Predicting Residential Energy Use Intensity from Constrained Open Data Using Embedding-Enhanced Machine Learning and Open Weather Data

Dominik Jan Mecko

Student Number: 189356954

Supervisor: Prof. Melih Çelik

Word count: 14,830

Master of Science in Business Analytics

December 20, 2025

Submitted as part of the requirement for completing an MSc in Business Analytics at the University of Bath

Abstract

Residential energy use is a major contributor to global CO2 emissions, yet scalable methods for predicting building-level energy use in data-sparse settings remain limited. This dissertation addresses this gap by combining publicly available U.S. building performance data with open weather records and Population Dynamics Foundation Model embeddings to predict long-term residential energy use intensity. Five machine-learning models are evaluated within a unified pipeline that includes systematic feature selection to identify key drivers of consumption and to test the incremental value of weather variables and embeddings. The best-performing model produces well-calibrated, low-bias predictions and explains a substantial share of the variance in EUI, although building-level point predictions remain noisy due to limited feature richness and inherently variable consumption behaviour. PDFM embeddings yield marginal accuracy gains at the building level, but the results indicate potential for downstream area-level inference, particularly where direct energy data are unavailable. Overall, the study demonstrates both the promise and constraints of open-data-driven residential energy analytics and motivates future work on targeted data enrichment and alternative embedding-enabled prediction tasks.

Acknowledgements

I would like to sincerely thank my supervisor, Professor Melih Çelik, for his thoughtful, understanding, and consistently supportive approach throughout this dissertation.

I am particularly grateful to Ian Hally and Antonin Samal for introducing me professionally to the topic of energy prediction, for providing invaluable insights, and for giving me the opportunity to engage deeply with this area of research.

I would also like to thank my family, friends, and my ice hockey team for their support throughout the completion of this dissertation.

Finally, I would like to thank my dog, who passed away during the writing of this dissertation, for providing a welcome distraction.

I acknowledge that this work is solely my own, and I used ChatGPT 5.2 (Open AI, https://chat.openai.com/) only for proofreading my final draft.

List of Abbreviations

ABS Agent-based simulation.

ANN Artificial neural network.

API Application programming interface.

ASHRAE American Society of Heating, Refrigerating and Air-Conditioning Engineers.

BTU British thermal unit.

CBECS Commercial Buildings Energy Consumption Survey.

CNN Convolutional neural network.

CO_2 Carbon dioxide.

CO_2eq Carbon-dioxide equivalent.

DNN Deep neural network.

EDA Exploratory data analysis.

EIA U.S. Energy Information Administration.

EPA U.S. Environmental Protection Agency.

ERQ Empirical research question.

ESG Environmental, social, and governance.

EUI Energy use intensity.

GHG Greenhouse gas.

HVAC Heating, ventilation, and air conditioning.

kWh Kilowatt-hour.

Lasso Least Absolute Shrinkage and Selection Operator.

LIDAR Light Detection and Ranging.

LightGBM Light Gradient Boosting Machine.

ML Machine learning.

NOAA National Oceanic and Atmospheric Administration.

OLS Ordinary least squares.

PCA Principal component analysis.

PDFM Population Dynamics Foundation Model.

RBECS Residential Buildings Energy Consumption Survey.

ReLU Rectified Linear Unit.

RF Random forest.

RNN Recurrent neural network.

SVR Support vector regression.

THERMOS THERMOS Project.

TRQ Theoretical research question.

U.S. United States of America.

XGBoost eXtreme Gradient Boosting.

Contents

List of Figures

List of Tables

1 Introduction

As scientists debate whether global warming has already exceeded 1.5°C, the world is under increasing pressure to rapidly scale down carbon dioxide emissions (Bernard, 2025). One of the most significant contributors to pollution is the built environment. In 2019, the global building stock accounted for 12 gigatonnes of CO_2eq emissions (21%). Slightly more than three-quarters (81%) of these emissions are related to energy generation (Reinhart, 2018; Weber, Mueller, and Reinhart, 2021; Cabeza, Bai, et al., 2022). Reducing emissions from energy generation, alongside improvements in energy efficiency, is therefore essential for lowering buildings' CO_2 emissions globally. Although there is evidence of efficiency improvements (Cabeza, Bai, et al., 2022), growth in total floor area offsets these gains, placing continued pressure on further emissions reductions. However, precise estimation of building energy demand constitutes a complex challenge, as it combines physical, technical, behavioural, and cultural factors and is inherently costly. The ability to predict buildings' energy demand remotely and at scale would reduce costs and enable more efficient solutions for sector-wide decarbonisation. Recent advances in machine learning, data availability, and academic research show promising results, yet further investigation remains necessary.

1.1 Problem Description

Several literature reviews and empirical studies have examined both short- and long-term energy prediction (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Cai et al., 2019; Miller et al., 2020; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b; C. Lu, S. Li, and Z. Lu, 2022; Njimbouom et al., 2022). A substantial share of the literature (57%) focuses on short-term prediction, which is primarily used for consumption optimisation, whereas only 12% of studies investigate long-term prediction and its applications in benchmarking and policy analysis (Amasyali and El-Gohary, 2018; Miller et al., 2020).

The dominant research trend involves the use of feature-rich datasets (Deng, Fannon, and Eckelman, 2018); however, there remains a clear gap in pragmatic studies that aim to predict long-term residential energy consumption using diversity-sparse data. Most existing studies rely on detailed survey or benchmark datasets (Deng, Fannon, and Eckelman,

2018; Njimbouom et al., 2022), often comprising tens or hundreds of variables that are difficult to obtain from public sources and frequently include highly linearly dependent covariates, resulting in models that are methodologically robust but often impractical to deploy. Although these approaches are well established, a portion of the variance in energy consumption remains unexplained. Recent advances in geospatial computation, such as Google's Population Dynamics Foundation Model (Agarwal et al., 2025), demonstrate promising potential for modelling latent drivers of energy consumption at the building level.

At the time of writing, there was limited published evidence on the application of population dynamics foundation models to building energy prediction. Accordingly, this dissertation addresses the identified research gap by enriching diversity-depleted building data with open and experimental datasets and by adopting a novel approach to modelling energy consumption.

1.2 Relevance

The urgency of climate change and the need to decarbonise the built environment through reductions in energy consumption highlight the social relevance of this dissertation. The academic relevance of the study lies in addressing limitations in the existing literature, which is predominantly focused on short-term energy prediction and offers limited evidence on scalable long-term approaches that account for data diversity sparsity and latent spatial factors. Finally, the dissertation has clear managerial relevance. Long-term energy prediction plays an important role in energy benchmarking, policymaking, retrofitting, portfolio management, and underwriting, particularly in regulatory environments where non-compliance may result in financial penalties.

1.3 Research Questions

The main research question addresses the problem definition by testing a novel data-enrichment approach to study long-term residential energy-demand prediction.

To what extent can residential building energy demand in the United States of America be predicted using constrained open-source data available at scale?

To support the main research question, both theoretical and empirical sub-questions were defined:

- **TRQ1:** What methods are used for calculating or predicting residential energy consumption?

- **TRQ2:** Which factors influence the energy demand of a residential building?

- **TRQ3:** How much energy does a typical U.S. residential building consume?

- **TRQ4:** What data challenges are encountered in residential building energy prediction?

The empirical sub-questions are as follows:

- **ERQ1:** Which insights from exploratory data analysis can inform modelling decisions?

- **ERQ2:** Which machine learning algorithm performs best for predicting residential energy consumption?

- **ERQ3:** Do population density foundation model embeddings improve the prediction accuracy of the baseline model?

- **ERQ4:** Do outlier removal, feature engineering, subset selection, and hyper-parameter tuning improve model accuracy?

1.4 Contributions

The methodological contribution of this dissertation lies in the development of a scalable pipeline that integrates three heterogeneous datasets and provides an efficient algorithm to retrieve third-party data through an application programming interface (API).

The empirical contribution involves a conceptually driven investigation of latent factor modelling through experimental data enrichment. The aim of the empirical analysis is to generate generalisable insights into residential building energy prediction using geospatial embeddings.

Finally, the dissertation offers a practical contribution by assessing the effectiveness of the proposed methods for policymakers and analytics-focused businesses, including applications such as benchmarking, portfolio assessment, and underwriting.

1.5 Proposed Methodology

Theoretical sub-questions are addressed through a literature review. The empirical component consists of building a data-collection and preprocessing pipeline, conducting exploratory data analysis, and training multiple machine-learning models with systematic comparisons and performance optimisation. Models are evaluated under a consistent design using comparable metrics, residual diagnostics, and uncertainty quantification.

1.6 Ethical Considerations and Limitations

This study presents several potential ethical issues and limitations. From an ethical perspective, most of the data are sourced from open-access datasets, which cannot be fully verified for quality. The choice of datasets introduces selection bias and challenges the representativeness of the population. Embeddings data are computed from unpublished sources using black-box models and are not reproducible without access to Google's systems. To mitigate reproducibility concerns arising from dataset terms, a synthetic data file is provided. While two of the datasets are fully replicable with appropriate authentication and the provided code, the Population Density Foundation Model (PDFM) embeddings are accessible only by contacting Google.

1.7 Structure

The dissertation comprises six chapters: Introduction 1, Literature Review 2, Research Methodology 3, Analysis 4, Discussion 5, and Conclusion 6.

2 Literature Review

2.1 Conceptualisation of Terms

2.1.1 Energy Consumption

Energy consumption of a residential building is defined as the total annual energy used on site, resulting from the building's operation, characteristics, and climate conditions, and originating from any energy source. It is expressed in kilowatt-hours (kWh). Reinhart (2018) notes that this metric and its unit primarily describe electricity demand, whereas thermal demand is commonly expressed in British thermal units (BTU). Although Reinhart (see 2018, p.22) further argues that summing electricity and thermal demand met with gas or fuel constitutes a "mixing of apples and oranges", BTU can be converted to kWh, allowing thermal and electricity demand to be combined for benchmarking purposes. This approach is adopted in other energy prediction studies (Y. Chen et al., 2022b), and therefore the definition of energy demand in this dissertation is the total demand expressed in kWh. The terms *energy demand* and *energy consumption* are used interchangeably, both referring to the site energy consumed.

2.1.2 Energy Use Intensity

To normalise energy demand by building size, energy use intensity (EUI) is computed. It is calculated by dividing total annual energy consumption by the gross floor area[1] of the building, and is expressed in kWh/m^2/year. This normalisation allows for comparisons of building performance and is particularly useful for describing thermal and lighting energy use. Following the energy consumption definition, EUI refers to site EUI, meaning that any transfer losses are ignored (which would otherwise correspond to "source" EUI). EUI is the prediction target of this study because absolute energy consumption is linearly dependent on floor area, which can lead to overfitting and predictive issues due to large range of values. As a continuous variable, EUI defines the machine learning task as a supervised regression problem.

[1]Gross floor area refers to the total building area measured from external walls. A full definition is available in the Glossary E.5.

2.2 Approaches to Estimating Building Energy Consumption

Four main approaches are commonly adopted for energy estimation or prediction, with several possible target variables. The choice of approach and target variable depends on the end goal and the prediction horizon, which may be hourly, weekly, monthly, or annual. Although the majority of existing research focuses on short-term prediction methods, these are not the focus of this dissertation (Amasyali and El-Gohary, 2018; Miller et al., 2020; C. Lu, S. Li, and Z. Lu, 2022). In addition to absolute energy demand and energy use intensity (EUI), other target metrics include heating and cooling loads. Several studies and literature reviews have examined annual energy prediction and compared different modelling approaches, reporting strong predictive performance (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Deng, Fannon, and Eckelman, 2018; Bourdeau et al., 2019; Cai et al., 2019; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b; Njimbouom et al., 2022). Beyond temporal frequency, the literature (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Bourdeau et al., 2019; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b) commonly classifies methods into four categories: engineering methods, statistical models, data-driven approaches, and hybrid methods. The classification of statistical models varies across studies, as they are sometimes grouped with engineering or data-driven approaches and and in other cases defined as a standalone category. In this dissertation, statistical models are categorised as a form of data-driven approach. The following section addresses the first theoretical research question.

2.2.1 Traditional Methods

The first approach to energy prediction involves a physical on-site assessment conducted by an accredited engineer, who collects the necessary data and estimates annual energy demand using physical equations, models, and simulations. This approach, often referred to as a "white-box" method, benefits from individual adjustments, as the visiting engineer can observe nuances in materials, construction quality, and even patterns of occupant use. However, in practice, this method is costly and time-consuming when high levels of accuracy are required. In many cases, inaccuracy remains a significant challenge (Zhao and Magoulès, 2012; Y. Chen et al., 2022b). Scaling such assessments is difficult due to constraints related to the availability of qualified professionals and budget limitations, rendering this approach unsuitable for addressing the urgency of climate change at scale. Engineering methods also include physical simulation models that estimate energy demand using thermodynamic equations. When specified correctly and supported by valid inputs, these methods can be highly accurate, but they are computationally intensive and time-consuming to implement. In addition, several heuristic approaches and simple

mathematical formulations exist for estimating building energy demand and related characteristics. For example, Reinhart (2018) describes a formula that approximates heating or cooling demand by multiplying a constant by the building's floor area. In summary, traditional methods are valuable for the energy optimisation of highly inefficient individual assets, but they are not suitable for large-scale application unless substantially simplified and supported by manually collected data.

2.2.2 Data-Driven Methods

The second approach is the data-driven approach, which encompasses both statistical methods and machine learning (ML) techniques. Using large datasets containing information on buildings, their characteristics, occupant behaviour, and climate conditions, statistical and machine learning models apply sophisticated algorithms and probabilistic structures to predict outcomes based on a given set of input variables and historical observations. In essence, patterns within the data are learned and subsequently used for prediction. The current body of research has successfully applied a wide range of machine learning algorithms to long-term energy prediction (Deng, Fannon, and Eckelman, 2018; Cai et al., 2019; Setyantho and Chang, 2020; Njimbouom et al., 2022). This class of methods is often described as "black-box" approaches, as many of the underlying algorithms are difficult to interpret and are primarily evaluated based on predictive performance rather than model transparency.

The baseline model within the data-driven approach in the literature is ordinary least squares (OLS) linear regression (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Bourdeau et al., 2019; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b). In some reviews, linear regression models are categorised as statistical models within engineering methods, primarily due to their interpretability, ease of use, and intuitive structure compared to more complex ML algorithms (Fumo and Rafe Biswas, 2015). While these characteristics explain its widespread adoption in energy prediction, it appears to be driven by practical simplicity rather than by suitability for capturing the complexity of building energy use. Key limitations include a restricted ability to model non-linear relationships (typically only through polynomial terms), sensitivity to input variable selection, and multicollinearity among predictors.

As a result, linear regression is best positioned as a baseline model, offering good interpretation of linear terms but relying on assumptions that may be overly restrictive for complex energy systems (James et al., 2023; Kashnitsky, 2025b). Extensions such as lasso and ridge regression introduce regularisation to mitigate overfitting and reduce the influence of unstable predictors. Although not discussed explicitly in the energy prediction literature, the performance of ridge regression is well-established in the broader

ML literature (Setyantho and Chang, 2020; James et al., 2023) and is therefore included for comparison against the baseline linear model.

According to Amasyali and El-Gohary (2018), only 4% of the reviewed literature focuses on decision trees and random forests (RF), despite their widespread adoption across applied machine learning. This limited representation is notable given evidence that RF models outperform autoregressive and linear approaches in predicting energy demand (Y. Chen et al., 2022b). In addition, RF models demonstrate greater robustness to overfitting. Their ability to compute feature importance measures also partially addresses interpretability concerns commonly associated with black-box models (Y. Chen et al., 2022b). Key strengths of RF models include the ability to compute feature importance measures and partially address interpretability concerns (Y. Chen et al., 2022b). More broadly, RF models often outperform linear regression due to their capacity to capture non-linear relationships and automatically select relevant features (James et al., 2023). Nevertheless, in many ML applications, RF performance has been surpassed by boosting-based tree ensembles.

Tree-based boosting algorithms, such as LightGBM and XGBoost, were introduced into ML practice over the past decade and have rapidly gained popularity in competitive contexts (Y. Chen et al., 2022b). However, within the energy prediction literature, boosting algorithms have been examined only sporadically, with relatively few studies focusing on their performance (Setyantho and Chang, 2020; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b). Given their modelling assumptions, the limited exploration of tree-based methods shows a notable gap in the literaure. Existing research suggests that XGBoost and LightGBM are well suited for handling large tabular datasets and lead to accuracy improvements, which positions tree methods as ideal candidates for diversity-sparse, big datasets with non-linear structures. In contrast, Sun, Haghighat, and Fung (2020) argue that XGBoost is more computationally efficient but less accurate, interpretable, and robust than RFs. Nonetheless, Y. Chen et al. (see 2022b, p.2663) specifically note that XGBoost is "good at long-term prediction", suggesting a suitable candindate for the research objectives. Considering a relatively limited adoption in energy prediction research, there is valuable potential for comparative insights.

The next non-linear method, support vector regression (SVR), offers several advantages in energy prediction, including the ability to find global extrema, insensitivity to noise, and computational complexity that is independent of the feature space dimensionality (Sun, Haghighat, and Fung, 2020). However, the algorithm's objective, which involves mapping inputs into a high-dimensional space, renders it less suitable for the research objectives of this study compared to other methods. Although reviews indicate that SVR has been extensively applied in building energy prediction (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Sun, Haghighat, and Fung, 2020; Y. Chen et al.,

2022b) and achieves strong results in certain contexts, its predictive performance is generally inferior to tree-based methods. Consequently, SVR is omitted from the empirical analysis in this dissertation.

The final algorithm discussed is the artificial neural network (ANN), which is the most widely used method in the literature (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Bourdeau et al., 2019; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b). Inspired by biological neural networks, ANNs are capable of modelling highly complex relationships in large datasets by adjusting weights across individual neurons, multiple layers, and numerous training epochs. Several subtypes of neural networks exist, categorised according to their architecture. Deep neural networks (DNNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs) represent architectural adaptations of ANNs tailored to different prediction tasks. All neural network types have been successfully applied in energy prediction research; however, their primary applicability appears to be in short-term prediction contexts, where the algorithm can exploit complex and hidden patterns in the data (Zhao and Magoulès, 2012; Amasyali and El-Gohary, 2018; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b). For long-term prediction, ANNs may encounter challenges due to sparsity in feature diversity. Furthermore, the high computational costs and extensive data requirements can result in performance that is inferior to tree-based methods when applied to tabular datasets typical of residential building energy characteristics. Overall, ANNs have strong theoretical potential for non-linear modelling of energy demand through high-dimensional data such as weather features and embeddings, but their assumptions pose a challenge in addressing the research objectives. Therefore, the main inclusion of the algorithm is to assess performance against tree-based methods.

In the context of long-term energy prediction, tree-based ensemble methods generally achieve the best results on the studied error metrics, compared to decision trees, neural networks, and support vector regression (Y. Chen et al., 2022b). Contrary to practice, RF energy prediction models often demonstrate slightly better performance than boosting methods such as XGBoost. This may be due to data limitations, as RFs are less prone to overfitting and their architecture favours datasets with a smaller number of significant features, which is often the case in the long-term energy prediction. The literature provides substantial evidence that ML models can perform well in the research objectives with sufficient data quality and explanatory variables. The apparent discrepancy between the success of methods in ML competitions and in energy prediction research will be further examined in Analysis 4. However, in practice, limited data diversity poses a key challenge, contrasting with the controlled conditions often assumed from survey or benchmarking datasets.

2.2.3 Hybrid Methods

The final approach, referred to in the literature as the "grey-box" method, combines elements of both physical and data-driven approaches, using a simplified physical simulation informed by data. These models often achieve the best predictive performance; however, their application in the literature is limited and primarily focuses on short-term energy prediction (Y. Chen et al., 2022b). In addition, several novel ML approaches, such as physics-informed ML, have been developed. Kashinath et al. (2021) applied this approach to climate and weather modelling, suggesting potential for future application in building energy prediction research. Another emerging hybrid approach is agent-based simulation (ABS), which integrates operational and physical aspects to study emergent behaviour in energy systems using real data and a bottom-up methodology. J. Chen et al. (2022a) combined clustering of behavioural profiles with ABS to predict short-term energy consumption of U.S. residential buildings, while Ding et al. (2022) applied ABS to predict short-term consumption in commercial office assets. Both studies report improved accuracy attributable to occupant-specific modelling.

Table 2.1 summarises and compares the methods.

2.3 Factors Influencing Building Energy Consumption

Factors that drive energy consumption can be divided into internal and external categories. Internal factors originate from the building itself and its operational use, whereas external factors depend on the physical and social environment surrounding the building. Their interaction forms complex and inherently varying relationships across building types that may be difficult to measure, presenting challenges for prediction. The following section examines both categories in detail to provide a foundation for addressing the second theoretical research question.

The primary exogenous factor influencing energy consumption is the physical environment in which a building is situated. Humans are physiologically adapted to specific conditions, and consequently, climate and weather have a substantial impact on energy demand for heating and cooling. Other environmental variables such as precipitation, wind, humidity, and snowfall play an essential role in determining energy requirements. Microclimatic conditions can further alter energy demand in less obvious ways. These including shade, local air quality and currents, and urban heat islands may be influential, yet they are challenging to measure or obtain at scale without targeted studies (Kamal et al., 2021). Weather is therefore critical for predicting short- and mid-term energy consumption, whereas climate and microclimate exert influence over the long term. Good spatial and temporal resolution in the data determines the success of weather related

Table 2.1: Synthesis of Building Energy Prediction Methods

Category	**Representative methods**	**Strengths**	**Limitations**	**Best for**
Traditional (white-box)	On-site audits; physics-based simulations; heuristic formulas	High interpretability; high accuracy	Costly and time-consuming; scaling; sensitive accuracy	Asset-level; retrofit diagnostics
Data-driven (black-box): Linear / Statistical	OLS; Lasso, Ridge regressions	Interpretable; computationally efficient	Non-linearities modelling; multicollinearity	Benchmarking; explaining linear drivers
Data-driven (black-box): Tree Ensembles	RF; XGBoost, LightGBM	Non-linearities and interactions; accuracy; scalable	Overfitting; interpretability	Long-term prediction; large tabular datasets
Data-driven (black-box): Kernel and Related Methods	SVR	Robust to noise; finds global optimum	Outperformed by tree methods; tuning is non-trivial	Medium-sized datasets with complex but smooth relationships
Data-driven: Neural Networks	ANN; DNN; CNN; RNN	Complex patterns modelling; short-term forecasting	Low interpretability; data and computational requirements	Short-term forecasting; large datasets complex signals
Hybrid (grey-box)	Physics-informed ML; ABS	Integrates frameworks; strong short-term accuracy; occupant dynamics	Limited evidence for long-term performance; implementation; data requirements	Short-term forecasting or scenario analysis

explained modelling variance. These findings signify the importance of including annual weather patterns in longitudinal datasets, as variations in key variables can substantially affect a given year's energy demand (Rafsanjani, 2016; Sun, Haghighat, and Fung, 2020; Y. Chen et al., 2022b).

The most apparent internal factor influencing a building's energy consumption is the building itself. Construction materials and their physical properties, the shape and size of the building, its positioning within the environment, and ventilation all play a major role in determining energy efficiency and overall energy demand (Rafsanjani, 2016). Many of these physical characteristics can be measured directly or indirectly at scale using satellite imagery, on-street photography, or surveying techniques, for example through LIDAR as in THERMOS Project (n.d.). However, other features, such as window types and positioning, insulation materials, or microclimatic conditions within the building, are difficult to obtain without detailed analysis or on-site visits, yet they can have a substantial effect on energy consumption (Y. Li et al., 2021). The limited accessibility of some internal features highlights a challenge for large-scale energy modelling, as omitting or approximating them may reduce model accuracy.

The type of use of a building constitutes another endogenous factor affecting its energy demand. Whether a building is commercial or residential, and the types and quantities of amenities it contains, directly influence energy consumption (Setyantho and Chang, 2020). For example, a data centre is likely to consume dramatically more energy than a single-family house. As this study focuses on residential buildings, the primary categories considered are single- and multi-family houses. Further subdivisions depend on the context in which buildings are classified. From an energy efficiency perspective, it is most logical to consider groupings of buildings, as shared walls can improve heat retention and reduce energy losses (Y. Li et al., 2021). Single-family houses are usually categorised as detached, semi-detached, or terraced. Multi-family buildings are typically divided according to the number of units, ranging from small multi-unit buildings to high-rise apartments. Although absolute energy consumption increases with the number of units, energy efficiency per floor area can improve in larger buildings because the heat loss area relative to volume is reduced, considering material choice (Zhigulina and Ponomarenko, 2018). It is also important to note that energy consumption per unit of floor area initially rises with occupancy, but after a certain threshold, the dilution effect can increase overall efficiency (Gao et al., 2019). Building type is commonly available in most datasets and as a categorical variable can proxy lots of latent effects, but cause some granularity loss due to aggregation and grouping.

Another important internal factor is the building's equipment, which can be divided into two categories: amenities in common spaces and privately owned equipment. Both categories primarily affect energy consumption through electricity use (Rafsanjani, 2016),

with heating and cooling systems typically accounting for the largest shares. The type of fuel, system efficiency, and whether the system is private or shared all influence energy consumption. In addition, shared amenities such as swimming pools, saunas, appliances, and lighting contribute additional energy demand. Private appliances, including cooking equipment, also directly impact the building's energy use. However, data on these features are difficult to obtain at scale, except for general indicators such as heating fuel type (Setyantho and Chang, 2020). Cultural and societal factors further affect energy consumption patterns; with anectodal evidence on different preferences for shared appliances (**Wikipedia2025**). In the U.S., household penetration of home appliances is high, and typical appliance usage is expected to be relatively uniform across residences (Cabeza, Űrge-Vorsatz, et al., 2018). Consequently, these factors are often omitted from large-scale modelling efforts, both due to the practical difficulties of data collection and because high appliance penetration suggests they may contribute only minor variations to overall energy consumption.

Related to building equipment, occupant behaviour constitutes the final significant factor influencing energy consumption. The manner in which residents use appliances and the amount of time they spend at home directly determine energy demand. In practice, modelling individual behaviour is challenging, although some studies have attempted to characterise occupant energy use patterns or identify the most relevant behavioural factors (Rafsanjani, 2016; Setyantho and Chang, 2020; J. Chen et al., 2022a). Key determinants include socio-economic indicators such as income, education, employment type, and occupant age, as well as room temperature preferences, household size, and energy-saving habits. While behavioural theories explain certain consumption patterns, these factors are costly to collect and nearly impossible to link to specific buildings in public datasets. Consequently, they are not comprehensively reviewed in this dissertation. The PDFM (Agarwal et al., 2025) contains analogous information in an aggregated form at the postcode level, embedded within latent features, and thus offers potential for modelling behavioural influences indirectly. However, Huebner (2015) found that building characteristics and climate conditions account for most of the variation in energy consumption, and after controlling for multicollinearity, only household size remained statistically significant. Furthermore, Huebner (2015) argued that more than half of the variability remains unexplained, even when increasing the number of predictors. This suggests a critical limitation of traditional data collection methods, highlighting that capturing occupant behaviour accurately is essential but practically difficult for effective energy prediction.

The identified factors are summarised in Table 2.2.

Table 2.2: Synthesis of Factors Influencing Residential Building Energy Consumption

Factor group	**Example factors**	**Expected impact**	**Observability**
External: climate and weather	Temperature; precipitation; wind; climate patterns	Heating and cooling; weather affects short- and medium-term, climate acts long-term	Measurable and widely available
External: microclimate and local environment	Shading; air quality; air flows; heat islands	Energy requirements relative to climate	Difficult to measure and obtain at scale
Internal: building characteristics	Construction materials; size and shape; ventilation; insulation	Determines efficiency and demand through heat loss	Some attributes available / proxied at scale
Internal: building use and typology	Residential/commercial; single/multi-family	Medium impact	Mostly available
Internal: equipment and systems	Heating and cooling; fuel; systems; amenities	Detailed driver of energy use	Fuel type often available, rest usually sparse
Internal: occupancy patterns and socio-economic factors	Appliance-use; income, education, household size	Significant at the individual level;	Costly to obtain and challenging to link

2.4 Energy Consumption of a Typical Residential Building in the U.S.

The U.S., thanks to its size, encompasses heterogeneous climates and socio-economic conditions, meaning that the median or average energy consumption of a representative building varies by region. Nevertheless, defining a typical residential U.S. building is useful for exploratory data analysis (EDA) and outlier detection and answers the third theoretical sub-question.

Typical buildings fall into two incomparable categories: single-family and multi-family houses. In the U.S., single-family homes are significantly more prevalent than multi-family dwellings (Potter, 2020; Statista, 2025). A representative single-family home is a detached house, typically constructed of wood, built in the second half of the twentieth century, with 1–2 stories and approximately 200 m^2 of floor area. Common multi-family houses are low- to mid-rise buildings (up to five stories) with smaller units than single-family homes. Larger multi-family buildings are typically constructed from concrete or masonry, whereas smaller units often have wooden frames, particularly for buildings erected around the early 2000s (Potter, 2020). The equipment characteristics of both building types were described in the previous section.

A typical household consumes approximately 11,000 kWh per year. For single-family homes, consumption is slightly higher (Potter, 2022; U.S. Energy Information Administration (EIA), 2023), translating to around 200 kWh/m^2/year, depending on the building and environmental factors. The typical range is 150–250 kWh/m^2/year (U.S. Environmental Protection Agency (EPA), 2025). Multi-family buildings tend toward the lower end of this range due to the non-linear relationship between floor area and energy use. Most energy is consumed for heating and cooling, followed by refrigeration, lighting, and other household uses (Reyna et al., 2022).

2.5 Data Challenges in Building Energy Prediction

As discussed in the previous sections, only 12% of studies examined annual energy consumption, potentially reflecting challenges related to data availability and quality (Amasyali and El-Gohary, 2018). The literature identifies several key issues in long-term energy prediction, including the requirement for extensive temporal data, the inherent non-linearity of building energy use, uncertainties in both input and outcome variables, and limitations in data availability and quality. The following section provides an answer to the fourth theoretical sub-question (Morewood, 2023; Qiao, 2023).

Amasyali and El-Gohary (2018) argue that good predictive performance depends on larger amounts of time series data for annual energy consumption compared to shorter

periods. However, their argument assumes that annual predictions are derived from higher-resolution data, such as daily or weekly frequencies, which may not reflect realistic scenarios for true annual prediction due to data size. The need for rich data is valid, as extending the dataset from single-year to multi-year time series enables the model to capture year-specific variations in weather conditions or building renovations. This issue is closely related to uncertainty in energy prediction. The aggregation process inherent in annual predictions can erase important variance on both the supply and demand sides of energy use, leading to smoothening and failures in proxy variables. Moreover, certain influential events occur only over the long term, necessitating specific modelling approaches. These findings underscore the nuance, complexity, and fragmented nature of energy prediction research. The limitations are critical for asset managers, but less so for policymakers. Non-linearity further complicates long-term prediction compared to short-term analysis (Amasyali and El-Gohary, 2018). Consequently, traditional approaches such as linear regression are largely inapplicable, and even more sophisticated models may yield inaccurate estimates when faced with non-linear relationships.

Data availability and quality constitute the final discussed set of challenges in long-term energy prediction. Qiao (2023) report that many datasets and individual buildings contain missing features or values. A substantial portion of research achieving low prediction errors (Deng, Fannon, and Eckelman, 2018; Setyantho and Chang, 2020) relies on comprehensive surveys and benchmark datasets, such as the Residential Building Energy Consumption Survey (RBECS), which provide complete and detailed feature spaces. The diversity in these datasets allows for precise relationship modelling, rarely achievable in practice. The deployed prediction models often depend on public datasets or user-provided inputs, which may be less detailed than benchmarking surveys. Legal, organisational, or situational constraints may prevent the collection of comprehensive data, resulting in incomplete and limited feature sets. Even when richer datasets can be obtained, issues such as sensor malfunctions, outliers, and suboptimal data processing often reduce data quality (Morewood, 2023; Qiao, 2023). As noted previously, highly influential factors such as occupant behaviour and building equipment are practically unobtainable at scale, yet their inclusion could substantially improve predictive accuracy (Huebner, 2015). The combined diversity of these factors, along with variations in building types and usage, creates a particularly challenging environment for achieving low-error long-term energy predictions (Bourdeau et al., 2019).

2.6 Emerging Technologies: Feature Embeddings

The discussion in the previous section regarding data challenges highlights the need for latent variable modelling. Recent computational developments have led to breakthroughs in foundational models capable of compressing high-dimensional datasets into much smaller

matrices while preserving embedded relationships. These matrices encode dense information in fewer dimensions and are particularly useful in data-sparse contexts, as they can capture complex patterns learned from data-rich environments.

One such model is the Population Density Foundational Model (PDFM), developed by Google (Agarwal et al., 2025). The PDFM employs a graph neural network (GNN) spanning the entire U.S., with counties and postcodes as nodes, and incorporates a substantial amount of data collected from diverse sources, including human-centric, environmental, and local characteristics. The GNN produces feature embeddings, which are dimension-reduced matrices containing dense and complex relationships extracted from the original features. These embeddings can be linked to geo-spatial identifiers, such as postcode or county codes, and used for various downstream tasks. Agarwal et al. (2025) identify four primary applications: interpolation, extrapolation, forecasting, and super-resolution.

In the context of long-term energy prediction, several applications are feasible. Embeddings can be used to extract postcode-level aggregated preferences that differ across space and incorporate them into downstream models to account for latent variables affecting energy use at the individual building level. They may also be used for geo-spatial prediction in underrepresented regions or to infer input variables from lower-resolution data, such as county-level income. Such approaches may help overcome population representativeness issues and enable predictions for all U.S. postcodes. However, the spatial mismatch between aggregated embeddings and building-level targets may introduce errors or oversimplify local variability, which forms one of the empirical objectives of this study.

Agarwal et al. (2025) report performance improvements of the PDFM over traditional methods, although evidence of its practical application in energy prediction remains limited. Adopting this novel approach therefore offers a new perspective on long-term energy prediction, extending beyond its current uses in large language models, geo-spatial inference (Tucker, 2024), house price prediction (Das et al., 2022), and other context-specific inference tasks (Rodriguez, Spirling, and Stewart, 2023). Nonetheless, the effectiveness of such embeddings depends on the quality and representativeness of the underlying data, and their aggregation may obscure important local heterogeneity.

The density of information across numerous domains motivates the hypothesis that enriching a diversity-limited dataset with embeddings may allow control for spatial differences and relationships that affect energy consumption but are otherwise difficult to observe. These latent factors may include environmental variables, behavioural patterns, micro-climate conditions, policies, or socio-economic characteristics.

However, the hidden nature of the embeddings raises significant ethical and practi-

cal issues. Low interpretability limits the ability to derive actionable policies or justify decisions based on the model outputs. Moreover, Google retains exclusive access to the underlying data and computational methodology, restricting transparency and reproducibility. While these issues are beyond the scope of this study, they remain important considerations for future research if embedding-based approaches in energy prediction prove successful.

2.7 Conceptual Research Model

The literature review focused on reviewing methods for energy prediction, identified factors affecting energy consumption, described typical U.S. homes, explained data challenges in energy prediction and its research and justified the potential for prediction improvements from experimental data. Based on the literature review, a synthesis Table 2.3 and a conceptual research model were constructed (Figure 2.1).

Table 2.3: Synthesis of the Literature Review

Topic	**Core claims**	**Implications for modelling**	**Limitations / gaps**
ML Methods	Linear models and non-linear models (RF, boosting, ANN/DNN)	Linear models as baselines; tree ensembles expected to be the best	Mixed evidence on RF vs boosting; data and overfitting sensitivity
Energy Consumption factors	External / internal	EDA, Feature engineering; feature importance, feature selection and interpretation	Some features hard to obtain; inherent variance
Typical U.S. home	Typical EUI $\approx$ 150–250 kWh/m^2	EDA, outlier handling	Values vary by climate and socio-economic context
Data challenges	Long-term work limited; missingness, nonlinearity, uncertainty, temporal richness	Imputation, feature selection, method selection, implications	Gap between research-grade and real-world sparse data
Embeddings	PDFM embeddings might help model latent variables	Benchmark performance improvements	Interpretability and reproducibility concerns; restricted data access

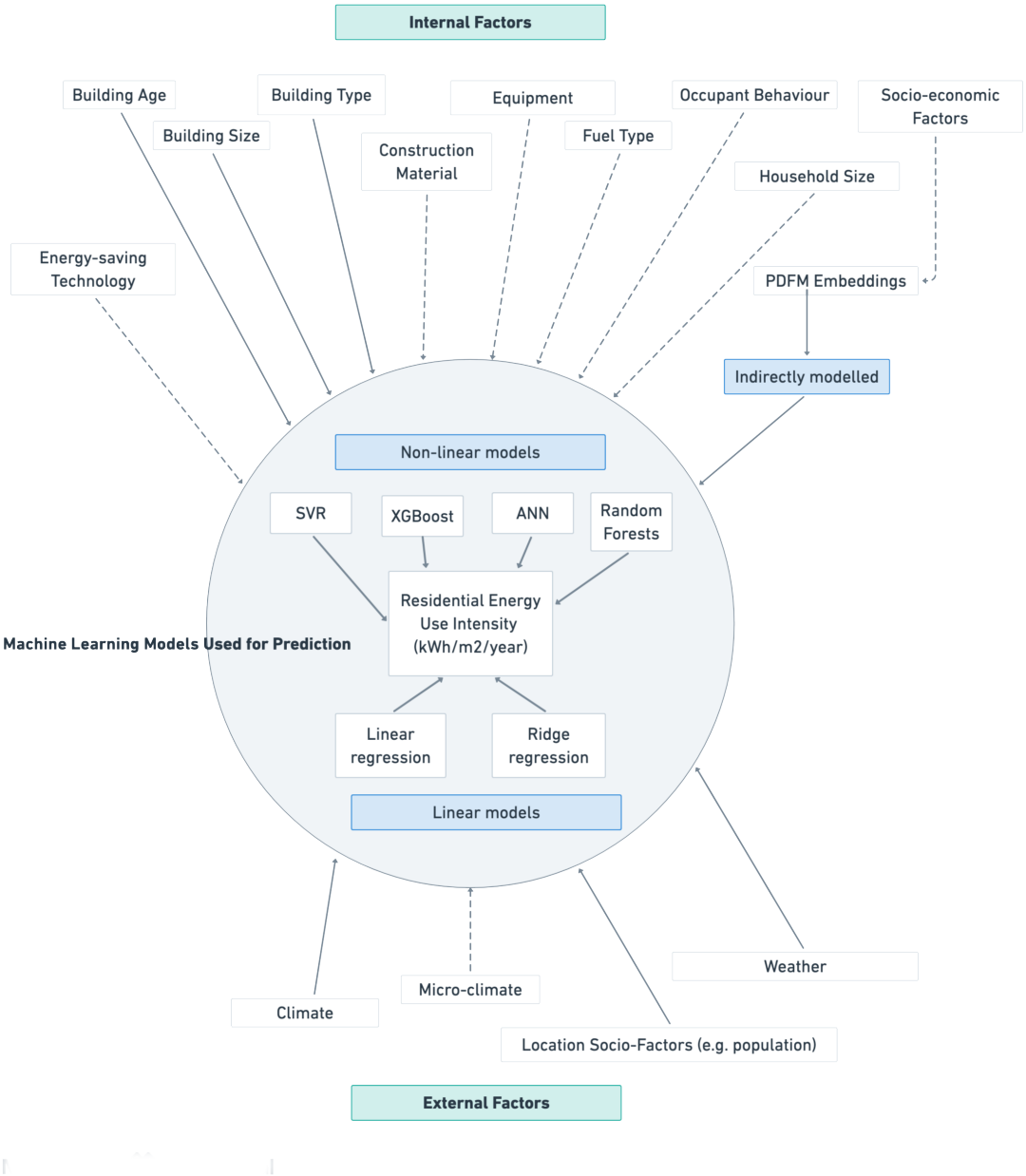


Figure 2.1: Conceptual Research Model

3 Research Methodology

The following chapter discusses the data collection and feature engineering, ML methods employed, performance evaluation and bias prevention.

3.1 Data

Secondary data from several sources were collected and processed in this study. The primary dataset was obtained from the Building Performance Database (BPD). Although the source is reputable, the data were collected by local governmental bodies in selected U.S. counties, which inherently introduces selection bias. The heterogeneous nature of the data collection process also results in variation in data quality. While the features are standardised, the temporal structure varies and is inconsistent across the underlying datasets; however, this limitation does not pose a threat to internal validity. Data quality and descriptive statistics are discussed in greater detail in the exploratory data analysis (EDA) section of the next chapter (Section 4.1). The BPD underlying datasets were collected and processed on 10 June 2025 as CSV files downloaded from the official website. The final dataset contained anonymised information on 60,855 buildings observed over multiple years, resulting in 325,611 records and 32 columns. The variables in this dataset served as baseline fixed-effects features. Almost a half was unusable, resulting in a diversity-limited, constrained building characteristic feature set. The metadata and variable descriptions are presented in Appendix D.3.

The first enhancement of the dataset consisted of adding weather variables for each building location in the corresponding year of observation. These data were collected from the National Oceanic and Atmospheric Administration (NOAA), a U.S. government institution, via its application programming interface (API) on 4 October 2025. The dataset included 88 monthly weather variables recorded at the relevant stations, primarily temperature, precipitation, snowfall, heating degree days (HDD), cooling degree days (CDD), wind speed, atmospheric pressure, and humidity. These variables with the embeddings accounted for the high-dimensionality of the feature space. The source dataset was the Global Summary of the Month (GSOM). Due to the dataset size and API rate limits, a heuristic approach was required to constrain the computational complexity

of the data collection process. The heuristic aimed to minimise the number of API requests by identifying the highest unique spatial resolution available in the GSOM for each building in the BPD in a given year. This approach reduced the number of required API requests from 60,855 to 576 while ensuring that each building in the BPD was matched with the highest-quality weather data without duplicating location-year queries. To further reduce computational overhead, three API keys were generated, and data requests were executed using a simple loop across three parallel Python environments. The full implementation is described in Appendix B.

The second data enhancement consisted of incorporating PDFM embeddings provided by Google. The dataset was requested for research purposes via a formal application on 9 May 2025. Although the provider is reputable, the data are not publicly available, which introduces replicability concerns for this study. The embeddings were matched to individual building records using postcode identifiers.

3.2 Exploratory Data Analysis

The exploratory analysis pursued three main objectives. The first objective was to understand the structure of the dataset through descriptive statistics, the identification of invariant features, the visualisation of distributions, and the removal of outliers. The second objective was to examine relationships between variables using scatter plots and correlation analysis. The third objective was to investigate potential feature engineering approaches, including transformations, interactions, and dimensionality reduction. Principal Component Analysis (PCA) was employed to reduce the number of features. The exploratory analysis did not follow a rigid structure. Instead, the methodology focused on addressing the primary objectives through a sequence of questions guided by preliminary results. The analytical framework was adopted from Kashnitsky (2025a) and Polusmak (2025). The complete analysis notebook, including commentary and exploratory questions, is available in the research repository.

3.3 Machine Learning Algorithms

The predictive component of the analysis was conducted using a selection of machine ML algorithms identified and discussed in the literature review. The models were implemented using the `scikit-learn`, and `tensorflow` ML libraries. All models shared an identical data pre-processing pipeline. Default hyperparameters were used where available to establish baseline performance, followed by parameter tuning of the best-performing model.

The baseline models used for evaluation were OLS linear regression and its L2-

penalised variant, ridge regression. These models served as performance benchmarks alongside the best-performing model. Both methods minimise a least-squares objective, with ridge regression introducing a penalty term with $\lambda = 1$ as a baseline for tuning in the optimisation formula to mitigate the influence of noise-inducing variables.

The second model was a RF regressor, implemented with default parameters from the `sklearn` package to establish a baseline for comparison, with only the random state specified to ensure reproducibility. The model was selected to capture non-linear relationships in the data and was expected to perform similarly to XGBoost, albeit with higher computational cost.

Although the literature reports the use of SVR, it was excluded from this study due to the large dataset size and evidence of inferior performance in comparable studies, to prioritise algorithms with more favourable scalability and interpretability trade-offs.

The next model was XGBoost, implemented with default parameters from the `xgboost` package to establish baseline performance. Based on the literature, the model was expected to achieve the strongest predictive performance and the highest computational efficiency among the evaluated algorithms. As XGBoost outperformed the other models, it was subsequently tuned and used for subset selection. Although feature importance scores were computed using gain, the model largely operates as a black box and therefore offers limited interpretability.

The final model constructed was a sequential ANN comprising an initial activation layer with 512 neurons and a rectified linear unit (ReLU) activation function, followed by two hidden layers with the same activation function containing 256 and 128 neurons, respectively. The output layer consisted of a single neuron providing the predicted value. Regularisation was applied via two dropout layers with a rate of 0.2. The model was compiled using the Adam optimiser and the mean squared error (MSE) loss function.

3.4 Data Pre-processing, Outlier Removal and Feature Engineering

Prior to model training, outliers were removed, engineered features were added, and the data were pre-processed to conform to model-compatible structures (Zhang et al., 2021; Kravchenko, 2025; Radchenko and Kashnitsky, 2025). All variables reported in U.S. units were converted to the metric system, namely British thermal units to watt-hours and square feet to square metres. Building type and ASHRAE climate codes were dummy-coded, with 1 indicating presence and 0 absence. All data types were set to `float32`. Detailed pre-processing procedures are provided in Appendix D.4.1.

Outlier treatment was applied solely to the prediction target, $siteEUI$, to account

for the high dimensionality of the feature space and to minimise data loss in the remaining features. As discussed in the Exploratory Data Analysis and Literature Review chapters, buildings with extremely high or low energy usage are atypical in the residential sector. For some buildings, such values may reflect measurement, unit or data entry errors, while for others, such consumption is consistently recorded. These extreme values are sparse in reality and were therefore removed. Low outliers were defined using an absolute threshold of 15 kWh/m^2/year, to account for unstability of the mean absolute percentage error (MAPE) near zero values. High outliers were identified using the following formula:

$$\text{site\_eui\_kWh/m}^2 \leq Q_3 + 1.5\,\text{IQR},$$

where

$$\text{IQR} = Q_3 - Q_1$$

and Q_1 and Q_3 are the first and third quartiles of site_eui_kWh/m^2.

The engineered features consisted of four groups: ratio, transformed, interaction, and dummy variables. The ratio variable was defined as electric EUI divided by fuel EUI. The interaction variable was the product of *building age* and $floor\ area$. The transformed variables included the square of *building age*, the square of the *Energy Star rating*, and the base-10 logarithm of $floor\ area$.[1] The dummy variables comprised climate categories and facility types. An overview of the engineered features is presented in Table D.2.

Given the skewness in the distributions of many variables and the robustness of median statistics to outliers, missing values were handled using median imputation. The dataset, excluding the embeddings, was standardised using Z-score scaling (see Appendix D.4.2) to ensure that features on differing scales contributed equally to the prediction and to mitigate the influence of skewness. Features with more than 55% missing values, or missing values in critical features such as the prediction target or embeddings, were filtered, along with invariant features. The models were contrained by omitting fuel and electric EUI as predictors. It was deemed inappropriate, as these variables are linearly dependent on *site EUI* and already contain almost all of its variation from other variables. Their inclusion is effectively trivial and introduces potential noise from other features. Together with the outlier removal, these steps resulted in the exclusion of approximately 2,000 buildings and 210 variables from the dataset.

[1]Energy Star rating refers to an energy efficiency score defined by U.S. Environmental Protection Agency (EPA) (2025) as a metric on a scale from 0 to 100, designed to compare relative energy efficiency performance among buildings of similar use, and normalised for weather and operational factors. Its main purpose is comparison, and participation in the programme is voluntary.

3.5 Training, Performance Evaluation and Tuning

The temporal structure of the dataset required the adoption of strategies to mitigate data leakage. The dataset contained multiple annual observations per building, which were inconsistent, with both the range and year of observation varying across buildings. To address data leakage, a custom function was implemented to randomly split the sample into training (75%) and test (25%) sets based on the building identifier, ensuring that each building appeared exclusively in one set (James et al., 2023). The same function was used in the cross-validation. Potential leakage from weather variables within the same region was considered, but was judged non-threatening to internal validity, as the objective was to model EUI conditional on weather, with no fixed-effects leakage and no generalisation to future years. Alternative datasets, generated by averaging quantitative features, were produced to enable experimental testing of aggregate models and performance comparison. These models included building-level and postcode-level aggregations.

To evaluate the performance of the ML algorithms, four standard metrics were selected based on the literature review and best practices (James et al., 2023; Sergeyev, 2025). As the target variable is continuous, mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) were selected. Since the target cannot reach zero, mean absolte percentage error (MAPE) was also considered a useful metric for recommendations and assessing practical applicability, in particular for businesses and policymakers. Explained variance (R^2) was employed to compare the effects of different feature set additions. Model diagnostics, including underfitting, overfitting, and unexplained variance, were assessed through analysis of residuals and observed-versus-predicted value plots. Uncertainty was measured with errors by decile, calibration curve and prediction intervals.

The real-world performance was estimated through cross-validation (CV) of the dataset splits using different random seeds. Five simple k-folds were employed, with the performance metrics averaged across folds. The CV procedure utilised the same custom train/test split function. Although this increases the computational time, it ensures that no data leakage occurs. For hyperparameter tuning of the best-performing algorithm, Bayesian CV – a probabilistic search algorithm for optimal parameters – was employed due to the size and complexity of the dataset. Other standard methods, such as grid search or random search, are inefficient or suboptimal in this context. The search space for the best model, XGBoost, is summarised in Table 3.1.

To assess the performance of the subsets, the same performance metrics were applied. In addition, feature importance by gain was calculated and and top 20 features were analysed (Radchenko and Kashnitsky, 2025). The performance of the best and tuned model was also evalued through residuals analysis – a scatterplot of predicted and true

Table 3.1: Search Space of XGBoost Hyperparameters

Parameter	Range	Meaning
max_depth	$2 \rightarrow 12$	Controls how deep each tree can grow; deeper trees model more complex patterns.
learning_rate	$0.001 \rightarrow 0.3$ (log-uniform)	Step size for boosting updates; smaller values slow learning but reduce overfitting.
n_estimators	$100 \rightarrow 2000$	Number of boosting rounds (trees); more trees improve fit but increase training time.
subsample	$0.5 \rightarrow 1.0$	Fraction of training samples used per boosting round; lower values add regularization.
colsample_bytree	$0.5 \rightarrow 1.0$	Fraction of features used per tree; sampling fewer features helps prevent overfitting.

observations, predicted values and residuals and residuals distribution.

To assess the performance of the feature subsets, the same performance metrics were applied. In addition, feature importance based on gain was calculated, and the top 20 features were analysed (Radchenko and Kashnitsky, 2025).

3.6 Research Design

Finally, to answer the research questions, the performance was compared using the following design:

1. **Identification of the best-performing algorithm:** All models were trained and evaluated on the same random splits using five-fold CV and all available features, excluding engineered and PCA-produced ones. The best performing model was compared to alternative models – one-building entry and averaged features model, and a model using embeddings to predict post-code level average energy.

2. **Identification of performance improvements**: The best model was optimised through hyper-parameter tuning.

3. **Identification of performance improvements through feature addition:** Comparisons of explained variance, performance and feature importance across feature subsets. The comparison included the following feature subsets: baseline, embeddings, weather, baselineembeddings, baselineweather, baselineweatherembeddings, all features without engineered features, baseline + feature engi-

neered, baseline without climate dummies and energy features, and all features

The full design of the research process is described in Figure 3.1.

3.7 Bias Prevention

The data leakage and temporal inconsistencies were previously discussed, along with proposed mitigation strategies.

The second problematic aspect of the dataset is its representativeness. The sample was found to be non-representative due to selection bias. A detailed analysis of data representativeness is presented in the EDA 4.1. This poses a threat to the external validity of the model, which can only generalise to buildings similar to those in the BPD dataset. However, Walter and Mathew (2019) indicate that the target remains reasonably representative. Given the sufficient dataset size and the non-causal research objectives of this study, sampling biases were not corrected and are considered a research limitation. Measurement bias was addressed through the removal of outliers and handling of missing values.

Lastly, the PDFM embeddings suffer from limited availability: neither the model nor the dataset are publicly accessible and can only be obtained by contacting Google. These reproducibility and transparency issues were discussed in the literature review and are acknowledged as a limitation of this study. Nevertheless, the research objectives focus on producing generalisable insights from experimental evaluation, and while this concern reduces reproducibility, it does not invalidate the research findings.

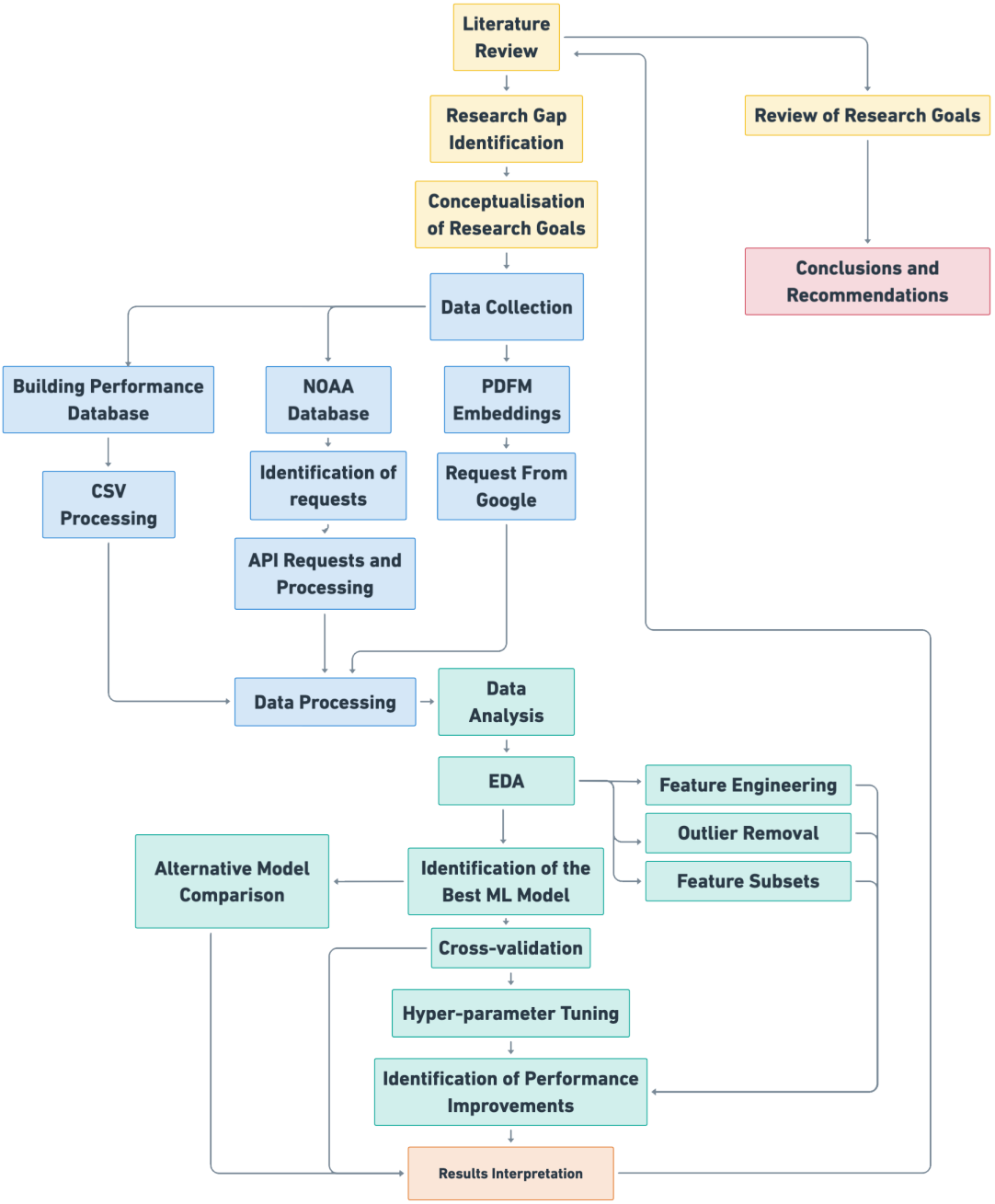


Figure 3.1: Research Design.

4 Analysis

This chapter presents the results of the analysis. First, the outcomes of the EDA are summarised, and the main insights from the exploratory analysis are discussed. The second part of the chapter focuses on the modelling results, including the identification of the best-performing model, its comparison with alternative models, hyperparameter tuning, the evaluation of different feature subsets and model diagnostics with uncertainty quantification.

4.1 Exploratory Data Analysis

4.1.1 Structure Analysis

The initial steps of the EDA aimed to understand the structure of the dataset, identify uninformative columns, describe baseline and weather variables (see Appendix D.3), and assess sample representativeness. The exploratory analysis was conducted on data prior to outlier removal, unlike the dataset used for modelling. It contained 243,815 entries corresponding to 50,302 buildings. After outlier removal, the final dataset comprised 204,639 entries (48,038 buildings) and 910 variables. Descriptive statistics (Table 4.1) of the sample indicate a broadly typical buildings sample with a few extreme points and skewed distributions, which are later discussed.

Table 4.1: Descriptive Statistics of the Sample

Variable	Count	Mean	Std	Min	Median	Max
site_eui_kwh_m2	243,815	185.31	118.75	3.15	155.16	3,134.99
year_built	242,775	1,967.62	31.72	1,649	1,972	2,022
floor_area_m2	243,815	8,606.99	13,965.31	47.01	47,53.57	185,617.21
energy_star_rating	148,091	57.99	33.30	0	65	100
electric_eui_kwh_m2	222,233	67.52	44.29	0	59.04	2,456.48
fuel_eui_kwh_m2	214,860	120.73	111.12	0	81.09	2,953.89
ghg_emissions_m2	227,767	48.69	26.27	0.63	45.73	886.28
population	243,310	42,986.02	23,865.06	28	34,193	116,469

Naturally, the descriptive statistics raised the question of whether the sample was

representative. Upon reflection on the literature review, the answer appears to be negative, due to the building type distribution differing from expected patterns. Single-family buildings, which are typically predominant in the U.S., were underrepresented relative to multi-family buildings. Another factor limiting the external validity of the dataset is building location, as the majority of entries are concentrated in Florida, New York, and California, which is not representative of the entire U.S. building stock (Figure 4.1).

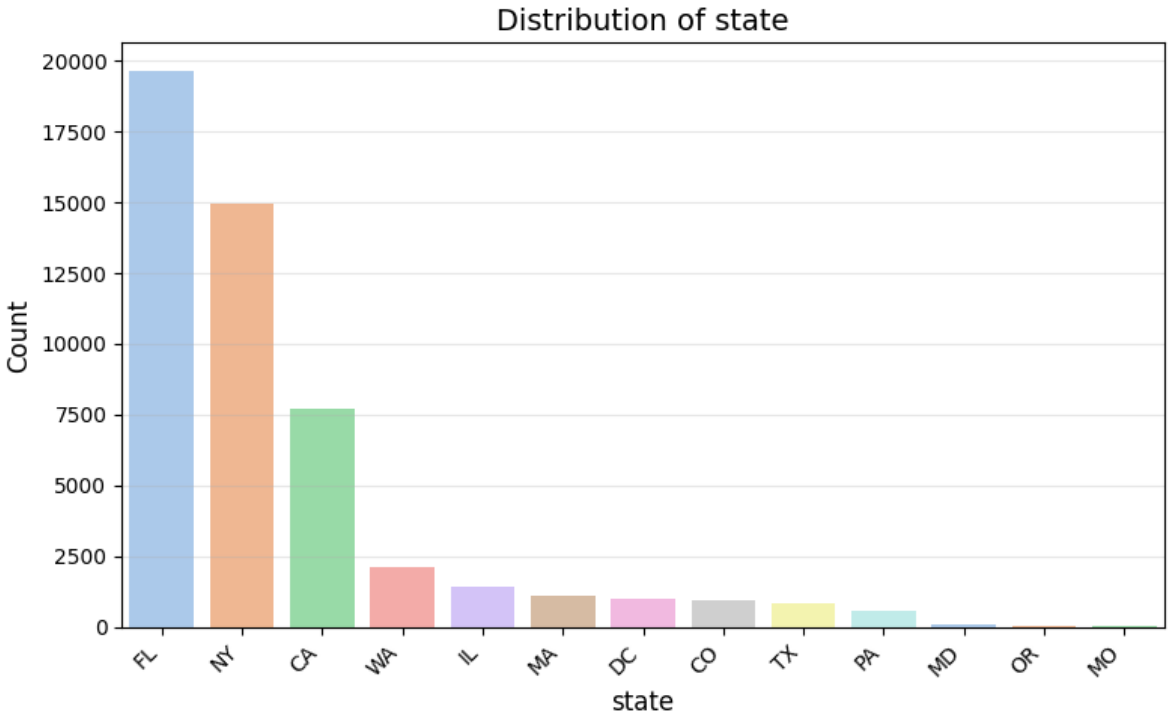


Figure 4.1: Distribution of Buildings Across States

However, the evidence in the literature is conflicting. Walter and Mathew (2019) report that the BPD sample is representative with respect to EUI, whereas Berger, Mathew, and Walter (2016) argue that the sample is not representative of the U.S. building stock. The EDA and insights from the literature review support the view of non-representativeness, leading to the conclusion that the dataset's external validity is limited and this is acknowledged as a research limitation.

Given the dimensionality of the dataset, outlier treatment was applied solely to the prediction target. As shown in Figure 4.2, the site EUI variable contains numerous outliers. Notably, many of these outliers exceed 500 kWh/m^2/year, a value already considered exceptionally high.

Therefore, the possibility of incorrect entries or local characteristics was examined. Time series analyses of five randomly selected buildings sharing the same postcode as one random outlier, and five random buildings with outlier values, indicated that outliers should be removed. For some buildings, a single incorrect entry may have caused the

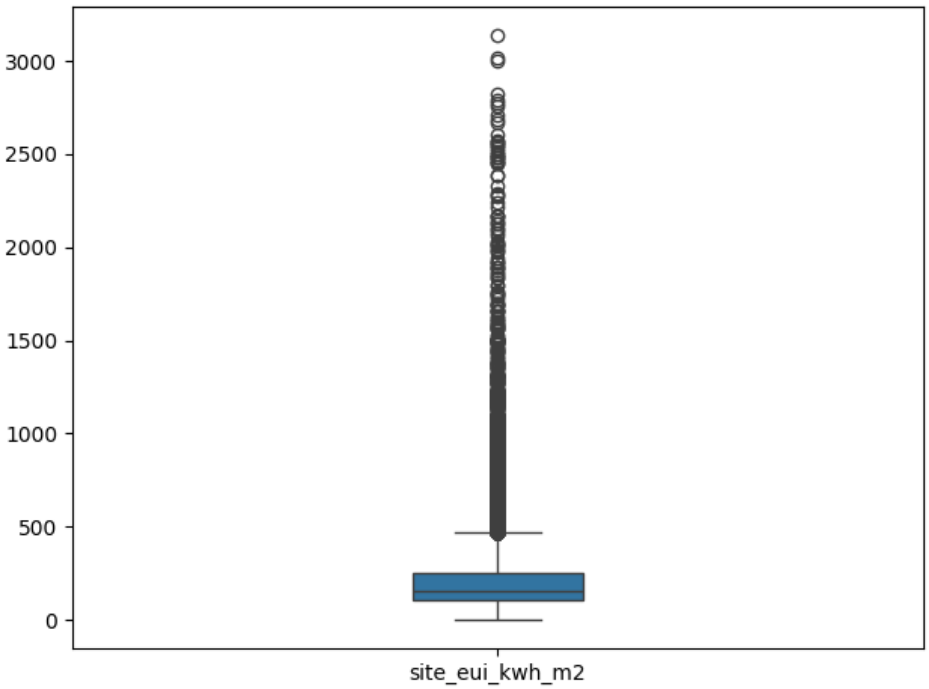


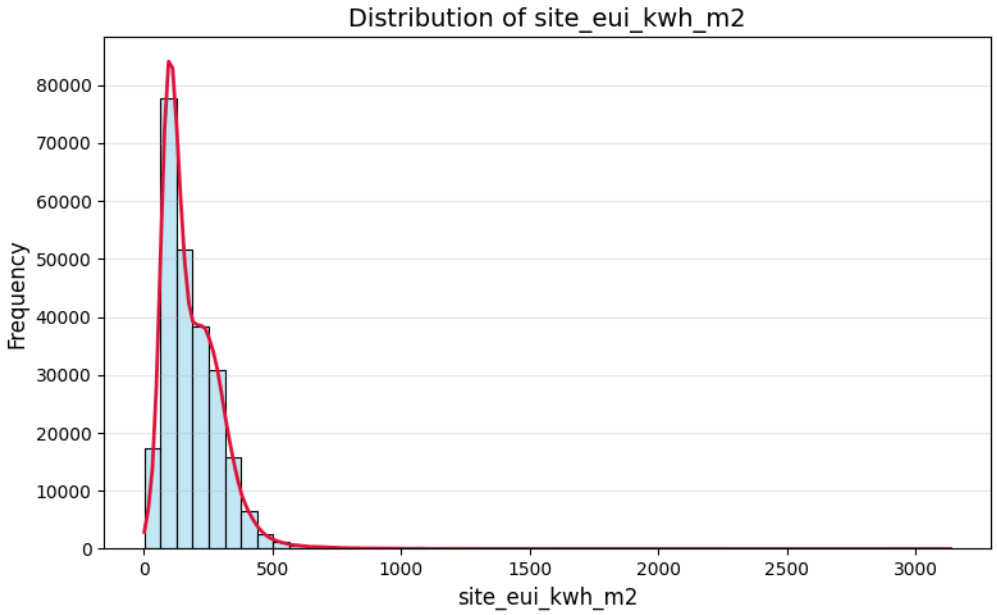
Figure 4.2: Histogram of Site EUI

extreme value, whereas others displayed consistently high measurements. Consequently, values exceeding $Q3 + 1.5 \times IQR$ kWh/m^2/year were removed.

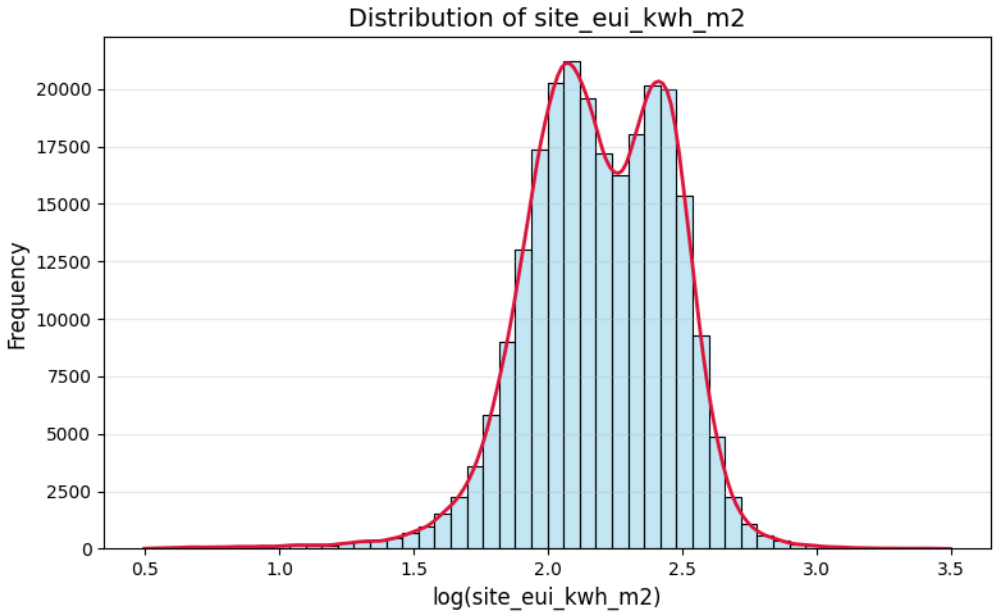
4.1.2 Distributions and Non-Linear Transformations

After understanding the basic structure of the dataset, the distributions of the most important building characteristics were assessed. These included *site EUI*, *floor area*, *year built*, *climate*, and *facility types*. The most critical aspect was the analysis of the *site EUI* distribution. As Figure 4.3 shows, the distribution is clearly skewed and resembles a log-normal distribution. Indeed, the the $\log_{10}$(site EUI) transformed scale of Figure 4.3, reveals a bi-modal distribution of the prediction target. This was a key finding of the EDA, as it suggested the presence of two sub-populations within the sample, and if the predictor variables do not adequately model this structure, predictions may be inaccurate for the overlapping region. The overlapping region was relatively large, which increased the challenge of accurate prediction. Descriptive statistics for the two groups, based on a $\log_{10}$(EUI) threshold, revealed differences primarily in average *building age*, *floor area*, *energy star rating*, and *fuel EUI*, but no clear boundary to construct groups. The most notable observation was the variation in fossil fuel use; since electricity consumption was comparable across groups, fuel use could serve as a proxy for differences in occupancy patterns and equipment. However, as described in the Methodology 3, *site EUI* is linearly dependent on both electricity and fuel use, and therefore only a ratio of these variables was included in the model. This finding provided insight into modelling

requirements and informs potential model diagnostics.



(a) Original scale



(b) Log-transformed scale

Figure 4.3: Distribution of Site EUI

A similar pattern was observed in the *floor area* variable. The non-transformed variable followed a log-normal distribution, while the $\log_{10}(x)$ transformation revealed two distinct peaks. This was expected, as the two sub-populations corresponded to single- and multi-family buildings.

Interestingly, a bi-modal distribution, although less pronounced, was also observed in the *year built / age* variable (Figure 4.4). The bi-modality gap was explained by

the Second World War. Given that older buildings generally contain less energy-efficient technologies, it was important to assess how well this distribution aligned with the target variable and whether older buildings had been renovated. As noted previously, the mean age differed between groups based on the log-transformed EUI threshold, but the renovation rate remained unknown.

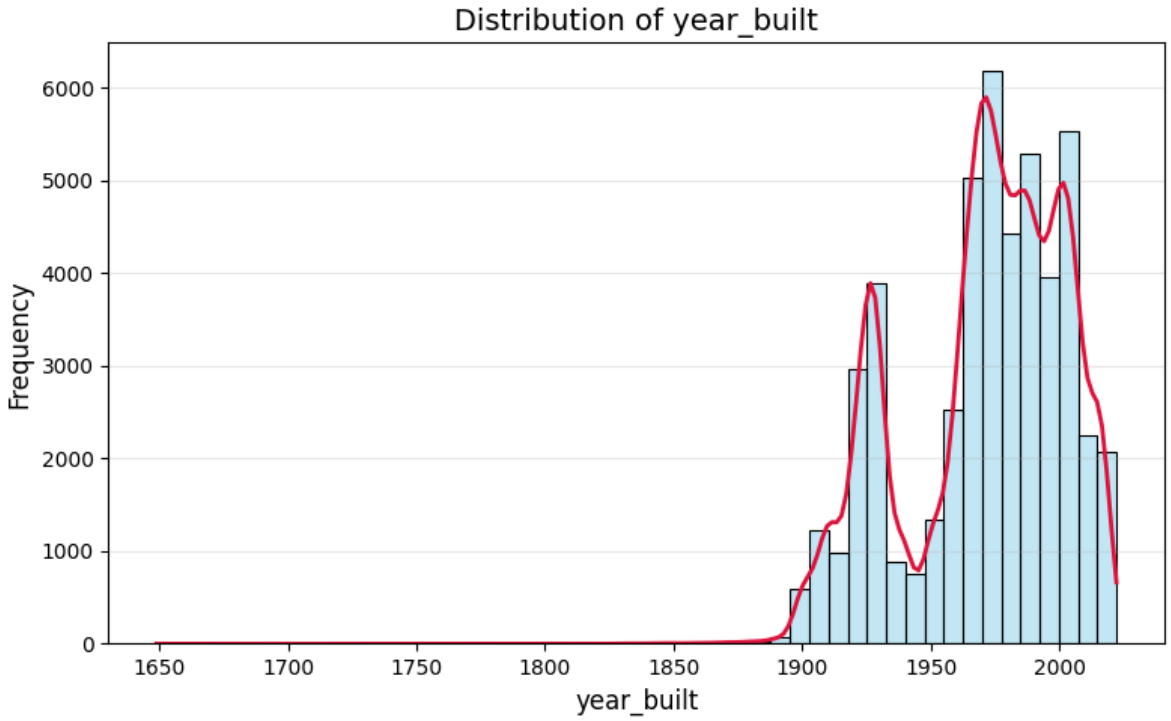


Figure 4.4: Distribution of Year Built

The distribution of the *energy star rating* (Figure 4.5) in the sample resembled a linearly increasing distribution, with a peak at 100 (ignoring missing values encoded as 0). Given that this variable is intended to rank buildings among peers, the observed distribution suggests the presence of sampling bias. Building owners of more energy-efficient buildings may have had a higher incentive to rate their building, or such buildings may have been more likely to be included in the BPD dataset. This sampling bias likely reduced the predictive power of the variable.

Lastly, the assessment of spatial distribution yielded mixed results. Hotter and more humid climates were well represented in the sample, whereas other climate types were underrepresented. Greater climate diversity would likely improve the predictive power of this variable. Although spatial distribution has not been extensively examined in the literature, the observed pattern is not ideal for modelling. Figure 4.6 shows the distribution of climate types, and Figure 4.7 presents a map of average site EUI based on the postcodes included in the sample.

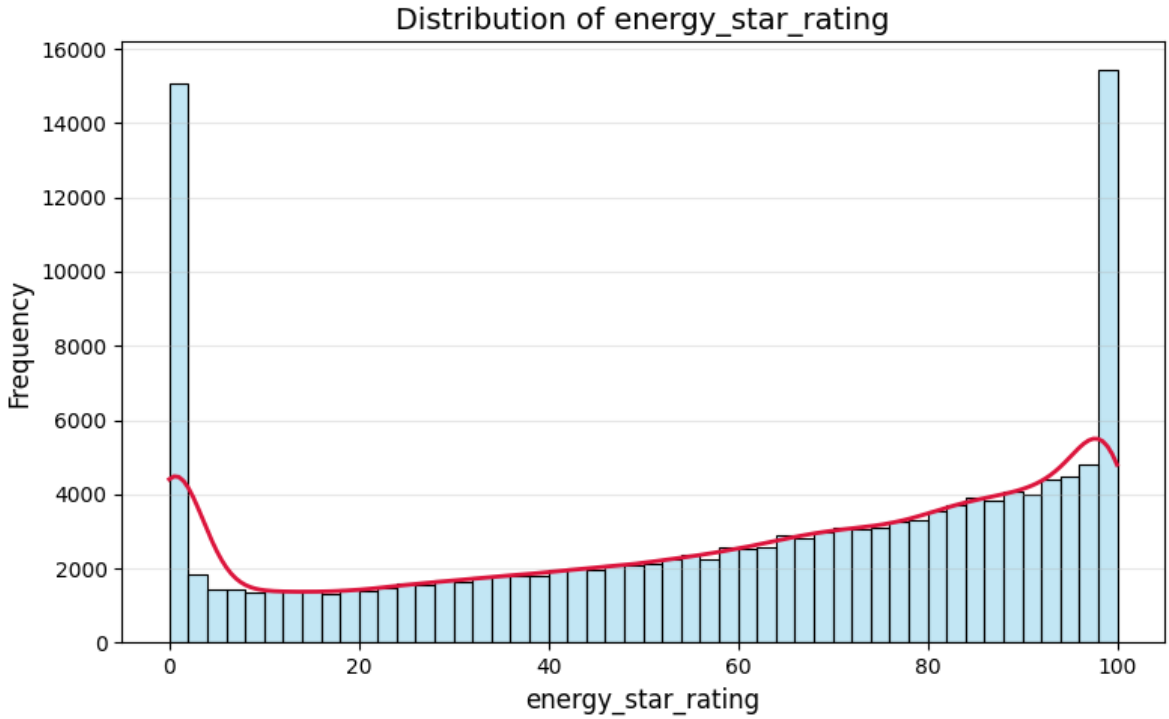


Figure 4.5: Distribution of Energy Star Rating

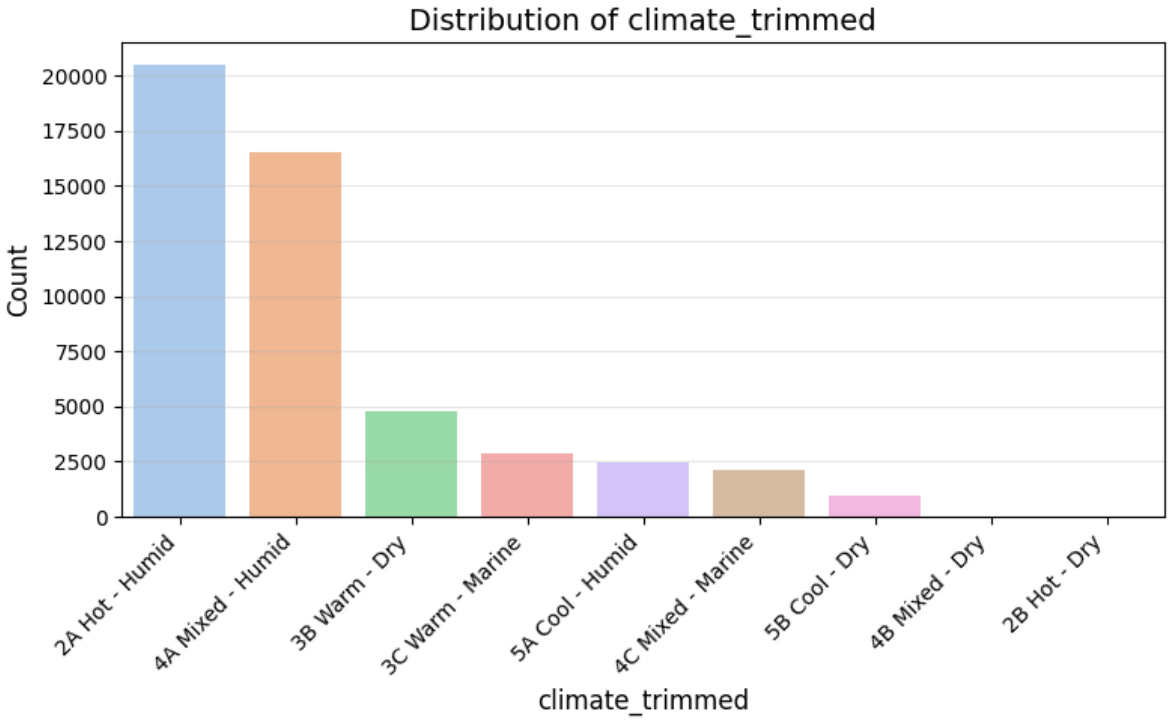


Figure 4.6: Distribution of ASHRAE Climate Types

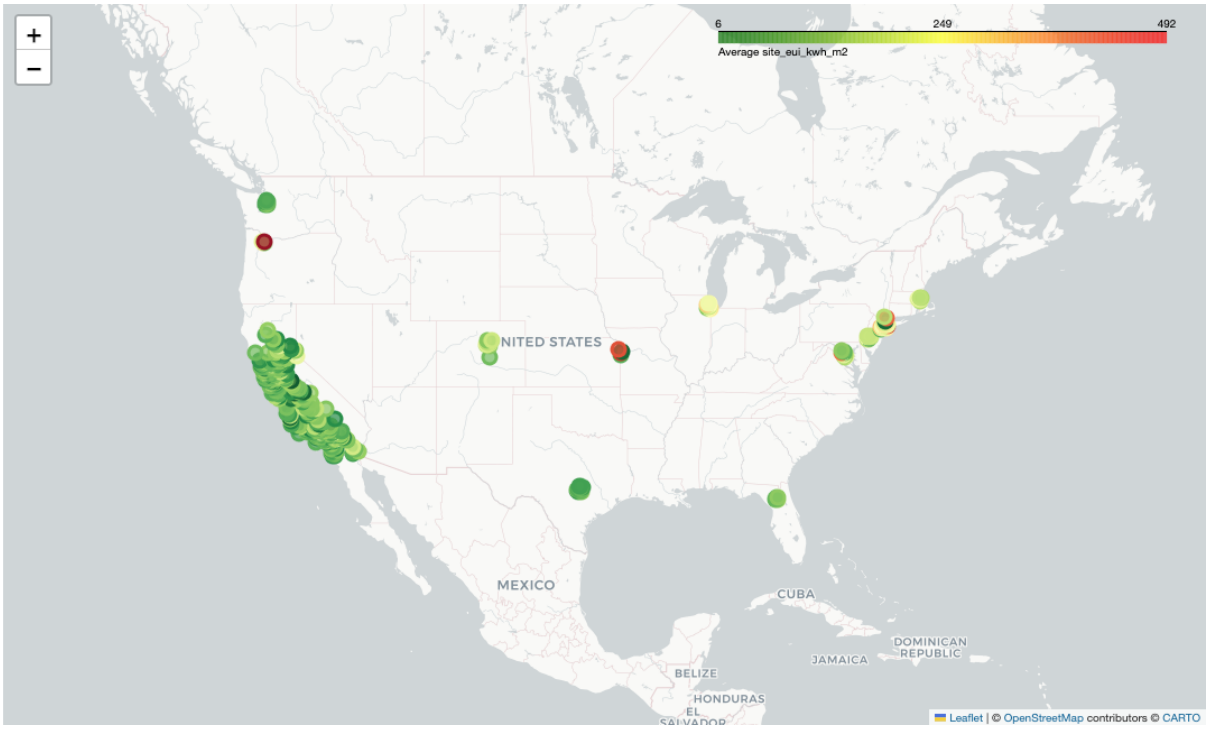


Figure 4.7: U.S. Map of Site EUI in the Sample

4.1.3 Relationships

Due to the large feature set, the assessment of linear relationships was limited to building characteristics, which were plotted on a pairwise scatterplot grid. Relationships among the most important variables were also visualised after applying non-linear transformations. A quantitative assessment of linear relationships was conducted through correlation analysis.

The scatterplot analysis did not reveal any meaningful patterns for modelling purposes. Most variables showed no clear relationships in the data. The only notable association was between the $\log_{10}$-transformed *site total energy* and the $\log_{10}$-transformed *floor area*. Figure 4.8 illustrates a positive linear trend, indicating that a percentage increase in floor area corresponds to a proportional percentage increase in total energy demand.

The correlation analysis provided several useful insights. Only correlations between the target variable and the remaining features were computed; consequently, multicollinearity among explanatory variables was not assessed. Embedding features were excluded from this analysis. As expected, the strongest positive correlation was observed with *fuel EUI*. Several weather variables exhibited moderate to strong positive or negative correlations with the target. These included heating and cooling degree days, their cumulative sums, average wind speed, precipitation, temperature variables, and humidity. Other explanatory variables, such as *building age*, *population*, *electric EUI*, and

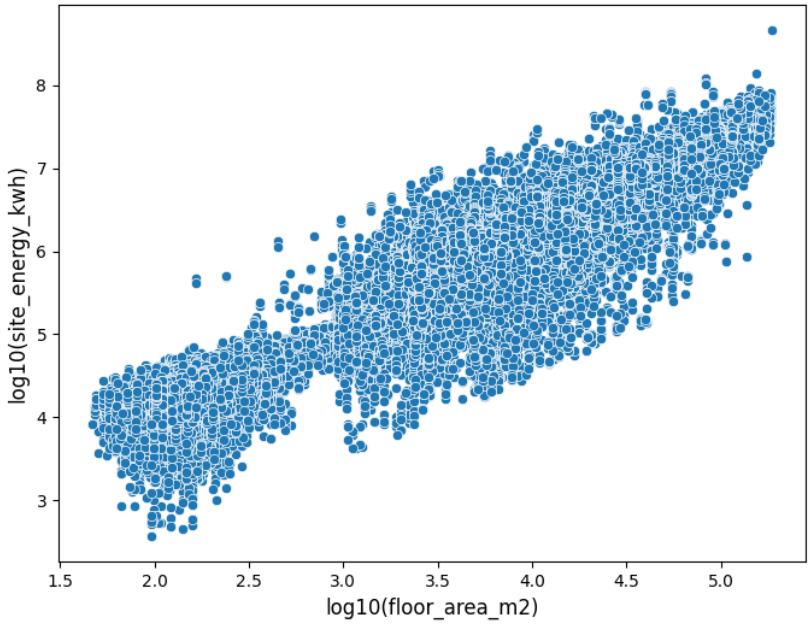


Figure 4.8: Scatterplot of $log10(floor\ area)$ and $log_{10}(site\ total\ energy)$

$\log_{10}(floor\ area)$, showed moderate positive correlations with the target, whereas the *energy star rating* was strongly negatively correlated. Out of all analysed variables, 172 had correlations greater than or equal to $|0.3|$, while 367 exhibited correlations below $|0.3|$. Figure 4.9 presents the ordered correlations, showing that most variables are only weakly associated with the target.

4.1.4 Dimensionality Reduction

To assess the linear dependence among columns and identify latent factors, PCA was conducted. The data were analysed using three PCA components, which explained approximately 60% of the total variance. The first component accounted for about 42% of the variance. A biplot of the first two components revealed distinct clustering of the data into approximately 7–9 clusters. The most influential variables for each component were weather-related: the first component was primarily associated with temperature, the second with atmospheric conditions, and the third with a mixture of other atmospheric conditions (see Appendix C). These results indicate that the clusters correspond to climate patterns. Since climate codes are already included in the BPD data and all variables were used in modelling to capture non-linear relationships, the clusters were not analysed further. Figure 4.10 presents the biplot, with loadings scaled to illustrate the influence of variables on cluster assignment.

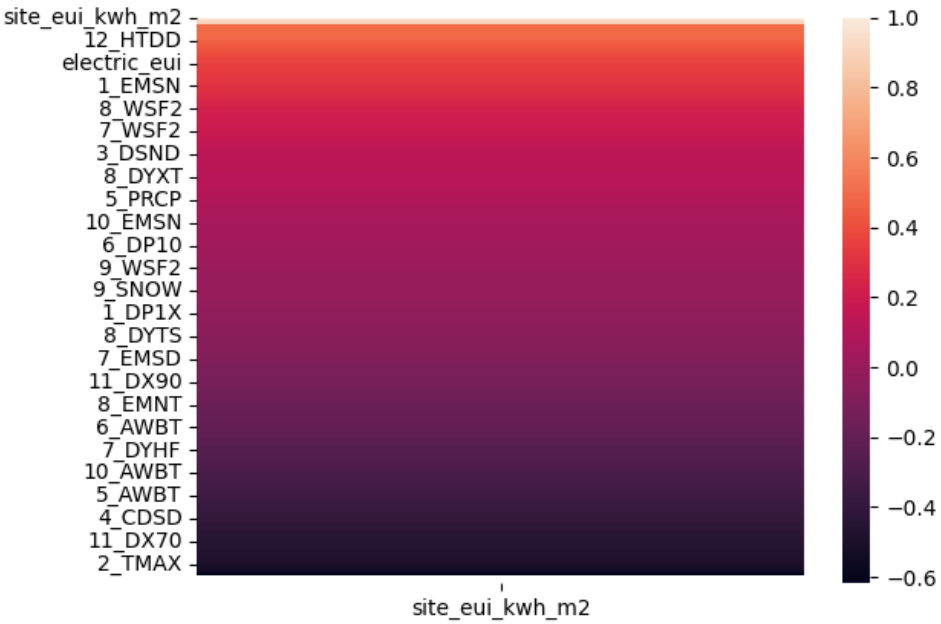


Figure 4.9: Heatmap of Ordered Correlations

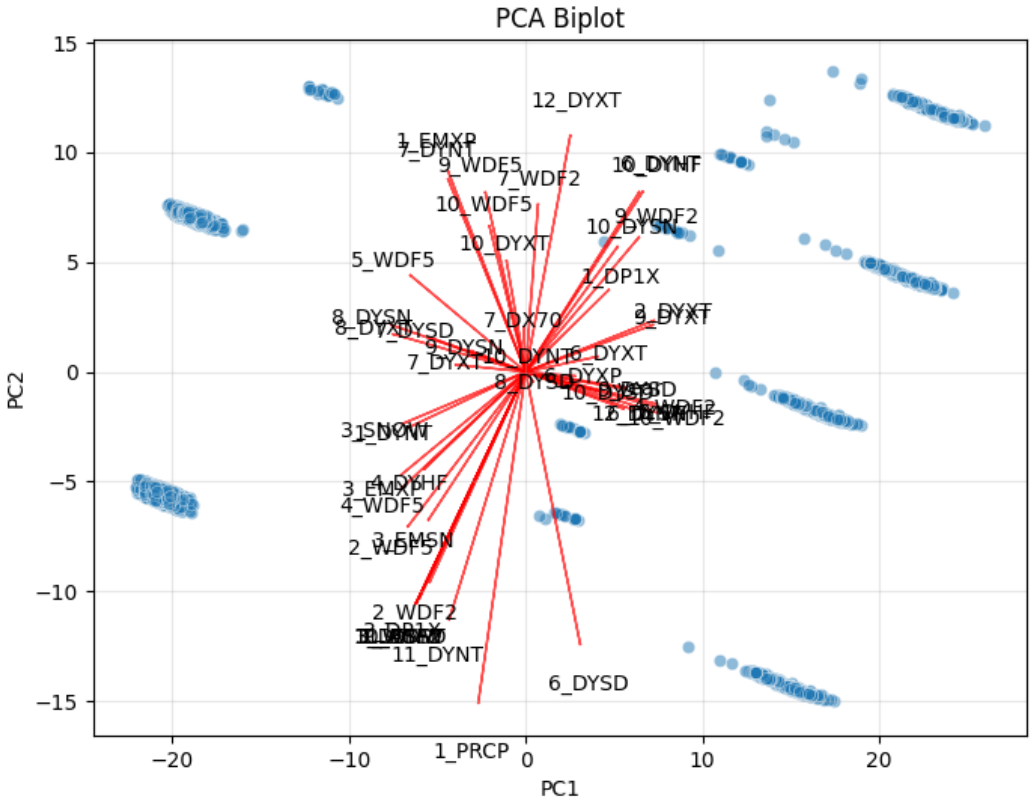


Figure 4.10: PCA Biplot with Components

4.1.5 Summary of Key Results of the EDA

The EDA revealed three key findings that were relevant for the modelling phase of this study. First, the target variable appeared to comprise two distinct populations. Comparison of descriptive statistics between these groups indicated some differences. However, the overlap between populations presented a challenge for accurate modelling. Second, the target variable contained numerous outliers, with values exceeding the typical ceiling for residential EUI. The causes of these extreme values were inconclusive, and they were therefore removed. Third, weather variables showed correlation with the target and exhibited linear dependence among themselves. It was expected that weather variables would contribute meaningfully to the predictive performance of the models.

4.2 Identification of the Best-Performing ML Model

The analysis of model performance identified XGBoost as the best-performing algorithm across all evaluated metrics. The model achieved consistent scores across five folds (see Figure 4.11 and Table 4.2). The second-best model was the RF regressor, which exhibited marginally lower performance in the tracked metrics but required a substantially longer training time. A baseline XGBoost model trained in approximately 10 seconds, whereas the RF models required around 10 minutes. The ANN ranked third, showing spikes in performance metrics, followed by the linear models. This underperformance was expected, as linear models are not well-suited to capturing the non-linear patterns present in the data. The standard deviation of scores was very low (<1) for all models, indicating stable performance on unseen data and a low risk of overfitting.

Table 4.2: Comparative model performance across evaluation metrics (mean ± standard deviation)

Model	**MAE**	**RMSE**	**R^2**	**MAPE**
ANN	35.75 ± 1.31	48.64 ± 1.41	0.690 ± 0.017	0.292 ± 0.017
Linear Regression	38.20 ± 0.16	51.34 ± 0.29	0.654 ± 0.003	0.329 ± 0.003
Ridge Regression	37.71 ± 0.24	59.58 ± 7.68	0.527 ± 0.117	0.326 ± 0.002
Random Forest	31.24 ± 0.21	43.58 ± 0.39	0.751 ± 0.004	0.258 ± 0.002
XGBoost	30.81 ± 0.18	42.75 ± 0.33	0.760 ± 0.003	0.254 ± 0.001

Furthermore, the performance of the best-performing model was compared with alternative data representations: one using time-averaged data at the building level and another using postcode-level averaged energy demand combined with feature embeddings. Averaging the building-level time series resulted in a slight performance improvement, which was unexpected. Given the temporal richness of the weather and building characteristics, it was anticipated that time-varying features would better explain variation

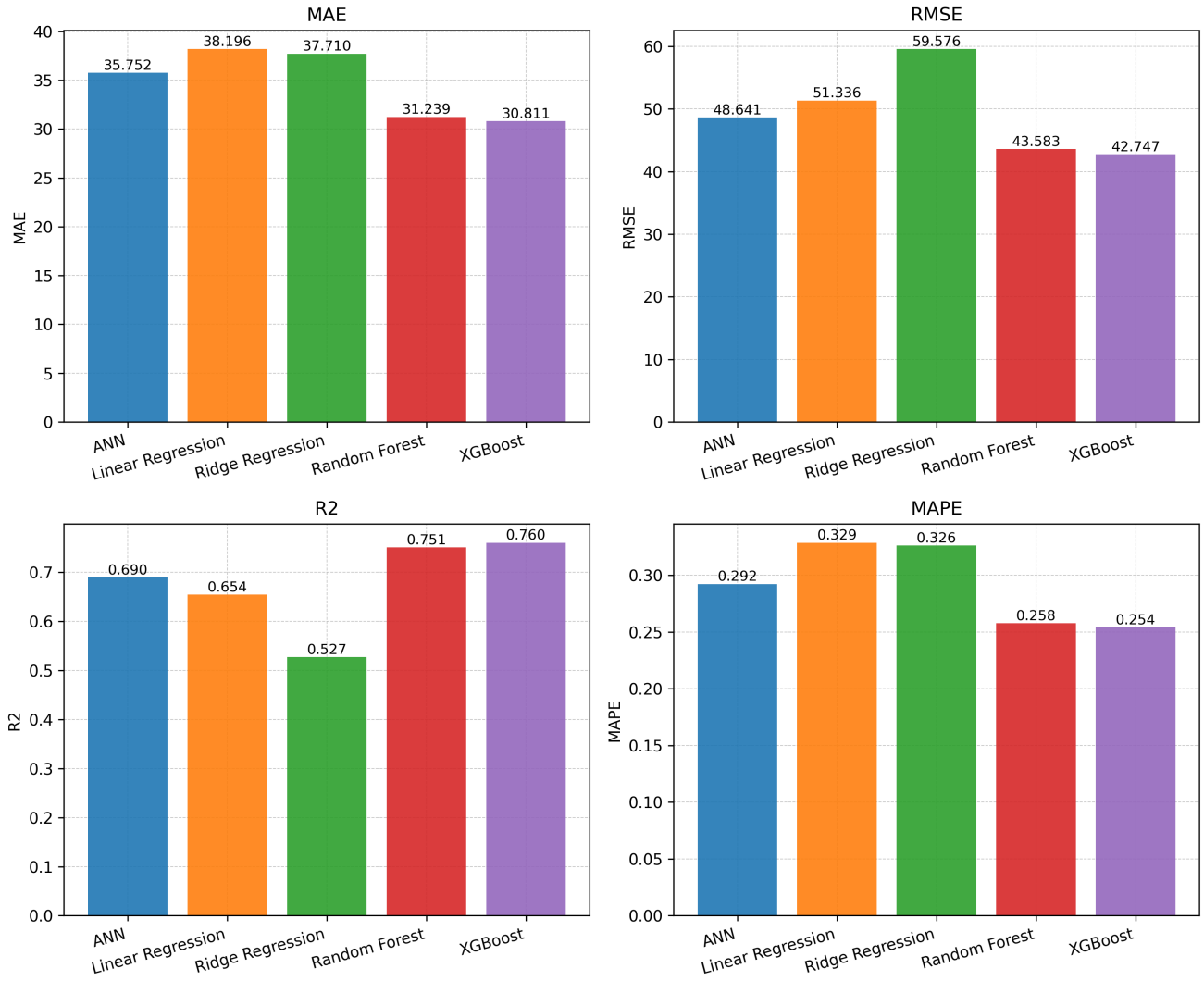


Figure 4.11: Comparison of Models Performance Across Five Folds

in energy consumption. However, the results suggest the opposite, as time-smoothening produced more stable inputs, potentially reducing the influence of measurement error or inherently variant effects. Consistent with this observation, the explained variance (R^2) also increased. More notably, the postcode-level embeddings model achieved superior performance across all error-based metrics, although it underperformed in terms of R^2. This outcome is plausible by design, as the embeddings are constructed at the postcode level, but it remains notable that the error metrics outperform those of the building-level models. Although the comparison is not strictly like-for-like, given the differing levels of resolution, the results nonetheless provide meaningful evidence of the potential value of embeddings in modelling pipelines and alternative predictive settings. The results are summarised in Figure 4.12.

4.3 Hyperparameter Tuning of the Best Model

The hyperparameter tuning of the XGBoost model yielded no substantial performance improvements over the baseline specification (Figure 4.13). Although all error metrics showed marginal improvements, the tuning proved detrimental in practice due to a substantial increase in training time. The baseline model trained in approximately 10 seconds, whereas the optimised model required almost 3 minutes and 2 hours to find the best parameters. This outcome reveals an additional insight regarding the baseline model. Not only does the model generalise well, but the minimal improvement achieved through hyperparameter tuning suggests that the baseline specification already captured most of the learnable signal present in the data. The sample size appears sufficient, and further performance improvements are likely to require additional or more informative features rather than optimisation.

4.4 Model Diagnostics

Model diagnostics and uncertainty revealed stable but noisy predictive performance. The analysis of predicted and observed values (Figure 4.14) revealed a clear linear trend with substantial dispersion, indicating that the model captures overall trends but fails to produce reliable point predictions. This suggests that the available variables are sufficient to model aggregated patterns, such as higher energy use in older buildings or in certain climates, but are insufficient for accurate individual-level prediction. For example, variance in occupant behaviour or renovation may result in substantially lower actual usage. This finding is consistent with earlier insights from the analysis, which indicate that the absence of detailed occupancy-related factors limits the model's ability to generate accurate individualised predictions.

Uncertainty analysis based on error deciles and the calibration plot confirmed these

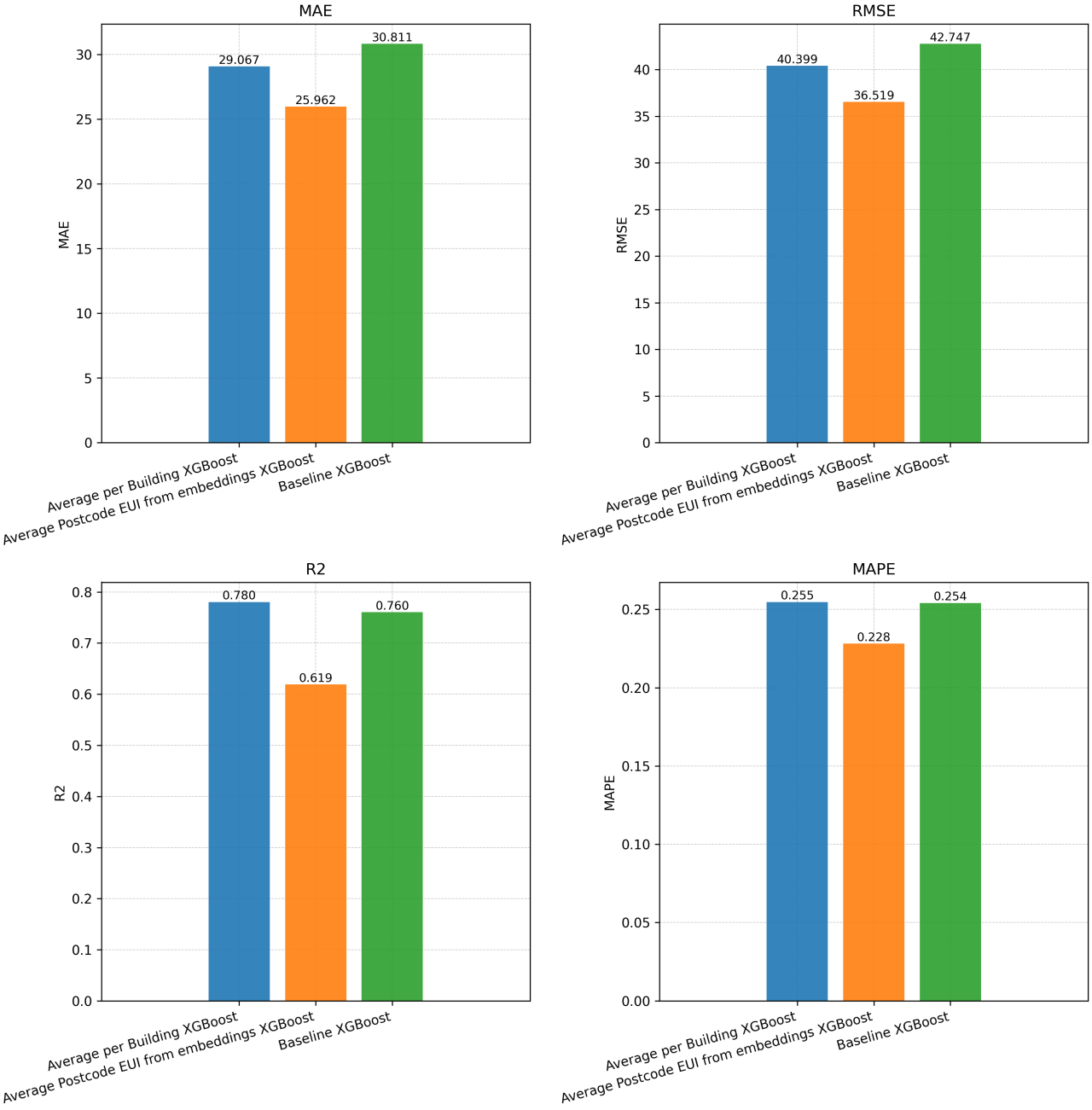


Figure 4.12: Comparison of Alternative Models

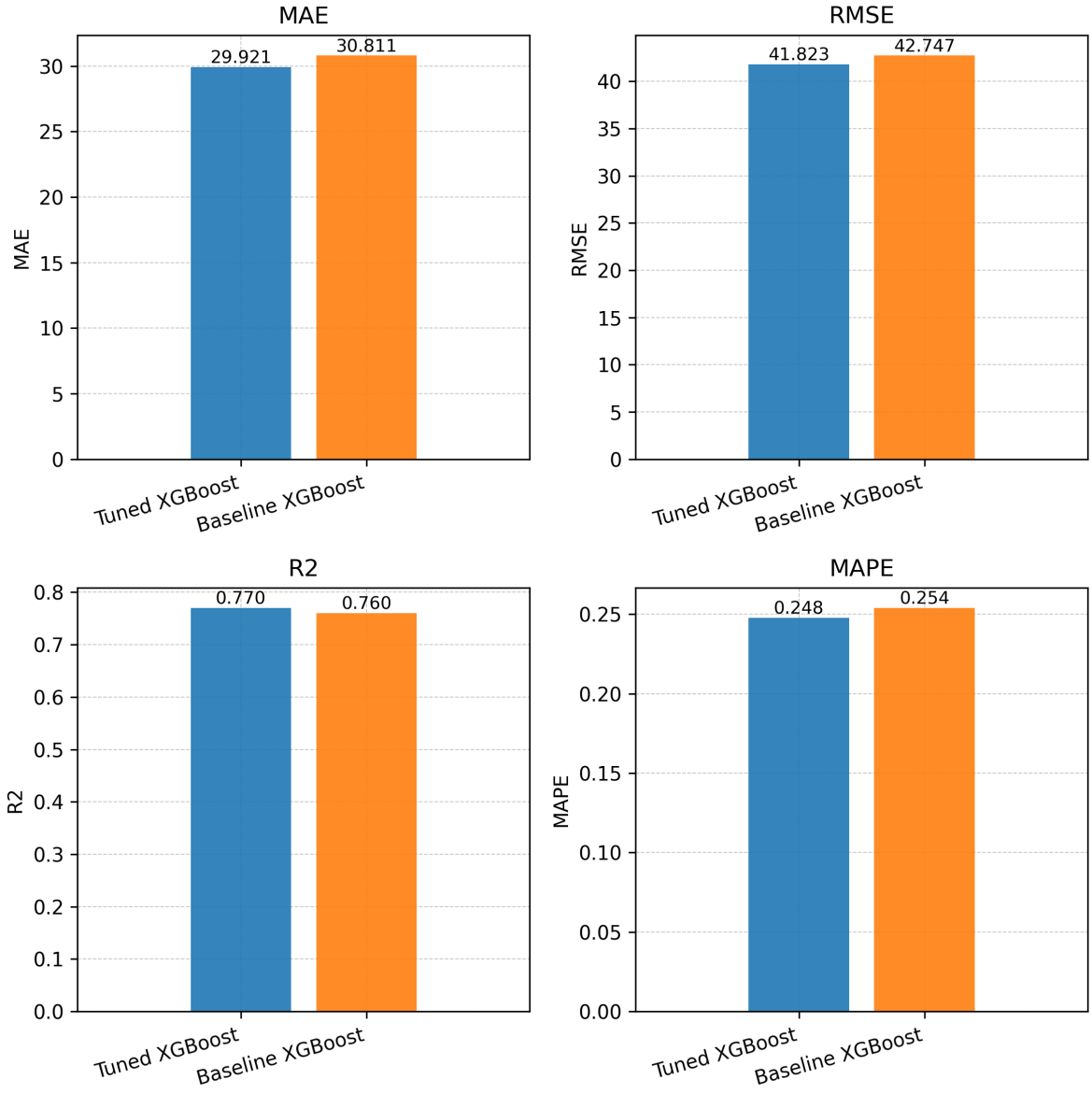


Figure 4.13: Comparison of Tuned and Baseline XGBoost Performance

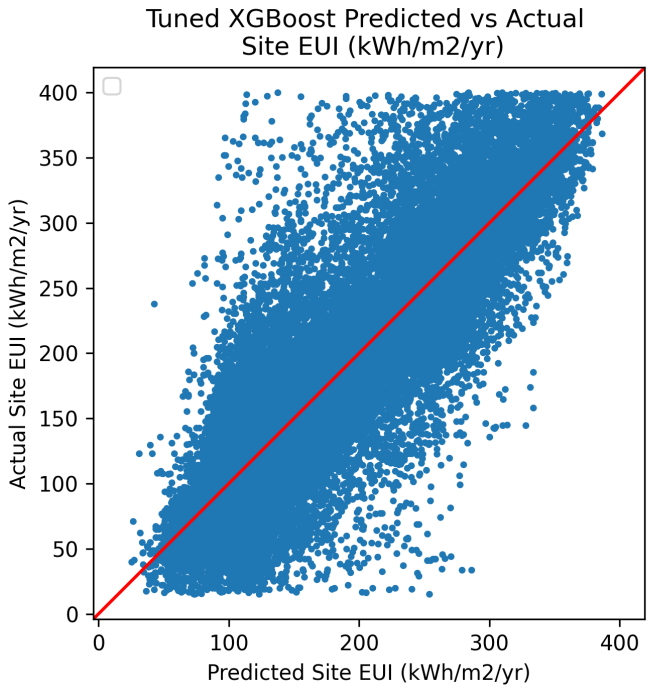


Figure 4.14: Scatterplot of Predicted and True Observations

findings and quantified the magnitude of prediction error. Errors were highest in the middle deciles (Table A.1), but relatively high for the low tail and relatively low for the high tail of predicted EUI. The calibration plot (Figure 4.15) revealed peak uncertainty in the mid-range of predicted EUI values. The increasing spread at low to medium EUI levels indicates greater intrinsic variability, while the lower spread at the extremes may partly reflect a smaller number of observations. These results indicate heteroskedastic errors, with prediction uncertainty varying systematically with the level of EUI. Prediction intervals further showed that 90% of observed values fall within the model-generated interval, although the interval width remains substantial (105-237 kWh/m^2/year).

The residual analysis (Figure 4.16) further contextualises the findings of the uncertainty assessment and provides insight into the sources of model error. Overall, the model performs well given the available data, and substantive performance improvements are likely to require additional explanatory variables rather than further model refinement. Consistent with the uncertainty analysis, the residuals exhibit clear heteroskedasticity, which is evident in the residual scatterplot. The violation of homoskedasticity is characterised by diagonal upper and lower bands, reflecting structural constraints in the data rather than deficiencies in the model. Specifically, site EUI is bounded below by zero and bounded above through the applied outlier removal, which mechanically restricts the magnitude of residuals. As a result, the variance of errors increases approximately linearly with the predicted value, indicating predominantly random error rather than systematic bias. The residual distribution is marginally right-skewed but remains ap-

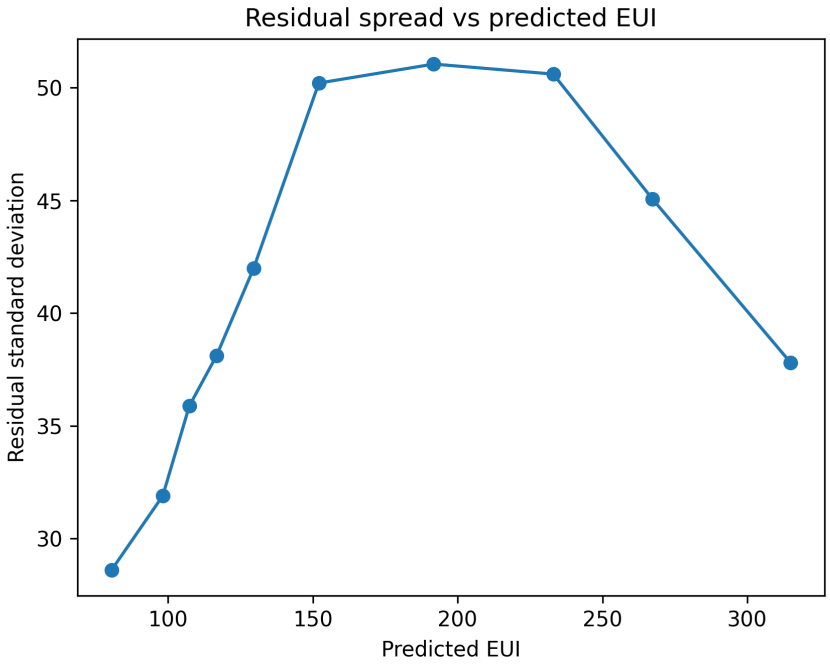


Figure 4.15: Calibration Plot

proximately normal, suggesting that the model errors are largely randomly distributed. This indicates that the model captures the dominant systematic patterns present in the data. Group analysis by location, building and climate types revealed no significant differences in residual error. Consequently, the remaining unexplained variance appears to be stochastic and unpredictable given the available feature set. This finding further supports the interpretation that site EUI exhibits substantial inherent variability and that the model does not suffer from systematic bias.

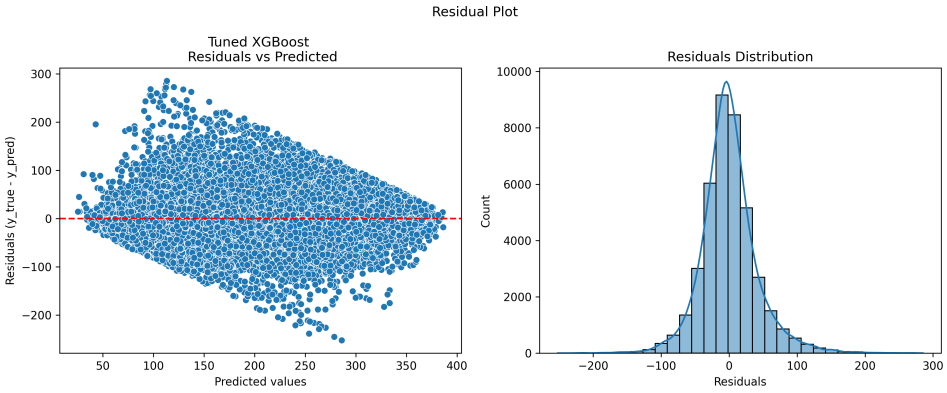


Figure 4.16: Scatterplot of Predicted Values and Residuals (left) and Distribution of Residuals (right)

4.5 Comparison of Performance Across Feature Subsets

The final stage of the analysis involved a performance comparison across variable subsets (Figure 4.17). As expected, the highest performance was achieved using the full feature set; however, the results reveal several additional insights. The first key finding is that weather variables and embeddings, when used independently, achieved comparable performance across all evaluation metrics. Given that embeddings are time-invariant while weather varies temporally, this suggests that the two groups capture distinct effects, despite environmental information being encoded within the embeddings. When combined with the baseline features, which already include climate data, the performance of both subsets improved; however, the incremental gain relative to the baseline-only model was modest. In light of the earlier results from the averaged single-entry and postcode-level embedding models, this finding further supports the interpretation that temporal variability in weather is subordinate to unobserved occupational variability in driving energy demand. The removal of climate dummies and the *energy star rating* lead to performance deterioration across all metrics, highlighting their importance for model accuracy. Additionally, the embeddings might be failing at building-level prediction due to spatial resolution mis-match and time invariance. The error by decile difference of the embeddings model compared to the baseline feature models was also negligible. This suggests that embeddings encode postcode characteristics, which are most informative for typical buildings but insufficient to explain high-consumption behaviour. The second key result is that engineered features consistently improved predictive performance. A comparison between baseline and full feature subsets, with and without engineered features, shows a decline in performance across all metrics when these features are excluded. This improvement is plausible, as the engineered features introduce additional information, capturing meaningful structural differences in energy use patterns.

The analysis of individual feature importances yielded inconclusive results. Several features were consistently important across models, while others appeared influential in one specification but not in closely related models. For instance, *energy star rating* and *climate zone* $4A$ lost their importance following the inclusion of embeddings. This pattern indicates multicollinearity and the use of multiple variables as proxies for similar underlying effects. In relation to earlier findings, this may explain why the weather-only subset achieved an R^2 of approximately 0.50, whereas removing climate dummies from the baseline feature set reduced its R^2 only marginally, to around 0.45. Weather variables may therefore serve not only as direct predictors but also as proxies for spatial location or other latent characteristics. The most consistently important variable across models was an embedding component, `feature270`, which is likely associated with environmental or spatial factors. Several models exhibited a strong reliance on a single dominant feature,

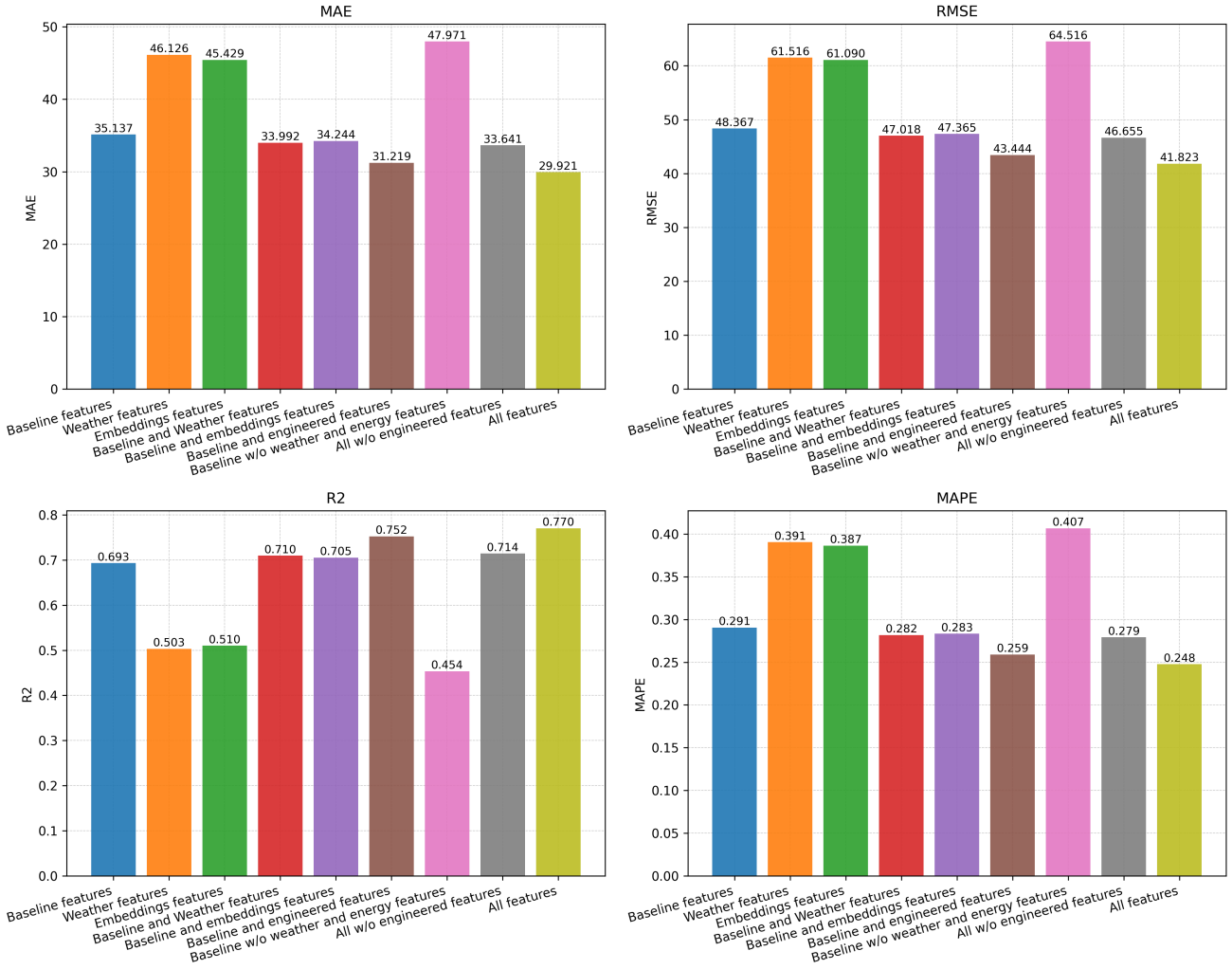


Figure 4.17: Comparison of Feature Subsets Performance

suggesting a degree of overfitting. Overall, the feature importance analysis indicates that a portion of the remaining unexplained variance is likely driven by unobserved occupational factors or by variables capturing building-level energy efficiency characteristics that are not directly observed in the dataset (Figure 4.18).

4.6 Summary of Analysis Results

The analysis confirmed XGBoost was the best-performing model, and while hyperparameter tuning was applied, it produced negligible improvements. Alternative models suggested that embeddings alone, aggregated at the postcode level, have predictive potential. The feature subset analysis demonstrated that engineered features improve model performance, and that weather and embeddings subsets yield comparable predictive power. Analysis of feature importances indicated some overfitting, but also revealed stable dominant predictors. Model diagnostics and uncertainty quantification confirmed that the available signal in the data was effectively captured, the models were well-calibrated but with relatively high errors, and that additional variables would be required to explain the unexplained variance.

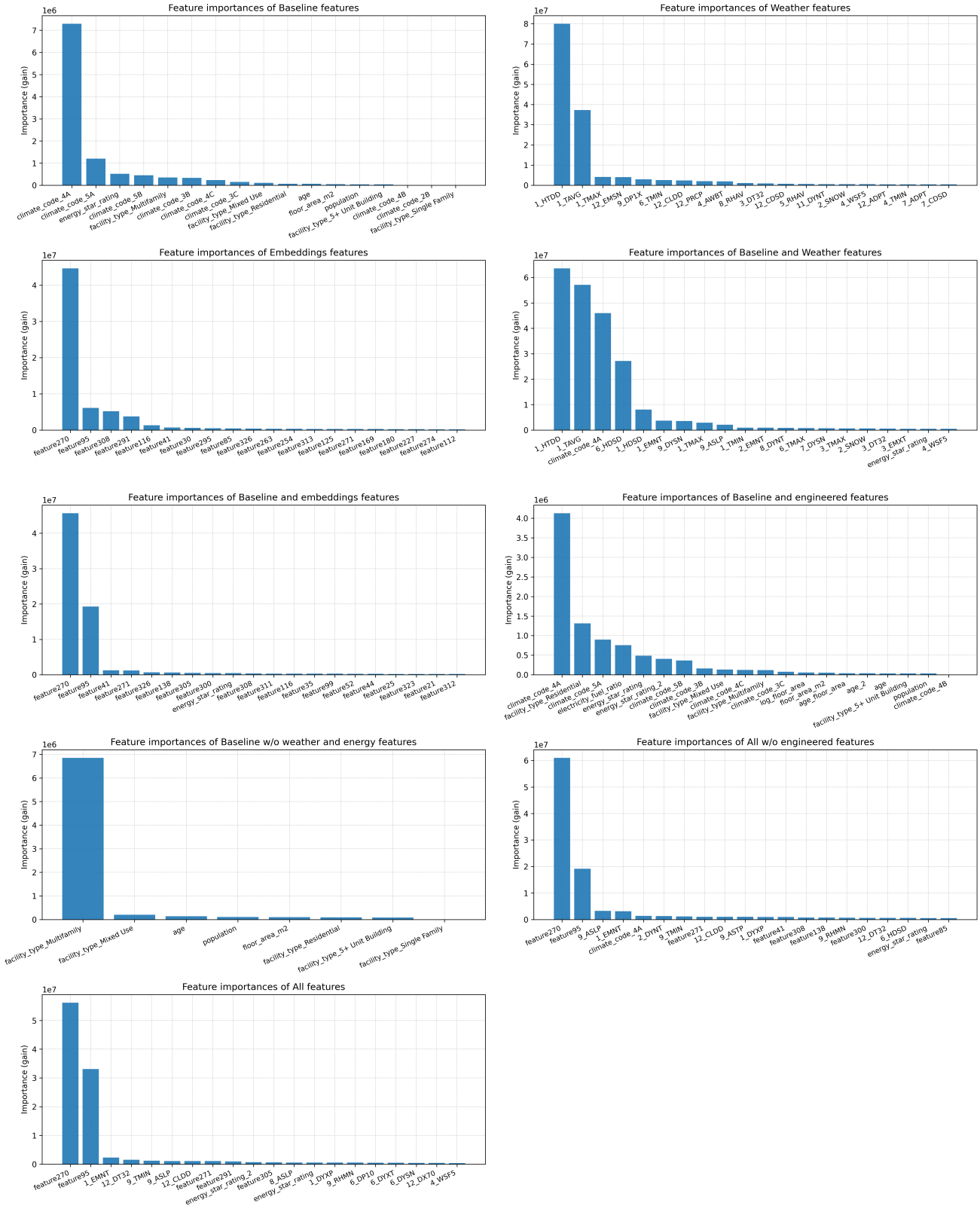


Figure 4.18: Feature Importances by Gain of Individual Subsets

5 Discussion

5.1 Discussion of Theoretical Findings

Three approaches to estimating building energy consumption were identified in the literature: traditional, data-driven, and hybrid methods. Traditional approaches include detailed engineering assessments based on simulations and physical equation-based modelling. Statistical methods, which specify causal relationships and manually selected predictors within linear models, bridge traditional approaches and data-driven methods. Machine learning models leverage large-scale datasets to learn statistical patterns without explicitly defined physical relationships. The hybrid approach combines physical and data-driven methods and includes advanced techniques such as physics-informed ML and agent-based modelling. The data-driven approach was examined in depth, and the literature identified several successful applications of ML algorithms for long-term energy consumption prediction. This addressed the first theoretical research question concerning how residential energy consumption can be predicted. The empirical analysis challenged the prevailing assumption that existing studies operate in data environments rich in diversity. The results confirmed the dominance of tree-based algorithms and demonstrated that long-term data-driven energy prediction depends less on algorithmic choice and more on data quality, latent variable representation, and feature availability.

The comprehensive overview of influencing factors provided a theoretical foundation for the analysis and addressed the second research question, which examined which factors affect residential building energy demand. The drivers were grouped into two categories: internal and external. Internal factors include physical characteristics of the building, usage patterns, installed equipment, and occupational behaviour. External factors relate to environmental conditions, such as climate, microclimate, and weather. Both the literature review and the empirical analysis highlighted that while some of these factors are readily available, others – particularly occupancy patterns – are influential yet difficult to observe or obtain.

The second to last theoretical sub-question aimed to describe the energy use of typical residential building in the U.S. The review described that the population of U.S. residential assets is dominated by single-family homes and identified some differences

with multi-family homes. A typical residential EUI was found in the range of 150-250 kWh/m^2/year. The obtained sample contrasted this population description and subsequent analysis revealed population representativeness issues, a limitation threatening external validity of the study. Outlier removal was also grounded in these theoretical findings.

The penultimate theoretical sub-question aimed to characterise the energy use of a typical residential building in the United States. The literature review indicated that the U.S. residential building stock is dominated by single-family homes and identified differences relative to multi-family buildings in terms of size, occupancy, and energy use. Reported values of typical residential EUI were found to lie in the range of 150-250 kWh/m^2/year. The empirical sample contrasted with this population description, exhibiting an over-representation of multifamily homes, which led to representativeness issues and constitutes a limitation threatening the external validity of the study. This implies that model performance and inferred relationships should be interpreted as conditional on the observed sample rather than as generalisable to the population. The outlier removal procedure was grounded in these theoretical benchmarks, ensuring that extreme observations outside plausible residential EUI ranges did not distort model estimation or predictive evaluation.

Data challenges formed the focus of the final theoretical question. The literature highlighted limitations related to sparse long-term temporal coverage, non-linearity, uncertainty, and issues of data availability and quality. The empirical analysis strongly confirmed these constraints, demonstrating that limited feature diversity and missing data restrict achievable predictive accuracy. However, the implied necessity of high-frequency temporal data was partially challenged, as temporally smoothed representations produced more accurate and stable predictions. This suggests that, for long-term residential energy demand modelling, structural and occupational signals dominate short-term weather variability. While targeted data enhancements led to incremental performance gains, multicollinearity and the absence of key latent variables remained constraints on point prediction accuracy. These findings reinforce the theoretical argument that advances in long-term energy demand prediction are driven improvements in data richness, representativeness, and modelling of latent behavioural factors.

5.2 Discussion of Empirical Analysis

5.2.1 Exploratory Data Analysis

The findings of the EDA directly informed the first empirical sub-question: which insights from EDA can guide modelling decisions. Although prior literature suggests reasonable

population representativeness, the empirical analysis identified clear over-representation of multi-family buildings and hot climates, thereby challenging external validity. This imbalance is most plausibly explained by selection bias arising from the data collection. Evidence of a bi-modal target distribution motivated further investigation through uncertainty quantification, which confirmed elevated variance in the mid-range of EUI values. However, causal interpretation remained limited, as this variance could not be disentangled from sample density effects or heteroskedasticity induced by model constraints. Consequently, stratification of the data into sub-populations proved unreliable, a limitation later reinforced by modelling results that indicated the presence of unobserved latent variables. While advanced approaches such as generalised variational inference may offer a principled way to approximate such latent structure, their application would require different data and modelling assumptions. Feature transformations, interactions, and engineered ratios improved predictive performance, despite the EDA offering limited formal guidance for their construction beyond domain-informed intuition. Contrary to expectations set by the literature, variables such as building age and floor area exhibited weak and noisy relationships with site EUI. By contrast, strong associations between weather, climate indicators, and the target variable were consistently observed, aligning with prior findings. Finally, informed outlier removal stabilised model behaviour in the distribution tails, an effect later corroborated through residual diagnostics and uncertainty analysis.

5.2.2 Best-Performing Model and Alternative Data Representations

One of the most important empirical sub-questions sought to identify the best-performing machine learning model. The analysis confirmed prior expectations from the literature that tree-based ensemble methods outperform alternative approaches in long-term energy prediction, with XGBoost emerging as the strongest model in this sample. Its underlying assumptions align well with the structure of the data, and the algorithm remains a suitable benchmark for comparable future studies.

Alternative data representations produced results that partially contradicted findings in the literature, as spatial aggregation led to reduced prediction error. In the case of building-level temporal aggregation, this suggests that occupational patterns (e.g. vacancy rates or usage intensity) are more influential in explaining annual energy variation than weather, which is more relevant at higher temporal resolutions. For postcode-level aggregation, strong performance was more expected, given that the embeddings were constructed at this spatial resolution; however, their outperformance relative to other models remained notable.

The reduction in R^2 provides a plausible explanation, as spatial averaging stabilises

idiosyncratic building-level variation and shifts predictions closer to national or regional means. Nevertheless, the ability of embeddings alone to achieve competitive error metrics is encouraging. The tuned and extended versions of this approach could therefore be valuable for planners, policymakers, geospatial researchers, and organisations with limited data resources. In particular, areas of elevated energy use could be identified, or energy demand inferred in regions with sparse or missing observations.

Given that the BPD dataset is spatially limited, extending postcode-level predictions to the full U.S. building stock could yield additional insights into national energy patterns. However, such an application lies beyond the scope of this study and is left as a recommendation for future research. Finally, the observed mismatch introduced by spatial aggregation at finer resolutions highlights an important trade-off between predictive stability and granularity, which should inform the design of future geospatial modelling efforts beyond this specific application.

5.2.3 Performance Improvements

The analysis of feature subsets lead to the conclusion to the third empirical question and found that embeddings enhance model accuracy, but they embed relationships at misaligned granularity, and therefore, lead to marginal gains. Evidence across the empirical analysis showed that EUI is inherently variant due to occupancy patterns and embeddings fail to model these effects because of their low resolution that aggregates these trends. The feature importance analysis revealed that the additive effects of embeddings were related to environmental factors that are naturally smoother in space that other latent factors and thus more aligned with their structure. The hypothesis that certain socio-economic patterns can be modelled with embeddings, could not be fully rejected but the evidence point in this direction. However, embeddings worked well with postcode-aggregated data, which is aligned with their spatial resolution and it is recommended to explore similar modelling options in energy prediction.

The analysis of feature subsets addressed the third empirical question, revealing that embeddings improve model accuracy only marginally, primarily due to a mismatch between their spatial granularity and the intrinsic variability of site EUI. Across the analysis, site EUI was found to exhibit both systematic and inherently variable components, with a portion of the variation attributable to occupancy patterns and other latent building-specific factors that embeddings, aggregated at the postcode level, were unable to fully capture. Quantitative comparison showed that the inclusion of embeddings increased the R^2 by only 0.01 over the baseline feature set, while reductions in error metrics were similarly modest, confirming their limited incremental predictive power. Feature importance analysis indicated that embeddings predominantly contributed environmental informa-

tion, which exhibits smoother spatial variation and is thus more aligned with the coarser resolution of the embeddings. While embeddings may theoretically encode socio-economic patterns at a coarse spatial level, the empirical evidence shows only marginal gains and no clear capture of similar effects. Notably, embeddings performed effectively when applied to postcode-aggregated data, consistent with their spatial resolution, suggesting that their utility is maximised when aligned with the scale of the target aggregation. In other sparse data contexts, such as climate risk or urban analytics, embeddings might prove futile. These findings indicate that further improvements in predictive accuracy are hard to obtain solely from embeddings.

The final empirical sub-question aimed to determine whether outlier removal, feature engineering, and hyperparameter tuning improve model accuracy. Hyperparameter tuning yielded minimal improvements across all metrics. In contrast, feature engineering and subset selection consistently enhanced model performance. The limited effect of tuning is interconnected with the performance gains from feature selection. As discussed previously, the model captured the available signal effectively due to the sample size, leaving little room for optimisation improvements through parameter tuning. The remaining prediction errors were largely random; although they caused mis-predictions at the individual level, the model generalised well to unseen sample data and accurately captured overall trends. Additional variables introduced into the BPD dataset provided only marginal improvements – climate dummy variables alone were sufficient to explain a substantial portion of the predictable variation in the target. This observation was confirmed through the coverage metric, which indicated a well-calibrated model given the data. From a bias-variance perspective, long-term energy prediction exhibited low bias but high variance. Therefore, further improvements require the inclusion of additional explanatory variables. Some of this was achieved by adding engineered features such as ratios, interactions, and transformations. This interpretation aligns with findings from the EDA and uncertainty analysis. The evidence of bi-modality in the target, wide prediction intervals, increased spread in mid-range EUI, and large errors in middle deciles indicate the presence of structural variance attributable to unobserved variables. Furthermore, literature showed that certain variation in energy consumption is inherently unexplainable even after the inclusion of additional predictors. Overall, the analysis supports the understanding that long-term energy prediction performance is less dependent on algorithm choice and more constrained by data quality, feature richness, and latent variable modelling.

5.3 Conclusion on the Main Research Question

In summary, all research sub-questions were addressed, and the main research question – *To what extent can residential building energy demand in the United States of America*

be predicted using open-source data available at scale? – was answered. While the empirical research did not achieve individual-level accuracy comparable to current literature, it successfully demonstrated solid predictive performance under constraints of linearly dependent variables and sparse, diversity-rich covariates. The analysis showed that over 70% of the variance in site EUI can be explained using building characteristics, climate, and proxy variables, with an average point prediction error of 25.4%. The remaining noise was not attributable to modelling techniques, indicating that open-source data are sufficiently reliable to construct an analytical solution for residential energy demand, with potential for further improvement given additional high-resolution features. Thus, the final conclusion is that basic building and climate characteristics allow prediction of site EUI with 25.4% average error, which is higher than state-of-the-art models but represents robust performance under sparse feature conditions.

Key insight 1: Open-source data can support useful building-level EUI prediction within BPD-like populations, explaining approximately 0.76

Key insight 2: In diversity-sparse, long-term residential energy prediction, PDFM embeddings do not improve building-level accuracy because energy use variance is dominated by idiosyncratic behavioural factors below the spatial resolution of the embeddings.

5.4 Limitations

This paper had limitations identified both prior to the analysis and through the data:

- **Reproducibility Issues**: Two data sources are publicly available with conditional authentication, but PDFM embeddings are accessible only upon request, limiting the reproducibility of this study.
- **Limited Interpretability of the Best Models**: XGBoost models provide limited interpretability due to their design, and feature importance results offer only partial insight; the embeddings are unitless numerical features with no direct interpretation.
- **Validity Threats**: Population non-representativeness due to selection bias limits external validity, while internal validity is potentially affected by outlier removal and measurement errors; the model is well-calibrated for the sample but should be used cautiously for generalised inference.
- **Noisy Predictions**: Even though the model extracted all available signal and provided unbiased estimates, predictions remain relatively variable for individual buildings.
- **Limitations Due to Inherent Variance**: The analysis shows that prediction

accuracy is constrained by inherent variation in building energy use rather than algorithm choice, and additional data may not substantially improve performance.

5.5 Recommendations for Policymakers and Businesses

The analysis surfaced several insights that can be translated into actionable recommendations for policymakers and analytics businesses, while acknowledging the limitations of the data and model:

- **Use Prediction Intervals Rather Than Points**: The model is well-calibrated and unbiased for interval predictions, but point predictions are noisy. In commercial or policy contexts, it should be used for trend estimation and bounded forecasts rather than exact values, particularly given unobserved occupant and system-level variability.
- **Identify Energy Hotspots**: The aggregated postcode-embeddings model can identify regions with high energy consumption, providing a basis for targeted interventions, while recognising that building-level heterogeneity limits precision.
- **Predict for Regions with Limited Data**: The postcode-embedding model can extrapolate energy use to regions where observational data are sparse, but predictions are more reliable at aggregated spatial levels and may not fully capture idiosyncratic behavioural patterns.
- **Integrate Commercial and Research Applications**: Additional data collection, including occupancy, system-level, and temporal features, can improve predictive performance and reduce uncertainty, supporting urban planning, analytical products, or research applications.
- **Acknowledge Data Limitations and Generalisability**: Insights derived from the BPD-like sample may not generalise to all U.S. residential buildings. Policymakers and analysts should account for potential sampling biases and missing latent variables when applying the model to new contexts.

5.6 Recommendations for Future Research

Several key findings from the analysis suggest avenues for future research:

- **Test PDFM Embedding Models on Other Metrics**: The potential of postcode-level predictions leads to the hypothesis that similar models could be applied to other energy-related tasks, such as forecasting heat demand, renovation rates, or energy hotspots.

- **Enhance Alternative Models**: Incorporating additional features into alternative models may improve their point prediction accuracy.
- **Use of PDFM Embeddings for Covariate Prediction**: Embeddings could be trained on external data to serve as inputs for energy prediction models, including socio-economic factors, population dynamics, macro-economic variables, or aggregated building trends.
- **Identify Instrumental Variables for Energy Use**: Finding reliable and easily obtainable proxies or instrumental variables remains a challenge and could benefit from the increasing availability of open data.
- **Test Other Modelling Approaches**: Other techniques, such as physics-informed machine learning, agent-based simulations, or complex parametric methods like generalised variational inference, could be explored.
- **Validate Business and Policy Applications**: Models should be tested in real-world policy and commercial contexts to assess scalability, usability, and practical effectiveness.

6 Conclusion

This dissertation examined how accurately long-term residential energy use intensity in the United States can be predicted using diversity-constrained building data, and whether adding open weather variables and Population Dynamics Foundation Model embeddings meaningfully improves performance. A scalable pipeline was developed to systematically integrate Building Performance Database records with weather data and postcode-level embeddings, and five machine learning models were evaluated within a consistent training and feature-comparison framework.

Empirically, the results show that building-level EUI can be predicted with useful explanatory power from limited building characteristics and climate-related information, but that individual point predictions remain noisy because important drivers are not directly observed. XGBoost performed best and generalised well on unseen sample data under the leakage-aware splitting strategy, while hyperparameter tuning added little improvements beyond feature engineering and subset selection optimisations.

The open data weather variables and PDFM embeddings provided only incremental gains over baseline building characteristics and simple climate indicators at the building level, consistent with a scale and time-resolution mismatch: embeddings are spatially aggregated and time-invariant, while building EUI contains substantial within-building and temporal variability. However, embeddings were more promising for postcode-level tasks where the prediction target better matches the their granularity, suggesting a practical use-case for screening, benchmarking, and hotspot identification rather than asset-level forecasting. Their failure and modelling results highlighted key findings that modelling energy prediction from diversity-sparse data can be achieved to a reasonable extent, but suffers from noisy point preidctions, and that the variance is dominated by endogeneous factors below the resolution of the embeddings.

Overall, the findings indicate that, in diversity-sparse settings, long-term energy prediction is constrained less by the model choice than by feature availability, and the ability to proxy unavailable determinants. Future improvements are therefore most likely to come from better covariates and from aligning modelling objectives with the spatial and temporal resolutions of available data appropriate.

References

Agarwal, M., Sun, M., Kamath, C., Muslim, A., Sarker, P., Paul, J., Yee, H., Sieniek, M., Jablonski, K., Mayer, Y., Fork, D., Guia, S. de, McPike, J., Boulanger, A., Shekel, T., Schottlander, D., Xiao, Y., Manukonda, M.C., Liu, Y., Bulut, N., Abu-el-haija, S., Perozzi, B., Bharel, M., Nguyen, V., Barrington, L., Efron, N., Matias, Y., Corrado, G., Eswaran, K., Prabhakara, S., Shetty, S., and Prasad, G., 2025. General geospatial inference with a population dynamics foundation model. *Arxiv preprint* [Online], arXiv:2411.07207. [Accessed 21 June 2025]. Available from: `https://arxiv.org/abs/2411.07207`.

Amasyali, K. and El-Gohary, N.M., 2018. A review of data-driven building energy consumption prediction studies. *Renewable and sustainable energy reviews* [Online], 81. [Accessed 10 July 2025], pp.1192–1205. Available from: `https://doi.org/10.1016/j.rser.2017.04.095`.

Berger, M., Mathew, P., and Walter, T., 2016. *Big data analytics in the building industry* [Online]. (technical report LBNL-1005983). [Accessed 10 October 2025]. Lawrence Berkeley National Laboratory. Available from: `https://escholarship.org/uc/item/03m2g5tp`.

Bernard, S., 2025. How we made it: the chances of 2c of global warming within five years. *Financial times* [Online]. [Accessed 15 June 2025]. Available from: `https://www.ft.com/content/364bf19d-3e81-4a16-b55e-212d493f7290`.

Bourdeau, M., Zhai, X., Nefzaoui, E., Guo, X., and Chatellier, P., 2019. Modeling and forecasting building energy consumption: a review of data-driven techniques. *Sustainable cities and society* [Online], 48. [Accessed 10 July 2025], p.101533. Available from: `https://doi.org/10.1016/j.scs.2019.101533`.

Cabeza, L.F., Bai, Q., Bertoldi, P., Kihila, J.M., Lucena, A.F.P., Mata, É., Mirasgedis, S., Novikova, A., and Saheb, Y., 2022. Buildings. In: P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, and J. Malley, eds. *Climate change 2022: mitigation of climate change. contribution of working group iii to the sixth assessment*

report of the intergovernmental panel on climate change [Online]. [Accessed 20 June 2025]. Cambridge and New York, NY: Cambridge University Press, pp.953–1048. Available from: https://doi.org/10.1017/9781009157926.011.

Cabeza, L.F., Űrge-Vorsatz, D., Palacios, A., Ürge, D., Serrano, S., Barreneche, C., and Õrge, S., 2018. Trends in penetration and ownership of household appliances. *Renewable and sustainable energy reviews* [Online], 82. [Accessed 5 September 2025], pp.4044–4059. Available from: https://doi.org/10.1016/j.rser.2017.10.055.

Cai, H., Shen, S., Lin, Q., Li, X., and Xiao, H., 2019. Predicting the energy consumption of residential buildings for regional electricity supply-side and demand-side management. *Ieee access* [Online], 7. [Accessed 11 July 2025], pp.30386–30397. Available from: https://doi.org/10.1109/ACCESS.2019.2901257.

Chen, J., Adhikari, R., Wilson, E., Robertson, J., Fontanini, A., Polly, B., and Olawale, O., 2022a. Stochastic simulation of residential building occupant-driven energy use in a bottom-up model of the u.s. housing stock. *Applied energy* [Online], 325. [Accessed 10 July 2025], p.119890. Available from: https://doi.org/10.1016/j.apenergy.2022.119890.

Chen, Y., Guo, M., Chen, Z., Chen, Z., and Ji, Y., 2022b. Physical energy and data-driven models in building energy prediction: a review. *Energy reports* [Online], 8. [Accessed 10 July 2025], pp.2656–2671. Available from: https://doi.org/10.1016/j.egyr.2022.01.162.

Das, S.S.S., Ali, M.E., Li, Y.F., Kang, Y.B., and Sellis, T., 2022. Boosting house price predictions using geo-spatial network embedding. *World wide web* [Online], 25(5). [Accessed 20 August 2025], pp.1801–1824. Available from: https://doi.org/10.1007/s11280-022-00989-0.

Deng, H., Fannon, D., and Eckelman, M.J., 2018. Predictive modeling for us commercial building energy use: a comparison of existing statistical and machine learning algorithms using cbecs microdata. *Energy and buildings* [Online], 163. [Accessed 10 July 2025], pp.34–43. Available from: https://doi.org/10.1016/j.enbuild.2017.12.031.

Ding, Y., Pan, X., Chen, W., Tian, Z., Wang, Z., and He, Q., 2022. Prediction method for office building energy consumption based on an agent-based model considering occupant–equipment interaction behavior. *Energies* [Online], 15(22). [Accessed 20 July 2025], p.8689. Available from: https://doi.org/10.3390/en15228689.

Fumo, N. and Rafe Biswas, M.A., 2015. Regression analysis for prediction of residential energy consumption. *Renewable and sustainable energy reviews* [Online], 47. [Accessed 10 July 2025], pp.332–343. Available from: https://doi.org/10.1016/j.rser.2015.03.035.

Gao, J., Zhong, X., Cai, W., Ren, H., Huo, T., Wang, X., and Mi, Z., 2019. Dilution effect of the building area on energy intensity in urban residential buildings. *Nature communications* [Online], 10. [Accessed 20 July 2025], p.4944. Available from: https://doi.org/10.1038/s41467-019-12852-9.

Huebner, G.M., 2015. Explaining domestic energy consumption. *Applied energy* [Online], 159. [Accessed 5 September 2025], pp.589–601. Available from: https://doi.org/10.1016/j.apenergy.2015.01.036.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J., 2023. *An introduction to statistical learning: with applications in python* [Online]. [Accessed 10 August 2025]. Cham: Springer. Available from: https://doi.org/10.1007/978-3-031-38747-0.

Kamal, A., Abidi, S.M.H., Mahfouz, A., Kadam, S., Rahman, A., Hassan, I.G., and Wang, L.L., 2021. Impact of urban morphology on urban microclimate and building energy loads. *Energy and buildings* [Online], 253. [Accessed 25 July 2025], p.111499. Available from: https://doi.org/10.1016/j.enbuild.2021.111499.

Kashinath, K., Mustafa, M., Albert, A., Wu, J.-L., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., Manepalli, A., Chirila, D., Yu, R., Walters, R., White, B., Xiao, H., Tchelepi, H.A., Marcus, P., Anandkumar, A., Hassanzadeh, P., and Prabhat, 2021. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences* [Online], 379(2194). [Accessed 20 July 2025], p.20200093. Available from: https://doi.org/10.1098/rsta.2020.0093.

Kashnitsky, Y., 2025a. *Exploratory data analysis with pandas* [Online]. [Accessed 13 October 2025]. Available from: https://mlcourse.ai/book/topic01/topic01_pandas_data_analysis.html.

Kashnitsky, Y., 2025b. *Topic 4. linear classification and regression - part 1: ordinary least squares; mse, likelihood, bias-variance* [Online]. [Accessed 10 August 2025]. mlcourse.ai. Available from: https://mlcourse.ai/book/topic04/topic4_linear_models_part1_mse_likelihood_bias_variance.html.

Kravchenko, A., 2025. *Topic 6: feature engineering & feature selection* [Online]. [Accessed 13 October 2025]. Available from: https://mlcourse.ai/book/topic06/topic6_feature_engineering_feature_selection.html.

Li, Y., Wang, D., Li, S., and Gao, W., 2021. Impact analysis of urban morphology on residential district heat energy demand and microclimate based on field measurement data. *Sustainability (switzerland)* [Online], 13(4). [Accessed 25 July 2025], p.2070. Available from: https://doi.org/10.3390/su13042070.

Lu, C., Li, S., and Lu, Z., 2022. Building energy prediction using artificial neural networks: a literature survey. *Energy and buildings* [Online], 262. [Accessed 11 July 2025], p.111718. Available from: https://doi.org/10.1016/j.enbuild.2021.111718.

Miller, C., Arjukan, P., Kathirgamanathan, A., Fu, C., Roth, J., Park, J.Y., Balbach, C., Gowri, K., Nagy, Z., Fontanini, A., and Haberl, J., 2020. The ashrae great energy predictor iii competition: overview and results. *Science and technology for the built environment* [Online], 26(10). [Accessed 9 July 2025], pp.1427–1447. Available from: https://doi.org/10.1080/23744731.2020.1795514.

Morewood, J., 2023. Building energy performance monitoring through the lens of data quality: a review. *Energy & buildings* [Online], 279. [Accessed 20 August 2025], p.112701. Available from: https://doi.org/10.1016/j.enbuild.2022.112701.

Njimbouom, S.N., Lee, K., Lee, H., and Kim, J., 2022. Predicting site energy usage intensity using machine learning models. *Sensors* [Online], 23(1). [Accessed 12 July 2025], p.82. Available from: https://doi.org/10.3390/s23010082.

Polusmak, E., 2025. *Visual data analysis* [Online]. [Accessed 13 October 2025]. Available from: https://mlcourse.ai/book/topic02/topic02_visual_data_analysis.html.

Potter, B., 2020. *Every building in america – an analysis of the u.s. building stock* [Online]. [Accessed 20 August 2025]. Construction Physics. Available from: https://www.construction-physics.com/p/every-building-in-america-an-analysis.

Potter, B., 2022. *Looking at energy use in us residential buildings* [Online]. [Accessed 21 August 2025]. Construction Physics. Available from: https://www.construction-physics.com/p/looking-at-energy-use-in-us-residential.

Qiao, Q., 2023. *Development of a holistic machine learning-based approach for building energy consumption prediction under limited data conditions* [Online]. [Accessed 10 July 2025]. PhD thesis. University of Manchester. Available from: https://research.manchester.ac.uk/en/studentTheses/development-of-a-holistic-machine-learning-based-approach-for-bui.

Radchenko, V. and Kashnitsky, Y., 2025. *Topic 5 – part 3: feature importance* [Online]. [Accessed 13 October 2025]. Available from: https://mlcourse.ai/book/topic05/topic5_part3_feature_importance.html.

Rafsanjani, H.N., 2016. Factors influencing the energy consumption of residential buildings: a review. *Construction research congress 2016 (crc 2016)* [Online]. [Accessed 20 July 2025]. Reston, VA: American Society of Civil Engineers, pp.1133–1142. Available from: https://doi.org/10.1061/9780784479827.114.

Reinhart, C., 2018. *Environmental technologies in buildings: lecture 2* [Online]. [Accessed 20 July 2025]. Available from: https://ocw.mit.edu/courses/4-401-environmental-technologies-in-buildings-fall-2018/resources/mit4_401f18_lec2/.

Reyna, J., Wilson, E., Parker, A., Satre-Meloy, A., Egerter, A., Bianchi, C., Praprost, M., Speake, A., Liu, L., Horsey, R., Dahlhausen, M., CaraDonna, C., and Rothgeb, S., 2022. *U.s. building stock characterization study: a national typology for decarbonizing u.s. buildings* [Online]. (NREL/TP-5500-83063). [Accessed 21 August 2025]. Golden, CO: National Renewable Energy Laboratory. Available from: https://www.nrel.gov/docs/fy22osti/83063.pdf.

Rodriguez, P.L., Spirling, A., and Stewart, B.M., 2023. Embedding regression: models for context-specific description and inference. *American political science review* [Online], 117(4). [Accessed 20 August 2025], pp.1255–1274. Available from: https://doi.org/10.1017/S0003055422001228.

Sergeyev, D., 2025. *Topic 9 – part 1: time-series analysis in python* [Online]. [Accessed 13 October 2025]. Available from: https://mlcourse.ai/book/topic09/topic9_part1_time_series_python.html#id4.

Setyantho, G.R. and Chang, S., 2020. Identification of primary factors influencing energy consumption patterns of commercial and residential buildings. *Kieae journal* [Online], 20(6). [Accessed 5 September 2025], pp.21–30. Available from: https://doi.org/10.12813/kieae.2020.20.6.021.

Statista, 2025. *Single-family vs multifamily homes in the u.s.* [Online]. [Accessed 20 August 2025]. Available from: https://www.statista.com/statistics/1042111/single-family-vs-multifamily-homes-usa/.

Sun, Y., Haghighat, F., and Fung, B.C.M., 2020. A review of the-state-of-the-art in data-driven approaches for building energy prediction. *Energy and buildings* [Online], 221. [Accessed 11 July 2025], p.110022. Available from: https://doi.org/10.1016/j.enbuild.2020.110022.

THERMOS Project, n.d. *Demand model* [Online]. [Accessed 2 August 2025]. Available from: https://tool.thermos-project.eu/help/demand/demand-model.html.

Tucker, S., 2024. *A systematic review of geospatial location embedding approaches in large language models: a path to spatial ai systems* [Online]. [Accessed 23 August 2025]. arXiv: 2401.10279. Available from: https://arxiv.org/abs/2401.10279.

U.S. Energy Information Administration (EIA), 2023. *Electricity use in homes* [Online]. [Accessed 20 August 2025]. Available from: https://www.eia.gov/energyexplained/use-of-energy/electricity-use-in-homes.php.

U.S. Environmental Protection Agency (EPA), 2025. *Energy star certification* [Online]. [Accessed 10 July 2025]. Available from: https://www.energystar.gov/about/how-energy-star-works/energy-star-certification.

Walter, T. and Mathew, P., 2019. *Is the bpd nationally representative? a comparison of the building performance database to the commercial buildings energy consumption survey* [Online]. (technical report LBNL 2001198). [Accessed 10 October 2025]. Lawrence Berkeley National Laboratory. Available from: https://escholarship.org/uc/item/8m85301c.

Weber, R.E., Mueller, C., and Reinhart, C., 2021. Building for zero: the grand challenge of architecture without carbon. *Ssrn electronic journal* [Online]. [Accessed 11 July 2025]. Available from: https://doi.org/10.2139/ssrn.3939009.

Zhang, Y.-Y., Hu, Z.-Z., Lin, J.-R., and Zhang, J.-P., 2021. Data cleaning for prediction and its evaluation of building energy consumption. *Proceedings of the 38th international symposium on automation and robotics in construction (isarc 2021)* [Online]. [Accessed 9 September 2025]. Tsinghua University / Shenzhen International Graduate School. Available from: https://www.iaarc.org/publications/fulltext/057%20ISARC%2021%20Paper131.pdf.

Zhao, H. and Magoulès, F., 2012. A review on the prediction of building energy consumption. *Renewable and sustainable energy reviews* [Online], 16(6). [Accessed 10 July 2025], pp.3586–3592. Available from: https://doi.org/10.1016/j.rser.2012.02.049.

Zhigulina, A.Y. and Ponomarenko, A.M., 2018. Energy efficiency of high-rise buildings. *E3s web of conferences* [Online], 33. [Accessed 25 July 2025], p.02003. Available from: https://doi.org/10.1051/e3sconf/20183302003.

A Theoretical Details

A.1 Model Specifications

This section includes formal details of employed models. Let y_i denote the observed value, x_i the feature vector for observation i, and $\hat{y}_i$ the predicted value.

a) **Linear Regression (OLS)**
A standard linear model from sklearn API estimating coefficients β by minimising the sum of squared residuals:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{A.1}$$

b) **Ridge Regression**
A linear model with ℓ_2 regularization from sklearn API to prevent overfitting. The coefficients β are estimated by minimising:

$$\mathcal{L}_{\text{ridge}} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{A.2}$$

where $\lambda > 0$ is the regularization parameter.

c) **Random Forest Regression**
An ensemble of T decision trees from sklearn API, where the prediction is the average of individual tree outputs:

$$\hat{y}_i = \frac{1}{T} \sum_{t=1}^{T} f_t(x_i) \tag{A.3}$$

Each f_t is trained on a bootstrap sample with feature subsetting.

d) **XGBoost Regressor**
A gradient boosting tree model from XGBoost package that sequentially fits new

trees to the residuals of previous trees with squared error as the objective:

$$\hat{y}_i^{(m)} = \hat{y}_i^{(m-1)} + \eta\, f_m(x_i), \quad m = 1, \ldots, M \tag{A.4}$$

where f_m is the m-th tree, η the learning rate, and M the total number of trees.

e) **Artificial Neural Network**
A feedforward neural network from Tensorflow package with $L = 4$ layers that applies non-linear transformations through Rectified Linear Unit:

$$\hat{y}_i = f_\theta(x_i) = W_L\, \phi\Big(W_{L-1}\, \phi(\ldots \phi(W_1 x_i + b_1) \ldots) + b_{L-1}\Big) + b_L \tag{A.5}$$

where $\phi(\cdot)$ denotes the ReLU activation function defined as:

$$\phi(z) = \max(0, z) \tag{A.6}$$

Here, W_ℓ and b_ℓ represent the weight matrices and bias vectors of layer ℓ, respectively, and θ denotes the full set of network parameters.

A.2 Hyperparameter Tuning

Bayesian cross-validation search was used. Formally, let $\mathcal{H}$ denote the hyperparameter space and $f : \mathcal{H} \to \mathbb{R}$ a validation performance metric (e.g., cross-validated RMSE). The goal is to find the hyperparameters $h^* \in \mathcal{H}$ that minimize f:

$$h^* = \arg\min_{h \in \mathcal{H}} f(h) \tag{A.7}$$

The Bayesian optimisation iteratively updates a surrogate model $\hat{f}(h)$ and selects new hyperparameters by maximizing an acquisition function $\alpha(h \mid \hat{f})$, which balances exploration and exploitation:

$$h_{\text{next}} = \arg\max_{h \in \mathcal{H}} \alpha(h \mid \hat{f}) \tag{A.8}$$

This process continues until a stopping criterion is reached, such as a maximum number of evaluations or convergence of the validation metric.

- **Number of folds:** 3
- **Number of iterations:** 10
- **Scoring:** negative room mean squared error

A.3 Cross-validation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote the full dataset. The data were split into $K = 5$ folds $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. For each fold $k \in \{1, \ldots, K\}$, the model is trained on the training set $\mathcal{D} \setminus \mathcal{D}_k$ and evaluated on the validation fold $\mathcal{D}_k$.

The cross-validated performance metric was computed as the average validation error across all folds:

$$\mathcal{L}_{\text{CV}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}\left(f^{(-k)}, \mathcal{D}_k\right) \tag{A.9}$$

A.4 Performance Evaluation Metrics

The following error and goodness-of-fit metrics were employed. Let y_i denote the observed value, $\hat{y}_i$ the predicted value, and n the number of observations.

a) **Mean Absolute Error (MAE)**

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{A.10}$$

b) **Root Mean Squared Error (RMSE)**

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{A.11}$$

c) **Coefficient of Determination (R^2)**

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{A.12}$$

where $\bar{y}$ denotes the sample mean of the observed values.

d) **Mean Absolute Percentage Error (MAPE)**

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{A.13}$$

A.5 Model Diagnostics and Uncertainty Quantification

a) **Error by decile of predicted energy use**

Test observations were grouped into deciles based on the predicted value. Let $\hat{y}_i$ denote the predicted EUI for observation i, and y_i the corresponding observed value. The absolute prediction error is defined as

$$e_i = |y_i - \hat{y}_i|. \tag{A.14}$$

Observations were split into $K = 10$ groups $\mathcal{D}_k$ according to the empirical deciles of $\hat{y}_i$. For each decile k, the MAE and RMSE were computed to evaluate heteroskedasticity of model errors.

The errors are summarised in Table A.1.

Table A.1: Prediction error by decile of predicted energy use intensity

Decile	**Mean predicted EUI**	**MAE**	**RMSE**	**Observations**
0	80.44	20.24	28.63	4082
1	98.15	23.61	31.90	4081
2	107.34	26.19	35.90	4081
3	116.76	28.21	38.13	4081
4	129.53	30.09	41.99	4082
5	152.06	36.06	50.20	4081
6	191.73	36.21	51.04	4081
7	233.02	36.32	50.62	4081
8	267.30	33.35	45.07	4081
9	315.00	28.94	37.82	4082

b) **Calibration of residual dispersion**

Prediction uncertainty variance in the magnitude of predicted EUI residual dispersion was analysed conditional on $\hat{y}_i$. The residual is defined as

$$r_i = y_i - \hat{y}_i. \tag{A.15}$$

Test observations were grouped into deciles with the residual standard deviation computed as

$$\sigma_k = \sqrt{\frac{1}{|\mathcal{D}_k| - 1} \sum_{i \in \mathcal{D}_k} (r_i - \bar{r}_k)^2}, \tag{A.16}$$

where $\bar{r}_k$ denotes the mean residual within decile k.

Then σ_k was plotteed against the mean predicted EUI in each decile as a calibration curve for residual spread.

c) **Prediction intervals and empirical coverage**

Prediction intervals were constructed to quantify uncertainty around point predictions. For a trained point prediction model $\hat{f}(x)$

$$\mathrm{PI}(x) = \left[\hat{f}(x) - q,\ \hat{f}(x) + q\right], \tag{A.17}$$

where $q \geq 0$ is a data-driven interval half-width.

The parameter q was obtained using a calibration set that was not used for model training. Let $(x_j^{\mathrm{cal}}, y_j^{\mathrm{cal}})$ denote calibration observations, and define the nonconformity scores as

$$s_j = |y_j^{\mathrm{cal}} - \hat{f}(x_j^{\mathrm{cal}})|. \tag{A.18}$$

For a desired nominal coverage level $1 - \alpha$, the interval width was chosen as the $(1 - \alpha)$ quantile of $\{s_j\}$:

$$q = \mathrm{Quantile}_{1-\alpha}\left(s_1, \ldots, s_{n_{\mathrm{cal}}}\right). \tag{A.19}$$

Empirical coverage on the test set was then defined as

$$\mathrm{Coverage} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathbf{1}\left\{y_i \in \mathrm{PI}(x_i)\right\}, \tag{A.20}$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

B Weather Data API Requests Details

The simplified pseudocode of the algorithm:

Algorithm 1 Weather Data Enhancement Procedure

Require: Dataset with Year, City, County or State ID
Ensure: API Access to NOAA GSOM Dataset

```
for each building b in Dataset do
    if b.city exists and data available then
        Download data                                   ▷ use city data
        continue to next entry
    else if b.county exists and data available and b.city data unavailable then
        Download data                                   ▷ use city data
        continue to next entry
    else if e.state exists and data available and both b.city and b.county data unavailable then
        Download data                                   ▷ use state data
        continue to next entry
    else
        Assign default values for further removal
    end if
end for
```

The following steps describe the required steps to efficiently request all available data:

1. Fetch all available locations and their data and time coverage from the NOAA database.
   Store cities, counties and U.S. states in data object.
2. Process locations data to ensure data objects can be merged successfuly (e.g. Washington has a different code, add county codes to BPD data)
3. Get unique cities, counties and states in the BPD dataset.
4. Find cities, counties and states that are matching in both datasets.
5. Create arrays of unique combinations of cities, counties and years with their minimum and maximum years of data collection.

Then add years to the matching locations objects.

6. Filter the locations in each object for the dates and data coverage.
7. For each building try to find the highest resolution spatial identifier for a given year, i.e. try a matching city, then a county, then a state.
   Store the matching identifier in a *request_id* column.
8. Create a list of unqiue combinations of location identifiers and years.
9. Calculate the number of requests to estimate the request size.
10. (Optional): Get approximate number of requests for each unique location to estimate required time.
11. Split request list into batches of 50 locations.
12. (Optional): Set up other environments to process data in parallel.
13. Request data and process it for storage.
14. (Optional): Due to NOAA API being buggy, consider exports in smaller batches, e.g. every 25 locations.

C Exploratory Data Analysis Details

The full EDA notebook with commentaries and detailed analysis steps is available in the repository or on this link: `https://github.com/dominikmecko/MSc-Thesis/blob/d124fc0a7cb61095546bfe3ec317704fb985011b/src/data-exploration/exploratory-analysis.ipynb`

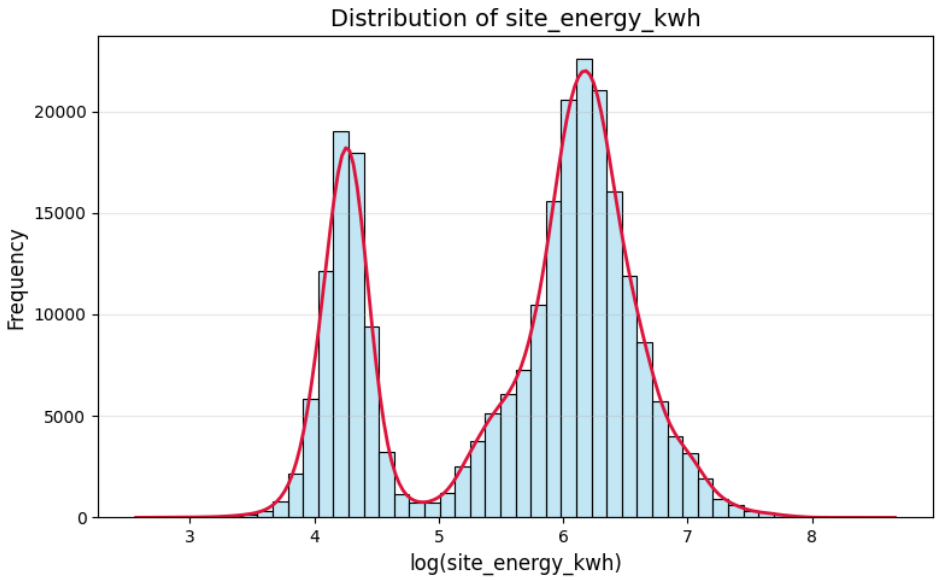
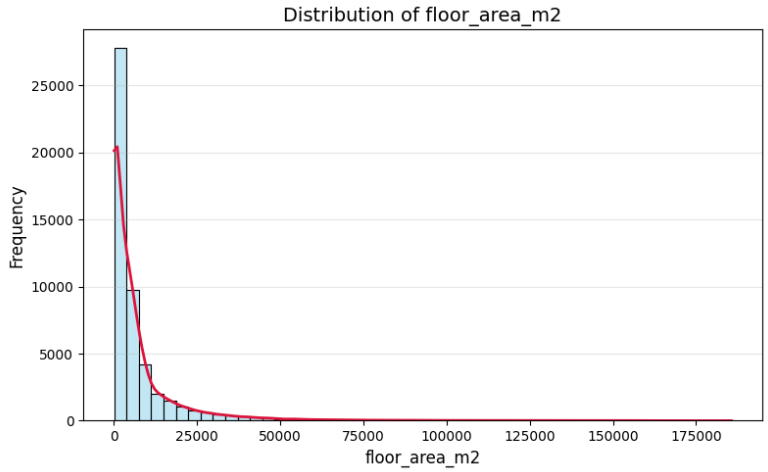
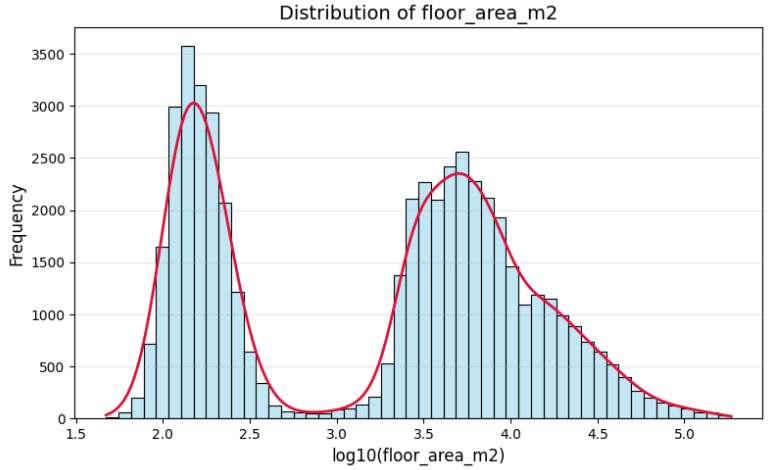


Figure C.1: Distribution of Total Energy (kWh) on a Log-transformed Scale



(a) Original scale



(b) Log-transformed scale

Figure C.2: Distribution of Floor Area (m^2)

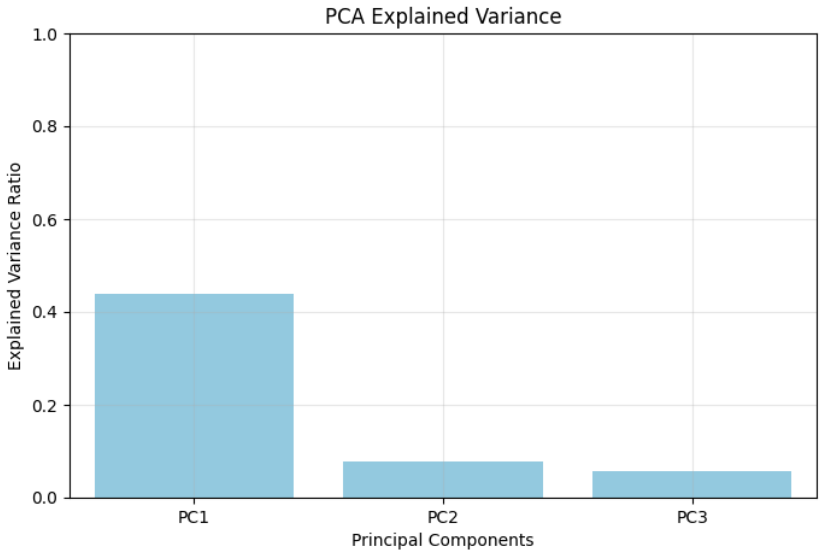


Figure C.3: PCA Variance Explained

Table C.1: Top 5 Contributing Features to Principal Components

Variable	PC1	PC2	PC3
8_DP10	0.0510		
8_DP01	0.0510		
4_HDSD	0.0509		
8_DX90	0.0508		
3_HDSD	0.0508		
2_WDF2		0.1114	
4_WDF5		0.1111	
10_WDF5		0.1078	
2_WDF5		0.1076	
5_DYHF		0.1061	
11_EMSD			0.1369
11_DSND			0.1354
9_AWBT			0.1293
12_ASLP			0.1266
10_DYTS			0.1242

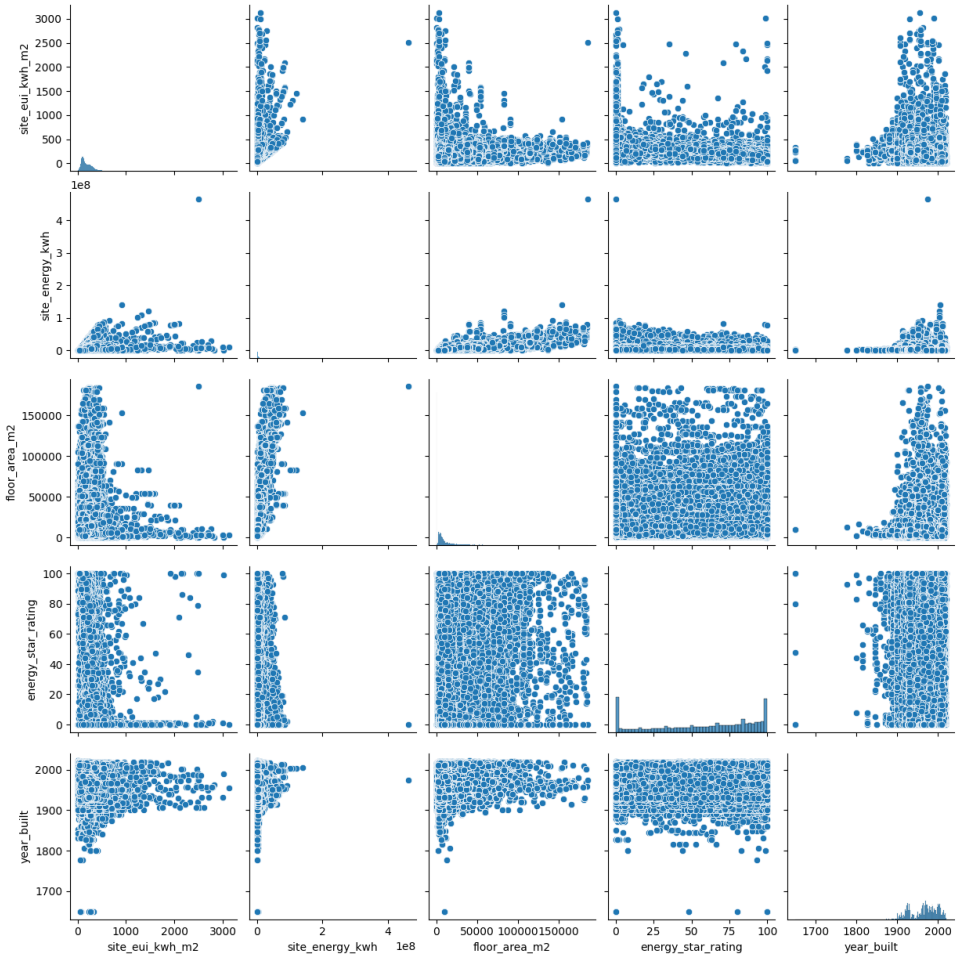


Figure C.4: Pairplot of Selected Variables

Table C.2: Top 10 Positive and Negative Correlations with site_eui_kwh_m2 (filtered)

Top Positive	**Correlation**	**Top Negative**	**Correlation**
fuel_eui_kwh_m2	0.9269	energy_star_rating	-0.6126
ghg_emissions_m2	0.9013	energy_star_rating_2	-0.6187
1_HTDD	0.5346	1_ADPT	-0.5560
3_AWND	0.5122	1_AWBT	-0.5529
2_HTDD	0.5047	1_TMAX	-0.5341
1_AWND	0.4983	1_TAVG	-0.5283
2_DT32	0.4977	2_AWBT	-0.5257
3_HDSD	0.4971	2_ADPT	-0.5198
2_AWND	0.4962	3_ADPT	-0.5132
1_DT32	0.4956	3_AWBT	-0.5129

Table C.3: Summary statistics of site_eui by state

State	**Count**	**Mean**	**Std**	**Min**	**50%**	**Max**
CA	39,192	127.13	82.77	3.15	115.77	1,762.24
CO	5,401	197.61	93.00	26.18	183.91	2,825.06
DC	12,500	190.12	83.77	4.52	178.91	1,065.71
FL	71,569	115.00	46.39	3.16	108.38	529.79
IL	8,941	228.90	108.95	3.96	214.77	2,756.78
MA	3,105	236.46	183.90	3.49	209.04	3,014.14
MD	166	323.92	293.64	8.04	233.70	1,693.15
MO	30	270.27	153.34	18.21	258.84	507.80
NY	86,153	271.49	120.04	3.39	264.32	3,135.00
OR	65	362.25	272.86	50.79	273.82	1,340.39
PA	1,983	200.00	135.73	3.22	174.85	1,630.75
TX	1,612	108.76	31.48	10.71	109.51	2,15.93
WA	13,098	129.48	111.35	9.23	101.28	2,531.32

Table C.4: Summary statistics of site_eui by ASHRAE climate zone

Climate Zone	**Count**	**Mean**	**Std**	**Min**	**50%**	**Max**
2A Hot - Humid	73,181	114.86	46.13	3.16	108.40	529.79
2B Hot - Dry	40	114.40	61.23	20.40	114.04	212.02
3B Warm - Dry	22,809	118.08	61.75	3.15	113.88	1,62.24
3C Warm - Marine	16,245	139.86	104.22	3.15	119.62	1,488.33
4A Mixed - Humid	100,832	260.08	120.36	3.22	253.63	3,135.00
4B Mixed - Dry	76	133.75	64.07	30.64	139.44	286.37
4C Mixed - Marine	13,163	130.63	113.87	9.23	101.47	2,531.32
5A Cool - Humid	12,046	230.85	132.42	3.49	213.55	3,014.14
5B Cool - Dry	5,423	197.28	93.07	8.97	183.71	2,825.06

D Data Sources and Descriptions

D.1 Data Sources and Collection

D.1.1 Building Performance Dataset

The datasets were downloaded and processed as individual `.csv` files obtained via a web application interface. The web application is available at `https://bpd.lbl.gov/`. Access requires user authentication; however, at the time of writing, the data were available free of charge upon provision of the required registration details. The datasets are also accessible programmatically through an API.

D.1.2 National Oceanic and Atmospheric Administration

Weather data were obtained through an API of the NOAA Global Summary of the Month Dataset. The documentation is available at `https://www.ncdc.noaa.gov/cdo-web/webservices/v2`. Access requires authentication; at the time of writing, the access token was provided free of charge upon provision of required details.

D.1.3 PDFM Embeddings

PDFM Embeddings were provided by Google after filling a request form available at `https://github.com/google-research/population-dynamics?tab=readme-ov-file`.

D.2 Loading of the Data

The data were loaded from processed `.csv` files. Due to data providers terms and conditions, original data cannot be distributed but a synthetic sample containing 1,000 rows and randomly generated data of numerical variables preserving mean and st.deviation is stored in `processed` folder in the repository.

Table D.1: Baseline Variables from BPD Descriptions

Variable	**Decription**
id	Building ID
year	Year of data record
zip_code	Postcode of the building
city	City of the building
state	State of the building
climate	ASHRAE climate code of the building
facility_type	Build type of the building (e.g. single/multi family).
floor_area_m2	Gross floor are in m2
year_built	Year in which the building was constructed
energy_star_rating	Energy Star rating of the building
electric_eui	EUI of electricity in kWh/m2/year
fuel_eui	EUI of fuel use in kWh/m2/year
site_eui	Total site EUI of the building in kWh/m2/year
ghg_emissions_int	GHG emissions of the building
population	Population of the postcode area

Table D.2: Engineered Features for Analysis

Variable	**Formula / Computation**
electricity_fuel_ratio	$\frac{\text{electric\_eui\_kWh/m}^2}{\text{fuel\_eui\_kWh/m}^2}$, replace ∞ with NaN
age_floor_area	age $\times$ floor_area$_{\text{m}^2}$
age_2	$(\text{age})^2$
energy_star_rating_2	$(\text{energy\_star\_rating})^2$
log_floor_area	$\log_{10}(\text{floor\_area}_{\text{m}^2})$
climate_code	$\begin{cases} 1, & \text{if building belongs to category } i \\ 0, & \text{otherwise} \end{cases}$
facility_type	$\begin{cases} 1, & \text{if building belongs to category } j \\ 0, & \text{otherwise} \end{cases}$

Table D.3: Weather variables descriptions

Variable	Decription
ADPT	Monthly Average Dew Point Temperature. Average of daily dew point temperatures given in Celsius or Fahrenheit depending on user specification. Missing if more than 5 days within the month are missing or flagged or if more than 3 consecutive values within the month are missing or flagged. DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
ASLP	Monthly Average Sea Level Pressure. Average of daily sea level pressures given in hPa or inches of mercury depending on user specification. Missing if more than 5 days within the month are missing or flagged or if more than 3 consecutive values within the month are missing or flagged. DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
ASTP	Monthly Average Station Level Pressure. Average of daily station level pressures given in hPa or inches of mercury depending on user specification. Missing if more than 5 days within the month aremissing or flagged or if more than 3 consecutive values within the month are missing or flagged.DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
AWBT	Monthly Average Wet Bulb Temperature. Average of daily wet bulb temperatures given in Celsius or Fahrenheit depending on user specification. Missing if more than 5 days within the month are missing or flagged or if more than 3 consecutive values within the month are missing or flagged. DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
AWND	Average wind speed
CDSD	Cooling Degree Days Season to Date
CLDD	Cooling Degree Days
DP01	Number of days with greater than or equal to 0.1 inch of precipitation
DP10	Number of days with greater than or equal to 1.0 inch of precipitation
DP1X	Number of days with >= 1.00 inch/25.4 millimeters in the month
DSND	Number days with snow depth > 1 inch(25.4mm) for the period.
DSNW	Number days with snow depth > 1 inch.
DT00	Number days with minimum temperature less than or equal to 0.0 F
DT32	Number days with minimum temperature less than or equal to 32.0 F
DX32	Number days with maximum temperature < 32 F.
DX70	Number days with maximum temperature > 70 F (21.1C)
DX90	Number days with maximum temperature > 90 F (32.2C)

Table D.4: Weather variables descriptions (cont.)

Variable	Description
DYFG	Number of Days with Fog
DYHF	Number of Days with Heavy Fog (visibility less than 1/4 statute mile)
DYNT	Day Extreme minimum temperature occurred for the period.
DYSD	Day the extreme maximum daily Snow Depth for the period occurred.
DYSN	Day the extreme maximum daily snowfall for the period occurred.
DYTS	Number of Days with Thunderstorms
DYXP	Day Extreme maximum daily precipitation for the period.
DYXT	Date Extreme maximum temperature occurred for the period.
EMNT	Extreme minimum temperature for the period.
EMSD	Extreme maximum snow depth for the period.
EMSN	Extreme maximum snowfall for the period.
EMXP	Extreme maximum precipitation for the period.
EMXT	Extreme maximum temperature for the period.
HDSD	Heating Degree Days Season to Date
HTDD	Heating degree days
PRCP	Precipitation
RHAV	Monthly Average Relative Humidity. Average of daily relative humidity values given in percent. Missing if more than 5 days within the month are missing or flagged or if more than 3 consecutive values within the month are missing or flagged. DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
RHMN	Monthly Average of Minimum Relative Humidity. Average of daily minimum relative humidity values given in percent. Missing if more than 5 days within the month are missing or flagged or if more than 3 consecutive values within the month are missing or flagged. DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
RHMX	Monthly Average of Maximum Relative Humidity. Average of daily maximum relative humidity values given in percent. Missing if more than 5 days within the month are missing or flagged or if more than 3 consecutive values within the month are missing or flagged. DaysMissing: Flag indicating number of days missing or flagged (from 1 to 5).
SNOW	Snowfall
TAVG	Average Temperature.
TMAX	Maximum temperature
TMIN	Minimum temperature
WDF2	Direction of fastest 2-minute wind
WDF5	Direction of fastest 5-second wind
WSF2	Fastest 2-minute wind speed
WSF5	Fastest 5-second wind speed

Table D.5: ASHRAE Climate Zones in the United States (ASHRAE Standard 169)

Climate Code	**Description**
1A	Very Hot, Humid, e.g. southern Florida
1B	Very Hot, Dry. e.g. southern Arizona and southeastern California.
2A	Hot, Humid. e.g. U.S. Gulf Coast and southeastern states
2B	Hot, Dry. e.g. desert regions
3A	Warm, Humid. e.g. southeastern U.S. and parts of the Mid-Atlantic
3B	Warm, Dry. e.g. parts of California and the Southwest
3C	Warm, Marine. e.g. coastal California, Oregon, and Washington.
4A	Mixed, Humid. e.g. Midwest and Northeast.
4B	Mixed, Dry. e.g. parts of the interior western U.S. such as Colorado and Utah.
4C	Mixed, Marine. e.g. Pacific Northwest coastal regions.
5A	Cool, Humid. e.g. northern Midwest and Northeast regions such as Minnesota and Maine.
5B	Cool, Dry. e.g. interior Northwest and high plains regions.
5C	Cool, Marine. Rare in the U.S.; limited coastal regions.
6A	Cold, Humid. Much of the northern U.S. such as Wisconsin and Michigan.
6B	Cold, Dry. Includes high-altitude western regions.
7	Very Cold. Includes northern Minnesota and parts of Alaska.
8	Subarctic. Includes interior Alaska.

D.3 Data Descriptions

D.4 Data Cleaning and Processing

D.4.1 Pre-Processing Steps

a) **Missing values handling**

- Placeholder labels (*No Value*, *Unknown*, *NaN*, *Null*) were converted to `np.nan` for imputation.
- Variables with more than 55% missing observations were removed.

b) **Data type standardisation**
All numerical variables were cast to `float32`.

c) **String normalisation**
Climate codes and facility-type labels were shortened to improve readability and debugging.

d) **Unit translation and feature construction**

- Floor area was converted from square feet to square metres:

$$\texttt{floor\_area\_m2} = \texttt{floor\_area} \times c_{\mathrm{ft}^2 \to \mathrm{m}^2}$$

- Total site energy consumption was computed and converted to kilowatt-hours:

$$\texttt{site\_energy\_kWh} = (\texttt{floor\_area} \times \texttt{site\_eui}) \times c_{\mathrm{kBTU} \to \mathrm{kWh}}$$

- Energy use intensities were expressed in $\mathrm{kWh/m^2}$:

$$\texttt{site\_eui\_kWh\_m2} = \frac{\texttt{site\_energy\_kWh}}{\texttt{floor\_area\_m2}}$$

$$\texttt{fuel\_eui\_kWh\_m2}, \texttt{electric\_eui\_kWh\_m2} = \texttt{EUI} \times c_{\mathrm{kBTU/ft}^2 \to \mathrm{kWh/m}^2}$$

- Greenhouse gas emissions were normalised by floor area:

$$\texttt{ghg\_emissions\_m2} = \frac{\texttt{ghg\_emissions\_int} \times \texttt{floor\_area}}{\texttt{floor\_area\_m2}}$$

- Building age was derived as the difference between the observation year and

the year of construction:

$$\texttt{age} = \texttt{year} - \texttt{year\_built}$$

D.4.2 Modelling Transformations

a) **Imputation**: The imputation strategy was median imputation for all missing values besides `site_eui_kwh_m2` and `embeddings`.

b) **Scaling**: All numerical columns except `embeddings` were scaled using the standard scaler of the sklearn API with this formula:

$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sigma}$$

D.4.3 Train-Test splits

Custom function designed to avoid data leakage:

Let $\mathcal{D} = \{(x_i, y_i, b_i)\}_{i=1}^{n}$ denote the full dataset, where $b_i \in \mathcal{B}$ represents the building identifier. A fraction $s \in (0, 1)$ of buildings is randomly assigned to the training set using a fixed random seed to ensure reproducibility:

$$|\mathcal{B}_{\text{train}}| = \lfloor s \cdot |\mathcal{B}| \rfloor \tag{D.1}$$

The remaining buildings form the test set:

$$\mathcal{B}_{\text{test}} = \mathcal{B} \setminus \mathcal{B}_{\text{train}} \tag{D.2}$$

The corresponding index sets are defined as:

$$\mathcal{I}_{\text{train}} = \{i \mid b_i \in \mathcal{B}_{\text{train}}\} \tag{D.3}$$

$$\mathcal{I}_{\text{test}} = \{i \mid b_i \in \mathcal{B}_{\text{test}}\} \tag{D.4}$$

E Technical Specifications

E.1 Code

All code required to obtain weather data, run the EDA and modelling is stored in a GitHub repository available at `https://github.com/dominikmecko/MSc-Thesis-Public`. The repository will remain uploaded and maintained for at least 5 years.

E.2 Reproducibility

To ensure reproducibility random seeds were generated and used in all steps with random values:

a) **Synthetic Data Generation**: Random seed: 42

b) **Train-test Split Function**: Random seed: 3

c) **Cross Validation**:

- Initial random seed: 3
- CV seeds: random sample of $\{1, 2, 3, \ldots, 10000\}$ with a random seed 5

d) **Modelling Seeds**:

- All models random state: 3
- Train-test split seed sampled from random sample of $\{1, 2, 3, \ldots, 10000\}$ with a random seed 1
- Hyperparameter tuning seed sampled from random sample of $\{1, 2, 3, \ldots, 10000\}$ with a random seed 3
- Bayesian search model random state: 42
- Comparison of hyperparameter performance model random state: 42
- Subset predictive effects model random state: 42

E.3 Environment Requirements

The full environmental requirements with versions are stored in a `environment.yaml` file in the repository. The basic libraries in the environment are as follows:

- **Core:** Python 3.10
- **Computation**: numpy, pandas
- **Modelling**: sklearn, tensorflow, xgboost
- **Visualisation**: mathplotlib, seaborn,
- **Computation**: numpy, pandas

E.4 Hardware

Modelling was done on a computer with the following specifications:

- **CPU:** Apple M2 Pro Silicion
- **RAM:** 16GB
- **Operating system:** macOS Version 26.1 (25B78)

E.5 Approximate Computational Requirements

a) **Weather Data Collection**: Three parallel environments with three tokens ≈ 2-3 hours.

b) **Model Training:**

- **Individual Baseline Model Times:** Ranging from a few seconds to a few minutes.
- **Cross-validation:** ≈ 2 hours.
- **Hyperparameter Tuning:** ≈ 3 hours.
- **Subset Comparison:** ≈ 30 minutes.

Glossary

Annual energy consumption (site) Total annual energy consumed at the building/site across energy sources; expressed in kWh, combining electric and thermal demand after unit conversion.

Energy Use Intensity (EUI) Annual energy consumption normalised by floor area; expressed as kWh/m^2/year.

Site EUI vs Source EUI Site EUI counts energy consumed at the building, while source EUI includes upstream generation/transmission losses; site EUI refers to EUI in this paper.

Gross floor area / Floor area Area metric used to compute EUI; serves as the area basis for normalisation. This paper uses groos floor area and floor area interchangeably. Gross floor area refers to the total amount of floor space measured from outside the external walls.

Features / variables / covariates / columns All terms refer to available independent variables present in the dataset and are used interchangeably throughout the paper.

Baseline variables (BPD extract) Core columns listed in Appendix Table D.1 (e.g. building characteristics).

Engineered features Derived variables listed in Appendix Table D.2.

Building age Derived as observation year minus year built.

Imputation Missing-value approach: median imputation for missing values (exceptions like target and embeddings).

Scaling (standardisation) Normalisation of data to a fixed interval.

Train–test split without leakage Splitting of data into the set for training the ML model and evaluating on the test split. The split is done by the building identifier so the same building does not appear in both train and test sets.

Weather variables (NOAA GSOM Dataset) Monthly aggregates of observations used

as features, e.g., TAVG/TMAX/TMIN, PRCP, RH*, degree days, wind, snow/fog/thunderstorm day counts. Definitions in Table D.3 and Table D.4.

ASHRAE climate zones Climate-zone categories defined by the The American Society of Heating, Refrigerating and Air-Conditioning Engineers. Full definitions available at `https://openei.org/wiki/ASHRAE_Climate_Zones`

Embedding(s) Matrix of vector representations used as features; in this paper referred to PDFM embeddings used as additional input.

Regularisation (ridge/lasso) Linear-model penalty approaches to reduce overfitting and aid feature selection.

Random forest regression Ensemble of decision trees whose predictions are averaged.

Gradient boosting / XGBoost Sequential tree-ensemble method fitting trees to residuals of previous trees.

Bayesian hyperparameter optimisation Hyperparameter search approach using a surrogate model and acquisition function.