Energy Forecasting in Commercial Buildings Using Property Features and Natural Resources

MSc Research Project
Data Analytics

Basil Varghese
Student ID: X23213574

School of Computing
National College of Ireland

Supervisor: Dr. Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Basil Varghese
Student ID:	X23213574
Programme:	Data Analytics
Year:	2024 -2025
Module:	MSc Research Project
Supervisor:	Dr. Noel Cosgrave
Submission Due Date:	12/12/2024
Project Title:	Energy Forecasting in Commercial Buildings Using Property Features and Natural Resources
Word Count:	9506
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Basil Varghese
Date:	11th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Energy Forecasting in Commercial Buildings Using Property Features and Natural Resources

Basil Varghese
X23213574

Abstract

This study aimed at using machine learning algorithms to forecast energy consumption in commercial buildings based on CBECS 2018 data. Principal component analysis brought the feature sets down from 1249 to 257 components while retaining 95% of the data variance. Four types of machine learning was used to the target variables with logarithmic transformation. The best hyper tuning done using RandomizedSearchCV, although applying 500 fits and determining the hyperparameter tuning with an MSE of 0.8954 and an MAPE of 5.67% was the XGBoost model applied with a learning rate of 0.1,a max depth of 3,and 300 no. of estimators out of 50 fits. Random Forest, tuned via RandomizedSearchCV across 50 fits, showed competitive results (MSE: 0.75, RMSE: 0.8651). Linear and Ridge Regression, two main models, gave baseline performance with an MSE of 0.9280 and 0.9314 respectively. This study also found that as far as renewable energy is concerned, natural gas is the most used source (70.6% of adoption across buildings), but electricity remains the most accepted traditional resource, Adoptions of renewable energy sources remain restricted to solar energy source, which was 4.4%. The study has shown that complex pattens of energy consumption are better represented by ensemble methodologies and that the penetration of renewable energy has high potential in existing commercial spaces.

Keywords: Commercial buildings, Machine learning algorithms, Principal Component Analysis, CBECS 2018

1 Introduction

Energy consumption forecasting in various sectors, particularly commercial buildings,is fundamental in managing commercial buildings and enhancing construction sustainability. The CBECS 2018 reveals that there are approximately 5.9 million buildings, mainly in the United States, that use about 6.8 quads of energy. The global commercial building energy consumption cost rose by 14% to about $141 billion from 2012 to 2018.The high levels of required energy control and minimization of energy consumption created high operational costs and reduced monetary effectiveness. There was an increasing need to consider the negative consequences of burning fossil fuels and the emissions of carbon dioxide (Ahmad et al.; 2019). The COVID-19 pandemic has further complicated this landscape, causing fluctuations in energy usage patterns due to reduced occupancy, while simultaneously introducing new operational procedures that has increased energy consumption (Li et al.; 2023). The Energy Use Intensity (EUI) has shown improvement through benchmarking of the buildings and green ratings such as LEED and ENERGY

STAR, though significant challenges remain in modifying existing structures and financing improvements.

There are many reasons why energy use in commercial buildings is not constant throughout the year. These factors include the size of the building, its age, and the usage patterns of the occupants, which involve the types of heating, ventilation, and air conditioning systems are fundamental to energy consumption (Deng et al.; 2014). The combination of natural resource information with current property characteristics offers the benefit of improving the assessment of energy use, and subsequently, a possible improvement in energy management and conservancy in existing buildings. These are some of the challenges that current prediction methods do not capture well enough, making it crucial to align building operational requirements with sustainable energy management. Using data on the quantity of natural resources with the features of the property offers a main possible approach to address this gap and potentially will lead to enhanced antecedent and more building energy efficiency.

To these challenges, it is possible to respond that one of the most promising developments is the integration of natural resource information with the characteristics of the property to improve energy management strategies. This approach could also provide possible solutions to some of the current problems derived from the traditional method of energy prediction for buildings, which fails to fully incorporate the dynamics of building energy systems (Hong et al.; 2017). Additionally, the fast pace of development of smart technologies applied to buildings, increased pressure from regulations, and rising expectations of consumers and investors to reduce energy consumption and greenhouse emissions have prompted both opportunities and challenges for new energy management solutions.

Most research has utilized the 2012 CBECS dataset, with little to no research conducted using the 2018 dataset. The 2018 survey possesses improved technology, more accurate energy usage tracking, and enhanced data collection methods compared to the 2012 survey. It underwent several of notable revisions, with the following key modifications: modified variables, such as PUBCLIM (climate zone - updated climate zone classifications are now based on ASHRAE climate zone standards), and recoded NFLOOR (number of floors); removed several activity-related variables and building characteristic variables.It also highlights changes in fuel sources, emissions data, and building types to better align with modern trends. Major revisions were implemented in the sections related to heating and cooling systems as well.

The importance of this research will be particularly significant given the present trends in automating building systems, increased environmental conservation laws, and necessities for sustainable construction projects (Raza and Zhong; 2024). In light of the changing use and design of commercial buildings adapting to smart technologies and changing operational paradigms, the requirement for complex energy-predicting strategies is crucial. The difficulty lies not only in identifying present energy usage but also in designing heuristics to evaluate future alterations in building utilization and emerging technologies (Miller et al.; 2022).

1.1 Objectives

According to the provided gaps in commercial building energy consumption, this research will endeavour to develop machine learning and statistical analyses to forecast energy use patterns. The dataset used in this study is CBECS 2018, and the data will undergo data cleaning and data pre-processing programmes before being used in modeling. Three different models will be trained and tested: XG Boost, which is used to fit models for multivariate non-linear econometric relationships; Random Forest, which is used to fit models to capture interactions between features; and Generalized Linear Model (GLM), to fit models with linear relationships other than linear regarding building energy consumption. The MSE, RMSE, MAE, and R^2 value of each model will be computed on a cross-validation set to compare and pass tolerance. Lastly, the study will evaluate the applicability of these models to more fully capture energy use characteristics by building type and climate region, as well as offer constructive recommendations for improving building energy efficiency and decision-making.

1.2 Motivation and Purpose

Commercial buildings have inherent difficulty in controlling and forecasting energy use due to various interrelations between architectural features, occupancy routines, and climatic conditions. The aim of this study is to assist building managers and investors in improving energy consumption prediction models based on building features, natural sources, and various operational parameters. Most facility managers consider it necessary to effectively manage the energy consumed by buildings, as it is more important than simply identifying its utilization. In response, studying patterns of commercial building energy consumption is crucial for analysis using the CBECS 2018 dataset, and this work investigates commercial building energy consumption patterns in various building types and climate regions.

Most organizations currently believe that achieving the highest level of energy efficiency in commercial facilities is nearly impossible due to numerous interactions. However, considering aspects that determine energy usage can simplify this task. Therefore, the objective of this paper is to develop energy consumption profiling models using XG Boost, Random Forest,Linear regression and Ridge Regression methodologies to assist in energy efficiency management.

Altogether, the choice of these models offers unique advantages for studying building energy use. XG-Boost is highly capable of handling non-linear dependencies whereas Random Forest strongly performs in terms of feature dependency and relevance. Linear Regression induces easily interpretable starting-line models and Ridge Regression presents a solution to multi-colinearity issues regarding building characteristic, also preserving stability in predictions. In this regard, the proposed model-integrated approach combines a high modeling capacity to represent complex facility interactions, while at the same time providing sufficient interpretability for facility managers.

In addition, this paper seeks to improve knowledge of the factors thought to greatly influence Building Energy Consumption (BEC), making energy use clearer and easier to understand at every stage of the process. Finally, this work aims to inform building managers in their decisions on how to manage energy in buildings. The goal of this work is to improve the perfomance of energy predictions for commercial buildings and decrease the complexity of applying energy conservation measures.

1.3 Research Question

How does the integration of the renewable energy sources and building properties improve the performance of energy consumption forecasting in commercial buildings?

1.4 Research Objectives and Deliverables

1. To analyze the significance of the CBECS dataset and to perform the pre-processing.
2. In order to reduce the factors which impact on energy consumption estimation by applying the Dimensionality Reduction technique PCA.
3. To estimate energy usage in commercial buildings based on characteristics of the property and natural resources, four algorithms Random Forest, XG-Boost, Logistic Regression, Ridge Regression.
4. Performance Analysis and Comparing the most finest algorithm for energy consumption prediction

2 Related Work

This section of the report examines existing research on the application of machine learning and deep learning in forecasting energy consumption in commercial buildings, focusing on the studies that investigate aspects of energy usage prediction, with an emphasis on methodology, results, and identified weaknesses, which will be discussed in the following three sub-sections.

2.1 Evaluation of Previous Studies for Energy Consumption Prediction in Commercial Buildings using Renewable and Non-renewable sources.

The 2012 Commercial Building Energy Consumption Survey (CBECS) microdata is used by (Deng et al.; 2018) to experiment on various predictive modeling approaches to calculate Energy Use Intensity (EUI) for commercial buildings. The research involves preliminary statistical analysis which is followed by the application of different regression and machine learning techniques. These models are then compared to find their effectiveness in predicting energy use.

Among the techniques employed, Support Vector Machine and Random Forest models gives both accuracy and stability in their predictions. But the results show that machine learning algorithms only minorly outperform traditional linear regression models, with a 10-15% reduction in prediction errors for Total EUI. Meanwhile linear regression models perform better in estimating Plug Loads EUI. The paper also points out that the variables available in CBECS might not have required predictive power to precisely map actual energy consumption. Therefore, filling information gaps related to occupant behavior, power management, and building thermal performance could enhance predictive modeling efforts.

(Bourdeau et al.; 2019) also reflects an understanding that human factors play a critical role in energy consumption patterns and should be considered in forecasting models. The paper points out the need for effective energy management strategies as the urban populations are rising and so the energy demands. The paper focuses on different methods in building energy consumption modeling. A thorough examination of the methodologies currently in use and their effectiveness was also done by the authors. A significant aspect mentioned is the integration of occupants' behavior into data-driven modeling.

(Seyedzadeh et al.; 2018) in his study talks about various unsupervised and machine learning techniques. Neural networks are one of the key Machine Learning techniques for energy calculation and have been successfully applied to model non-linear problems and complex systems. At the same time, statistical models offer superior performance as they can be trained with a limited number of samples, making them appropriate for modelling study cases with no historical data.

(Dinh et al.; 2023) find multivariate multilayered long short-term memory (LSTM) as an appropriate prediction method, comparing its performance against methods such as linear regression, bidirectional LSTM, and LSTM. This study uses a random forest model to predict complex behaviour because of its learning capacity.

Gradient Boosting that combines different models that are comparatively weak to create a robust predictive model by recognising and rectifying errors is another technique. Gradient Boosting strengthens prediction methods and bring down errors by harmonising weak models with the loss function. A variety of gradient boosting called XGBoost, which is a versatile ensemble technique and one of the most productive machine-learning methods.

With the popularity of renewable energy, (Benti et al.; 2023) reviews the models and approaches that have been employed in renewable energy forecasting, like lack of certainity energy generation, the interpretability of models, data availability etc. Natural gas is a primary heating source in most buildings, majorly in low temperature climatic conditions. (Kamath; 2020) predicts natural gas heating use in buildings using electricity consumption data from the CBECS dataset. He built linear models to predict natural gas heating energy intensity (EI) based on cooling electric EI and electric energy use intensity (EUI). Despite his efforts, the results were disappointing, with R^2 values never exceeding 0.1.

2.2 Assessing the Effectiveness of Different Methods for Predicting Energy Consumption.

Forecasting performances of data-driven algorithms are tested using accuracy metrics. The most common are the mean absolute percentage error (MAPE), the root mean square error (RMSE), and the mean average error (MAE).

(Bourdeau et al.; 2019) provides a thorough review of various data-driven building energy modeling techniques, including autoregressive models (AR), statistical regressions, k-nearest neighbors (k-NN), decision trees (DT), support vector machines (SVM), and artificial neural networks (ANN). The paper discusses how understanding human factors can enhance the accuracy of energy consumption forecasts, which is often overlooked in traditional models. The paper focuses on the progresses made in integrating traditional methods with machine learning techniques. The AR model proposed by (Bourdeau et al.; 2019) shows the best MAE of 2.01 kW. OLS performances ranged between a MAE of 2.05 kW for all variables and a MAE of 3.74 kW with temperature only. SVM performances ranged between a MAE of 1.94 for all variables and "recency" only, and a MAE of 3.46 kW with temperature only. Results for electricity consumption forecasting showed that the DT performed slightly better for summertime with RMSE of 39.36 kWh, compared to the ANN (RMSE of 39.53 kWh) and the regression (RMSE of 39.42 kWh). However, higher accuracy was achieved for wintertime with the ANN (44.14 kWh) and with the regression (44.18 kWh) than with DT (44.40 kWh).

(Tsanas and Xifara; 2012) compared an iterative reweighted least square regression method with a RF model for forecasting heating and cooling loads of housing properties simulated using the Ecotect tool, utilising a database of eight passive systems variables. Results shows that RF outperformed the regression, with a MAE of 0.51 kW/1.42 kW and 2.14 kW/2.21 kW, for heating/cooling loads and for both models, respectively. For cooling load assessment results (MAPE), methods ranked as follows: SVM with 2.99%, ANN with 4.40% and regression with 4.96%. For heating load (MAPE), the method ranking is SVR with 1.13, ANN with 2.36% and regression with 4.59%.

(Kamath; 2020) in his findings indicate that certain building types, such as retail and education buildings, exhibit higher levels of distinguishability in energy use. In contrast, food service, food sales, and inpatient healthcare buildings show the highest mean EUIs. The author employing the Jackknife method extracts meaningful information from the noisy dataset. The $\mathbf{R^2}$ values from the linear models never exceeding 0.1, points to a very low predictive power for the models implying that the models might showcased a good performance to produce a promising MAPE, RMSE, or MAE values.

(Fu et al.; 2021) indicates that while artificial neural networks (ANNs) have been widely used and can provide accurate predictions, cleverly assembled statistical regression methods have outperformed ANNs in certain cases. This suggests that traditional statistical approaches still hold significant value in energy modeling.

Gradient boosting regressor (GBR) quite often yields the best outcomes and gradient descent optimization provides more accurate predictions. The extreme gradient boosting (XGBoost) algorithm performs a sampling of a subset of data, fitting a single predictor to minimize the loss function. Thus, XGBoost aims to accelerate decision tree training. Since finding the best data split takes up the most time, XGBoost simplifies the process by removing the need to manually determine the optimal data split.

(Deng et al.; 2018) points out that Random Forest Models and Support Vector Machine (SVM) exhibit stability and accuracy in predicting energy use. The authors employ Random Forest to understand the role of individual variables in predicting energy consumption and finds that the variables reported in the CBECS dataset does not possess adequate predictive power, pointing to the importance of factors like building thermal performance and occupant behaviour.

In predicting individual energy end-uses, linear regression showed improved performance than some advanced machine learning methods. The analysis of the importance of variables by (Deng et al.; 2018) shows none of the predictors have a better influence on Total EUI, as all predictors show less than 15% mean decrease in accuracy. The CBECS data was randomly split into three portions: 50% for training, 25% for validation, and 25% for testing. The validation set was primarily used to tune the parameters of each regression model, while the testing set was used to calculate the final MAE for each model. The paper included both RMSE and MAE. The results indicated that different models exhibited varying levels of MAE and RMSE. The error terms were normalized to compare these values by showing the errors as a percentage of the mean of the real data, such that the forecasting errors were made scale free.

(Benti et al.; 2023) points out involving external data sources like grid data and weather. In case of renewable energy forecasting, an SVR model performed better on the validation dataset with an MAE of 32.57 and RMSE of 43.16. The ANN model gives a MAPE of 16.45%, proving its efficiency in predicting solar radiation. Meanwhile, the RMSE of Random Forest Regression model gives an RMSE of 86 and an MAE of 69. Therefore, SVR performed better than RFR in terms of both RMSE and MAE proving to be more effective in renewable energy forecasting.

Meanwhile, (Han et al.; 2018) points out that random forest outperforms the comparative classifiers in terms of robustness to features, accuracy and stability. Random Forest is less susceptible to environmental noise. The author says that random forest is a suitable pattern recognition method for the intelligent diagnosis of rotating machinery.

2.3 Feature Analysis and Identification of Gaps in different approaches.

(Kamath; 2020) through a series of subsets to build linear models tried to predict natural gas heating energy intensity (EI). He progressively excluded buildings that did not meet specific criteria, such as those that heated less than 90% of their space or did not use electricity for cooling. This rigorous data cleaning process reduced the sample size to 152 buildings, which was crucial for ensuring the reliability of the results. The linear models built by Kamath (2020) to predict natural gas heating energy intensity (EI) using cooling electric EI or electric EUI as predictors yielded very low R^2 values, never exceeding 0.1 which points to the need of a better and strong relationship between the variables analyzed. Therefore, he used the Kaiser-Meyer-Olkin (KMO) test, giving a value of 0.5, pointing that the dataset is "miserable" for factor analysis. Therefore, he suggested modifications to the dataset, like a larger sample size, a greater number of predictors etc. Predictors like system type, occupant behavior, building envelope insulation, lighting wattage and detailed climatic information are put forth by the author.

The problems linking to data quality is also discussed by (Benti et al.; 2023). According to the authors many ML and DL models operate as "black boxes," implying the difficulty in analyzing the approach to the predictions. Advanced models that can simultaneously work with several renewable energy sources is an area to be researched on. More advanced models like random forests, support vector machines (SVMs), and XGBoost shows better. Also, many ML and DL models are not transparent enough to analyse their predictions making this an area for future research.

(Fu et al.; 2021) talks about the lack of literature on including occupancy as an independent variable pointing to the need more research in this area. With a high-quality data, statistical models could be further developed, including domestic water consumption and electric demand. M. (Bourdeau et al.; 2019) points to the need of future research on hybrid models that combine data-driven techniques with physical models to improve accuracy.

In a random forest, decision trees can produce numerous trees. This can be handled by restricting tree growth by way of maximum depth or minimum instance requirements. By executing methods to expedite decision tree training, concentrating on minimising computational complexity and utilising randomization techniques, XGBoost avoid overfitting and increase training speed. XGBoost's randomization techniques include training individual trees on random subsamples and performing column sub-sampling at each tree and tree node levels.

Most of the energy used in buildings is still from non-renewable, fossil fuel resources. On the other hand, the building sector also holds the highest energy efficiency potential. Fossil fuels being in high demand and limited in its presence and availability points to a concern over energy security in the near future.

3 Methodology

This research uses the KDD (Knowledge Discovery in Databases) process to examine and analyze energy usage in commercial buildings from the CBECS 2018 information set. Such an approach is appropriate when constructing and enhancing assumption models, which is why this type of method is applied to the integration of complex machine learning models, including XG-Boost, Random Forest, Linear Regression and Ridge Regression, in the study. The purpose of these models in this research is to attempt to forecast building energy use or usage patterns based on building characteristics such as building type, operational data, climate information, and occupancy trends. The remaining subsections describe the general strategy applied to pre-process the data and train several models for building energy consumption prediction.

3.1 Data Collection

The selected dataset is the 2018 Commercial Building Energy Consumption Survey (CBECS) which can be freely downloaded from the U.S. Energy Information Administration (EIA) website. This analysis is done in a way that does not breach any ethical principle, as the dataset utilized is available to the public from the EIA website. The database contains 6436 observations and 1249 variables and is representative of commercial buildings across all 50 states and the District of Columbia. Each record is a unique value representing one in-scope building from the sample of buildings. The sample displays an estimated about 5.9 million whole buildings in the United States.

3.2 Data Pre-processing

The data pre-processing stage is crucial as it prepares the CBECS 2018 dataset for analysis by algorithms. The pre-processing steps involve dealing with missing values for respective variables. A standard procedure would be used on the dataset for analysis, excluding outliers that would otherwise have a significant impact on the analysis of the dataset. It also converts categorical variables, such as building types or natural sources, into numerical variables. Additionally, feature scaling ensures that all numerical values, including building size, energy consumption values, and operating hours, are in a comparable range. After selecting features, correlations are checked to ensure that their selection is not duplicated, and at last the cleaned dataset is split into training and testing datasets for modeling and assessment.

3.3 Data categorization and Variable manipulation

Started by categorizing all data through Python tools for data analysis.

Identified 295 numerical columns in the dataset, with each column being classified into one of two data types.T he data types include 'int64' for integer data, meaning whole numbers, 'float64' for floating-point numbers with decimal points. This initial datatype recognition offered the basic framework to comprehend the essential characteristics of our numerical data. The dataset featured a total of 1,249 numerical columns out of which 612 were of type "int64" (integer) and 637 were of type "float64" (float point). Out of these, 294 were identified as continuous variables. These continuous variables of a single data type were further used for our statistical analyses.

From the summary statistics, the level of data completeness varies significantly for different variables, with some variables having a full complement of observations and some variables having only 2 to 3 data points. The variables are also reported to have different ranges of values, with some being used as values between 1 and 2 and some being used as square footage of the building, which has a range of 1,001 to 2,100,000. It can be noted that many variables have been found to be right-skewed because the mean is higher than the median, Some other variables are also documented to have large outliers, especially if the maximum values are significantly larger than the values at the 75th percentile. This combination of binary, categorical, and continuous variables, along with missing and outlier values adds complexity to the dataset and deserves a cautious approach during analysis.

The target variable ELCNS (energy consumption) was evaluated in both its natural and log forms. On the original scale which was presented, ELCNS exhibited a mean of 2,705,907.73 ± a Standard Deviation of 6,386,292 and a median of 716,335, with a range of 36 to 113,727,053, indicating that the distribution was prone to substantial skewing and outliers. The large standard deviation of the mean was 6,186,292, which was a result of the variation across the observations.

To eliminate this skewness, a log transformation was employed and this resulted in the formation of the variable ELCNS-log, which has a minimum value of 3.61 and a maximum value of 18.55, with a mean value of 13.19 and a median value 13.48. The slight skewness of the data and a standard deviation of 2.11 meant that the data was approximated to a normal distribution. This transformation reduced the effect of outliers and skewness, clearing the way for the identification of suitable data for predictive modeling.

Statistics, correlations, and relations were carried out, and a technical analysis of building characteristics was also conducted.

Mean values represent the average for each variable, while others exhibit a substantial spread. For instance, outliers in a few variables suggest averages that range from nearly 1 to thousands, indicating wide differences in scale across the various measurements.The standard deviations reflect large divergence in the data, with some variables being more distantly scattered while others comprise low central distributions.The median value (or 50%) of the data, as in the statistics ,considers the median value measurement of those variables, which is almost always different from the mean for a several variables due to the skewed nature of many of such distributions. Moreover, high values of the means in subject areas of the distance from the central tendencies also mean that the values are not normally distributed for these few selected variables.The quartiles, essentially the 25th and 75th percentiles, provide information regarding the inter-quartile range, thus enabling data dissemination of the middle fifty percent of the sample which differs across the variables in this case.

Handling Missing Values: Any column with 20% or more missing values was discarded to minimize excessive data elimination, as this would negatively affect most of the data and might give a skewed picture of the model. This threshold was selected because the following attributes would otherwise have a negative effect on the model: Wage Rate of Pay (83.27% missing data) and Quarter (48.92% missing data). That is why we decided to set this limit, it allows for maintaining a good balance in terms of the size of data that should be left and the true records that might be helpful in solving the case.

The data cleaning exercise was aimed at correcting two crucial attributes in the dataset: the missing values and duplicates. For dealing with the missing values, a systematic method was put in place where columns considered to be a case with missing values in excess of 4000 were targeted and eliminated from the dataset. This threshold was selected because such a high number of missing values (making an appreciable part of the 6436 total observations) would render these variables useless for proper and sound analysis. All columns exceeding the missing value thresholds were dropped from the data set using the drop function, thereby refining the data. Furthermore, the data set was also checked for duplicate entries where additional evidence for duplication was sought and all duplicates were removed to ensure data identity.

Checking correlation:The Pearson correlation coefficient was also used to detect the correlation coefficient of all variables used in the study with the dependent variables to determine which of the variables most influenced the model. To test multi-collinearity among the independent variables, a heatmap correlation was plotted using the Seaborn package in Python. The analysis of variables such as SQFT found a strong positive relationship with the dependent variable ELCNS ($r = 0.719$); $r = 0.920$ indicated a high positive relationship between MFBTU and MFEXP with ELCNS ($r = 0.932$). These variables help examine important aspects concerning energy use in commercial buildings.

Log Transformation: The distribution of electricity consumption, 'ELCNS', was negatively skewed, but applying a log transformation to the variable effectively addressed this issue, making the data distribution more normal. This transformation decreased skewness, increased variance stability, and enhanced suitability for analysis. During the modeling stage I reversed the log-transformed target variable to the normal variable to get the actual energy consumption values.

Outlier Treatment: After performing the log transformation on the dependent variable ELCNS and subsequent outlier removal for all variables, the outliers within ELCNS reduced drastically from 623 (9.68%) to 9 (0.14%). This also helped decrease skewness in variables, making the distribution more reasonable and stable with better inputs for modeling.

Feature Selection: To facilitate the selection process, Principal Component Analysis (PCA) was used. 257 principal components were required to capture 95% of the variance in the dataset.

For modelling, the following final dataset was extracted, containing 6436 instances and 257 features. To make the actual data more organized for training, validation, and testing of the model, the data was divided into training data, validation data, and testing data. so that the target variable y maintains the distribution of classes across the splits.

- Training Set: 70% of the data was used to train the model.
- Test Set: 30% of the data was used for evaluating the final model performance.

3.4 Exploratory Data Analysis

The most important part is EDA as this involves coming up with an understanding of distribution of the data set. It helps in understanding patterns, variations and trends. To extract information from the commercial building energy data, the data was summarized qualitatively through the use of simple figures and statistically through measures of central tendency.

Looking at the distribution of energy consumption by building type showed that electricity usage by fire stations/police stations and administrative/professional offices was the highest, with total consumption in excess of 50% of the total (17.4 billion kWh/year), which is attributed to over 24/7 operations and usage. The largest number of buildings included administrative and commercial spaces, 1,329; grocery stores, 936; and medical offices. The energy consumed by the smaller categories is much less; like recreation, which accounts for 35, and public storage units, which were at 23.

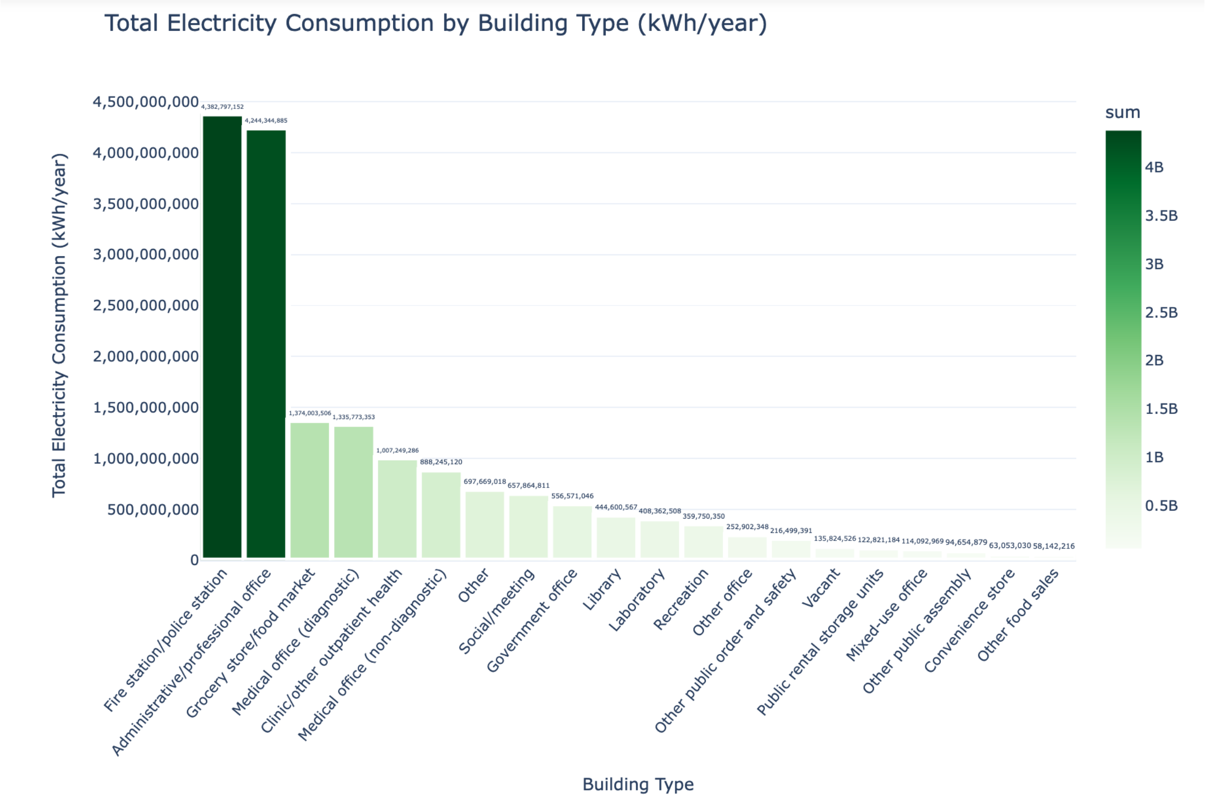

Figure 1: Total Energy Consumption by Different Building Types

These insights discuss where more focused energy conservation measures are required for high-consumption and high-frequency building types to achieve maximum energy efficiency and minimum utilization cost.

From Figure 2, the energy source distribution indicates that natural gas is the most common energy source for the buildings, at 70.6%, followed by fuel oil at 26.7% and propane at 9.6%. District energy heating systems, like steam heating are moderately used in organizations at 7.1%, followed by chilled water at 6.1%, but there is very little use of renewable energy sources, at 4.4% by solar. Others include biomass, comprised of wood, at 0.7,% while coal is used at 0% (-) 0.1, illustrating their rare use. This distribution points out that natural gas and fuel oil are major sources of energy used to meet energy needs, rather than renewable resources such as district systems, and this offers a chance to increase the use of sustainable energy resources in buildings.

4 Design Specification

The design structure of this project is divided into two layers, as shown in the figure 3. The first section is about the data layer, while the lower section of the chart combines both the application layer and the business logic layer. The information about energy consumption in the data layer mentioned above was obtained from the U.S. Energy

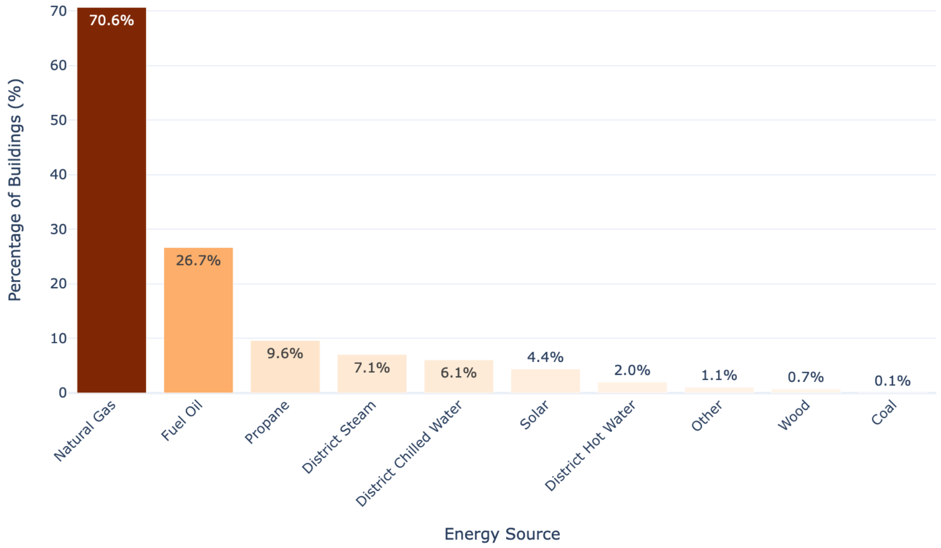

Figure 2: Building Energy Source Distribution

Information Administration website and then cleaned up using Python programming languages. Exploratory data analysis (EDA) was conducted in order to gain a better understanding of the distinct data features of the cleansed data. Then in the post EDA, data encoding was done using Python's scikit-learn library. The project follows different stages of process: Step 1 Preprocessing of data and Analysis - The given data was hereby processed in Jupyter notebook. The results of this analysis served as the basis for the modelling. Step 2: The four models used in the study under consideration are Random forest, XG-Boost, Linear Regression and Ridge regression.

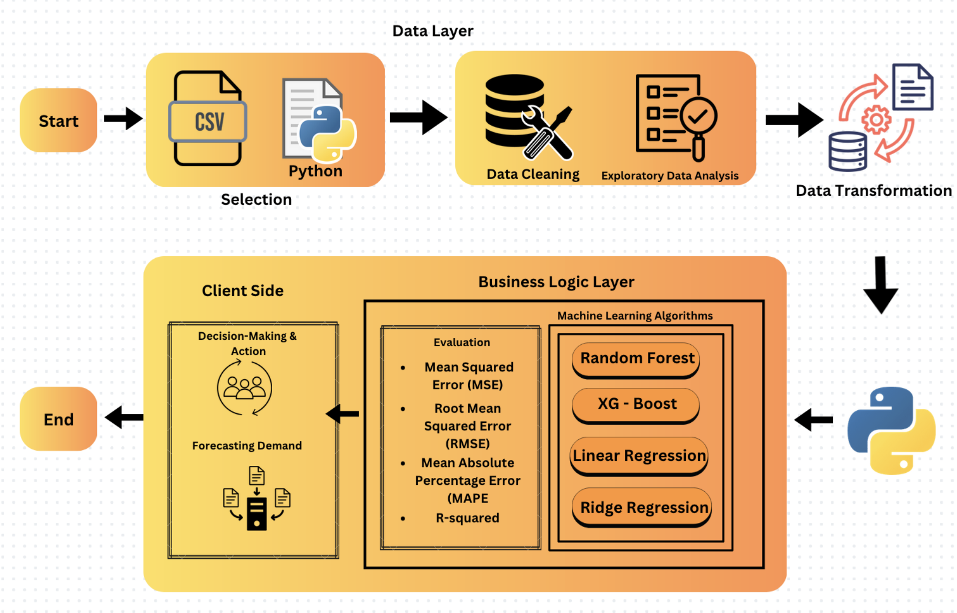

Figure 3: Design Architecture

- Random Forest: For the purpose of testing and forecasting various features that affect ELCNS, multiple decision tree learning classification was employed on the selected features that enhanced the parameters of the test model for a better and closer estimation of its actual consumption.

- XG-Boost: To assess energy predictions, an extreme gradient boosting algorithm was used and tested on numerous structured parameters. The proper choice of hyper-parameters was then optimized to give the best results.

- Linear Model: To achieve this a linear regression model was adopted to set the standard base for determining the amount of electricity used which is the ELCNS, as a way of determining direct variables and setting simple forecasts

- Ridge Regression: Linear regression was used as the main method of analysis, and an L2-regularized model was applied to minimize several major concerns of the study: multi-collinearity and over-fitting, which can lead to more stable predictions of electricity consumption in buildings.

On the Business logic layer four machine learning algorithms has been used on the on and the extent to which their performance outcomes had been assessed using the metrics of encoded/transformed data.On the Client Side, users are provided with processed elements of the Business Logic Layer such as the generated predictions or evaluation of energy consumption within buildings.For stakeholders, this knowledge is useful in observing how energy is used, where there are inefficiencies and how such behaviors can be altered to improve energy use. Clients can interact with this information by dashboards, reports or APIs and may give their comments to enhance the models.

5 Implementation

The implementation phase was conducted in the Python language and Jupyter Notebook environment while utilizing numerous important libraries, This resulted into final preprocessed dataset of 6436 instances consisting of 257 features after applying PCA transformation, and the data split into 70% training and 30% testing data with 4505 instances in training set and 1931 instances in testing set. The target variable ELCNS had outliers reduced from 623 to nine through the use of transformed log values. For this research, I applied four separate algorithms for the formation of machine learning models that predict commercial building energy consumption using the CBECS 2018 data set. Given the extent to which ELCNS is skewed to the right, I gained normality in distribution by taking a logarithm of the ELCNS. This allowed for a much better distribution of the data and the overall pre-processing step was found to be more useful for modeling.

- Outlier treatment: The implementation also entailed a complete data pre-processing step for overcoming skewness and for handling outliers present in the numerical data of the CBECS dataset. In the case of variables where skewness was greater than 2, log1p transformation was used and where values were negative the result was reciprocated. These changes enhanced the significance of distribution of such variables as PBA, and as for the second variable – SQFT, the skewness was decreased from 3.78 to -0.33. Outlier treatment was performed using the Interquartile Range (IQR) method of handling outliers by removing any values below Q1 - 1.5IQR and above

Q3 + 1.5IQR. This process was applied systematically to all numerical variables with the only modification being that the data structure was maintained by making new treated columns. Mean and median values were also computed for the hopes as well as skewness tests were performed to compare new data with the original data to check the efficiency of the transformations.

- Applying PCA on the Variables: The use of dimension reduction was done by implementing Principal Component Analysis (PCA) on the transformed and scaled feature set whereby the dependent variables ('ELCNS-log') were omitted. StandardScaler was used to scale the features in order to apply PCA. Cumulative variance analysis showed that 257 principal components accounted for 95% of the variance in the data set. This dimensionality reduction outcome was presented in the form of a cumulative variance plot that showed a connection between the number of principal components and the variance. The applied PCA transformation allowed to preserve the most important information about the variables, as well as lowering the dimensionality of the feature space, which in its turn will help in the next modeling steps.

6 Evaluation

Table 1: Performance Metrics of Machine Learning Algorithms.

Machine Learning Algorithms	**MSE**	**RMSE**	**MAPE**
Random Forest	0.9112	0.9546	5.6794%
XG-Boost	0.82	0.9037	5.43%
Linear Regression	0.9280	0.9634	5.95%
Ridge Regression	0.9314	0.9651	5.97%

6.1 Experiment 1

Random Forest

Random Forest has been selected on the grounds that it possesses strong capabilities in capturing intricate details and non linear relations in data, thus it is useful in analyzing patterns in energy usages of commercial building. It works through several decision trees for processing multiple building characteristics at the same time. Since model performance and training time are often inversely related, utilizing hyper-parameter tuning,In order to enhance the model used to predict the energy consumption of buildings, hyper-parameters tuning such as n-estimators, max-depth, and min-samples-split were fine tuned utilizing RandomizedSearchCV. It should be noted that RandomizedSearchCV is in contrast to GridSearchCV that evaluates all parameter combinations because GridSearchCV examines all possible outcomes, which is not always practical, that is why RandomizedSearchCV is more efficient while finding the best parameters.

Model Optimization and Performance

All in all, the model performance and errors will be assess throughout the model implementation processes to ensure an understanding of the effectiveness of the prediction model. For performance assessment of the Random Forest model, the predictions are on a par with the actual values through multiple errors. The Mean Squared Error (MSE) was computed to be 0.9112 and this represents the average of squared differences between estimated energy consumptions and the observed log- transformed values. such a relatively low MSE suggests good prediction accuracy on the dataset.

RMSE at 0.9546 gives us a general idea of the model's predictive capability and is in the same units as the Dependent Variable. This measurement is especially helpful in evaluating an extent of prediction errors where the lower measure represents more accurate models.

Most importantly, considering its practical application, Mean Absolute Percentage Error with the value 5.6794 is reasonable. Reporting to the realm of assessed variables, this metric shows that, in general, the predictive model differs from the real energy consumption values by about 5.68 percent. MAPE below 10 indicates that the model can provide good estimates for energy consumption of commercial buildings for practical application.

RandomizedSearchCV includes trying different combinations and it narrows the target variable's hyper-parameter space as well as its computational power. This method is ideal for energy consumption forecasting in buildings as it enables the model to learn non-linear relationships in the data without the risk of over-fitting, ensuring that the predictions are well-informed.

Cross validation folds were used for hyper-parameter tuning, which tested a great number of parameters and selected the best model configuration. The selected model produced a following performance measure on cross validation of model that are accuracy - -0.9272, mean squared error - 0.8571, root mean squared error - 0.9258, R squared of Y - 0.7827 which gives a good indication about its predictive capability.

Altogether, these indications suggest that the Random Forest model provides an appreciable capture of building energy consumption patterns and is equally accurate in generalizing across buildings type differences and variations.

6.2 Experiment 2

XG- Boost Model

XGBoost (Extreme Gradient Boosting) can be described as a very efficient and scalable application of the gradient boosting framework. Since CBECS data consists of multiple factors that contributed to the building energy consumption, the sequential decision making of XG-Boost model in growing trees was useful in capturing such fine grained relationship. This was further explored the hyper-parameter tuning with RandomizedSearchCV with some features such as learning-rate, max-depth, n-estimators. This optimization allowed the model to learn well from the relations that exist within the operations of low rise commercial buildings.

Model Optimization and Performance

The results further analyzed showed that the XG-Boost model had a low prediction error score, with a Mean Squared Error (MSE) of 0.82 on log scale. The Root Mean

Squared Error (RMSE) of 0.9037 reveals the average difference of the magnitude of our predictions from the target variable to the same degree.

Most important, the perfomance of the created model is calculated based on MAPE, which makes 5.43, that is, deviations from real value of energy consumption makes 5.43% in average. This low percentage error indicates high practical significance for realistic energy consumption prediction in commercial buildings.

These error metrics imply that the XG-Boost model is powerful in identifying implicit correlations between the general building characteristics and energy consumption patterns, moreover it has good generalization capacities. The results of model evaluation also show reasonable accuracy of prediction in case of various types of buildings and working modes.

XG-Boost model hyper-parameters utilizing Random Search with 5 fold cross validation. Among the parameters, learning-rate was set to 0.1, max-depth to 3 and there were 300 estimators in the best configuration. This tuned model proved to be effective by attaining slightly lower MSE of 0.75 and RMSE of 0.865 on the scale. Using Mean Squared Error as a measure, from the base model, the Mean Squared Error fell down to 0.75 in the tuned model and RMSE to 0.8651. Such improvement indicates that hyper-parameter optimization can indeed improve model performance for predicting building energy consumption. Low values of -0.7874 of score obtained from cross-validation proves that the model achieved an average accuracies on each sub-samples, hence shows good ability for generalization. The tuned model maintains reliable prediction accuracy while avoiding over-fitting through optimized regularization parameters (reg-alpha: 0.01, reg-lambda: 0.01).

6.3 Experiment 3

Linear Regression Model

Integrating Linear Regression whereby its utilization served as a basic model of developing fundamental relations in analyzing the patterns of energy consumption in buildings. Converting the target variable to log form also assisted to correct skewness in relations of features with energy consumption and this made this model more appropriate for capturing relations. This implement can be used as a reference for comparing with other complex models, while the simple consumption analysis derived from this implement is comprehensible.

Model Optimization and Performance

Linear Regression model evaluation accuracy profile shows moderate predictive capacity for applying log-based data sets. The model obtained a MSE of 0.9280 and, RMSE of 0.9634 thus showing that the prediction of the model was moderate in its accuracy given the simplicity of the model developed above.

Obtained value of MAPE = 5.95% indicates that the average deviation of measurement predictions from actual energy consumption values is about 6%. Although higher than the ensemble models this error rate is still low enough for practical use in baseline energy consumption estimates.

These metrics prove the capabilities of the simple linear model to capture the existing patterns in the data on building energy consumption if the features are transformed, despite the somewhat lower performance as compared to more complex models such as Random Forest and XG-Boost.

6.4 Experiment 4

Ridge Regression Model

Ridge Regression was specifically used to handle the multi-collinearity problem that arises when constructing data related to building energy. The L2 regularization is beneficial in handling cases where features are correlated while also working well due to managing high numbers of features in my processed set of data (257 features). Transferring the target variable through logarithm gave a positive view of tune and permitted to stabilize the prediction equation for the range of energy consumption values.

Model Optimization and Performance

Ridge Regression model exhibited similar error statistics as the basic Linear Regression model. Similarly, it has an MSE of 0.9314 and RMSE of 0.9651 for log transformed datasets and is equally accurate in predictions. According to the MAPE, it shows that the predicted values differ with the actual values by about 6 percent.

As for the L2 regularization that helps in solving the problem of multi-collinearity the performance indicators reveal the improvement. This can suggest that, after using PCA to reduce dimensionality, multi-collinearity it may not have been a major problem with the final selected features.

By making these implementations, A framework was developed for predicting commercial building's energy consumption based on the different modeling strategies and their effectiveness in managing the commercial building energy data for further development in the construction areas for making buildings more energy efficient.

6.5 Discussion

As is mentioned, there were some considerable difficulties during the work on this project. below:

- Missing values in the dataset are an important factor that had to be resolved. The handling of missing values was critical, aligning with approaches discussed by (Deng et al.; 2018).Any column that had more than 20% of its columns missing its values was excluded for cleansing the data. Furthermore, the study conducted a limited data experiment of dropping columns with more than 20 and 30 percent of missing values and listed a very close result in all of the four measurement standards. The last method aimed at reducing data loss while insuring reliability and involved excluding columns with missing values above 20%.

- The analysis of the energy usage in commercial structures exposed a number of patterns in the natural resource consumption. Above 70% usage of natural gas as a source of natural energy was reported while the utilization of solar which is a renewable energy source was low at 4.4%. Even with the high level of natural gas usage, electricity is still the main energy source of all structures, more so in administrative offices, grocery stores and medical facilities. Once again, gas, though quite available, gets dwarfed by electricity's dominance. In this case, the position occupied by (Benti et al.; 2023) is suggested,the authors identified the gaps in the US renewable resource forecasting and commercial buildings load models.

- The data preprocessing phase resolved some important issues of the distribution of the variates and outliers. The initial analysis of the target variable ELCNS, i.e. electricity consumption in log scale exhibited a significant range of 3.61 - 18.55 with the mean and the median being 13.19 and 13.48 respectively indicating that a fair degree of symmetry after transformation has been attained.

In this research project,four machine learning models were integrated after applying PCA dimensionality reduction that brought down the feature space to 257 components while maintaining 95% variance on the data. As the target variable,ELCNS-log has a distribution ranging from 3.61 to 18.55 on the logarithmic scale which is a range of 14.94. With this wide span in recorded numbers, the modeling approach used was required to serve for this large range of energy use patterns as the mean is 13.19 and median is 13.48.

Another tuning that was explored involved the RandomizedSearchCV and the parameters that were used are learning-rate = 0.1 max-depth = 3 and number of estimators = 300. The best model that was trained by choosing the above parameters was XG-Boost with an MSE of 0.75 and an RMSE of 8651. The results found from the Random Forest model are quite satisfactory with an MSE value of 0. 9112, an RMSE of 0. 9546, and an MAPE of 5. 6794%.

Linear Regression provided baseline metrics with an MSE of 0.9280 and RMSE of 0.9634, while Ridge Regression showed comparable results (MSE: 0.9314, RMSE: 0.9651). Both kept their MAPE values to within the range of about 6 percent.

The findings revealed that the XG-Boost had the highest prediction accuracy of the gradient boosting family, with the an MAPE of 5.43% that has a very good practical application for energy forecasting. This paper shows that ensemble methods offer more accurate predictions of the diverse relationships affecting the energy consumption of buildings than linear methods.

7 Conclusion and Future Work

This research was able to achieve the creation of machine learning models to forecast commercial buildings' energy consumption based on the CBECS 2018 data. In this study, the PCA implementation helped to reduce number of features from 1,249 down to 257 components while still attaining 95% variance in the data. In this study, through model development and fine-tuning steps, we also obtained fairly remarkable energy consumption prediction.

The study of the energy consumption data showed that energy used varied between 36 kW and 113,727,053 kW. Afterward, natural logarithm scaling was applied and the original values were transformed to a range of 3.6109 and 18.5493 (total range: 14.9384). As a result of this logarithmic transformation, the energy consumption data had a balanced distribution with a mean of 13.1922 and median of 13.4819, and with a standard deviation of 2.11 where most of the values were tightly clustered. These log values are however not the same as kilowatts rather they represent the same energy range but in a diffused pattern when they are translated back to kilowatts. All the models worked better on this transformed scale with XGBoost being the highest performing in this regard (MSE: 0.75, MAPE: 5.43%), which therefore suggests its predictions on energy consumption generally differed with less than 5.43% spread in kilowatts. In other words, a building with a kWh of 10,000 would have a prediction difference of about ± 543 kW. The predictions assuming other models had equal prowess included Random Forest (MSE: 0.9112, MAPE: 5.6794%) and linear where the arbitrage range was MAPE: 5.95-5.97%). This significant rate of accuracy across a broad set of building consumption levels shows the strength of the log transformation in obtaining reasonable energy predictions across the sizes of buildings.

Research Question: How does the integration of the renewable energy sources and building properties improve the performance of energy consumption forecasting in commercial buildings?

To answer the research question, the outcome provided essential information concern to energy usage in commercial buildings. Out of the natural resources, natural gas was identified as the predominant natural energy source with contribution of more than seventy percent while a limited use of renewable energy sources was identified as both a challenge and potential growth area. When evaluating the electricity consumption the correspondence of the heat consumption, the use of the heat got from the electricity was noticed which can be considered as the opportunity of the energy diversification for all the types of the building. After hyper-parameter optimization XG-Boost was found to have the smallest MSE (0.75) and MAPE (5.43%) and therefore has great practical use in energy forecasting.

A distinctive feature of this research is that it is one of the first to focus on energy consumption predictions for commercial buildings using the CBECS 2018 dataset. This novelty of using this recent dataset means that there are no expected comparative studies available to benchmark the results against which presents both opportunities and limitations. However, the models had good performance metrics. The absence of literature comparing the results on the same dataset makes it difficult to place these results in the research puzzle. This original component of this work provides additional axes of investigation, but it equally indicates that more thorough examination and validation studies were needed using this dataset.

Some real limitations include the fact that the dataset is most prominently U.S. dominated and does not contain sufficient information on renewable energy use. In particular, the building technology trends and energy patterns obtained in data may not accurately present the current state. More research should be directed toward temporal analysis to be able to establish the different seasonal fluctuations as well as a changing model of prediction. For the detection on specific building types and systems related to renewable energies the application of improved feature engineering could lead to an increase of the model's performance. and also, the development of new models for the different regions that takes into consideration,including the climate change projection into the future would be useful in the long-term management of building energy management.

These future directions would ideal for practical applications in support of viable building management approaches. These outcomes could also bring improvements to the energy consumption in commercial structures and further the cause of environmental sustainability.

References

Ahmad, M., Zhao, Z.-Y. and Li, H. (2019). Revealing stylized empirical interactions among construction sector, urbanization, energy consumption, economic growth and co2 emissions in china, *Science of The Total Environment* **657**: 1085–1098.

Benti, N. E., Chaka, M. D. and Semie, A. G. (2023). Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects, *Sustainability* **15**: 7087.

Bourdeau, M., Zhai, X. q., Nefzaoui, E., Guo, X. and Chatellier, P. (2019). Modeling and forecasting building energy consumption: A review of data-driven techniques, *Sustainable Cities and Society* **48**: 101533.

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* **163**: 34–43.

Deng, S., Wang, R. and Dai, Y. (2014). How to evaluate performance of net zero energy building – a literature research, *Energy* **71**: 1–16.

Dinh, T. N., Thirunavukkarasu, G. S., Seyedmahmoudian, M., Mekhilef, S. and Stojcevski, A. (2023). Predicting commercial building energy consumption using a multivariate multilayered long-short term memory time-series model, *Applied sciences* **13**: 7775–7775.

Fu, H., Baltazar, J.-C. and Claridge, D. E. (2021). Review of developments in whole-building statistical energy consumption models for commercial buildings, *Renewable and Sustainable Energy Reviews* **147**: 111248.

Han, T. S., Jiang, D., Zhao, Q., Wang, L. and Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery, *Transactions of the Institute of Measurement and Control* **40**: 2681–2693.

Hong, T., Yan, D., D'Oca, S. and Chen, C.-f. (2017). Ten questions concerning occupant behavior in buildings: The big picture, *Building and Environment* **114**: 518–530.
URL: *https://www.sciencedirect.com/science/article/abs/pii/S0360132316304851*

Kamath, S. M. (2020). Energy use intensities across building use types and climate zones using the cbecs dataset.
URL: $https://etd.ohiolink.edu/acprod/odb_etd/etd/r/1501/10?clear = 10p10_accession_num = case1586533755964739$

Li, H., Johra, H., de Andrade Pereira, F., Hong, T., Le Dréau, J., Maturo, A., Wei, M., Liu, Y., Saberi-Derakhtenjani, A., Nagy, Z., Marszal-Pomianowska, A., Finn, D., Miyata, S., Kaspar, K., Nweye, K., O'Neill, Z., Pallonetto, F. and Dong, B. (2023). Data-driven key performance indicators and datasets for building energy flexibility: A review and perspectives, *Applied Energy* **343**: 121217.

Miller, C., Picchetti, B., Fu, C. and Pantelic, J. (2022). Limitations of machine learning for building energy prediction: Ashrae great energy predictor iii kaggle competition error analysis, *Science and Technology for the Built Environment* **28**: 610–627.

Raza, M. H. and Zhong, R. Y. (2024). Integration of additive manufacturing, lean and green construction: A conceptual framework, *Procedia CIRP* **128**: 180–185.

Seyedzadeh, S., Rahimian, F. P., Glesk, I. and Roper, M. (2018). Machine learning for estimation of building energy consumption and performance: a review, *Visualization in Engineering* **6**.
URL: *https://link.springer.com/article/10.1186/s40327-018-0064-7*

Tsanas, A. and Xifara, A. (2012). Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, *Energy and Buildings* **49**: 560–567.