Imperial College
London

Estimations for Energy Consumption in Domestic and Public Buildings in London Using Machine Learning

Project Code: **MLBD_14**
Name: **NURUL FATIHAH BT. NOR AZLAN SHAH**
CID: **02544137**

Supervisor: **Dr. Heather Graven**
Assessor: **Professor Ralf Toumi**

Word Count: 8100
Date of Submission: **6th September 2024**

Estimations for Energy Consumption in Domestic and Public Buildings in London Using Machine Learning

The fluctuations of energy prices and demand between 2022 and 2023 are due to the post-impact of a conflict between Ukraine and Russia that disrupted Europe's energy supply. Moreover, a few factors affect the usage of energy consumption in a building, such as spatial factors, government policies and behavioural factors. Based on features available from the Energy Performance Certificate (EPC), Display Energy Certificate (DEC) and London Building Stock Model (LBSM), the pattern of energy consumption of domestic and public buildings in London is influenced by socioeconomic factors and spatial factors [1,2]. Existing studies may not account for the full impact of recently implemented energy efficiency policies, especially on heritage and older buildings, and how they are affected by spatial and socioeconomic factors. More research is needed to measure whether retrofitted buildings improve energy efficiency and constantly reduce CO_2 emissions. To address this gap, six Machine Learning models namely Extreme Gradient Boosting (XGBoost), Gradient Booting (GB), Random Forest (RF), Decision Tree (DT) and Deep Neural Network (DNN) for estimating energy consumption in domestic and public buildings. This study also investigates the model performance based on a dataset of EPC with LBSM versus a test dataset with LBSM only for domestic buildings. The dataset for public buildings is between DEC and LBSM versus the dataset with LBSM only. Further analysis is by studying the correlation of estimated energy consumption with CO_2 emission. GB (0.89) of R^2 emerged as the best model for domestic buildings, while RF (0.73) of R^2 was depicted as the best model for public buildings. Figure 1 represents estimated energy consumption of public and domestic buildings. Domestic building increased gradually from 150 kWh/m^2 and reached its peak in 2019, almost 300 kWh/m^2, and slightly reduced in 2020. meanwhile, estimated energy consumption for public buildings decreased gradually from 150 kWh/m^2 in 2018 to 100 kWh/m^2 in 2020 due to the lockdown from March 2020 until May 2020, consequences of the global coronavirus pandemic (COVID-19) where all public buildings are closed during the pandemic. This reduces the occupancy rate as most of the occupants work from home. In contrast, energy used in domestic building decreased slightly from 2019 to 2020, showing that the usage of energy consumption for domestic buildings was relatively the same during pre-COVID-19 and during lockdown, as the estimated energy usage did not increase significantly nor decrease steadily. In addition, the estimated energy consumption of public buildings increased gradually after the lockdown, while the energy consumption of domestic buildings decreased gradually. This illustrates that some policies implemented to improve the usage of energy efficiency for domestic buildings are successful. Based on Department of Energy and Climate Change (DECC) policies along with Energy Company Obligation (ECO) aimed to improve energy efficiency for domestic buildings and houses, especially for lower-income households, by rolling out smart meters, replacing inefficient boilers with boilers that save more energy and expected to cover 10.5 million homes by 2020. This explains the significant energy reduction from 2020 until 2023[3]. Next, a comparison of model performance using two different datasets. First, GB is training to compare its performance when EPC and LBSM are input and when only LBSM is input. The metric difference is R^2 (0.55), MAE (0.74), MSE (0.88) and RMSE(0.66). In contrast, RF is used to train DEC with LBSM and LBSM only with metric values is R^2 (0.20), MAE (0.16), MSE (0.33) and RMSE(0.18). The metric difference for RF is much lower compared to GB as input features for public buildings have more contribution from LBSM compared to input features of domestic building. Based on Figure 4.5.1 and Figure 4.5.2 shows a correlation between estimated energy consumption for domestic buildings with CO2 emissions per floor area from EPC using regression plot and polynomial fit plot. Referring to Figure 4 from appendices represents that the actual values of energy consumption with CO2 emissions per floor area are highly correlated. However, the estimated energy consumption using spatial and socioeconomic factors as input features shows that the estimated energy values is not correlated with the CO2 emissions per floor area. Polynomial fit is used to determine the non-linear relationship between the estimated energy consumption with the CO2 emissions per floor area. This also shows that this method is a low performance method to compare estimated energy consumption with CO2 emissions using spatial and socioeconomic factors as input features. Other features can be considered to estimate energy consumption to compare with CO2 emissions.

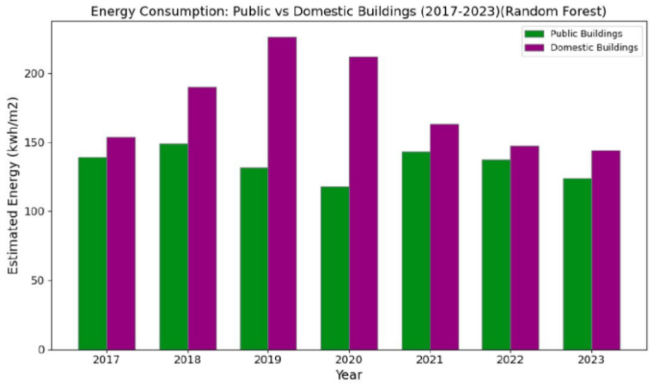


Figure 1: Estimated Energy used for Public and Domestic Buildings (2017 – 2023) (Random Forest)

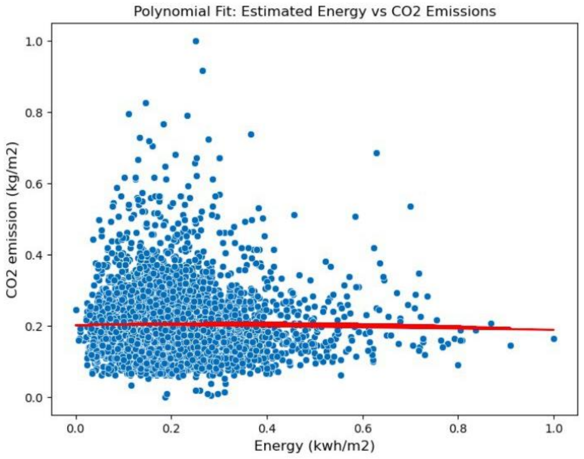


Figure 2: Correlation between estimated energy consumption with CO_2 emission of domestic building using polynomial plot

REFERENCES

1. Energy Performance Certificate Register. (n.d.). Domestic Energy Performance Certificates search. Open Data Communities. Retrieved [30th August 2024], fromhttps://epc.opendatacommunities.org/domestic/search

2. Steadman, P., Evans, S., Liddiard, R., Godoy-Shimizu, D., Ruyssevelt, P., & Humphrey, D. (2020). Building stock energy modelling in the UK: The 3DStock method and the London Building Stock Model. Buildings and Cities, 1(1), 100-119. https://doi.org/10.5334/bc.52

3. Department of Energy and Climate Change. (2013). *Policy impacts on prices and bills*.GOV.UK. https://www.gov.uk/guidance/policy-impacts-on-prices-and-bills

Abstract

The fluctuation of energy prices due to sentiment, market price, and other factors has influenced the pattern of energy consumption, especially in London, which has the highest population density in the largest city in the United Kingdom. Moreover, most of the types of buildings in London were built before 1944. These old buildings have less energy efficiency in terms of poor insulation, single glazed window, older building materials and outdated lighting. The recent method to estimate energy consumption is machine learning. Existing studies may not account for the full impact of recently implemented energy efficiency policies, especially on heritage and older buildings, and how they are affected by spatial and socioeconomic factors. More research is needed to measure whether retrofitted buildings improve energy efficiency and constantly reduce CO_2 emissions. To address this gap, six Machine Learning models namely Extreme Gradient Boosting (XGBoost), Gradient Booting (GB), Random Forest (RF), Decision Tree (DT) and Deep Neural Network (DNN) for estimating energy consumption in domestic and public buildings. This study also investigates the model performance based on an Energy Performance Certificate (EPC) dataset with London and Building Stock model (LBSM) versus a test dataset with LBSM only for domestic buildings. The dataset for the public building is between Display Energy Certificate (DEC) with LBSM versus dataset with LBSM only.

Further analysis involves studying the correlation of estimated energy consumption with CO_2 emission. XGBoost emerged as the best model for domestic buildings, while GB was depicted as the best model for public buildings. Input features from EPC and DEC play an important role in improving model performance based on evaluation of metric values. The correlation between estimated energy consumption and CO_2 emission is determined.

TABLE OF CONTENTS

LIST OF TABLES

Appendices

LIST OF FIGURES

Appendices

.

Chapter 1

Introduction

Department for Energy Security and Net Zero stated in their 2023 UK report that energy consumption of the UK domestic sector has reduced by up to 14% due to increasing temperature. [1]. This shows that the heating demand decreases as temperature increases, leading to lower energy consumption. Another external factor that affects energy consumption is the increment in energy prices [1]. As the energy price increases, the household member will save more energy usage by turning off the heater during summer or spring, especially in households with a lower income and is considered as 'fuel poverty'. This shows that environmental and socioeconomic factors play a role in energy consumption. However, the energy price reduction in 2023 has increased the energy demand to 8.4% [2]. The fluctuations in energy prices and demand between 2022 and 2023 are due to the post-impact of a conflict between Ukraine and Russia that disrupted Europe's energy supply. Moreover, few factors affect the usage of energy consumption in a building, such as spatial factors, government policies and behavioural factors. Based on features available from the Energy Performance Certificate (EPC), Display Energy Certificate (DEC) and London Building Stock Model (LBSM), the pattern of energy consumption of domestic and public buildings in London is influenced by socioeconomic factors and spatial factors [3,4].

The UK government is on its way to improving energy consumption in all sectors, including the domestic sector, to achieve Net zero carbon emissions by 2050. London has a high density of buildings and estimating energy consumption is essential to understanding the city's carbon footprint patterns. This will help policymakers implement more effective strategies to reduce emissions from buildings, such as enhancing the energy efficiency standard or promoting the use of renewable energy sources. Moreover, understanding energy consumption patterns in London can inform the development of smart cities and sustainable communities to ensure new developments contribute to a low-carbon future and modernize old buildings with energy-efficient technologies to reduce greenhouse gaseous

(GHG) emissions. Besides, it will help policymakers study the impact of energy patterns on other communities. It will lead to targeted efforts to reduce energy poverty as London has diverse populations and socioeconomic groups with varying energy needs. London is among the cities with the highest urbanization rate since it is the largest city in the UK. Thus, increased building construction activities and maintenance leads to higher energy consumption needs [5].

Moreover, some studies may need to consider the full impact of recently implemented energy efficiency policies on older buildings and how spatial and socioeconomic factors affect the energy consumption in a building|. More research is needed to determine whether the retrofitted buildings improve energy efficiency and successfully reduce CO_2 emissions in London. To fill this gap, six Machine Learning (ML) models, Extreme Gradient Boosting (XGBoost), Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM) and Deep Neural Network (DNN), were applied. Therefore, by developing best practices in energy management, London can influence global strategies for reducing energy consumption in urban areas, significantly contributing to global climate goals.

London has 1.9 a population with 1572 km2 with a total of 1.5 million houses,1.9 million flats and 250,000 non-domestic buildings stored in the London Building Stock Model (LBSM) [4,6]. This is a large-scale dataset. Thus, ML techniques are a famous approach widely used to estimate the energy consumption of buildings in London with a large-scale dataset. The algorithm was introduced in early 2000 but recently gained more attention due to the increased usage of the Internet of Things (IoT), Artificial Intelligence (AI) and data storage. Although there is a lot of statistical data on energy consumption from various sectors and sources, the data is limited to building energy consumption in London. The energy values gained for domestic buildings are from the Energy Performance Certificate (EPC), and for public buildings is from Display Energy Certificate (DEC). EPC has no energy values for non-domestic buildings. This is a downscale analysis of energy consumption, mainly for domestic and public buildings in London, using ML algorithms.

Machine Learning and deep learning are the new approaches to estimating building energy consumption. Estimating energy consumption at the early development phase is very useful and is proven to optimize almost 10% to 30% of total energy consumption in a building [7,8]. According to the American Society of Heating, Refrigeration and Air-

Conditioning Engineers (ASHRAE), there are two models to forecast building energy consumption: physics-based models and data-driven models [9]. Physics-based models required massive and detailed input about the buildings and their environment. Heating, ventilation, and air conditioning (HVAC) systems, thermal properties, occupancy numbers, weather and other environmental factors are examples of input used. Common models used for physics-based models are EnergyPlus, TRNSYS, and DOE-2 [8,10]. Thus, these models are the best choice for studying energy consumption efficiency for a building and optimizing a building design with high accuracy as the models have detailed inputs. However, it is impractical as these models are highly computational and too complex [10].

In contrast, data-driven models use historical data to estimate energy consumption where machine learning (ML) models are the best choice as some of the ML models use a statistical approach for estimating target variable that has linear relationships; meanwhile, some of the models can work for both linear and non-linear relationship between other features and the target variable. However, these models still need a large dataset as the model can learn the pattern from the data during training and validate the model's accuracy using the test dataset. In addition, there are a few ML models, such as supervised learning, unsupervised learning, batch learning, online learning, instance-based and model-based learning. The ML models combined supervised, online, and model-based learning in this project. Various ML models can be used to estimate energy consumption, as different algorithms will affect a model's computational efficiency and accuracy. For instance, ML models with ensemble methods such as XGBoost, GB, RF and DT are excellent approaches to determining feature importance and capturing features with non-linear relationship with energy consumption. As technology evolves, deep learning has been introduced as the advanced machine learning model. Most deep learning models are from neural networks that have grown from Artificial Neural Networks (ANN), Deep Neural Networks (DNN), and others, depending on the data type. Deep learning is considered a complex model as it is susceptible and can model large datasets and optimize the architectural model using regularization, dropout, early stopping, and other methods to improve the model.

In addition, a comparison between Extreme Gradient Boosting (XGBoost), Random Forest (RF), SVM and DNN is applied for predicting 24-hour ahead building cooling loads at The Hong Kong Polytechnic University. From this study, XGBoost performs the best among other models in terms of computational cost and computational efficiency. Longer time is taken for RF in capturing feature selection compared to

XGBoost; similarly, XGboost performs well in predicting energy consumption and has a shorter time compared to DNN [11]. In addition, a study on two buildings within a university campus in Tianjin, China, by integrating multiple models such as RF, GB, XGBoost, SVM and K-Nearest Neighbours (KNN) performed the best in terms of accuracy, generalization, and robustness compared to individual performance [12]. Therefore, this study aims to develop ML models that can estimate energy consumption by using linear and non-linear features based on building characteristics, spatial and socioeconomic factors, and the impact of GHG emissions. The objectives of this research are as follows:

- To determine the best Machine Learning models using spatial and socioeconomic factors from EPC, DEC and LBSM for domestic and public buildings
- To do a comparative analysis of the impact of feature selection from EPC, DEC, and LBSM, compared to using only LBSM data, on the accuracy of energy consumption predictions for domestic and public buildings in London
- To investigate the difference in estimated energy consumption between domestic and public buildings based on the features selected from EPC, DEC, and LBSM data
- To study the correlation of estimated energy consumption with CO_2 emissions from EPC and DEC

Chapter 2

Literature Review

2.1 Overview

Each ML model has advantages and disadvantages. It is essential to understand the data's input and output types before deciding which models to use for analysis. Moreover, some algorithms may perform very well in other studies but may not due to various types of data merging and pre-processing. The following sections are the related theories for the various types of ML models used in estimating energy consumption based on their advantages over other models in terms of handling spatial and socioeconomic features.

2.2 Gradient Boosting (GB) and Extreme Gradient Boosting (XGBoost)

Generally, both GB and XGBoost are functions of ensemble models. Two types of functions in Boosting are classifier and regressor for both models. The algorithm will fit the new function to the loss function made by previous function. These ensemble models of different algorithms are the best in handling complexity between linear and non-linear features with target variable and help to reduce overfitting [13,14]. Moreover, XGBoost has received lot of attention in the industry. It is improved in terms of design scalable system to handle a million to billion datasets and the algorithm's overall performance, especially in terms of speed and accuracy, by making the decision-making process during the tree construction. Also, a new algorithm that is aware of data sparsity distribution has been developed in XGBoost, which helps to improve the efficiency of constructing a decision tree. Finally, the author improves the algorithm's performance when handling data that exceeds the memory capacity by choosing the best block size, which is 2^{16} examples per block balance. This helps to process the distribution of computational tasks across multiple processors simultaneously instead of in sequence and reduces the computational time [15].

2.3 Random Forest (RF)

Random Forest is a combination method of decision trees trained through the bagging [13]. RF requires more time to train a dataset as it trains multiple trees. This shows that RF is computationally expensive when handling larger datasets or real-time predictions. However, RF did an excellent job of extracting the importance of features. Each tree in the algorithm is trained on a random batch of datasets. Then, the best feature is selected from a random subset of features at each node of a tree using Gini impurity index [15,17]. Moreover, RF shows its ability to handle null values during training and testing phases, which is an advantage for a real-time prediction [16]. Besides feature selection, RF also used bootstrap resampling, out-of-bag (OOB) estimation and full-depth decision tree growing for the prediction tasks. Bootstrap resampling is used to create multiple training datasets for each decision tree in the forest, where each tree is trained on a different bootstrap sample, which introduces diversity among the trees and helps reduce overfitting. Next, OOB data is predicted from each tree, where OOB error is calculated by summing up the prediction values across the trees and comparing the predicted values with the actual values. Finally, each decision tree in the forest will keep splitting nodes until it stops splitting after meeting a criterion or if all the data points at a node come from the same class. This technique helps reduce overfitting as RF capture detail inputs as much as possible and generalize well to the test dataset [16,17].

2.4 Decision Tree (DT)

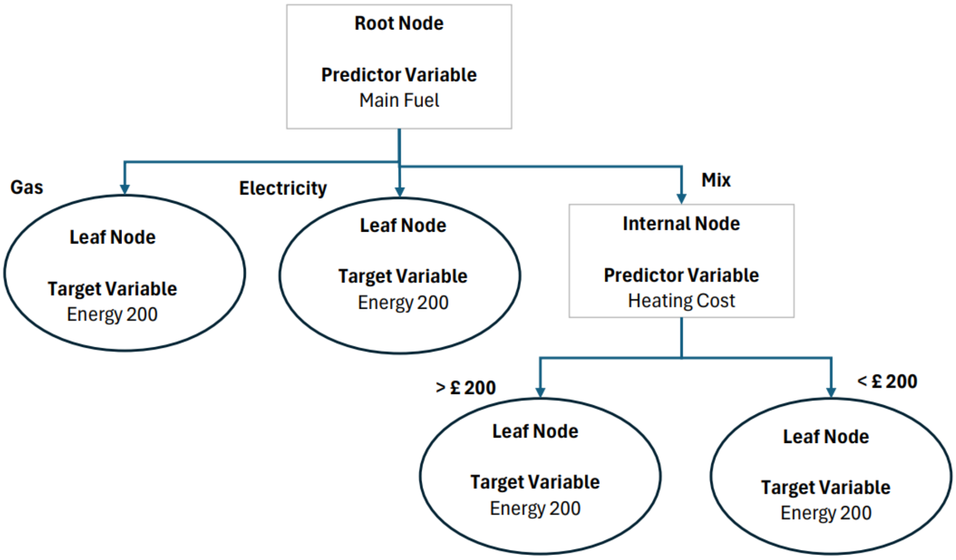


Figure 2.4.1: Illustrative Diagram of annual energy consumption based on the type of primary fuel [8,18]

Decision Tree (DT) is a technique to divide the data into groups using a flowchart like a tree as shown in Figure 2.4.1. The predictor variables are the input into the DT model and are independent of the target variable. The input data will be divided into a few categories based on the splitting criteria and dispersed into branches of nodes where the initial point is the root node and split into sub-nodes. Next, the sub-nodes will split further or stop splitting. The internal node is where the further data split is conducted to create new subcategories. Finally, the leaf nodes will conclude the splitting process and update the data as its final outputs. Figure 2.4.1 represents energy consumption based on the type of main fuel used in a building and divided into another internal node by heating cost. The predictor variable is the other feature used to predict the target variable, where energy consumption (kwh/m^2) is the target value [8,18].

2.5 Deep Neural Network (DNN)

Deep learning is a famous technique for handling complicated tasks requiring five layers or more. DNN is an improved version of an Artificial Neural Network (ANN) that, also called a shallow network, has fewer hidden layers. However, in alignment with its ability to handle complicated tasks, either it's a higher resolution image, natural language processing or time-series prediction, the model will face problems such as gradients fluctuation when moving backwards during training, insufficient training data for a larger network architecture or the time taken to train the data is very slow [13]. In addition, it's not a guarantee that adding more layers into a network can improve the model's accuracy, yet it is time-consuming and a waste if the dataset is smaller, even for ANN [19]. Thus, multiple trials must be tested to find the best tuning, such as the best learning rate, appropriate dropout, regularization, batch normalization and appropriate activation function [13].

2.6 Support Vector Machine (SVM)

Support Vector Machine can be used for classification and regressor tasks. Support Vector Regression (SVR) is widely used to estimate the relationship between non-linear input with and continuous data. For instance, $x_i \in R^N$ is the input while $Y_i \in R^N$ is the target or the continuous value to be targeted. During the training phase, SVR will construct a decision function $F(x_i)$, referring to the previous data. The decision function is assumed in the form of

$$F(x_i) = <w, \varphi (x_i) > + \mathrm{b} \tag{1}$$

where the bias $\mathrm{b} \in R$ and w is the dot product and weight defined in R^N while $\varphi (x_i)$ is a variable for multidimensional features non-linearly with the input data. It is a pre-requisite for the input not to deviate from the real target [8,18,20]. SVM is famous for smaller datasets, between hundreds to thousands of datasets, yet less performance on a bigger dataset [8].

2.7 Applications of Machine Learning in Estimating Building Energy Consumption

Energy consumption for buildings is divided into two categories: energy consumption for domestic and non-domestic buildings. A study has been done at the Northern part of England involving 5000 datasets of domestic buildings for a one-year dataset in 2020. This paper concludes that DNN is the best ML model in terms of R-squared (R^2), accuracy and computational efficiency [8]. ANN, GB, and SVM follow the model performance, while DT demonstrated the best result during training time. However, Moreover, the author emphasizes that the data for houses is the most dominant feature, and the type of building does not affect the model performance after feature selection [8]. In addition, a study on 3840 datasets of domestic buildings in a region of Qassim, Saudi Arabia, shows that ANN performs better compared to Multiple Linear Regression (MLR), where eight types of building characteristics such as glazed area, floor height, building size, wall area, window to wall ratio, Window glazing U-value, roof-U-values and External Wall U-value are the inputs and Heating Load (HL) and Cooling Load (CL) are the predicted outputs [21].

Furthermore, 100,000 data points are collected from domestic buildings in Victoria, Australia, to predict hourly energy consumption based on the occupancy rate, household behaviours, seasonality and the number and types of appliances used in a building. DNN is outperforming compared to MLR, XGBoost and ANN in terms of accuracy and robustness. The study also concludes that the energy consumption rate rose gradually as the number of occupants increased due to the increment in the number of appliances used and more extended usage [22].

A study has been conducted in five different area such as Phoenix (Very hot), Houston (Hot-humid), San Jose (Warm-marine), New York (Mix-humid) and Chicago (Cool-humid) in the United States, to investigate the influence of occupant behaviours and their impact on energy consumption specifically for office buildings. The data is tested with different sample size from 1000 data points to 1 million data points, different ML models and the time taken are observed. The ML models are DNN, ANN, Classification and Regression Trees (CART) and Ensemble Bagging Trees (EBT). DNN and ANN are computationally expensive, starting with 100,000 data points, as running the models took more than 20 minutes. However, it achieves the highest accuracy as the data size increases. However, CART and EBT also achieved more than 90% accuracy starting from the 2000 dataset with a shorter time to run the models. This shows that model selection is based on the objective of the task and the sample size of the datasets. The occupancy behaviours highlighted are the cooling setpoint and the window operation that affect the energy consumption in a building.

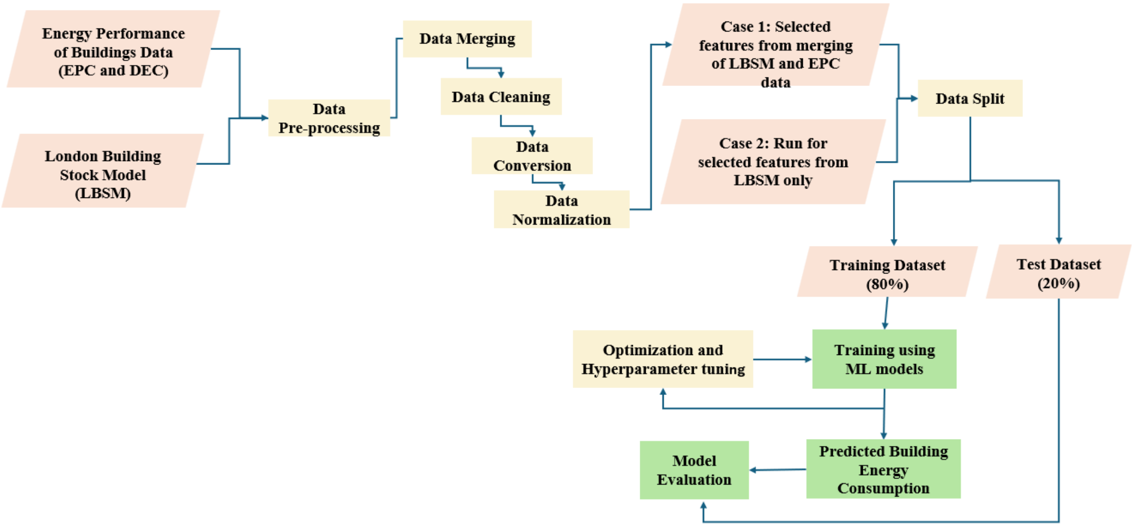
The cooling setpoint is the temperature where the air-conditioning system is set to maintain indoor temperature, whereas a lower cooling setpoint will increase energy consumption as the air-conditioning system needs more energy to maintain the indoor temperature. Similarly, if the window is opened, the energy consumption will increase for cooling and maintain the indoor temperature. Improving these behaviours will improve energy management and efficiency in non-domestic buildings [23]. Environmental factors such as temperature, relative humidity, time variables and chilled water supply system affect the energy consumption of the Hong Kong Polytechnic University, Hong Kong. Unsupervised Deep Learning is used to extract features, and XGBoost turns out to be the best model in terms of prediction energy consumption compared to other inputs of non-linear variables. In this research, DNN is underperforming compared to XGBoost due to the fewest datasets. The author used multiple time variables in terms of weekday and weekend, as well as 24-hour prediction, which helps understand the chiller demand during peak hours. This pattern can help to develop chiller control schemes to maximize energy efficiency for the chiller plants [11].

Chapter 3

Research Methodology

3.1 Overview

This research studies the estimation method to estimate energy consumption for domestic and public buildings in London, United Kingdom (UK). The ML models adopted are Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), Deep Neural Network (DNN) and Support Vector Machine (SVM). These models are selected due to their known advantages in extracting feature importance [11] to train larger datasets [22] and capture features with non-linear relationships with the target variable [24]. Each model is developed using the Python Jupiter environment. The data is split into 80% for training data and 20% for testing data for all models and domestic and public buildings. Referring to Figure 3.1, the flowchart is divided into four phases, which are i) Data collection, ii) Data Pre-processing, iii) Feature Importance selection, iv) Model Development and v) Model evaluation.



Energy Performance of Buildings Data (EPC and DEC)
London Building Stock Model (LBSM)
Data Pre-processing
Data Merging
Data Cleaning
Data Conversion
Data Normalization
Case 1: Selected features from merging of LBSM and EPC data
Case 2: Run for selected features from LBSM only
Data Split
Training Dataset (80%)
Test Dataset (20%)
Optimization and Hyperparameter tuning
Training using ML models
Predicted Building Energy Consumption
Model Evaluation

Figure 3.1.1: Flowchart process of estimating energy consumption for Domestic and Public Buildings in London

3.2 Data Collection

There are two types of data collected namely building data from Energy Performance Certificate (EPC) for domestic buildings and Display Energy Certificate (DEC) for public buildings and spatial data from London Building Stock Model (LBSM) . The building data is collected from the Ministry of Housing Communities and Local Government (MHCLG) repository while spatial data is collected from Greater London Authority (GLA) [3,4].

Building data covers almost 181,000 of domestic buildings and 4530 public buildings approximately across 32 boroughs in London. Both data were collected yearly for each building covering the period from 2017 to 2023. The variables from the EPC are helpful not only during the construction phase or for technology improvement, such as wall description and lighting description, but also include variables from socioeconomic factors such as heating cost, types of tenure, and types of main fuel used in a building. Next, building data from DEC mainly shows the types of primary heating fuel, the presence of air-conditioning, and occupancy level. Moreover, spatial data from LBSM are Output Area (OA), Lower Layer Super Output Area (LSOA), Middle Layer Output Area (MSOA), Output Area Classification (OAC) and Ward code from the Office of National Statistics (ONS) directory that have almost 4.7 million data points with latest update in 2023. Refer to Table 1 and Table 2 from Appendices for the input features for the ML models, meanwhile, the target variable (output) is the energy consumption (kwh/m^2) that comes from EPC for domestic buildings and DEC for public buildings. The analysis has been executed using Python programming language and performed using the following hardware specification (13th Gen Intel Core i7-1365U with 16 GB RAM and 10 cores).

3.3 Data Pre-processing

Data Pre-processing is crucial before data analysis. A few processes are involved in data pre-processing to get clean data. The processes are data merging, data cleaning, and data normalization [8].

Data merging: For domestic buildings, there is 1.034 million data points were collected from EPC. Later, the data from EPC is merged with the data from LBSM through the same column, namely Unique Property Reference Number (UPRN), which represents the unique identity of a building. Similarly, for public buildings, 42724 data points were collected from DEC and merged with data from LBSM through a column, namely UPRN. After merging, the total number of domestic buildings was reduced to 1.029 million data points, while public buildings were reduced to 23622 data points.

Data Cleaning: Each feature is cleaned based on the type of features. The features are continuous data, discrete data, and categorical data. For continuous data, the null value in a cell will be replaced by finding the mean average value between the value from the previous cell and value from the next cell. Besides, discrete data is cleaned using data binning, where the data is placed into grouping and converted the data into categorical data. Finally, the null values for categorical data are removed instantly. After cleaning, only 939,000 data points were available for domestic buildings, while it was reduced to 23586 data points for public buildings. Domestic buildings used 25,000 data points as input, while public buildings used 23586 data points as input.

Data conversion: Most of the data from EPC, DEC, and LBSM are categorical data, such as wall and roof descriptions, occupancy level, ward code, and administrative area. Thus, these features are converted into numerical values using the label encoder library since the categorical data is nominal.

Data Normalization: Data normalization is important to ensure all input features are on the same scale. For example, if a feature has a value between 1 and 100, then another feature has a range between 1000 and 10000. After normalization, all the input features will be on the same scale between 0 and 1 [8]. In this research, the StandardScaler() function is used to normalize all the input features.

$$Z = \frac{(\boldsymbol{x} - \boldsymbol{\mu})}{\boldsymbol{\sigma}} \qquad (2)$$

Where:

- Z is the normalized value.

- x is the actual value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

3.4 Feature selection

Multiple features are available from EPC and DEC, such as spatial, socioeconomic, energy efficiency, energy rating, and emission factors. In contrast, features from LBSM mainly focus on spatial factors. In this research, spatial and socioeconomic factors are merged from EPC, DEC and LBSM to study the relationship between these features with the target variable, energy consumption of a building (kwh/m^2). A study analysis to adopt the Green Deal Assessment (GDA) across 532 Westminster Parliamentary Constituencies (WPCs) in England proves that spatial and socioeconomic factors affect energy efficiency performance and CO_2 emission. However, a policymaker must know that only some policies fit the whole. For instance, GDA implementation is unsuitable for a modern property with energy efficiency technology [25]. Referring to Table 3 from the appendices, the cost of heating, lighting and energy tariff are the socioeconomic factors. In contrast, transaction type can be partially spatial and socioeconomic factors, while others are spatial. Moreover, Table 4 from appendices referring to features from public buildings only have spatial factors.

Feature selection is meaningful to remove irrelevant features that least impactful to the target variable [8,26]. It also helps to improve accuracy and reduce computational time [26]. XGBoost, GB, RF and DT are among the best models for retrieving feature importance [8,11,15,17]. Referring to Table 1 and Table 2 from Appendices, GB, with 90% of R-squared, is the best model to rank the most critical features for domestic buildings, while RF emerged as the highest score (69%) for R-squared of public buildings. Figure 3.4.1 selects the top 15 features as input to predict energy consumption. The features are total floor area, heating cost current, lighting cost current, main heat description, detailed address of a building (Address1), construction age band, roof description, latitude, lighting description, Census Middle Layer Super Output Area (MSOA), main fuel, mix class, number of heated rooms, northing and hot water description, and Gross external area of a self-contained unit (SCU footprint), floor description. Walls description, the height of a building (Mean Height of a object) and Similarly, from Figure 3.4.2, the top 15 features are property type, building category, Gross External Area of a self-contained unit (SCU footprint), total floor area, address, height of a building (Mean Object Height), occupancy level, Longitude, easting, air-condition system (aircon present) , Census Output Area Classification (OAC), Census Output Area (OA), latitude, northing and code for borough (ward code). The features description can be obtained from Table 3 and Table 4 in appendices.

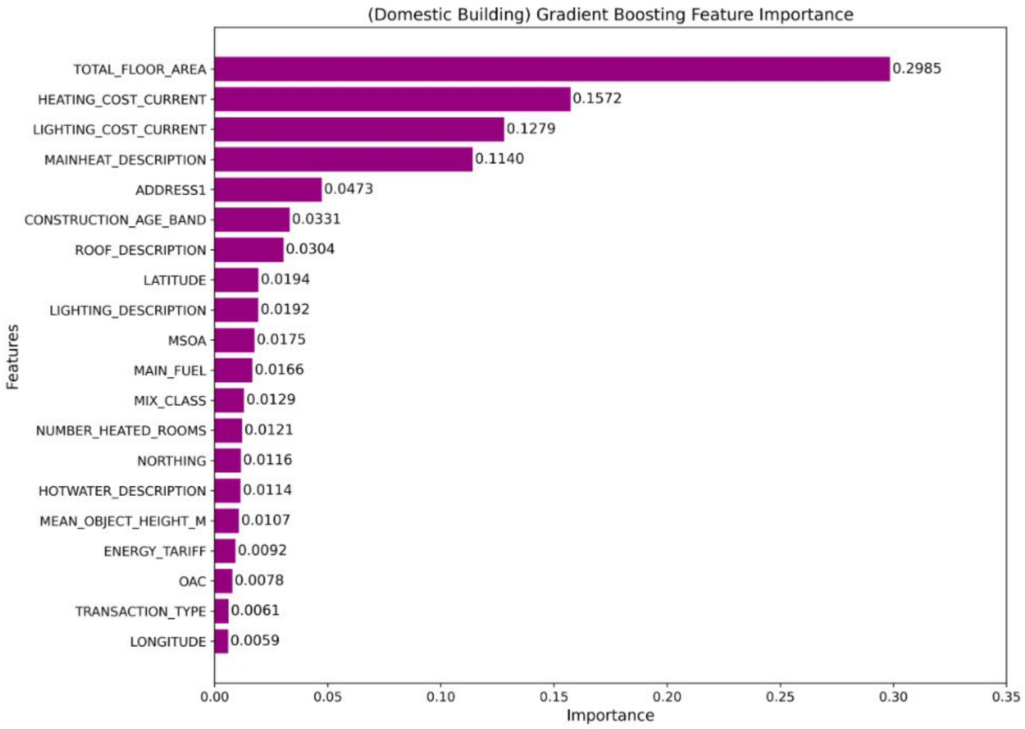


Figure 3.4.1: Feature Importance using Gradient Boosting (Domestic Building)

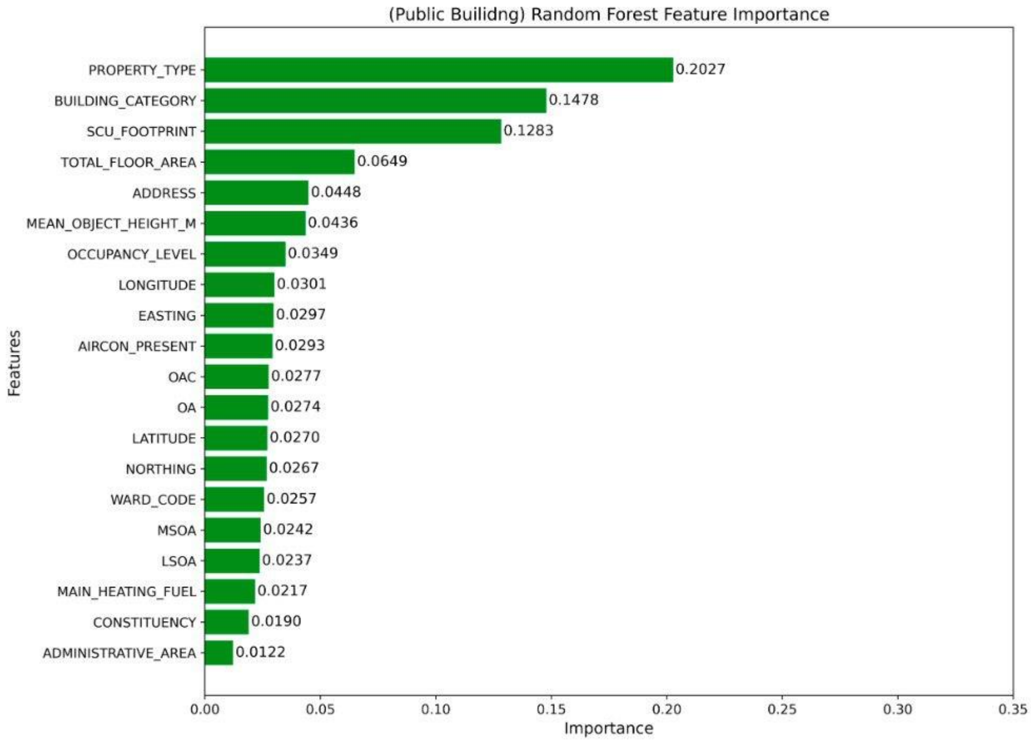


Figure 3.4.2: Feature Importance using Random Forest (Public Building)

3.5 Model Development

After normalization, the data is split into 80% for training and 20% for testing using the train_test_split function. To achieve the highest accuracy, both datasets from domestic and public buildings are trained on six different models in monthly average with different hyperparameters. The models are XGBoost, GB, RF, DT, SVM, and DNN.

For domestic buildings, the hyperparameters used for XGBoost are ‘squared error’ for the ‘objective’, 0.1 for the ‘learning_rate’, the ‘subsample’ is 0.8, the ‘colsample_bytree’ is 0.8, the ‘booster’ is ‘gbtree’ and the number of boosting stages (‘n_estimators’) is 500 same for GB and RF while for GB, 0.2 for the ‘learning_rate’ and 4 for ‘max_depth’ [24]. Next, the hyperparameters used for RF are ‘friedman_mse’ for criterion and 7 for ‘max_features' . The decision tree has 'absolute_error’ as criterion, 42 for the ‘random_state’ and ‘best’ for splitting at each node of a tree, ‘splitter’. The kernel for SVM is ‘rbf', 'C' is 10 and epsilon is 0.01. The architectural model for DNN consists of four layers with 512, 256, 128 and 1 number of neurons at the output layer. ‘Adam’ optimizer is chosen with ‘learning_rate' 0. 001. The activation function is 'relu’ with regularizers at 0.0005 at each hidden layer to avoid overfitting, and the 'linear' activation function is used at the output layer. The model is trained at 400 epochs with loss function is 'mean_absolute_error’.

For public building, the hyperparameters used for GB are 0.1 for the ‘learning_rate’, ‘max_depth' is 5, and 2 for both 'min_samples_leaf’ and ‘min_samples_split’.The hyperparameters for XGBoost are ‘squared_error’ for the ‘objective’, 0.1 for the ‘learning_rate', and the number of boosting stages is 500 for GB and XGBoost. RF hyperparameters are ‘squared_error’ for criterion, 5 for ‘max_features', and the number of boosting stages is 100. Then, DT has 'squared_error’ for criterion with ‘random_state’ is equal to 52. The kernel for SVM is ‘rbf', 'C' is 100, and epsilon is 0.5. The difference from the DNN model used for domestic building is the number of neurons for each hidden layers are 128, 64 and 32 and activation function ‘relu’ for all hidden layers and the output layer.

3.6 Model Evaluation

The performance for each model is evaluated using R-squared (R^2) and Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics where n is the number of observations, y_m is the actual value, $\widehat{y_m}$ is the predicted value and $\overline{y}_m$ is the average of actual values, y_m. The coefficient of determination, or R-squared, is represented by measuring the difference in the target variables and justified by the input features. It does not show the actual size of the prediction errors, but it shows how good the model is to learn the overall data. The value is between 0 and 1, where the goal is to get as near to 1 as it shows the model fits the data very well [8]. Next, MAE is the magnitude difference between the actual and the predicted values. The goal is to get the MAE value as close to 0 since it indicates the model is performed well while close to 1 indicates the model does not learn anything from the data. In addition, MSE calculates the average of the squared differences between predicted and actual values; meanwhile, RMSE is the squared root of MSE.

MSE and RMSE goals are to get values as close to 0 as the lower values of MSE and RMSE show that the model performs very well. The difference between MSE and RMSE is that RMSE has the same unit with the target variables, making it more interpretable compared to MSE. Equations (3) until (6) are the metric equations for the model performance used in this research,

$$R^2 = 1 - \sum_{m=1}^{n} \frac{(y_m - \widehat{y_m})^2}{(y_m - \overline{y}_m)^2} \tag{3}$$

$$MAE = \frac{1}{n} \sum_{m=1}^{n} |y_m - \widehat{y_m}| \tag{4}$$

$$MSE = \frac{1}{n} \sum_{k=1}^{n} (y_m - \widehat{y_m})^2 \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{m=1}^{n} (y_m - \widehat{y_m})^2} \tag{6}$$

Chapter 4

Results and Discussion

4.1 Overview

This chapter identifies four sections from the analysis of EPC, DEC and LBSM datasets. Firstly, the best ML model was determined from the selected features of EPC and LBSM for domestic buildings and DEC and LBSM for public buildings. Next, further analysis was done to compare the best model's performance with spatial data from LBSM for both domestic and public buildings. Next, the estimated energy consumption between domestic and public buildings is compared using Random Forest. Finally, the correlation between estimated energy consumption and CO_2 emission is determined.

4.2 Impact of EPC, DEC, and LBSM Data Integration on Model Performance

There are 15 features selected as input based on feature importance ranking in Figure 3.4.1 and Figure 3.4.2. The features are a mix of spatial and socioeconomic features from EPC, DEC and LBSM. Six models are trained for both datasets: EPC with LBSM and DEC with LBSM. Results shown from Table 4.2.1 show the model's performance for domestic buildings in a rank. GB has emerged as the best model, followed by XGBoost, DNN, RF, SVR and DT. GB depicted the best value for metric values of R-squared, MSE and RMSE. Moreover, it has also been proven that the new algorithm in XGBoost to improve data sparsity distribution by enhancing the construction of a decision tree performs very well with datasets from domestic buildings [15].

Regarding computational efficiency, DNN requires more time as the dataset is run on 400 epochs, as shown in Figure 1 in the appendices. However, although it is time-consuming compared to the other models, the training and validation loss plot in

Figure 1 from the appendices proves that the DNN model is learning effectively in the training dataset as the loss decreased rapidly and started to stabilize after 100 epochs. Although with only three hidden layers, DNN is not overfitting as the training loss (red line) and the validation loss (blue line) are very close to each other. This also explained that DNN generalizes well on the unseen data known as test dataset (validation dataset). Theoretically, SVM is used and perform better with smaller datasets [27]. However, it has also been proven to perform well with a larger dataset in estimating energy consumption [8]. In terms of score for R-squared, all models successfully performed above 0.7. This shows that selecting features based on ranking of feature importance is very important to achieve high accuracy and help to optimize by determining the right tuning to improve the model's performance. This also explained that selected models are well performed in capturing the relationship between non-linear relationship with the target variables, which is the energy consumption.

Table 4.2.1: Model performance based on selected features (Domestic Buildings)

Model	**Training Time (s)**	**R^2**	**MAE**	**MAP**	**RMSE**
Gradient Boosting	**0.307**	**0.898**	**0.132**	**0.063**	**0.251**
Extreme Gradient Boosting	0.270	0.886	0.147	0.070	0.265
Deep Neural Network	567.600	0.846	0.138	0.094	0.308
Random Forest	22.900	0.826	0.156	0.107	0.327
Support Vector Machine	176.620	0.827	0.159	0.106	0.326
Decision Tree	6.195	0.743	0.210	0.158	0.397

Based on Table 4.2.2, RF is depicted as the best model in terms of model evaluation based on metric values of R-squared, MAE, MSE, and RMSE, where MAE, MSE and RMSE for RF are the lowest compared to the other models and has the highest value for R-squared. Generally, GB, XGBoost and RF perform well in capturing non-linear features with the target variable in domestic and public building datasets. DNN take shorter time to run dataset from public building, which is 228.6 seconds. This is due to the lower number of neurons used for the DNN model of public buildings compared to the DNN model of domestic buildings. Figure 2 from the appendices shows that the DNN model generalizes well with the unseen data, while the training and validation lines close to each other indicate minimal overfitting. However, the MAE value is still high and constant after more than 100 epochs. This indicates that the model has stopped learning. However, some optimization has been applied to the model, such as dropout, reducing the learning rate, adding regularizers and reducing the model's complexity.

Table 4.2.2: Model performance based on selected features (Public Buildings)

Model	**Training Time (s)**	**R^2**	**MAE**	**MAP**	**RMSE**
Random Forest	**6.330**	**0.726**	**0.277**	**0.250**	**0.500**
Gradient Boosting	21.640	0.657	0.309	0.313	0.560
Extreme Gradient Boosting	0.390	0.635	0.340	0.333	0.577
Decision Tree	0.028	0.598	0.321	0.367	0.606
Support Vector Machine	19.310	0.563	0.400	0.399	0.632
Deep Neural Network	228.600	0.420	0.470	0.520	0.720

4.3 Comparative Analysis of Energy Consumption Patterns in Domestic and Public Buildings

The best ML model for domestic buildings is GB, while the best ML model for public buildings is RF. Figure 3 from the appendices compares actual versus estimated energy consumption for public and domestic buildings using RF from 2017 to 2023. To compare the estimated energy consumption between public and domestic buildings, same model is used to get the best outcome. Compared to the R-squared values of domestic building are higher than that of public building. Thus, the estimated energy consumption value in 2020 for bon from RF is used for both types of buildings for further analysis. The estimated energy consumption for both buildings are from the test dataset (20%) from the input data.

Based on Figure 4.3.1, the estimated energy consumption of domestic building increased gradually from 150 kwh/m^2 and reached its peak in 2019, almost 300 kwh/m^2, and slightly reduced in 2020. meanwhile, estimated energy consumption for public buildings decreased gradually from 150 kwh/m^2 in 2018 to 100 kwh/m^2 in 2020 due to the lockdown from March 2020 until May 2020 consequences of the global coronavirus pandemic (COVID-19), where all public buildings are closed during the pandemic. This reduces the occupancy rate as most of the occupants work from home.

In contrast, energy used in domestic buildings decreased slightly from 2019 to 2020, but not a steep decline compared to the estimated energy consumed by public buildings. This shows that the energy consumption usage for domestic buildings is relatively the same during pre-COVID-19 and lockdown, as the estimated energy usage does not increase significantly nor decrease steadily. Households with financial difficulties will install in-home displays (IHD), a device that is connected to a smart meter that provides real-time information for a household's energy consumption. Heating behaviour may vary depending on a person's behaviour pre-COVID-19 and during lockdown. For instance, those who used to work from home during pre-COVID-19 tend to consume less heating and other electrical appliances compared to those who just experienced working from home during the lockdown [28]. This shows that socioeconomic features such as the cost of heating, lighting, and rate of energy tariffs impact estimating energy consumption for domestic buildings.

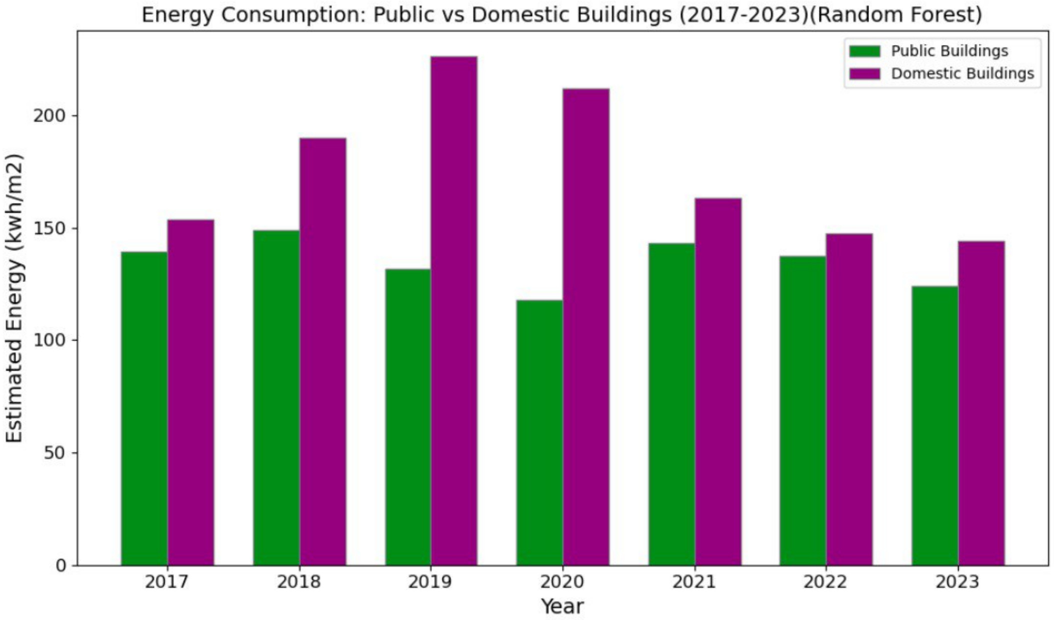


Figure 4.3.1: Estimated Energy Consumption for Public and Domestic Buildings (2017 – 2023) (Random Forest)

In addition, the estimated energy consumption of public buildings increased gradually after the lockdown, while the energy consumption of domestic buildings decreased gradually. This illustrates that some policies implemented to improve energy efficiency usage for domestic buildings are successful. Based on Department of Energy and Climate Change (DECC) policies along with Energy Company Obligation (ECO) aimed to improve energy efficiency for domestic buildings and houses, especially for lower-income households, by rollout smart meters, replacing inefficient boilers with boilers that are energy saver and expected to cover 10.5 million homes by 2020. This explains the significant energy reduction from 2020 until 2023 [31]. Moreover, the Green Deal policy and ECO allowed homeowners to improve their energy efficiency usage with no upfront costs. Instead, the cost is repaired by saving energy bills [31]. This shows that socioeconomic factors help policymakers create policies for targeted households and manage to reduce energy consumption for domestic buildings in London. After the lockdown, the energy consumption of public buildings was relatively the same as that of people returning to their normal routine activities. The estimated energy consumption is kept below than 150 kwh/m^2.

4.4 Analyzing Model Effectiveness Using Solely LBSM Data

Based on Table 3 and Table 4 from appendices, features number 1 until number 15 are chosen as input features for domestic and public buildings. From Table 3, the features from LBSM are Census Middle Layer Super Output Area (MSOA), Latitude, Mix class and Northing, while others are features from EPC. Figure 4.4.1 depicts the difference of metric values R-squared, MAE, MSE and RMSE from GB between two different datasets from domestic buildings. The blue bar represents metric values from LBSM only, while the orange bar represents metric values from EPC and LBSM. R-squared has improved significantly when the GB model used a combination of EPC and LBSM as input features.

Moreover, MAE, MSE, and RMSE have been reduced to 0.74, 0.88, and 0.66, respectively. It was also concluded that using the input features from EPC and LBSM is more accurate in improving model performance than input from LBSM alone. This explained the importance of spatial and socioeconomic factors from EPC influenced the target variable (energy consumption).

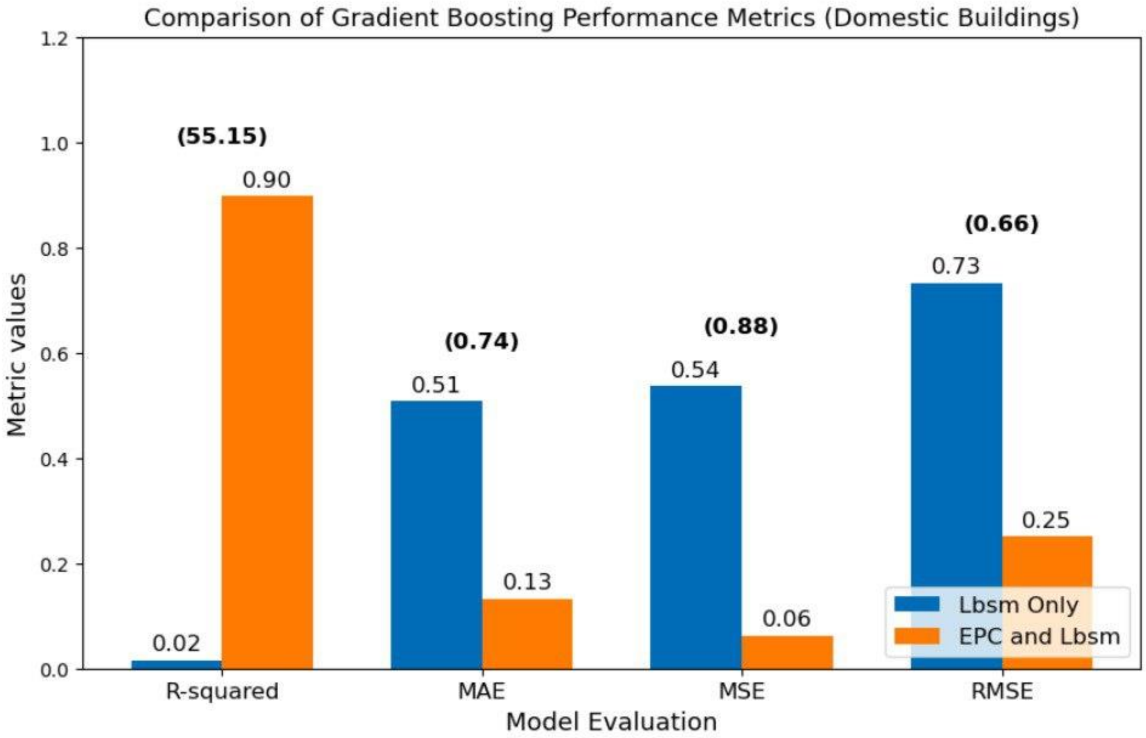


Figure 4.4.1: Metrics comparison with different dataset (Domestic Building)

Based on Figure 4.4.2, the significant difference is on MSE value which is 0.33, while others metric values have reduced to RMSE (0.18), R-squared (0.20) and MAE (0.16). The percentage difference for public buildings is lower for all metric values is due to more LBSM features as input when merged with features from DEC. The LBSM features from the merged are SCU footprint, Mean Object height, Census Output Area (OA), Census Output Area Classification (OAC), Easting, Air-condition presence, Northing, Ward code and Longitude. However, GB performed very well by combining DEC input features instead of LBSM alone. Metric values of the orange bar (DEC and LBSM) win overall metric values of the blue bar (LBSM only) where R-squared near one while MAE, MSE and RMSE are near zero for an orange bar in comparison with the metric values of the blue bar.

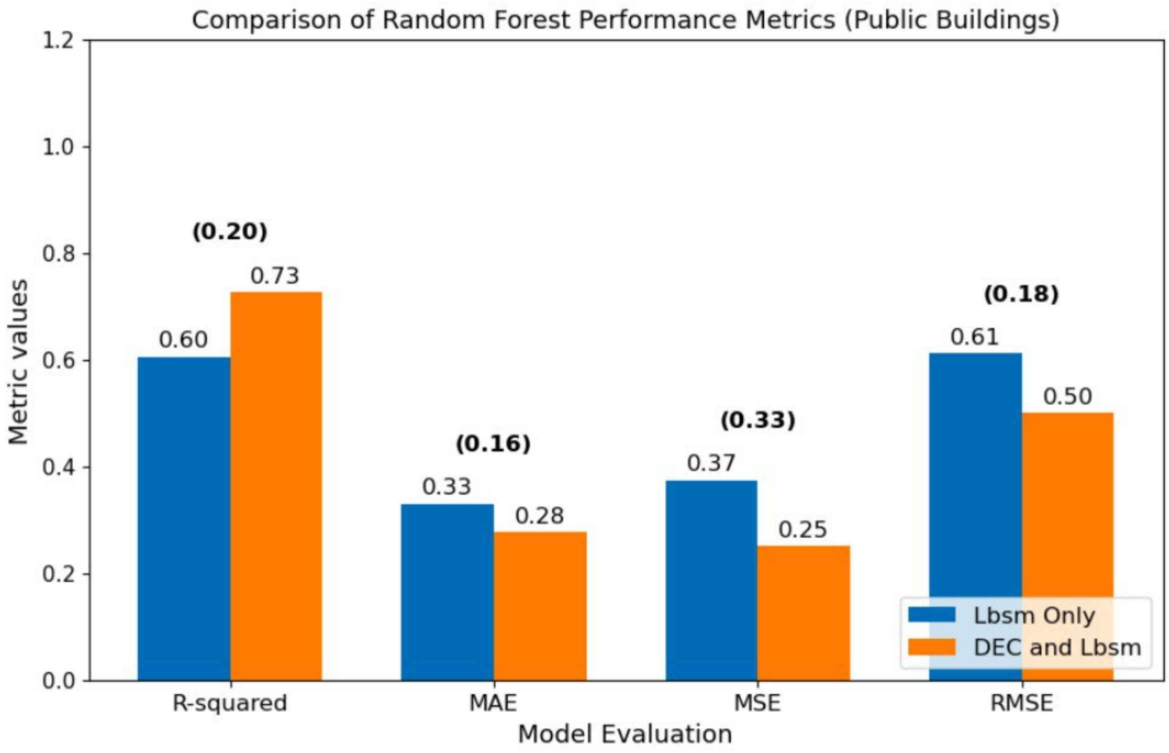


Figure 4.4.2: Metrics comparison with different datasets (Public Building)

4.5 A Comparative Analysis of Estimated Energy Consumption CO2 Emissions

Based on Figure 4.5.1 and Figure 4.5.2 shows a correlation between estimated energy consumption for domestic buildings with CO_2 emissions per floor area from EPC using a regression plot and polynomial fit plot. Figure 4 from the appendices shows that the actual energy consumption values with CO_2 emissions per floor area are highly correlated. However, the estimated energy consumption using spatial and socioeconomic factors as input features shows that the estimated energy values are not correlated with the CO_2 emissions per floor area. Polynomial fit determines the non-linear relationship between the estimated energy consumption and the CO_2 emissions per floor area. Moreover, Figure 5 and Figure 6 show that the estimated energy consumption of public buildings also has no significant correlation with heating CO_2 emissions and electricity CO_2 emissions from DEC. This also shows that this low-performance method compares estimated energy consumption with CO_2 emissions using spatial and socioeconomic factors as input features. Other features that can be considered when estimating energy consumption to compare with CO_2 emissions are the environmental features such as temperature, pressure and wind speed. However, since the temperature in London will be relatively the same, one should consider comparing with other regions.

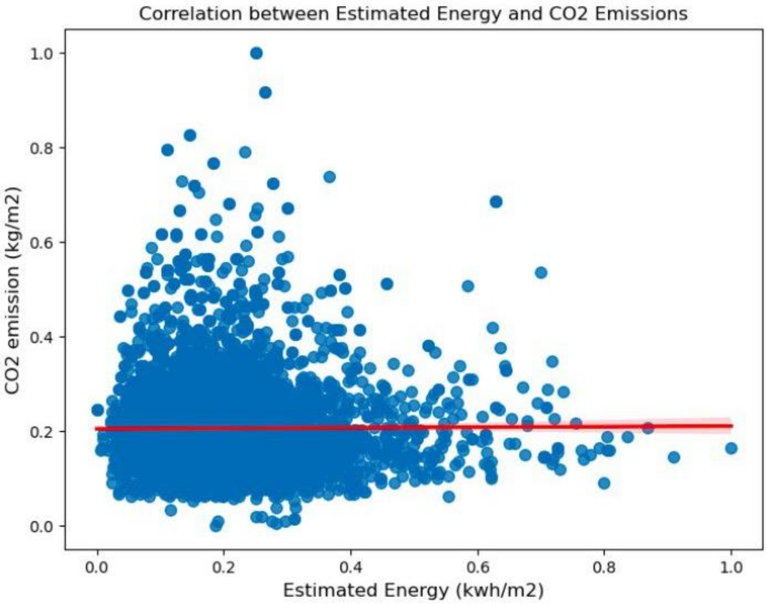


Figure 4.5.1: Correlation between estimated energy consumption with CO_2 emission of domestic building using regression plot

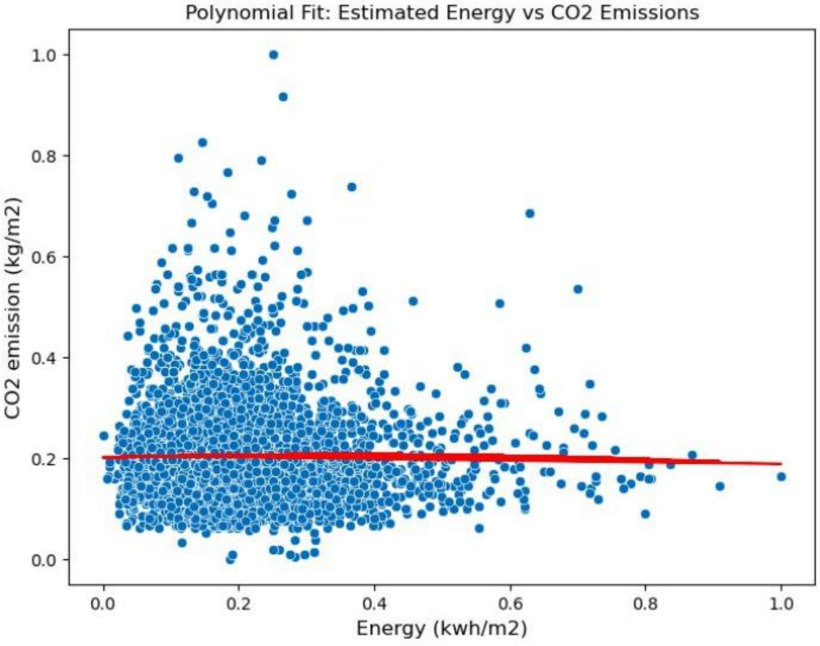


Figure 4.5.2: Correlation between estimated energy consumption with CO_2 emission of the domestic building using polynomial plot

Conclusion

This study compared energy consumption estimation for domestic and public buildings using six machine learning models: XGBoost, GB, DT, RF, DNN, and SVM. The data used in this study is the merging of data from EPC and DEC with LBSM that focus on spatial and socioeconomic factors that influenced the energy consumption for domestic and public buildings. The best model is identified for two different datasets. GB is the best model for computational efficiency and metric values for domestic buildings, while RF is the best model as it depicts the best scores for all metric values for public buildings. However, other models score more than 0.7 of R-squared for domestic buildings, indicating that the model performs well with the dataset, while DNN and SVM perform less with the dataset from the public buildings. Next, estimated energy consumption for both buildings is gained from RF model since RF performed well with an R-squared of more than 0.7 for both datasets.

Furthermore, the estimated energy consumption of domestic buildings increased gradually from 150 kWh/m^2 and reached its peak in 2019 before being reduced in 2020. meanwhile, estimated energy consumption for public building decreased gradually from 150 kWh/m^2 in 2018 to 100 kWh/m^2 in 2020 due to the lockdown from March 2020 until May 2020 consequences of the global coronavirus pandemic (COVID-19) where all public buildings are closed during the pandemic. This reduces the occupancy rate as most of the occupants work from home. The energy consumption for domestic buildings does not increase abruptly during COVID-19 as people who used to work from home during pre-COVID-19 get used to energy saving compared to those who experienced working from home during lockdown. This shows that socioeconomic factors do influence energy consumption. After the year 2020, the estimated energy consumption of both buildings kept decreasing due to the impact of policies implemented, such as DECC policies and the Green Deal policy that focus more on targeted groups to improve energy efficiency [31]. However, studies need to be done on the impact of retrofitting buildings after improvement has been implemented, such as adding solar panels, changing heat pumps or installing energy-efficient windows or doors. Besides, a study on to maintain the energy efficiency of retrofitting building is crucial as old or refurbished building always required proper maintenance and care to preserve its quality.

Based on Figure 7 from appendices, the highest construction age band for buildings in London is band 1900-1929 and almost more than 400 thousand of domestic buildings in a range of 1900-1949, while a study on spatial regression analysis in London shows almost 60% of domestic buildings in London were built before 1944 and only 5% of the buildings dated after 1999 [30]. This represents that most domestic buildings are old, and physically, they are less energy efficient than modern buildings. Moreover, households with higher incomes consume more energy due to multiple electrical appliances used in a household [29]. This indicates that spatial and socioeconomic features from EPC, DEC, and LBSM are vital to estimate domestic energy consumption, and they are a guide for policymakers to consider spatial and socioeconomic factors before implementing any policies to improve energy efficiency for domestic and public buildings. Moreover, the policymakers can implement the policies on a few targeted group first.

The limitation of this research is because the data is yearly data. Thus, it focuses more on spatial and other factors than temporal prediction. Therefore, only after a few years can one evaluate the impact of policies implemented to improve energy efficiency and reduce CO_2 emission production from domestic and public buildings in London. Moreover, the data on energy consumption from non-domestic buildings is not available to the public due to right issue. Future work can be done by using other data with lower temporal resolution, such as hourly to daily for energy consumption and CO_2 emission from different regions that have different climate conditions. This will help policymakers study the pattern of energy consumption and CO_2 emissions not only from spatial and socioeconomic factors but also from environmental factors without waiting for a few years to see the impact of the policies implemented.

Acknowledgements

I would like to thank my supervisors, Dr. Heather Graven and Dr. Fangxin, for guiding me to finish this research project. I would also like to thank my colleagues Enas, Iman, Raif, Palm, Liam, and Marcelo for all their support and fruitful discussions. Foremost, I thank my mum, my sister, and my family for their constant support throughout this journey.

Bibliography

1. Department for Energy Security and Net Zero. (2023). *Energy Consumption in the UK 2023*. UK Government.

   https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_ data/file/1187617/Energy_Consumption_in_the_UK_2023.pdf

2. Department for Energy Security and Net Zero. (2024). *Digest of UK Energy Statistics (DUKES) 2024: Chapters 1-7*. UK Government.

   https://assets.publishing.service.gov.uk/media/66a7e14da3c2a28abb50d922/DUKES_202 4_Chapters_1-7.pdf

3. Energy Performance Certificate Register. (n.d.). Domestic Energy Performance Certificates search. Open Data Communities. Retrieved [August 30 2024], from https://epc.opendatacommunities.org/domestic/search

4. Steadman, P., Evans, S., Liddiard, R., Godoy-Shimizu, D., Ruyssevelt, P., & Humphrey, D. (2020). Building stock energy modelling in the UK: The 3DStock method and the London Building Stock Model. *Buildings and Cities, 1*(1), 100–119. https://doi.org/10.5334/bc.52

5. Duan, H., Chen, S., & Song, J. (2022). Characterizing regional building energy consumption under joint climatic and socioeconomic impacts. *Energy, 245*, 123290. https://doi.org/10.1016/j.energy.2022.123290

6. Gardham, R. (2022, April 11). *The 25 largest cities in the UK (and their investment strengths)*. Investment Monitor. https://www.investmentmonitor.ai/features/largest-cities-uk-investment-strengths/

7. A. Colmenar-Santos, et al., Solutions to reduce energy consumption in managing large buildings, Energy Build. 56 (2013) 66–77, https://doi.org/ 10.1016/j.enbuild.2012.10.004.

8. Olu-Ajayi, R., Alaka, H., Sulaimon, I., Sunmola, F., & Ajayi, S. (2022). Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques. *Journal of Building Engineering*, *45*, 103406-. https://doi.org/10.1016/j.jobe.2021.103406

9. American Society of Heating, Refrigeration and Air-Conditioning Engineers, ASHRAE handbook: Fundamentals, in American Society of Heating, Refrigerating and Air-Conditioning Engineers, ASHRAE, Atlanta, GA, USA, 2009.

10. J. Runge, R. Zmeureanu, Forecasting energy use in buildings using artificial neural networks: a review, Energies 12 (17) (2019) 3254, https://doi.org/10.3390/

En12173254.

11. C. Fan, F. Xiao, Y. Zhao, A short-term building cooling load prediction method using deep learning algorithms, Appl. Energy 195 (2017) 222–233, https://doi.org/10.1016/j.apenergy.2017.03.064.
12. R. Wang, S. Lu, W. Feng, A novel improved model for building energy consumption prediction based on model integration, Appl. Energy 262 (2020) 114561, https://doi.org/10.1016/j.apenergy.2020.114561.
13. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
14. R.A. Berk, An introduction to ensemble methods for data analysis, Socio. Methods Res. 34 (3) (2006) 263–295, https://doi.org/10.1177/0049124105283119
15. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
16. M.W. Ahmad, M. Mourshed, Y. Rezgui, Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption, Energy Build. 147 (2017) 77–89, https://doi.org/10.1016/j.Enbuild.2017.04.038.
17. R. Jiang, W. Tang, X. Wu, W. Fu, A random forest approach to the detection of epistatic interactions in case-control studies, BMC Bioinform. 10 (1) (2009) 1.
18. Y. Wei, et al., A review of data-driven approaches for predicting and classif ying building energy consumption, Renew. Sustain. Energy Rev. 82 (2018) 1027–1047, https://doi.org/10.1016/j.rser.2017.09.108.
19. K. Amasyali, N. El-Gohary, Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings, Renew. Sustain. Energy Rev. 142 (2021) 110714, https://doi.org/10.1016/j.rser.2021.110714.
20. Dong, B., Cao, C., & Lee, S. E. (2005). Applying support vector machines to predict building energy consumption in tropical regions. *Energy and Buildings, 37*(6), 545-553. https://doi.org/10.1016/j.enbuild.2004.09.009
21. Ibrahim, D. M., Almhafdy, A., Al-Shargabi, A. A., Alghieth, M., Elragi, A., & Chiclana, F. (2022). I am using statistical and machine learning tools to accurately quantify the energy performance of residential buildings. *PeerJ Computer Science, 8*, e856. https://doi.org/10.7717/peerj-cs.856
22. Truong, L. H. M., Chow, K. H. K., Luevisadpaibul, R., Thirunavukkarasu, G. S., Seyedmahmoudian, M., Horan, B., Mekhilef, S., & Stojcevski, A. (2021). Accurate prediction of hourly energy consumption in a residential building based on the occupancy rate using machine learning approaches. *Applied Sciences, 11*(5), 2229. https://doi.org/10.3390/app11052229
23. K. Amasyali, N. El-Gohary, Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings, Renew. Sustain. Energy Rev. 142 (2021) 110714, https://doi.org/10.1016/j.rser.2021.110714.

24. AlShafeey, M., & Rashdan, O. (2023). Quantifying the impact of energy consumption sources on GHG emissions in major economies: A machine learning approach. *Energy Strategy Reviews, 49*, 101159. https://doi.org/10.1016/j.esr.2023.101159

25. Morton, C., Wilson, C., & Anable, J. (2018). The diffusion of domestic energy efficiency policies: A spatial perspective. *Energy Policy*, 114, 77–88. https://doi.org/10.1016/j.enpol.2017.11.057

26. Zhao H-X, Magoulès F. Feature Selection for Predicting Building Energy Consumption Based on Statistical Learning Method. Journal of Algorithms & Computational Technology. 2012;6(1):59-77. doi:10.1260/1748-3018.6.1.59

27. Liu, Y., Chen, H., Zhang, L., Wu, X., & Wang, X. (2020). Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: A case study in China. *Journal of Cleaner Production, 272*, 122542. https://doi.org/10.1016/j.jclepro.2020.122542

28. Huebner, G. M., Watson, N. E., Direk, K., McKenna, E., Webborn, E., Hollick, F., Elam, S., & Oreszczyn, T. (2021). Survey study on energy use in UK homes during Covid-19. *Buildings and Cities*, 2(1), 952–969. https://doi.org/10.5334/bc.162

29. Jones, R. V., Fuertes, A., & Lomas, K. J. (2015). The socioeconomic, dwelling and appliance-related factors affecting electricity consumption in domestic buildings. *Renewable and Sustainable Energy Reviews*, *43*, 901–917. https://doi.org/10.1016/j.rser.2014.11.084

30. Tian, W., Song, J., & Li, Z. (2014). Spatial regression analysis of domestic energy in urban areas. *Energy*, 76, 629-640. https://doi.org/10.1016/j.energy.2014.08.057

31. Department of Energy and Climate Change. (2013). *Policy impacts on prices and bills*. GOV.UK. https://www.gov.uk/guidance/policy-impacts-on-prices-and-bills

32. Fatihah (2024). RESEARCH PROJECT [GitHub repository]. GitHub. https://github.com/fatihiey/RESEARCH-PROECT.git

Appendices

Table 1: Model performance for feature importance (Domestic Buildings)

Model	Training Time (s)	R-squared
Extreme Gradient Boosting	1.80	0.89
Gradient Boosting	**126.93**	**0.90**
Random Forest	80.69	0.83
Decision Tree	0.18	0.49

Table 2: Model performance for feature importance (Public Buildings)

Model	Training Time (s)	R-squared
Extreme Gradient Boosting	0.33	0.58
Gradient Boosting	25.53	0.64
Random Forest	**32.74**	**0.69**
Decision Tree	0.06	0.55

Table 3: Features from EPC and LBSM (Domestic Buildings) [3,4]

No.	Features	Abbreviation	Unit	Type of data	Label
10	Energy Consumption	Energy Consumption	kWh/m^2/yr	Continuous	Output
2	Total Floor Area	Total Floor Area	m^2		Input
3	Heating Cost Current	Heating Cost Current	NA	Categorical	
4	Lighting Cost Current	Lighting Cost Current			
5	Mainheat Description	Mainheat Description			
6	Detail address of a building	Address1			
7	Age of a building	Construction Age Band			
8	Roof Description	Roof Description			
9	A value of Latitude location based on ETRS89 coordinate reference system from OS AddressBase	Latitude			
10	Lighting Description	Lighting Description			
11	Census Middle Layer Super Output Area	MSOA			
12	Main Fuel	Main Fuel			
13	Mix class	Mix class			
14	Number of heated rooms	Number Heated Rooms			
15	The x location based on OS AddressBase	Northing			
16 17	Hotwater Description	Hotwater Description			
18	Energy Tariff	Energy Tariff			
19	Census Output Area classification	OAC			
20	Type of transaction that triggered EPC (mandatory issue)	Transaction type			
21	Height of a building	Mean Object Height M	m		

Table 4: Features from DEC and LBSM (Public Buildings) [3,4]

No.	Features	Abbreviation	Unit	Type of data	Label
1	Energy Consumption	Energy Consumption	kWh/m²/yr	Continuous	Output
2	Total Floor Area	Total Floor Area	m²		Input
3	Gross External area of a self-contained units (SCU)	SCU Footprint	m²		
4	Height of a building	Mean Object Height M	m		
5	Type of property	Property Type		Categorical	
6	Mix uses in a building (retail store, office space, etc)	Building Category			
7	Detail address of a building	Address			
8	Total number of occupants	Occupancy Level			
9	A value of Longitude location based on ETRS89 coordinate reference system from OS AddressBase	Longitude			
10	The y location based on OS AddressBase	Easting			
11	Air condition system	Aircon present			
12	Census Output Area classification	OAC			
13	Census Output Area	OA			
14	A value of Latitude location based on ETRS89 coordinate reference system from OS AddressBase	Latitude			
15	The x location based on OS AddressBase	Northing			
16	The code for the current borough based ONS Postcode Directory.	Ward Code			
17	Census Middle Layer Super Output Area	MSOA			
18	Census Lower Layer Super Output Area	LSOA			
19	The main type of fuel used	Main Heating Fuel			
20	Constituency	Constituency			
21	Borough based on OS AddressBase	Administrative Area			

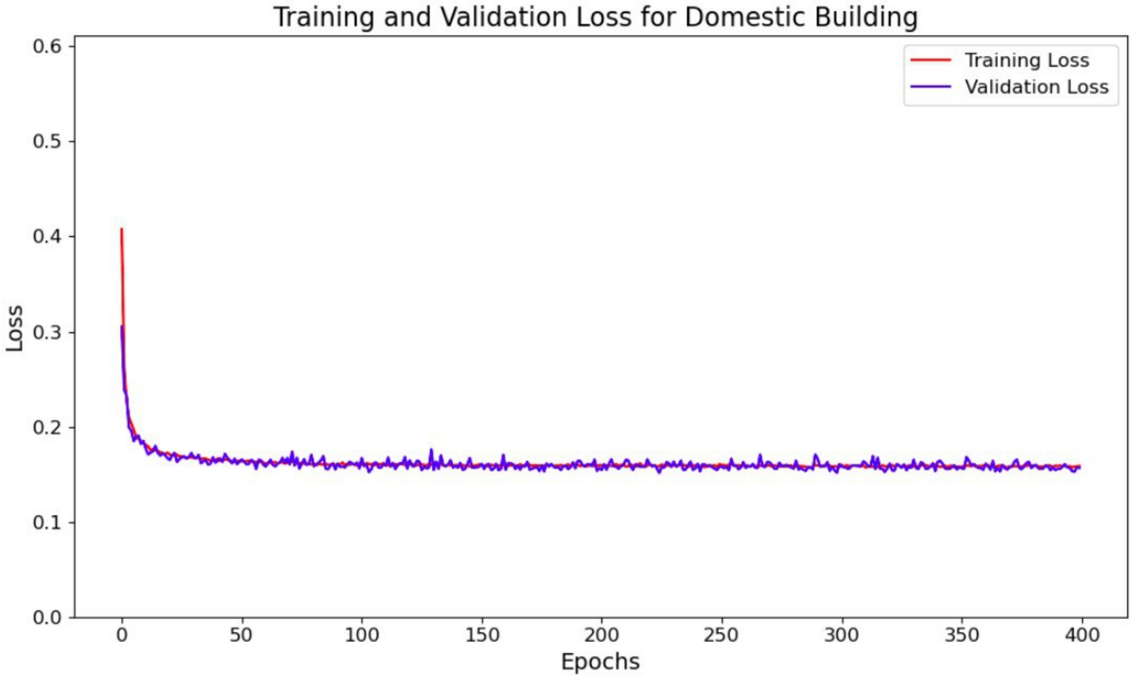


Figure 1: Training and Validation Loss for Domestic Buildings from DNN

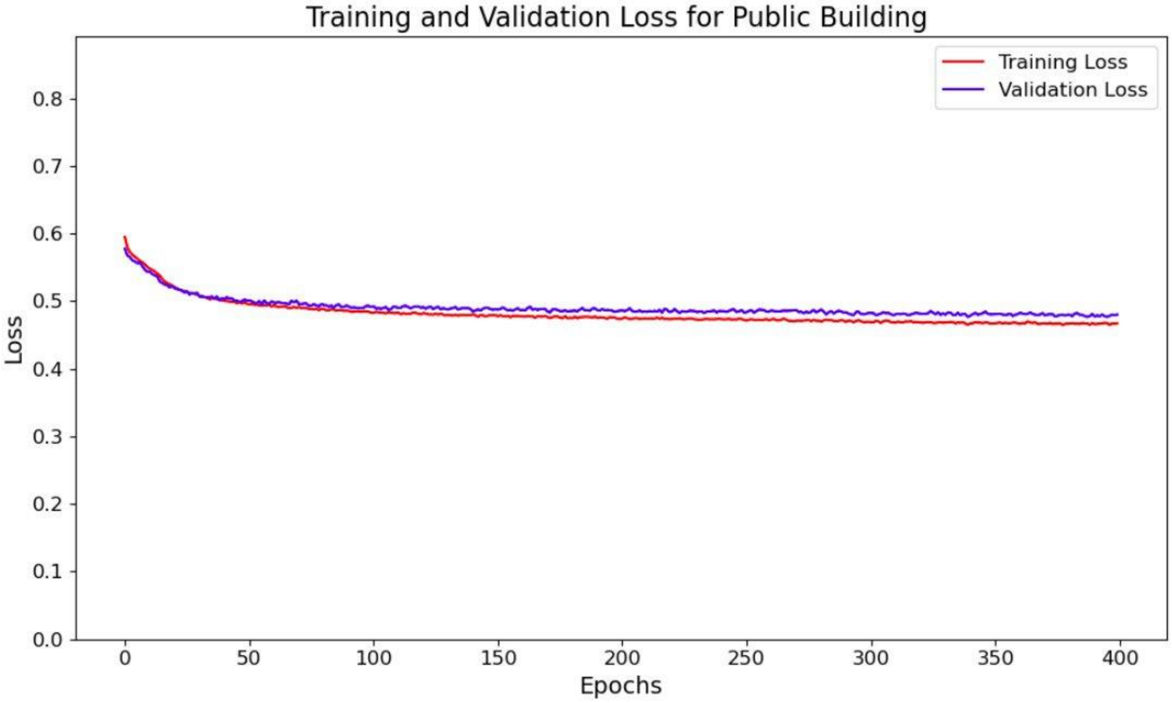


Figure 2: Training and Validation Loss for Public Buildings from DNN

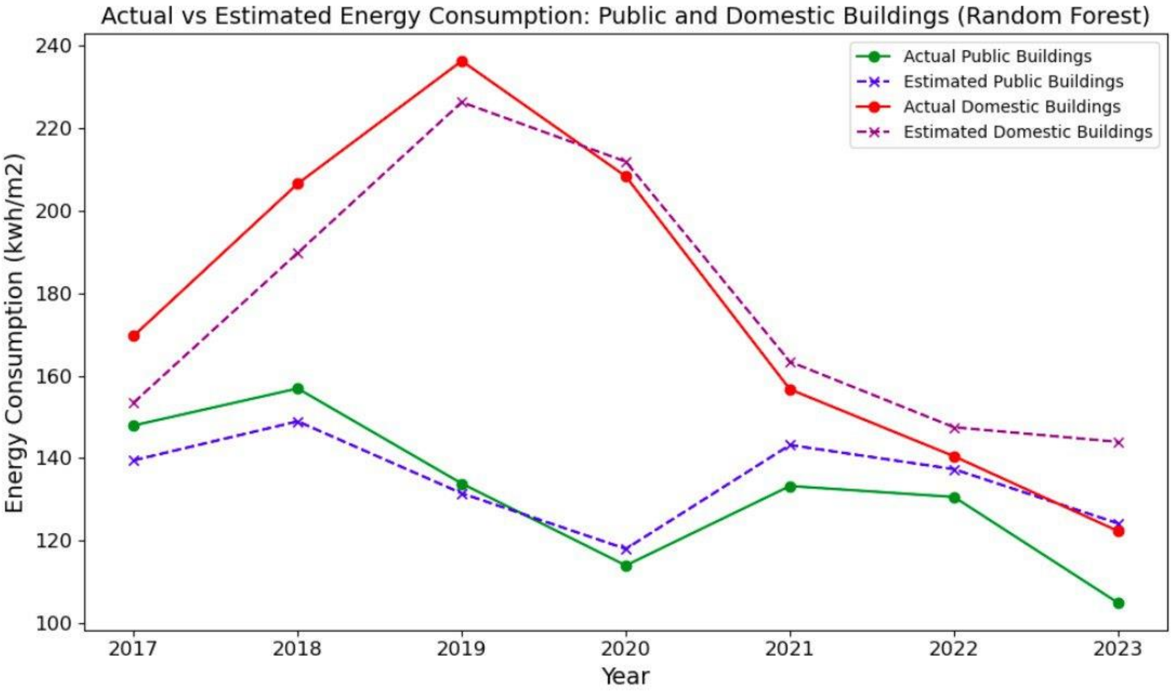


Figure 3: Comparison of Actual versus Estimated Energy Consumption for Public and Domestic Buildings (2017–2023) Using Random Forest

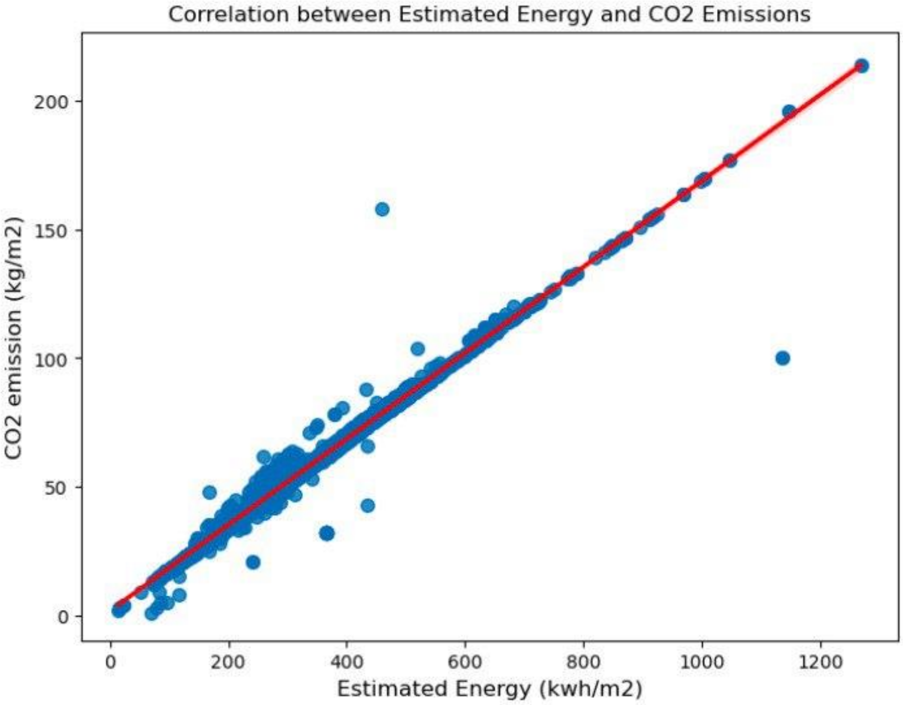


Figure 4: Correlation between actual values of energy consumption with CO_2 emission of domestic building

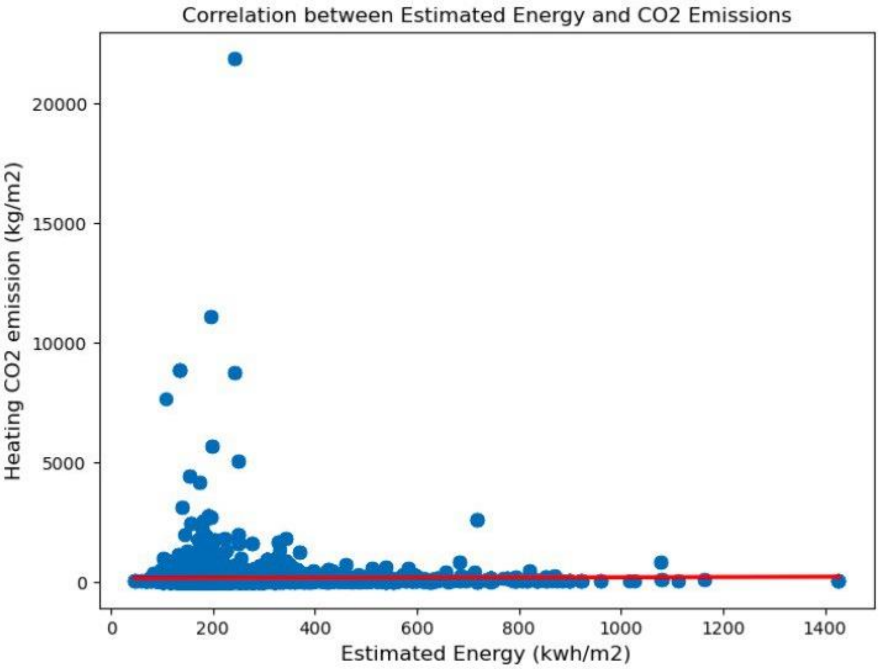


Figure 5: Correlation between estimated energy consumption with Heating CO_2 emission of public building

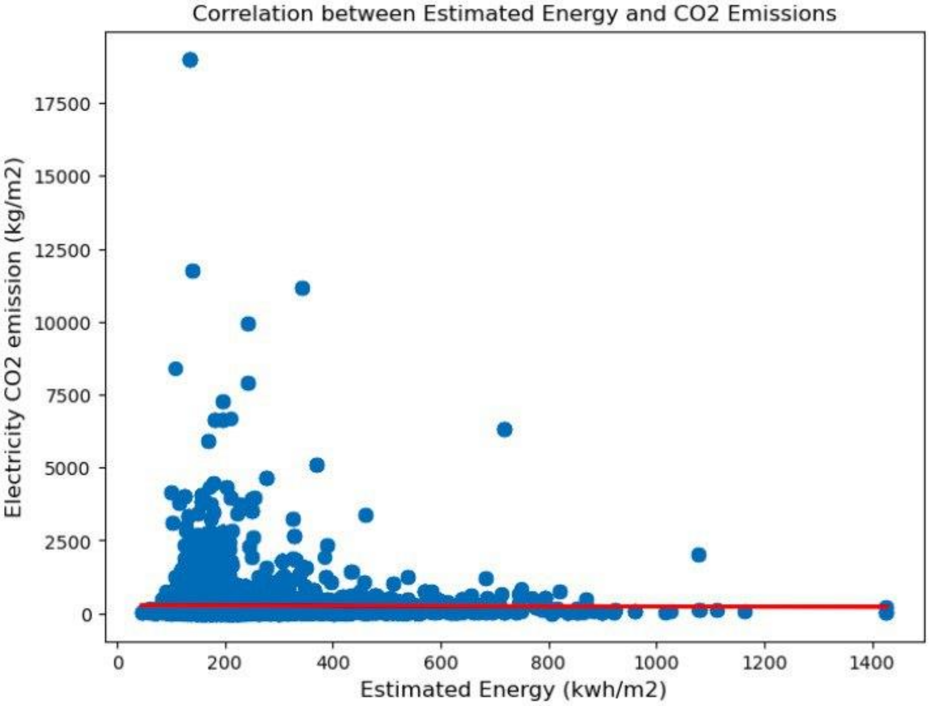


Figure 6: Correlation between estimated energy consumption with Electricity CO_2 emission of public building

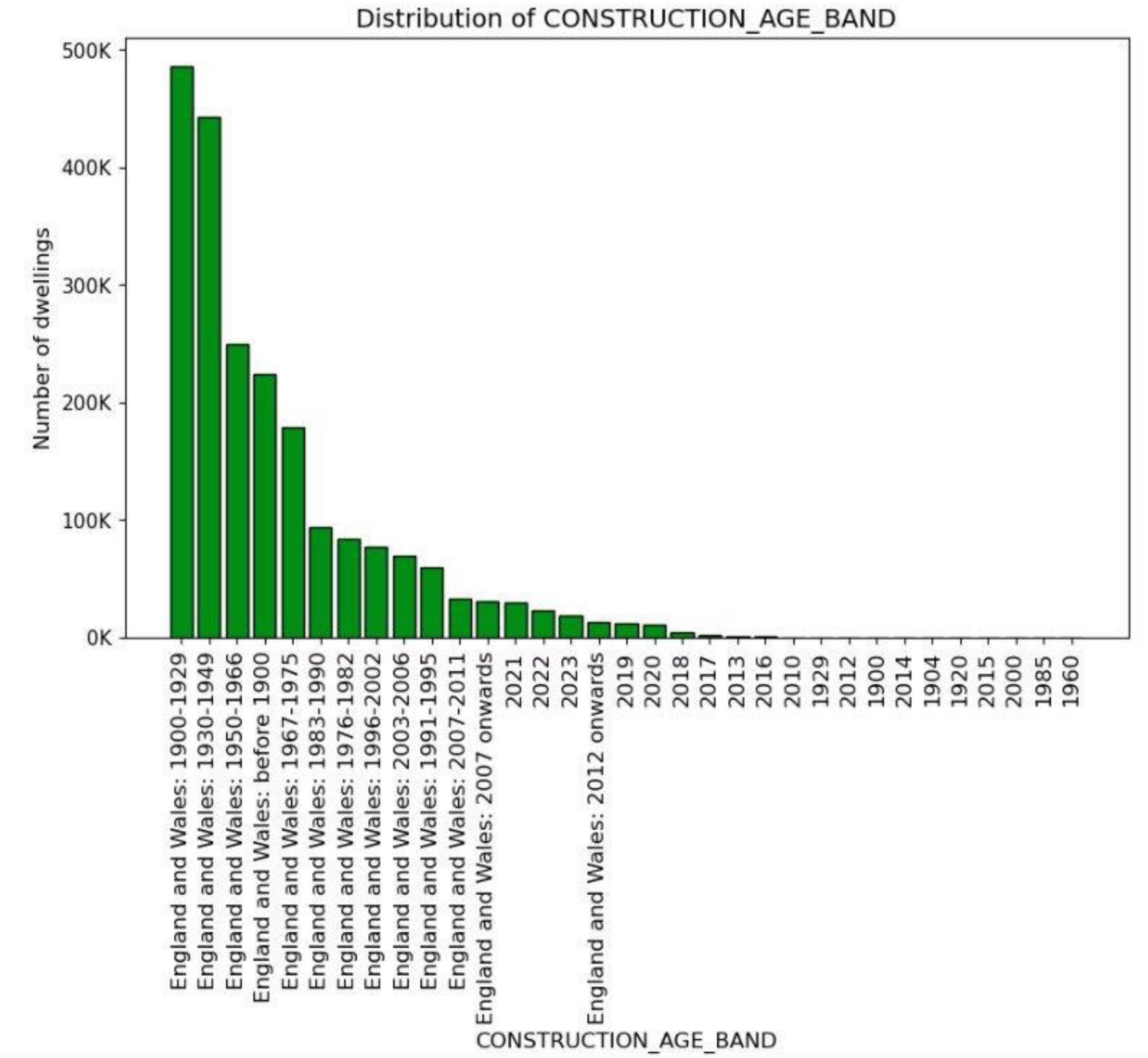


Figure 7: Distribution of Domestic Building Construction Age Bands in London from EPC