EN3300 Predict gas consumption of residential buildings using Random Forest models

By

Kaiyue Zhang

21086707

CARDIFF
UNIVERSITY

PRIFYSGOL
CAERDYDD

Supervised by

Prof Monjur Mourshed

School of Engineering

Cardiff University

Cardiff, Wales, United Kingdom

Declaration

I hereby declare:
that except where reference has clearly been made to work by others, all the work presented in this report is my own work;
that it has not previously been submitted for assessment; and that I have
not knowingly allowed any of it to be copied by another student.
I understand that deceiving or attempting to deceive examiners by passing off the work of
another as my own is plagiarism. I also understand that plagiarising the work of another
or knowingly allowing another student to plagiarise from my work is against the University
regulations and that doing so will result in loss of marks and possible disciplinary proceedings against me.

Signed …………………………………………

Date …………………………………………

Abstract

This study investigates the application of Random Forest models for predicting residential gas consumption in three cities in Wales: Cardiff, Newport, and Swansea. By leveraging detailed Energy Performance Certificate (EPC) data and local weather variables, the research demonstrates the model's effectiveness in providing accurate energy consumption predictions. The Random Forest model achieved R^2 values of 0.94, 0.96, and 0.96 for Cardiff, Newport, and Swansea, respectively, indicating strong predictive performance. The study highlights the variability in prediction accuracy due to differences in building characteristics and local climatic conditions. It emphasizes the importance of high-quality data and comprehensive variable selection in enhancing model robustness. Additionally, the research identifies limitations such as potential biases in data collection, limited geographical coverage, and the assumption of homogeneous residential buildings. Future directions include incorporating additional data sources, exploring advanced machine learning algorithms, and expanding the study geographically and temporally to improve model accuracy and applicability.

1. Introduction

1.1 Background

Many countries are facing the challenge of achieving carbon neutrality in their communities. The European Union has approved a "Roadmap for moving to a competitive low carbon economy in 2050," (European Telecommunications Network Operators' Association 2011) setting ambitious goals to reduce greenhouse gas emissions by 80-95% by 2050.

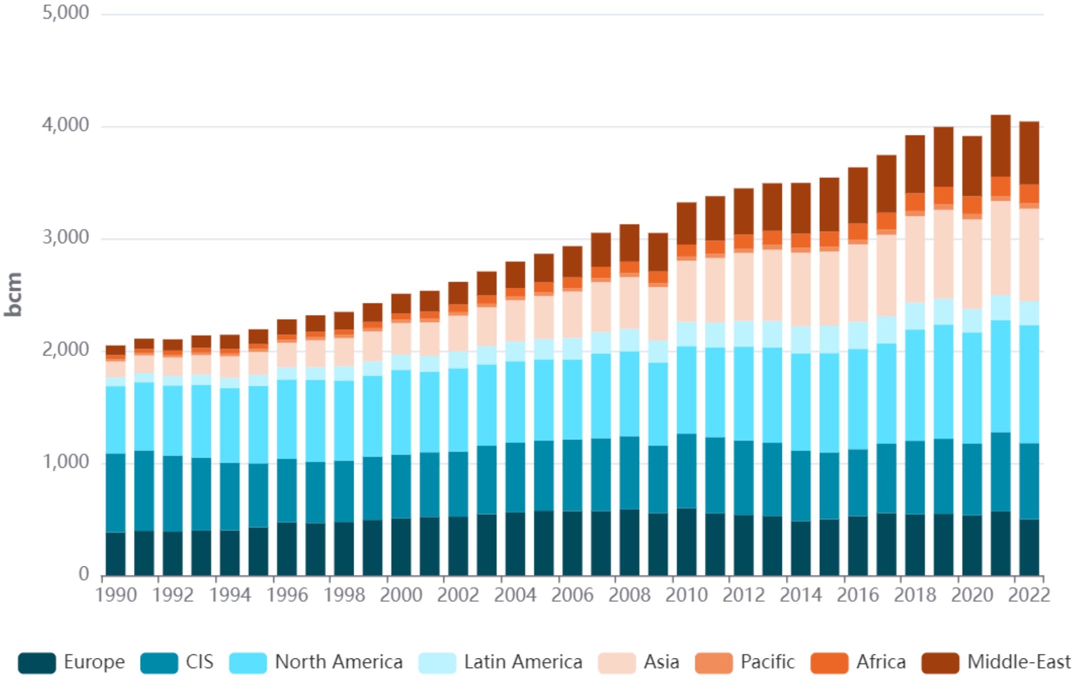

Figure 1. World Gas Consumption (Enerdata, [No date])

Research has shown that buildings in the EU are the most energy-intensive sector, accounting for 40% of total energy consumption and 36% of greenhouse gas emissions. (Ahamd et al. 2017) The rapid increase in greenhouse gases, primarily from burning fossil fuels, significantly contributes to global warming.

The Wales government is following the goals set by EU, archiving net zero by the 2050. The Energy Performance Certificate (EPC) database has become a crucial source for building energy data since the EU began collecting EPC data. (Pasichnyi et al. 2019).

In the report from the Decarbonisation of Homes in Wales Advisory Group to the Welsh Ministers, the group do have concerns about the accuracy of EPC data, however they also agreed that EPC data helps them to assessing the energy performance. (Wales Government 2019) This research will be focusing on three main cities in Wales: Cardiff, Newport, and Swansea.

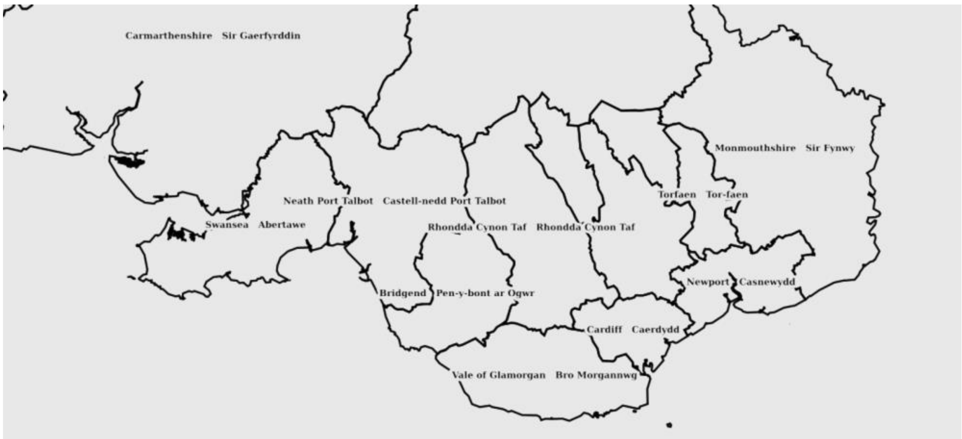
Carmarthenshire Sir Gaerfyrddin
Neath Port Talbot Castell-nedd Port Talbot
Swansea Abertawe
Rhondda Cynon Taf Rhondda Cynon Taf
Monmouthshire Sir Fynwy
Torfaen Tor-faen
Bridgend Pen-y-bont ar Ogwr
Newport Casnewydd
Cardiff Caerdydd
Vale of Glamorgan Bro Morgannwg

Figure 2. Locations of Cardiff, Newport and Swansea（Wales Government, [No date]）

1.2 Using Machine learning models to predict gas consumption

To improve building energy efficiency and gain a deeper understanding of residential building performance, two main methods are used to predict energy efficiency: physical models and data-based models, also known as white box and black box models respectively (Guo et al. 2018).

Physical models use a large amount of detailed information about the building, such as walls, roofs, windows, and weather conditions, to predict energy consumption. In contrast, data-based models primarily use machine learning algorithms to predict future energy use. The main Machine Learning (ML) models include Random Forest (RF), Artificial Neural Networks (ANN), Decision Trees (DT), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Support Vector Machines (SVM) (Pham et al. 2020).

Machine learning models, such as Random Forest (RF), can manage complex interactions between variables and improve prediction performance (Wang et al. 2018). This study employs the Random Forest algorithm to predict gas consumption in residential buildings in three main cities in Wales: Cardiff, Newport, and Swansea. The focus of this research is on using a data-based model for performance analysis.

2. Present research

2.1 What is EPC

The Energy Performance Certificate (EPC) was introduced following the European Union's adoption of the European Performance of Buildings Directive (EPBD) in 2003 (European Parliament, Council of the European Union 2010). By 2008, it became mandatory for all buildings in England and Wales to have an EPC, valid for 10 years (The National Archives, [No date]).

In England and Wales, the EPC is calculated rather than directly measured. An assessor gathers input data through an on-site assessment and supporting documents. This data is then processed using a standard model to calculate the building's energy efficiency under standard conditions (Yuan and Choudhary 2023). The UK government has made the EPC dataset publicly available from 2008 onwards, comprising over 15 million records (Crawley et al. 2019). This research utilizes a portion of this dataset to train machine learning models.

2.1.1 Does EPC reliable

In the early stages of EPC implementation, concerns were raised about the reliability of the data due to incorrect input information, which could lead to inaccurate energy efficiency calculations (Burman et al. 2014). The accuracy of EPC data has been questioned due to discrepancies between estimated and actual energy performance (Hårsman et al. 2016). In Wales, EPC coverage is limited to about half of all homes, with some inconsistencies observed in the ratings over time (Data Science Campus 2020).

2.1.2 Why Use EPC Data?

Despite these reliability issues, EPCs provide a standardized measure of building energy performance, essential for making informed decisions about energy-saving measures and investments (Jeong et al. 2017). The availability and comprehensiveness of the EPC dataset make it a valuable resource for large-scale analysis. Using EPC data allows researchers to leverage a vast amount of information, facilitating the identification of energy efficiency trends and the development of predictive models (Pampuri et al. 2017). Additionally, machine learning algorithms can mitigate some of the inaccuracies in EPC data by identifying patterns and correcting for biases (Data Science Campus 2020).

2.1.3 EPC Data Collection and Use in the UK

EPCs are collected by qualified energy assessors and are essential for buying, selling, or renting homes in the UK. The data includes information on building materials, heating systems, and insulation, which are used to calculate an energy efficiency rating from A (most efficient) to G (least efficient) (Paul, 2010).

2.1.4 Use of EPC Data in Other Countries

In Sweden, EPC data has been utilized to evaluate energy efficiency, though concerns about data quality persist. Studies have highlighted the need for improving the reliability of EPC data to ensure effective energy performance assessments (Cozza et al. 2020). Similar issues have been observed in other countries, emphasizing the importance of standardizing data collection methods to enhance accuracy and reliability (Tsoka et al. 2022).

2.2 Types of Models Used in Existing Research

Existing research on energy consumption prediction models typically falls into two broad categories: white box models and black box models. White box models often require a large amount of detailed information, and the data often are limited or related to experiments. These models, also known as physical models, use detailed parameters of buildings such as material properties, dimensions, and weather conditions to simulate energy consumption.

Black box models, on the other hand, are based on learning methods like linear regression and time series analysis. These models are often simpler and do not require detailed physical information. However, they may lack the flexibility and accuracy required to handle complex and large datasets.

Machine learning models, meanwhile, are increasingly preferred due to their ability to learn from data and improve prediction accuracy over time. Unlike traditional models, machine learning models can handle non-linear relationships and interactions between variables, making them more suitable for complex datasets. This approach has been shown to provide better performance in predicting energy consumption (Parizad et al. 2024).

2.3 Advantages of Machine Learning Models in Energy Consumption Prediction

Machine learning models offer several distinct advantages over traditional energy consumption prediction models, such as white box models:

Enhanced Accuracy: Machine learning models can capture complex patterns and interactions within the data that traditional models might miss, leading to more accurate predictions. These models use advanced algorithms that can model non-linear relationships, improving the overall predictive performance (Domingos 2012).

Scalability with Large Datasets: Designed to process and analyse large amounts of data efficiently, machine learning models can handle big datasets without significant performance degradation. This capability is crucial for modern applications where data volumes are continually increasing (Chen and Guestrin 2016).

Adaptability: Machine learning models can be retrained with new data, allowing them to adapt and improve over time. This continuous learning process makes them particularly suitable for dynamic environments where data patterns evolve.

In contrast, traditional models often rely on predefined equations and assumptions, which can limit their flexibility and accuracy. These models typically require extensive manual tuning and may not adapt well to new data or changing conditions. Machine learning models, however, can automatically adjust to new information, making them more robust and versatile for complex and dynamic datasets.

2.4 Types of Machine Learning Models

Machine learning models used in energy consumption prediction can be categorized into several types:

Decision Trees（DT）: These are simple and interpretable models that split the data into branches to make predictions. However, they are prone to overfitting, which can reduce their generalizability (Quinlan 1986).

Random Forests (RF): An ensemble method that builds multiple decision trees and averages their results to improve accuracy and robustness. Random Forests reduce the risk of overfitting compared to a single decision tree by aggregating the results of multiple trees (Breiman 2001).

Artificial Neural Networks (ANNs): These models are capable of capturing non-linear relationships and patterns in the data. ANNs consist of interconnected layers of nodes, or neurons, that process the input data through weighted connections. They require large datasets and significant computational power for training, but they can model complex phenomena effectively (Ahamd et al. 2017).

Support Vector Machines (SVMs): SVMs are effective for high-dimensional spaces and can be used for both classification and regression tasks. They work by finding the hyperplane that best separates the data into different classes. However, SVMs are less interpretable and can be slower to train compared to other models (Geysen et al. 2017).

Gradient Boosting Machines (GBMs): GBMs combine multiple weak models, usually decision trees, to create a strong predictive model. Each tree corrects the errors of the previous ones, which improves accuracy. GBMs offer high accuracy but at the cost of increased training time and complexity (Deng et al. 2017).

2.5 Advantages and Disadvantages of Random Forest

2.5.1 Advantages

High Prediction Accuracy: Random Forest models typically offer high accuracy by averaging the results of multiple decision trees. This ensemble approach helps to reduce the variance and improve the overall predictive performance (Huang and Huang 2020)

Ability to Handle Large Datasets: Random Forests can efficiently process large datasets. The algorithm's inherent parallelism allows it to handle extensive data volumes, making it suitable for tasks like energy consumption prediction (Breiman 2001; Chen and Guestrin 2016).

2.5.2 Disadvantages

Computational Resources: Random Forest models require significant computational power and memory, especially when dealing with large datasets. This can be a limitation for environments with constrained computational resources (Hengi et al. 2018).

Overfitting: Although Random Forests are less prone to overfitting compared to single decision trees, they can still overfit, especially when the data is noisy. Overfitting occurs when the model learns the noise in the training data rather than the actual signal, leading to poor generalization to new data (Dietterich 2000).

2.6 Why Choose Random Forest?

Random Forest is chosen for this research due to its balance between accuracy, robustness, and its suitability for handling large datasets like the EPC data.

High Prediction Accuracy and Robustness: Random Forest models offer high prediction accuracy by averaging the results of multiple decision trees, which helps to reduce variance and improve robustness. This is crucial for reliable energy consumption predictions (Breiman 2001).

Handling Large and Complex Datasets: Random Forests are designed to efficiently process large datasets. This capability makes them ideal for managing extensive data volumes and complex interactions within the data, such as those found in the EPC dataset (Chen and Guestrin 2016).

Reduced Overfitting: Compared to single decision trees, Random Forests are less prone to overfitting. By using multiple trees and averaging their predictions, Random Forests generalize better to new data, which is essential for accurate energy consumption forecasting (Dietterich 2000).

Feature Importance: Random Forests provide insights into the importance of different features in the prediction process. This ability helps to identify significant factors affecting energy consumption, guiding targeted energy-saving measures and policy decisions

3. Methods

3.1 Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and relevance of the data used for modeling. This study involved several key steps:

Data Selection: The selection of data was based on the availability and relevance of EPC and weather data. EPC data provides detailed information on building characteristics, while weather data includes variables such as temperature, humidity, and heating degree days.

Filtering and Cleaning: Non-numeric and irrelevant text data were removed from the dataset to focus on essential numerical features. This step ensured that only the most relevant data, such as building materials, heating systems, and insulation types from the EPC dataset, were included. This process is essential for eliminating noise and improving the model's performance (Han et al. 2011).

Selection of Important Numeric Data: Key numeric variables were selected based on their potential impact on energy consumption. These variables included:
Lodgement Date
Energy Consumption Current
Current Energy Efficiency
Environment Impact Current
CO2 Emissions Current
Lighting Cost Current

Heating Cost Current
Total Floor Area
Main Fuel

Handling Specific Data Requirements: During data processing, specific code was implemented to remove entries where the main fuel was electricity. This is because this research focuses solely on gas consumption. By excluding these records, the dataset is refined to better suit the study's objectives (Witten et al. 2016).

3.2 Theoretical Background of Random Forest

Random Forest is an ensemble learning method based on decision trees, widely used for classification and regression tasks. The core idea of Random Forest is to construct multiple decision trees during training and combine their outputs to improve overall prediction performance and robustness. This method was introduced by Breiman (2001) to address the overfitting problem often encountered with single decision tree models.

The technique consists of two main components: "randomness" and "forest." Firstly, in the construction of each tree, Random Forest employs bootstrap sampling (i.e., sampling with replacement) to generate multiple different training sets. This process, known as Bagging (Bootstrap Aggregating), effectively reduces the model variance. Secondly, at each node split, Random Forest randomly selects a subset of features rather than using all features, which further enhances the model's diversity and generalization ability.

Studies have shown that Random Forest has several advantages. It can handle high-dimensional data and a large number of missing values effectively.

Moreover, it performs well in capturing nonlinear relationships and complex interaction effects (Liaw and Wiener 2002). Additionally, since multiple trees are built, they can be processed in parallel, making Random Forest computationally efficient in big data environments (Geurts et al. 2006).

However, Random Forest also has some limitations. For instance, its complexity can make it challenging to interpret individual prediction paths. Furthermore, although diversity improves robustness, too many trees might lead to unnecessary computational resource consumption in some cases (Biau and Scornet 2016).

In summary, Random Forest is a powerful ensemble learning method that has become a crucial tool in machine learning due to its superior performance and robustness in handling various complex datasets.

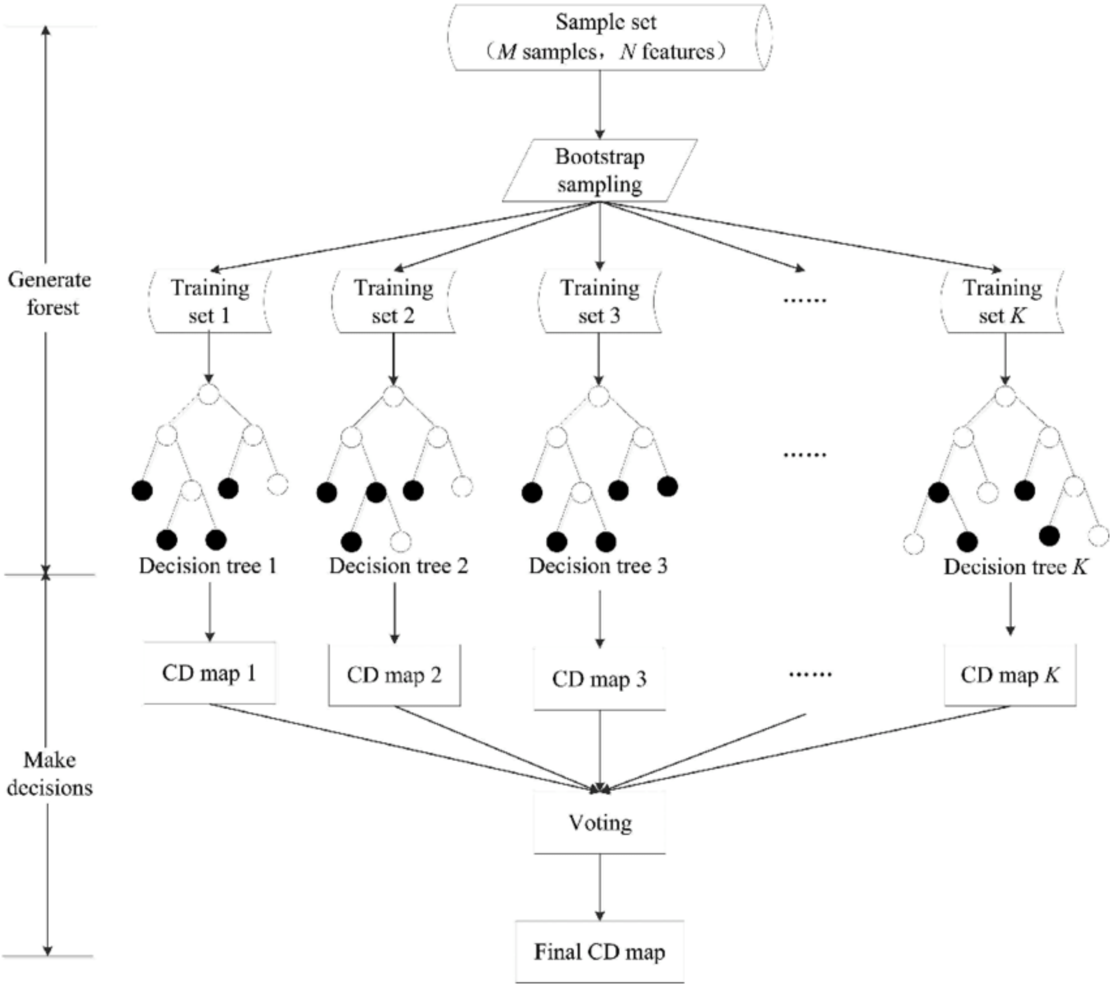

Figure 3. The Flow Chart of Random Forest (Feng et al., 2018)

3.4 Weather Data

Weather data significantly impacts the performance of Random Forest models in residential energy prediction. When forecasting residential energy consumption, weather conditions such as temperature, humidity, wind speed, and precipitation are considered crucial feature variables (Tso and Yau 2007). These weather factors directly affect heating and cooling demands, thereby significantly influencing energy consumption patterns.

Studies have shown that incorporating detailed weather data can substantially enhance the accuracy of prediction models. For instance, temperature variations lead to changes in the frequency of heating or cooling system usage, while humidity levels affect the efficiency of air conditioning systems (Fumo 2014). In this context, Random Forest models can effectively capture these complex nonlinear relationships by selecting the most relevant features and aggregating the results from multiple decision trees (Yu et al. 2011).

The weather data was selected based on its significant impact on energy consumption. The variables included in the weather dataset were:

Year

Month

Maximum Temperature (tmax)

Minimum Temperature (tmin)
Air Frost (af)
Rainfall (rain)
Sunshine (sun)
Heating Degree Days (hdd)

3.5 Variable Selection

The variables included in this study, based on their relevance and impact, are:

Building Characterises:
Lodgement Date
Total Floor Area
Main Fuel
Energy Consumption Current
Current Energy Efficiency
Environment Impact Current
CO2 Emissions Current
Lighting Cost Current
Heating Cost Current

Weather Data:
Year
Month
Maximum Temperature (tmax)
Minimum Temperature (tmin)
Air Frost (af)
Rainfall (rain)
Sunshine (sun)
Heating Degree Days (hdd)

3.6 Hyperparameter Tuning for Random Forest

In machine learning, the performance of a model largely depends on the hyperparameters chosen. It is crucial to test different parameters to determine the optimal ones, ensuring that the model performs best in practical applications. The process of parameter selection can significantly affect the model's accuracy, complexity, and computational efficiency (Bergstra and Bengio 2012).

In this study, we used grid search and cross-validation techniques to determine the optimal parameters for the Random Forest regression model. Grid search systematically explores a predefined set of parameter values to evaluate the performance of each combination. Cross-validation, a common model validation technique, involves dividing the training data into multiple subsets, using different subsets as the validation set while the remaining subsets are used as the training set, thereby assessing the model's robustness and generalization ability.

Specifically, we tested the following parameter ranges:

N estimators (number of decision trees): 30,50,70
Max depth (maximum depth of the trees): None, 20, 40
Min samples split (minimum number of samples required to split an internal node): 4,6,8
Min samples leaf (minimum number of samples required to be at a leaf node): 1,3,5
Max features (maximum number of features considered for splitting): sqrt, log2

To reduce computational cost and time, we selected a smaller parameter range and reduced the number of cross-validation folds from the default five to three. By plotting the effect of different parameter values on model performance (measured by the R² score), we can visually observe and compare the impact of each parameter on the model's performance, thus selecting the best parameter combination.

The optimal hyperparameters selected through this process are:

N estimators: 50
Min samples split: 6
Min samples leaf: 1
Max features: 'sqrt'
Max depth: None

The best cross-validation scores for each city are:

Cardiff: Best Cross-validation Score: 0.97
Swansea: Best Cross-validation Score: 0.91
Newport: Best Cross-validation Score: 0.90

3.6.1 Performance of each Hyperparameter

The following graphs will be included to illustrate the performance of each hyperparameter:

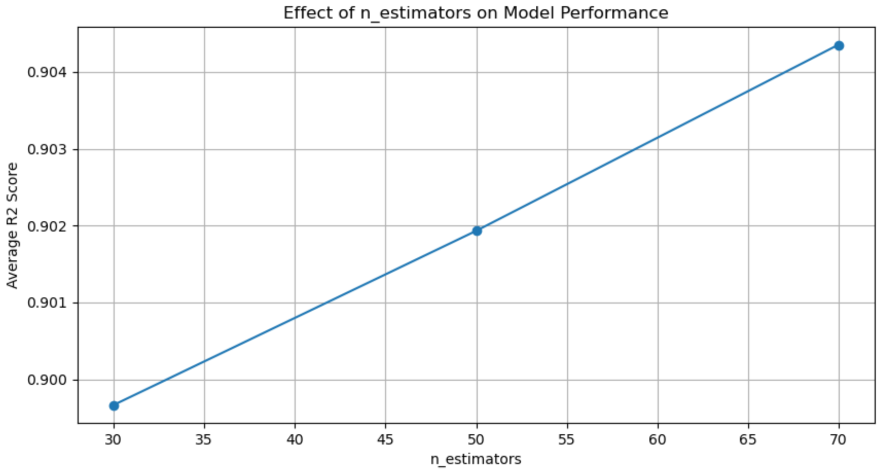

Figure 4. The effect of n estimators on model performance.

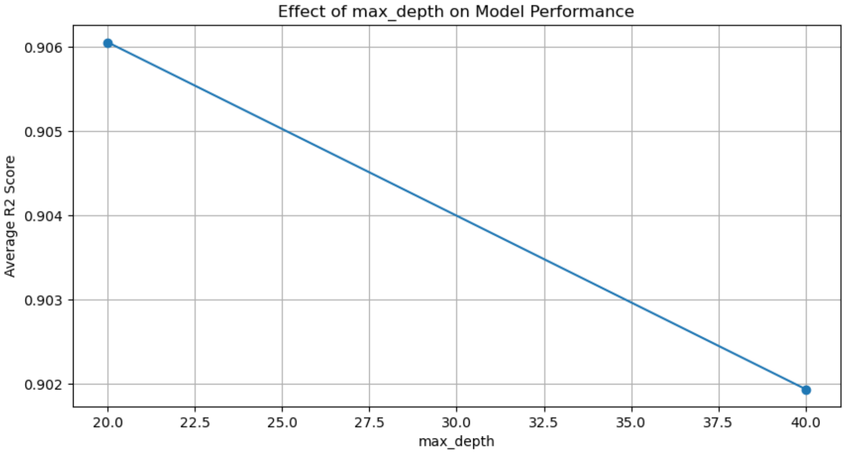

Figure 5. The effect of max depth on model performance.

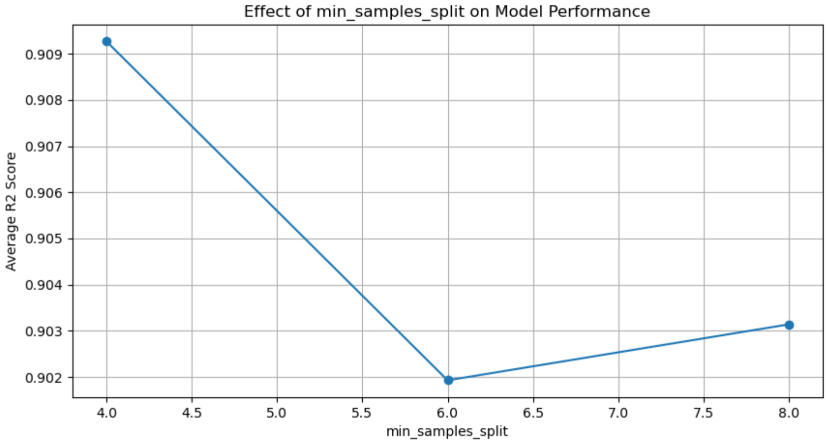

Figure 6, The effect of min samples split on model performance.

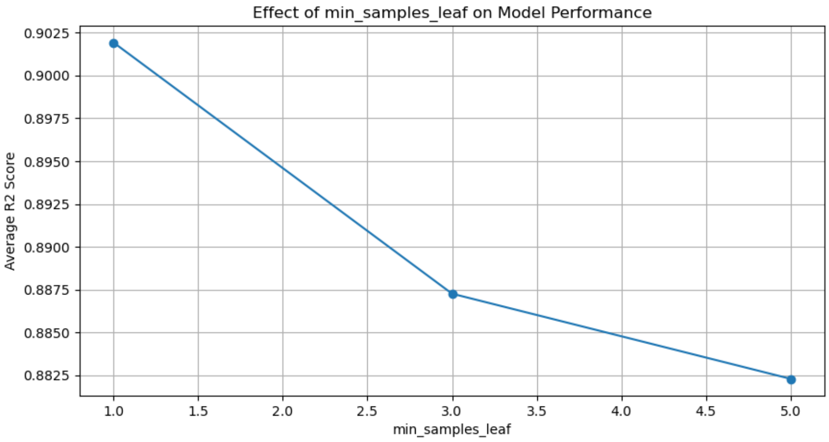

Figure 7. The effect of min samples leaf on model performance.

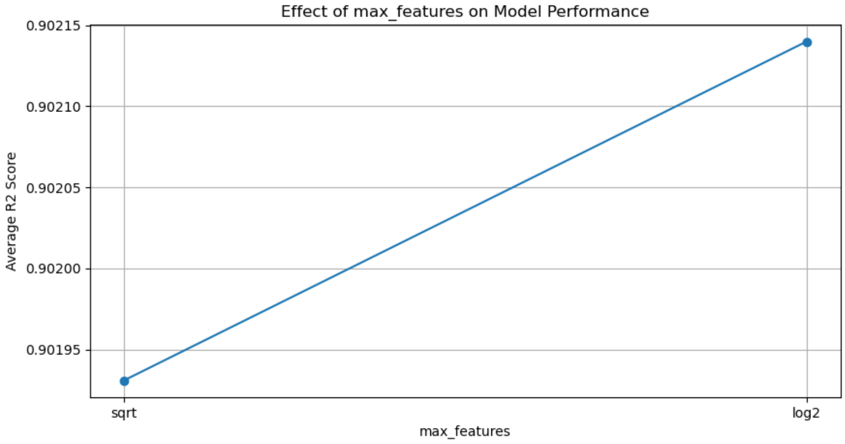

Figure 8. The effect of max features on model performance.

3.7 Training the Random Forest Model

In this study, a laptop equipped with an AMD Ryzen 7 5800H processor and an NVIDIA GeForce RTX 3060 Laptop GPU was used to train a model for predicting residential energy consumption. The following are the detailed model training steps, which were implemented using Python and the scikit-learn library.

3.7.1 Data Loading and Preparation

First, the columns to be loaded from the EPC and weather data were defined. These columns include energy efficiency, environmental impact, energy consumption, CO2 emissions, lighting cost, heating cost, hot water cost, total floor area, number of habitable rooms, number of heated rooms, floor height, main fuel, inspection date, and lodgement date.

Load and Prepare Data: A function ***load_and_prepare_data*** was defined to load and prepare the data:

Load EPC Data: Read the necessary columns from the EPC data file and convert ***LODGEMENT_DATE*** to date format.

Filter Data: Filter out rows where the ***MAIN_FUEL*** column contains the term "electricity".

Load Weather Data: Read the weather data file and combine the year and month into a date format.

Merge Data: Merge the EPC data with the weather data based on the year and month extracted from the lodgement date.

Create Interaction Features: Create new features as interactions between each weather feature and EPC features.

Drop Unnecessary Columns: Drop intermediate columns to simplify the dataset.

Model Training and Evaluation:

3.7.2 Model Training and Evaluation

Define Target Variable and Features: Set ***ENERGY_CONSUMPTION_CURRENT*** as the target variable and exclude non-numeric columns.
Split Data: Split the data into training and testing sets with a ratio of 80:20 using scikit-learn's *train_test_split.*
Feature Standardization: Standardize the numeric features using scikit-learn's *StandardScaler.*
Train Random Forest Model: Train the Random Forest regression model using predefined optimal parameters with scikit-learn's *RandomForestRegressor*. The parameters include:
50 decision trees (n_estimators)
A minimum of 6 samples required to split an internal node (min_samples_split)
A minimum of 1 sample at a leaf node (min_samples_leaf)
sqrt as the maximum number of features for splitting (max_features)
No restriction on the maximum depth of the trees (max_depth)
Model Evaluation: Evaluate the model performance using the R^2 score and mean squared error (MSE). If the R^2 score is greater than 0.8, plot a scatter plot of actual vs. predicted values using matplotlib.

3.8 Identifying and Correcting Errors

During the training phase, several issues were identified and corrected to enhance the model's performance and ensure accurate predictions:
Firstly, handling missing values was an essential step. Missing data points were addressed by imputing values based on the mean or median of the existing data. This approach ensures data integrity and prevents the model from being biased or erroneous due to missing data during the training process.
Secondly, the initial dataset size was too small, resulting in low model accuracy. A smaller dataset limits the model's ability to learn and generalize, thereby affecting its predictive performance. To address this issue, more sample data were added, enabling the model to better capture the patterns and trends in residential energy consumption.

4. Result

4.1 Definition and Formula of MSE and R^2

Mean Squared Error (MSE) and R-squared (R^2) are standard metrics used to evaluate the performance of predictive models.
Mean Squared Error (MSE): MSE measures the average squared difference between predicted and actual values. It is calculated using the formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where Y_i is the actual value $\hat{Y}_i$ is the predicted value, and n is the number of observations. A lower MSE indicates a better fit of the model to the data.

R-squared (R²): R² indicates the proportion of variance in the dependent variable that is predictable from the independent variables. It is calculated using the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

where $\bar{Y}$ is the mean of the actual values. R² values range from 0 to 1, with higher values indicating a better fit. An R² value closer to 1 suggests that the model explains a large portion of the variance in the outcome variable, while a value closer to 0 indicates that the model does not explain much of the variance

4.2 Why Use R² as the Final Metric?

R² is chosen as the final evaluation metric for several reasons:
Firstly, R² has high interpretability. It provides a direct and intuitive measure of how well the model explains the variability of the target variable. This characteristic makes R² easy to understand and communicate, especially when discussing results with stakeholders who may not have a technical background.
Secondly, R² is extremely useful in comparative analysis. It allows for the performance comparison of different models on the same dataset, aiding in model selection. When evaluating multiple regression models, comparing their R² values helps determine which model best explains the variability in the data, facilitating more informed decision-making
Furthermore, R² serves as a standardized metric, making it easier to compare results across different studies and datasets. This standardization is particularly beneficial for large-scale literature reviews or meta-analyses, as researchers can use R² values to consistently measure and compare the outcomes of various studies. (Groemping 2006)

4.3 Comparison of Prediction Results Across Cities

The performance of the Random Forest model was evaluated across three cities: Cardiff, Newport, and Swansea. The results indicated varying levels of prediction accuracy, influenced by differences in building characteristics and local weather patterns.

Cardiff: The model achieved an R² of 0.94 and an MSE of 574
Newport: The model achieved an R² of 0.96 and an MSE of 386
Swansea: The model achieved an R² of 0.96 and an MSE of 345

Although Cardiff and Newport are geographically close, Cardiff had the lowest R². This could be influenced by several factors:

Firstly, Cardiff's building characteristics may be more diverse compared to Newport and Swansea. As the capital of Wales, Cardiff features a wider variety of architectural styles and ages, which could increase data heterogeneity and affect the model's prediction accuracy (Kong et al. 2023).

Secondly, despite using Cardiff's weather data for all cities, microclimatic conditions may still vary. Cardiff's larger urban scale and higher building density could result in more significant local climate effects, such as urban heat islands, impacting energy consumption patterns (Li et al. 2019)

Furthermore, Cardiff's status as a major city brings higher population density and more diverse lifestyle patterns, which may also influence energy consumption. These factors can lead to more complex energy use patterns, making it harder for the model to predict accurately.

Newport's highest R² could be attributed to its relatively homogeneous building characteristics and residential patterns. The residential buildings in Newport might be more uniform, with less variation in architectural style and construction periods, making it easier for the model to capture energy consumption patterns and trends, thus enhancing prediction accuracy.

The following graphs will be included to illustrate the actual vs. predicted energy consumption for each city:

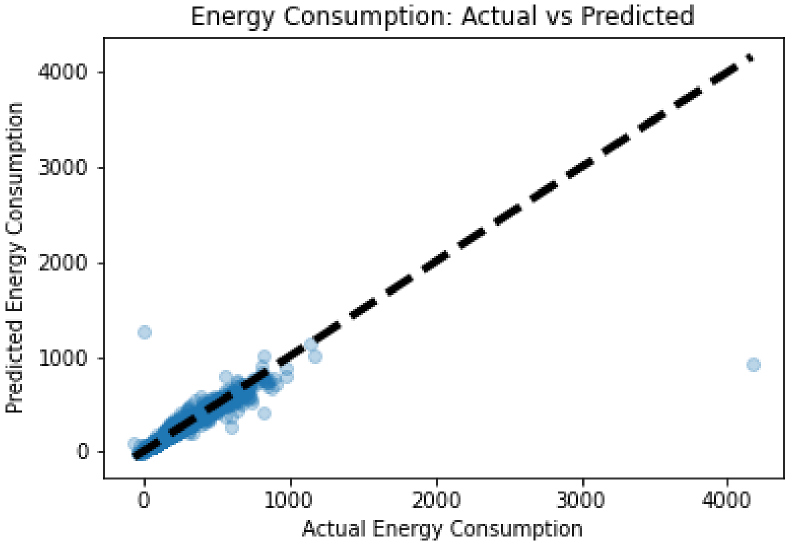

Figure 9. Actual vs. Predicted values for Cardiff.

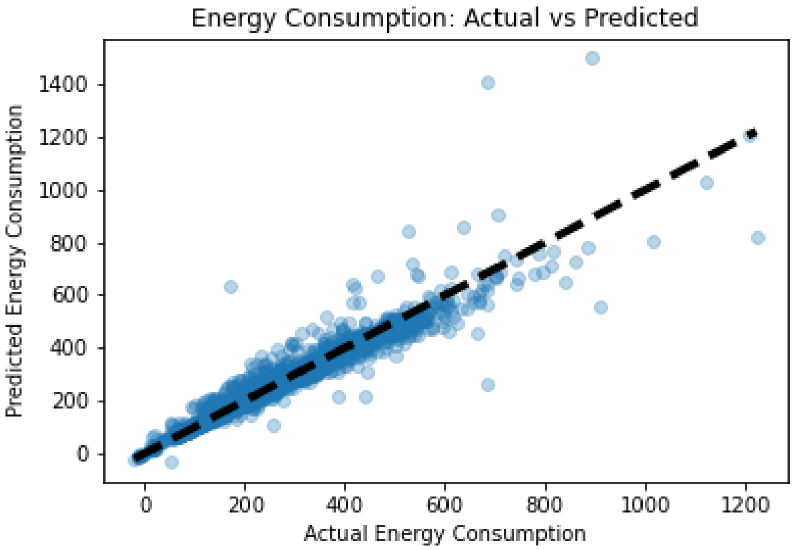

Figure 10. Actual vs. Predicted values for Newport.

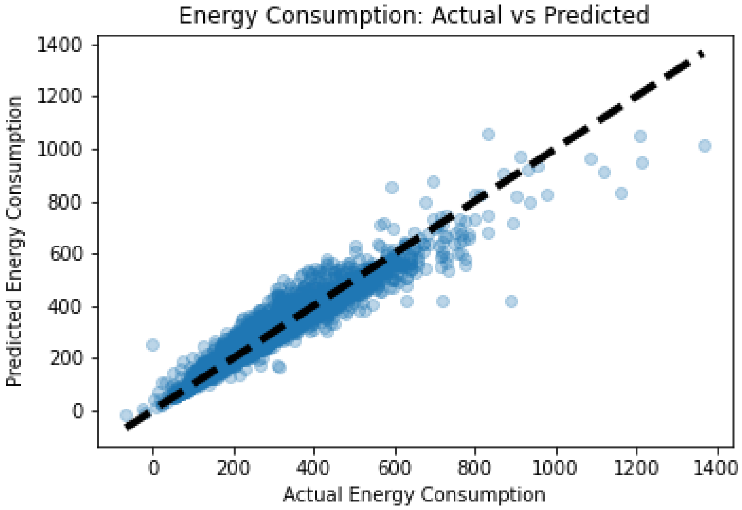

Figure 11. Actual vs. Predicted values for Swansea.

5. Discussion

5.1 Limitations of the Study

Limitations of the Study
This study has several limitations that should be considered when interpreting the results:

Limited Coverage of EPC Data: The EPC dataset used in this study covers only a subset of residential buildings in Cardiff, Newport, and Swansea. Consequently, the data may not fully capture the diversity of building characteristics and energy consumption patterns across these cities. This limitation could lead to a biased understanding of energy use trends, as buildings not included in the dataset might exhibit different behaviours

Potential Biases in Data Collection: The accuracy and consistency of EPC data can be influenced by biases in the data collection process. Energy assessors may vary in their methods of collecting and recording information, leading to discrepancies in the data. These inconsistencies can affect the model’s performance by introducing noise and reducing the reliability of the predictions

Variability in Local Conditions: Although Cardiff, Newport, and Swansea are geographically close, there are differences in local weather conditions and building types that can affect the model's performance. Cardiff's lower R^2 value may be attributed to a greater variety of building characteristics and more pronounced urban heat island effects, which are not as prevalent in Newport and Swansea. These factors add complexity to the model’s task of accurately predicting energy consumption

Uniform Weather Data: The study used Cardiff's weather data for all three cities, potentially overlooking localized climatic variations. Although Cardiff's weather data serves as a useful approximation, it may not account for specific microclimates in Newport and Swansea, which could influence energy consumption differently. This uniformity in weather data might have introduced some inaccuracies in the model's predictions

Assumption of Homogeneous Residential Buildings: The study assumes that all residential buildings have similar energy consumption patterns, which may not hold true in reality. Differences in building age, insulation quality, and occupancy behaviour can significantly impact energy use. Future studies should consider a more detailed classification of residential buildings to improve model accuracy

6. Conclusion

6.1 Summary of Findings

This study demonstrates the effectiveness of using machine learning models, specifically Random Forests, to predict residential gas consumption with high accuracy. The results from this research offer valuable insights that can inform policy decisions aimed at enhancing energy efficiency in the residential sector.

The Random Forest model showed varying levels of performance across different cities. In Cardiff, the model achieved an R^2 of 0.94 and an MSE of 574, indicating strong but slightly lower predictive accuracy compared to Newport and Swansea. This variation is likely due to Cardiff's greater diversity in building characteristics and more pronounced urban heat island effects, which add complexity to the energy consumption patterns

In Newport, the model performed exceptionally well, with an R^2 of 0.96 and an MSE of 386. The higher accuracy here can be attributed to the relative homogeneity in building types and energy consumption patterns, which allowed the model to better capture and predict the underlying trends

Swansea also exhibited high model accuracy, with an R^2 of 0.96 and an MSE of 345, similar to Newport. The consistency in prediction accuracy across Newport and Swansea highlights the potential for applying the model in cities with similar residential building characteristics

Overall, the study underscores the importance of comprehensive and high-quality data in achieving accurate predictions. The insights derived from this research can be utilized to develop more effective energy efficiency measures and policies, tailored to the specific characteristics of different urban environments. Future research should focus on expanding the dataset to include more diverse geographical areas and integrating additional data sources to further enhance model robustness and accuracy.

6.2 Future Research Directions

Future studies should explore several avenues to further enhance prediction accuracy and model robustness:

Incorporating Additional Data Sources: Future research should integrate other data sources, such as real-time energy usage data, socio-economic information, and more detailed weather data, to provide a more holistic view of energy consumption patterns. For instance, integrating smart meter data and socio-economic indicators can significantly improve model accuracy (Fan et al. 2015).

Application of Other Machine Learning Algorithms: Investigating the use of other advanced machine learning algorithms, such as Gradient Boosting Machines (GBMs) and Deep Learning models, could provide further improvements in prediction accuracy. Ahmad et al. (2014) compared the performance of Random Forest and neural networks in predicting building energy consumption, highlighting the different strengths of each approach.

Geographical and Temporal Expansion: Expanding the study to include more diverse geographical areas and longer time periods can help in understanding long-term trends and regional differences in energy consumption. Zhang et al (2023) emphasized the benefits of using deep learning and IoT technologies in smart city planning for predicting building energy consumption.

Reference

Ahmad, M. W., Mourshed, M. and Rezgui, Y. 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Building* 147(15), pp. 77-89. doi: 10.1016/j.enbuild.2017.04.038

Enerdata. [No date]. *Natural gas domestic consumption*. Available at: https://yearbook.enerdata.net/natural-gas/gas-consumption-data.html [Accessed: 16 May 2024]

European Telecommunications Network Operators' Association. 2011. *Roadmap for moving to a competitive low carbon economy in 2050.* Available at: https://etno.eu/news/8-news/36-roadmap-for-moving-to-a-competitive-low-carbon-economy-in-2050.html [Accessed: 16 May 2024]

Pasichnyi, O., Wallin, J., Levihn, F., Shahrokni, H. and Kordas, O. 2019. Energy performance certificates — New opportunities for data-enabled urban energy policy instruments? *Energy Policy* 127, pp. 486-499. doi: 10.1016/j.enpol.2018.11.051

Pham, A., Ngo, N., Truong, T. T. A., Huynh, N. and Truong, N. 2020. Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability. *Journal of Cleaner Production* 260. doi: 10.1016/j.jclepro.2020.121082
Get rights and content

Wales government. 2019. *Better Homes, Better Wales, Better World*. Available at: https://www.gov.wales/sites/default/files/publications/2019-07/independent-review-on-decarbonising-welsh-homes-report.pdf [Accessed: 16 May 2024]