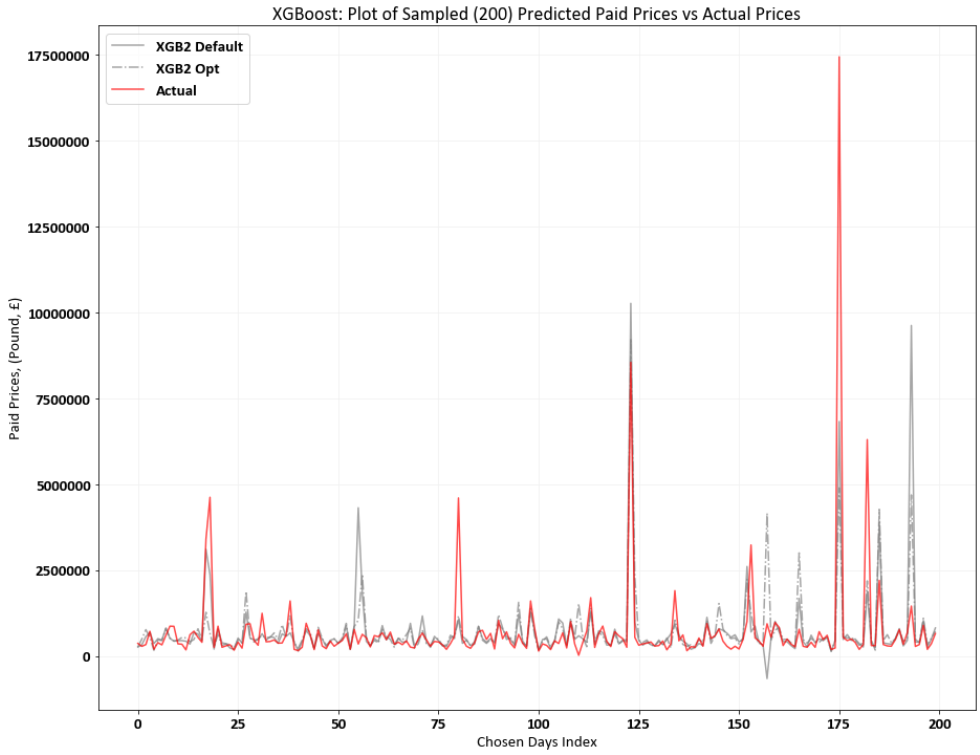


Figure 5.23: XGBoost Tier 2 – actual prices versus default prediction versus optimised prediction

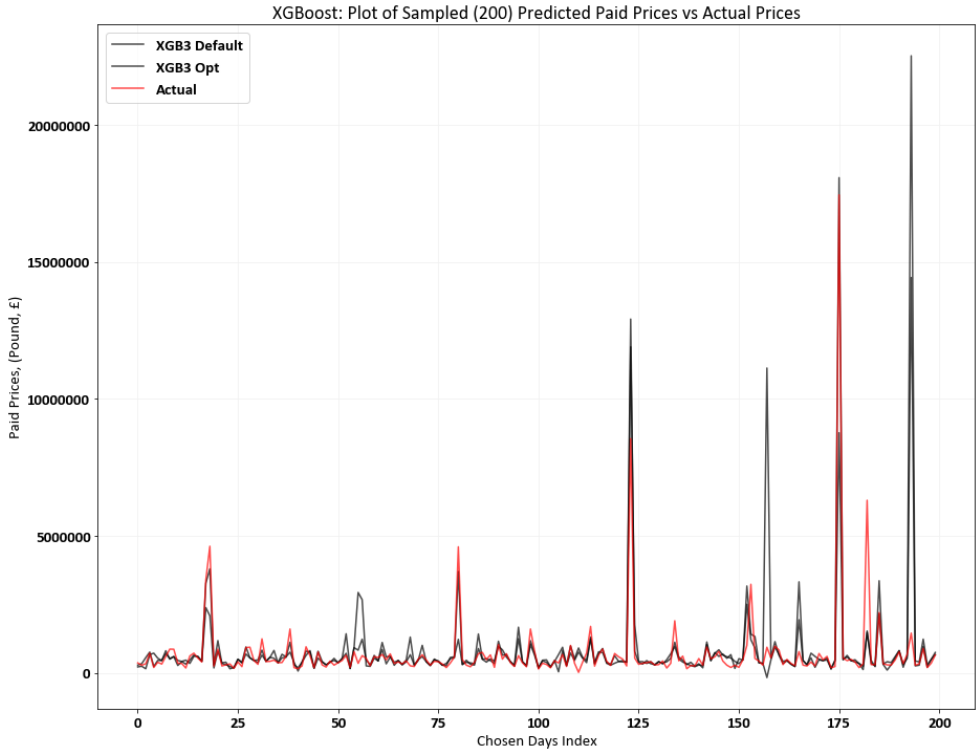


Figure 5.24: XGBoost Tier 3 – actual prices versus default prediction versus optimised prediction

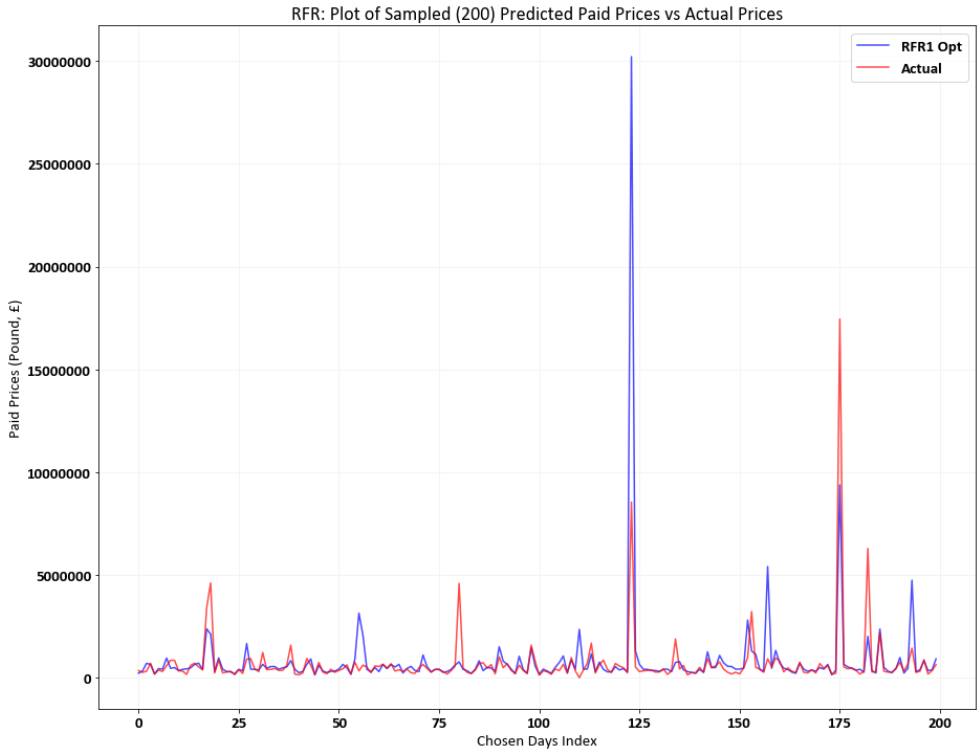


Figure 5.25: Random Forest Tier 1 – actual prices versus optimised prediction

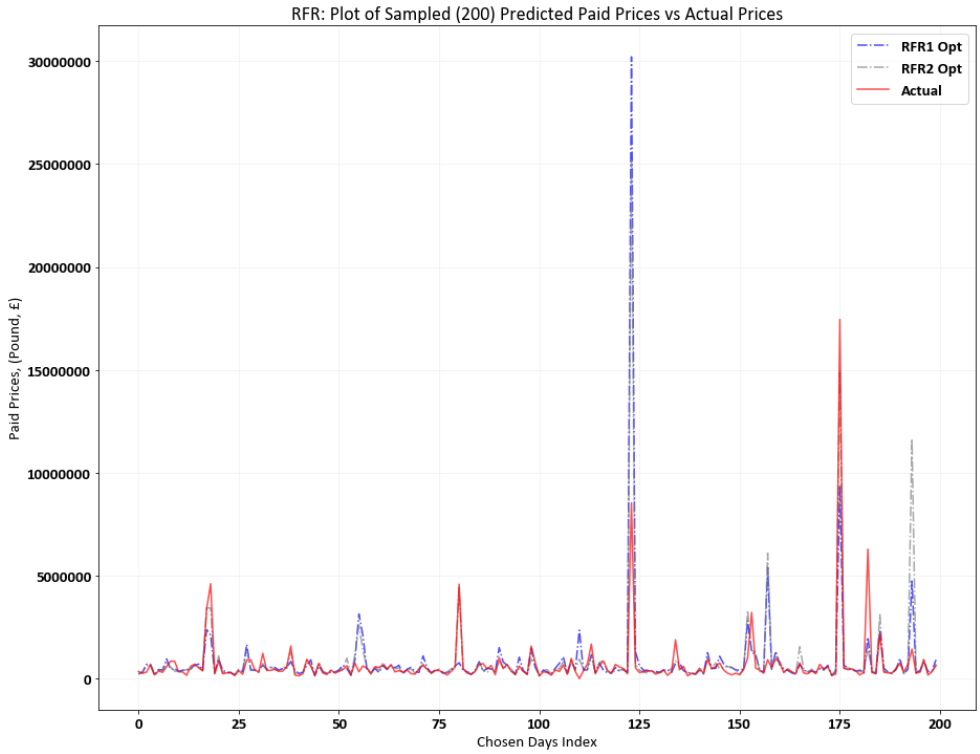


Figure 5.26: Random Forest Tier 2 optimised versus Tier 1 optimised prediction

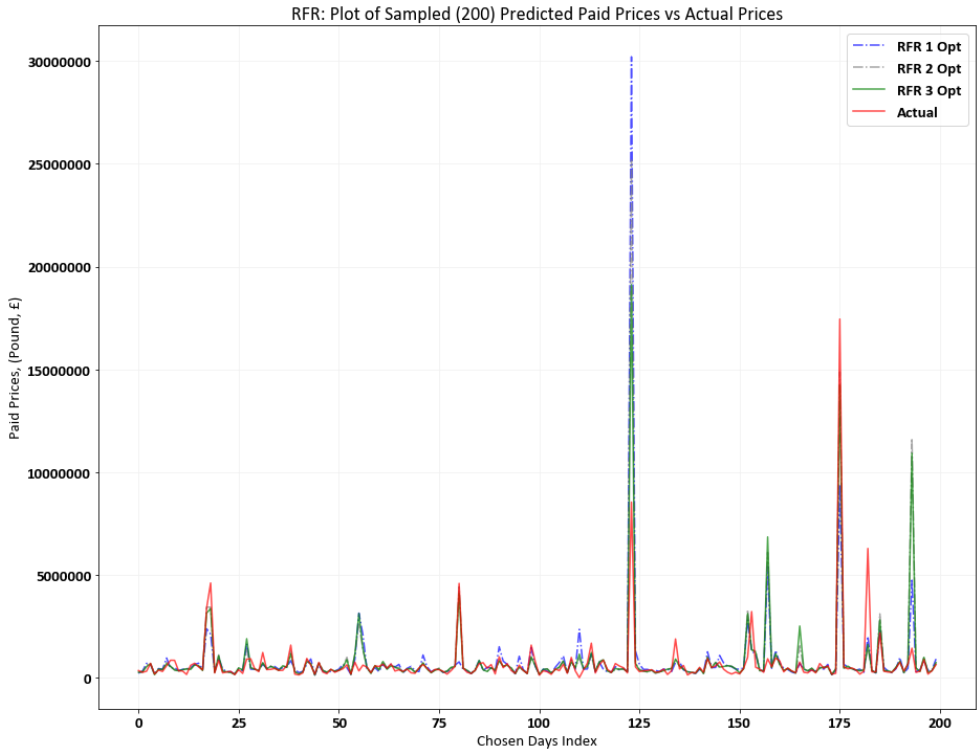


Figure 5.27: Random Forest Tier 3 optimised versus Tier 2 optimised versus Tier 1 optimised prediction

Tables 5.7, 5.8 and 5.9 show the results for models using the optimised parameters for each algorithm stated in Section 4.11 on Tier 1, Tier 2 and Tier 3 features. Across all tiers and all metrics, the Random Forest models are found to be the best performing models based on RMSE, MAE and R-squared scores.

Table 0.7: Modelling results using optimised parameters (Tier 1)

	Model	RMSE	MAE	R-squared
Tier_1_Optimised	**LightGBM**	3668646.78	472872.95	0.0092
	Random Forest	**3523322.99**	**358772.42**	**0.0862**
	XGBoost	3563082.33	425546.49	0.0654

Table 0.8: Modelling results using optimised parameters (Tier 2)

	Model	RMSE	MAE	R-squared
Tier_2_Optimised	**LightGBM**	3611587.57	415669.82	0.0398
	Random Forest	**3415804.71**	**299770.67**	**0.1411**
	XGBoost	3714580.93	555323.00	-0.0157

Table 0.9: Modelling results using optimised parameters (Tier 3)

	Model	RMSE	MAE	R-squared
Tier_3_Optimised	**LightGBM**	3659978.45	460920.87	0.0139
	Random Forest	**3410039.05**	**304249.60**	**0.1440**
	XGBoost	3537771.68	327800.72	0.0787

As discussed earlier, RMSE estimates the standard deviation of an observed value from the model prediction. According to (Moody, 2019), the observed value is equal to the sum of the predicted value and predictably distributed random noise with mean zero. If the noise is negligible, as estimated by RMSE, the model is assessed as good at predicting the observed data. This means that in the cumulative Multi-feature House Price Estimation Framework, if RMSE is low for models in a tier, or decreases when comparing model results across tiers, the model is assessed as good at predicting the observed data. However, when RMSE is large for models in a specific tier, or increases when comparing similar model results between tiers, it means that the model is not accounting for important features in the data.

Despite the performance of the models within the isolated tiers, the summary of optimisation results shown in Tables 5.10 to 5.12 reveals the impact of the concept cumulative multi-feature layering which this thesis has introduced, as it highlights the improvement in model performance across tiers based on RMSE scores.

A concurrent view of Figure 5.28 and Table 5.10 shows a reduction in RMSE scores for 66.6% of the models in Tier 2 compared with Tier 1, being an overall improvement in the accuracy of the models as a result of the introduction of Tier 2 features. Furthermore, the introduction of Tier 3 features then also shows a reduction in RMSE scores for 66.6% of the modules compared with the improvements from Tier 2. This means a further improvement in the accuracy of 60% of the models because of the introduction of Tier 3 features. The results also show that although the Random Forest model recorded the highest performance improvement with the cumulative introduction of Tier 2 features, the XGBoost model recorded the highest performance improvement at the cumulative introduction of Tier 3 features.

Table 0.10: Impact of cumulative multi-feature layer – optimised models (RMSE)

	RMSE				
Model	**Tier_1_Optimised**	**Tier_2_Optimised**	**% Improvement**	**Tier_3_Optimised**	**% Improvement**
LightGBM	3668646.78	3611587.57	1.6%	3659978.45	-1.3%
Random Forest	3523322.99	3415804.71	**3.1%**	3410039.05	0.2%
XGBoost	3563082.33	3714580.93	-4.3%	3537771.68	**4.8%**

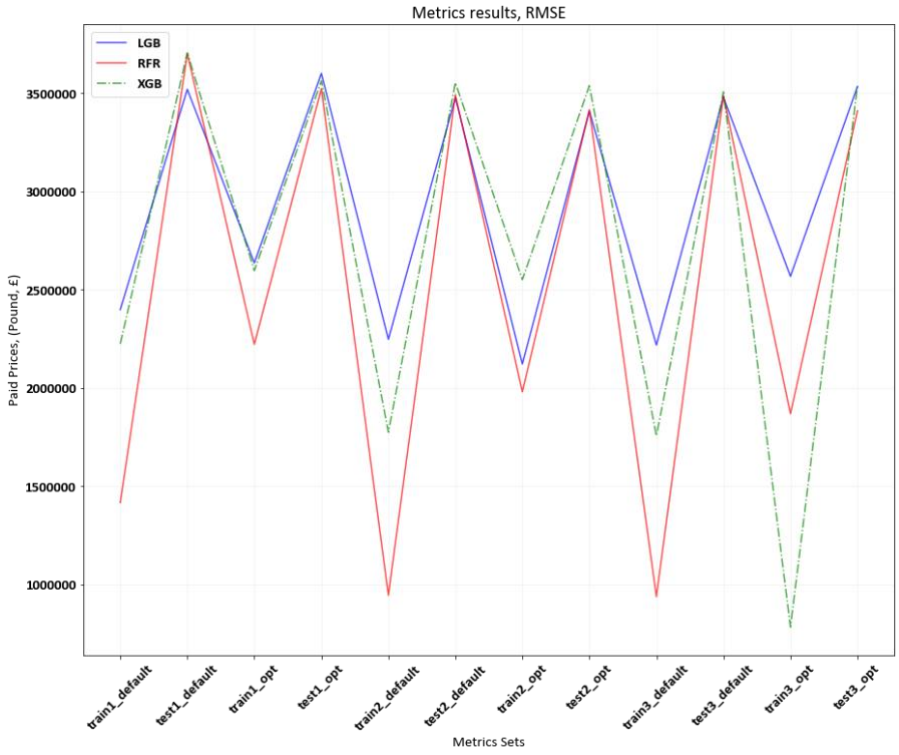


Figure 5.28: Root Mean Square Error – full results | all tiers | three algorithms

MAE is the average of the absolute value of the difference between the forecasted value and the actual value. In the context of machine learning models, it shows the size of the expected error from the predicted results (Watson, 2012). For a system like the cumulative MfHPE framework, with multiple levels of modelling, a reduction or lower MAE shows that the predicted house prices are close to the actual house prices captured by HM Land Registry price paid data (Micalizzi, 2020). As shown in Table 5.11 and Figure 5.29 concurrently, a reduction in MAE values was also recorded for 66.6% of the models in Tier 2 in comparison to Tier 1, showing improvements in the accuracy of the modules as

a result of the cumulative layering of Tier 2 features. The cumulative layering of Tier 3 features resulted in improved accuracy of the XGBoost model by 41%. These results show that although the Random Forest model recorded the highest performance improvement with the cumulative introduction of Tier 2 features, the XGBoost model recorded the highest performance improvement at the cumulative introduction of Tier 3 features.

Table 0.11: Impact of cumulative multi-feature layer – optimised models (MAE)

	MAE				
Model	**Tier_1_Optimised**	**Tier_2_Optimised**	**% Improvement**	**Tier_3_Optimised**	**% Improvement**
LightGBM	472872.95	415669.82	12.1%	460920.87	-10.9%
Random Forest	358772.42	299770.67	**16.4%**	304249.60	-1.5%
XGBoost	425546.49	555323.00	-30.5%	327800.72	**41.0%**

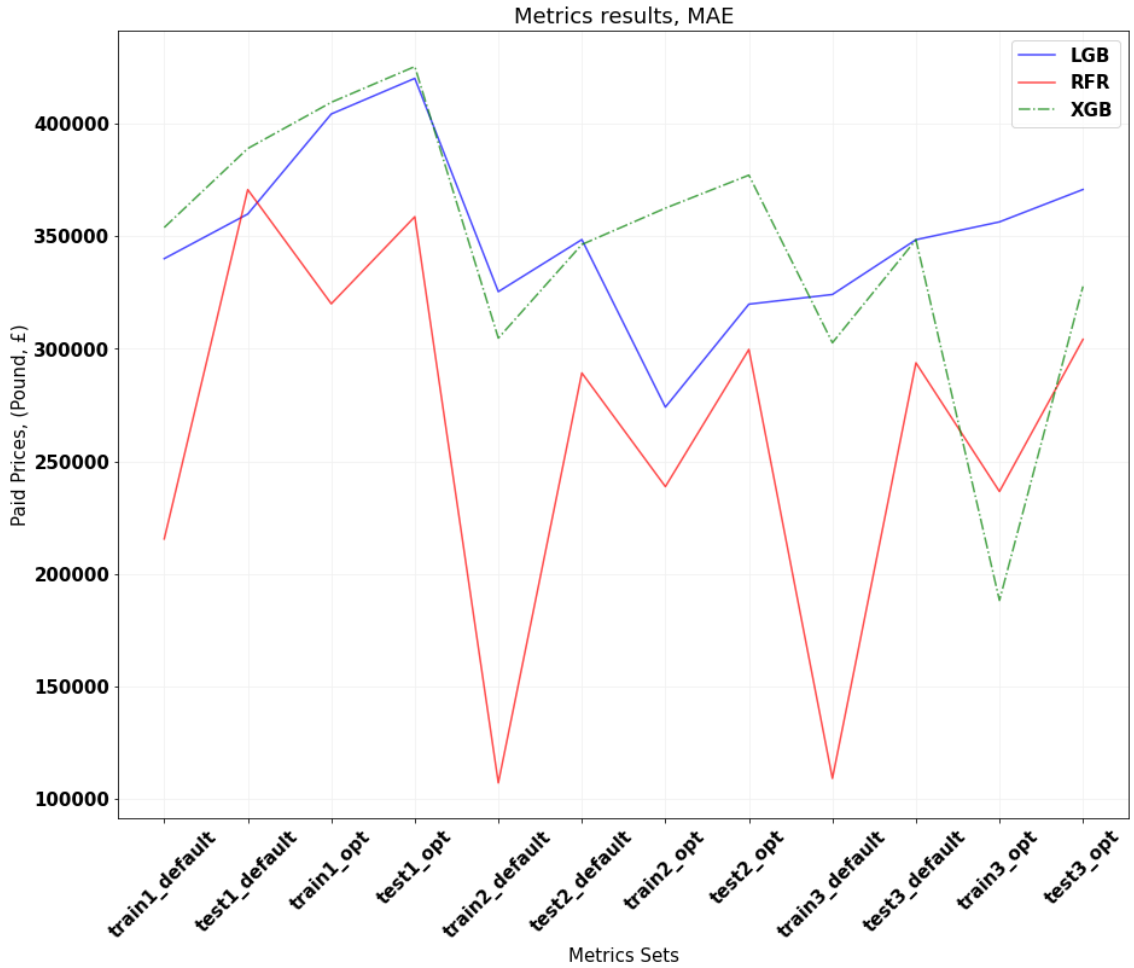


Figure 5.29: Mean Absolute Error – full results | all tiers | three algorithms

R-squared is described by (Fernando et al., 2021) as a statistical metric that shows the proportion of the variance for a target variable that is explained by a range of independent variables in a regression model. A high R-squared value shows that predicted values are close to the actual values. Therefore, in a process-based data driven machine learning based house price estimation framework like the MfHPE, Table 5.12 and Figure 5.30 concurrently show an improved performance for the XGBoost model after the cumulative layering introducing Tier 2 features into the framework. The cumulative introduction of Tier 3 features produced further improved performance.

Table 0.12: Impact of cumulative multi-feature layer – optimised models (R-squared)

	R-squared				
Model	**Tier_1_Optimised**	**Tier_2_Optimised**	**% Improvement**	**Tier_3_Optimised**	**% Improvement**
LightGBM	0.01	0.04	-331.6%	0.01	65.1%
Random Forest	0.09	0.14	-63.7%	0.14	-2.1%
XGBoost	0.07	-0.02	**124.1%**	0.08	**599.6%**

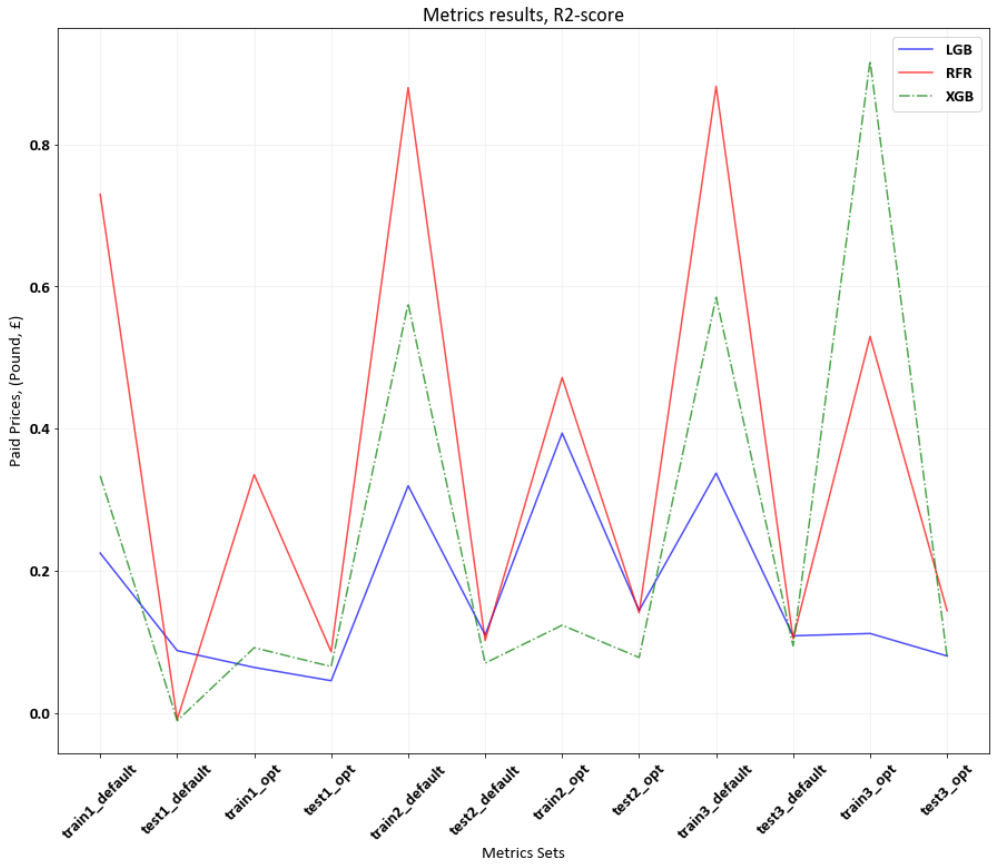


Figure 5.30: R-squared – full results | all tiers | three algorithms

Overall, the baseline model results established that the cumulative introduction of features through the tiers of the MfHPE framework leads to improved model accuracy. However, it also shows a change in the best performing algorithm as the multiple features are layered through the framework. Random Forest was the best performing algorithm for both Tier 1 and Tier 2, while XGBoost was the best performing at Tier 3. This also establishes that the performance of machine learning algorithms has a dependence on the variety of features ingested into the ML system.

5.4 Conclusion

The results from thirty baseline models and eighteen optimised models created to estimate or predict house prices have been discussed extensively in this chapter. Overall, these results show both an improvement and a decline in model performance, largely driven by the change in the variety of features.

Both the baseline and optimised modelling results established that the cumulative introduction of features through the tiers of the MfHPE framework leads to improved model accuracy. However, it also shows a change in the best performing algorithm as the multiple features are layered through the framework. The Random Forest model recorded the highest performance improvement with the cumulative introduction of Tier 2 features, while the XGBoost model recorded the highest performance improvement at the cumulative introduction of Tier 3 features. This also establishes that the performance of machine learning algorithms has a dependence on the variety of features ingested into the ML system. These improvements in the performance of the models are consistent for RMSE and MAE, but different for R-squared, thereby further emphasising one of the contributions to knowledge of this research, which states that evaluation metrics respond differently to features and machine learning algorithms.

Chapter 6: Discussion and Interpretation

6.1 Introduction

After instantiating the MfHPE Framework in Chapter 5, the focus of this chapter is to presents an overarching evaluation with advanced analytics and also to provide answers to the research questions, with a view to proving or disproving the research hypothesis. The structure of this chapter is as follows. First, a confirmation of the final model that was selected for the cumulative MfHPE framework based on the features exploited. It is important to note that with a change in the variety of the features, there could also be a change in the best performing algorithm and consequently, the model. Second, the provisioning of an overview on model explainability and then how it is applied in explaining the price prediction or estimation provided by the selected model. Third, the validation of model results by using the unseen data for a year that was not used in the modelling exercise at all.

6.2 Framework model selection

As observed in the model results shown in Chapter 5, although Random Forest was the best performing model for Tier 1 and Tier 2, XGBoost was the best performing model at Tier 3. Whilst model explainability can be implemented for all the models in the framework, as a result of computational limitations the focus will be on Tier 3, being the highest tier of the cumulative MfHPE framework at this time. Therefore, XGBoost becomes the model of choice, and its explainability will now be discussed further.

Whilst existing studies have provided insights on how different ML algorithms perform, how algorithms compare with one another, and what factors may have a positive, negative or neutral impact on the estimation of house prices. The challenge is that in a real-life scenario and based on geographic location, all or some of the various parameters used in machine learning models exist in a static or changing continuum. Parameters like house physical properties or features

and neighbourhood amenities mostly exist in a static continuum, while features like economic metrics vary every month or quarter in a changing continuum. Therefore, it becomes imperative that all known factors or parameters are considered through the creation of a framework in which multiple new or existing machine algorithms can be exploited on the various parameters that actually co-exist in real life for the estimation of house prices. Therefore, the MfHPE framework exploits features within both the static and changing continuum but in three tiers so as to gain an understanding of the impact of different groups of features.

6.3 Framework model explainability

The cumulative MfHPE framework mainly explores the global explainability approach, although a presentation of a local or individual record will be made for comparison. Also, in this thesis model explainability is based only on the XGBoost-based model, being the best performing model at the highest level of this cumulative MfHPE framework. The SHapley Additive exPlanations (SHAP) was proposed by (Lundberg and Lee, 2017) to provide an approach to understanding why machine learning models make certain predictions. The explainability of machine learning models is important because the reasons behind the predictions they make is as crucial as the accuracy of the models. Figure 6.1 shows the top 10 features that have an impact on Tier 1 estimation, with property type (other) contributing the most to the model. Other features in the top 5 include: (i) '*City of Westminster*', being a district with consistently the highest value for transactions year-on-year between 2011 and 2020 (Appendix 3); (ii) the '*Freehold*' status of houses also had a significant impact, meaning buyers are keener to own both property and land; (iii) the district '*Kensington and Chelsea*' must have also had a significant impact because, according to Appendix 3, the district was consistently in the top 3 of districts with the highest total value of transactions year-on-year. However, it is interesting to observe that the London Borough of Wandsworth, another contender among the districts, did not make enough impact to feature in the top 10 even though London Borough of Camden

and City of London did. Figure 6.2, the SHAP value plot, shows the relationship each of the top 10 features have with the target feature, price. This measures the feature importance, impact, original value and correlation. Table 6.1 is a summary of the SHAP explainability for the XGBoost model output based on Tier 1 features.

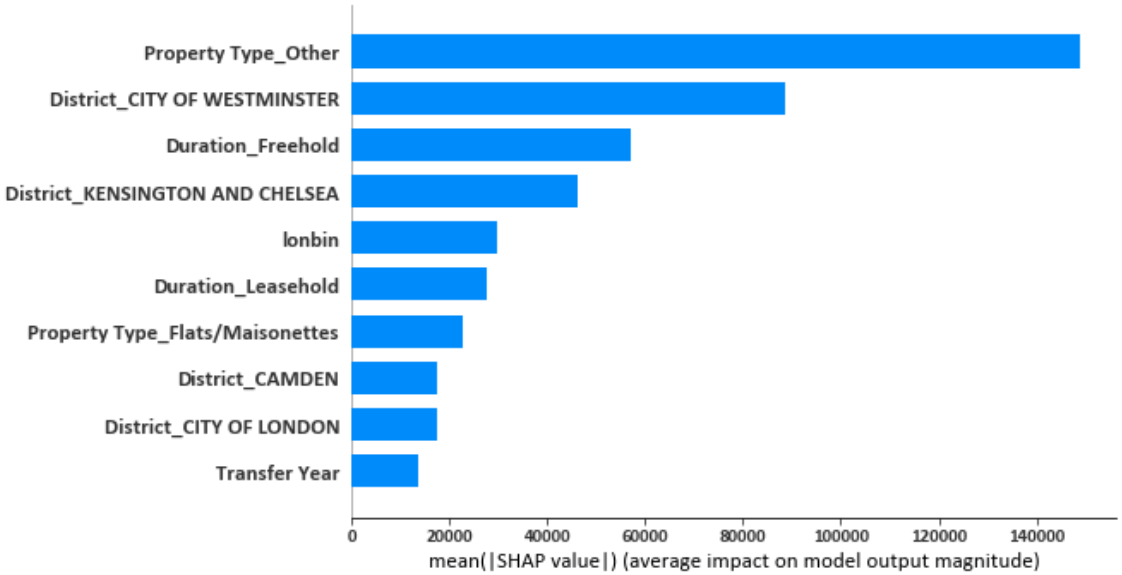


Figure 6.1: Feature importance – top ten features with an average impact on Tier 1 XGBoost model (note that 'lonbin' is longitude)

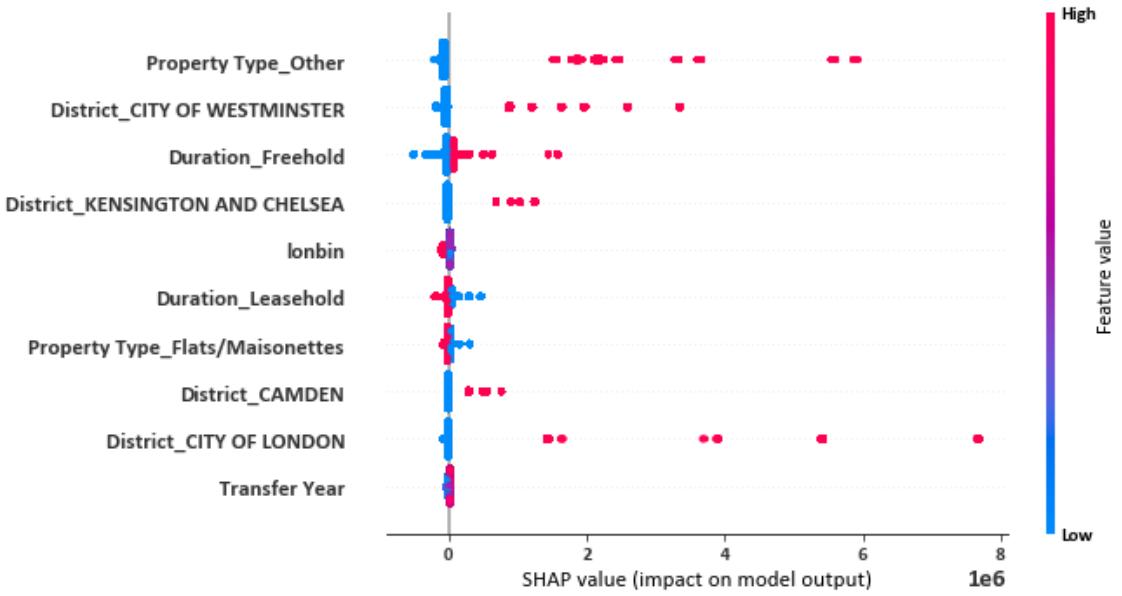


Figure 6.2: SHAP value – top ten features with LOW - HIGH impact on Tier 1 XGBoost model output (note that 'lonbin' is longitude)

Table 0.1: Summary SHAP explainability for Tier 1 XGBoost model output (note that 'lonbin' is longitude)

Feature	Importance	Impact	Correlation
Property Type_Other	High	Higher	Positive
District_City of Westminster	High	Higher	Positive
Duration_Freehold	High	Lower	Positive
District_Kensington and Chelsea	High	Lower	Positive
Lonbin	Low	Lower	Negative
Duration_Leasehold	Low	Lower	Negative
Property Type_Flats/Maisonettes	Low	Lower	
District_Camden	Low	Lower	Positive
District_City of London	Low	Higher	Positive
Transfer Year	Low	Lower	

After the cumulative introduction of neighbourhood features (Tier 2), Figure 6.3 shows the Tier 2 features ranked based on their importance in descending order. The 'Freehold' status of houses now has the most importance in the prediction, but this is closely followed by '*Property Type_Other*'. The SHAP value in Figure 6.4 shows that '*Freehold*' has a high and positive impact on house price prediction by this model. However, although the shortest distance between a house and a bus stop has a high feature importance, it has a low but positive impact on house price.

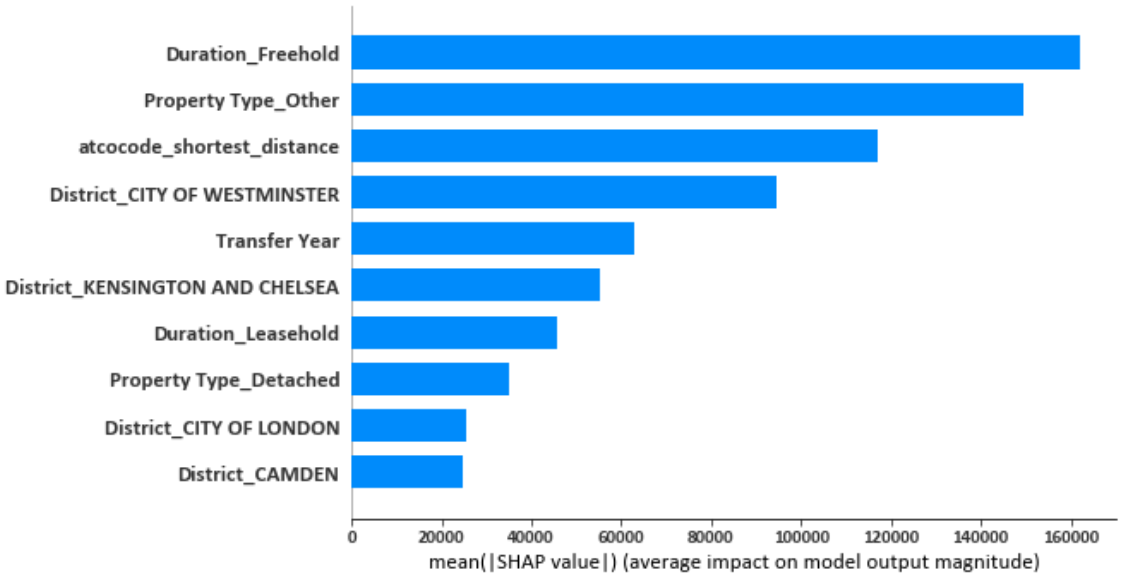


Figure 6.3: Feature importance – top ten features with an average impact on Tier 2 XGBoost model (note that atcocode is the code for bus stops, therefore 'atcocode_shortest_distance' is the shortest distance from a bus stop)

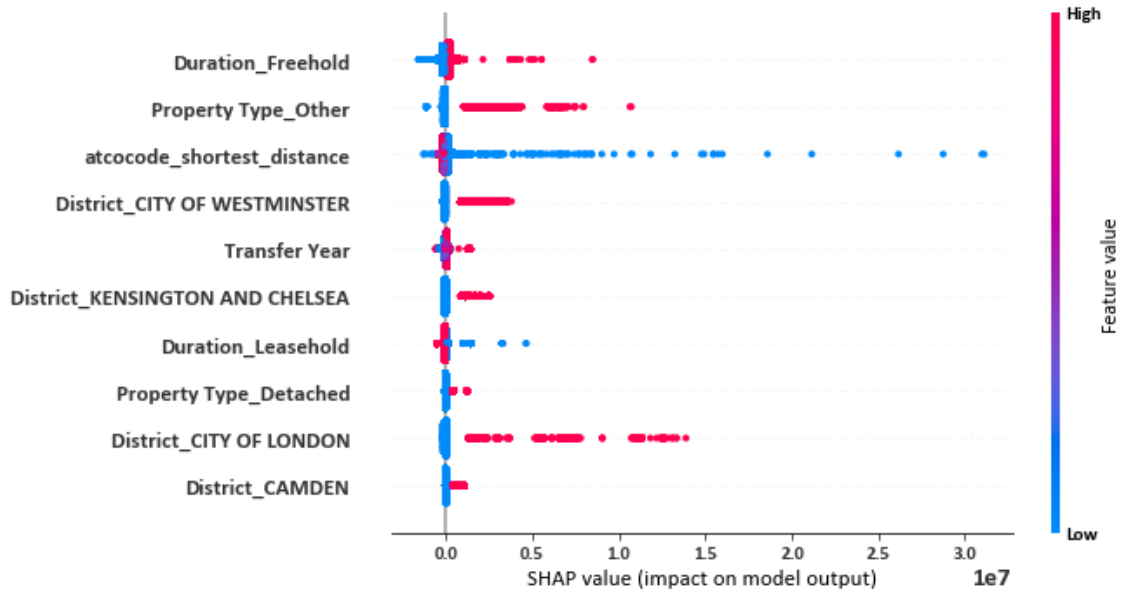


Figure 6.4: SHAP value of top ten features with LOW - HIGH impact on Tier 2 XGBoost model output (note that atcocode is the code for bus stops, therefore 'atcocode_shortest_distance' is the shortest distance from a bus stop)

The feature '*atcocode_shortest_distance*' is also observed to be the only Tier 2 feature in the top 10, whilst other Tier 2 features have potentially strengthened the importance of Tier 1 features like '*Duration_Freehold*', '*Transfer Year*' and '*Property Type_Detached*'.

In the model output for Tier 3, as shown in Figures 6.5 and 6.6, it is observed that more Tier 2 features have now made it to the top 10.

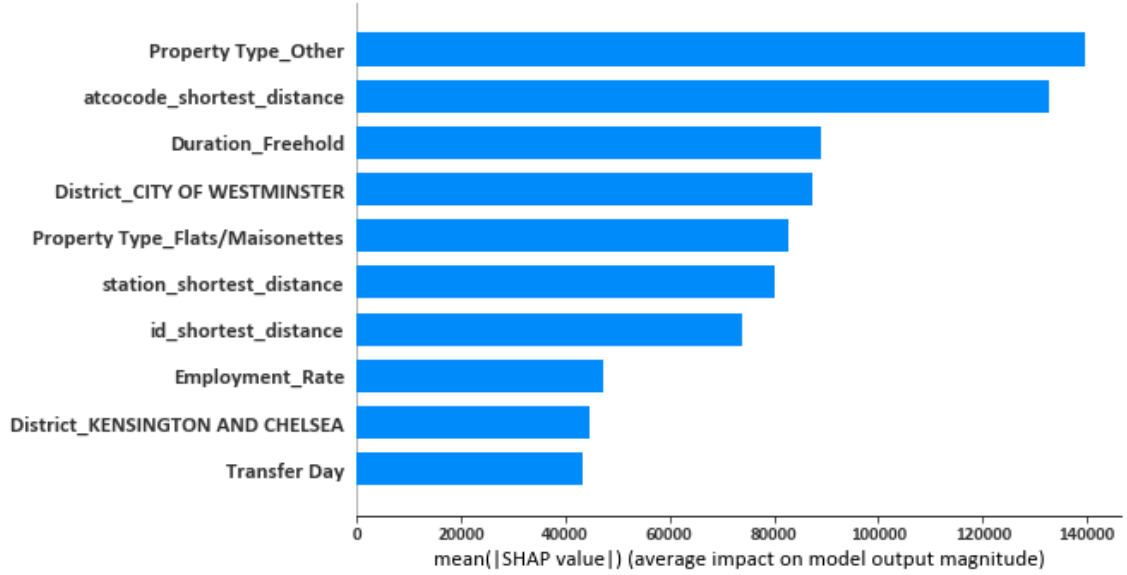


Figure 6.5: Feature importance – top ten features with an average impact on Tier 3 XGBoost model

As Tier 3 means the introduction of macroeconomic indicators, the most important of the newly introduced features is '*Employment_Rate*'.

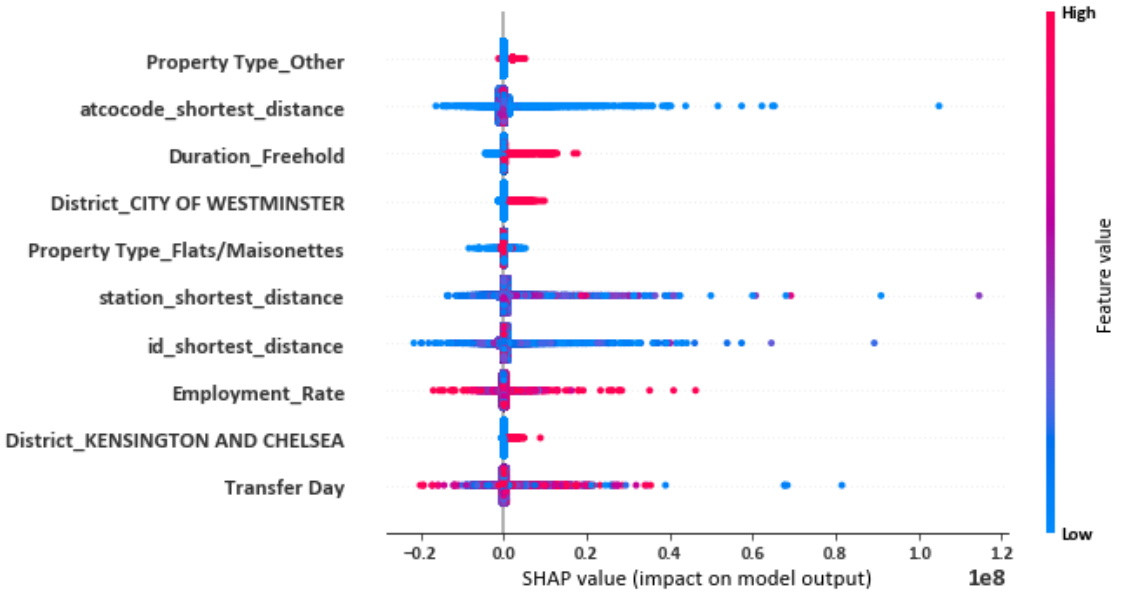


Figure 6.6: SHAP value of top ten features with LOW - HIGH impact on Tier 3 XGBoost model output

However, '*Property Type_Other*' is observed to be the feature with the most importance, as seen in Tier 1. Some of the interesting revelations about the features driving the model output for Tier 3 include the fact that although the shortest distance from a bus stop has a high feature importance, it is less positively correlated to the target variable compared with '*Duration_Freehold*'.

explainability

As stated above the of model outputs can be presented globally or locally. The global explainability provides an explanation for the features that made an impact on the overall prediction, while the local explainability provides an explanation for the features that influenced the prediction for a single record in the dataset, i.e. a single house price. Figures 6.7 to 6.9 show the local SHAP explainability for a single record. The house is an old flat or maisonette located in the London Borough of Hackney with a last transaction date of 16th April 2018 and exchanged for £437,000.00. As observed in Figure 6.7, the estimated house price for a house with such a profile at the time the model was run on 15th December 2021 is £445,202.49. The local explainability plot also flags the features of high impact.

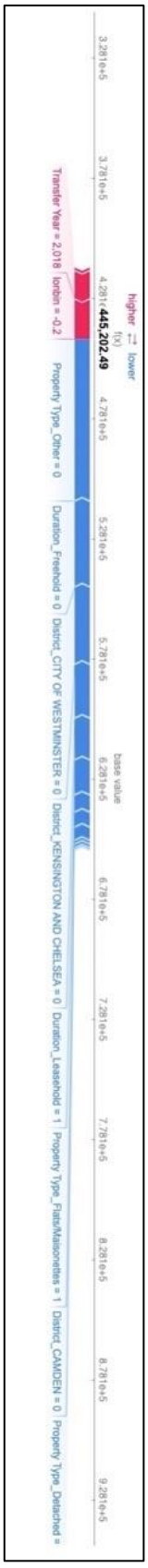


Figure 6.7: Local SHAP explainability for a single record – Tier 1

The Tier 2 local explainability plot for the same property in Figure 6.8 shows an estimated price of £540,274.38, and the features representing location and distance of bus stops had the highest impact.

Figure 6.8: Local SHAP explainability for a single record – Tier 2

Finally, the Tier 3 local explainability plot in Figure 6.9 shows macroeconomic features like employment rate and consumer price index making a high impact on the estimated or predicted house price alongside the neighbourhood features.

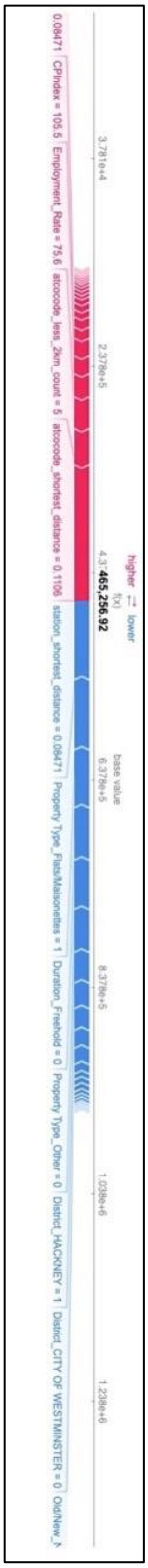


Figure 6.9: Local SHAP explainability for a single record – Tier 3

The estimated house price at Tier 3 is £465,256.92, which is a 6% appreciation in value of the property in 32 months despite the impact of Covid-19.

6.4 Framework model validation

This thesis has exploited HM Land Registry Price Paid Data as the primary dataset showing house characteristics and the price paid. A total of 48 models were created using transactions between 1st January 2011 and 31st December 2020 for training, testing and optimisation. The validation of the framework was then initiated with 100% unseen data comprising 2021 transactions. The validation focuses on the performance of the framework model across a range of segments of the Multi-feature House Price Estimation framework. The segments include performance based on: (i) different house price bands; (ii) the unique dynamics of the 31 London boroughs including the City of Westminster and City of London; (iii) property type; (iv) age of property; (v) duration/tenure; (vi) transfer month; (vii) transfer quarters; (viii) inflation rate; (ix) employment rate; and (x) proximity to public transportation.

The validation data is comprised of eighty-four thousand and fifty-one (84,051) records and seventy-two (72) features. Two evaluation metrics are in focus for the framework validation and these are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). RMSE and MAE are just metrics that are used for the evaluation of regression models by showing deviation from actual. They are both an average of errors, being the difference between predicted and actual values (Acharya, 2021).

6.4.1 Framework validation based on house price bands

The validation data was split into twelve bands based on house price as shown in Table 6.2. The results show each of these bands being evaluated across all tiers using both RMSE and MAE. In Tier 1, and evaluating with RMSE, the model shows better estimation for houses with price ranging between £200,000 and £1M than houses worth over a million or under £200,000, but performs best for the estimation of house prices with value between £400,000 and £500,000, with

a relatively good performance score for houses with value up to £800,000, which aligns with the band for an average house in London. According to (Rightmove, 2022), the average price for flats in London was £528,621, terraced houses £742,938, and semi-detached houses £725,384. The introduction of neighbourhood features, Tier 2, did not seem to make a significant impact on the price bands, for which the model has performed better. However, in this cumulative Multi-feature House Price Estimation Framework, the introduction of macroeconomic indicators, Tier 3, shows the model performing best for houses within the price range of £400,000 to £500,000.

Table 0.2:Results for framework validation based on house price bands

			Tier 1		Tier 2		Tier 3	
	No of Records	**%**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Over 5M	873	1.040	37.609	15.321	37.266	14.665	34.823	13.451
1M - 5M	9747	11.600	1.284	0.974	1.357	0.818	2.879	1.044
900K - 1M	2403	2.860	0.823	0.483	1.264	0.452	2.226	0.567
800K - 900K	3251	3.870	0.616	0.384	0.745	0.368	1.502	0.454
700K - 800K	4792	5.700	0.498	0.273	0.601	0.291	1.147	0.354
600K - 700K	6868	8.170	0.434	0.191	0.533	0.236	0.977	0.285
500K - 600K	10649	12.670	0.372	**0.131**	0.443	0.169	1.402	0.257
400K - 500K	16216	19.290	**0.365**	0.133	**0.426**	**0.156**	**0.701**	**0.195**
300K - 400K	16108	19.160	0.407	0.203	0.434	0.175	0.724	0.211
200K - 300K	8752	10.410	0.564	0.294	0.602	0.214	1.104	0.267
100K - 200K	2730	3.250	1.107	0.607	1.172	0.530	2.247	0.713
Under 100K	1662	1.980	3.088	2.666	3.466	2.710	6.497	3.562

It is important to note that there is a slight variation in the evaluation results produced by MAE compared with RMSE. In Tier 1, and evaluating with MAE, the model shows best performance for houses with price ranging between £300,000 and £400,000, alongside the best performing price bands based on RMSE. Both RMSE and MAE were relatively on par for Tier 2, although MAE also produced good performance scores for houses valued between £100,000 and £200,000 and between £900,000 and £5M, but in Tier 3, RMASE and MAE showed the best performance for houses with price range between £400,000 and £500,000. This means that the product owners of data science and machine learning based projects can be flexible in their choice of evaluation metrics, depending on the

machine learning algorithm, variety of features and the expected goal of the project. Figure 6.10 is a plot showing the alignment of the predicted price and the actual price. The actual price is the transaction price for each house as captured in the HM Land Registry Price Paid Data, while the predicted price is the price based on prediction results by the XGBoost model on Tier 3 features.

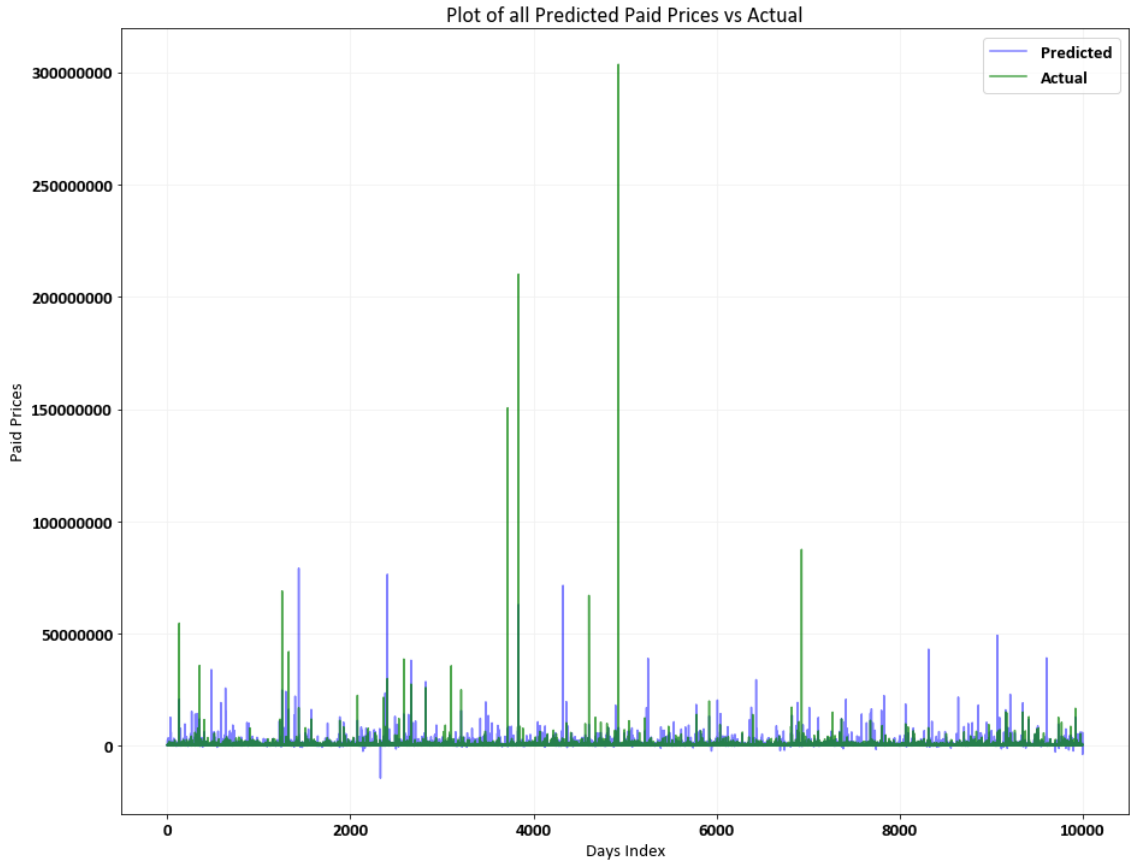


Figure 6.10: Alignment plot of predicted and actual price

6.4.2 Framework validation based on London boroughs

The lower the RMSE and MAE value, the better. Based on the validation data, the RMSE and MAE evaluation in Table 6.3 shows that the model estimation is more accurate for house price estimation in the London Borough of Havering than it is for City of London and City of Westminster. It also shows that within the context of the current state of the Multi-feature House Price Estimation (MfHPE) framework, the model performs better for the London boroughs of Bromley, Redbridge and Lewisham that it does for the London Borough of Bexley when

evaluated using RMSE. However, when evaluation is based on MAE, the model is observed to perform better for the London Borough of Bexley than it does for the London boroughs of Bromley, Redbridge and Lewisham. This further reinforces the argument that the choice evaluation metrics should be driven by machine learning algorithm, variety of features and the expected goal of the project.

Table 0.3: Results for framework validation based on London boroughs

	RMSE	**MAE**
Redbridge	716876.980	278024.610
Haringey	1434005.690	381076.430
Barking and Dagenham	1401597.480	231605.230
Newham	7788107.310	777545.160
Bromley	702459.780	237607.140
Enfield	1586598.200	308431.700
Tower Hamlets	3705544.290	702614.320
Waltham Forest	1434195.260	248460.290
Hackney	3569664.510	685913.580
Havering	**685753.430**	**179600.520**
Barnet	963180.680	305099.280
Camden	8878168.030	178461.400
City of Westminster	14868287.580	3824202.210
Hillingdon	1274273.870	275299.680
Brent	1231299.820	432186.990
Harrow	1104653.840	314335.020
Hounslow	2017187.970	402558.960
Hammersmith and Fulham	2417088.770	819460.010
Kensington and Chelsea	4056425.810	1499238.190
Sutton	1083429.890	281996.300
Lambert	3447359.710	545469.240
Greenwich	969642.360	309653.980
Lewisham	854336.740	276086.190
Richmond upon Thames	1141411.310	369330.110
Islington	8059242.700	1029113.430
Bexley	871895.110	206769.270
Ealing	3375141.060	405565.530
Wandsworth	1381542.870	415560.180
Kingston upon Thames	1723734.190	305596.390
Southwark	3263439.990	643275.270

Merton	947983.910	317231.930
Croydon	1011097.650	231429.570
City of London	16441998.510	8004872.990

Figures 6.11 and 6.12 are plots of the total actual value of all transactions for each London borough and how this compares with the predicted value respectively. There is a significant variance observed for City of Westminster while London boroughs like Barking and Dagenham, Ealing, Greenwich, Hounslow, Richmond upon Thames and Waltham Forest are quite aligned.

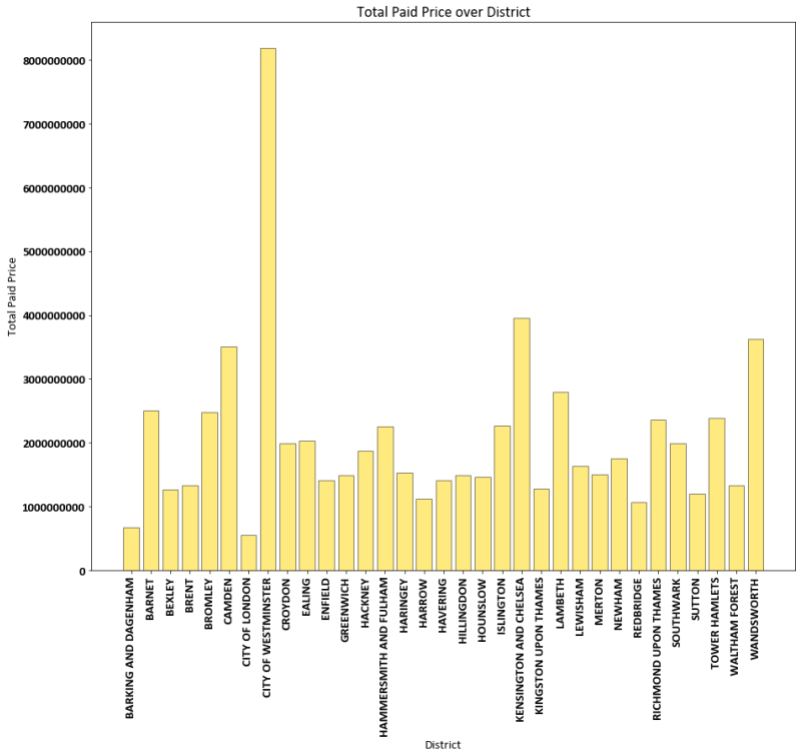


Figure 6.11: Total price paid per London borough

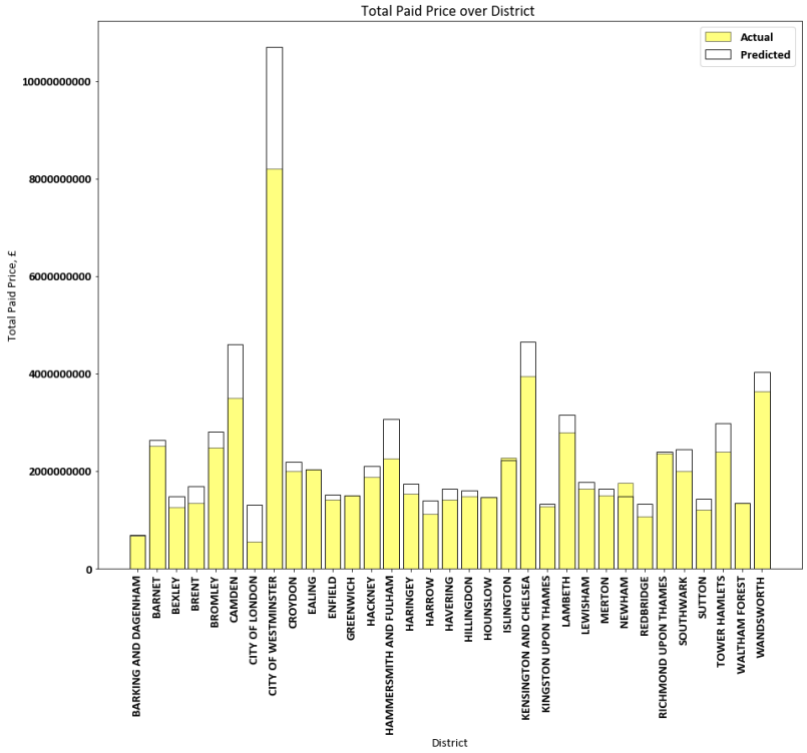


Figure 6.12: Total price paid per London borough – predicted vs actual

6.4.3 Framework validation based on property type

The cumulative Multi-feature House Price Estimation (MfHPE) framework in its current state shows different performance levels based on property type. It is observed from RMSE and MAE results in Table 6.4 that the model of this framework estimates the price of semi-detached houses more accurately than terraced house, which are then more accurately estimated than flats or maisonettes, which are then more accurately estimated than detached houses. However, if the goal of a data science or machine learning driven project is to estimate the price of a particular house type, either the MAE or RMSE metrics could be used because their evaluation based on this feature has produced the same performance results. Figures 6.13 and 6.14 show the total price value per house type, and how the model has performed in predicting house price based on this feature.

Table 0.4: Results for framework validation based on property type

Property Type	RMSE	MAE
Terraced	576816.010	238399.050
Flats/Maisonettes	701937.390	257975.460
Semi-detached	**477432.510**	**201791.490**
Detached	905674.790	416480.200
Other	15714324.380	5451280.220

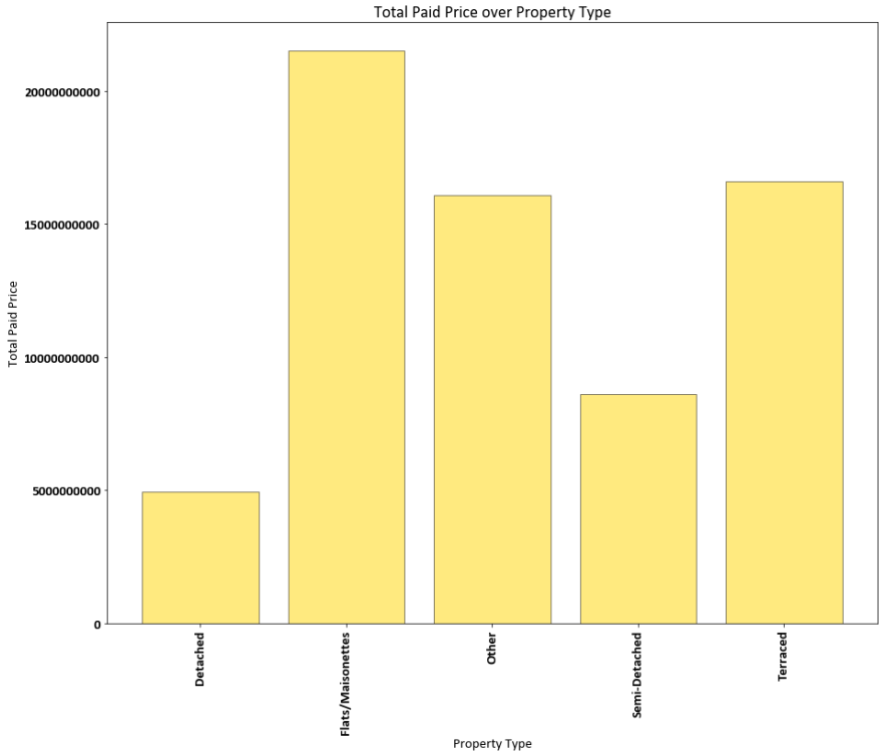


Figure 6.13: Total price paid per property type

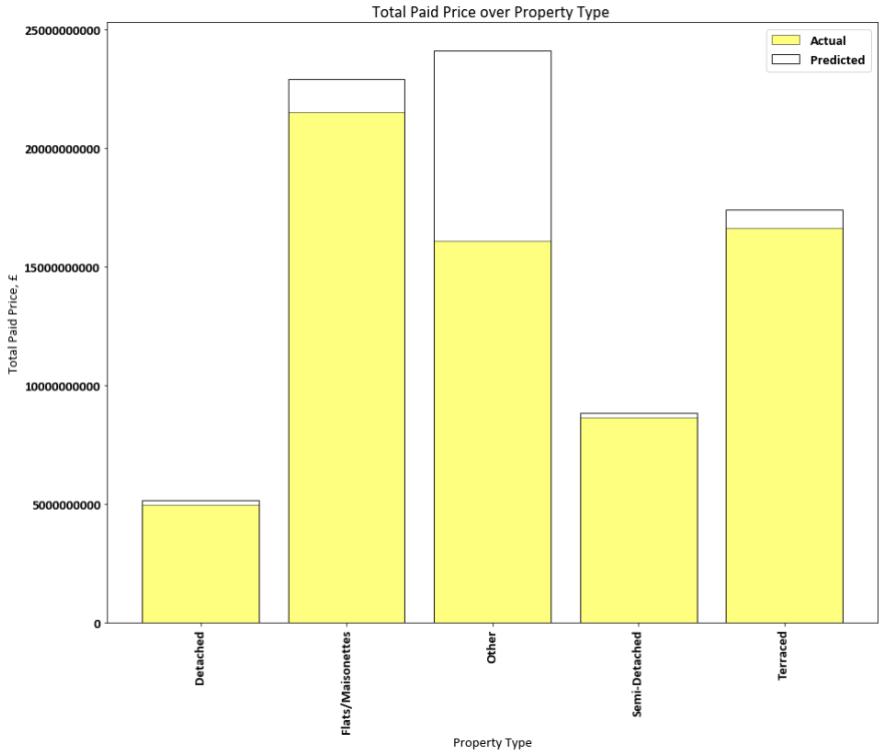


Figure 6.14: Total price paid per property type – predicted vs actual

6.4.4 Framework validation based on age of property

Table 6.5 shows results for the evaluation of the validation data based on the age of the property. It is observed that the performance for both RMSE and MAE evaluations are alike in that the model performs better when estimating the prices of new houses than it does for old houses. In the context of the dataset, 'Old' means the house is not a new build at the time of exchange.

Table 0.5: Results for framework validation based on age of property

Age	RMSE	MAE
Old	4139511.300	591106.420
New	**1076591.820**	**384846.310**

Figures 6.15 and 6.16 show the total price paid based on whether the house is brand new at the time of exchange or not, and how the model has performed in predicting house price based on this feature. It is observed that the model performs better when predicting or estimating the price of new houses compared

to old houses. Figure 6.15 especially shows the lack of investment in new houses in London.

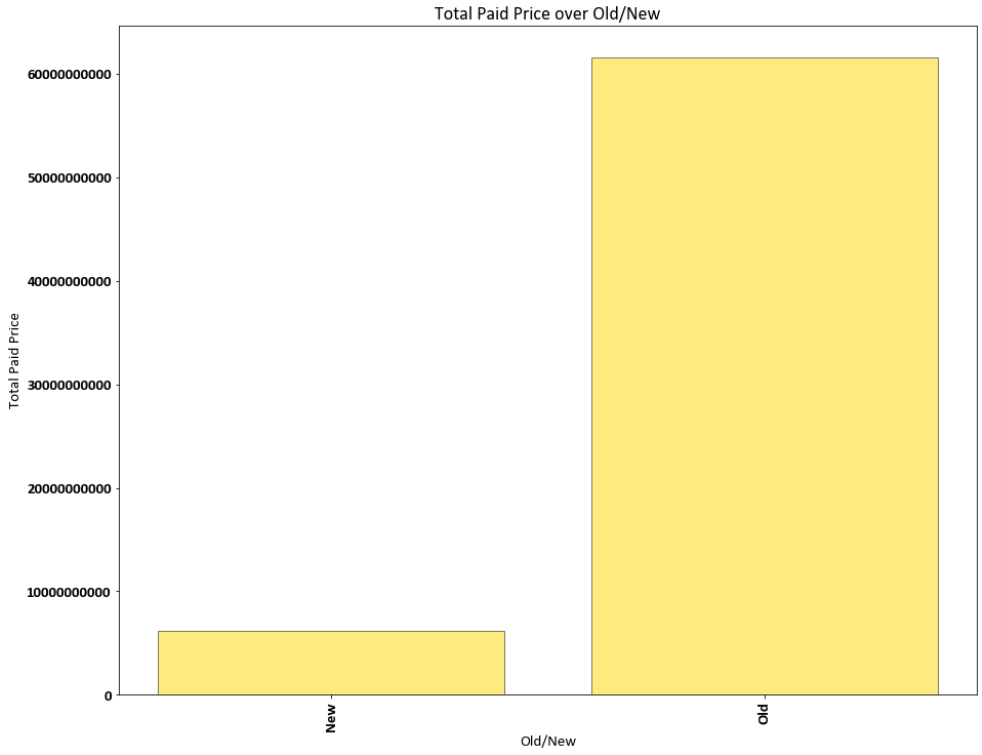


Figure 6.15: Total price paid based on property age

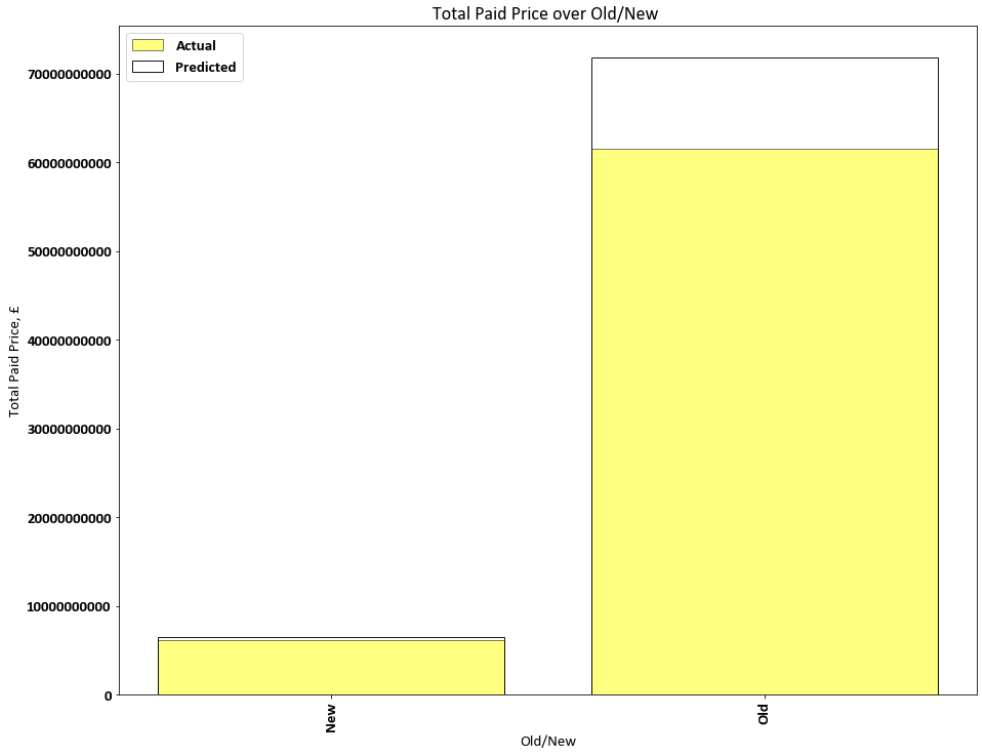


Figure 6.16: Total price paid based on property age – predicted vs actual

6.4.5 Framework validation based on duration

The duration of a house describes whether the buyer owns both the house and the land it is built on or just the house. If the buyer owns both, it is freehold, and if the buyer only owns the house and the land is leased, its leasehold (Cobb Farr, 2021). In Table 6.6, it can be seen that the performance of the model for both evaluation metrics are similar as results show that the model performs better when estimating the price of houses that are leasehold as compared with those that are freehold.

Table 0.6: Results for framework validation based on duration

Duration	RMSE	MAE
Freehold	4378421.930	661539.540
Leasehold	**3432322.960**	**477025.570**

Figures 6.17 and 6.18 further emphasise the total value of transactions based on duration, and especially that the model performs better for leasehold houses than freehold houses, although the difference in the actual and predicted prices is not significant.

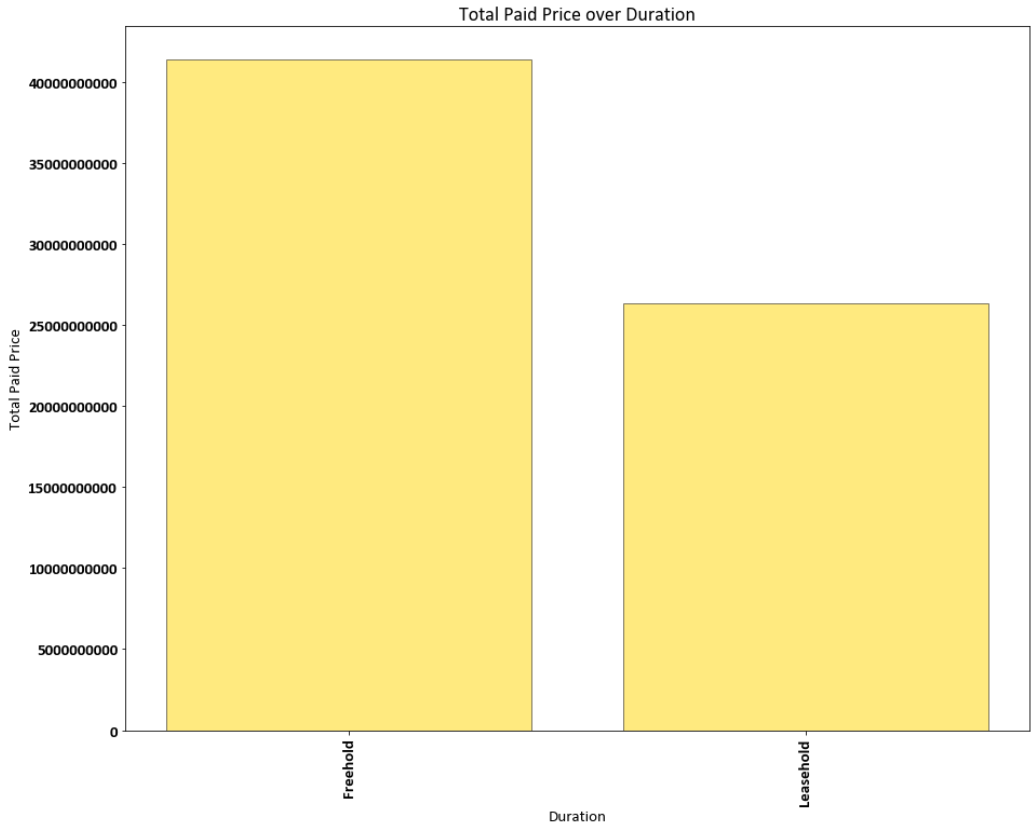


Figure 6.17: Total price paid per duration type

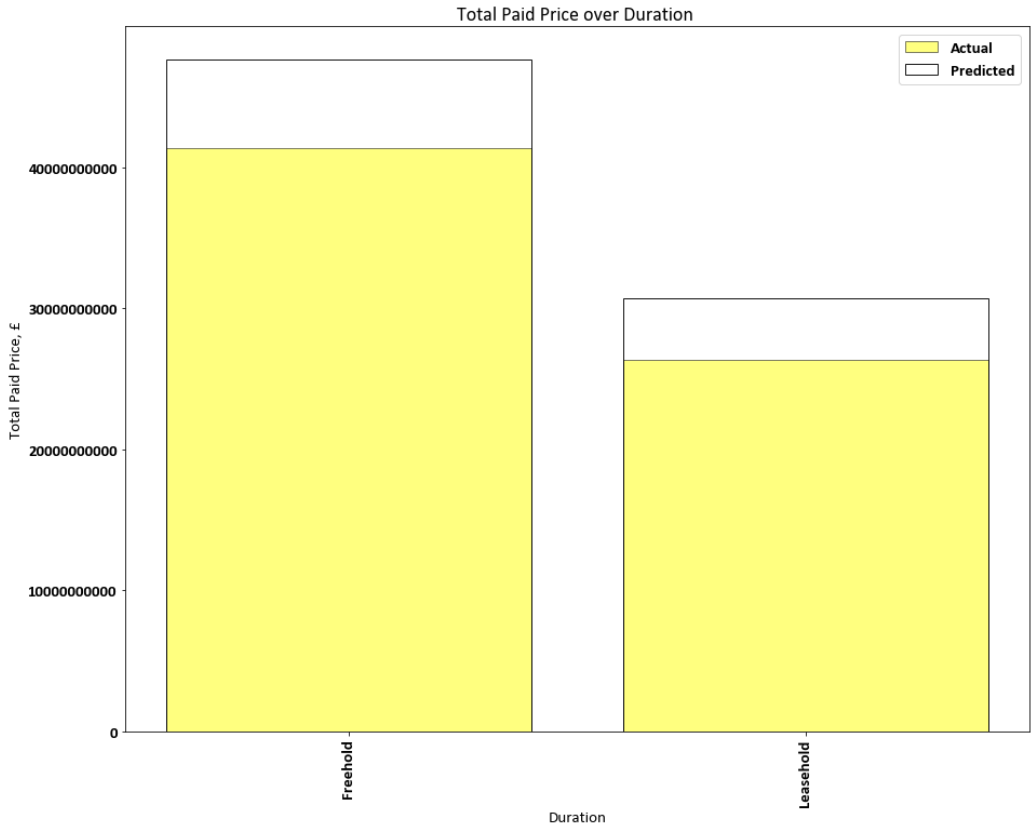


Figure 6.18: Total price paid per duration type – predicted vs actual

6.4.6 Framework validation based on transfer month

The performance of the model to accurately predict the price of a house is observed to vary over the course of a year. Based on the current state of the cumulative MfHPE framework, the models perform better for transactions that occur in August (the eighth month) than for transactions in September. However, based on the RMSE evaluation of the model performance, Table 6.7 shows that estimated house prices for April are like likely to be more accurate than transactions in June, the MAE evaluation presents a reverse outlook with the model performing better for transactions in June compared with April.

Table 0.7: Results for framework validation based on month of transfer

Month	RMSE	MAE
1 - January	2596389.590	641343.200
2 - February	2384264.040	498687.130
3 - March	4044070.310	675542.210
4 - April	3942841.440	652246.740
5 - May	3630347.510	565040.070
6 - June	5168423.790	583850.170
7 - July	3214640.580	572550.530
8 - August	**1867642.790**	**469394.420**
9 - September	3663407.910	562152.630
10 - October	6439268.700	569999.960
11 - November	3406176.680	500342.500
12 - December	4280932.110	568234.390

6.4.7 Framework validation based on transfer quarter

Similarly to Section 6.4.6, the performance of the cumulative MfHPE framework was validated with a view to understanding model performance for each quarter of the year. The model, based on the RMSE evaluation, is observed to perform better with transactions in the first quarter compared to the second quarter, while MAE evaluation is the reverse. However, the model performs best for transactions in the third quarter for both MAE and RMSE.

Table 0.8: Results for framework validation based on quarter of transfer

Quarters	RMSE	MAE
Q1	3131678.070	608297.220
Q2	4411916.140	599588.670
Q3	**3049658.250**	**536256.490**
Q4	4899194.670	547306.220

The summation of most studies will usually focus on the global performance of a machine learning model, which based on the metrics exploited here is an average rather than a detailed insight on each feature and their subcategories. Overall, this section is designed for the user of the cumulative Multi-feature House Price Estimation (MfHPE) framework in its current state to assess its suitability for their project goal or business problem.

6.5 Conclusion

This chapter has provided an overarching evaluation as well as a segment-based evaluation of results by assessing the impact of the models across multiple dimensions including property type, duration, age, transfer month or quarter and the London borough in which the house is located.

A review of the top ten features that made the most impact for each tier reveals the dominance of Tier 1 features across the models. This means that house characteristics play a significant role in the price of a house, while other features can be described as value-adding. However, it is important to note that this may be true for London and similar UK cities, but may not remain true in other UK locations.

The validation of the framework model was completed both at a generic level and across various verticals so as to get an understanding of model performance at more granular levels. One of the findings here highlights that the choice of evaluation metrics could be influenced by a blend of business problems or project goals, the variety of features and the machine learning algorithms used. This tripartite view to choosing an evaluation metric is more likely to provide the best-fit insights that enable decision making.

Chapter 7: Conclusion and Recommendations

7.1 Introduction

This research investigated the possibility of making a contribution to knowledge based on the foundations that have been laid for how machine learning systems can be designed to learn and exploit multiple data points so as to produce insights that enable the decision-making process and/or behaviours of stakeholders in the housing market. The reference point for decisions such as where to build or invest in a house for the best return on investment ultimately depends on the price of the house. If the price of a house can be accurately estimated based on a number of known factors and possible future infrastructural or neighbourhood or economic changes, these stakeholders will have a baseline that enables the choices or decisions they make. The investigations have led to the design of the Multi-feature House Price Estimation framework (MfHPE) which is process-oriented, modularised, data-driven and machine learning enabled. This design was implemented using the Design Science Research methodology (DSRM), where the MfHPE Framework has been developed and evaluated. The next sections of this chapter are organised as follows: *Section 7.2* provides a summary of the contributions to knowledge; *Section 7.3* provides answers to the research hypothesis and questions; *Section 7.4* highlights the limitations of this research while *Section 7.5* sets some scope for future directions by making some recommendations.

7.2 Contributions to knowledge

The overarching contribution to knowledge in this research is to minimise the existing research gap by taking a cumulative multi-feature layering approach to the development of the MfHPE framework that is process-based, modularised, data-driven and machine learning enabled in order to produce insights that enhance decision making for a range of stakeholders. The summary captured

below is a snapshot of the leading contributions of this research ordered by their significance.

First, the **modularised Multi-feature House Prices Estimation framework**: This is the main design output of this research based on design science methodology. It is described as '**modularised**' because it is made up of nine different modules. The modularity of the MfHPE framework makes it robust and enables (i) update of existing datasets, (ii) introduction of new datasets, and (iii) exploitation of other interesting machine learning algorithms.

Second, the **Cumulative Multi-feature Layering** of groups of multiple parameters throughout the model development. This led to the creation of 48 machine learning models that exploit five different machine learning algorithms and three groups of features, being the baseline features followed by a layer of neighbourhood features and then a layer of macroeconomic features. The MfHPE is described as (i) '**cumulative**' because the layering approach allows the introduction of new layers without the removal of existing layers in each model, (ii) '**multi-feature**' because the framework has leveraged features from multiple datapoints and can have more introduced by design, and (iii) '**layering**' because groups of features are introduced as new layers into the framework.

Third, the **Research Dataset** which leverages the HM Land Registry Price Paid Data as a baseline for transactions data. This is then geo-coded by blending it with the ONS NSPL product. The variables of the geo-coded HM Land Registry Price Paid Data are used to create new variables that enable a further data-blend with neighbourhood and macroeconomic datasets to create the complete research dataset. Therefore, this research dataset comprises Price Paid transaction data for London boroughs published by HM Land Registry blended with ONS NSPL product being Tier 1, then with bus stops, retail locations, national rail and underground stations being Tier 2 features, and then macroeconomic indicators including GDP, inflation rate, employment rate, unemployment rate and consumer price index being Tier 3 features. It is a whole new dataset that can be analysed for insights.

Fourth, **Response of Machine Learning Algorithms to changes in Data Variety**. As compared with over 200 papers reviewed on the subject of house price estimation, the cumulative multi-feature layering approach explored in this

research unveiled how machine learning algorithms respond to a changing landscape of features as multiple tiers of features were introduced into the framework, thereby scaling the variety of features used in creating the machine learning models. The focus of existing studies ranges from comparison between algorithms to model accuracy, algorithm performance, predictors and model explainability.

Fifth, **Evaluation Metrics respond differently to Features and Machine Learning Algorithms:** As discussed in Section 3.9 of the literature review, the choice of evaluation metrics is usually driven by the type of machine learning problem. For example, Classification Problem (F1-Score, ROC, Precision), Regression Problem (RMSE, MAE, MSE), Ranking Problem (NDGG, MRR) and Statistical Problem (Correlation). However, beyond these the choice of evaluation could be influenced by a blend of business problems or project goals, the variety of features and machine learning algorithms. The analysis and evaluation of the validation data in this thesis as detailed in Section 6.4 proposes that this tripartite view to choosing an evaluation metric is more likely to provide the best-fit insights that enable decision making.

These contributions have the potential short-term to life-long as the research data could be extended the next set of studies while the contributions based on modularity, cumulative multi-feature layering, machine learning algorithm response to data variety and the behaviour of evaluation metrics can be life-long.

7.3 Fulfilment of the research hypothesis and questions

The instantiation of the MfHPE framework focusing on the research hypothesis, followed by the evaluation of the framework designed, has led to the conclusion that the use of standalone and ensemble machine learning (ML) algorithms on a publicly available dataset can create a deeper understanding of how different algorithms perform based on variation in datasets, and can also produce insights that enhance the decision making process for a range of stakeholders through the estimation of house prices, as designed and demonstrated in Chapter 4 and 5 respectively. Therefore, the outcomes of the design and development of the

framework are now discussed within the context of the research questions as follows:

Research Question 1: What data led methods have been used for estimating house prices?

Existing literature shows that there are various data led methods that have been exploited for the estimation of house prices. These range from statistical models, to hedonic regression models and machine learning models.

The proposed framework is machine learning enabled and has leveraged multiple features from ten datasets, five machine learning algorithms creating over 48 models, and three evaluation metrics so as to estimate house prices.

Research Question 2: What are the house characteristics, neighbourhood factors, macroeconomic indicators and other factors that influence the value of house prices?

Previous academic and commercial studies have investigated the various factors that have an impact on the value of houses, and these include: (i) house characteristics such as overall floor space, number of bedrooms, bathrooms, reception rooms, floors, driveway, balcony, storage, garden, energy efficiency, carpet area, age and design/layout; (ii) neighbourhood factors such as schools, retail, seaside, road network, public transport network, recreation, cafes, restaurants, hospitals, fire stations, police stations and crime rate; (iii) macroeconomic indicators such as GDP, interest rate, inflation rate, mortgage rate, disposable income, employment rate and unemployment rate; and (iv) others such as class, diversity, location, population, previous sale price and history.

The MfHPE framework has leveraged the HM Land Registry's Price Paid data, which is a compilation of all house sale transactions for England and Wales. This dataset is limited as it does not provide enough context on the house characteristics. For full details of all attributes see Table 4.1. The framework also exploits neighbourhood factors like rail stations, retail and bus stops by calculating the distance between postcodes and the existing factors. It also gives some consideration to the volume of each of these neighbourhood factors within

a specific distance from the postcode of each house. Finally, the framework then leverages five macroeconomic indicators including GDP, inflation rate, employment rate, unemployment rate and consumer price index.

Research Question 3: Can machine learning be used to understand the influence different groups of factors have on the estimate house prices?

Further to a review of existing research, the findings of some studies show how some factors have a positive correlation and others a negative correlation to house price. Park and Bae (2014) in their study found that although previous studies have leveraged hedonic-based methods, 'machine learning algorithms can enhance the predictability of house prices'.

The MfHPE framework has made a provision for providing insights on the prevailing features that drive the prediction/estimation made by the models created. The sections on model explainability in 4.14 and 6.3 helps provide an overview of the SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) which provides an approach to understanding of why machine learning models make certain predictions. The explainability of machine learning models is important because the reasons behind the predictions they make is as crucial as the accuracy of the models.

Research Question 4: What evaluation approaches exist in this industry and how will this research work be evaluated?

The evaluation of machine learning models is an integral part of the development of the models because it helps to identify the models that perform the best with the data in use, and therefore informs future development. 90% of existing research reviewed used Root Mean Square Error (RMSE), R-squared, Mean Absolute Error (MAE) or Mean Square Error (MSE), while others used evaluation metrics including Precision, Root Mean Square Logarithmic Error (RMSLE). The proposed framework has leveraged R-squared, MAE and RMSE as metrics to evaluate the errors in the 48 models created as shown and discussed extensively in Chapter 4. Furthermore, the framework validation in Section 6.4 then unveils insights about the dynamics of evaluation metrics when model performance measurement takes a more granular approach than the global approach, which

is the trend. It is observed that different evaluation metrics may produce different performance results and this leads to the recommendation that the choice of evaluation metrics should be driven by (i) business problems or project goals, (ii) features, and (iii) machine learning algorithms.

Research Question 5: Can multiple datapoints be integrated so as to improve the accuracy of house price estimation?

All existing research on the subject of prediction or estimation of house prices using machine learning has used datasets with multiple features to develop models, but not all studies have exploited multiple data points. Kuvalekar et al. (2020), in their examination of how to predict house prices in Mumbai city, exploited a single 'real estate' dataset with relevant house characteristics.

The MfHPE framework has exploited six data points, ten datasets and multiple features for the development of a machine learning system that is modularised, process-based and data-driven. The data points are Bank of England, HM Land Registry, Office of National Statistics, Geolytix, Doogal and GOV.UK, as seen on the data profiles in Chapter 4.

Research Question 6: What is the impact data variety on the accuracy of house price estimation?

Research Question 7: How do different machine learning algorithms respond to changes due to data variety?

Summarising the MfHPE framework in the context of questions 6 and 7, the development of the ML models based on the results shown and discussed in Chapters 5 and 6, reveals that as the variety of the research data scaled from Tier 1 through Tier 2 to Tier 3, the performance of the five machine learning algorithms used changed. Random Forest was the best performing algorithm for Tiers 1 and 2, while XGBoost became the best performing algorithm at Tier 3 as a result of the introduction of new features that increased the variety of data in the models. As a result of the variation in the performance of models in the MfHPE framework, the best performing model overall based on the focus on Tier 3

features is XGBoost because the results show the lowest value for RMSE and MAE as well as the highest value for R-squared.

7.4 Research limitations

This research is limited to a single geographic location, London, United Kingdom, and the width or variety of features captured in the datasets exploited. Even though the combination of features that make up the research data are essential in any machine learning system designed to estimate house prices, additional supportive features are required, including: (i) house characteristics such as energy efficiency, overall floor space, number of bedrooms, number of bathrooms, number of reception rooms, floors, driveway, balcony, storage, garden, flooring material and layout; (ii) neighbourhood features such as schools, seaside, road network, recreation, cafes, restaurants, hospitals, fire stations, police stations and crime rate; (iii) macroeconomic indicators such as the stock market, producer price index, balance of trade, housing starts and interest rate, and other unwritten factors such as class, diversity, location, population, previous sale price and sentiments. Expanding the scope of this research to cover these features will lead to more complexity as the number of features in existing tiers will increase and new tiers of features created, modelling and tuning of models will become more computationally expensive, processing will take longer, and a budget will be required for datasets that are not free as well as a cloud platform providing scalable computing resources, etc. These are beyond the scope of a self-funded professional doctorate programme but could definitely be explored with some industry funding.

7.5 Future research recommendations

The subsections that follow will outline a list of the fundamental future research recommendations that emerged during the life cycle of this research:

7.5.1 Expanding the scope of the cumulative MfHPE framework

The scope of the cumulative Multi-feature House Price Estimation Framework can be expanded in various ways. First, through the introduction of more features that fit into the existing tiers. For house characteristics as a baseline, these will include features such as house energy efficiency score, number of bedrooms, reception rooms, bathrooms, full floor space, building materials, maintenance history and lots more. Neighbourhood factors will include features like schools, recreation, hospitals, petrol stations, dentist, pet shops, etc. Expanding the scope will make the tiers richer and ultimately improve the accuracy of the algorithms, although they may not necessarily respond in the same way. Second, through the creation of more tiers for groups of features that do not fit into the existing tiers, such as crime rate, perception, population, reduced stamp duty rates during the pandemic, etc. Third, through the introduction of other regression algorithms for the assessment of their performance in terms of accuracy as well as response to data variety.

7.5.2 Modelling the impact of layered features individually

The cumulative layering approach explored in this version of the MfHPE framework introduces tiers of features as layers. However, conducting the investigation at a more granular level by introducing features one-by-one and cumulatively rather than as tiers will provide deeper insights into the practical impact of each feature on the estimation of house prices. It will be great to see the impact the volume of commuters passing through a train station has on the value of houses within a specific radius, or even the impact a particular brand of retailer has on house prices in the area. This is going to be more computationally more expensive but worth it.

7.5.3 Identification of best-fit models and evaluation metrics for unique features and sub-categories

As a result of the framework validation initiated in Section 6.4, it has become imperative that a global approach to model evaluation may not suffice in fully optimising the performance of machine learning models. Further, studies on the path of this thesis are encouraged to discover the combination of algorithms,

optimisation techniques and evaluation metrics that will produce the best performance for the estimation of house prices, as the results discussed in this thesis shows there is not yet a single best approach.

7.6 Conclusion

This chapter has focused on the five contributions of this thesis to knowledge with a focus on UK house price estimation, though with some learnings from other geographic domains, how the completion of this thesis answers the seven research questions raised, discussed some research limitations, and provided some recommendations for future work.

References and Bibliography

ABDULAI, R. T. & OWUSU-ANSAH, A. 2011. House Price Determinants in Liverpool, United Kingdom. *Current Politics and Economics of Europe,* 22**,** 26.

ACHARYA, S. 2021. *What are RMSE and MAE?* [Online]. Towards Data Science: Towards Data Science. Available: https://towardsdatascience.com/what-are-rmse-and-mae-e405ce230383 [Accessed 19/01/2021 2022].

ADAMS, Z. & FÜSS, R. 2010. Macroeconomic Determinants of International Housing Markets. *Journal of Housing Economics,* 19**,** 13.

AGRAWAL, N. 2020. *Hyperparameter Tuning for Machine Learning Explained* [Online]. Great Learning: Great Learning. Available: https://www.mygreatlearning.com/blog/hyperparameter-tuning-explained/ [Accessed 28/06/2021 2021].

AGUIAR, R. 2019. *An Overview of Model Explainability in Modern Machine Learning* [Online]. Available: https://towardsdatascience.com/an-overview-of-model-explainability-in-modern-machine-learning-fc0f22c8c29a [Accessed 12-Sep-21 2021].

ALBOUY, D. 2016. What are Cities Worth? Land Rents, Local Productivity, and the Total Value of Amenities. *Review of Economics and Statistics.*

ALBUQUERQUE, A. M., BENNETT, B., CUSTODIO, C. & CVIJANOVIC, D. 2021. CEO Compensation and Real Estate Prices: Pay for Luck or Pay for Action?

ANTIPOV, E. A. & POKRYSHEVSKAYA, E. B. 2012. Mass appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics. *Expert Systems with Applications,* 39**,** 7.

ARIETTA, S. M., EFROS, A. A., RAMAMOORTHI, R. & AGRAWALA, M. 2014. City Forensics: Using Visual Elements to Predict Non-Visual City Attributes. *IEEE transactions on visualization and computer graphics,* 20**,** 10.

AVANIJAA, J. 2021. Prediction of House Price Using XGBoost Regression Algorithm. 5.

AWONAIKE, A., GHORASHI, S. A. & HAMMAAD, R. 2021. A Machine Learning Framework for House Price Estimation. *International Conference on Intelligent Systems Design and Application.* Online.

BALCILAR, M., GUPTA, R. & STEPHEN, M. 2015. The Out-of-Sample Forecasting Performance of Nonlinear Models of Regional Housing Prices in the US. *Applied Economics,* 47**,** 19.

BALDAUF, M., GARLAPPI, L. & YANNELIS, C. 2020. Does Climate Change Affect Real Estate Prices? Only If You Believe In It. *The Review of Financial Studies*

BARKSENIUS, A. & RUNDELL, E. 2013. House Prices for Real–The Determinants of Swedish Nominal Real Estate Prices.

BELCHER, R. N. & CHISHOLM, R. A. 2018. Tropical Vegetation and Residential Property Value: A Hedonic Pricing Analysis in Singapore. *Ecological Economics*

BENIGNO, G. & THOENISSEN, C. 2002. Equilibrium exchange rates and supply-side performance. *In:* ENGLAND, B. O. (ed.). London: Bank of England.
BHATT, U., XIANG, A., SHARMA, S., WELLER, A., TALY, A., JIA, Y., GHOSH, J., PURI, R., MOURA, J. M. & ECKERSLEY, P. 2020. Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.
BHATTACHARYA, S. 2019. Model Evaluation Techniques for Machine Learning Classification Model. *GreatLearning Blog: Free Resources what Matters to shape your Career!* [Online]. Available from: https://www.mygreatlearning.com/blog/model-evaluation-techniques-for-machine-learning-classification-models/ [Accessed <td align="right" class="xl65" height="21" style="height:15.75pt; width:56pt" width="74">12-Jun-21 2021].
BIBAL, A., LOGNOUL, M., DE STREEL, A. & FRÉNAY, B. 2021. Legal Requirements on Explainability in Machine Learning. *Artificial Intelligence and Law,* 29**,** 21.
BISHOP, C. 2013. Model-based machine learning. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences,* 371**,** 201-222.
BISWAS, S. & RAJAN, H. 2021. Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline. *arXiv preprint arXiv:2106.06054* 52.
BROWNLEE, J. 2020. *Ordinal and One-Hot Encodings for Categorical Data* [Online]. Available: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/ [Accessed 30-Jun-21 2021].
BURKART, N. & HUBER, M. F. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research,* 70**,** 73.
BURLUTSKIY, N., PETRIDIS, M., FISH, A., CHERNOV, A. & ALI, N. 2016. *An Investigation on Online versus Batch Learning in Predicting User Behaviour* [Online]. Available: https://cris.brighton.ac.uk/ws/portalfiles/portal/453691/offline_vs_online1.pdf [Accessed 09/09/2021 2021].
CARIOU, B. 2020. What Is Data Modeling & Why Does It Matter? Available from: www.trifacta.com/blog/what-is-data-modeling.
CASE, K. E., QUIGLEY, J. M. & SHILLER, R. J. 2013. Wealth Effects Revisited 1978–2009 (No. w16848). *LES DÉTERMINANTS DES PRIX DE L'IMMOBILIER AUX ÉTATS-UNIS APRÈS.*
ČEH, M., KILIBARDA, M., LISEC, A. & BAJAT, B. 2018. Estimating the Performance of Random Forest Versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS international journal of geo-information,* 7.
CERDA, P. & VAROQUAUX, G. 2020. Encoding High-Cardinality String Categorical Variables. *IEEE Transactions on Knowledge and Data*.
CHAUHAN, N. S. 2020a. *Hyperparameter Optimization for Machine Learning Models* [Online]. Available: https://www.kdnuggets.com/2020/05/hyperparameter-optimization-machine-learning-models.html [Accessed 28/06/2021 2021].

CHAUHAN, N. S. 2020b. *Model-Evaluation-Metrics-Machine-Learning* [Online]. Available: https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html [Accessed 2021].
CHEN, N.-K., CHENG, H.-L. & MAO, C.-S. 2014. Identifying and Forecasting House Prices: A Macroeconomic Perspective. *Quantitative Finance,* 14**,** 16.
CHEN, S. & JIN, H. 2019. Pricing for the Clean Air: Evidence from Chinese Housing Market. *Journal of Cleaner Production,* 206**,** 10.
CHEN, T. & GUESTRIN, C. 2016. Xgboost: A Scalable Tree Boosting System. *International Conference on Knowledge Discovery and Data Mining.*
CHI, B., DENNETT, A., OLÉRON EVANS, T. & MORPHET, R. 2019. Creating a New Dataset to Analyse House Prices in England. *CASA Working Paper,* 213.
CHICA-OLMO, J., CANO-GUERVOS, R. & CHICA-RIVAS, M. 2019. Estimation of Housing Price Variations Using Spatio-Temporal Data. 11.
CHO, K., MERRIËNBOER, B. V., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. & BENGIO, Y. 2014. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.
CHOUDHURY, A. 2019. 10 Model Evaluation Techniques Every ML Enthusiast Must Know.
CHOUDHURY, A. 2021. *Top 8 Approaches For Tuning Hyperparameters Of Machine Learning Models* [Online]. Available: https://analyticsindiamag.com/top-8-approaches-for-tuning-hyperparameters-of-machine-learning-models/ [Accessed 28-Jun-21 2021].
CHU, Y. 2014. Credit Constraints, Inelastic Supply, And the Housing Boom. *Review of Economic Dynamics,* 17**,** 18.
CICCERI, G., INSERRA, G. & LIMOSANI, M. 2020. A Machine Learning Approach to Forecast Economic Recessions—An Italian Case Study. *Mathematics* 8.
COBB FARR. 2021. *Difference between Freehold vs Leasehold Property in the UK* [Online]. Available: https://www.cobbfarr.com/the-knowledge/the-differences-between-freehold-and-leasehold-property-in-the-uk/#:~:text=What's%20the%20difference%20between%20Freehold,property%20but%20not%20the%20land. [Accessed 20/01/2022 2022].
COHEN, V. & KARPAVICIUTE, L. 2017. The Analysis of the Determinants of Housing Prices. *Independent journal of management & production,* 8**,** 15.
CORDERA, R., COPPOLA, P., DELL'OLIO, L. & IBEAS, Á. 2019. The Impact of Accessibility by Public Transport on Real Estate Values: A Comparison Between the Cities of Rome and Santander. *Transportation Research Part A: Policy and Practice,* 125**,** 12.
DAS, I. & GUPTA, R. 2012. Relationship Between Price and Rent in the Real Estate Market and Stability Analysis: A Theoretical Approach. *Indian Economic* Review**,** 18.
DE MELO JUNIOR, G., OLIVEIRA, S., FERREIRA, C., FILHO, E., CALIXTO, W. & FURRIEL, G. 2017. Evaluation Techniques of Machine Learning in Task of Reprovation Prediction of Technical High School Students.

DENNIS, J. B. 2003. Fresh Breeze: A Multiprocessor Chip Architecture Guided by Modular Programming Principles. *ACM SIGARCH Computer Architecture News,* 31**,** 9.
DOHAIMAN, M. S. 2017. The Impact of Stock Market and Macroeconomic Variables on Real Estate Prices Dynamics: Evidence from Saudi Arabia. *International Journal of Sustainable Real Estate and Construction Economics,* 1**,** 13.
DONOVAN, H. G., LANDRY, S. & WINTER, C. 2019. Urban Trees, House Price, and Redevelopment Pressure in Tampa, Florida. *Urban Forestry & Urban Greening,* 38**,** 7.
DRAKE, L. 1993. Modelling UK House Prices Using Co-Integration: An Application of the Johansen Technique. *Applied Economics,* 25**,** 4.
DUGAR, D. 2018. *Skew and Kurtosis: 2 Important Statistics Terms You Need to Know in Data Science* [Online]. Available: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa [Accessed 2021].
DURGANJALI, P. & PUJITHA, M. V. 2019. House Resale Price Prediction Using Classification Algorithms. *In 2019 International Conference on Smart Structures and Systems (ICSSS).* Chennai, India: IEEE.
EBRAHIM, M. & MAHMOUD, G. 2014. Project-Database Normalization.
EESSAAR, E. 2016. The Database Normalization Theory and the Theory of Normalized Systems: Finding a Common Ground. *Baltic Journal of Modern Computing,* 4.
ELLISON, M. & SCOTT, A. 2000. Sticky Prices and Volatile Output. *Journal of Monetary Economics,* 46**,** 12.
FENG, L., LV, X.-L., YAN, T.-H. & ZHOU, H.-C. 2019. Modular Programming of Hierarchy and Diversity in Multivariate Polymer/Metal–Organic Framework Hybrid Composites. *Journal of the American Chemical Society,* 141**,** 8.
FERLAN, N., BASTIC, M. & PSUNDER, I. 2017. Influential Factors on the Market Value of Residential Properties. *Engineering Economics,* 28**,** 10.
FERNANDO, J., SMITH, A. & PEREZ, Y. 2021. *R-Squared* [Online]. Investopedia: Investopedia Available: https://www.investopedia.com/terms/r/r-squared.asp [Accessed 04/01/2022 2022].
FEURER, M. & HUTTER, F. 2019. Hyperparameter Optimization. *Springer***,** 30.
FORD, C. 2015. Understanding Q-Q Plot. *University of Virginia Library Research Data Services + Sciences.*
GALLENT, N., DURRANT, D. & MAY, N. 2017. Housing Supply, Investment Demand and Money Creation: A Comment on the Drivers of London's housing Crisis. *Urban Studies,* 54**,** 13.
GALLIN, J. 2008. The Long-Run Relationship Between House Prices and Rents. *Real Estate Economics,* 36**,** 24.
GAO, G., BAO, Z., CAO, J., QIN, A. K., SELLIS, T. & WU, Z. 2019. Location-Centered House Price Prediction: A Multi-Task Learning Approach. *arXiv preprint*
GARBACZ, M. 2021. Model Explainability — How to Choose the Right Tool? Available from: https://medium.com/ing-blog/model-explainability-how-to-choose-the-right-tool-6c5eabd1a46a [Accessed 12/09/2021 2021].

GAUR, C. 2020. *Data Modelling - Understanding Tools and Techniques Involved* [Online]. Available: www.xenonstack.com/insights/data-modelling [Accessed 2021].
GAWALI, S. 2021. Shape of Data: Skewness and Kurtosis. Available from: https://www.analyticsvidhya.com/blog/2021/05/shape-of-data-skewness-and-kurtosis/ [Accessed 04/07/2021 2021].
GE, C., WANG, Y., XIE, X., LIU, H. & ZHOU, Z. 2019. An Integrated Model for Urban Subregion House Price Forecasting: A Multi-Source Data Perspective. *In 2019 IEEE International Conference on Data Mining (ICDM).* IEEE.
GREEN, L. 2018. *The Effect of Immigration on UK House Prices.* Bachelors, The University of Nottingham.
GREENE, D., CUNNINGHAM, P. & MAYER, R. 2008. *Unsupervised Learning and Clustering* [Online]. ResearchGate: Springer. Available: https://www.researchgate.net/publication/235328198_Unsupervised_Learning_and_Clustering [Accessed 05/07/2021 2021].
GROLEMUND, G. 2021. *7 Exploratory Data Analysis* [Online]. Available: https://r4ds.had.co.nz/exploratory-data-analysis.html [Accessed 2021].
GU, Y. 2018. What are the Most Important Factors that Influence the Changes in London Real Estate Prices? How to Quantify Them?
GUALDI, S., TARZIA, M., ZAMPONI, F. & BOUCHAUD, J.-P. 2015. Tipping Points in Macroeconomic Agent-Based Models. *Journal of Economic Dynamics and Control,* 50**,** 33.
GUZMAN, I. & JUAN SILVA, E. 2018. Copper Price Determination: Fundamentals Versus Non-Fundamentals. *Mineral Economics,* 31**,** 18.
HALDANE, G. A. & TURRELL, A. E. 2017. An Interdisciplinary Model for Macroeconomics. *In:* ENGLAND, B. O. (ed.) *Bank of England.* Bank of England.
HAMMAD, R. K. M. 2018. *A Hybrid E-Learning Framework: Process-Based, Semantically-Enriched and Service-Oriented.* Doctor of Philisophy, University of West England.
HAMNETT, C. 2003. Gentrification and the Middle-class Remaking of Inner London, 1961–2001. *Urban Studies,* 40**,** 2401-2426.
HAO, J. & HO, T. K. 2019. Machine Learning Made Easy: A Review of Scikit-Learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics,* 44**,** 14.
HARGREAVES, B. 2008. What do Rents tell us about House Prices? *Pacific Rim Real Estate Society Conference.* Fremantle: Massey University.
HERLAND, M., KHOSHGOFTAAR, T. M. & BAUDER, R. A. 2018. Big Data Fraud Detection Using Multiple Medicare Data Sources. *Journal of Big Data.*
HILBERS, P., BANERJI, A., SHI, H. & HOFFMAISTER, A. 2008. House Price Developments in Europe: A Comparison. IMF Working Papers: IMF Working Papers.
HM LAND REGISTRY. 2020. *UK House Price Index for March 2020* [Online]. https://www.gov.uk/government/news/uk-house-price-index-for-march-2020: HM Land Registry. [Accessed 5 January 2021 2021].
HM LAND REGISTRY 2021. pp-complete. *In:* REGISTRY, H. L. (ed.). gov.uk.
HO, W. K., TANG, B.-S. & WONG, S. W. 2021. Predicting Property Prices with Machine Learning Algorithms. 23.

HONG, J., CHOI, H. & KIM, W.-S. 2020. A House Price Valuation Based on The Random Forest Approach: The Mass Appraisal of Residential Property in South Korea. *International Journal of Strategic Property Management,* 24.
HOWE, P. D., MILDENBERGER, M., MARLON, J. R. & LEISEROWITZ, A. 2015. Geographic Variation in Opinions on Climate Change at State and Local Scales in the USA. *Nature climate change.*
HUANG, P. & HESS, T. 2018. Impact of Distance to School on Housing Price: Evidence from a Quantile Regression. *The Empirical Economics Letters,* 17**,** 8.
HUME, M. & SENTANCE, A. 2009. The Global Credit Boom: Challenges for Macroeconomics and Policy. *Journal of international Money and Finance* 28**,** 36.
INVESTOPEDIA. 2021a. *Gross Domestic Product* [Online]. Available: https://www.investopedia.com/terms/g/gdp.asp [Accessed 30/05/2021].
INVESTOPEDIA. 2021b. *Inflation* [Online]. Available: https://www.investopedia.com/terms/i/inflation.asp [Accessed 30/05/2021].
IQBAL, A. & WILHELMSSON, M. 2018. Park Proximity, Crime and Apartment Prices.
JIANG , Z. & SHEN, G. 2019. Prediction of House Price Based on The Back Propagation Neural Network in The Keras Deep Learning Framework. *2019 6th International Conference on Systems and Informatics (ICSAI).* IEEE.
JORDAN, J. 2017. *Evaluating a Machine Learning Model* [Online]. Available: https://www.jeremyjordan.me/evaluating-a-machine-learning-model/ [Accessed 12/06/2021 2021].
JUNG, J.-K., PATNAM, M. & TER-MARTIROSYAN, A. 2018. An Algorithmic Crystal Ball: Forecasts-based on Machine Learning. *International Monetary*
KARAGÖZ, K. & ÖZKUBAT, G. 2019. Impact of Macroeconomic Factors on Housing Prices: An Analysis for Aegean Region. *Yaşar Üniversitesi E-Dergisi,* 16**,** 23.
KE, G., MENG, Q., FINLEY , T., WANG, T., WEI, C., MA, W., YE, Q. & LIU, T.-Y. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems,* 30**,** 3146-3154.
KIM, Y., CHOI, S. & YI, M. Y. 2020. Applying Comparable Sales Method to the Automated Estimation of Real Estate Prices. *Sustainability,* 12.
KOKTASHEV, V., MAKEEV, V., SHCHEPIN, E., PERESUNKO, P. & TYNCHENKO, V. V. 2019. Pricing Modeling in the Housing Market with Urban Infrastructure Effect. *Journal of Physics: Conference Series*.
KOMOROWSKI, M., MARSHALL, D. C., SALCICCIOLI, J. D. & CRUTAIN, Y. 2016. Exploratory Data Analysis.
KUMAR, G. K., RANI, D. M., KOPPULA, N. & ASHRAF, S. 2021. Prediction of House Price Using Machine Learning
Algorithms. *International Conference on Trends in Electronics and Informatics.* IEEE Xplore.
KUNTZ, M. & HELBICH, M. 2014. Geostatistical Mapping of Real Estate Prices: An Empirical Comparison of Kriging and Cokriging. *International Journal of Geographical Information Science,* 28**,** 18.

KUVALEKAR, A., MANCHEWAR, S., MAHADIK, S. & JAWALE, S. 2020. House Price Forecasting Using Machine Learning. *International Conference on Advances in Science & Technology (ICAST) 2020.* SSRN.
LAVY, I. & RAMI, R. 2018. The Circumstances in which Modular Programming becomes the Favor Choice by Novice Programmers. *International Journal of Modern Education and Computer Science,* 11.
LAW, S., PAIGE, B. & RUSSELL, C. 2019. Take a Look Around: Using Street View and Satellite Images to Estimate House Prices. *ACM Transactions on Intelligent Systems and Technology (TIST),* 10**,** 19.
LAWRYNOWICZ, A. & TRESP, V. 2014. *Introducing Machine Learning* [Online]. ResearchGate. Available: https://www.researchgate.net/publication/268804320_Introducing_Machine_Learning [Accessed 07/07/2021 2021].
LECCIS, F. 2019. Regeneration Programmes: Enforcing the Right to Housing or Fostering Gentrification? The Example of Bankside in London.
LIAO, Y. 2017. Machine Learning in Macro-Economic Series fFrecasting. *International Journal of Economics and Finance.*
LIGHTGBM. 2021. *Parameters* [Online]. Available: https://lightgbm.readthedocs.io/en/latest/Parameters.html [Accessed 08/06/2021 2021].
LIU, Q. & WU, Y. 2012. *Supervised Learning* [Online]. ResearchGate. [Accessed 05/07/2021 2021].
LOY, A., FOLLETT, L. & HOFMANN, H. 2016. Variations of Q–Q plots: The Power of Our Eyes! *The American Statistician,* 70**,** 13.
LU, S., LI, Z., QIN, Z., YANG, X. & GOH, R. A hybrid Regression Technique for House Prices Prediction. International Conference on Industrial Engineering and Engineering Management, 2017. 319-323.
LUNDBERG, S. M. & LEE, S.-I. A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems, 2017 Long Beach, CA, USA. ACM Digital Library, 4768–4777.
MADHURI, C. R., ANURADHA, G. & PUJITHA, M. V. 2019. House Price Prediction Using Regression Techniques: A Comparative Study. *International Conference on Smart Structures and Systems.* Chennai, India: IEEE.
MALI, P., PATIL, S., GUJAR, P. & TIWARI, M. 2021. Prediction of House Sales Prices Using Machine Learning Algorithm.
MANASA, J., GUPTA, R. & NARAHARI, N. S. 2020. Machine Learning based Predicting House Prices using Regression Techniques. *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).* IEEE.
MARSDEN, J. 2015. House prices in London–an economic analysis of London's housing market. *Greater London Authority Economics*: *Greater London Authority Economics*.
MASROM, S., MOHD, T., JAMIL, N. S., RAHMAN, A. S. A. & BAHARUN, N. 2019. Automated Machine Learning based on Genetic Programming: A Case Study on a Real House Pricing Dataset. *In 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS).* IEEE.
MCGINNIS, W. 2016. *Useful Data Science: Feature Hashing* [Online]. Available: https://www.kdnuggets.com/2016/01/useful-data-science-feature-hashing.html [Accessed 30/06/2021 2021].

MCNEESE, B. 2008. *Are the Skewness and Kurtosis Useful Statistics?* [Online]. Available: https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics [Accessed 2021].
MICALIZZI, P. 2020. *How to assess air quality sensor accuracy: MAE* [Online]. clarity: clarity. Available: https://www.clarity.io/blog/how-to-assess-sensor-accuracy-mae [Accessed 04/01/2022 2021].
MOHAMMED , M., KHAN , M. B. & BASHIER , E. B. M. 2020. *Machine Learning: Algorithms and Applications* CRC Pr I Llc.
MONNERY, N. 2011. *Safe as Houses? A Historic Analysis of Property Prices*, London Publishing Partnership.
MONTERO, J.-M., FERNÁNDEZ-AVILÉS, G. & MATEU, J. 2015. Spatial and Spatio-Temporal Geostatistical Modeling and Krigin. *John Wiley & Sons,* 998.
MOODY, J. 2019. *What does RSME really mean?* [Online]. Towards Data Science. Available: https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e [Accessed 11/09/2021 2021].
NASTESKI , V. 2017. *An overview of the supervised machine learning methods* [Online]. ResearchGate. Available: https://www.researchgate.net/publication/328146111_An_overview_of_the_supervised_machine_learning_methods [Accessed 05/07/2021 2021].
NDEGWA, J. N. 2018. Determinants of Apartment Prices Within Housing Estates of Nairobi metropolitan Area. *International Journal of Economics and Finance,* 10**,** 8.
NGUYEN, M.-L. T. 2020. The Hedonic Pricing Model Applied to the Housing Market. *International Journal of Economics & Business Administration (IJEBA),* 8**,** 13.
NIGHANIA, K. 2018. *Various Ways to Evaluate a Machine Learning Model's Performance* [Online]. Available: https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15 [Accessed 12/06/2021 2021].
NIST.GOV. 2021. *Measures of Skewness and Kurtosis* [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm [Accessed 2021].
NOVAKOVIĆ, J. D., VELJOVIĆ, A. & S, S. 2017. Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science,* 7**,** 8.
OECD EMPLOYMENT OUTLOOK. 2020. *Employment Rate* [Online]. Available: https://data.oecd.org/emp/employment-rate.htm [Accessed 30/05/2021].
OFFICE OF NATIONAL STATISTICS. 2021a. *Employee Earnings in the UK: 2020* [Online]. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/2020: ONS. [Accessed 30 May 2021 2021].
OFFICE OF NATIONAL STATISTICS. 2021b. *Inflation and Price Indices* [Online]. Available: https://www.ons.gov.uk/economy/inflationandpriceindic [Accessed 30-May-21 2021].
OFFICE OF NATIONAL STATISTICS. 2021c. *Unemployment* [Online]. Available: https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment [Accessed 30/05/2021].

OSHIRO, T. M., PEREZ, P. S. & BARANAUSKAS, J. A. 2021. How Many Trees in a Random Forest? *Springer, Berlin*, 14.
PAI, P.-F. & WANG, W.-C. 2020. Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. *Applied Sciences,* 10, 5832.
PARGENT, F. 2019. *A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling* [Online]. Available: https://osf.io/6fstx/ [Accessed 30/06/2021 2021].
PARK, B. & BAE, J. K. 2014. Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. *Expert Systems with Application,* 42.
PATIL, P. 2018. *What is Exploratory Data Analysis?* [Online]. Available: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15 [Accessed 2021].
PEFFERS, K., TUUNANEN, T., ROTHENBERGER, M. A. & CHATTERJEE, S. 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems,* 24, 45-77.
PENG, Z., HUANG, Q. & HAN, Y. 2019. Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm. *In 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT).* IEEE.
PHAN, T. D. 2018. Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. *2018 International Conference on Machine Learning and Data Engineering (iCMLDE).* IEEE.
PONCELA, P. & GARCÍA-FERRER, A. 2014. The effects of disaggregation on forecasting nonstationary time series. *Journal of forecasting*, 15.
POTDAR, K., PARDAWALA, T. S. & PAI, C. D. 2017. A comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International journal of computer applications.*
PROBST, P. & BOULESTEIX, A.-L. 2017. To Tune or Not to Tune the Number of Trees in Random Forest. 18, 18.
QUITADADMO, A., JOHNSON, J. & SHI, X. 2017. Bayesian Hyperparameter Optimization for Machine Learning Based eQTL Analysis. *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatic.*
REDDY, Y., PULABAIGARI, V. & B, E. 2018. Semi-supervised learning: a brief review. *International Journal of Engineering & Technology,* 7, 81.
REED, R. 2016. The relationship between house prices and demographic variables: An Australian case study. *International Journal of Housing Markets and Analysis,* 9, 520-537.
RIBEIRO, A., SILVA, A. & SILVA, A. R. D. 2015. Data Modeling and Data Analytics: A Survey from a Big Data Perspective. *Journal of Software Engineering and Applications,* 8.
RICO-JUAN, J. R. & TALTAVULL DE LA PAZ, P. 2021. Machine Learning with Explainability or Epatial Hedonics Tools? An Analysis of the Asking Prices in the Housing Market in Alicante, Spain. *Expert Systems with Applications,* 171.
RIGHTMOVE. 2022. *House Prices in London* [Online]. Rightmove. Available: https://www.rightmove.co.uk/house-prices-in-London.html [Accessed 20/01/2022 2022].

RIZWAN, A. 2020. *Standardization and Normalization* [Online]. Available: https://towardsdatascience.com/normalization-vs-standardization-explained-209e84d0f81e [Accessed 04/07/2021 2021].
ROSCHER, R., BOHN, B., DUARTE, M. F. & GARCKE, J. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *Ieee Access,* 8**,** 17.
ROSEN, S. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy,* 82**,** 22.
SAWANT, R., JANGID, Y., TIWARI, T., JAIN, S. & GUPTA, A. 2018. Comprehensive Analysis of Housing Price Prediction in Pune using Multi-Featured Random Forest Approach. *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).* IEEE.
SCIKIT LEARN. 2021. *sklearn.ensemble. RandomForestRegressor* [Online]. Scikit-Learn. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Accessed 07/07/2021 2021].
SCIKIT-LEARN. 2021a. *Stable* [Online]. Available: https://scikit-learn.org/stable/ [Accessed 25/03/2021 2021].
SCIKIT-LEARN. 2021b. *Pipeline and FeatureUnion: Combining estimators* [Online]. Available: https://scikit-learn.org/0.16/modules/pipeline.html [Accessed 25/03/2021 2021].
SEGER, C. 2018. An Investigation of Categorical Variable Encoding Techniques in Machine Learning: Binary Versus One-Hot and Feature Hashing.
SELECT COMMITTEE ON ECONOMIC AFFAIRS 2016. Building more homes. *In:* LORDS, H. O. (ed.). House of Lords, UK: House of Lords.
SHAIER, S. 2019. *ML Algorithms: One SD (σ)- Instance-based Algorithms* [Online]. Available: https://towardsdatascience.com/ml-algorithms-one-sd-%CF%83-instance-based-algorithms-4349224ed4f3 [Accessed 05/07/2021 2021].
SHILLER, J. R. 2015. *Irrational Exuberance*.
SHIMIZU, C. 2014. Estimation of Hedonic Single-Family House Price Function Considering Neighborhood Effect Variables. *Sustainability,* 6**,** 15.
SHINDE, N. & GAWANDE, K. 2018. Survey on predicting property price. *2018 International Conference on Automation and Computational Engineering (ICACE).* Greater Noida, India: IEEE.
SMEEKES, S. & WIJLER, E. 2018. Macroeconomic Forecasting Using Penalized Pegression Methods. *International journal of forecasting,* 34**,** 23.
SNELLING, C., COLEBROOK, C. & MURPHY, L. 2016. Homesharing & London's housing market. Institute for Public Policy Research: Institute for Public Policy Research.
SUBHAN, S., ARIFUDIN, R., EFRILIANDA, D. A. & PRASETIYO, B. 2021. Development of Automatic Spam Detection Application Based on Modular Programming. *In Journal of Physics: Conference Series.* IOP Publishing.
SUGIMURA, P. & HARTL, F. 2018. Building a Reproducible Machine Learning Pipeline.
SVIDELOC. 2020. *Target Encoding Vs. One-hot Encoding with Simple Examples* [Online]. Available: https://medium.com/analytics-vidhya/target-encoding-vs-one-hot-encoding-with-simple-examples-276a7e7b3e64 [Accessed 30/06/2021 2021].

TAJIK, M., ALIAKBARI, S., GHALIA, T. & KAFFASH, S. 2015. House prices and credit risk: Evidence from the United States. *Economic Modelling,* 51**,** 123-135.
TARANTO-VERA, G., GALINDO-VILLARDÓN, P., MERCHÁN-SÁNCHEZ-JARA, J., SALAZAR-POZO, J., MORENO-SALAZAR, A. & SALAZAR-VILLALVA, V. 2021. Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature. *The Journal of Supercomputing,* 77**,** 11481-11513.
THAMARAI, M. & MALARVIZHI, S. P. 2020. House Price Prediction Modeling Using Machine Learning. *International Journal of Information Engineering & Electronic Business,* 12.
TIFFIN, M. A. 2016. Seeing in the Dark : A Machine-Learning Approach to Nowcasting in Lebanon.
TRAUER, J. M., RAGONNET, R., DOAN, T. N. & MCBRYDE, E. S. 2017. Modular Programming for Tuberculosis Control, the "AuTuMN" Platform. *BMC infectious diseases,* 17**,** 12.
TRINH, H. H., KHAN, M. K., SQUIRES, G. & MAREIC, M. 2021. Housing Price, Financial Development, Energy Intensity,
FDI inflows: Global Evidence. New Zealand Association of Economists.
TROJANEK, R., GLUSZAK, M. & TANAS, J. 2018. The Effect of Urban Green Spaces on House Prices in Warsaw. *International Journal of Strategic Property Management.*
TRUONG, Q., NGUYEN, M., DANG, H. & MEI, B. 2020. Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science,* 174**,** 433-442.
UK GOVERNMENT. 2018. Analysis of the Determinants of House Price. *Ministry of Housing, Communities and Local Government* [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/699846/OFF_SEN_Ad_Hoc_SFR_House_prices_v_PDF.pdf.
VARMA , A., SARMA , A., DOSHI , S. & NAIR , R. House Price Prediction Using Machine Learning and Neural Networks. International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018 International Conference on Inventive Communication and Computational Technologies IEEE, 1936-1939.
VEEN, B. V. & JONGMANS, S.-S. 2018. Modular Programming of Synchronization and Communication amongTasks in Parallel Programs. *In 2018 IEEE International Parallel and Distributed Processing SymposiumWorkshops (IPDPSW).* IEEE.
VRBKA, J. 2016. Predicting Future GDP Development by Means of Artificial Intelligence. *Littera Scripta [online],* 9**,** 14.
WALLACE, A., RHODES, D. J. & WEBBER, R. 2017. Overseas Investors in London's New Build Housing Market.
WANG, C. & WU, H. 2018. A New Nachine Learning Approach to House Price Estimation. *New Trends in Mathematical Sciences,* 6**,** 165-171.
WANG, F., ZOU, Y., ZHANG, H. & SHI, H. 2019. House Price Prediction Approach Based on Deep Learning and ARIMA Model. *In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT).* IEEE.

WATSON, N. 2012. *Using Mean Absolute Error for Forecast Accuracy* [Online]. Contemporary Analysis. Available: https://canworksmart.com/using-mean-absolute-error-forecast-accuracy/ [Accessed 04/01/2022 2022].
WENG, B., LU, L., WANG, X. & M, F. 2018. Predicting Short-Term Stock Prices Using Ensemble Methods and Online Data Sources. *Expert Systems with Applications,* 112**,** 16.
WILHELMSSON, M. 2009. Construction and Updating of Property Price Index Series: The Case of Segmented Markets in Stockholm. *Property Management*.
WOLPERT, D. H. 1992. Stacked Generalization. *Neural Networks: The Official Journal of the International Neural Network Society,* 5**,** 241-259.
WRIGHT, R. 2018. Interpreting Black-Box Machine Learning Models Using Partial Dependence and Individual Conditional Expectation Plots. SAS Institute Inc.
XGBOOST, D. 2021. *XGBoost Parameters* [Online]. Available: https://xgboost.readthedocs.io/en/latest/parameter.html [Accessed 13/09/2021 2021].
XU, L. & TANG, B. 2014. On the Determinants of UK House Prices. *International Journal of Economics and Research,* 5**,** 8.
XU, Y.-H., XU, K.-X. & CHEN, Y.-Q. 2016. A Study on the Influence Factors of Real Estate Prices Based on Econometric Model: A Case of Wuhan. *DEStech Transactions on Social Science, Education and Human Science*.
YEARSLEY, J. 2015. *Quantile-Quantile Plots* [Online]. Available: https://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html [Accessed 25-Aug-21 2021].
YILDIRIM , M. O., GRIMA, S., ÖZEN, E. & BOZ, H. 2021. Financial Development and House Prices in Turkey. *Contemporary Issues in Social Science,* 106**,** 205-220.
YUSOF, A. M. & ISMAIL, S. 2012. Multiple Regressions in Analysing House Price Variations. *Communications of the IBIM*.
ZHANG, A. 2012. *Evaluating Machine Learning Models* [Online]. Available: https://www.oreilly.com/library/view/evaluating-machine-learning/9781492048756/ch04.html [Accessed 28/06/2021].
ZHAO, Y., CHETTY , G. & TRAN , D. Deep Learning with XGBoost for Real Estate Appraisal. 2019 IEEE Symposium Series on Computational Intelligence, SSCI 2019, 2019 United States of America. IEEE, Institute of Electrical and Electronics Engineers, 1396-1401.
ZHENG, X. & HAO, T. 2018. House Price Forecast Based on Dynamic Model Averaging Model Combined With Web Search Index. *In 2018 International Conference on Cloud Computing, Big Data and Blockchain (ICCBB).*

Appendix

Appendix 1: Year on year bottom three and top three London boroughs based on total number of transactions

Transfer Year	District	No of Transactions	District	No of Transactions
	Bottom 3		Top 3	
2011	City of London	254	Barnet	3,950
2011	Barking and Dagenham	1,465	Bromley	4,367
2011	Newham	1,617	Wandsworth	5,274
2012	City of London	197	Barnet	4,089
2012	Barking and Dagenham	1,350	Bromley	4,860
2012	Newham	1,707	Wandsworth	5,303
2013	City of London	470	Lambert	5,036
2013	Barking and Dagenham	1,820	Bromley	5,544
2013	Newham	2,446	Wandsworth	6,379
2014	City of London	415	Croydon	6,284
2014	Barking and Dagenham	2,388	Bromley	6,329
2014	Harrow	2,899	Wandsworth	6,365
2015	City of London	401	Bromley	6,412
2015	Islington	2,624	Croydon	6,472
2015	Barking and Dagenham	2,640	Wandsworth	6,573
2016	City of London	241	Barnet	5,855
2016	Kensington and Chelsea	2,341	Croydon	6,069
2016	Kingston Upon Thames	2,416	Wandsworth	6,447
2017	City of London	360	Bromley	5,286
2017	Islington	2,065	Croydon	5,948
2017	Kensington and Chelsea	2,254	Wandsworth	6,012
2018	City of London	518	Bromley	5,324
2018	Kensington and Chelsea	1,941	Croydon	5,402
2018	Islington	2,252	Wandsworth	5,557
2019	City of London	284	Bromley	5,122
2019	Kensington and Chelsea	2,083	Croydon	5,227
2019	Barking and Dagenham	2,118	Wandsworth	5,341
2020	City of London	188	Croydon	4,124
2020	Barking and Dagenham	1,564	Wandsworth	4,311
2020	Kensington and Chelsea	1,713	Bromley	4,522

Appendix 2: Total number of transactions per London borough between 2011 and 2020

District	Total No Transactions
City of London	3,328
Barking and Dagenham	20,640
Kensington and Chelsea	23,714
Kingston upon Thames	24,770
Islington	25,451
Harrow	26,386
Haringey	26,399
Camden	26,476
Hammersmith and Fulham	27,027
Brent	27,398
Hackney	28,035
Merton	28,684
Hounslow	28,789
Newham	30,018
Sutton	30,244
Redbridge	30,397
Waltham Forest	31,571
Richmond upon Thames	31,865
Enfield	33,535
Bexley	34,722
Ealing	35,000
City of Westminster	35,952
Hillingdon	35,993
Havering	36,870
Greenwich	37,577
Southwark	39,258
Lewisham	39,793
Tower Hamlets	43,163
Lambert	43,797
Barnet	47,843
Croydon	51,612
Bromley	53,433
Wandsworth	57,562

Appendix 3: Year on year bottom three and top three London boroughs based on total value of transactions

Transfer Year	District	Total Value of Transactions (£)	District	Total Value of Transactions (£)
	Bottom Three		Top Three	
2011	City of London	136,590,045.00	Wandsworth	2,680,364,086.00
2011	Barking and Dagenham	263,925,175.00	City of Westminster	3,166,848,519.00
2011	Newham	347,635,128.00	Kensington and Chelsea	3,226,933,368.00
2012	City of London	105,820,478.00	Wandsworth	2,823,256,160.00
2012	Barking and Dagenham	245,194,545.00	Kensington and Chelsea	3,408,991,846.00
2012	Newham	373,599,931.00	City of Westminster	3,727,171,842.00
2013	Barking and Dagenham	355,662,717.00	Wandsworth	3,805,371,806.00
2013	Newham	675,548,744.00	Kensington and Chelsea	4,918,113,270.00
2013	City of London	806,332,253.00	City of Westminster	6,132,699,096.00
2014	Barking and Dagenham	521,439,104.00	Wandsworth	4,511,802,047.00
2014	Newham	943,746,691.00	Kensington and Chelsea	6,334,326,267.00
2014	Sutton	1,148,246,614.00	City of Westminster	8,275,219,052.00
2015	Barking and Dagenham	729,321,999.00	Wandsworth	4,614,668,511.00
2015	Newham	1,204,284,842.00	Kensington and Chelsea	6,187,424,414.00
2015	City of London	1,279,860,601.00	City of Westminster	9,540,117,764.00
2016	Barking and Dagenham	774,329,643.00	Kensington and Chelsea	4,957,373,513.00
2016	City of London	1,276,612,837.00	Wandsworth	5,372,238,804.00
2016	Bexley	1,364,447,135.00	City of Westminster	9,945,387,772.00
2017	Barking and Dagenham	757,892,162.00	Wandsworth	5,049,731,939.00
2017	Bexley	1,393,458,777.00	Camden	5,120,315,512.00
2017	Kingston upon Thames	1,463,206,293.00	City of Westminster	11,474,690,254.00
2018	Barking and Dagenham	882,732,227.00	Wandsworth	4,713,773,244.00
2018	Sutton	1,323,028,052.00	Kensington and Chelsea	5,061,100,221.00

2018	Bexley	1,357,308,281.00	City of Westminster	11,653,320,507.00
2019	Barking and Dagenham	780,195,069.00	Camden	4,427,574,024.00
2019	Redbridge	1,378,190,203.00	Kensington and Chelsea	5,157,090,310.00
2019	Bexley	1,428,974,207.00	City of Westminster	9,700,904,307.00
2020	City of London	555,425,752.00	Wandsworth	3,629,473,998.00
2020	Barking and Dagenham	669,141,534.00	Kensington and Chelsea	3,949,856,465.00
2020	Redbridge	1,069,940,153.00	City of Westminster	8,193,502,379.00

Appendix 4: Total value of transactions per London borough between 2011 and 2020

District	Total Value of Transactions (£)
Barking and Dagenham	5,979,834,175.00
Bexley	11,402,257,646.00
Sutton	11,693,196,047.00
Redbridge	12,358,038,379.00
Harrow	12,404,612,652.00
Waltham Forest	12,610,773,242.00
Havering	12,820,475,521.00
Kingston upon Thames	13,281,177,160.00
Newham	13,340,908,426.00
Enfield	14,379,001,426.00
City of London	14,621,494,833.00
Hounslow	14,981,363,390.00
Brent	15,279,895,402.00
Haringey	15,512,162,433.00
Greenwich	16,174,664,517.00
Lewisham	16,573,319,052.00
Merton	16,737,653,443.00
Hillingdon	17,018,650,416.00
Hackney	17,968,007,269.00
Croydon	19,845,653,601.00
Ealing	20,691,686,420.00
Islington	21,162,772,251.00
Bromley	24,365,509,763.00
Richmond upon Thames	24,387,508,055.00
Hammersmith and Fulham	26,055,759,347.00
Lambert	26,252,394,525.00
Tower Hamlets	27,018,218,218.00

Southwark	27,536,200,436.00
Barnet	28,405,490,108.00
Camden	35,713,596,460.00
Wandsworth	41,384,401,292.00
Kensington and Chelsea	47,852,006,919.00
City of Westminster	81,809,861,492.00