City, University of London MSc in Data Science

Project Report 2021

Developing a property investment appraisal tool using alternative data sources and input from industry experts

Babak Hessamian

Supervised by: Dr Aidan Slingsby

02/04/2022

Declaration

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work, I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Babak Hessamian

Table of Contents

Table of Figures

Table of Tables

Abstract

The aim of this project was application-oriented in that it set out to create a valuation model to serve a business purpose with input from industry experts. A mixed approach was adopted combining qualitative (in-depth interviews) and quantitative (predictive models) data generation and analysis methods. Two objectives were defined: *Can the use of alternative data sources improve or match the price and rental prediction models in a property investment appraisal context?* and *How can the use of alternative data sources impact, and potentially benefit, the industry experts who use property appraisal tools?*. The interview themes were usability, predictive power, and explainability. These were address by building robust models using XGBoost algorithm, which was trained and tested with alternative data sources (not price-paid). The models predict the price and the rent for a given full postcode in the London Borough of Lambeth (predictive power). The best performing models were used in a prototype of an appraisal system. In the front-end, the features that were used in the model training were ranked and scored in order of their contribution to the prediction (explainability). The metrics displayed put the investment opportunity into context for the appraiser (usability). Creating more explainable tools for industry experts requires their participation in the development of those tools.

Keywords: Residential Property Investment Appraisal, Role of the Expert, Explainability, Qualitative with Quantitative, XGBoost

1 Introduction

1.1 Background

There are 2.7 million dwellings that are rented privately to 4.4 million households in the UK [1]. This sector is known as the Private Rented Sector or PRS, worth billions of pounds. Since the introduction of Buy to Let (BTL) mortgages in early 1990s, and the decline of social housing provision by the local councils in the 2000s, PRS has been almost entirely dominated by private landlords, whose number stood at 2.66 million in 2020 [2]. More recently residential property markets have been under scrutiny from the government, as with increased demand the rents may rise to quickly and by too much, making certain areas unaffordable. In 2016 the UK Chancellor of the Exchequer George Osborne introduced a series of measures designed to curb this. With an emerging Build to Rent sector (BTR), the UK's residential rental sector is changing with an increasing number of larger organisations getting involved. Most recently Lloyds Banking Group has set up its own rental homes brand Citra, with an aim to build a portfolio of 50,000 homes by 2030 [1].

With the professionalisation of the PRS sector and the emergence of the BTR as a sector, there is an accelerated need for a whole host of software, including property investment appraisal systems. The process of appraising an investment property or a property developing opportunity can be unstructured, made more difficult by an at-times opaque world of private firms. Gauging both rental and sales demand identifying the drivers of that demand is time-consuming and only done through subjective means, such as speaking to estate agents. Many firms rely on the 'gut-feeling' of a handful of individuals. These methods are broadly known as comparable models, where historic transaction prices, property characteristics, and subjective opinions take centre stage.

More recently there has been a push to go beyond the comparable model and harness alternative data sources to build house price and rental value prediction models. For instance, a McKinsey & Co research on predicting house prices found that the proportion of the predictive power attributed to features derived from proximity to points of interest (POI) and the quality of those POIs is between 26% and 32% respectively [3]. This was as opposed to 14% attributed to market performance (income, etc) and 18% and 12% to property performance (vacancy rate, etc) and property features (number of bedrooms, etc). This highlights the potential benefits of going beyond historic price data.

1.2 Objectives & Beneficiaries

Too often the research on technology in a real estate context focuses solely on building the best model. While this is essential for developing better valuation and appraisal models, the input of the industry experts should be carefully considered. This project's aim was to bridge the gap between industry and research by focusing on appraisal from a property investor's perspective. It does not seek to build a tool for potential homebuyers and home movers to help them in their decision making.

The property investment decision process can vary from one investor or investment manager to the next. It can be used: (1) deciding which portfolio to hold, (2) deal sourcing, (3) purchase decisions, (4) hold/sell decisions, (5) asset management decisions, (6) use of financial leverage [4]. In each of these stages knowing the value of the property or the portfolio of properties is key. Once the market and rental values are predicted, it can be used in helping the investor decision makers in several of the elements mentioned above.

The first objective (OB1) to be addressed was:

Can the use of alternative data sources, as opposed to price paid datasets, match, or improve the performance of price and rental prediction models in a property investment appraisal context?

The second objective is more focused on the role that property professionals who currently use property investment appraisal tools played in this project. The interviews guided the project throughout and is mainly based around the objective of finding out (OB2):

How can the use of alternative data sources impact, and potentially benefit, the industry experts who use property appraisal tools?

Outcomes: a prototype of a tool, developed with contribution from industry experts, that takes property information as input and outputs valuable information for the decision-makers to aid in the property investment appraisal process. The contribution of this project is less methodological (i.e. to compare and seek to improve machine learning algorithms that are used to predict property values), and more empirical (i.e. addressing the needs of property investment decision-makers).

Beneficiaries: Real Estate Advisory and Investment firms to serve their existing and future client base better. Property decision-makers looking to identify assets in low value areas that are rising in popularity, mortgage lenders that are looking to build a risk profile for a property in a specific area, those interested in the predictive power of alternative data in building models. All property experts who seek a more transparent and less 'black-box' approach to the use of technology in the real estate sector.

1.3 Methods

Two data gathering and analysis methods were used that can broadly be described as qualitative and quantitative. The outcome of the analysis was shown in a front-end that was deployed on to the cloud.

The intention of the literature review was to find the state-of-art models to use in a price and rent prediction task while incorporating the results of in-depth interviews with industry experts, which were in turn informed by the literature review in qualitative data analysis methods. This was then showcased in the form a simple a workable front-end tool deployed to the cloud with.

Qualitative: The qualitative phase was carried out before the qualitative phase. To avoid subjectivity, systematic data gathering, and analysis methods were followed. This project has not sought to develop new techniques or analyse data in a new way. The contribution to knowledge has come from turning data from in-depth interviews with experts into an actionable tool with the aim of bridging the gap between highly technical teams and non-technical teams.

Quantitative: In the quantitative phase the interviews were analysed using methods identified in the literature review to minimise subjectivity. State-of-art prediction models were explored and tested to be used in the price and rent prediction tasks while incorporating the results of the in-depth interviews with industry experts. This project is not a comparative study of different machine learning models. The intention of the literature review was to.

Front-end: The front end did not focus on user experience design, but on deploying a simple prototype that demonstrates the findings of the qualitative and quantitative phases to allow for discussion and reflection on the approaches used. The focus is much broader and includes issues around the transparency, explainability, and usability.

1.4 Work plan

The work plan was divided into the following segments:

- Conduct in-depth interviews with experts
- Perform a systematic analysis of the interviews to devise themes and categories
- Using the themes and categories:
    - explore, identify, and evaluate data sources
    - feature engineer using a selection of the data sources
    - run state-of-the-art models identified in the literature review with a combination of features
    - design and deploy an actionable tool using the best performing models

It is important to reiterate that the outcome of the qualitative phase informed the quantitative phase. Documenting the findings and iterating from one stage to another were done throughout the process.

1.5 Report structure

In **Section 2** the report puts the two objectives of the projects into context. The body of literature on automated valuation models is explored along with that of qualitative methods of research in information technology. Research on property centric and machine learning approaches to valuing residential property is discussed and the role that choosing data sources plays in this project is explored.

The literature review was split into three sections. Since valuation is a key part of investment appraisal, a review of the literature on how different valuation models have been adopted and assessed was carried out (2.1). To relate the literature to the research question which partly focuses on the use of different data sources, a review of the literature on data availability and accessibility in the context of property valuations was conducted (2.2). As the project also addresses the real business needs of industry experts, a review of the literature on analysing qualitative data (in-depth interviews in the case of this project) was also conducted (2.3).

In **Section 3** the methods that were used to address the two objectives are explained. The section is split to three sub-sections of qualitative, qualitative, and prototyping. In the qualitative sub-section (3.1), the details about the approach to conducting and analysing the in-depth interviews with experts in the field of property investment is explained. In the quantitative sub-section (3.2), the data and feature extraction, feature engineering, and modelling pipelines that were used in this project are explained. The prototyping sub-section (3.3) explains how the best performing prediction models and the data sources they use are translated into a usable property investment appraisal tool that is deployed on the google cloud platform.

Section 4 includes the results of what was produced after using the methods from section 3 to address the objectives of this project. The themes and categories from the interviews are explored. These informed the quantitative phase of the project; the final dataset used, and the results of the best performing models are given. The prototype built was the culmination of the quantitative and qualitative phases; a through explanation along with screenshots and model evaluation metrics is presented.

In **Section 5** the degree to which the two objectives were addressed is discussed and reflected upon. And finally, **Section 6** concludes the report by answering a key question and suggesting future work.

2 Context

In this section the aim was to identify what research has been done in the field of property price valuations, qualitative approaches in the context of information technology. The weaknesses of the research were identified, as well the methods that were used in this project. In section 3 more details are given of how this project is put in the context of this literature review.

This project focused on the appraisal methods for residential property investment decision makers. The task for these individuals or groups of individuals is to determine an investment value for a given real estate asset [5] . Investment value was defined by the International Valuation Standards (IVSC) in 2019 as:

> …the value of an asset to a particular owner or prospective owner for individual investment or operational objectives. [6]

The *individual investment* or *operational objectives* in this definition is subjective and dependent on the investment criteria of the individual investor or investment company. Few of these objectives, include deciding whether to hold an asset or sell it, or whether to use financial leverage for existing assets or new purchases. On a basic level property investors and investment companies can benefit from the difference between investment value and the market and rental value of an asset [7], which can have indications for potential gains or losses. Hence the key component of appraising a potential investment opportunity is determining the market value, or the price, and rental value, or the rent, for a given property. Once these two elements are determined, the investment value will be decided according to the needs and expectations of the investors, whose input played the key role in guiding this project.

Traditionally the Direct Capital Comparison (DCC) has been used to value residential property and by accredited surveyors. Valuation by DCC can be broken down into four steps: select comparables and extract, confirm and analyse comparable sale prices, adjust sale prices for noted differences, formulate an opinion of open market value for subject property, present the results in a report [8]. This process is shrouded in mystery. The DCC valuation process is commonly known as the comparable model amongst industry professionals and is sued by estate agents, land buyers, and property investor decision makers. After observation of and discussions with valuers, Jenkins et al [9] compare the DCC valuation process to a black box since the four steps are not committed to paper. Diaz et al. [10] thought numerous studies has found that valuers do not always follow the correct procedure which may result in anchoring, which is when the valuer arrives at an estimation for the valuation by adjusting from an initial starting figure [5]. Gallimore [11] found valuers showing precipitance, in other words forming an opinion very early on in the process rather than following correct procedure and keeping an open mind until the end of the process. All this is evidence to the subjectivity of the valuation process using DCC, which is prone to be systematically biased and prone to uncertainty and lagging [5].

In the UK valuations are carried out by professional surveyors, who are often accredited members of the Royal Chartered Institute of Surveyors (RICS). These often use the DCC method mentioned above. RICS is a UK-based global professional body that published the RICS Valuation Global Standards ('Red Book Global Standard') which contains mandatory rules, best practice guidance and related commentary for all members undertaking asset valuations. The professional surveyors may act for a purchaser, vendor, mortgage lender, or any other institution with a stake in the purchase. In the December 2021 independent review of real estate investment valuations, Peter J. Pereira Gray [12] mentions The Red Book states that the role of valuer does not go beyond providing an assessment of the exchange price at which an asset would likely sell or be acquired. They also need to provide suitable supporting evidence and a rationale for their decision. Crosby et al [13] find that the RICS supports one method while the financial institutions, major property companies and their advisors, in other words property investment decision makers, seem to prefer an alternative approach.

In principle this project assumed a widely adopted convention in the property valuation literature [8]:

*price is the actual exchange price in the marketplace
*value is an estimation of that price were the property to be sold in the market

The literature on valuations revealed the clear need for a more robust data driven approach toward valuations that is less subjective and more cost-effective while being more tuned to the nuances and complexities of the property market and all that affects it. This was the direction of the next step in the literature review which revealed the concept of Automated Valuation Models, or AVMs, which is explored in 2.1 below.

It is reiterated that the role of the expert is key in the investment decision making process. Qualitative approaches were explored as a potential method to gather and analyse data that captures the views of the experts. Quantitative data analysis uses mathematical approaches and statistics to address the research questions. Qualitative data analysis uses themes in the words people use to talk about a certain topic [14]. This project uses both analyses to address the research questions. The literature for the quantitative approaches that are relevant to this project fall under AVMs and are explored in 2.1 and the literature for the quantitative approaches that are relevant to this project are explored in 2.3.

2.1 Automated Valuation Models in Real Estate

To put the term AVM into content it is possible to think of what each term is referring to. *Automated* related to the use of BigData, analytical statistics, and IT integration. *Valuation* refers to estimating and predicting the market or investment value, or rental value of a given property or a group of properties. *Model* refers to the methods and techniques used to determine the value. The outputs of an AVM is often interpreted by an expert with knowledge of the property market who can consider its possibilities and limitations. So, predicting house price and rental value is the key component of AVMs. Existing work on this topic can roughly be split into two approaches: property feature centric and machine learning centric [15].

2.1.1 Property centric approaches

Properties are inherently heterogenous, meaning no two properties are the same, making valuations more complicated. One way of overcoming this problem is to decompose properties into different characteristics or features. This approach assumes that property value is driven by property characteristics such as the number of bedrooms and bathrooms, square footage, the year built etc. The comparable approach seeks to compare properties with similar features of a known sold price to from an opinion on the value of the property. A commonly used model that adopts this approach is the hedonic regression model [16]. Rosen [16] in 1974 first applied hedonic regression in real estate research. Since then it has been studied in the context of what property features influence the price [17] [18] [19]. Some literature on hedonic regression in the real estate context takes locational attributes into consideration as well as property characteristics. These locational attributes are often incorporated into the models as the distance to nearest amenities [20].

These models rely on pure statistical methods such as ordinary least squares (OLS) that rely on price paid data and assume linearity. As such they fail to capture the heterogenous aspects that property market. The equation below shows a typical such model:

$$V(P) = \eta + \sum_{1 \leq i \leq n} \beta_i \times f_i(P) + \epsilon(P)$$

where V(P) is the value of property P, which depends n the corresponding property features [$f_1(P)$, ..., $f_n(P)$], β_i is the "i-th" coefficient, $\epsilon(P)$ is some correction to be applied to this particular property, and η is a property-independent correction that applies to some application or geographical context.

There are shortcomings associated with the property centric approach. Firstly, the earlier versions of the main property centric approach, hedonic regression, do not consider the spatial dependency of the variables. There has been attempts to consider the spatial awareness in hedonic pricing models, but the research has found that spatial dependence cannot be handled well in regression residuals [21] [22]. Secondly, hedonic regression assumes linearity between the features and the target variable, the property value and fails to discover complex relationships between variables [23] Thirdly, some research has found that the results can lack robustness [23] [24]. Some of these limitations have been overcome by applying machine learning and using a range of datasets to better predict the value of a property.

2.1.2 Machine learning approaches

Machine learning algorithms have become increasingly popular in the past few years and applied to regression and classification problems. Property valuation has also been researched substantially. This research shows there are huge benefits to using machine learning algorithms in estimating property value and rental value over the conventional models such as hedonic regression. Machine learning models allow the handling of large amounts of data containing variable with often with complex and nonlinear relationships. Some of these variables were not accessible by traditional methods as they lacked quantitative measurements [25] (such as user reviews of nearby restaurants). They also do not always require the training data to be normally distributed. Hyperparameter tuning can be used in an iterative process like grid search in the case of XGBoost to improve the models at speed given the advances both in algorithm and computational power development. All the above has allowed to drive down the estimation error in models that are built.

There are learning algorithms that have been used to spot trends and predict price. For example Baldominos et al [26], who use four techniques of support vector regression, *k*-nearest neighbours, ensemble of regression trees, and multi-layer perceptron to identify opportunities in the housing market, defined as those houses listed on the market at a lower price substantially lower than the market price. Satish et al [27] use linear regression, LASSO and gradient boosting algorithms, to predict property prices. Their results show LASSO outperforming other algorithms in terms of accuracy.

There are learning algorithms that, as well as being able to predict the price, allow the measurement of the importance of different features such as Random Forests [28] [29], and Gradient Boosting [30]. This is of particular interest to this project given the qualitative analysis phase and the research objectives, hence this section following on such algorithms. A deep-dive was carried out into the working of these algorithms and examples of their use in the literature studied to help determine which one was most appropriate for this project.

Random forest is a supervised learning algorithm with an ensemble learning method. It works by combining n decision trees and can be used for classification as well as regression problems. Decision trees lack accuracy and are prone to overfitting if they grow deep [Elements of Statistical Learning Second edition 2009 by Trevor Hastie] [31], in other words if the number of decision nodes. Random forests seek to reduce this high variance by averaging the learning from multiple decision trees that have been trained on different parts of the dataset to make a final prediction. The different parts of the dataset that the training is based on is determined by a random sampling with replacement from the training data, which is called bootstrapping. The aggregation of the bootsrapped dataset which is performed many times and training an estimator for each bootsrapped dataset is called Bootstrap Aggregation or bagging [27].

Boosting is an ensemble machine learning algorithm with the aim of making weak learners perform better [Kearns and Valiant 1989] [32]. An early example of a gradient boosting algorithm is Adaptive Boosting or AdaBoost which uses decision trees as weak learners and works by adding weights to each decision tree considering their prediction accuracy [33]. It then makes a final prediction by majority vote for the learners' prediction that were weighted according to their individual accuracy.

Gradient Boosting was introduced by Friedman [30] by developing this concept further where the weak learners are added using a gradient descent procedure. New weak learners are added one at a time while the existing weak learners are left unchanged, making this machine learning technique a stage-wise additive model. The trees added with each iteration seek to correct the prediction errors of prior models. With each iteration the model uses the error rate to compute the gradient of the error function and adopts the gradient to tune the parameters and reduce the error rate for that iteration [27]. Since Friedman views the function estimation task as a numerical optimisation problem, gradient boosting can be used for regression and multi-class classicisation predictions, as opposed to only binary classification problems.

Extreme Gradient Boosting, or XGBoost, was first introduced by Tinaqi Chen [34]. It is classified as a boosting integration model since it combines gradient booting algorithm and decision trees. These models are called tree boosting algorithms. In addition to aggregating the accuracy scores of previous learning trees to reduce errors, XGBoost has many other features like parallel and distributed computing, allowing it to achieve faster execution speed and superior model performance [35]. XGBoost seeks to build a group of weak regressors. The formula includes a penalising function that keeps the leaf scores small and minimises the number of leaves.

An example of literature that has found Random Forest (RF) to outperform linear regression is [36] . Other research has also shown RF has proved a robust algorithm with accurate predictions of property prices. For instance, Koktashev et al. (2019) [37] use number of rooms, total area, floor, parking, type of repair, number of balconies, type of bathroom, number of elevators, garbage disposal, year of construction as attributes and compare the performance of ridge regression, linear regression, and random forest in predicting property prices. Based on the mean absolute error (MAE), they find that random forest gave a better performance. Lacosta et al. [38] found that random forest outperforms random subspace while Borde et al. [39] compared the performance of linear regression, k nearest neighbours regression and random forest with random forest being the best performer. To predict rental values, Hu et al [23] used random forest, extra-trees, gradient-boosting, support vector regression, multi-layer perceptron neural network, and *k*-nearest neighbours. They concluded RF and extra trees to be the best performing models.

XGBoost has been used in a variety of price prediction contexts, for instance predicting crude oil prices [40]. Peng et al. [41] find XGBoost to be the best algorithm in their study of comparing multiple linear regression, decision trees and XGBoost. De Nadai et al. [42] researched the economic impact of neighbourhood characteristics on property values and found XGBoost algorithm to produce highly accurate predictions of house prices. Also De Nedai [42]: Moreover, XGBoost allows to "open" its black box, by explaining variables' importance and their contribution to the predicted values.

Another important benefit of using machine learning is the ability to leverage location centric, or geospatial, data and consider the spatial dependency between residuals in the regression models [43]. By geospatial it is meant the location centric data in the context of the property that is being valued, in other words asking questions such as "how close is the nearest restaurant with the best rating?". This question not only quantified the vibrancy of the neighbourhood, but it also seeks to quantify the quality of the POI as well. Bourassa et al [44] define spatial submarkets taking spatial dependence into account using principal component analysis and cluster analysis and conclude that location plays a key role in explaining the reasons why housing submarkets matter when it comes to price prediction and market analysis.

Considering the geospatial aspect of research into machine learning approaches to property price predictions, Cortright et al. [45] found a positive correlation between walkability and housing prices in almost all the analysed US cities. Several studies have measured the level of influence of the distance between the closest POI and a property on the price [46] [47]. Other studies have explored the quality of the POIs as an influencing factor on house prices [47]. Both geospatial proximity and POI quality were of particular interest in this project.

A novel method is presented in [15] which the authors call geo-spatial network embedding. The method uses graph neural networks in the geospatial context of the points of interest and their quality. This is very close to what this project has set out to achieve. However, the authors do not look at building models based on the input from the industry experts. In fact, they do see feature engineering and property appraiser input as an impudence. This is exactly the opposite of the approach of this project that puts the appraiser and expert input at its core.

In short machine learning approaches prioritise predictive power over inference since the algorithms can model complex non-linear relationships. However, the interpretation of the outputs only are useful if they add value to the experts who use them.

2.2 Data sources

Majority of the models that form part of the appraisal tools not only are propriety, but they may be using data that is not publicly available. For instance, in a company that owns a thousand residential units there will be internal data relating to rental price and vacancy rates which are both very important to appraising future investment opportunities. Despite this, a review of the literature revealed a variety of often open sources of data that have been researched in the context of property price prediction. These were sought out considering the results of the qualitative research phase of this project.

Park et al [48] use school ratings and mortgage rates as additional features to enhance their predictive performance. Pavloc et al [49] develop features using recent immigration data to find the effects of immigration on property prices. Rafei et al [50] build a novel model using a wide range of economic features, such as wholesale price of building material and stock market index. Brooks et al [51] investigate the effects of sexually oriented businesses on house prices. These modern learning algorithms can effectively capture the spatial dependence from location attributes which highlights the need of explicit feature engineering.

POI is unstructured data, and as such feature engineering had to be carried out. Thinking in the geospatial context variable based on factors such as "distance from the nearest subway station", "distance from the nearest school", "number of shops within one kilometre", and so on as mentioned in [52]. the POI data was used with ratings and pricing bands as the assumption is that different amenities will have different levels of impact in the vibrancy and desirability of neighbourhood [53] [54] [55] [56], which in turn affects the property and rental values.

De Nadai et al. [42] used multi modal data from OpenStreetMap, Good Street View, property tax, census data, and listings data to train XGBoost algorithm that reached a very high accuracy in prediction house prices. They found that using a varied data source lowers the model performance by 60%. Lacosta et al [38], Borde et al. [39] and Park et al. [48] used a dataset of 5,000 properties. Perez-Rave et al. [57]

use a larger dataset of 60,000 properties use regression trees and achieve similar results to hedonic price models. All this demonstrates that the size of the data source is dependent of the algorithm.

The data for the qualitative analysis phase of this project was generated by conducting in-depth unstructured interviews with property professionals. Details of how this data was gathered is given in section 3.1.1. below.

2.3 Role of experts

While algorithms can be powerful, it is important that the outputs are actionable insight into the data [58], which are understood by experts who will use the outputs to make decisions. The professionals play a key role in the decision-making process behind property investments. They may make decision on: (1) deciding which portfolio to hold, (2) deal sourcing, (3) purchase decisions, (4) hold/sell decisions, (5) asset management decisions, (6) use of financial leverage [4]. Increasingly their decisions are made using tools developed by technology companies. This is the junction where the industry experts meet technology. There are several UK and US based companies that provide an appraisal system, such as LandInsight, REalyse and Yuvoh Analytics, Foxy AI, StreeBees, Credifi, SpaceQuant, Skyline AI, Mashvisor, Oreeva. Access to the models and algorithms of Automated Values Service providers is limited due to the proprietary nature of these software. To understand the relationship between these tools, often developed by highly technical teams, and the professionals who use them, the decision was made to do interviews of professionals who use the tools. As such a review of the literature on gathering and analysing quantitative data was carried out.

The quantitative approach is generally based on a preconception that knowledge of the research questions was situated with professionals and could be best accessed by personal interview or 'close dialogue' [59].

Oats has written extensively on different approaches to gathering and analysing qualitative data in the information systems and computing [60]. Oates [14] mentions the benefits of qualitative data and its analysis as being rich and detailed as well as possibly revealing alternative explanations, while acknowledging that the interpretation of the data being closely tied to the researcher compared to quantitative data analysis. As such the background of the author of this project becomes relevant. The author has experience in the property sector as an investment advisor to several High Net Worth Individuals (HNWIs) with substantial UK residential property portfolios. It is also noteworthy that a major problem with analysing qualitative data, as Oates points out, is that sometimes it looks as if conclusions appear by magic. This highlighted the need to have a robust way of analysing the data from the interviews.

The ‘grounded theory’ approach relies on sequential discovery whereby details such as sample size, and data generation methods are not determined in advance. Instead, researchers are open to discovering new themes and patterns in the data and without any pre conceived theories or ideas in mind. This project did not adopt the grounded theory approach as the interviewees were selected in advance as opposed to doing the research and collecting data to allow a theory to emerged form field data.

Another body of literature that was considered to aid with linking the qualitative phase of this project with the quantitative phase was ‘expert systems’. Expert systems can be defined as a programme that uses knowledge and inference procedures to solve problems that are difficult enough for the machine to need human expertise. It can be a very complex topic involving decision theory, fuzzy logic, uncertainty theory, and more specifically to the property industry, portfolio theory where models are used to suggest allocating and pricing of assets [61]. Expert systems don’t come in one size fits all as they do require some knowledge of modelling on behalf of the appraiser**.** After reviewing the literature on expert systems, it was concluded that this project does not fall under this umbrella and exploring all the theories mentioned shifts the focus away from the objectives.

Studying expert systems and mass appraisals in the real estate context, Kilpatrick [62] finds a the disjoint between academic research and practice in the property industry. He points to the appraisal methods having developed heuristically on a ‘two-track system’ where academics do research using state-of-the-art models and techniques, but experts rely on professional organisations like RICS for methodological guidance. This is in contract to professions such as accounting where accountants have a integration between their academics and practitioners. One of the objectives of this project is trying to bridge this gap by ‘listening’ to the experts in developing a tool.

The purpose of analysing qualitative data is often to abstract patterns that are relevant to the research focus. This can involve using quantitative methods such as counting the number of times a word appears in a body of text. Oates acknowledges that the while quantitative data can generate robust research conclusions, qualitative data often relies on the researcher’s ability to extract the themes. However, there are some inherent problems with this approach which Barton [63] points out in the context of finding gentrified neighbourhoods; ‘qualitative strategies for identifying gentrified neighbourhoods may overlook areas that experienced similar changes to those more widely recognised as gentrified’. As such extra caution has to be taken and robust methods followed to avoid bias.

3 Methods

The aim of this section is to explain in detail the processes that were implemented in relation to the research objective of the project. Two data generation methods were used to corroborate the findings and enhance their validity. Oates [14] refers to this as method triangulation. The first was a qualitative approach which included five interviews with industry experts. The second was a qualitative approach of gathering data and [64] feature engineering to build models. The first approach informed the second process. The results were incorporated presented in a usable prototype of a property investment appraisal tool.

3.1 Qualitative

This project has not sought to develop new techniques or analyse data in a new way. The contribution to knowledge has come from turning data from in-depth interviews with experts into an actionable tool with the aim of bridging the gap between highly technical teams and non-technical teams. As such the qualitative research phase is a core part of this project. It was carried out before the quantitative phase. The tools of appraisal and AVMs are only useful if they can be used effectively and add value to the workflow of their users. As such the views and knowledge of four experts were sought. To avoid subjectivity, we will need to be systematic in the methods used. The approach used is described in 3.1.1 and 3.1.2.

3.1.1 Interviews

The interviews were on average one hour long and recorded on the author's mobile phone. They were scheduled for a set time and conducted in the private residence of the author to minimise any distractions. The consent form was presented and signed by all the interviewees. They all wish to remain anonymous with the recordings and transcripts of their interviews remaining private. Each interview was then transcribed in a separate Microsoft Word document. Another word document was created with all the five transcripts included to allow analysis. Copies were made of all the word documents to prevent data loss.

The data generation method for the qualitative phase of this project was in-depth interviews for two reasons; 1) the aim was to obtain detailed information from industry experts, of whom there are a limited number, 2) the questions were often complex and open-ended with the order being different for different interviewees. Initially three scaled questions were to be included in the interview schedule to gauge the importance of appraisal systems within the property investment sector, however this was scrapped. First, the importance of the systems became clear while conducting the literature review (section 2.1). Second, the number of interviewees was below 30, so no meaningful statistical analysis (e.g. calculating the mean etc.) was possible [14].

Five individuals with extensive experience in the real estate sector who perform property investment appraisals were interviewed. The reason for a small sample was two-fold: (1) the questions that were asked were detailed, and (2) there are a limited number of experts in the field of residential property investment appraisal with access to them being limited. The format of the interviews was semi-structured where a list of 10 questions and 4 themes were used by the interviewer to steer the flow of the conversation. This format was chosen with the intention to allow the experts speak freely and allow the author to possibility to discover new themes not included in the interview schedule.

Table 1 shows the interview questions, the 4 purposes, and the objective that each purpose addresses: 1) Current practice, 2) Current appraisal tools, 3) Understanding what is useful to investment professionals, 4) Establishing areas of improvement with current tools. These 'purposes' were chosen to guide in the interview process, but not necessarily for the analysis of the interview data. The methods used to carry out that analysis is described in section 3.1.2.

As a reminder the objectives of the project are:

> OB1: Can the use of alternative data sources, as opposed to price paid datasets, match, or improve the performance of price and rental prediction models in a property investment appraisal context?
>
> OB2: How can the use of alternative data sources impact, and potentially benefit, the industry experts who use property appraisal tools?

A brief background and description of each of the interviewees is given in this paragraph. They were chosen to represent different 'types' of professionals involved in the UK residential property investment ecosystem. They all use various tools to appraise potential investments, and are leaders in their own 'type' of investor/appraiser. The interviewees are referred to as IA, IB, IC, ID, IE and a brief background for each one is given below. The reasons and the outcome of the interviewee selection choices are discovered in section 5, discussion.

#	Question	Purpose	Objective
1	How do you appraise potential investments?	Current practice	OB1
2	What data do you rely on when appraising a potential investment?		
3	If you use tools to appraise potential investments, what do you like about them	Current appraisal tools	OB1
4	What do you think is missing from the tools/software you use?		
5	How important is being able to tell what data is driving the appraisal tool's decision making process?	Understanding what is useful to investment professionals	OB2
6	What information do you include in your presentation of your appraisal results to your team or outside investors?		
7	What do you do to keep track of the effects of the different data sources on your appraisals?		
8	What qualifiers/metrics might help you weigh up investments better?	Establishing areas of improvement with current tools	OB1 OB2
9	What kind of problems do you usually encounter when appraising potential investments?		
10	What would you like to see improved in the appraisal systems and processes?		

Table 1. Interview questions, purposes, and objectives

IA is the head of acquisition and investments for a small property company that connects private high net worth individuals (HNWIs) with property developers who need funds to buy and complete property development projects. They specialise in deal origination, financial modelling, valuation, due diligence & financial analysis and studied a MSc in Real Estate: Planning & Development, Valuations & Funding, Asset Management. They personally use various valuation and appraisal tools and are part of duo who make the final decisions

IB is the founding member and advisor of a new fund help investors to acquire, manage and dispose of UK residential property, with a focus on sustainability as well as profits. Previously they developed the strategy and built the seed portfolio for a HNWI backed fund targeting a £100m+ housing portfolio, was involved with c. £2bn+ transactions as a Strategist at a Big Four (the nickname used to refer to the four largest accounting firms in the United States, as measured by revenue: Deloitte, EY, PwC, and KPMG) and studied real estate at the University of Cambridge. They personally use valuation and appraisal tools and negotiate the pricing for such tools and are part of duo who make the final decisions. Their motto of their company is property consultancy with a social conscience.

IC has been involved in the Student Accommodation market, Private Rented Sector (PRS) and Specialist Supported Housing (SSH). They have worked with universities and councils throughout the UK to develop and deliver quality accommodation on transactions worth £150m+ and have a small team of employees, who carry out valuations and appraisals using various tools, reporting to them. They are solely responsible for making the final investment decisions.

ID is a founder of a property data analytics company that offer an appraisal tool as a service. Before founding the company they worked for two large property developers as portfolio and development analyst. Their company provides data and analytics in the residential property sector via a web application. Their services includes an AVM which uses billions of data-points and the most extensive real estate database for the UK. They are very numbers and data oriented with no time for ‘gut feeling’. They do not make investment decisions but run the strategy of all the data and analytics that are used in their appraisal products.

IE has worked in the hedge fund sector for the past 25 years and been investing his own personal funds into UK residential property since he began work. They have built a large private portfolio of properties across London, specifically in Lambeth. They are financially astute and their decisions have been made throughout the years through ‘gut-feeling’, as opposed to derived from robust data analytics or using appraisal tools. Including this interviewee adds a non-technical perspective that may inform the purely qualitative aspects of successful investing in UK residential property. They solely make the final decisions to do with investment.

3.1.2 Analysing the interviews

An attempt to carry out a thematic analysis was made; however, this was deemed not useful due to the limited number of interviewees. Instead, the decision was made to focus on the depth of the interviews and combining this with the knowledge of the authors to draw out themes and categories.

The analysis began by reading through the transcripts and using three colours to highlight the text and devise three initial themes: irrelevant sections were greyed out, yellow was used for segments that relate to the general discussion about the use of technology in the property sector, and segments that were directly relevant to data and appraisal tools were highlighted green. Majority of segments were no more than two sentences.

Each relevant segment highlighted green was then looked at closely and divided into four categories. The categories were those used by respondents and those that occurred to the author while reading the transcripts. As such an *inductive approach* was adopted to devise the categories, as defined by Oates [14]. The categories were changed and refined after several readings of the transcripts and copying and pasting all the texts from each category into one merged document to pick out inconsistencies. The number of times a phrase was mentioned was counted in these merged documents to confirm the existence of the themes and categories. The final four categories are presented in section 4.1 below, along with an explanation of the themes.

Amongst other suggestions for evaluating qualitative data analysis, Oates says asking questions such as is if alternative explanations can be explored for the data. It is also important to recognise the limitations of the qualitative data analysis phase. These points are address in section 4.1 below.

3.2 Quantitative

Objective 1 of this project revolves around the use of different data sources to build a useful tool for the property investment decision makers. This was going to be informed by the qualitative phase which sought their input in this process. After interviews were conducted and analysed the decision was made to focus on three independent variables, i.e. features that were going to be predicted using machine learning algorithms. These were the mean *"price_per_sqf_price", "rent"* and *"price"* of a given property in a postcode. This is further discussed in sections 4 and 5 below.

The focus was on one borough of London for the quantitative research phase: London Borough of Lambeth. This was for two reasons. First due to the density of the data making it more computationally expensive to do analysis on all the boroughs. For the same reason of high data density, we found that enough data was present in the highly populated borough of Lambeth for processing and building models. Second, the author has domain knowledge through appraising investment properties in the borough of Lambeth and is aware of the wide range of property types with varying price points, which allowed for back testing of models' performance.

Location centric, or geospatial, factors in appraising property price and rent emerged as an important factor in the review of the literature as well as the in-depth interviews. As such the bedrock of all the data and modelling pipelines in this project based on geospatial factors with proximity being a key consideration. To achieve this latitude and longitude data was needed for all variables. Throughout this project full postcodes were used. UK postcodes are described as full (e.g. 'W14 9JH'), district (e.g. 'W14') or sector (e.g. 'W14 9'). UK postcodes allowed for accommodating the need for granularity as well as the need for a common geospatial reference to link all the data sources together and build a coherent data pipeline.

A consideration was given to use the Lower Layer Super Output Area, which is a geographical grouping provided by Office for National Statistics (ONS) [65] containing between 1,000 and 3,000 inhabitants living in between 400 and 1,2000 households [66]. After an initial investigation it was decided to use postcodes instead for two reasons. First, it would have been impossible to aggregate the data from the other chosen data sources as some of the attributes are only available at full postcode level. Second, the

LSOA for London still use the 2001 census boundaries as opposed to the latest 2011 census. This could have made aggregation more difficult if the plan was to visualise the datapoint on a map and calculate geographical distances.

3.2.1 Data acquisition pipeline

The data acquisition pipeline is defined by a set of data acquisition processes, one for each selected data source and shown in **Error! Reference source not found.**. Those processes were defined by scripts to load data to a local repository from different APIs and external datasets. The data acquisition pipeline feeds a feature extraction pipeline, a feature engineering pipeline and the data modelling pipeline described in sections 3.2.3 and 3.2.4 respectively.

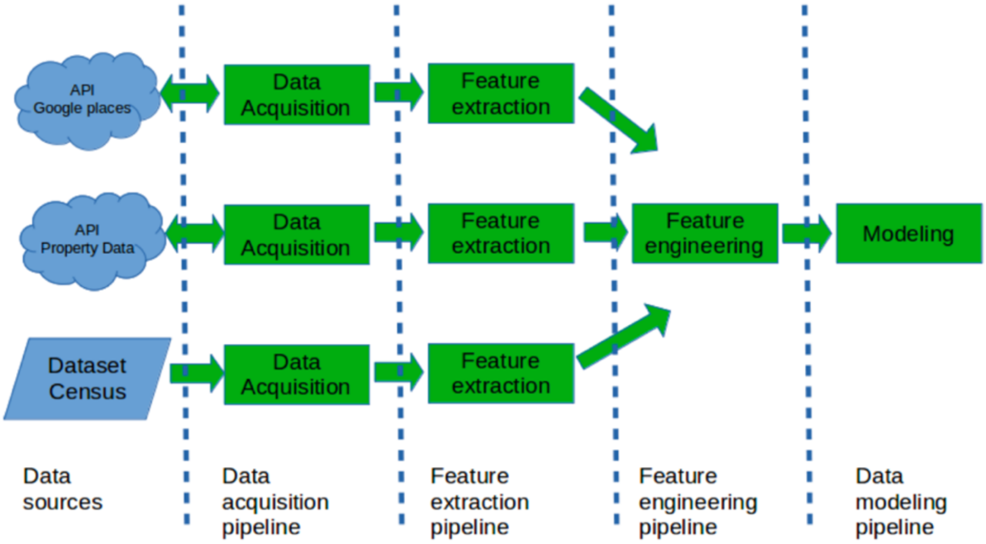


Figure 1. Overview of the data acquisition pipelines

The feature extraction pipeline is used to select and process the loaded data and it is specific for each data source. It also consists of a set of scripts used to aggregate and transform the original variables. The processed data is integrated into a single dataset through the feature engineering pipeline to be used by the machine learning algorithm in the modelling pipeline, described in section 3.2.4.

3.2.2 Data sources

The datasets were chosen bearing in mind the four categories that emerged in the interviews. The categories that emerged from the interviews are discussed and evaluated in sections 4 and 5. For this section it will suffice to mentioned that they informed the data source selection. The grouping in this section do not correspond to the four categories as some were combined (e.g. census incorporated into ‘demographics’, or EPC used to develop ‘housing stock’ variables). In section 3.2.1 the data acquisition and the feature extraction pipelines used in this project are described. The data acquisition explanations are mainly around the methods used to access the data, such as connecting to APIs. The feature extraction pipeline descriptions are around which features were used from the datasets.

In Figure 1 the two APIs and two websites that were used as the main data sources in this project are shown: GooglePlaces API [67], PropertyData API [68], Energy Performance Certificate (EPC) data, and UK Census data. All data sources used are publicly available, however the first two, the APIs are not free, while the last two, the websites, are free. The payments necessary for API calls were made by personal means of the author of this project. According to the Google Cloud Platform dashboard a total 11,785 chargeable calls were made to the Places API at a cost of £197.42. This was for 21 of the 98 'types' available form this API. For PropertyData API a 15,000 calls per month plan was chosen at a cost of £80 for one month. 7,000 calls were made to build the data ac pipelinquiesit.io n and feature extraction pipelines. To minimize data acquisition costs, only API's endpoints were called when the required data to train the model for a postcode was not already on the local repository.

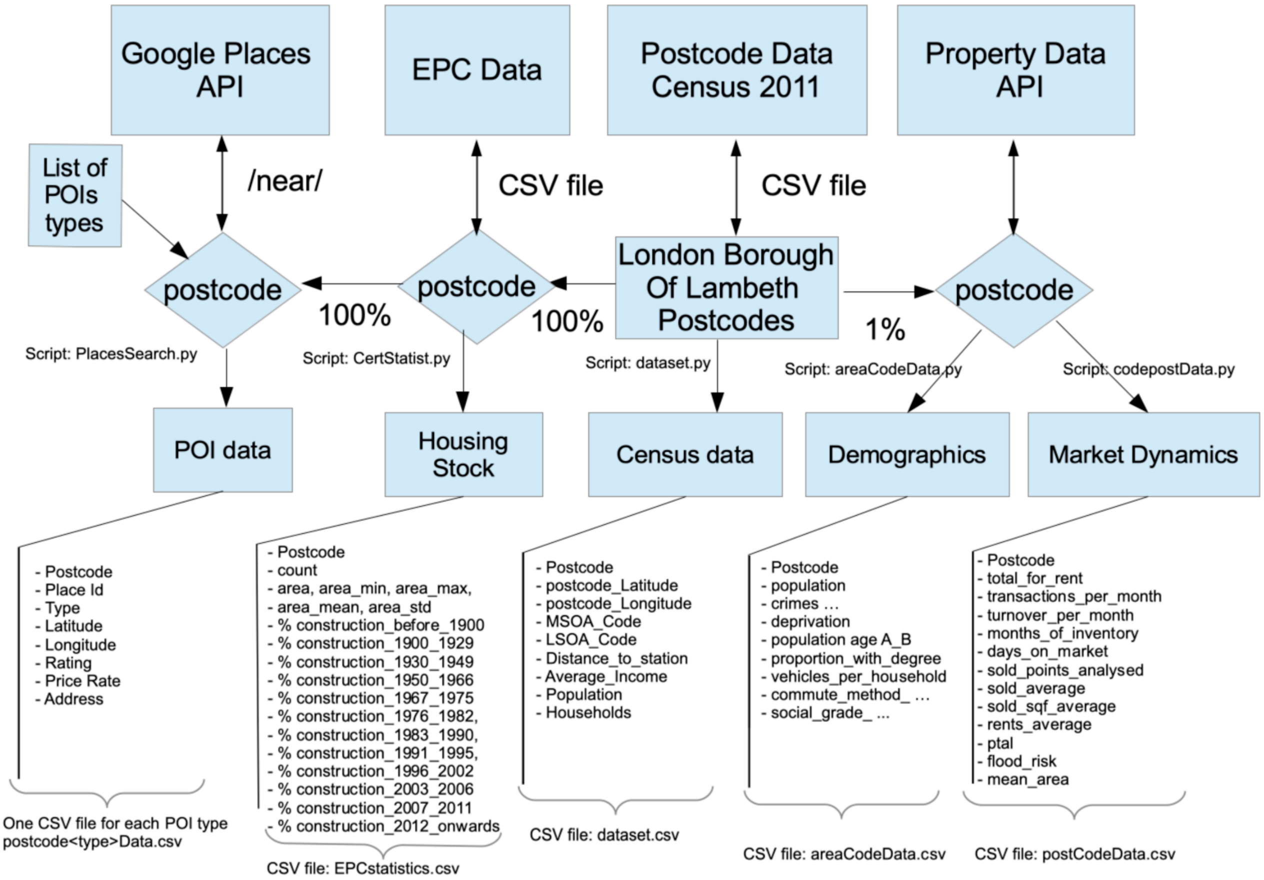


Figure 2. Data gathering and feature extraction process with their relevant Python scripts

Figure 2 shows the data acquisition pipeline in more detail. The relevant Python scripts are also mentioned. It also shows the process of feature engineering which is describe in depth in Section 3.2.4 below. Details of how the data used in feature engineering pipeline were collected is given for each data source below.

Census Data: what is referred to as census data in the figures in this project was derived from the LondonPostcodes.csv from www.doogal.co.uk [69] file mentioned above. As mentioned in section 3.2 above, full UK postcode was chosen as the link for all data sources. A csv file, LondonPostcodes.csv, of full postcodes for the UK was downloaded from www.doogal.co.uk [69]. All full UK postcodes and their coordinates, latitude, and longitude, within Lambeth were extracted and exported as a csv file and stored on a local repository. Latitude of centroid of a postcode was in decimal format i.e. 51.50205 and longitude of centroid of a postcode for this was in decimal format. i.e -0.07864 (negative values are those to the West of the zero (Greenwich) meridian). This file was then used to aggregate the relevant geospatial data from other sources for London Borough of Lambeth and create various datasets for the feature engineering pipeline. The feature engineering process is described in more detail in Section 3.2.3.

Apart from using the latitude and longitude four other variables were chosen from the LondonPostcodes csv file. At postcode level *distance_to_station*, *Average_Income*, *Population*, *Households* were also downloaded within the dataset and used as variable. The *distance_to_station* is the nearest train station to the postcode, which for London, also includes Underground and tram stops. *Average_Income* is the average household income of the MSOA that the postcode is located in. *Population* is the population of the area covered by the Postcode (from the 2011 census). *Households* is the number of Households in the area covered by the Postcode (from the 2011 census).

PropertyData: to access the API the PropertyData website (www.propertydata.co.uk) was visited to sign up for an API account and obtain an API key. A Python script, areaCodeData.py, was written using the *requests* library to send get requests to the API and save the responses. As it was demonstrated in Figure 2 that two groups of datasets were derived from the PropertyData API: demographics and market dynamics. These were grouped after analysing the in-depth interviews (more details in the Analysis section below). The main sources of all the data from the PropertyData API that relate to this project is shown in Table 2 below.

Land Registration	Monthly property transaction data and land title boundary data for all of England & Wales.
Office of National Statistics	Statistical data on house prices, economic metrics and demographics for the whole of the UK
Police.UK	Monthly crime data for England & Wales
Independent Schools Inspectorate	Data on independent schools for the whole of the UK
Food Standards Agency	Local restaurant hygiene data for England, Wales, and Northern Ireland
The Office for Standards in Education, Children's Services and Skills (Oftsed)	State school performance and review data for England
Ministry of Housing, Communities & Local Government	Property internal areas & energy scores, quarterly housing supply data, and Council Tax bands & rates.

Table 2. Original sources of data presented in PropertyData API and their description.

Extra caution was observed while using these variables since many of them that are not directly derived from other sources (such as census or property transactions) could have been processed in ways that we are not aware of.

Of particular interest was the data from Office of National Statistics (ONS) which includes demographics. These are in turn derived from census 2011 data. Accessing census data through PropertyData API was an easier and cleaner option compared to using the government data APIs. Socio economic demographics such as age, social grade, health, deprivation indices, and highest academic degree were used as well as crime, vehicle per household, and commute mode. The full list is included in the next section, 3.2.5. Feature engineering was performed on the age, some crime and commute variables of the demographics data and this is presented in section 3.2.3 below. The *social_grade, proportion_with_degree, vehicles_per_household, population, crimes_last_12m crimes_per_thousand, crime_rating,* were used as they were. A full list is given in the next section.

The /growth end point of the PropertyData API returns the five-year capital growth figures for a given UK postcode (full, district or sector). These, as well as transport related variables like the *ptal* score for a given postcode, which is a score that range from 0(worse) to 6a and 6b (the best) and assesses connectivity (level of access) to the transport network, combining walk time to the public transport network with service wait times were also used for different interval [70] for London postcodes. The full list of used variables is given in the next section and the descriptions are given in the Glossary, Section 7.

The target variables for each postcode were also extracted from the PropertyData API. The /sold-prices end point returns statistical average and confidence intervals of property sold prices. These were used to extract the *dy_sold_average* for a given postcode. The /rents end point returns statistical average and confidence intervals of live property asking rents (long-let on property portal www.rightmove.co.uk), from the smallest radius at which there is reasonable data. All rents are expressed as per week (for monthly values, multiply by 4.333). These were used to extract the *dy_rent_average* for a postcode. The /sold-prices-per-sqf end point returns statistical average and confidence intervals of property sold prices per square foot. These were used to extract the *dy_sold_sqf_average* for a given postcode. The PropertyData API bases the prices per square foot on 449,000 datapoints from sold prices in the past 12 months.

GooglePlaces API: Google Maps Platform has a Places API that includes POIs. To access the API, Google Maps API website was visited to sign up for an account and acquire a free API key. GooglePlaces API was used to build the POI dataset for model training and testing.

'Places' are defined within this API as establishments, geographic locations, or prominent points of interest. They are referred to as 'types' and there are 98 of them, ranging from libraries, locksmiths, schools, to universities and zoos [67]. The advantage of using the Places API was that these 'types' are seen as businesses to Google and come with reviews and star ratings. Not all 'types' available on Places API were chosen for this project for two reasons: first, the spiralling cost of making calls to the Places API, second the analysis of the data from the qualitative phase of the research (see Section 3.2.3). All the 'types' from the Places API are referred to as POIs in this project.

A Python script (PlacesSearch.py) was created to connect to the googlemaps library and access the Places API. The script uses the full Lambeth postcodes from the LondonPostcode.csv file as well their corresponding latitude and longitude attributes because there was a need, informed by qualitative research to calculate the distance of POIs to a given property address in a postcode. Feature engineering was performed on the POI dataset that is described in section 3.2.3 below. The initial data extracted from the API before feature engineering were stored in a local repository and included the following variables*: place_id*, *name*,*formatted_address*, *latitude*, *longitude*, *types*, *rating*,*price_level*. The *rating* attribute were of particular interest since these allowed for determining the 'quality' of a given POI and its potential impact on the price predictions by building features. A rating of 1 indicates a poor quality POI, while 5 is the best rating that can be given.

Energy Performance Certificate (EPC) Data: was accessed using the UK government website [71]. A csv file of all the available EPCs for the Borough of Lambeth was downloaded (certificates.csv). The EPC dataset includes the *address*, *postcode*, *local_authority*, *property_type*, *floor_area, construction_year* and several other variables relating to energy efficiency (e.g. CO2 emissions, heating cost, hot water, glazing type). Upon investigating the results of the in-depth interviews, it was observed that the energy efficiency variables were not of interest and were not chosen to be included in the final dataset. Instead *floor_area* and *construction_year* variables were used to aggregate and create new variables. More details given in the feature engineering section below.

EPC data was missing for some postcodes. For these a mean of a higher level of postcode was used, as such for a missing full postcode a mean of the sector level postcode, and for a missing sector level postcode the mean of the district level postcode were used. For example, if EPC data was not available for SW15 5PT, a mean for SW15 5 was used, if not available then the mean for SW15 was used. The code for this is included in CertStatis.py. The EPC data was used to engineer features, which are described in section 3.2.3 below.

3.2.3 Feature engineering

Since the second objective of the project is to incorporate the needs and insights of industry experts, there is a need to build a tool simple enough to be used by non-technical individuals. As such the basis of the feature engineering stage was informed by the qualitative phase of the project. The main categories that emerged were Points of Interest (POI), Housing Stock, Demographics, Market Dynamics. A through explanation on the interview analyses is given in Section 4.2 below.

The feature engineering pipeline was used process the loaded data and it is specific for each data source. This consists of a set of scripts used to aggregate and transform the data original variables. The processed data is integrated into a single dataset through to be used in the modelling pipeline.

POI data is unstructured and as such feature engineering was essential. This was informed by the in-depth interviews with the experts. This is discussed in more depth in Section 5 below. For each postcode the nearby POIs (types as referred to by Places API) were identified using three distance- based filters: 500 meters, 1,000 meters, 1,500 meters. The distance groupings of <500m, 500m<x<1000m, and 1000m<x<1500m were chosen for the following reason. Based on an average walking speed of 3.5 miles, or 5,632 meters, per hour, these ranges roughly translate to less than 5 minutes' walk, to about 10 minutes' walk, and 15 to 20 minutes' walk respectively [72].

In an attempt to deliver on one of the objectives of the project to quantify the quality of the POIs, the POI *rating* attribute was used. A separate script was written for each POI type (e.g. placeCafePostcode.py) where the following calculations were made to build nine features. If the distance between a given postcode and a group of POIs was less than 500 meters, the number of a specific POI (e.g. café) was counted, the mean and the standard deviation for the ratings of that specific type of POI (e.g. café) calculated. The same was done for each POI type at between 500 meters and

1,000 meters from a centre of a given postcode. And again, the same was done for each POI type at between 1,000 meters and 1,500 meters from a centre of a given postcode. The proximity grouping can also be represented as: POI_type<500m, 500m<POI_type<1000m, 1000m<POI_type<1500m. As such for each POI (e.g. café) a separate csv file was created with the following features:

> <POI_type>_rating_mean_1500 mean rating for this type of POI at >1000m and <1500m
> <POI_type>_rating_std_1500 rating std for this type of POI at >1000m and <1500m
> <POI_type>_count_1500 number of this type of POI at >1000m and <1500m
> <POI_type>_rating_mean_1000 mean rating for this type of POI at >500m and <1000m
> <POI_type>_rating_std_1000 rating std for this type of POI at >500m and <15000m
> <POI_type>_count_1000 number of this type of POI at >1000m and <1500m
> <POI_type>_rating_mean_500 mean rating for this type of POI at <500m
> <POI_type>_rating_std_500 rating std for this type of POI at <500m
> <POI_type>_count_500 number of this type of POI at <500m

Additionally, the minimum distance from the given postcode to the nearest POI was calculated and stored a variable called *<POI_type>_min minim distant to a POI of this type in meters*.

The **PropertyData API** was used to build demographic and market dynamics features. The data was very dispersed with some variables being too scattered. So, armed with the results of the in-depth interviews, improvements were made through aggregations to help with making meaningful interpretations once the model is built and tested. These choices are discussed in the results and discussion sections below.

The crimes were aggregated into four groups as shown in Figure 3 according to their potential impact on property prices and rental values. These were partly based on a body of literature on geographical profiling for different crime types and partly based on the outcome of the interviews. was studied to develop these features. The literature suggests that geographical profiling can be done for certain crime types such as murder (e.g., Canter et al [73]), robbery (e.g., Harries [74]), and burglary (e.g., Sarangi & Youngs [75]). However there seems to be a significant disagreement between those advocating for complex computer-based algorithms (e.g., Rossmo [76]) and those who prefer simple statistical and geometric methods and even human judges (e.g. Paulsen [77]). It is worth mentioning that the literature mentioned here did not explore crime types in a property price context. A thorough explanation, based on the literature, for choosing the following types is beyond the scope of this report, as such the categories should be considered with caution.

Type_A contains more violent crimes which impact the community and potentially the desirability of the area, *Type_B* crimes relate to theft, which may mean a more affluent area, *Type_C* contains crimes with an anti-social element, which may indicate a neighbourhood with a population who take pride in their neighbourhood, *Type_D* crime is vehicle crime, where, according to Tonkin et al [78] home of the person committing the crime is not as much a determining factor in predicting the location of the crime, than for some other types of crime. This may in turn imply that areas with more expensive vehicles, and potentially higher house prices, see higher degree of vehicle crime.

The age groups were aggregated as follows, again as shown in Figure 3. The younger ages were kept into their original bands with 0-4 being babies, 5-9 pre-school, 10-14 primary school, 15-19 secondary school. These can have indications for school catchments areas and their affect on the price and rental values. A group was created for the age group of 30 to 64 year olds as essentially the core work force. *TotalModPop* was also created which is the percentage of the active population of the whole population. Ages 20 to 29 were group together as the students and recent graduates. These variables are all presented in percentage terms (i.e. 0-4 is the percentage of the population in that age band in that postcode).

The commute type was aggregated into three types as shown in Figure 3. Commuters by train, bus, and underground were put in the ‘public’ category. Those travelling by motorcycle, taxi, car (either as passenger or drive) were put in the ‘private’ category and those travelling by bicycle and on foot put in the ‘open’ category. It was expected for the public category to have the biggest impact as Lambeth is an inner London borough with many stations and bus routes.

Social grades were not altered. Social grades are a socio-economic classification produced by the ONS (UK Office for National Statistics) by applying an algorithm developed by members of the MRS Census & Geodemographics Group. They have been constructed to measure the employment relations and conditions of occupations [79].

The /growth end point from PropertyData API, which returns the five-year capital growth figures was used to extract and create market performance variables of *Year_1_Value, Year_2_Value, Year_2_growth%, Year_3_Value, Year_3_growth%, Year_4_Value, Year_4_growth%, Year_5_Value, Year_5_growth%*. PropertyData API uses price paid data from the UK Land Registry to calculate these growths for each postcode.

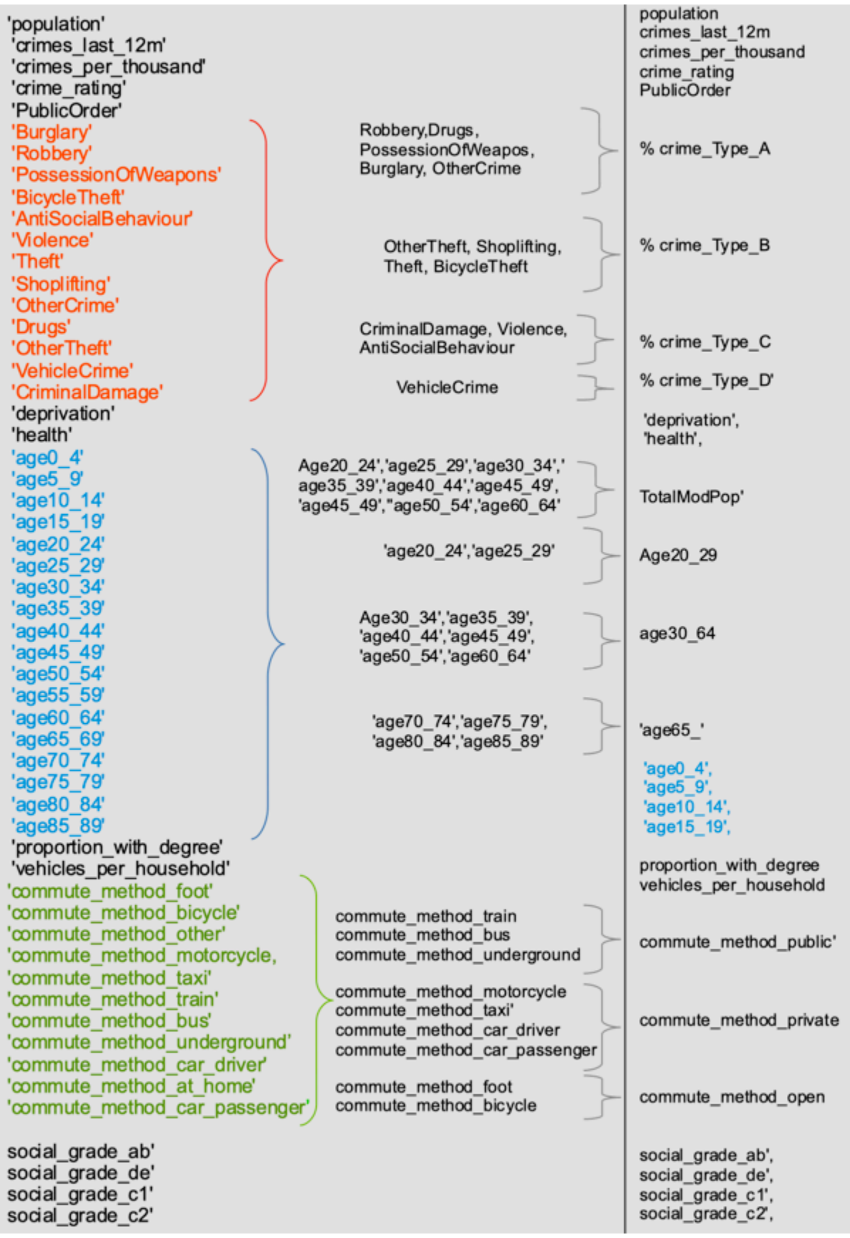


Figure 3. Feature engineering of the census data acquired via PropertyData API

From the **EPC** dataset for each postcode the number of building with certificated were counted and the variables ‘count’ was then combined with the ‘area’ variable to construct four other variable: *area_min*, *area_max*, *area_mean*, *area_std*. A Python script, CertStatist.py, was written to this effect. The scripts selects the relevant postcode and then created the variables. These variables describe the minimum, maximum, mean, and standard deviation of the floor area for a given postcode.

The percentage of the buildings that fall within certain construction period was also calculated for different periods. These were grouped in the following bands and shown in Figure 4: before 1900, 1900 to 1929, 1930 to 1949, 1950 to 1975, 1976 to 1982, 1983 to 1995, 1996 to 2006, and 2007 to 2012 and above. This aggregation was informed by the housing stock type in Britain. The housing stock before the 1900 is usually referred to as 'period' properties [80]. Period houses in Lambeth are often thought of as not very well insulated and expensive to maintain, yet desirable if well-located, for instance in a conservation area or in certain school catchment areas. Period flats are often converted from houses in Lambeth and sometimes without proper planning permission or building regulations sign-f which means they are not as desirable as period houses.

The period between 1900 to 1929 saw the boom in construction before the collapse of the great depression of 1929. 1930 to 1949 is the bracket that falls roughly in the second world war period where construction was limited. From 1950 to 1970 construction of social housing properties boomed. From 1983 to the mid 1990s a new type of housing stock was on the rise, with more modern blocks being built for private sale as opposed to social housing. During this period tenants in social housing got the 'right to buy' their properties from the council. Mid 1990s to mid 2000s was the first wave of 'open-plan' living apartments where the kitchen and the living area are not separate. Since mid 2000s there has been efforts to improve the quality of the housing stock in London by introducing measures such as the 'minimum space standards' where new-build apartments can not be smaller than a certain prescribed size. These aggregations were stored in variables named *% 'construction_before_1900'*, *'%construction_1900_1929'*, etc. The aggregation process is described in Figure 4. The full list is included in the results section below.

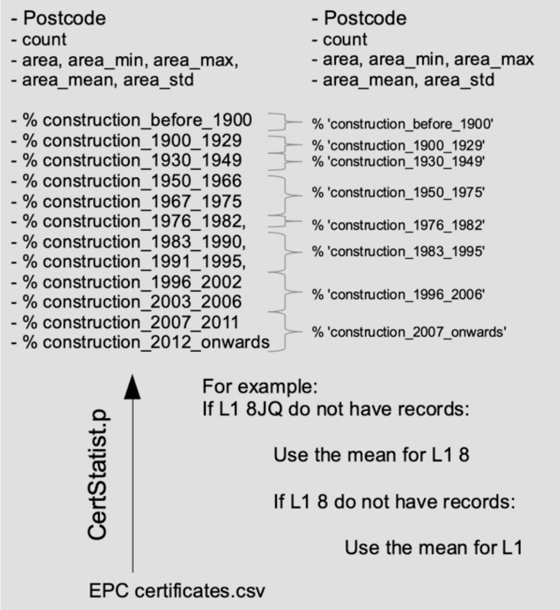

- Postcode
- count
- area, area_min, area_max,
- area_mean, area_std
- % construction_before_1900
- % construction_1900_1929
- % construction_1930_1949
- % construction_1950_1966
- % construction_1967_1975
- % construction_1976_1982,
- % construction_1983_1990,
- % construction_1991_1995,
- % construction_1996_2002
- % construction_2003_2006
- % construction_2007_2011
- % construction_2012_onwards
- Postcode
- count
- area, area_min, area_max
- area_mean, area_std
% 'construction_before_1900'
% 'construction_1900_1929'
% 'construction_1930_1949'
% 'construction_1950_1975'
% 'construction_1976_1982'
% 'construction_1983_1995'
% 'construction_1996_2006'
% 'construction_2007_onwards'
CertStatist.p
For example:
If L1 8JQ do not have records:
Use the mean for L1 8
If L1 8 do not have records:
Use the mean for L1
EPC certificates.csv

Figure 4. Feature engineering on the EPC dataset with an example of handling the missing values.

No feature engineering was done using the **Census** data extracted from the LondonPostCode csv file.

3.2.4 Model selection

The literature review in section 2.1.2 above clearly demonstrates that machine learning methods have notable advantages to the hedonic price models. It also be deduced that random forest and similar models that are derived from it perform better. Hence the decision was made to use XGBoost for the purposes of this project. The literature review also highlighted the higher speed of learning and predicting tasks of XGBoost, which is useful given the large datasets used in this project. More importantly Noh et al [60] point out that XGBoost model allows a wide range of hyperparameter tuning that in turn reduces model overfitting in imbalanced datasets, like those used for this project. XGBoost also has the ability to measure feature importance, i.e. analysing the contribution of variables and calculating their influence on a model's prediction [81]. This ability to a certain extent allows the user to open the 'black-box' of machine learning algorithms making it more transparent. Transparency became of particular interest in this project after conducting the in-depth interviews. These relationships are discussed in the results and discussion sections below.

This project is not a comparative study of different machine learning models. The intention of the literature review was to find the state-of-art models to use in the prediction task while incorporating the results of the in-depth interviews with industry experts. Despite this other machine learning algorithms that featured in the literature review were tested to make sure that the choice to use XGBoost was a good one.

LASSO, Support Vector Regressor (SVR), Ridge, Random Forest and PCA were all tested. The Lasso model, which performs feature selection with an L1 penalty, was fitted with a high regularisation parameter alpha of 0.5. The RandomForestRegressor from the sklearn library has a built-in feature_importances function that shows how often a feature has been used in each tree of the forest. PCA was tested in a script called SQLMATTselectionPCA.ipynb to see if the speed of the training could be increased and also if the principle components could explain the variance and what the most important variables are. Here, the data was decomposed into 10 principal components with the first two components *containing* 95% of the data variance. Those two components were dominated by variables describing the market dynamics, in other words price related variables. SVR was also tested in SQLMATTselectionPCA.ipynb with an Radial Basis Function (rbk) kernel. With inferior prediction results/

The modelling pipeline can generate three types of models to predict the mean property price, the mean price per sq foot, and the mean property rental value for a given postcode. The python scripts are called NewSaleModelPlacesV2.ipynb, NewSFAModelPlaceV2.ipynb, and NewRentalModelPlacesV2.ipynb.

The feature engineering described in section 3.2.3 was carried out in these scripts. Those models were trained and evaluated using data set described in section 3.2.4. and labelled datasetV6.csv.

The only categorical variables PTAL score (0, 1, 2, 3, 4, 5, 6a, 6b), flood risk (none, very low, low, high, very high), and crime rating (Very high, High crime, Average, Low, Very low) were hot one coded as XGBoost can not deal with categorical variables.

Hyperparameter tuning was performed as this was needed to achieve better results. At a most basic level, in tree-based models like the XGBoost the learnable parameters are the choice of decision variables at each node and the numeric thresholds used to decide whether to take the left or right branch when generating predictions. From a computational perspective this mean minimising the loss function, which in this case is the mean square error (MSE) as XGBoost is a supervised machine learning algorithm.

The parameters that were tuned are: number of trees (n_estimators), learning rate (eta), minimum split loss in each node (min_split_loss), minimum weight of the each of the child nodes in the tree (min_child_weight), lamba which is the L2 regularisation term on weights (reg_lambda) and alpha which is the L1 regularization term on weight (reg_alpha) which are both to reduce overfitting, column sample by tree which denotes the fraction of columns to be randomly samples for each tree (colsample_bytree), and sub sample which denotes the fraction of observations to be randomly samples for each tree (subsample).

The optuna library was used to help with hyperparameter optimisation and automate hyperparameter search. To begin with, this library is much faster than scikit-learn's grid search. It uses a state-of-the-art algorithm for sampling hyperparameters and efficiently pruning unpromising trials, while grid search is a brute force strategy which tries all possible values in the grid and takes longer. The SuccessiveHalvingPruner() pruner of the optuna library uses the Asynchronous Successive Halving Algorithm as presented in Li et al [82]. Cross validation with 4 folds was used with optuna to reduce the tunning time.

The xgboost library was used in Python to fit the XGBoost models. Description statistics was carried out on the target variables to find a benchmark mean to be used to calculate the mean absolute error of the models. Metrics were defined for both train and test sets to monitor the performance of the models. These were the Mean Absolute Error or MAE, Mean Absolute Percentage Error or MAPE, and Mean Square Error or R2. Of particular interest was the MAE %, which is the relative error and expressed in percentages, and defined by the ratio of the MAE to the mean value measured, in this case price or rent this is referred to as MAE% in the results section.

To further improve the results a model evaluation method was defined by using a repeated stratified 10-fold cross-validation with 3 repeats. Results for the three models mentioned in this report all use this model evaluation method.

Figure 5 shows the overall process of model training as well as the chosen hyperparameters for the price prediction model. It also shows the data sources categorised based on Figure 5 data acquisition pipeline]. The same process of hyperparameter tuning was carried out for the rent model and the hyperparameter values were very different. The final choice is given in the results section below.

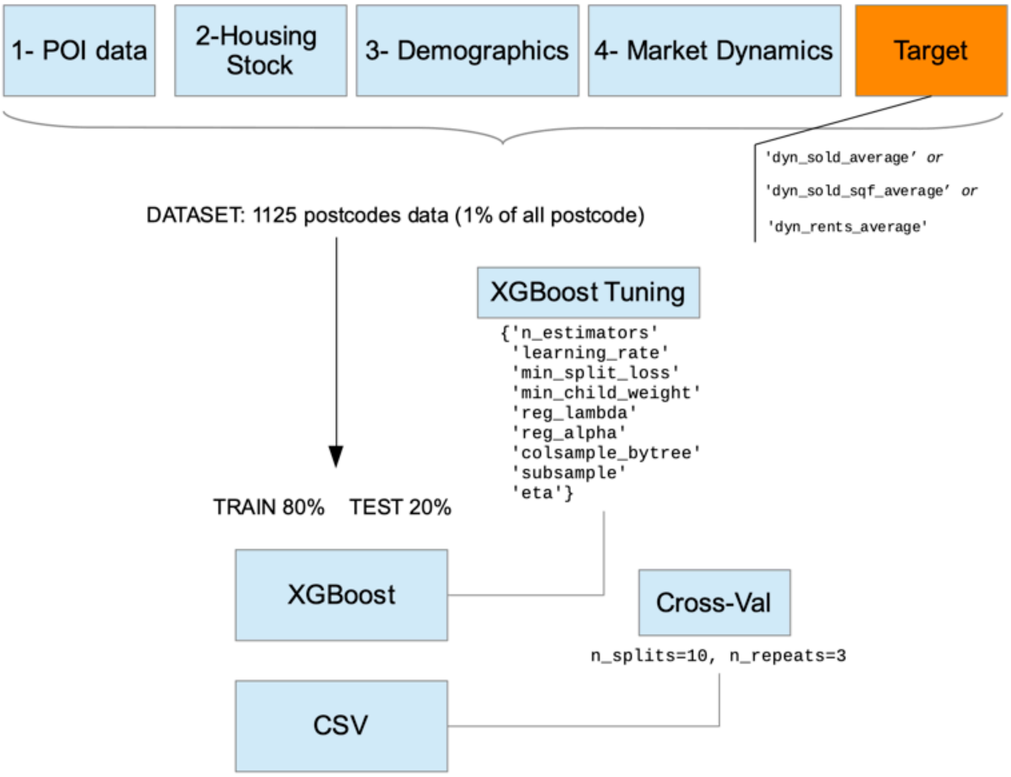


Figure 5. Modelling pipeline with XGBoost hyperparameters that were tuned, the data and the cross validation used.

After all the training and testing in this phase, the hyperparameters for the best performing models were used in the prototype model training (section 3.3) and ultimately the deployment (section 3.4).

3.3 Prototyping

Since one of the objectives of the project was to incorporate the needs and insights of industry experts, there was a need to build a tool simple enough to be at least tested by non-technical individuals. As such the final models were incorporated and deployed in a simple prototype to enable the testing of the model performance as well as assessing the degree to which expert insights informed the process of creating the appraisal tool. The themes that emerged from the in-depth interviews informed the decision of the front-end design. Although it is important to note that purpose of this design and deployment was to demonstrate the models and a very basic user interface in the quickest and most cost-effective way as opposed to advancing user research or data visualisation research objectives. The inner workings of the model training and testing for the front end is described in 3.3.1 below. Decisions and the process in terms of what to include in the front-end are explained in 3.3.2 below. How the front-end user interface was deployed to test the models is described in section 3.3.3 below.

3.3.1 Model training

The best performing XGBoost was found in the data modelling pipeline. The best hyperparameters differed for the price prediction models and the rent prediction model; these were noted. For the front-end the same scripts were used to incorporate the same modelling pipeline was incorporated: NewRentalModelPlacesV2.ipynb, NewSaleModelPlacesV2.ipynb, NewSFAModelPlaceV2.ipynb. however the data that is used is different. For a given postcode and address, the input put in by the user, the script finds all the variables described in section 3.2.4 within a 2 kilometres (km) radius of that address and postcode (for choice for distance refer to section 3.2.3). It then splits this into train 80% and test 20% sets and uses the same evaluation metrics, MAE, as those used in the modelling pipeline were the best performing models were selected. As mentioned before, the final dataset was for 1% of the postcodes in the borough of Lambeth. If the postcode inputted by the user is not in this 1% dataset, the prototype modelling and forecasting pipeline makes a call to the relevant APIs to obtain the relevant data. In the prototype modelling pipeline the optuna library is not used and there is no cross-validation since these were used to obtain the best hyperparameters in the modelling pipeline.

An overview of the data modelling pipeline that was used is presented in Figure 1, which also mentions elements of the data related pipelines. Model training was postcode oriented to serve the purpose of the prototype development. One model was trained on-demand for a given postcode using data on the postcode vicinity, which was stored on a local repository (2). For missing data on the local repository in each postcode, a set of API end-points were called to update the repository with the required data (3). A tabular dataset was created to feed the machine learning algorithm after applying feature engineering and attribute selection (4). This dataset was split into training and test sets. The training data was used to train various models. The models were applied to the input postcode (1) to generate predictions (5) and test data used to estimate the prediction error (5). The model structure was analysed to construct a rank of the impact of the attributes on the model prediction (6). This produced output is presented on the front-end (7). When the relevant attributes are related to POI information, we can locate the associated POI location using its coordinates (8).

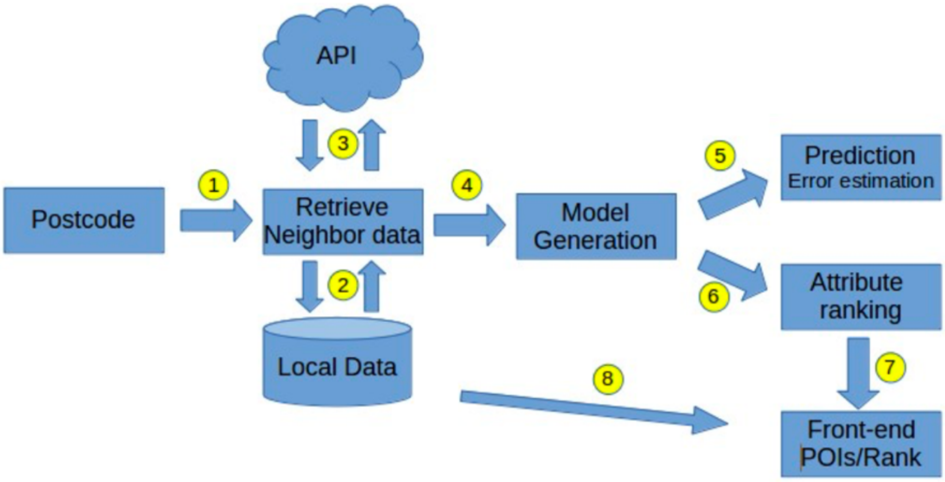


Figure 6. overview of the data modelling and deployment pipeline

Tree based models, like the chosen algorithm for this project XGBoost, the 'importance' of a variable or feature can be defined using various importance metrics such as Gain, Coverage, and Weight. The script for the models in predicting the price, price per square foot, and rental value all include a section to rank the features (variables that were used as input for the best performing model) by importance. The script used in the prototype also includes a section to rank the models.

The *Gain* implies the relative contribution of the corresponding feature to the model calculated by taking each feature's contribution for each tree in the model. A higher value of this metric, when compared to another feature, implies it is more important for generating a prediction.

The *Coverage* metric means the relative number of observations related to this feature. For example, there are 100 observations, 4 features and 3 trees, and suppose feature1 is used to decide the leaf node for 10, 5, and 2 observations in tree1, tree2 and tree3 respectively; then the metric will count cover for this feature as 10+5+2 = 17 observations. This will be calculated for all the 4 features and the cover will be 17 expressed as a percentage for all features' coverage metrics.

The *Weight* is the percentage representing the relative number of times a particular feature occurs in the trees of the model. In the above example, if feature1 occurred in 2 splits, 1 split and 3 splits in each of tree1, tree2 and tree3; then the weight for feature1 will be 2+1+3 = 6. The frequency for feature1 is calculated as its percentage weight over weights of all features.

Gain is the most relevant metric to interpret the relative importance of each feature for the purposes of this project, hence the chosen method to rank the variables by, was their gain in the XGBoost algorithm.

3.3.2 Front-end

For the front-end were driven by the needs of the insights gained through the interviews.

1- Data input fields. The choices here were driven by the information that anyone who is considering buying a particular property usually has at hand, in other words an address and a postcode. This was one of the reasons why a full postcode was chosen as the common thread in this project to connect all different data sources together, as opposed to MSOA or LSOA. The specific MSOA or LSOA is not usually known to any of the parties, be it buyer, seller, or lender.

2- Map with the property vicinity. The decision to include a map was made after trying the interfaces of two of the existing property investment appraisal tools available in the market: Realyze and Yuvoh Analytics. Also, during the in-depth interviews it emerged that most of the property investment appraisers do use maps in conjunction with their own knowledge or insights gained through their networks to 'place' a potential investment property during the appraisal process. The POIs and the radii were displayed.

3- Model output fields. The main aim was to predict the price and the rent for a given property, so these were included: *Mean property price* and *Property rental price* which are given for that particular full UK postcode. In addition to these, the evaluation of the model performance had to be translated into understandable and usable language for the property investment professionals. It was determined from the interviews that two useful metrics would be a *Mean price variance* in square foot and a *property price interval*. *Mean price variance* per square foot was calculated by using the MAE of the price per square foot model. The *property price interval* was calculated using the MAE as well by finding where intervals where the real value should be: if the error has a lower standard deviation this interval is shorter, an error with a big standard deviation produces larger intervals. *Mean area in postcode* in square foot *(sqf)* was also included to give an idea about the size of the properties in the postcode. *Gross rental yield* was also calculated and included as Annual rental income (weekly rental income x 52) / property value x 100. This is a standard metric used by property investment professional to assess income producing assets. The final output is described and shown in section 4.3.

4- Feature rank. One of the main themes that emerged from the in-depth interviews was the need for a more ‘explainable’ and ‘transparent’ systems where by the importance of the driving factors in the model’s decision making process are known. It was suggested that this in turn allows the investment decision maker to access the inner working of the model and decide whether they agree or not based on their own knowledge.

To display the variables in order of preference a dictionary was created with pair of variables name as used in the model and a more meaningful name to communicate as much information as possible about the variable to the user. This dictionary was saved a separate file called dictionary.docx in the WebService folder. For example while crime types were grouped and labelled Type_, Type_B etc, they were renamed in the dictionary to convey a better meaning: crime_Type_A used by the model is translated as 'Crime: Robbery, Drugs, Weapons, Burglary'.

Based on the analysis of the in-depth interviews, the contributing variables were divided into four categories from the interviews of POI, Housing Stock, Demographics, and Market Dynamics. These groupings were only made for the front-end display and did not have any impact on the model training and testing.

3.3.3 Deployment

Figure 7 demonstrates how the appraisal tool was deployed. The datasets were uploaded to a Docker container where a Django web service runs. The service in the Django uses the csv files created in the project, namely googlePlaces, datasetV7, and ListOfIDs_<POI>. datasetV7 is the same as datasetV6 which was used for model training and testing minus the googleplaces data. This data was trained on 1% of postcodes as mentioned in section 3.1.2. Using the two separate files here served the purpose which was to include the location of POIs on the map and more importantly rank the variables in order of importance in the front-end. The googlemaps API is called to show a map and the location of POIs. If the data that is queried does not exist in the csv files, then an API call is made to the ProeprtyData API and the dataset is updated.

A folder called Webserver contains all the files that relate to the deployment. The Django code can be found in the ‘root’ folder. Dockerfile contains all the API keys necessary to call PropertyData API and GooglePlaces API. Only changes were made to the urls.py, where the location of the two html files that were written by modifying templates is stored: landing_page.html and output_rank.html. The back-end code is in views.py, where it was copied from NewRentalModelPlacesV2.ipynb and NewSFAModelPlaceV2.ipynb which contains all the pipelines described in 3.3.1 and the dictionary that is used for displayed for ranking the variables. The function ‘def POI(request)’ is used to read the information inputted by the user. The Docker container was then deployed on the Google Cloud Platform.

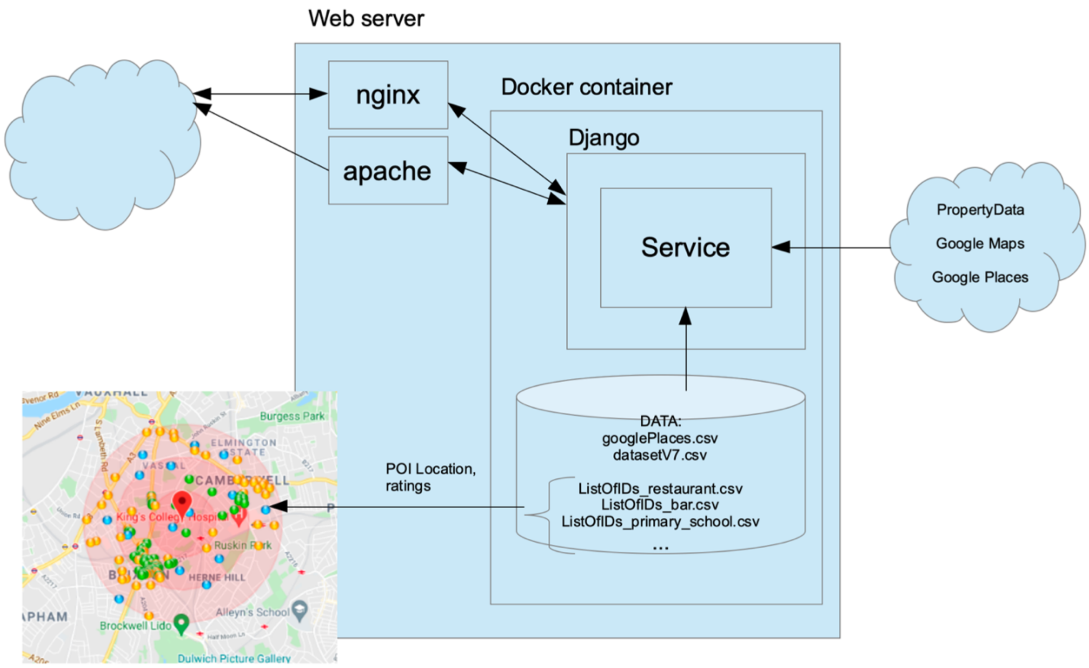


Figure 7. Deployment details with the docker container and web server.

The implementation can be tested on https://ben-5g54k44v2a-nw.a.run.app/ and by using the following username and password:

username: ben
password: ghssdjjk1234hv

The following postcodes and addresses are already included in the file that has been stored in the Docker container and can be used to test the prototype:

76, Kempshott Road SW16 5LH
Flat 1, 24, Greyhound Lane, SW16 5SB
10, Polperro Mews, SE11 4TY
248, Brixton Road, SW9 6AQ
Flat 15, Union Grove, SW8 2QR

The service is running serverless, which meant a cost saving. It has the following settings:

CPU allocation: CPU is only allocated during request processing

CPU: 1

Memory: 512MiB

The CPU is stopped after 300s without requests. To speed up the service response and model training more CPU cores and memory can be allocated.

4 Results

This section link the results of the qualitative and quantitative phases with aim of achieving the research objectives:

OB1: *Can the use of alternative data sources, as opposed to price paid datasets, match or improve the performance of price and rental prediction models in a property investment appraisal context?*

OB2: *How can the use of alternative data sources impact, and potentially benefit, the industry experts who use property appraisal tools?*

The results of the prototype development are also included with screenshots of the appraisal tool and instructions to access the deployed version.

4.1 Analysis of interviews

4.1.1 Themes

In this section it is important to remember the background of all the interviewees. As a reminder a table is provided with their abbreviations and meaning so that the reader can follow this section according to the experience of the interviewees if they so wish to.

Interviewee ID	Interviewee Code	Interviewee code description
IA	SBA	small sized business advisor
IB	MBO	medium business owner
IC	LBO	large business owner
ID	TBO	technical business owner
IE	PLL	portfolio landlord

The interviews were chosen for their experience of using tools to make key investment decisions, and their implicit knowledge of the UK residential property market. With the sample size of five interviewees, we were not able to carry out independent coding of the interviews. It is acknowledged that the inferences from the qualitative phase of the project may be biased. All five experts use several sophisticated AVMs such as Realyse, Yuvoh Analytics, LandInsight as well as more simple tools such as Microsoft Excel spreadsheets. The most sophisticated tool used by IE is Zoopla and Rightmove property price estimates, although they do rely heavily on a mature base knowledge and leveraging their network in an area that they know well. It also transpired that everybody in the property investment process, such as the valuer, buyer, seller, lender do their own financial models as they don't trust others.

In the process of deducing the themes a decision had to be made to focus the project in one area, which was the use of geospatial data sources. This was done by counting the number of times a topic was mentioned. The themes of what was coloured grey and deemed not relevant to this project (not mentioned as much as the other yellow and green themes) included the lower cost of the AVM tools compared to bespoke data solutions, streamlining the process to allow staff to do high value tasks rather than looking for comparable sales on the internet, red tape (bureaucracy, politics), and regulatory concerns. These amongst other themes were dismissed as they do not have any bearing on the use of data and the AVM's performance.

Yellow was used for segments that relate to the general discussion about the use of technology in the property sector. The yellow segments highlighted the importance and growing need of using technology, relating mostly to objective 1 of the project. Themes emerged included the need for using a vast number of *alternative data sources*, *developing accurate models*, and *seeking more transparency* (not black box).

IA mentioned that using technology can be part of building a good decision-making process that overcomes bias as much as possible and offers alternative perspectives which can possibly lead to alternative courses of action.

The main them here though was the quest to find reliable data. Especially ID who runs a property data company emphasised the constant quest to bridge the lag that exists in certain datasets (such as census) and build features using alternative datasets such as companies house (for information about companies registered in a postcode) and UK finance (data on household finances). Not all the important trends in price and rent at a neighbourhood level can be spotted using only price paid data – they certainly can not be explained just by looking at historic price data. Furthermore, some contributing factors and fluctuations take longer to show up in the data, especially the price paid data. So this further exacerbates the need to use various data sources rather than solely relying on historic transactions.

Alternative data sources were also mentioned by IA who would like to look at what variables and features are being used by the company that provides the appraisal tools. IB mentions that they would like to receive the actual data in a manageable and structured way that is usable for non-technical people, rather than pdf reports, as the reports can be vague and clearly lack depth and at times accuracy. IB said that one of the tools they use uses 50 property related features, 100 demographics, income, crime, transport, schools, supermarket, etc…, with a few engineered variables like the chicken shop to coffee shops ratio. IB finds this useful as an appraiser as it gives a clearer picture as to what data is at least being used. In some cases this is particularly useful, for instance new energy efficiency standards that have recently come into force are affecting specific property type rather than whole area, looking at the data at postcode level has the potential to highlight these points. This call for further transparency in the form of knowing where the data that drives the models comes from was echoed by IA and ID.

There is significant asymmetry of information between buyer and seller in the property market, where the seller knows more about the property than the buyer. IC pointed out that in the lack of specific information about a particular property, one must rely on every other bit of data that one can including geospatial data on demographics and points of interest. IB mentioned that not everyone in the value chain (RICS valuer, bank, investor, developer, tenant, seller) can benefit unless they all have access to the same data. They also pointed out that this will never happen in real estate it will result in losing your advantage.

IE who has relied on their 'gut-feeling' throughout the years admit that the overall positive market conditions has helped them with the success they have seen. This has not been based on robust data analysis up until recently when the market got more competitive, and the sellers are also using the appraisal tools that once only the lenders and some buyer had access to. IE, despite relying on their network and gut-feeling, as well as an element of luck, also admit to limitations on relying on these means if the intention is to invest in property at scale.

IC mentioned that the conversation is always about data, how to harvest it, how to understand its importance

> 'we don't have any input in that usually as an investment firm, we don't know what's under the bonnet of the AVM and that's frustrating!' *IC*

These findings from the yellow segments were pointed to the value of **good quality and often updated geospatial data** transparent. Hence the green themes were organised so to reflect this.

4.1.2 Categories

Green was used to highlight relevant sections in the body of the interview transcripts. These were mostly answers to the questions about what the experts like or don't like or would like to see improved in the appraisal process and the tools they use. To begin with the mindset of the author was focused on traditional price paid data versus alternative data sources. An open mind was kept as to what range of alternative data sources were going to be discussed. This was left to the interviewees to mention in answer to the question: 'How important is being able to tell what data is driving the appraisal tool's decision-making process?' or 'How do you keep track of the effects of the different data sources on your appraisals?'. The segments highlighted green, mainly in the answers to these questions, saw four categories of important factors emerge. The categories were chosen based on the number of times the interviewees mentioned that certain topic in the context of the interview questions and what is helpful for the experts in ‘informing them to build a picture of the investment as opposed to just looking at the numbers’, as IC pointed out.

To best capture the essence of the chosen variables, both those readily available and those engineered, the data is presented in four separate categories of**:**

POI
Housing Stock
Demographics
Market Dynamics

IB gave London as an example where several sub-markets exist and the capital growth mantra does not apply to all of them. This is reference to the importance of the geospatial variables from various data sources to help bring out trends in those submarkets at a granular level.

The problems with inferior predictive power came up in the interview with ID in the context of banks/lenders having to meet regulatory obligations, hence using more explainable but inferior in predictive power models. This has caused many problems for other decision makers like IA, IB, IC, and IE, all of whom mostly rely on bank lending to acquire investment project.

4.2 Features selected and engineered

4.2.1 Techniques used

To best capture the essence of the in-depth interviews and the themes of **usability, predictive power, explainability**, the four categories that emerged during the qualitative phase were used to make all the decision in the data sourcing, data acquisition, feature extraction, and feature engineering pipelines. It was for this reason that the data sources were split into POI, housing stock, demographics, and market dynamics. Census data and those variables that related to demographic from the PropertyData API were put in the demographics category. The EPC data was used mostly for the housing stock category. The market dynamic variables were mostly built from the PropertyData API. Section 4.2.2 contains table with all the variables and their sources.

As part of the feature selection and engineering the algorithms that were tested but dismissed were LASSO, SVR, Random Forest and PCA. These were initially tested for dimensionality reduction, and then for prediction task. The scripts are all attached to this project in LassoRidgeRandomForest.ipynb and SQLMATTselectionPCA.ipynb. Although, as an example the results for PCA were as followed:

```
Train MAE :  54.555787 --- Test MAE :  51.889463
Train MAPE :  42.324315 --- Test MAPE :  42.466602
Train R2 :  0.409572 --- Test R2 :  0.464242
```

Using any of these algorithms would have defeated an important theme that emerged the in depth interviews and the in the literature review: ranking the variables in order of their contribution to the model and ultimately making the prediction more transparency and exploitability. Random forest, although had similar accuracy to XGBoost, two variables repeatedly came up as the most important contributor: tube distance and crime rate, which indicated overfitting as is the case with most random forest algorithms with a problem similar to that in this project. The stacking ensemble used for the SVR yielded similar accuracy scores to PCA, which were all inferior to XGBoost. PCA was dismissed as a feature reduction or prediction model because when the variables are projected on to a vector and transformed, they lose their meaning.

Variables losing their meaning or being eliminated, even to achieve better model performance, was actively prevented in this project since an important aspect is to preserve the variables and rank them in order of importance. Using XGBoost as the main machine learning method to perform feature engineering and the prediction was the best balance. Given the pivotal role that the input of the experts played in this project, the final variables included in the final dataset were decided based on the in-depth interviews, as opposed to dimensionality reduction techniques, some of which were tested and describe above.

4.2.2 Final Dataset

The final dataset chosen included 1% of the postcodes in the Borough of Lambeth, which amounted to 1125 postcodes. The 1% was determined by the number of EPC certificates available for that postcode. Only 1% of the postcodes had at least one house with an EPC certificate. It was important to have this dataset available as it was used to extrapolate information about the housing age, which has implications for housing type, in the postcode vicinity. The data set had 348 variables.

The script dataset.py was used to join all the generated csv files together based on the postcode column and create the final dataset in csv format, called datasetV7.csv. As an overview, the collected data, and the engineered features, describe a postcode using:

Data	*Source*
POIs descriptions near the postcode centre at <500m, <1000m <1500m	GooglePlacesAPI
Housing stock construction statistics in the postcode or near the postcode	EPC
Demographics	PropertyDataAPI + Census Data
Market Dynamics	PropertyDataAPI

The aggregations of crime, market dynamics, and age data mentioned in the previous section are all contained in the dataset.py script.

Table 3 below shows all the variables that were selected for training the model. The importance of attribution selection was two-fold: they had to be based on the literature, and they had to reflect the analysis of the in-depth expert interviews. These are discussed in sections four and five below. To best capture the results of the in-depth interviews in the qualitative phase of this project, the data is presented in four separate categories of **POI**, **housing stock**, **demographics**, and **market dynamics**. These categories were informed by the qualitative phase of this project. The variable names below do not correspond exactly to the variable names in the Data in the python script dataset.py, but there is a very close correlation between the two that it is possible to relate this report to the code with no effort. Market Dynamics section is the only price related data that was used in this project as the focus has been on using alternative data sources in the in. Not directly price-paid data was used in the end to distinguish this project from all the papers and research that use historic transaction data as their main data source.

POI

Point of Interest (POI) from GooglePlaces API

Bar
Cafe
Cemetery
Church
Fire Station
Hindu Temple
Hospital
Park
Parking
Police
Primary School
Restaurant
School
Secondary School
Shopping Mall
Subway Station
Supermarket
Synagogue
Train Station
University

For each of the POIs from GooglePlaces API

POI Rating Mean & Standard Deviation at 500m, 1000m, and 1500m
POI Price Rate

POIs from PropertyData API

Distance to station
School Average Score
Schools Rating
Restaurants with Average Hygiene Rating of 5
Restaurant's Proportion Bad

Housing Stock

Features constructed from Energy Performance Certificate data

'area_min' min area of property on the postcode
'area_max' max area of property on the postcode
'area_mean' mean area of property on the postcode
'area_std' area std of property on the postcode

% 'construction_pre_1900' in the postcode
% 'construction_1900_1929' in the postcode
% 'construction_1930_1949' in the postcode
% 'construction_1950_1975' in the postcode
% 'construction_1976_1982' in the postcode
% 'construction_1983_1995' in the postcode
% 'construction_1996_2006' in the postcode
% 'construction_before_1900' in the postcode
% 'construction_2007_onwards' in the postcode

Demographics

Census data from PropertyData API (originally ONS census) & Census UK (2011)

Census
Deprivation
Population
Households
House per Population
Health
Social Grade AB
Social Grade C1
Social Grade C2
Social Grade DE
Proportion with Degree
Average Household Income
Vehicles per Household
PTAL Score

Commute Method Public
Underground/Light Rail
Train
Bus

Commute Method Private
Taxi
motorcycle
Car Driver
Car Passenger

Commute Method Open
Bicycle
Foot

Age
Age_0-4
Age_5-9
Age_10-14
Age_15-19
Age_20_64 (% Active Population)
Age_65+

Crime data from PropertyData API (originally www.police.uk) - all in % of total crime to aid with meaningful prototype development

Crimes per Thousand
Crime Rating
Crimes last 12months

Type A	*Type B*	*Type C*	*Type D*
Robbery	Bicycle Theft	Anti-Social Behaviour	Vehicle-Crime
Drugs	Shoplifting	Criminal Damage / Arson	
Possession of Weapons	Theft	Violence / Sexual Offences	
Burglary	Other Theft		
Other Crime			

Market Dynamics

From PropertyData API to describe market growth on the last 5 years for a given postcode

Year 1 Value this is the total of transactions in 2016
Year 2 Value this is the total of transactions in 2017
Year 2 growth% the growth from 2016-2017
Year 3 Value this is the total of transactions in 2018
Year 3 growth% the growth from 2017-2018
Year 4 Value this is the total of transactions in 2019
Year 4 growth% the growth from 2018-2019
Year 5 Value this is the total of transactions in 2020
Year 5 growth% the growth from 2019-2020

Market dynamics for a given postcode for a property are also described by

Total Number of Property Listed to Rent
Transactions per Month
Months of Inventory
Days on the Market
Rental Demand Rating

Target variables

Target variables from PropertyData API (originally from the house price index & property portals in the case of the rent)

Average_Rent
Average_Sale_Per_Sqf_Price
Average_Sale_Price

Table 3. List of all the variables used.

For the data acquired from the EPC dataset, the mean of the data points (e.g. area, construction year) from the surrounding postcodes was used to impute missing values. This process of data completion might be improved by defining a Poisson model to predict the distribution of the construction ages on postcodes with no available information.

The age groups aggregation was defined by people with similar needs. The proximity of certain POIs such as primary schools, secondary schools, universities, and hospitals had a strong correlation with price and age groups. Modelling with the original age groups was tested but yielded worst prediction results. The results of the in-depth interviews also put an emphasis on the working population versus school aged children and the retired population. As such a variable called *TotalModPop*, which contains the % of active population, those between 30 and 65 who are of working age, can be indicative of demand in the borough. The commute types aggregation was based on intuition.

The performance improved significantly when using the GooglePlaces API variables. For instance to begin with the strongest predictor of the price was the restaurant variable used from the PropertyData API. After using the restaurant ratings from the Google Places API with the grouping that was described in the methods section above the results improved. This highlights the precaution that needs to be taken when using 'ready-made' data solutions from APIs. In this case the inner working of the restaurant rating variable from PropertyData API is not made public hence the black box phenomenon is replicated here. The decision was made to drop this variable.

4.3 Model performance

Three models were developed to predict the price, price per square foot, and rental value of a property. The methods used were all describe in section 3.3.1 above. Since the price model was not as accurate as the price per square foot model, the decision was made to calculate the estimated price using the mean of the area of properties in each postcode. High degree of caution in terms of using mean property area is to be taken as it can vary greatly or missing from the EPC or the PropertyData API datasets. Implementing the price prediction model directly was tested, however while slightly improving the predictions, this significantly reduced the speed in the deployment. Using virtual machines, which comes at a potentially high cost, would address this issue.

Furthermore, to improve the prototype performance it was decided that while the price per square foot model is generated on demand for a given postcode, the rental model is pre-trained for all the district and saved. The feature ranking used in the prototype is obtained from the Price SFA model alone, which is trained each time a postcode is inputted.

Figure 8 show the process of training the two models to predict the price per square foot and rental price on the 2km-vicitinity data stored in the Django container. The figure also shows how the error of the price per square foot model is presented as mean price variance. Additionally, the figure shows how the prototype was designed to compute the mean property price from the mean price per square foot, and the mean area in the given postcode. It is important to remember that the hyperparameters for these best performing models here were used for the two models that were used in the deployment: Model 1 - Price per square foot and Model 2 - Rent prediction models.

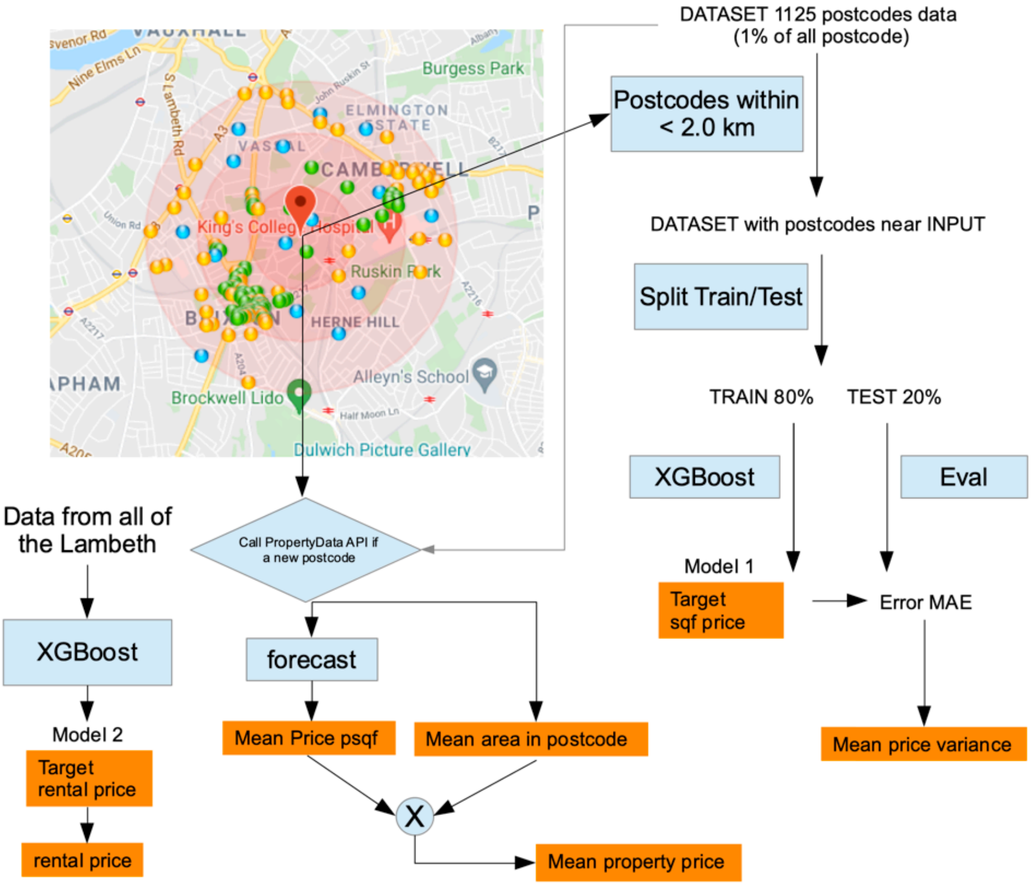


Figure 8. The two models used how they relate to the metrics displayed in the front end.

The reasons for choosing XGBoost versus other machine learning algorithms that were tested were given in 4.2.1 and 3.2.4. One method that came close to XGBoost was Ridge Regression which had an R2 score on the test data of 0.80 versus that of the XGBoost of 0.84. Table 4 shows the chosen hyperparameters for each of the three XGBoost models that were trained and tested in the model development pipeline.

Parameter	Model		
	Price	*Price per Square foot*	*Rent*
n_estimators	501	502	492
learning_rate	0.066	0.086	0.05
min_split_loss	1.50	2.40	1.01
min_child_weight	3	2	1
reg_lambda	0.10	0.016	4.60
reg_alpha	0.33	8.11	9.61
colsample_bytree	0.78	0.66	0.69
subsample	0.48	0.96	0.73

Table 4. Hyperparameter values for all three XGBoost models used in the project.

Figure 5 shows which hyperparameters were tuned and chosen and the cross validation used. The results mentioned here relate to that modelling pipeline. Vargas – Calderon [83] states that the magnitude of the error has to be compared with the natural price variance offered by the market. This was the most intuitive way of putting the error and accuracy of the models in the context that the experts understand. Hence the decision to include two metrics to evaluate model performance: *MAE* in percentage terms and the *standard deviation of the error* in pounds sterling. The Mean Absolute Error in percentage terms is the relative error expressed in percentages and defined by the ratio of the MAE to the mean value that is being measured. The mean values were calculated using the available dataset; for price the mean value was 5.939672e+05 or £593,967.2, for price per square foot the mean value was £715.21 per square foot, and for rent the mean value was £409.96 per week. The standard deviation in

Table 5 is the error dispersion of the models; a low standard deviation means errors are clustered around the mean error, and high standard deviation indicates errors are more spread out. The MAE in % and the standard deviation for the models used in the prototype are given in

Table 5 below.

Error type	Model		
	Price	Price per Square foot	Rent
MAE %	5.39%	2.23%	2.27%
Standard Dev	£32,019	£16.21	£9.29

Table 5. Performance of the best models

Since the MAE % was calculated using the mean price and rent form the data, it was decided to compare the results with another source of the mean price and rent.

The MAEs for the models are considered acceptable when looked at alongside the standard deviation. A deviation of £32,019 in the London borough of Lambeth with an average price of £525,268 in July 2020 [84] translates to a 6% margin of error. Put into a price growth context, Lambeth saw a year-on-year growth of 3.7% in July 2020, which is roughly half of the error margin of the price prediction model. Given the very diverse range of housing and the wide distribution of the price in Lambeth, this margin of error would be acceptable. However, there is room for improvement, which can be made through using the price prediction model that was developed in the modelling pipeline, in the deployment. As mentioned, this doubled the processing time and was not considered the right decision at this stage.

Assessing the rental prediction model is more difficult as the data for private rental values in the UK is not recorded as rigorously as that of sold prices. The Office for National Statistics, ONS, publishes the median monthly rental prices for the private rental market in England [85]. The mean monthly rent for Lambeth was £1,731 between October 2020 to September 2021, which is £20,772 per annum, or £399.46 per week. The standard deviation from the model is £9.29 which translates roughly to a 2.32% margin of error. This a good result as this variation in this region are not likely to make a material difference in the decision making.

It was observed that the XGBoost models to be not very stable in the sense that the structure of the trees that created each time the model was run were different. The reason for this is that these models are stochastic. While the results were consistently satisfactory once the hyperparameter tuning was performed, this instability caused problems when it came to ranking the variables in terms of the contribution to the models output. To overcome this a process of ranking was used to arrange the variables based on the number of times that variables appear in the trees.

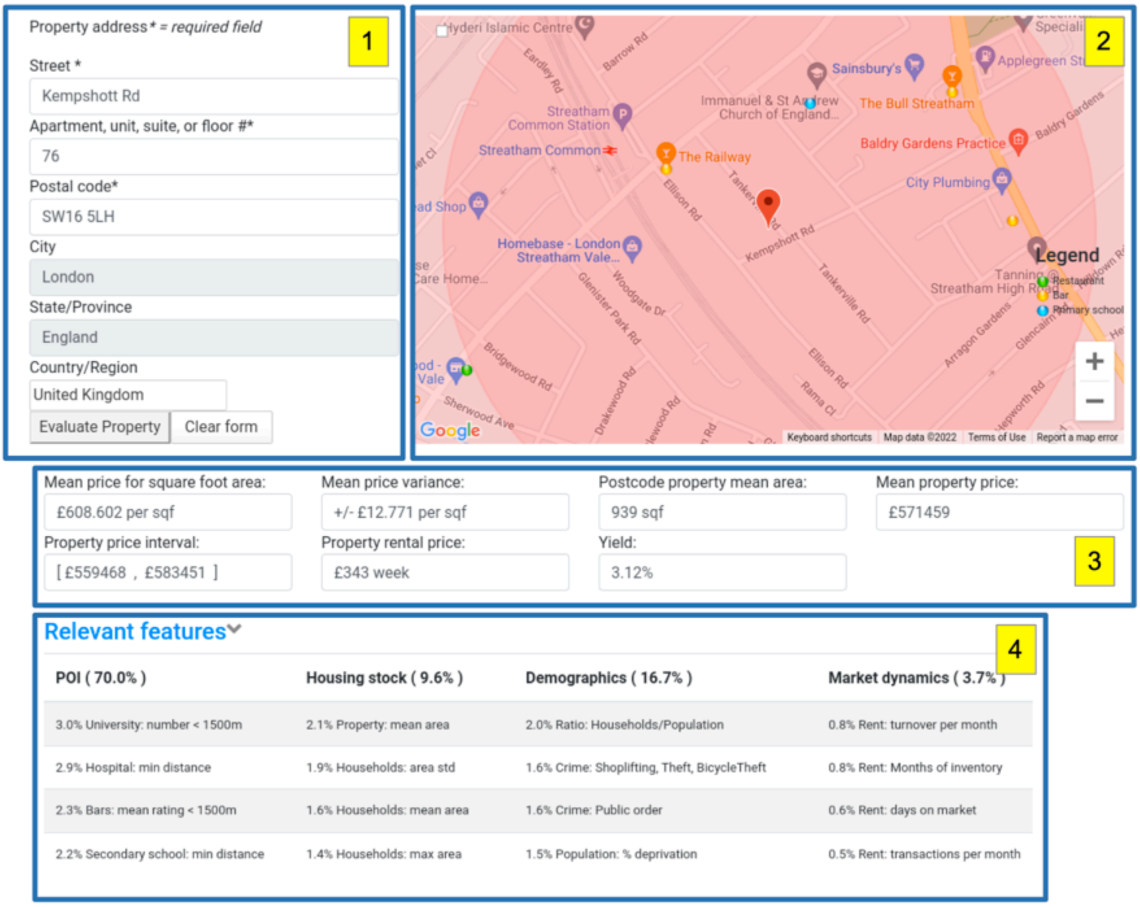
While developing the modelling pipeline, it was decided that the gain was the best metric to use when ranking the variables in order of their contribution to the prediction. When two regression trees, for the same problem, are trained with different training data the resulting trees can very different. They can have similar perdition accuracy and error while at the same time having very different structures. Theoretically, this instability can be mitigated by using the gain to rank the features. However, it is important to point out that in practice with the limited data size in this project, this instability cannot be mitigated. As proof, the rank generated using the data for all the districts is more stable than that using data <2km from a postcode. On a practical level, the instability can be hidden if a seed is fixed using a random number generator. The reason why the models differ so much is the result of splitting the data into train and test sets randomly. Fixing the splitting point of the data into train and test could ensure that the same model generated in all the runs on the prototype. However, this 'quick-fix' can not be misused to hide the model's inefficiencies in deployment and other model improvement measures have to be tried.

4.4 Front end

There are several points that need addressing before such system can be deployed properly. The most obvious is that local repositories of data were form to for model training and testing in the approach used. This is common in developing and testing stage, however for deployment a policy has to be defined to update the local database.

Although this data acquisition pipeline was defined to download the data into local repositories for the development phase, going forward it can be adopted to update on-demand for production.

The front-end was in the end defined by 4 components: 1- Data input fields, 2- Map with the property and POI in the 2km vicinity, 3- Model output fields, 4- Variables rankings grouped by the four categories. These can be seen with their labels in Figure 9.



Property address * = *required field*

Street *

Kempshott Rd

Apartment, unit, suite, or floor #*

76

Postal code*

SW16 5LH

City

London

State/Province

England

Country/Region

United Kingdom

Mean price for square foot area: £608.602 per sqf

Mean price variance: +/- £12.771 per sqf

Postcode property mean area: 939 sqf

Mean property price: £571459

Property price interval: [£559468 , £583451]

Property rental price: £343 week

Yield: 3.12%

Relevant features

POI (70.0%)	Housing stock (9.6%)	Demographics (16.7%)	Market dynamics (3.7%)
3.0% University: number < 1500m	2.1% Property: mean area	2.0% Ratio: Households/Population	0.8% Rent: turnover per month
2.9% Hospital: min distance	1.9% Households: area std	1.6% Crime: Shoplifting, Theft, BicycleTheft	0.8% Rent: Months of inventory
2.3% Bars: mean rating < 1500m	1.6% Households: mean area	1.6% Crime: Public order	0.6% Rent: days on market
2.2% Secondary school: min distance	1.4% Households: max area	1.5% Population: % deprivation	0.5% Rent: transactions per month

Figure 9. Front-end screenshot with the four different sections.

Property information is inputted in (1). The address field is auto-completed and it is submitted by pressing the button “Evaluate Property”. After the submission, two models are trained, using the property postcode as a reference, a first model to predict the mean price per square foot and a second model to predict the mean property rental price. Both models use the same hyperparameters that were determined in the model development phase (section 3.2.4). The property is located and shown on the map which uses Google Maps API (2) and model perdition is used to populate the output fields. Model 1 is the price per square fit model, Model 2 is the rent model, as per Figure 8.

1- “Mean price for square foot area” is model 1’s prediction

2- “Mean price variance” is defined by model 1’s error interval

3- “Postcode property mean area” is the mean area of properties in the given postcode

4- “Mean property price” is “mean area” times the “price per square foot” predicted by model 1

5- “Property price interval”

6- “Property rental price” by week is model 2’s prediction

7- “yield” in percentage terms calculated as: annual rental income (weekly rental income x 52) / property value x 100 ·

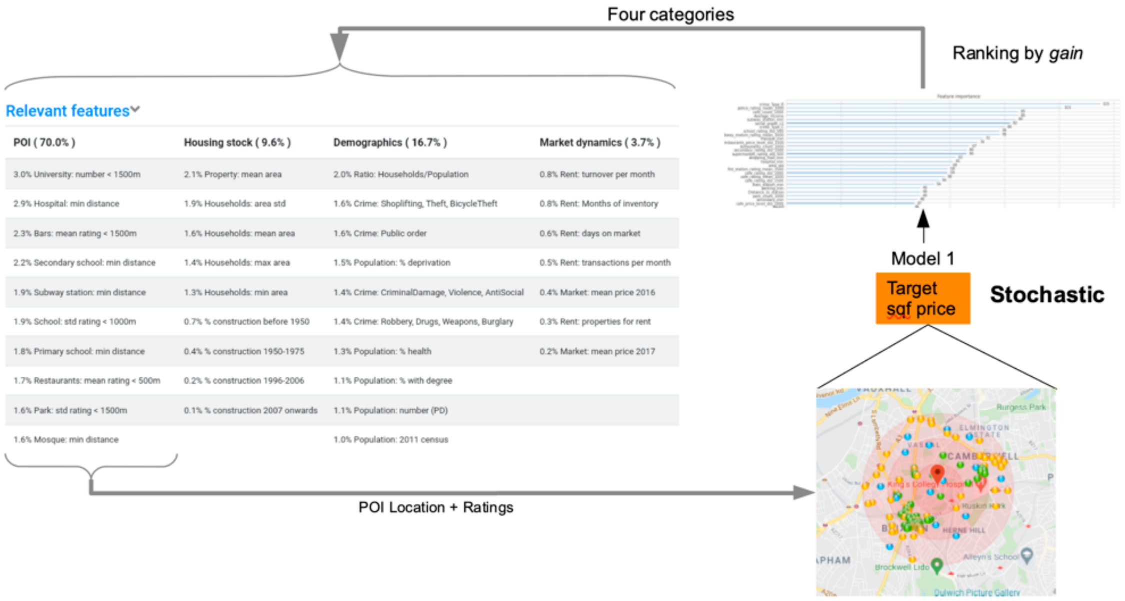
All but the yield were outlined in Figure 8 and explained in 4.3.

The rank for the features used in the model to predict the price per square foot, are classified into four categories that correspond to the green categories from the in-depth interviews.

1- POI (point of interest)
2- Housing Stock
3- Demographics
4- Market Dynamics

For each of the variables or features in these categories a percentage is given which presents the contribution that the specific feature is making to the predicted mean price per square foot and the price. Also, an overall percentage is given for each of the four categories which represents the total contribution that the specific category is making to the predicted mean price per square foot.

Figure 10 shows the how the variables are ranked and grouped into the four categories, before being displayed on the map. Hovering over these POIs will reveal their sub-categories (e.g. bar, school, etc) and their rating (from 1 to 5).



Relevant features

POI (70.0%)	Housing stock (9.6%)	Demographics (16.7%)	Market dynamics (3.7%)
3.0% University: number < 1500m	2.1% Property: mean area	2.0% Ratio: Households/Population	0.8% Rent: turnover per month
2.9% Hospital: min distance	1.9% Households: area std	1.6% Crime: Shoplifting, Theft, BicycleTheft	0.8% Rent: Months of inventory
2.3% Bars: mean rating < 1500m	1.6% Households: mean area	1.6% Crime: Public order	0.6% Rent: days on market
2.2% Secondary school: min distance	1.4% Households: max area	1.5% Population: % deprivation	0.5% Rent: transactions per month
1.9% Subway station: min distance	1.3% Households: min area	1.4% Crime: CriminalDamage, Violence, AntiSocial	0.4% Market: mean price 2016
1.9% School: std rating < 1000m	0.7% % construction before 1950	1.4% Crime: Robbery, Drugs, Weapons, Burglary	0.3% Rent: properties for rent
1.8% Primary school: min distance	0.4% % construction 1950-1975	1.3% Population: % health	0.2% Market: mean price 2017
1.7% Restaurants: mean rating < 500m	0.2% % construction 1996-2006	1.1% Population: % with degree	
1.6% Park: std rating < 1500m	0.1% % construction 2007 onwards	1.1% Population: number (PD)	
1.6% Mosque: min distance		1.0% Population: 2011 census	

Figure 10. Relationship between the ranking, front-end and the displayed map.

5 Discussion & Reflection

The aim of this project was application-oriented in that it set out to create a valuation model to serve a business purpose with input from industry experts.

A major hurdle was finding people with in-depth industry knowledge, as well as some data literacy, who use appraisal tools to make investment decisions. Anything to bridge this gap is welcome. The aim was to bring qualitative insights into a qualitative industry through a systematic and robust research process. One way is to make the AVM more usable and explainable while maintain a high predictive power.

5.1 Objective 1

The first objective of the project, OB1: *Can the use of alternative data sources, as opposed to price paid datasets, improve the price and rental prediction models in a property investment appraisal context?*

Upon reflection on the project, to address OB1, the approach was three-fold:

1) doing the in-depth interviews to identify useful alternative data sources which add value to the experts' decision-making process. The useful data sources were all geospatial.

2) making a conscious decision to exclude the most commonly used data source for making predictions, i.e. historic price paid data. The only market related data source was growth percentages, which were used to address OB2 (see 5.2)

3) using state-of-the-art models for the prediction task based on a thorough literature review. This was done partly based on the need for transparency in OB2 (see 5.2).

It was observed that there are advantages and disadvantages to using 'alternative' data sources. The advantage is how it can overcome the lag that exits with price paid data availability. It takes in some cases up to six months for this data to update and made available for analysis. In a fast-moving market where the expert have to make decisions quickly, this lag is unhelpful to say the least. On the other hand, POI data from GooglePlaces API or Police UK are updated on a daily or weekly basis. It is also noted that not all alternative data sources can overcome this lag. For instance, the census data is updated every 10 years.

Data quality emerged as a constant battle too. The importance of validating the data source was highlighted in using a few of the end-points from PropertyData API, where the data providers had already processed and aggregated the data with minimal disclosure as to how. To address this issue, trial and error was adopted throughout the project to find the best variables to achieve best results with the model.

Some alternative data sources require to be benchmarked against other metrics to make them more robust. For instance, for this project the distances for the POIs were chosen to be x<500m, 500m<x<1000m, and 1000m<x<1500m based on the in-depth interviews and intuition. To make this assumption more robust, benchmarking this against already developed walkability indices like one developed by Stockton et al [86] could help with developing a more reliable and less subjective data source. Also, more POIs can be included from the Places API if there are no financial restrictions.

XGBoost yielded accurate results i*n the context of property prices and rental values in the London Borough of Lambeth,* as discussed in section 4.3.

5.2 Objective 2

The second objective of the project was, OB2: *How can the use of alternative data sources impact, and potentially benefit, the industry experts who use property appraisal tools?*

Working towards achieving this objective, in-depth interviews were carried out and qualitative analysis carried out. The need for a robust predictive model that is explainable, and outputs useful information was highlighted. The following themes and categories emerged:

Three themes: alternative data, accurate models, transparency (move away from black box)

Four categories: POI, Housing Stock, Demographics, Market Dynamics

The culmination of the yellow themes and green categories can be summarised as the need for a tool that is robust in its predictive power, is explainable, and outputs useful information; hence an appraisal system that addresses **usability, predictive power, explainability**. This was address by building robust accurate models that predict the price and the rent for a given postcode (predictive power). The features or variables that were used in the model training are ranked in order of their contribution to the prediction and grouped by each of the four categories (explainability).

The metrics displayed in on the front-end: calculated using the average price per square foot and average rent in the vicinity of a postcode, which immediately puts them into context for the appraiser (usability). The mean property price variance and the property price interval metrics are calculated using the error metrics of the predictive models (explainability).

The themes and four categories were incorporated in every step of the methods used for this project: the four categories were incorporated into the data acquisition, feature extraction, and feature engineering pipelines to maintain coherence. The theme of *predictive power* was incorporated in the model building pipeline, and the deployment pipeline incorporates usability and *explainability themes*.

It is important to acknowledge the pitfalls and shortcoming of the qualitative approach adopted. Small sample size was the main one, which left us open to bias. But also inevitable in this case since the number of individuals with such in-depth knowledge of the property market from different perspective, and some knowledge of using data tools is limited. Although an inductive approach was adopted, it has to be noted that the data generated in the qualitative phase is inevitably part author-led and part expert-led. Interviewee selection was very carefully considered to minimise some of these shortcomings. The research questions were examined from the perspective of those with extensive knowledge of the property investment landscape who are also 'users' of valuation tools (IA, IB, IC), those with knowledge and experience of being the 'creator' of the valuation tools (ID), and those with long term experience in investing in property but no/limited experience of leverage technology or data to make decisions (IE).

6 Conclusion

6.1 Reflecting on project plan

To assess whether the project has achieved what it set out to achieve, one key question could be considered: do the results from the quantitative phase confirm or deny what the experts said in the qualitative phase? The method of combining qualitative research and quantitative research helped come up with the conclusion that the results from the second phase confirms the findings of the first phase.

Using two data generation and analysis methods, proved useful and allowed to corroborate the findings and enhance their validity, which is referred to as method triangulation by Oates [14]. The findings here were the advantages of using alternative data in the context of building a usable and explainable tool that accurately predicts price and rent for a property. Specifically mixing robust qualitative and quantitative data generation and analysis methods in research can be a useful tool to bridge the gap between the data science and property industry experts. Creating more explainable tools for industry experts requires their participation in the development of the tools.

6.2 Future work

Future work could see a vast number of alternative data sources being explored to enrich the theme of explainable. Building 'neighbourhood profiles', where neighbourhoods are classified based on the key variables that were used to predict the price would be of interest. Using *k*-means clustering and using different mixes of variables could be one solution to building such an area classifier. The outcome would be a useful tool to identify 'up and coming' areas. Furthermore, significant visualisation improvements can be made in the front end, as well as building different data gathering and deployment pipelines for the appraisal tool to cover the whole of the UK.

7 References

[1] G. Hammond, “Financial Times,” 31 August 2021. [Online]. Available: https://www.ft.com/content/662d3839-2c58-4b88-8df8-71d8f0af85d5. [Accessed 02 September 2021].

[2] Hamptons, “Number of landlords falls to 7 year low,” Hamptons International, February 2020. [Online]. Available: https://www.hamptons.co.uk/research/articles/2020/lettings-index-january-2020.pdf/. [Accessed 21 August 2021].

[3] G. M. Asaftei, S. Doshi, J. Means and A. Sanghvi, “Getting ahead of the market: How big data is transforming real estate,” 2018. [Online]. [Accessed 21 July 2021].

[4] Investment Property Forum, “Mitigating behavioural influences on decision making within the property investment process,” Investment Property Forum, London, 2021.

[5] A. Baum, N. Crosby, P. Gallimore, P. McAllister and A. Gray, “The Influence of Valuers and Valuations on the Workings of the Commercial Property Investment Market,” Research funded by the Education Trusts of the Investment Property Forum, Jones Lang LaSalle and the Royal Institution of Chartered Surveyors, Reading, 2011.

[6] International Valuation Standards Council, “Annual Report 2019-20,” IVSC , 2020. [Online]. Available: https://www.ivsc.org/wp-content/uploads/2021/08/IVSCANNUALREPORT2019-20.pdf. [Accessed 2021].

[7] A. E. Baum, N. Crosby and S. Devaney, Property Investment Appraisal, Wiley-Blackwell, 2021.

[8] E. Pagourtzi, V. Assimakopoulos and N. French, “Real estate appraisal: a review of valuation methods,” *Journal of Property Investment & Finance,* vol. 21, no. 4, p. 2003, 2003.

[9] D. Jenkins, O. Lewis, N. Almond, S. Gronow and J. Ware, “Towards an intelligent residential appraisal model,” *Journal of Property Research,* vol. 16, no. 1, pp. 67-90, 1998.

[10] P. G. D. L. Julian Diaz III, “Residential valuation behaviour in the United States, the United Kingdom and New Zealand,” *Journal of Property Research,* vol. 19, no. 4, pp. 313-326, 2002.

[11] P. Gallimore, “Confirmation bias in the valuation process: a test for corroborating evidence,” *Journal of Property Research* , vol. 13, no. 4, pp. 261-273, 1996.

[12] P. P. Gray, “Independent review of real estate investment valuations,” Royal Society of Chartered Surveyors , London, 2021.

[13] N. Crosby and J. Henneberry, “Financialisation, the valuation of investment property and the urban built environment in the UK,” *Urban Studies Journal,* vol. 53, no. 7, p. 1424–1441, 2016.

[14] B. J. Oates, Researching Information Systems and Computing, Sage Publications Ltd., 2006.

[15] S. S. S. Das, M. E. Ali, Y.-F. Li, Y.-B. Kang and T. Sellis, “Boosting house price predictions using geo-spatial network embedding,” *Data Mining and Knowledge Discovery,* 2021.

[16] S. Rosen, “Hedonic prices and implicit markets: product differentiation in pure competition,” *Journal of Political Economy,* vol. 82, no. 1, pp. 34-55, 1974.

[17] A. Owusu-Ansah, “A review of hedonic pricing models in housing research,” *A Compendium of International Real Estate and Construction,* vol. 1, p. 17–38, 2013.

[18] A. C. Goodman and T. Thibodeau, “Housing market segmentation and hedonic prediction accuracy,” *Journal of Housing Economics,* vol. 12, no. 3, pp. 181-201, 2003.

[19] G. Lisi and M. Iacobini, “Estimation of a Hedonic House Price Model with Bargaining: Evidence from the Italian Housing Market,” *Economics and Econometrics Society,* vol. 56, no. 1, pp. 61-73, 2013.

[20] R. Chau, C. H. Yeh and K. A. Smith, “Developing a personal multilingual Web space,” in *IASTED International Multi-Conference on Applied Informatics*, 2003.

[21] S. C. Bourassa, M. Hoesli and V. S. Pengd, “Do housing submarkets really matter?,” *Journal of Housing Economics,* vol. 12, pp. 12 - 28, 2003.

[22] T. Fik, D. Ling and G. Mulligan, “Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach,” *Real Estate Economics* , vol. 31, no. 4, pp. 623-646, 2003.

[23] L. Hu, S. He, Z. Han, H. Xiao, S. Su, M. Weng and Z. Cai, “Monitoring housing rental prices based on social media:An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies,” *Land Use Policy,* vol. 82, pp. 657-673, 2019.

[24] G. Li, Z. Cai, X. Liu and J. L. &. S. Su, “A comparison of machine learning approaches for identifying high-poverty counties: robust features of DMSP/OLS night-time light imagery,” *International Journal of Remote Sensing,* vol. 40, no. 15, pp. 5716-5736, 2019.

[25] F. Z. W. P. S. G. J. R. F. D. C. R. Yuhao Kang, “Understanding house price appreciation using multi-source big geo-data and machine learning,” *Land Use Policy,* vol. 111, p. 104919, 2021.

[26] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez and C. Afonso, “Identifying Real Estate Opportunities Using Machine Learning,” *Applied Sciences,* vol. 8, no. 2321, 2018.

[27] W. K. Ho, B.-S. Tang and S. W. Wong, “Predicting property prices with machine learning algorithms,” *Journal of Property Research,* vol. 38, no. 1, pp. 48-70, 2020.

[28] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, pp. 5-32, 2001.

[29] L. W. A. S. P. G. Gilles Louppe, “Understanding variable importances in forests of randomized trees,” in *Gilles Louppe, Louis Wehenkel, Antonio Sutera, Pierre Geurts*, Lake Tahoe, 2013.

[30] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *The Annals of Statistics,* vol. 29, no. 5, pp. 1189-1232, 2001.

[31] T. H. F. Tibshirani, he elements of statistical learning: data mining, inference, and prediction, New York: Springer, 2001.

[32] M. Kearns and L. Valiant, “Cryptographic limitations on learning Boolean formulae and finite automata,” *Journal of the ACM,* vol. 41, no. 1, pp. 67-95, 1994.

[33] R. E. Schapire, “The strength of weak learnability,” *Machine Learning,* vol. 5, p. 197–227, 1990.

[34] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 2016.

[35] Y. Zhao, G. Chetty and D. Tran, “Deep Learning with XGBoost for Real Estate Appraisal,” in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Xiamen, 2019.

[36] H. Wu and C. Wang, “A new machine learning approach to house price estimation,” *New Trends in Mathematical Sciences,* vol. 4, no. 6, pp. 164 - 171, 2018.

[37] V. Koktashev, V. Makeev, E. Shchepin and V. Tynchenko, “Pricing modeling in the housing market with urban infrastructure effect,” *Journal of Physics Conference Series* , vol. 1353, no. 1, p. 1353, 2019.

[38] T. Lasota, M. S. Bogdan Trawinski and Z. Telec, “Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms,” *International Journal of Applied Mathematics and Computer Science,* vol. 22, no. 4, 2012.

[39] S. Borde, A. Rane, G. Shende and S. Shetty, “Real Estate Investment Advising Using Machine Learning,” *International Research Journal of Engineering and Technology* , vol. 4, no. 3, p. 1821–1825, 2017.

[40] J. Liu and X. Huang, “Forecasting Crude Oil Price Using Event Extraction,” *IEEE Access,* vol. 9, p. 149067 – 149076, 2021.

[41] Z. Peng, Q. Huang and Y. Han, “Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm,” in *IEEE 11th International Conference on Advanced Infocomm Technology*, Jinan, 2019.

[42] M. De Nadai and B. Lepri, “The economic value of neighborhoods: Predicting real estate prices from the urban environment,” in *Proceedings of IEEE DSAA*, Turin, 2018.

[43] R. K. Pace and O. Gilley, “Generalizing the OLS and grid estimators,” *Real Estate Economy,* vol. 23, no. 2, pp. 331-347, 1998.

[44] S. C. Bourassa, M. Hoesli and V. S. Pengd, “Do housing submarkets really matter?,” *Journal of Housing Economics,* no. 12, pp. 12-28, 2003.

[45] J. Cortright, “Walking the walk: How walkability raises home values in us cities,,” CEOs for Cities, Durham, NC, 2009.

[46] Y.-H. Chen and T. Fik, “Housing-market bubble adjustment in coastal communities - A spatial and temporal analysis of housing prices in Midwest Pinellas County, Florida,” *Applied Geography,* vol. 80, pp. 48-63, 2017.

[47] Y. Huang, “A study of sub-divided units (SDUs) in Hong Kong rental market,” *Habitat International,* vol. 62, pp. 43-50, 2017.

[48] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applicaations,* vol. 42, pp. 2928-2934, 2015.

[49] A. Pavlov and T. Somerville, "Immigration, Capital Flows and Housing Prices," *Real Estate Economics,* vol. 48, no. 3, pp. 915-949, 2020.

[50] M. H. Rafiei and H. Adeli, "A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units," *Journal of Construction Engineering and Management ,* vol. 142, no. 2, 2016.

[51] T. J. Brooks, B. R. Humphreys and A. Nowak, " Strip Clubs, "Secondary Effects" and Residential Property Prices," *Real Estate Economics,* vol. 48, no. 3, pp. 850-885, 2020.

[52] R. F.-Y. Lin, C. Ou, K.-K. Tseng, D. Bowen, K. Yung and W. Ip, "The Spatial neural network model with disruptive technology for property appraisal in real estate industry," *Technological Forecasting and Social Change,* vol. 173, p. 121512, 2021.

[53] Y. Kang, F. Zhang, W. Peng, S. Gao, J. Rao, F. Duarte and C. Ratti, "Understanding house price appreciation using multi-source big geo-data and machine learning," *Land Use Policy,* vol. 111, p. 104919, 2021.

[54] Y. Liu, F. Wang, Y. Xiao and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai," *Landscape and Urban Planning,* vol. 106, no. 1, pp. 73-87, 2012.

[55] S. Gao, K. Janowicz and H. Couclelis, "Extracting urban functional regions from points of interest and human activities on location-based social networks," *Transactions in GIS,* vol. 21, no. 3, pp. 446-467, 2017.

[56] Yang Yue, Y. Zhuang, A. G. O. Yeh, J.-Y. Xie, C.-L. Ma and Q.-Q. Li, "Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy," *International Journal of Geographical Information Science,* vol. 31, no. 4, pp. 658-675, 2017.

[57] I. Pérez-Rave, J. C. Correa-Morales and F. González-Echavarría, "A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes," *Journal of Property Research,* vol. 36, no. 1, pp. 59-96, 2019.

[58] U. I. I. S. T. H. E. M. Jørgen Ødegård, "Large-scale genomic prediction using singular value decomposition of the genotype matrix," *Genetics Selection Evolution,* vol. 50, no. 6, 2018.

[59] G. l. Clark, "Stylized Facts and Close Dialogue: Methodology in Economic Geography," *Annals of the Association of American Geographers,* vol. 88, no. 1, pp. 73-87, 1998.

[60] S.-C. Noh and J.-H. Park, "Café and Restaurant under My Home: Predicting Urban Commercialization through Machine Learning," *Sustainability,* vol. 13, pp. 56-99, 2021.

[61] S. M. S. C. S. A. E. Sam Mosallaeipour, "A robust expert decision support system for making real estate location decisions, a case of investor-developer-user organization in industry 4.0 era," *Journal of Corporate Real Estate,* vol. 22, no. 1, pp. 21-47, 1029.

[62] J. Kilpatrick, "Expert systems and mass appraisal," *Journal of Property Investment & Finance,* vol. 29, no. 4/5, pp. 529-578, 2011.

[63] M. Barton, "An exploration of the importance of the strategy used to identify gentrification," *Urban Studies,* vol. 53, no. 1, pp. 92-111, 2016.

[64] F. Bergadano, R. Bertilone, D. Paolotti and G. Ruffo, "Learning Real Estate Automated Valuation Models from Heterogeneous Data Sources," *International Journal of Real Estate Studies,* vol. 15, no. 1, pp. 71-85, 2021.

[65] Office for National Statistics, " Census geography," Office for National Statistics, 2021. [Online]. Available: https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography. [Accessed 2021].

[66] J. Reades, J. D. Souza and P. Hubbard, "Understanding urban gentrification through machine learning," *Urban Studies,* vol. 56, no. 5, pp. 922-942, 2019.

[67] Google, "Google Maps Platform - Places API," Google, 2021. [Online]. Available: https://developers.google.com/maps/documentation/places/web-service/overview. [Accessed 9 10 2021].

[68] PropertyData, "Property Analytics API - PropertyData," Liberty Tech Limited, [Online]. Available: https://propertydata.co.uk/api. [Accessed 10 November 2021].

[69] C. Bell, “Postcodes in Lambeth, London Borough,” Doogal, 2021. [Online]. Available: https://www.doogal.co.uk/AdministrativeAreas.php?district=E09000022#google_vignette. [Accessed 29 August 2021].

[70] Transport for London (TFL), “Public Transport Accessibility Levels,” Tranport for London (TFL), 2017. [Online]. Available: https://data.london.gov.uk/dataset/public-transport-accessibility-levels. [Accessed 2021].

[71] Department for Levelling Up, Housing & Communities, “Energy Performance of Buildings Data: England and Wales,” Department for Levelling Up, Housing & Communities, 2021. [Online]. Available: https://epc.opendatacommunities.org/. [Accessed 2021].

[72] Y. Yang and A. V. Diez-Roux, “Walking Distance by Trip Purpose and Population Subgroups,” *American journal of preventive medicine,* vol. 43, no. 1, pp. 11-19, 2012.

[73] T. C. M. H. C. M. David Canter, “Predicting Serial Killers' Home Base Using a Decision Support System,” *Journal of Quantitative Criminology,* vol. 16, no. 4, pp. 457-478, 2000.

[74] K. D. Harries, Mapping crime: principle and practice, University of Michigan Library , 1999.

[75] S. Sarangi and D. Youngs, “Spatial patterns of Indian serial burglars with relevance to geographic profiling,” *Journal of Investigative Psychology and Offender Profiling,* vol. 3, no. 2, pp. 105-115, 2006.

[76] K. Rossmo, “Geographic heuristics or shortcuts to failure?: response to Snook et al.,” *Applied Cognitive Psychology,* vol. 19, no. 5, pp. 651-654, 2005.

[77] D. Paulsen, “Human versus machine: a comparison of the accuracy of geographic profiling methods,” *Journal of Investigative Psychology,* vol. 3, no. 2, pp. 77-89, 2006.

[78] M. Tonkin, J. Woodhams and J. W. Bond, “A Theoretical and Practical Test of Geographical Profiling with Serial Vehicle Theft in a UK Context,” *Behavioral Sciences & the Law,* vol. 28, no. 3, pp. 442-460, 2009.

[79] Office for National Statistics, “ The National Statistics Socio-economic classification (NS-SEC),” Office for National Statistics, 2010. [Online]. Available: https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010. [Accessed 2021].

[80] W. D. Rubinstein, Twentieth-Century Britain: A Political History, Basingstoke: Palgrave, 2003.

[81] S. Li, Y. Jiang, S. Ke, K. Nie and C. Wu, “Understanding the Effects of Influential Factors on Housing Prices by Combining Extreme Gradient Boosting and a Hedonic Price Model (XGBoost-HPM),” *Land,* vol. 10, no. 5, p. 533, 2021.

[82] K. J. A. R. E. G. J. B.-T. M. H. B. R. A. T. Liam Li, “A System For Massively Parallel Hyperparameter Tuning,” in *The 3rd MLSys Conference*, Austin, 2020.

[83] V. Vargas-Calderon and J. E. Camargo, “Towards robust and speculation-reduction real estate pricing models based on a data-driven strategy,” *Journal of the Operational Research Society ,* vol. DOI: 10.1080/01605682.2021.2023672, 2022.

[84] Land Registry, “House Price Statistics,” UK Hous Price Index, 2020. [Online]. Available: https://landregistry.data.gov.uk/app/ukhpi/browse?from=2019-07-01&location=http%3A%2F%2Flandregistry.data.gov.uk%2Fid%2Fregion%2Flambeth&to=2020-07-01&lang=en. [Accessed 2021].

[85] Office for National Statistics, “Private rental market summary statistics in England,” Office for National Statistics, 2021. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/privaterentalmarketsummarystatisticsinengland. [Accessed 2021].

[86] J. C. Stockton, O. Duke-Williams, E. Stamatakis, J. S. Mindell, E. J. Brunner and N. J. Shelton, “Development of a novel walkability index for London, United Kingdom: cross-sectional application to the Whitehall II Study,” *BMC Public Health volume ,* vol. 16, 2016.

[87] B. Chi and A. Dennett, “Shedding new light on residential property price variation in England: A multi-scale exploration,” *Urban Analytics and City Science,* vol. 28, no. 7, pp. 1895-1911, 2020.

[88] Ordinance Survey, “Points of Interst Classification Scheme Official,” December 2020. [Online]. Available: https://www.ordnancesurvey.co.uk/documents/product-support/support/points-of-interest-classification-scheme.pdf. [Accessed August 2021].

[89] J. Yee and A. Dennett, “Unpacking the Nuances of London’s Neighbourhood Change & Gentrification Trajectories,” Centre for Advanced Spatial Analysis University College London, London, 2021.

[90] OpenStreetMap, “OpenStreeMap - Map Features,” OpenStreetMap, [Online]. Available: https://wiki.openstreetmap.org/wiki/Map_features. [Accessed October 2021].

[91] J. K. B. Byeonghwa Park, “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data,” *Expert Systems with Applications,* vol. 42, p. 2928–2934 , 2015.

[92] M. Steurer, R. J. Hill and N. Pfeifer, “Metrics for evaluating the performance of machine learning based automated valuation models,” *Journal of Property Research,* vol. 38, no. 2, pp. 99-129, 2021.

8 Appendix 1 – Project Proposal

Developing an Appraisal System for Property Investment – a study of traditional versus non-traditional data sources

Babak Hessamian
Project Proposal

October 2021

1. Introduction

There are 2.7 million dwellings that are rented privately to 4.4 million households in the UK [1]. This sector is known as the Private Rented Sector or PRS, worth billions of pounds. Since the introduction of Buy to Let (BTL) mortgages in early 1990s, and the decline of social housing provision by the local councils in the 2000s, PRS has been almost entirely dominated by private landlords, whose number stood at 2.66 million in 2020 [2].

More recently residential property markets have been under scrutiny from the government, as with increased demand the rents may rise to quickly and by too much, making certain areas unaffordable. In 2016 the UK Chancellor of the Exchequer George Osborne introduced a series of measures designed to curb this. With an emerging Build to Rent sector (BTR), the UK's residential rental sector is changing with an increasing number of larger organisations getting involved. Most recently Lloyds Banking Group has set up its own rental homes brand Citra, with an aim to build a portfolio of 50,000 homes by 2030 [1].

Property as an investment is more complex than other types of investments like stocks and bonds, and equities. Due to the lack of minute-by-minute sales data, like in the stock market, property sales data is more infrequent and therefore scattered. Another layer of complexity is the fact that property is heterogenous, meaning no two properties are the same, or the same property can have different values for different buyers. Investments are valued based on the monetary returns they generate, which in the case of a property investment would be the income (rent) and capital return, if any, for the holding period [7]. For all investment types, the most basic measure of the return is the investment yield or rate of return. Yields reflect the market perception of risk and are used for valuation purposes in all asset classes. A 'good investment' can be seen as one that produces the highest total return in terms of income and capital return.

With the professionalisation of the PRS sector and the emergence of the BTR as a sector, there is an accelerated need for a whole host of software, including property investment appraisal systems. The process of appraising an investment property or a property developing opportunity can be unstructured, made more difficult by an at-times opaque world of private firms. Gauging both rental and sales demand is time-consuming and only done through subjective means, such as speaking to estate agents. Many firms rely on the 'gut-feeling' of a handful of individuals. These methods are broadly known as comparable models, where historic transaction prices, property characteristics, and subjective opinions take centre stage.

More recently there has been a push to go beyond the comparable model and harness nontraditional data to build house price prediction models. A McKinsey & Co research on predicting house prices found that the proportion of the predictive power attributed to features derived from proximity to points of interest (POI) and the quality of those POIs is between 26% and 32% respectively [3]. This was as opposed to 14% attributed to market performance (income, etc) and 18% and 12% to property performance (vacancy rate, etc) and property features (number of bedrooms, etc).

The first Research Question (RQ1) we are aiming to answer is:

> Can the use of non-traditional data improve price and rental prediction models in a property investment appraisal context?

We will expand the research question further after carrying out qualitative research in the form of interviews with property professionals who currently use property investment appraisal tools. This will allow us to explain the results and answer the second Research Question (RQ2):

> 'Why' does our prediction rely on certain features more than others?

Objectives: build a model that uses a bespoke combination of traditional and non-traditional data to help property investment decision-makers gain actionable insight.

Outcomes: a prototype of a tool that takes property information as input and outputs valuable information for the decision-makers to aid the property investment appraisal process.

Beneficiaries: Real Estate Advisory and Investment firms to serve their existing and future client base better. Property decision-makers looking to identify assets in low value areas that are rising in popularity, mortgage lenders that are looking to build a risk profile for a property in a specific area, those interested in the predictive power of non-traditional data in building models.

2. Critical Context

2.1. Literature Review

Investment Value was defined by the International Valuation Standards (IVSC) in 2019 as:

> …the value of an asset to a particular owner or prospective owner for individual investment or operational objectives. (IVSC 2019)

Our project focuses on the investment appraisal from a property investor's perspective. Investors can benefit from the difference between Investment Value and the Market Value of an asset [7], and this can have indications for potential gains or losses. Since valuation is a key part of investment appraisal, the focus of our literature review will be on research on predicting market value of properties. Once we predict the market and rental values of a property, we can put this in the investment context by considering the desired investment yield of an investor. Such valuation models are known in the literature as Automated Valuation Models (AVMs).

Existing work on house price prediction can roughly be split into three approaches:

a) Property feature centric, such as hedonic regression [16] which has been studied extensively and seeks to model the relationship between the features and the price of a property. These rely on pure statistical methods such as ordinary least squares (OLS) that take into account historic transaction data, but fail to capture the heterogenous aspects that property market.

b) Machine learning, such as Baldominos et al [26], who use four techniques of support vector regression, *k*-nearest neighbours, ensemble of regression trees, and multi-layer perceptron to identify opportunities in the housing market. They define opportunities as those houses listed on the market at a lower price substantially lower than the market price. This definition of opportunity can be translated into the investment appraisal world in terms of the need to achieve certain price level to as identifying potential investment opportunities

c) Location centric, which considers the spatial dependency between residuals in the regression models [43]. Bourassa et al [44] define spatial submarkets taking spatial dependence into account using principal component analysis and cluster analysis. A novel method is presented in [15] which the authors call geo-spatial network embedding. The method uses graph neural networks in the geospatial context of the points of interest and their quality.

There has been a growing trend of using alternative data in the machine learning approach. By alternative data we mean any data that is not directly related to the property and its features (historic transaction prices, number of bathrooms, year built, etc). According to McKinsey & Co's report [3] on big data in real estate, the authors define two categories for alternative data, which they call non-traditional data: dispersion of points of interest, or POI (how far is the nearest school, supermarket, café, etc.), and the quality of the POI (school ratings, supermarket brand, reviews for a café, etc.). Their research has shown that not only proximity to a POI can be a driver of property values, but the access to the right quantity, mix and quality of those POIs matter as well.

There are numerous examples of research using non-traditional data with machine learning techniques to predict house prices.

Therefore creating features to use in our models is a key part of this project. Addressing the "curse of dimensionality" is important as this might lead to overfitting as well as sparse data. As such, a more in-depth review of the literature on feature engineering and dimensionality reduction in the context of building predictive is to be carried out to find suitable state-of-the-art machine learning methods.

The literature mainly focuses on residential price valuations, and not on predicting rental values. We will apply the same methods to predict rental values, as an investment can only be appraised if the returns (rent in this case) can be predicted and modelled along with the price.

3. Approaches: Methods, Data, Tools, Implementation

We use a mixed-method approach, in that we will use a combination of qualitative and quantitative approaches. Our qualitative approach consists of interviewing people from the industry who perform property investment appraisals as part of their job. Our quantitative approach will consist of comparing predictive performance of various machine learning methods to those of the state-of-the-art and measure the contribution of the features we develop to the prediction. This will allow us to triangulate our results to get to the 'why' in our second questions. The artefact that we are looking to produce is not a software, but a model that uses IT to aid and enrich a property investment opportunity appraisal.

3.1. Methods

3.1.1. Qualitative:

There are several companies that provide an appraisal system, such as LandInsight, REalyse and Yuvoh Analytics. Access to the models and algorithms of Automated Values Service providers is limited due to the proprietary nature of these software. To avoid subjectivity, we will need to be systematic in our methods. The most obvious method would be to design a survey and distribute it amongst a statistically significant number of users of the tools. This would take too long and is outside the scope of this project. Instead, we will interview five individuals with experience in the real estate sector who perform property investment appraisals. These will be structured interviews and include open-ended discussion questions as well as scaled questions.

The scaled questions will focus on gauging the importance of the appraisal systems within the property investment sector. The open-ended questions will focus on: 1) Current practice, 2) Current appraisal tools, 3) Understanding what is useful to investment professionals, 4) Establishing areas of improvement with current tools. Together these will guide us in developing features for our prediction models in the quantitative stage of the project. The questions are presented in the table below.

#	Question	Purpose

Scaled questions (1 = not very, 2 = somewhat, 3 = neutral, 4 = very, 5 = crucial)		
1	How important is appraisals in your decision making process?	Gauging importance of topic
2	How important is being able to track what drives your decision making based on your appraisal?	
3	How important is to understand what drives a automated valuation service's (appraisal tool) decisions making?	
Open-ended questions		
4	How do you appraise potential investments?	Current practice
5	What data do you rely on when appraising a potential investment?	
6	If you use tools to appraise potential investments, which ones are they?	Current appraisal tools
7	What do you like about the tools/software you use?	
8	What do you think is missing from the tools/software you use?	
9	How important is being able to tell what data is driving the appraisal tool's decision making process?	Understanding what is useful to investment professionals
10	What do you include in your presentation of your appraisal results?	
11	How do you keep track of the effects of the different data sources on your appraisals?	
12	What qualifiers/metrics might help you weight up investments better?	
13	What kind of problems do you usually encounter when appraising potential investments?	Establishing areas of improvement with current tools
14	What would like to see improved in the appraisal systems?	

We will carry out a thematic analysis of the responses to our open-ended questions by initially dividing the data into three themes: segments with no relation to our overall Research Questions, segments that provide information to use as context for our project, segments that directly relate to our Research Questions [14]. Then, we will carefully study our responses to devise categories guided by existing theory or observations in the data. Finally, we look at themes and interconnections between our segments and categories. Throughout we will use visualisation to guide us.

3.1.2. Quantitative

There are two quantitative aspects to this project: (1) identifying and qualifying the data sets to use for feature engineering, (2) implementing state-of-the-art spatially based house prediction models. Our approaches will be drive by our qualitative research and literature review on state-of-the-art methods.

1) Our aim is to compare models that use traditional and non-traditional data sources. We will use the nontraditional data sources to develop features to use in our model. In developing features to use in the model we have to be systematic and use unsupervised methods such as principal component analysis and clustering, to reduce the dimension and select features [44].

2) We will discover all the state-of-the-art machine learning algorithms currently used in the domain of price prediction and real estate market analysis. These algorithms will be implemented using features engineered using both traditional and non-traditional data.

We elaborate on our qualitative methods in the implementations section below (3.4).

3.2. Data:

Since our project is partly based on developing features using different data sources, the starting point for our research method is to identify and assess the data sets. As mentioned in 2.1. above, we distinguish between traditional and non-traditional data.

3.2.1. Traditional

Traditional property data, for the purposes of this project, can be divided into market performance, property features, property performance. For market performance we will use the Price Paid Data (PPD) from Land Registry. There are other data sources available such as the regional house price index (HPI) from the Office for National Statistics (ONS) that we may use as additional features. Bin Chi et al [87] developed a price per square meter dataset by linking PPD to data from Ordnance Survey and Energy Performance Certificates. This will be one other source of market performance data for our project.

Property features data is available from two possible sources: land registry, which includes property type (detached house, semi-detached house, flat, etc), estate type (leasehold, freehold) or Zoopla property portal, which includes a vast amount of information about each property listing (e.g. number of bedrooms, floors, build type, garden, parking). Zoopla has an archive of millions of listings, but the listings go as far back as Zoopla started trading. On the other hand, PPD dates much further back, but lacks the richness of Zoopla's features data. Zoopla has a very comprehensive API, which makes it more attractive for use in our project. Additionally, we will be using rental values, which is only obtainable from current and historic.

Property performance data is not readily available and has to be constructed using market performance data, property features data, and domain knowledge (e.g. rental yield can be calculated as rental value/price paid in percentage terms). Features based on property performance can easily be constructed using widely used Python libraries.

3.2.2. Nontraditional

Nontraditional data sources in relation to property, for the purposes of this project, can be divided into two categories: Dispersion of POIs, Quality of those POIs. As such, we need the find POI datasets.

The Ordinance Survey POI data is comprehensive. For academic research, access to OS data is licenced through EDINA's Digimap service. Through our City, University of London account, access to the POI OS data has been obtained. OS POI classification scheme has a 3-tier hierarchy: 9 Groups, 52 categories, 600 classes, on levels one, two, and three respectively. [88] (see Fig. 1.). Users can choose POIs from Group and Category levels.

Each POI has several variable including:
"unique_reference_number"|"name"|"pointx_classification_code"|"distance"|"address_detail"|"street_name"|"locality"|"geographic_county"|"postcode"|"verified_address"|"administrative_boundary"|"telephone_number"|"url"|"brand"|"qualifier_type"|"qualifier_data"|"provenance"|"date_of_supply"
[89]The quality of the data can be examined using some of these variables such as the `provenance` and `date_of_supply`. OS team updates the POIs every 6 months.
An alternative data source is the community maintained open source OpenStreetMap (OSM). The list of 'Map Features' [90] reveals a comprehensive collection of POI data which we may use.

PropertyData is a UK based company that has compiled many useful property performance metrics that are accessible through an API, but for a fee. Most of their API methods require a supply of UK postcode. One parameter of interest is the planning permission data source that scrapes planning applications across the UK. This could help with developing features.

Google Maps Platform has a Places API that includes POIs. There are 98 'types' of POI ranging from libraries, locksmiths, schools, to universities and zoos. Although the categorisation is different, the 'types' are similar to those of the OS POI dataset. The advantage of using the Places API is that majority of the establishments that fall under the 'types' are seen as businesses to Google, and come with reviews and star ratings. Scraping this data, or using it through Places API, will help with our quest to establish the quality of our POI and its potential impact on the price predictions.

3.3. Tools:

We will use Python as our coding language. We will use various libraries to implement the models such as pandas, numpy, statsmodels, scikit-learn, matplotlib, seaborn. We may use geopandas, Folium and Altair packages for geospatial analysis and building interactive maps. We will use Multiple Geographically Weighted Regression (MGWR) module with the functionality to calibrate multiscale (M)GWR as well as traditional GWR to perform geographically weighted statistical analyses.

For exploratory data analyses during the qualitative and the gather-parse-clean stages in our implementation plan, we will use visualisations libraries matplotlib and seaborn to inform our analytical decisions. We have obtained API keys for the Google Developer Platform, Zoopla, and PropertyData. We will use request, BeautifulSoup, googlemaps, to connect to APIs and scrape data from relevant websites.

Once we establish which model performs best and which features are useful to be communicated according to our qualitative research findings, we plan to build a simple web application to demonstrate the workings of the models. For the back-end we will use a Flask Framework, and for the front-end a simple CSS or HTML template. The app will then be deployed on Heroku.

3.4. Implementation:

Find what is important from the survey/interviews. To understand the analytical domain, we must be led by people in the industry. Our interview surveys will help us identify what property professionals value in an appraisal system.

Gather + Parse - Find and evaluate the data: trad + non-trad. Through the Digimap interface a specific area can be drawn on a map and the OS POIs file downloaded. The process for using OpenStreetMap requires more coding as it uses its own formats than need unpacking. All requests to various APIs such as PropertyData and Google will be made using the GET method of the HTTP protocol. We will document how the data was collected/generated and qualified.

Clean - Explore and clean the data. We will primarily build regression models. We will use visualisations such as confusion and correlation matrices to see if variables contain autocorrelation and/or multicollinearity, or that correlation between variables is spurious. We also need to get the data ready for analysis by finding anomalies and imputing missing values in our data sources.

Build features using the data. Our feature engineering will be informed by domain knowledge, gained through our qualitative research, as well as heuristic measures and methods informed by the literature. We will use exploratory and visual analyses methods to derive new features. To confirm the existence of spatial dependence between different variables derived from nontraditional datasets, we will use machine learning methods such as principal component analysis and cluster analysis to look for patterns, group records based on their dimensions. We may use regression techniques such as Lasso and stepwise regression to aid with feature selection, for example Park et al [91] use t-test and stepwise logistic regression for this purpose.

We will then need to establish how the relationship between our nontrditional features and house prices varies spatially. To do this we will use geographically weighted statistics with different kernel bandwidths and types. This will produce a list of features based on nontraditional data and POIs, which will be used as model inputs.

We will also determine the quality of the POI using various methods and develop customised scales to use as model inputs. For example, Natural Language Processing (NLP) techniques ad sentiment analysis might be appropriate for putting the reviews for a restaurant on a scale, before using them as a feature. As another example, to categorise different planning permissions that might influence the property prices (informed by qualitative research), we might use NLP to build a picture of development velocity in each location. These qualifier scales and ratings will be used alongside their relevant POI as model inputs.

Run a series of state-of-the-art models with a combination of features. Once we have developed and selected our features we will build models. We will have a predefined radius for a given property address and include the features that fall within that radius as inputs for the models. To choose which models to use, we will rely on state-of-the-art. Baldominos [26] use support vector regression, *k*-nearest neighbours, ensemble of regression trees, and multi-layer perceptron. Zhao et al [35] use deep learning techniques combined with eXtreme Gradient Boosting (XGboost) for property appraisals and see an improvement in house price predictions. They use this on unstructured data, i.e. images from property listings. We will experiment with all these models based on the literature review.

Evaluate, communicate, and document findings (more details in Analysis and Evaluation section below). We will use appropriate metrics and visualisation, where applicable, to communicate the predictive performance. We will also use maps to communicate and evaluate geospatial findings. Documenting our findings throughout the process is imperative, as we are working towards writing a project reporting where all useful interim results and findings will be included.

Iteration is key to the data science process. We will be going back from latter stages to former stages regularly, with the aim of tailoring the process to answer our Research Questions.

4. Analysis and Evaluation

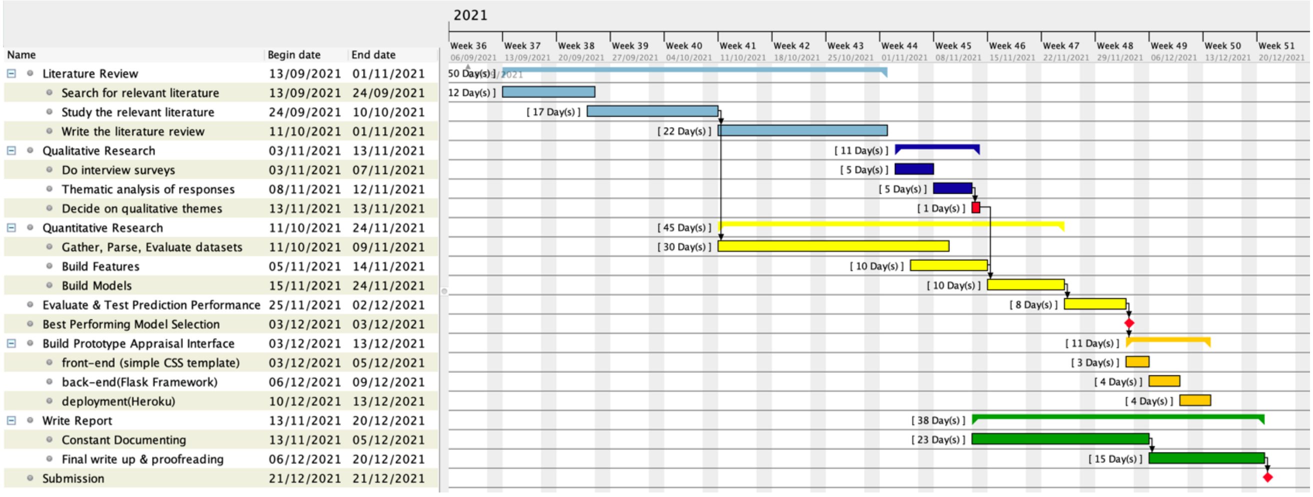
Our first and broad research question is to find whether the use of non-traditional data improves housing value and rent prediction models in a property investment appraisal context. To appraise a property investment, the first step is to value that property. Hence the prediction performance becomes the preliminary target of evaluation. To measure the model performance two type of errors will be used: relative error to see the model performance and mean absolute error for hyper-parameter selection.

The second step is to find and quantify the effect of the top contributing features built using non-traditional data that explain the changes in market and rental values of a property. It is also important in the domain of investment appraisal to note the direction of effect of various features as these may lead to completely different investment decisions. To evaluate these effects, we will compare the model performance given different features. Steurer et al [92] provide a short list of 7 metrics that are well suited to evaluate AVMs, which we will explore.

Throughout the project we will be guided by statistical model outputs and visualisations. We will document every finding at every stage and reflect on whether it is contributing to our two research questions and insights we gained through our interviews. Of particular interest is relating the qualitative and quantitative stages to answer RQ2: ‘Why’ does our prediction rely on certain features more than others?

5. Work Plan

The gantt chart in Figure 2. shows the workplan for this project. There are overlaps between the Qualitative Research and Quantitative Research phases as most of the preparatory coding (searching, gathering and parsing data) can be done before the interviews.



Name	Begin date	End date
⊟ Literature Review	13/09/2021	01/11/2021
Search for relevant literature	13/09/2021	24/09/2021
Study the relevant literature	24/09/2021	10/10/2021
Write the literature review	11/10/2021	01/11/2021
⊟ Qualitative Research	03/11/2021	13/11/2021
Do interview surveys	03/11/2021	07/11/2021
Thematic analysis of responses	08/11/2021	12/11/2021
Decide on qualitative themes	13/11/2021	13/11/2021
⊟ Quantitative Research	11/10/2021	24/11/2021
Gather, Parse, Evaluate datasets	11/10/2021	09/11/2021
Build Features	05/11/2021	14/11/2021
Build Models	15/11/2021	24/11/2021
Evaluate & Test Prediction Performance	25/11/2021	02/12/2021
Best Performing Model Selection	03/12/2021	03/12/2021
⊟ Build Prototype Appraisal Interface	03/12/2021	13/12/2021
front-end (simple CSS template)	03/12/2021	05/12/2021
back-end(Flask Framework)	06/12/2021	09/12/2021
deployment(Heroku)	10/12/2021	13/12/2021
⊟ Write Report	13/11/2021	20/12/2021
Constant Documenting	13/11/2021	05/12/2021
Final write up & proofreading	06/12/2021	20/12/2021
Submission	21/12/2021	21/12/2021

Figure 2. Work plan

6. Risks

Description	Likelihood 1- 3	Consequence 1 - 5	Impact L x C	Mitigation
No access to individuals to interview	1	5	5	Develop own set of criteria to rate 4 or 5 appraisal tools. Use metrics to guide qualitive research
No tangible insight from interviews	2	4	8	Develop own set of criteria to rate 4 or 5 appraisal tools. Use metrics to guide qualitive research
Not finding the correct data sources to build features	2	4	8	Use PropertyData features (readily available API)
Project overrun the specific timeline	1	5	5	The work plan is devised with a 20% extra leeway in the Quantitative Research phase
Coding difficulties	2	4	8	Seek help from members of staff at the University and do online courses

Not understating the model's behaviours and output in full	2	5	10	Go back to literature review and try different models. Also speak to Supervisor
No improvement in predictive performance after using nontraditional data	2	1	2	Report the findings
Laptop lost or stolen	2	1	2	All files are hosted on my personal Google Drive account that is regularly accessed form other computers

7. Ethics

It is essential that all ethical and legal concerns are addressed. At this stage we are planning to conduct interviews with people in the property industry. As such we will need ethics approval.

8. References

[1] G. Hammond, "Financial Times," 31 August 2021. [Online]. Available: https://www.ft.com/content/662d3839-2c58-4b88-8df8-71d8f0af85d5. [Accessed 02 September 2021].

[2] Hamptons, "Number of landlords falls to 7 year low," Hamptons International, February 2020. [Online]. Available: https://www.hamptons.co.uk/research/articles/2020/lettings-index-january-2020.pdf/. [Accessed 21 August 2021].

[3] A. E. Baum, N. Crosby and S. Devaney, Property Investment Appraisal, Wiley-Blackwell, 2021.

[4] G. M. Asaftei, S. Doshi, J. Means and A. Sanghvi, "Getting ahead of the market: How big data is transforming real estate," 2018. [Online]. [Accessed 21 July 2021].

[5] S. Rosen, "Hedonic prices and implicit markets: product differentiation in pure competition," *Journal of Political Economy,* vol. 82, no. 1, pp. 34-55, 1974.

[6] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez and C. Afonso, "Identifying Real Estate Opportunities Using Machine Learning," *Applied Sciences,* vol. 8, no. 2321, 2018.

[7] R. K. Pace and O. Gilley, "Generalizing the OLS and grid estimators," *Real Estate Economy,* vol. 23, no. 2, pp. 331-347, 1998.

[8] S. C. Bourassa, M. Hoesli and V. S. Pengd, "Do housing submarkets really matter?," *Journal of Housing Economics,* no. 12, pp. 12-28, 2003.

[9] S. S. S. Das, M. E. Ali, Y.-F. Li, Y.-B. Kang and T. Sellis, "Boosting house price predictions using geo-spatial network embedding," *Data Mining and Knowledge Discovery,* 2021.

[10] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applicaations,* vol. 42, pp. 2928-2934, 2015.

[11] T. S. Andrey Pavlov, "Immigration, Capital Flows and Housing Prices," *Real Estate Economics,* vol. 48, no. 3, pp. 915-949, 2020.

[12] M. H. Rafiei and H. Adeli, "A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units," *Journal of Construction Engineering and Management ,* vol. 142, no. 2, 2016.

[13] T. J. Brooks, B. R. Humphreys and A. Nowak, " Strip Clubs, "Secondary Effects" and Residential Property Prices," *Real Estate Economics,* vol. 48, no. 3, pp. 850-885, 2020.

[14] B. J. Oates, Researching Information Systems and Computing, Sage Publications Ltd., 2006.

[15] B. Chi and A. Dennett, "Shedding new light on residential property price variation in England: A multi-scale exploration," *Urban Analytics and City Science,* vol. 28, no. 7, pp. 1895-1911, 2020.

[16] Ordinance Survey, "Points of Interst Classification Scheme Official," December 2020. [Online]. Available: https://www.ordnancesurvey.co.uk/documents/product-support/support/points-of-interest-classification-scheme.pdf. [Accessed August 2021].

[17] OpenStreetMap, "OpenStreeMap - Map Features," OpenStreetMap, [Online]. Available: https://wiki.openstreetmap.org/wiki/Map_features. [Accessed October 2021].

[18] J. K. B. Byeonghwa Park, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Systems with Applications,* vol. 42, p. 2928–2934 , 2015.

[19] Y. Zhao, G. Chetty and D. Tran, "Deep Learning with XGBoost for Real Estate Appraisal," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, Xiamen, 2019.

[20] M. Steurer, R. J. Hill and N. Pfeifer, "Metrics for evaluating the performance of machine learning based automated valuation models," *Journal of Property Research,* vol. 38, no. 2, pp. 99-129, 2021.

[21] C. W. Dawson, Projects in Computing and Information Systems: A Student's Guide (2nd ed.), Addison Wesley: London, 2009.

[22] M. McCord, P. Davis, M. Haran and D. McIlhatton, "Understanding rental prices in the UK: a comparative application of spatial modelling approaches," *International Journal of Housing Markets and Analysis,* vol. 7, no. 1, pp. 98-128, 2014.

Ethics From

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

PART A: Ethics Checklist. All students must complete this part.
The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)?	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act?	**NO**

1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation?	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - **https://ethics.city.ac.uk/**		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent?	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects?	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study?	**NO**
2.6	Does your research involve invasive or intrusive procedures?	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - **https://ethics.city.ac.uk/** **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)?	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London?	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK.		*Delete as appropriate*

If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form. **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		
4	Does your project involve human participants or their identifiable personal data?	**YES**

B.1 The following questions must be answered fully. **All grey instructions must be removed.**		*Delete as appropriate*
1.1.	Will you ensure that participants taking part in your project are fully informed about the purpose of the research?	**YES**
1.2	Will you ensure that participants taking part in your project are fully informed about the procedures affecting them or affecting any information collected about them, including information about how the data will be used, to whom it will be disclosed, and how long it will be kept?	**YES**
1.3	When people agree to participate in your project, will it be made clear to them that they may withdraw (i.e. not participate) at any time without any penalty?	**YES**
1.4	Will consent be obtained from the participants in your project? Consent from participants will be necessary if you plan to involve them in your project or if you plan to use identifiable personal data from existing records. "Identifiable personal data" means data relating to a living person who might be identifiable if the record includes their name, username, student id, DNA, fingerprint, address, etc.	**NO**
1.5	Have you made arrangements to ensure that material and/or private information obtained from or about the participating individuals will remain confidential?	**YES**

B.2 If the answer to the following question (B2) is YES, you must provide details		*Delete as appropriate*
2	Will the research be conducted in the participant's home or other non-University location?	**YES**

9 Appendix 2 – Data gathering code

The python notebooks used in the data acquisition pipeline are included in a separate submission as one pdf file called: **Modelling_DataGathering_Deployment_Codes.pdf**. The pdf file combines the.py files mentioned in this report, with each one labelled accordingly:

PlacesSearch.py
placeCafePostcode.py (included as an example, one script similar to this per POI exists, included in submitted files).
areaCodeData.py
CertStatist.py
dataset.py

Individual pdfs were also submitted.

The function of these scripts are explained extensively in the report and shown in Figure 2.

This is the link to the OneDrive where all the files are located https://cityuni-my.sharepoint.com/:f:/r/personal/babak_hessamian_city_ac_uk/Documents/Babak Hessamian - Individual Project - INM363?csf=1&web=1&e=fr04dZ

10 Appendix 3 – Modelling code

The python notebooks used in the modelling pipeline are included in a separate submission as one pdf file called: **Modelling_DataGathering_Deployment_Codes.pdf**. The pdf file combines the .ipynb files mentioned in this report, with each one labelled accordingly:

NewSaleModelPlacesV1.ipynb
NewSFAModelPlaceV1.ipynb
NewRentalModelPlacesV1.ipynb
SQLMATTselectionPCA.ipynb
LassoRidgeRandomForest.ipynb

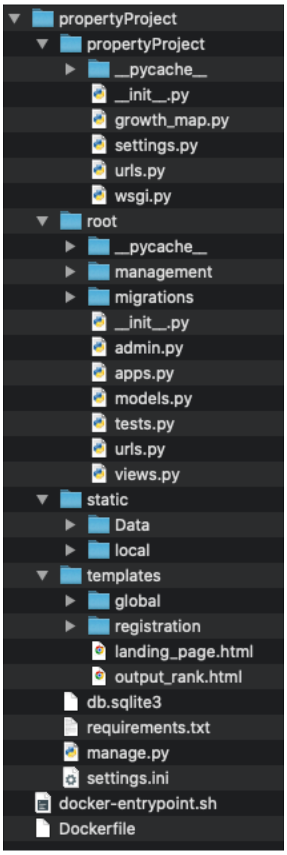
Individual pdfs were also submitted.

The scripts also include the feature engineering pipeline as explained in the report.

This is the link to the OneDrive where all the files are located https://cityuni-my.sharepoint.com/:f:/r/personal/babak_hessamian_city_ac_uk/Documents/Babak Hessamian - Individual Project - INM363?csf=1&web=1&e=fr04dZ

11 Appendix 4 – Deployment code

The structure for the docker folder is shown in the screenshot below.



```
▼ propertyProject
    ▼ propertyProject
        ▶ __pycache__
          __init__.py
          growth_map.py
          settings.py
          urls.py
          wsgi.py
    ▼ root
        ▶ __pycache__
        ▶ management
        ▶ migrations
          __init__.py
          admin.py
          apps.py
          models.py
          tests.py
          urls.py
          views.py
    ▼ static
        ▶ Data
        ▶ local
    ▼ templates
        ▶ global
        ▶ registration
          landing_page.html
          output_rank.html
      db.sqlite3
      requirements.txt
      manage.py
      settings.ini
  docker-entrypoint.sh
  Dockerfile
```

These are included in a separate submission as one pdf file called:

Modelling_DataGathering_Deployment_Codes.pdf. The pdf file combines the.html and .py files mentioned in this report, with each one labelled accordingly

urls.py
views.py
output_rank.html
landing_page.html

Individual pdfs were also submitted.

This is the link to the OneDrive where all the files are located https://cityuni-my.sharepoint.com/:f:/r/personal/babak_hessamian_city_ac_uk/Documents/Babak Hessamian - Individual Project - INM363?csf=1&web=1&e=fr04dZ