Building features driving city-scale residential energy consumption: A multimodal approach

Yulan Sheng, Hadi Arbabi, Wil OC Ward, Mauricio Álvarez and Martin Mayfield

Abstract

The important role of buildings in tackling climate change has been globally recognised. To avoid unnecessary costs and time wasted, it is important to understand the conditions and energy usage for existing housing stock to identify the most important features affecting housing energy consumption and to guide the relevant retrofit measures. Existing data-driven and statistical studies that use machine learning for energy consumption usually develop models using all available variables relevant to building, which can be redundant. This paper investigated how the spatial, morphological and thermal characteristics of residential houses contribute to energy consumption predictions by utilising a state-of-the-art automated machine learning (autoML) tool for properties' construction age bands and energy consumption prediction. A case study has been conducted with around 143,000 residential properties in Sheffield. The autoML model successfully estimated the energy consumption with a mean absolute percentage error of 18.1% and a R^2 score of 0.828. Variables used were ranked by their permutation feature importance. Housing sizes and conditions of the external walls are found to be the most important features when estimating energy consumption of residential buildings in Sheffield. Relatively larger houses developed in neighbourhood with higher density may benefited the most from home upgrading projects for more significant energy consumption reduction.

Keywords Residential Energy Consumption Prediction; Automated Machine Learning (autoML); Energy performance certificates (EPC); Permutation Feature Importance.

1 Introduction

1.1 Background

Residential buildings have become one of the largest consumers of energy around the world (BEIS, 2022b). The recent years have witnessed the growing pressure residents feel in paying energy bills, caused in part by the worldwide COVID-19 pandemic and the rapid increase in energy prices (BEIS, 2022a). In the UK, the residential sector is the only sector that rose in energy consumption since 2019, while other sectors: transport, industry and services, all decreased (BEIS, 2020). This increasing trend hints at the difficulties the UK government is currently facing to achieve its net-zero emissions goals by 2050 to tackle the climate crises.

Incentives have been introduced to mitigate the energy and environmental crisis. The UK government has proposed to raise the minimum energy standards for domestic buildings, especially privately rented houses, from energy rating E to C by 2030 (BEIS, 2019). According to the latest English Housing Survey, 53.8% of existing housing stocks are rated below energy rating C and therefore require retrofitting under the new proposals (DLUHC, 2021). In order to meet the new standard, UK government is investing nearly £4 billion during 2022 to 2026 to support home upgrading and retrofitting (BEIS, 2022a).

Retrofitting homes is relatively expensive and time-consuming compared to demolition and then constructing new buildings. BEIS studied the potential costs for home retrofitting projects and summarised that the most common retrofitting measure used is upgrading the fabric insulation, including the walls, lofts and floors, which can cost up to £15,000 per home (BEIS, 2017). Existing studies have implemented machine learning techniques to develop data-driven models to estimate the buildings' energy performance and identify the elements that are most in need for retrofitting. However, most of these studies chose the input variables and algorithms based on researchers' knowledge or the ones previous studies have used. This paper investigated how important each building feature is related to its energy prediction, by utilising automated machine learning (AutoML) to estimate the year of construction and energy consumption of residential buildings. Publicly available data was used to extract multi-modality features representing buildings' spatial, morphological and thermal characteristics. The marginal effects of features with relatively high permutation feature importance in the designed models were further examined using a series of partial dependence plots. The results provide a hint on what are the most essential features for energy consumption estimation when data is limited, and what are the essential housing characteristics should be considered for selecting target homes for retrofitting.

1.2 Related Work

When estimating residential buildings' energy performance, there are three approaches commonly found in the existing literature, either a data-driven approach, a physics-based approach or a hybrid method that combines the previous two approaches. Both the physics-based and hybrid approaches rely on detailed information on buildings' thermal characteristics, such as the thermal transmittance of the building material (Foucquier et al., 2013). They are usually applied in relatively small-scale studies focusing on a single building. When access to meter readings and buildings' internal space is limited, a data-driven approach is usually applied to develop statistical or machine learning models, based on historical energy consumption data and building morphology. It has been found that, in general (Rosser et al., 2019;

Kontokosta and Tull, 2017):

1. Buildings constructed in similar periods tend to have similar building characteristics; and

2. Buildings with similar characteristics tend to have similar energy needs.

Each rule has suggested one main feature affecting the buildings' energy performance. The first rule indicates the year of construction is important in energy estimation. One of the potential reasons is that, housing legislation changes regularly to comply with the housing and environmental concerns at that time and also what might be needed in the future, for instance, the Town and Country Planning Act issued in 1947 (Gallent and Tewdwr-Jones, 2007) prioritised developing single apartment blocks. The construction sector then develops homes accordingly, hence the second rule (Gallent and Tewdwr-Jones, 2007).

Despite the importance of building age in inferring building energy needs, no easily accessible complete database is available (Rosser et al., 2019). Existing studies have attempted to infer building age from its physical features (Sousa et al., 2017; Kontokosta and Tull, 2017). Rosser et al. (2019) proposed a methodology to predict the year of construction using map data and historical satellite images. Their machine learning model used the random forest algorithm achieved 77% prediction accuracy (Rosser et al., 2019). However, their model was trained based on a relatively small number of properties (1,096) in Nottingham to predict 5 aggregated age bands covering a rather wide time span. The testing samples they used were derived from a single neighbourhood, which tends to have similar building features and construction age.

The second rule, the relationship between building characteristics and energy needs, provides insight into how housing features can be used to estimate energy using the data-driven approach. Existing literature has experimented with a wide range of different data inputs providing such information, including data either in 2D or 3D, e.g. LiDAR point cloud (Dino et al., 2020), text-based (Wang et al., 2018) or image-based (Despotovic et al., 2019; Ali et al., 2019). One widely used database is the Energy Performance Certificates (EPC). EPC is an official document of buildings' energy performance required for every property in the UK. It ranks the building energy performance from G, the least efficient, to A, the most efficient calculated using the Standard Assessment Procedure (SAP) (DECC and BRE, 2014). Ali et al. (2019) developed a workflow that uses existing EPC data to predict buildings' energy ratings when such information is not available. Their best-performing machine learning model has achieved 88% accuracy in predicting building EPC ratings for properties in Ireland. However, there are issues with EPCs that the above studies did not take into consideration. For instance, Crawley et al. (2019) have summarised that there are around 1.6 million properties found to be associated with multiple valid EPCs in the system.

Existing energy prediction studies, including the aforementioned, usually develop the machine learning model without performing an exhaustive search and fine-tuning. One of the potential reasons is that doing an exhaustive search and fine-tuning with dataset at city-scale may require heavy computing power. This is one of the main reasons why the trend of implementing autoML tools is growing. The autoML approach can be considered as a complete "black box". It offers a combined algorithm selection and hyper-parameter optimisation tool to reduce the costs of machine learning model development (Feurer et al., 2015). It takes care of raw data input from the beginning to the final step, offers a tool that reduces development costs, and at the same time provides optimal estimation accuracy (He et al., 2021; Hutter et al., 2019).

1.3 Main Contributions of the Work

This paper investigated the ranking of housing features on building age and energy consumption prediction, based on a systematic approach utilising open-sourced data and autoML, this work

- Identified the most important features for building age and energy consumption estimation;
- Investigated the marginal effects of most important features on building age and energy consumption.

The paper is structured as follows. Section 2 provides a detailed description of what data has been utilised and what pre-process has taken place in this work. Due to the nature of open-source data, the limitations of the used data are listed, followed by how these limitations may hinder the overall model performance. Section 3 presents the methodology this study followed, detailing how the data is aggregated and sub-sampled, how autoML system implemented and robustness tested using a comparative study. A case study was conducted based on residential properties in Sheffield with results and discussion offered in Section 4.

2 Data

This paper mainly used text-based data from two sources: Ordnance Survey (OS) and EPC. The map data is used to describe the spatial and morphological characteristics of the houses, while the EPC provides information relating to housings' material and insulation conditions. The following sections will explain the procedures of the data collection and pre-processing conducted before model development.

2.1 Spatial and Morphological Data

The spatial and morphological data this paper used is the OS MasterMap Building Height Attribute products (Ordnance Survey, 2021). Table 1 has listed all the features extracted and used to describe the buildings' morphology.

Variables 1, 3 and 4 are values provided in the OS MasterMap, while the rest are calculated using ArcGIS.

Table 1: List of features based on OS MasterMap, with brief descriptions of what they represent of and how they are calculated

No.	Variables	Description
1	Total floor area	Area of the building footprint (a)
2	Perimeter	Total length of building polygon outline (p)
3	Relh2	Relative height from ground to the base of the roof
4	Relhmax	Relative height from ground to the highest part of the building
5	NPI	Normalised Perimeter Index (NPI) calculated by $\frac{2\sqrt{a\pi}}{p}$
6	Vxcount	Number of vertices in building polygon
7	Builtrate	Ratio between all property footprint and postcode area

Perimeters and Vxcount are calculated using the field calculator in Arcmap. Variables 5 and 6 are metrics adapted to describe the complexity of the building shape. Normalised Perimeter Index (NPI) is a shape metric measuring the roundness. A NPI value further departed from 1 suggests the building has a more complex shape (A. Wirth, 2004). Three properties are highlighted in Figure 1 as example. Property A is a primary school in Sheffield, while B and C are terraced houses that can be commonly found in the UK. Each property has been marked with its area, total perimeter length and the calculated NPI. By comparing these values, it can be seen that, buildings with more irregular shapes have smaller NPI values. On the other hand, B and C are the same type of houses, so similar values are found for NPI and building perimeter because they are more similar in building shapes.

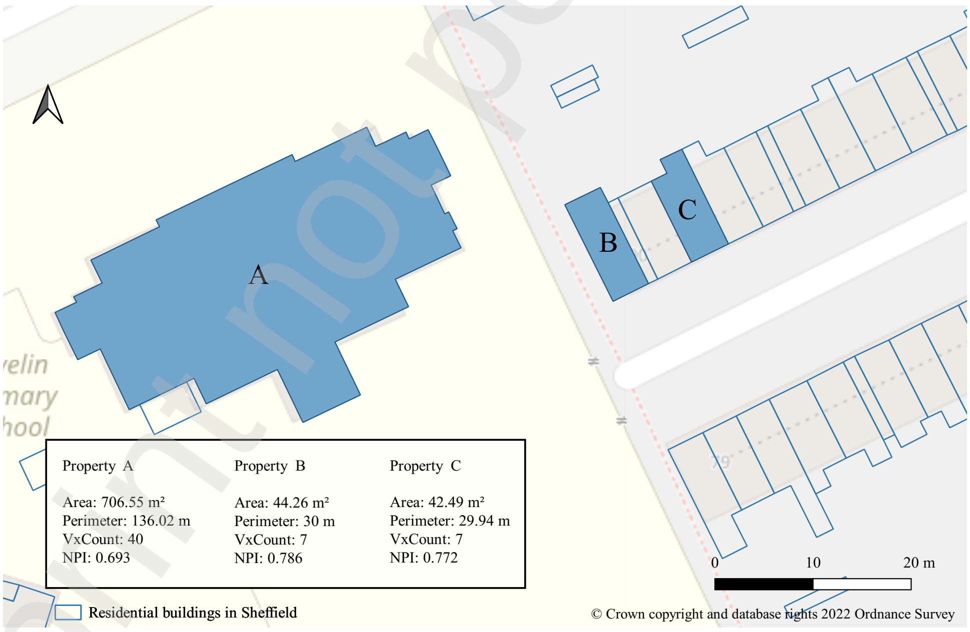


Figure 1: Illustration of example map data

2.2 Energy Performance Certificates

The UK government provides an online database for users to access and download EPC records as spreadsheets. In this study, the EPC is used to provide variables relating to buildings' energy performance. As discussed in Section 1.2, studies show that multiple EPC records can be found associated with the same property (Crawley et al., 2019). This study examined the downloaded EPC, if the property address or reference number occurred multiple times, it means that the property is associated with multiple EPC records. These redundant EPCs are filtered based on when the record was created. The single latest-issued EPC is used as the data input.

Overall, the EPC contains 92 categories offering building-related information from three perspectives: spatial and reference information to identify where the property is (e.g. Unique Property Reference Number (UPRN) and address); the current property characteristics and energy performance; and potential characteristics and energy performance if recommended retrofit implemented. Therefore, a data selection process is essential to filter unnecessary information and avoid high costs in time and computational power. The selected variables and their brief descriptions are listed in Table 2.

Table 2: List of data extracted from the EPC, with brief description of what the represent of and example classes in categorical data

No.	Variables	Description
8	Property type	Type of property (e.g. house)
9	Built form	Type of built-form (e.g. detached)
10	Transaction type	Status in the housing market (e.g. marketed sale)
11	Number habitable rooms	Number of rooms in the property
12	Number heated rooms	Number of rooms that are heated in the property
13	Roof description	Type of roof and its insulation conditions (e.g. pitched)
14	Walls description	Type of walls and its insulation conditions (e.g. filled cavity)
15	Floor description	Type of floor and insulation conditions (e.g. solid, insulated)
16	Lighting description	Percentage of low energy lighting used
17	Mainheat description	Type of main heating options used (e.g. boiler)
18	Main fuel	Type of main fuel used for central heating (e.g. mains gas)
19	Ageband	Construction age grouped in 12 bands (e.g. before 1900)
20	Energy consumption	Energy consumption (kWh per year)

Variables 8 to 12 are features describing the general characteristics of the buildings, while variables 13 to 18 provide more detailed descriptions to the conditions of specific building elements. The original energy consumption recorded in the EPCs are measured in kWh/m^2 per year. Total floor area for each house is taken into consideration here to produce the variable 20, which is used as the ground truth data for training the energy prediction model.

Inconsistencies and abnormal entries are found for the categorical variables. This may be caused by the fact that the records were created by multiple inspectors and may have also followed different versions of guidance on creating EPCs. All variables are preprocessed following two steps. The first step is to replace blank or abnormal entries. For example, if the entry is marked as 'INVALID!' or 'NO DATA', these entries are combined as 'unknown'. This process also ensures the records only contains English records.

The second step is reorganising the categorical data (variables 13-19). Similar descriptions in the categories are found and merged. For instance, 'some double glazed' and 'partial double glazed' used to describe the window insulation conditions are combined into one category.

Once the data from OS and EPC are prepared separately, they are matched using the Unique Property Reference Number (UPRN). The UPRN is a reference system commonly found in the UK geospatial data such as OS map data. It was recently introduced to EPC in November 2021 (Roberts and DLUHC, 2021), which enables this paper to match the map data with its relative EPC. The combined dataset is then used for training the machine learning models for age and energy prediction, which will be explained in the methodology section.

3 Methodology

This section presents the development of supervised machine learning models for age and energy prediction. The overall workflow is illustrated in Figure 2. The first model trains an autoML to predict construction age bands for properties with no age specified in the EPC. This step ensured the data for energy consumption prediction is complete. The second model then predicts energy consumption based on properties' and thermal characteristics.

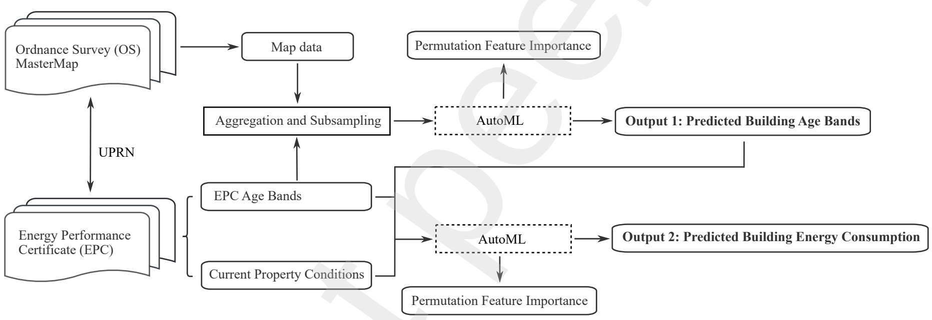


Figure 2: The designed workflow this study follows, including data inputs (OS and EPC), information extraction and pre-processing, model training by autoML and outputs.

3.1 Age Bands Aggregation and Subsampling

The ground truth data used in training the age prediction model is variable 19, the age band recorded in the EPC. The EPC has 12 age bands in total: before 1900; 1900-1929; 1930-1949; 1950-1966; 1967-1975; 1976-1982; 1983-1990; 1991-1995; 1996-2002; 2003-2006; 2007-2011; 2012 on-wards. These age bands are classified following the changes in regulation for building construction, which mainly are amendments for the conservation of fuels and power (DECC and BRE, 2014). The way the age bands are classified suggests it may not be the best representation of how buildings' physical shapes and designs change over time. Relatively lower prediction accuracy is expected when conducting the age detection. However, this is the only open-sourced data that can be found offering adequate spatial coverage and level of detail for property age. There are other age data, such as the products from Verisk (Verisk Analytics Inc, 2022), which interprets building age from imagery, but classified the age in a very generic way (i.e. historic, postwar and modern).

Although the uneven distribution is a representation of the number of properties constructed in the real world, it can poorly affect the performance of machine learning models. Machine learning models usually try to maximise the prediction accuracy by assigning more weights to classes with more occurrences (Appice et al., 2015). To reduce the bias caused by the imbalanced distribution, age bands with fewer records are aggregated into one class, as explained in section 2.2, and then the simple random sampling method is used to randomly select 4,000 properties from each age band for prediction.

3.2 Automated Machine Learning

3.2.1 Auto-Sklearn

After initially processing the raw input data, the workflow then proceed to the next stage to train and perform prediction using autoML. Auto-sklearn was selected as the automated model development tool for this study. Auto-sklearn is developed based on the Scikit-learn, a popular python library offering a wide range of machine learning algorithms (Feurer et al., 2015). As illustrated in Figure 3, Auto-sklearn can be considered as a pipeline with three main steps. The first step is meta-learning, where the input data is compared with pre-stored benchmark data (Feurer et al., 2015). The algorithms that performed well on the benchmark data that is similar to the user inputs are selected as target algorithms. The second stage then trains, fine-tunes and evaluates all target algorithms. The Bayesian optimisation simultaneously calculates the correlations between the hyper-parameter settings and the prediction accuracy. This correlation is the main criteria the Auto-sklearn used for algorithm selection. The pipeline also tests whether building an ensemble of multiple algorithms will achieve better prediction performance.

Two models were separately trained using Auto-sklearn, a classification model for age bands prediction, and a regression model for energy consumption prediction. To minimise the effects of multi-collinearity, the input data were divided into two sets based on the rules stated in Section 1.2. Building age bands were predicted primarily based on the spatial and morphological features of buildings, and the energy consumption was predicted with more thermal-related features. When training, all the input data was randomly split, 80% is used for training and 20% for testing. The trained model performance on the new dataset was examined using the testing data.

The performance of all the trained algorithms were evaluated. Model accuracy score and F1-Macro score were used for the age classification model. The accuracy score calculates the proportion of predicted label that exactly matched with the 'true' labels (Buitinck et al., 2013). F1-Macro score is calculated using the following equations (Geron, 2017), where TP stands for true positives, FP is false positives, and FN is false negatives:

$$
\begin{aligned}
precision &= \frac{\mathrm{TP}}{TP + FP} \\
recall &= \frac{\mathrm{TP}}{TP + FN} \\
F1 - Macro &= \frac{2 \times precision \times recall}{precision + recall}
\end{aligned}
$$

The most optimal algorithm for age band prediction was then used to predict the construction year band and complete the information for houses without age bands recorded. Regression models for energy consumption prediction was evaluated by R^2 and the mean absolute percentage error.

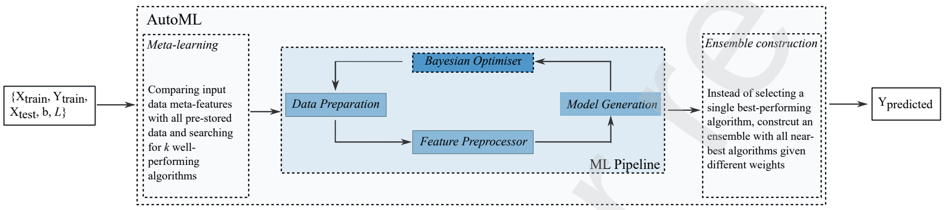


Figure 3: An overview of the Auto-sklearn system. The input data follows the pipeline to construct the most optimal model and then perform prediction. The pipeline involves meta-learning, data preparation, feature preprocessor, model generation, Bayesian optimisation and ensemble construction.

3.2.2 Comparison study between Auto-sklearn and traditional ML pipeline

This work also conducted a comparison study as a robustness test to examine whether Auto-sklearn outperforms a traditional machine learning pipeline, one algorithm selection and fine-tuning are conducted in separate steps. Similar to how Auto-sklearn behaves, the input data was preprocessed. Numeric data, variables 1-7, 11, 12, 16 and 19 (in the energy prediction model), was normalised to be unit invariant. Categorical data, variables 8-9, 13-15, 17 and 18, was processed using the one-hot encoding. This encoding process converts each class in the categorical data into a separate features in a binary format. If the sample falls into this feature, then 1 is marked, otherwise 0.

A list of algorithms that have either been used by existing studies or are potentially suitable for the input data was selected. The four most common machine learning model structures, K-Nearest Neighbours, Random Forest, Decision Tree, and Gradient Boosting, were tested for both age and energy consumption predictions (Geron, 2017; Murphy, 2012). F1-Macro score and R^2 score were also used for evaluating the models and comparing with the models trained using auto-Sklearn.

As shown in Table 3, the traditional pipeline provided a result different from what auto-Sklearn concluded. Among the four algorithms, random forest estimators achieved the best performance for both prediction tasks. It is also the algorithm that most of the existing studies have applied for residential building energy estimation (Rosser et al., 2019; Kontokosta and Tull, 2017). The resulted predictions are also less accurate than the Auto-sklearn computes.

Table 3: Comparison among model training scores for all predictions to check the robustness of using autoML. Different algorithm and better training accuracy were concluded by applying autoML.

		Age bands classification		Energy consumption regression	
	Algorithm	Model Score	F1-Macro	R^2	MAPE
AutoML	**Gradient Boosting**	**0.543**	**0.540**	**0.828**	**18.1%**
Manual	K-neighbours	0.412	0.583	0.758	19.1%
	Decision Tree	0.445	0.901	0.554	22.5%
	Random Forest	**0.468**	**0.991**	**0.776**	**18.7%**
	Gradient Boosting	0.446	0.473	0.767	20.9%

3.3 Permutation Feature Importance

Permutation feature importance (PFI) was used to rank how each variable can affect the overall model performance. The PFI is calculated by randomly shuffling or permutating each input data. The resulting prediction accuracy before and after the shuffling are calculated and compared. Larger difference in accuracy score suggests the variable is relatively more important to the model (Molnar, 2020). Comparing with the gini feature importance used in existing study (Rosser et al., 2019), the PFI performs better in dealing with categorical variables, especially if they are processed with one-hot encoder. For example, after one-hot encoding procedure, the feature class 'Property type', will be expended into four separate variables: property type: bungalow, property type: flat, property type: house, and property type: maisonette. The gini feature importance can only provides individual measures on the four sub-classes; while the PFI is able to store and permute before they are processed with the one-hot encoding system. More useful hints on what input data in their original class are necessary for the predictions can be offered.

4 Case Study: Residential Houses in Sheffield

4.1 Overview

This paper has conducted a case study focusing on all residential buildings in Sheffield, UK. Following the steps explained in the data and methodology sections, EPC records for all residential buildings in Sheffield available as of December 2021 were downloaded. All these records were first filtered so every property only contains the latest record. Among all EPCs downloaded, there were 23.5% properties found to be associated with multiple records which add up to 34.3% EPC records. The resulting dataset comprised 142,973 homes and their associated EPC records for the following study. According to the EPC, the residential properties in Sheffield have an average energy consumption of around 274.50 kWh/m^2 per year or 22219.42 kWh per year, if the footprint for each property recorded in the EPC is used for calculation.

As illustrated in Figure 4, before aggregation, the original records from EPCs show that most of the residential buildings in Sheffield were developed between 1900 and 1966, and few were built after 2012. There are also 10,392 (7.3%) properties' construction age remains unknown. Without pre-processing, this uneven distribution will lead to a biased model. Based on the number of properties each age band contains, the age band '1991-1995' and '1996-2002' were combined into the new class '1991-2002'; '2002-2006', '2007-2011' and '2012 on-wards' were aggregated into the new class 'post-2002'. The aggregation process ensured all age bands have enough data to follow the sampling process for model training.

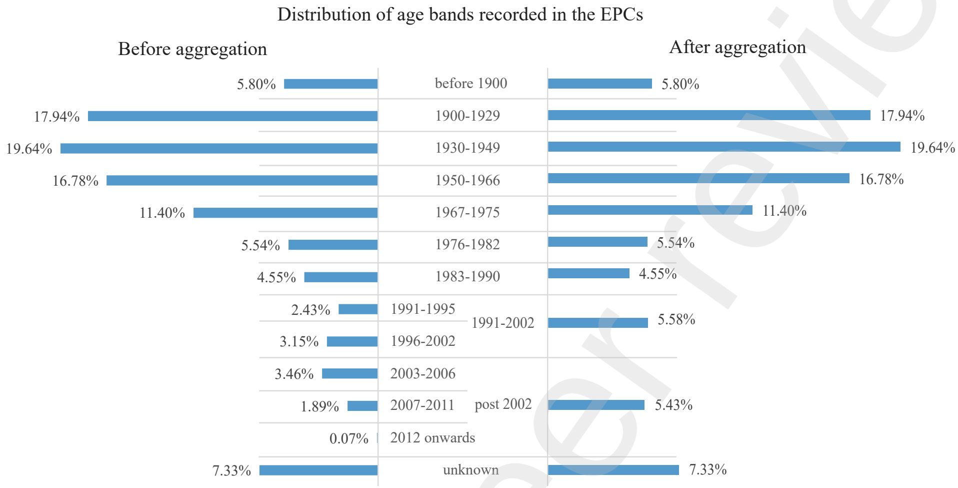


Figure 4: Distribution of construction age recorded in the EPCs before (left) and after aggregation (right)

Table 4 summarises the basic statistics of the numeric data and their subsets used in predictions, including their average, standard deviation (std) and coefficient of variance (cv). The summary of categorical data used in this paper is included in the Appendix. The last four variables in Table 4 are only used for energy prediction so no subsamples were generated. The coefficient of variance is calculated as the ratio between the std and the mean. Among all the numerical data used in this study, it is not surprising to find that, except for built rate, all the variables have cv less than 1. As more than 70% of residential properties in Sheffield are houses, they tend to have relatively similar physical features, the same as the example map illustrated in Figure 1. The only variable that has a cv larger than 1 is the built rate, this is also common because properties in the more rural areas of the city are less densely built than neighbourhoods around the city centre. By comparison, the subsets generated using the sampling method can to some extent be considered representative of all the data collected, as there is no significant difference between the statistics of original and subsampled data.

Table 4: Statistics of numeric data used for model prediction, before and after applying the simple random sampling approach

Variables	All Samples			Subsamples		
	Mean	Std	cv	Mean	Std	cv
Total floor area	81.45	38.16	0.47	81.02	40.11	0.49
Perimeter	41.82	26.01	0.62	45.84	32.83	0.72
Relh2	6.33	3.26	0.52	6.78	3.95	0.58
Relhmax	8.17	3.40	0.42	8.73	4.16	0.48
NPI	0.78	0.04	0.05	0.77	0.05	0.06
Vxcount	12.57	7.29	0.58	9.96	5.00	0.50
Builtrate	0.21	0.28	1.33	0.23	0.36	1.57
Number habitable rooms	4.06	1.77	0.44			
Number heated rooms	3.96	1.76	0.44			
Lighting description	0.53	0.34	0.64			
Energy consumption (kWh)	22219.42	14149.90	0.64			

4.2 Results and Discussion

4.2.1 Age Detection

The age detection model was trained on the processed dataset. The auto-Sklearn detected 37 algorithms that might be optimal for predicting building age bands. The most optimal model used a gradient boosting algorithm, which trains the model by sequentially adding input variables to the ensemble of decision trees and refit the model based on the errors made by the previous added inputs (Murphy, 2012).

For testing data, the most optimal model Auto-Sklearn trained achieved a accuracy score of 0.543 and an F1-Macro score of 0.540. The model performance was further evaluated by comparing the predicted age bands for the test data with their true class in EPC records, Figure 5. Although the majority of the age band were correctly predicted, especially for the aggregated age bands, as expected, a few remain mispredicted. Apart from the reason explained in the data section, that the age bands are classified based on the changes in energy regulations, other potential reason for this misprediction might be because developers tend to design houses that fit into the general building styles nearby.

The PFI plotted in Figure 6a ranked how important each input feature is to the age prediction model. The x-axis is plotted in its log form, to offer clearer visualisation for variables with less feature importance The importance rank suggested that, the built-up rate is the most important features when predicting the age bands of residential buildings in Sheffield, floor area and property types are also relatively importance. Excluding the variable builtrate caused a 23.9% decrease in model accuracy score, and a 25.6% decrease in F1-Macro score.

The NPI and the number of vertices are found relatively less important. As the example properties illustrated in Figure 1, when predicting the age of residential buildings, buildings tend to have little difference in shapes and thereby less sparsity in values can be found. Excluding NPI and the number of vertices only caused decrease in accuracy score and F1-Macro by 0.37% and 0.56% respectively. In

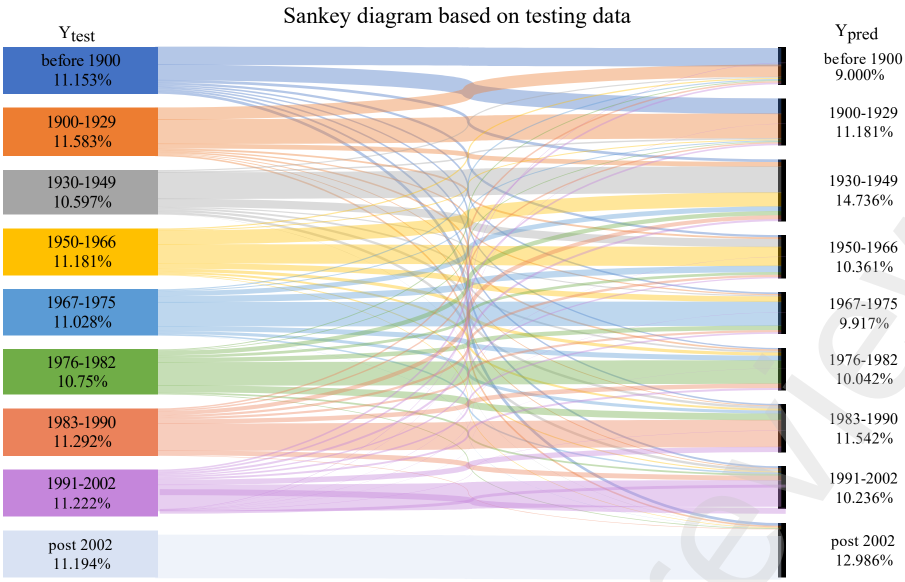
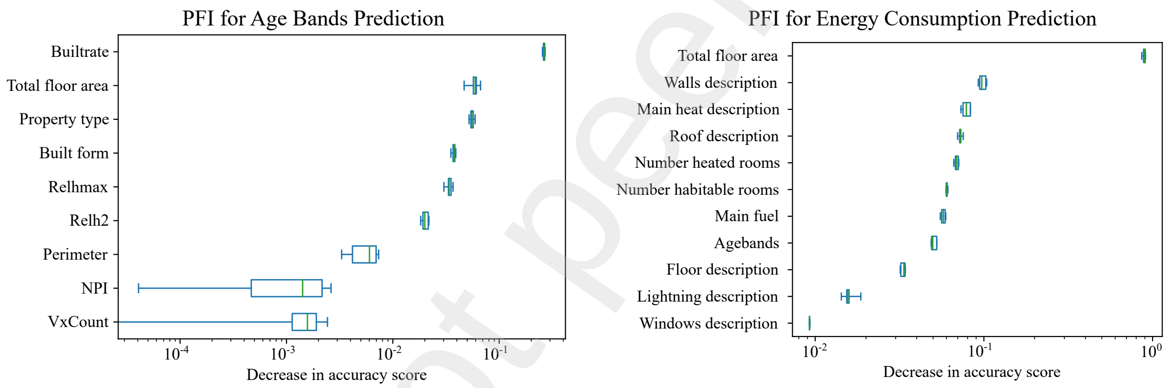


Figure 5: Sankey diagram showing the link between the true (left) and predicted age bands (right) using the random forest classification.



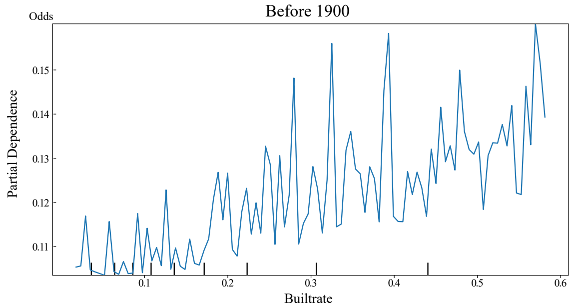
(a) PFI for age detection

(b) PFI for energy detection

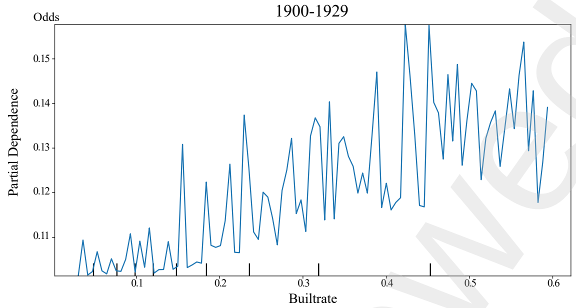
Figure 6: PFI for variables used in the two machine learning models, x axis in log form. (a) is for age detection and (b) is for energy consumption prediction

overall, when data availability is limited, the age band of the housing can be estimated by understanding the housing size, the building type, and how densely the postcode area it located at is developed.

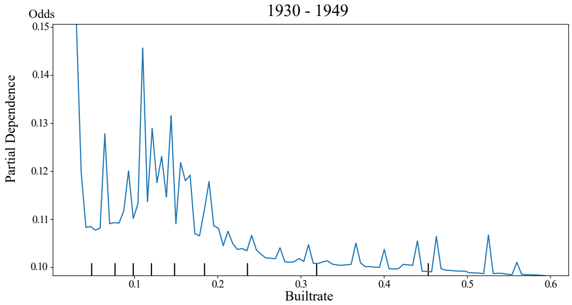
To further investigate how the variable 'Builtrate' contributes to the prediction of each age band, partial dependency plots (PDP) are adopted. The partial dependence calculates the average marginal effects a target feature has towards the prediction outcome, by considering all the other features as constant (Molnar, 2020). As illustrated in the series of charts in Figure 7, the relationships between the builtrate and each age band are complex. In general, in Sheffield, if the houses located in more densely developed postcode area, the houses have higher possibility of being built before 1929 or after 2002. Houses built in areas with less builtrate is more likely being built between 1967 and 1982.



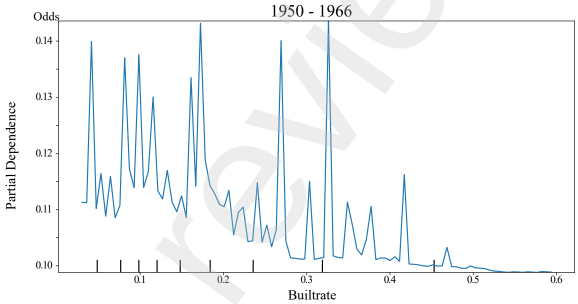
(a) PDP between builtrate and before 1900



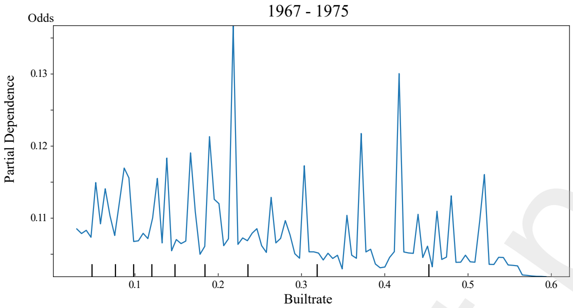
(b) PDP between builtrate and 1900-1929



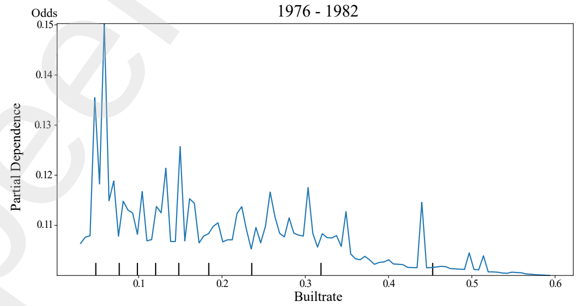
(c) PDP between builtrate and 1930-1949



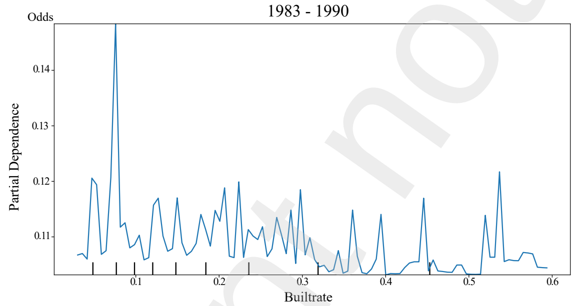
(d) PDP between builtrate and 1950-1966



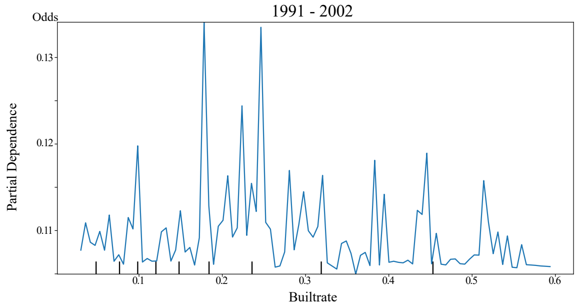
(e) PDP between builtrate and 1967-1975



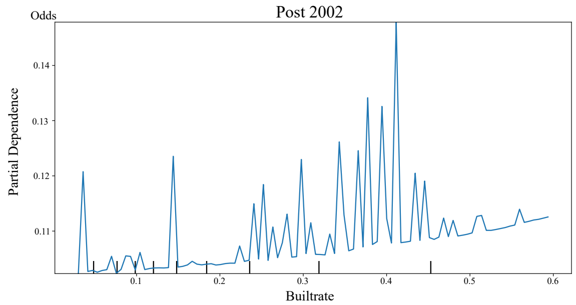
(f) PDP between builtrate and 1976-1982



(g) PDP between builtrate and 1983-1990



(h) PDP between builtrate and 1991-2002



(i) PDP between builtrate and post 2002

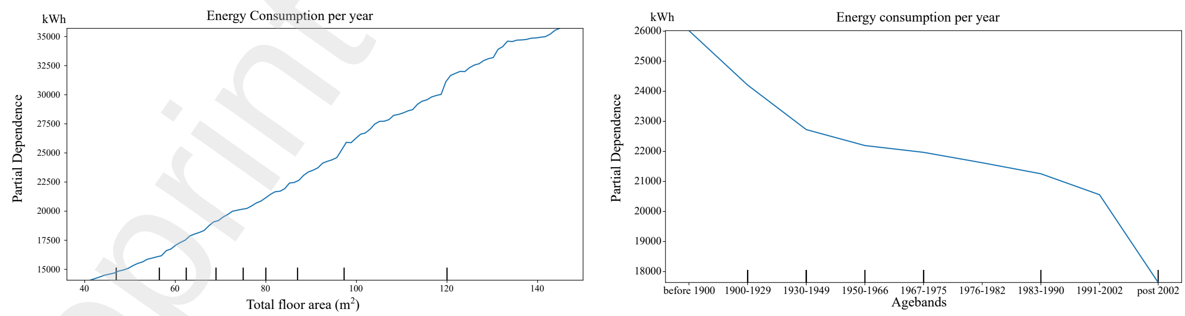
Figure 7: PDP for variable Builtrate and age bands. Each plot shows how 'Builtrate' (the x axis) contributes to the odds of houses developed in each age band (the y axis).

4.2.2 Energy Consumption Prediction

The energy consumption prediction was then conducted after age bands classified for each housing. The age prediction results from the first model were used to train the model. Auto-sklearn determined the best-performing algorithm used data preprocessors based on feature type, feature agglomeration as feature processors and gradient boosting as the regressor. The trained model achieved a R^2 score of 0.828, and a mean absolute percentage error (MAPE) of 18.1%. The results suggest that overall, around 82.8% of the test data can be explained by the trained algorithm; and the prediction results based on the test data have an average difference of 18.1% compared with the ground truth.

The PFI plotted in Figure 6b ranked how the input data may affect the model performance on estimating energy consumption. The total floor area is the dominating feature in this estimation. Excluding this feature from model training led to a 15.3% decrease in R^2 score and a 26.0% increase in MAPE value. The partial dependence plot 8a suggests that, a linear relationship can be found between house sizes and energy consumption. In general, larger houses in Sheffield usually have higher energy consumption. Apart from variables related to the housing size, the the type and condition of the walls is the most important feature when estimating residential housing energy. How different types and conditions of housing material may affect the housing energy needs are intensively researched (Tingley et al., 2015; Government, 2022). The external walls are also where most retrofitting projects target at.

On the other hand, window and lightning conditions are less important in estimating housing energy consumption, excluding these features only resulted in 1.20% decrease in R^2 score and 2.76% incease in MAPE. Houses' age bands ranked the eighth among all features, which indicates that it has relative less impacts for energy consumption prediction. The PDP plots in Figure 8b suggests that in overall, a declining linear relationship can be found between housing age and its energy consumption. Houses newly built tends to have less energy consumed.



(a) PDP for total floor area against energy consumption (b) PDP for age bands against energy consumption

Figure 8: PDP for the marginal effects of total floor area and age bands (the x-axis) towards residential energy consumption in kWh (the y-axis) in Sheffield.

5 Conclusion

This paper examines how spatial, morphological and thermal characteristics of residential houses contributes to housing age and energy consumption prediction, by applying an automated approach in machine learning model development. The trained model achieved a R^2 score of 0.828 in predicting residential building energy prediction. The PFI plots offer hints the essential information required for each model when data availability is limited to perform prediction. That means, when SAP calculation is not available, this approach can be followed to obtain a relatively accurate understanding of the building energy demands using variables with higher rank of feature importance: house size, material and conditions of the external walls, and also the main heat options used. By further examining how individual variable correlates with the amount of building energy consumption, the series of PDP plots suggests that, energy savings may be largely made by targeting at larger houses. For houses in similar sizes, improving the insulation conditions of the building walls will lead to the most significant changes in residential energy efficiency.

However, EPC is not reliable or accurate. Future work can be done to investigate potential alternative data sources to describe the building's thermal and physical conditions. For instance, photos for the target properties and scanned LiDAR 3D models. Multi-modal prediction can also be conducted to overcome the limitations caused by using only one type of data. This paper only utilised text-based data, but for future work, deep multi-modal learning may be developed to jointly take images and text data for prediction.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Yulan Sheng: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Visualisation, Writing - original draft, Writing - review & editing. **Hadi Arbabi:** Conceptualisation, Methodology, Supervision, Writing - review & editing. **Wil OC Wards:** Supervision, Writing - review & editing. **Mauricio Álvarez:** Methodology, Supervision. **Martin Mayfield:** Supervision, Writing - review & editing.

Acknowledgements

This work was supported by the University of Sheffield University Energy Flagship Institute Scholarship. WOCW was supported by EPSRC Active Building Centre [EP/V012053/1] and Towards Turing 2.0 under EPSRC [EP/W037211/1] and The Alan Turing Institute. Neither EPSRC nor The Alan Turing Institute had any involvement in study design; execution; or in the writing of this article.

References

M. A. Wirth. Shape analysis & measurement, 2004. URL http://www.cyto.purdue.edu/cdroms/micro2/content/education/wirth10.pdf.

U. Ali, M. H. Shamsi, F. Alshehri, E. Mangina, and J. O'Donnell. Application of intelligent algorithms for residential building energy performance rating prediction. *Building Simulation Conference Proceedings*, 5(September):3177–3184, 2019. ISSN 25222708. doi: 10.26868/25222708.2019.210232.

A. Appice, P. P. Rodrigues, V. S. Costa, C. Soares, and J. Gama. *Machine Learning and Knowledge Discovery in Databases*, volume 9284 of *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-23527-1. doi: 10.1007/978-3-319-23528-8. URL http://link.springer.com/10.1007/978-3-319-23528-8.

BEIS. What Does It Cost To Retrofit Homes?, 2017. URL https://www.gov.uk/government/collections/buildings-energy-efficiency-technical-research#full-publication-update-history.

BEIS. Setting long-term energy performance standards for the private rented sector in England and Wales, 2019. URL https://www.gov.uk/guidance/domestic-private-rented-property-minimum-energy-efficiency-standard-landlord-guidance.

BEIS. Energy Consumption in the UK (ECUK) 1970 to 2019, 2020. URL https://www.gov.uk/government/collections/digest-of-uk-energy-statistics-dukes.

BEIS. Energy company obligation - eco4: 2022-2026, 4 2022a. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1065823/eco4-government-response.pdf.

BEIS. National statistics: Energy consumption in the uk 2021, 2022b. URL https://www.gov.uk/government/statistics/energy-consumption-in-the-uk-2021.

L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

J. Crawley, P. Biddulph, P. J. Northrop, J. Wingfield, T. Oreszczyn, and C. Elwell. Quantifying the measurement error on england and wales epc ratings. *Energies*, 12(18):3523, 2019.

DECC and BRE. The Government ' s Standard Assessment Procedure for Energy Rating of Dwellings, 2014. URL https://www.bre.co.uk/filelibrary/SAP/2012/SAP-2012_9-92.pdf.

M. Despotovic, D. Koch, S. Leiber, M. Döller, M. Sakeena, and M. Zeppelzauer. Prediction and analysis of heating energy demand for detached houses by computer vision. *Energy and Buildings*, 193:29–35, 2019. ISSN 03787788. doi: 10.1016/j.enbuild.2019.03.036. URL https://doi.org/10.1016/j.enbuild.2019.03.036.

I. G. Dino, A. E. Sari, O. K. Iseri, S. Akin, E. Kalfaoglu, B. Erdogan, S. Kalkan, and A. A. Alatan. Image-based construction of building energy models using computer vision. *Automation in Construction*, 116:103231, aug 2020. ISSN 09265805. doi: 10.1016/j.autcon.2020.103231.

DLUHC. National Statistics: English Housing Survey 2020-21, 2021. ISSN 03787788.

M. Feurer, A. Klein, K. E. Jost, T. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pdf.

A. Foucquier, S. Robert, F. Suard, L. Stéphan, and A. Jay. State of the art in building modelling and energy performances prediction: A review, 2013. ISSN 13640321.

N. Gallent and M. Tewdwr-Jones. *Decent Homes for All: Planning's evolving role in housing provision.* Routledge, Abingdon, 2007. ISBN 978-0-415-27446-3.

A. Geron. *Hands-on machine learning with scikit-learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* Sebastopol, 2017. ISBN 9781491962268.

H. Government. The building regulations 2010: The merged approved documents, 2022. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082748/Merged_Approved_Documents__Jun2022_.pdf.

X. He, K. Zhao, and X. Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212, 1 2021. ISSN 09507051. doi: 10.1016/j.knosys.2020.106622.

F. Hutter, L. Kotthoff, and J. Vanschoren. *Automated Machine Learning Methods, Systems, Challenges.* 2019. ISBN 3-030-05318-0. URL http://www.springer.com/series/15602.

C. E. Kontokosta and C. Tull. A data-driven predictive model of city-scale energy use in buildings. *Applied Energy*, 197:303–317, 2017. ISSN 03062619. doi: 10.1016/j.apenergy.2017.04.005.

C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. URL http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=3339490.

Ordnance Survey. Os mastermap topography layer – building height attribute – overview, 2021. URL https://www.ordnancesurvey.co.uk/documents/product-support/user-guide/osmm-topography-layer-building-height-attribute-overview-v1.4.pdf.

B. Roberts and DLUHC. Energy performance certificates now include the Unique Property Reference Number (UPRN), 2021. URL https://news.opendatacommunities.org/energy-performance-certificates-now-include-uprn/.

J. F. Rosser, D. S. Boyd, G. Long, S. Zakhary, Y. Mao, and D. Robinson. Predicting residential building age from map data. *Computers, Environment and Urban Systems*, 73:56–67, 2019.

G. Sousa, B. M. Jones, P. A. Mirzaei, and D. Robinson. A review and critique of uk housing stock energy models, modelling approaches and data sources. *Energy and Buildings*, 151:66–80, 2017.

D. D. Tingley, A. Hathway, and B. Davison. An environmental impact comparison of external wall insulation types. *Building and Environment*, 85:182–189, 2 2015. ISSN 03601323. doi: 10.1016/j.buildenv.2014.11.021.

Verisk Analytics Inc. UKBuildings Reference Guide, 2022. URL https://www.verisk.com/en-gb/3d-visual-intelligence/products/ukbuildings/#form.

Z. Wang, Y. Wang, R. Zeng, R. S. Srinivasan, and S. Ahrentzen. Random Forest based hourly building energy prediction. *Energy and Buildings*, 171:11–25, 2018. ISSN 03787788. doi: 10.1016/j.enbuild.2018.04.008.

Appendix: Statistics of categorical data used in case study

Table 5: Property type

Property type	Proportion
Bungalow	4.54%
Flat	22.09%
House	70.84%
Maisonette	2.54%

Table 6: Built form

Built form	Proportion
Detached	17.40%
Enclosed End-Terrace	1.17%
Enclosed Mid-Terrace	0.83%
End-Terrace	14.22%
Mid-Terrace	29.21%
Semi-Detached	34.53%
unknown	2.66%

Table 7: Floor description

Floor description	Proportion
(another dwelling below)	16.20%
Conservatory	0.00%
insulated	0.00%
no insulation	0.00%
Solid, insulated	3.37%
Solid, no insulation	18.67%
Suspended, insulated	2.92%
Suspended, uninsulated	47.67%
To external air, insulated	0.11%
To external air, uninsulated	0.11%
To unheated space, insulated	1.15%
To unheated space, uninsulated	4.51%
Average thermal transmittance 0-1.33	5.23%
unknown	0.04%

Table 8: Windows description

Windows description	Proportion
Double glazing	90.39%
High performance glazing	5.47%
Multiple glazing	0.13%
Multiple glazing	0.00%
Secondary glazing	0.41%
Single glazing	3.38%
Triple glazing	0.14%
unknown	0.08%

Table 9: Walls description

Walls description	Proportion
Cavity wall, insulated	52.82%
Cavity wall, no insulation	12.92%
Cob, as built	0.01%
Granite or whin, insulated	0.01%
Granite or whin, no insulation	0.13%
Sandstone or limestone, insulated	0.45%
Sandstone or limestone, no insulation	4.78%
Solid brick, insulated	0.97%
Solid brick, no insulation	16.15%
System built, insulated	1.73%
System built, no insulation	1.04%
Timber frame, insulated	1.41%
Timber frame, no insulation	0.08%
Average thermal transmittance 0-2.1	7.46%
unknown	0.04%

Table 10: Roof description

Roof description	Proportion
(another dwelling above)	14.63%
Flat, insulated	2.06%
Flat, no insulation	1.38%
Pitched, insulated	58.37%
Pitched, no insulation	14.92%
Roof room(s), insulated	1.79%
Roof room(s), no insulation	1.40%
Thatched	0.00%
Thatched, insulated	0.01%
Average thermal transmittance 0-2.4	5.37%
unknown	0.06%

Table 11: Main fuel

Main fuel	Proportion
biogas	0.01%
biomass	0.02%
coal	0.10%
dual fuel	0.04%
electricity	7.19%
from heat network	0.00%
gas	91.43%
LPG	0.10%
no heating	0.29%
oil	0.12%
unknown	0.54%
waste combustion	0.15%
wood	0.01%

Table 12: Main heat

Main heat description	Proportion
Air source heat pump	0.14%
Boiler	86.96%
Community scheme	4.00%
Electric heat pumps	0.00%
Electric heaters	3.02%
Ground source heat pump	0.01%
Micro-cogeneration	0.00%
Room heaters	5.22%
unknown	0.11%
Warm air	0.53%
Water source heat pump	0.00%