# An Artificial Intelligence Approach to Modelling Household Energy Efficiency


Oisín Brannock (20235671)
School of Computer Science
National University of Ireland Galway

*Supervisors*
Dr. Karl Mason

In partial fulfillment of the requirements for the degree of
*MSc in Computer Science (Artificial Intelligence)*
August 6, 2022

**DECLARATION**

I, Oisín Brannock, do hereby declare that this thesis entitled "An Artificial Intelligence Approach to Modelling Household Energy Efficiency" is a bona fide record of research work done by me for the award of MSc in Computer Science (Artificial Intelligence) from National University of Ireland, Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: _____

# Abstract

Many studies have reported successful machine learning models using sensor data. However, generation of a successful model using generally accessible home owner data, like energy consumption and location, remains to be accomplished. Here I report that a combination of 13 general factors can be inserted into a pipeline that can: impute the missing values in features, encode categorical features and under/over sample any minority classes before modelling. The result post optimisation is a new data pipeline, simply named Building Energy Rating Predictor (BERP), that encapsulates the principles of CRISP-DM and MLOPs in the delivery of a user-friendly model that is able to successfully identify and correctly predict building energy efficiency scores to an F1 score of 0.82. Modelling of this kind can be used to underpin an application interface that home owners can make use of to find energy efficiency ratings for free, without on site inspection, as well as personalised benefits of improvements to energy costs. Likewise, it can underpin government level initiatives to identify the areas of within the Republic of Ireland at most in need of home improvements, in order to reach the 2030 B2 household target set. BERP has the ability to meet both of these criteria with excellent accuracy.

**Keywords:** Energy Efficiency, BER, F1, Accessibility, Cost Savings

# Contents


**1 Introduction** **1**

1.1 Motivation . . . . . 1

1.2 Thesis Structure . . . . . 2

1.3 Research Questions . . . . . 3

**2 Background** **4**

2.1 Machine Learning . . . . . 4

2.2 CRISP-DM . . . . . 5

2.3 Missing Data - Imputation . . . . . 8

2.3.1 Missing Forest Imputation . . . . . 8

2.4 Data Encoding . . . . . 9

2.5 Sampling . . . . . 11

2.6 Supervised Learning . . . . . 13

2.6.1 Classification . . . . . 13

2.6.1.1 K-Nearest Neighbours (kNN) . . . . . 13

2.6.1.2 Decision Trees . . . . . 15

2.6.1.3 Random Forests . . . . . 17

2.7 Neural Networks . . . . . 19

2.7.1 Artificial Neural Networks . . . . . 19

**3 Literature Review** **22**

3.1 Research Method . . . . . 22

3.2 Literary Review . . . . . 23

3.3 Review Conclusions . . . . . 30

**4 Data Processing & Analysis** **32**

4.1 Data Processing . . . . . 32

4.2 Analysis . . . . . 34

**5 Experimental Setup** **36**

5.1 Data Preparation . . . . . 36

5.2 ML Pipeline . . . . . 37

**6 Results** **39**

6.1 Model Evaluation . . . . . 39

6.1.1 Classification Evaluation . . . . . 40

6.1.2 Confusion Matrix . . . . . 41

6.1.3 Area Under the Receiver Operating Characteristics (AUROC) . . . . . 42

6.1.4 Feature Importance . . . . . 43

6.1.5 Precision Recall Evaluation . . . . . 44

**7 Discussion** **45**

7.1 Model Usage . . . . . 45

7.1.1 Interpretation of Confusion Matrix Importances . . . . . 46

7.1.2 Expectations . . . . . 46

**8 Conclusion** **49**

**References** **60**

# List of Figures


2.1 Simple Machine Learning Workflow . . . . . 5
2.2 CRISP-DM Methodology . . . . . 7
2.3 Distance Formulae . . . . . 14
2.4 Example Elbow Plot . . . . . 14
2.5 Classification Tree Flow Example (Taken from [1]) . . . . . 16
2.6 Random Forest Flow . . . . . 18
2.7 Linear Regression Network: $y = wx + b$ . . . . . 19
2.8 Feed Forward Neural Network . . . . . 21

5.1 BERP Pipeline . . . . . 38

6.1 Confusion Matrix . . . . . 41
6.2 ROC Curve . . . . . 42
6.3 Feature Importance . . . . . 43
6.4 Precision-Recall Curve . . . . . 44

# List of Tables

6.1 Model Results . . . . . 40

# List of Acronyms

**AI** Artificial Intelligence. 4, 30, 31, 49

**ANN** Artificial Neural Network. 19, 28, 35, 39, 47, 50

**AUC** Area Under the Curve. 35, 40, 42

**AUROC** Area Under the Receiver Operating Characteristics. iv, 42

**BER** Building Energy Rating. ii, 1, 2, 13, 24, 29–32, 36, 37, 39, 40, 42, 43, 45–47, 50, 51

**BERP** Building Energy Rating Predictor. ii, v, 26, 28, 38, 39, 41, 42, 44, 45, 47–50, 60

**CART** Classification and Regression Tree. 15

**CatBoost** Category Boosting. 37

**CNN** Convolutional Neural Network. 28, 29

**CPU** Central Processing Unit. 49

**CRISP-DM** Cross-Industry Standard Process for Data-Mining. ii, iii, 5–7

**DNN** Deep Neural Network. 47, 50

**ELM** Extreme Learning Networks. 28

**FLNN** Functional Link Neural Networks. 28

**GA** Genetic Algorithm. 28

**GDPR** General Data Protection Regulation. 32

**IEEE** Institute of Electrical and Electronics Engineers. 22, 23

**kNN** K-Nearest Neighbours. iii, 2, 8, 11, 13, 14, 33, 34, 39, 47

**LSSVM** Least Squares Support Vector Machine. 28

**ML** Machine Learning. 37

**MLOPs** Machine Learning Operations. ii

**MSE** Mean Squared Error. 20

**PCA** Principal Component Analysis. 14, 25

**PLS** Partial Least Squares. 25

**PR** Precision Recall. 44

**ReLU** Rectified Linear Unit. 20

**RGS** Regression Gradient Guided Feature Selection. 27

**ROC** Receiver Operating Curve. 42

**ROI** Republic of Ireland. 30, 32, 50

**SEAI** Sustainable Energy Authority Of Ireland. 32, 60

**sklearn** Scikit-Learn. 35

**SMOTE** Synthetic Minority Oversampling Technique. 11, 12, 37

**STL** SMOTETomekLinks. 12, 34, 37, 40, 47, 49

# Chapter 1

# Introduction

## 1.1 Motivation

**The BER of households can be predicted using general household characteristics to an accuracy of 80%.**

The topic of energy efficiency is one of the most prevalent topics in our society today. The move towards environmentally friendly practices has seen a dramatic rise in popularity, in both the private and public sectors, as well as within the general public. A variety of approaches have been taken in this regard. Journals published as early as the 1980s have broached different approaches in search of the ideal model [2]. The scope of this topic governs every industry in the world, and is imperative to a sustainable future for humankind. Initially, the studies in this field focused on efforts to improve the efficiency of buildings during their construction period [3]. This has evolved in the age of information through the use of computational models to predict what can impact energy usage in a building, whether it be a school, office or water treatment plant [4; 5; 6; 7].

The goal of this thesis is to model energy usage for households using generally accessible household data, and to highlight the uses for this model, not just for the home owners, but also general sustainability bodies in government, in the focus on driving to a net zero target by 2050 [8]. Home owners would gain an awareness of how much they spend, easily find their BER for free using the model and see potential savings given a new BER change. This analysis can help home owners save money, and most importantly help raise awareness and inspire action for a sustainable environment. However, this raises questions in regard to which approach is preferred. There are concerns around data availability, reliability and optimal approaches.

This thesis seeks to determine which method is the most viable and apply a new solution to an ever increasing concern. Methods such as neural networks, decision trees and kNNs have been shown to be effective in this space using sensor data, but how will they fare using generally accessible household data, like type of structure, insulation type, dwelling type etc. Sensor data is far more reliable, but not practical for home owners to simply gather on their own. Every feature used in this model is something people can access on their own.

## 1.2 Thesis Structure

Chapter 1 outlines the motivation for this paper.
Chapter 2 focuses on the methods used to perform research and analysis of the data.
Chapter 3 is an in depth literature review of the work done in the space of energy and sustainability to get an idea of a baseline to work with.
Chapter 4 delves into the data used for analysis; where it comes from, how it is processed for use and what considerations had to be taken into account in its

usage. The analysis portion deals with the modelling of the data; what models were chosen and why, how each is optimised for the dataset at hand.
Chapter 5 highlights the experimental method from start to finish; steps on how analysis was done, settings of cross validation.
Chapter 6 examines the model results, and goes into depth on post analysis to outline the results and what they mean.
Chapter 7 provides an in depth discussion of the results presented in the Chapter 6; what significance they have in real world use cases? Are they as expected, or is there some flaw highlighted?
Chapter 8 presents a summary of the thesis, outlines the contributions of the research, provides answers to the research questions posed in Chapter 1, and finally discusses the implications and impact of this work.
All diagrams used in this thesis have been created by the author and are only referenced when required.

## 1.3 Research Questions

This research will answer the following research questions (RQs):

$RQ_1$ - What are the data requirements for modelling energy usage for households?

$RQ_2$ - How should machine learning be applied to best predict household energy efficiency?

$RQ_3$ - How can users effectively interact with and gain insight about the household base from this model?

# Chapter 2

# Background

This chapter gives context to the methods used in this thesis for analysis. It also ties into how each one is relevant to the idea of energy saving.

## 2.1 Machine Learning

Machine learning is a branch of AI that allows machines to solve a vast range of problems faster, and more often than not, more accurately than a human can. Machine learning is concerned with designing programs that can learn rules and patterns from data, and adapt to new scenarios based on this training [9]. This allows machines to infer answers rather than having to be explicitly programmed. Many of the challenges we wish to overcome in today's world are not straightforward and cannot be simply programmed for a computer to solve in a binary manner. There are a plethora of techniques machines can use to solve problems, which are constantly being improved upon through research and analysis. Machine learning consists of 3 sub-categories: supervised learning, unsupervised learning and reinforcement learning. This thesis focuses on the use of supervised learning methods. In Fig. 2.1, the general flow of a machine learning process is

outlined. It starts by first gathering, cleaning and splitting a dataset into training and testing datasets, and normalising/transforming data to suits the means of the task at hand. Next the model is trained on the training data. The model is then evaluated using statistical tests to check if the results are statistically significant and if they make sense in reference to the hypothesis. The model can then be deployed for use and improved using new incoming data. This is a very simple idea of how models are created. Section 2.2 expands on this for a more cohesive structure.

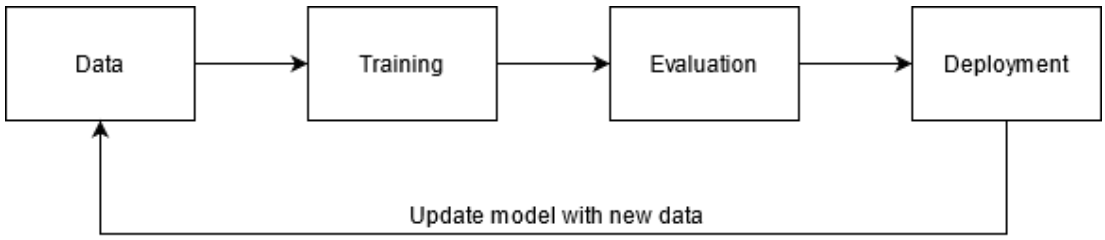


Figure 2.1: Simple Machine Learning Workflow

The availability of data today has led to a huge uptake of machine learning by new practitioners, expanding beyond the realms of scientific research into everyday life. This has included the creation of models that cover the likes of fraud detection in the financial industry, allow cars to be autonomous on the road, or facial recognition that is used in phone and security technology [10; 11; 12]. The use cases are endless and constantly being expanded upon.

## 2.2 CRISP-DM

The Cross-Industry Standard Process for Data-Mining (CRISP-DM) is a methodology that seeks to standardise how models are developed and maintained across industries. This is to keep a consistent and thorough approach that captures

all aspects of a problem [13; 14]. CRISP-DM is important in relation to the analysis of this thesis as it serves as a base layer of guidance. This thesis has been structured in a way that follows this format, ranging from the "business understanding" aspect in the first couple of chapters, all the way to modelling and deployment in latter chapters. This is especially important here, as the data used is household data, and therefore the process should reflect this to the best possible standard in order to make the best possible impact for this cohort.

Business understanding involves developing a hypothesis for the problem at hand. Is this really a problem that could benefit from machine learning? Is the problem well defined? Is there data available to make a model? What tools and technologies will be used? What does success look like? Break down the project into phases, like chapters of a thesis.

Data understanding comes when the problem has been analysed and hypothesised accordingly. What kind of data do we have? Is it structured in the form of comma separated values? Does it have one row per observation? Is it in wide or long format [15]?

Data preparation involves making sure the data is in an appropriate format for the modelling technique to be used. Does the data need to be normalised? Does it need to be in wide or long format? Do we need to fill in null values? How do we fill in these nulls if they exist; with 0, with the mode or with the mean? Could we create a model to predict what these nulls may be? Could we engineer new features based on the ones we already have at out disposal?

Modelling involves choosing and creating models and determining the one that best suits the data. The best model at first may not be the best model after cross validation. If model performance is lower than desired, one may need to go back a step or two to determine if the data has been understood correctly or has been prepared for analysis correctly.

Evaluation takes the best model and determines if it meets the needs of the project to an acceptable threshold. How would one explain the findings to someone in the simplest terms? If they cannot do this, they do not fully grasp what they have done and must review the work thoroughly to ensure accuracy. The key question in this phase is: **Have we solved the problem we outlined in the initial phase of this project?** If the answer is yes then one can begin devising how to move forward with the desired model.

Deployment involves putting a model into a state where it can be accessed by people outside of the project to solve the problem at hand. How does one plan on achieving this deployment? What resources will be used? Who will maintain the model for incoming data? When these questions have been addressed, a final report can be outlined on the project and reviewed for where things could have been improved in hindsight. If deployment needs to be updated or improved in some capacity, the process can always be restarted.

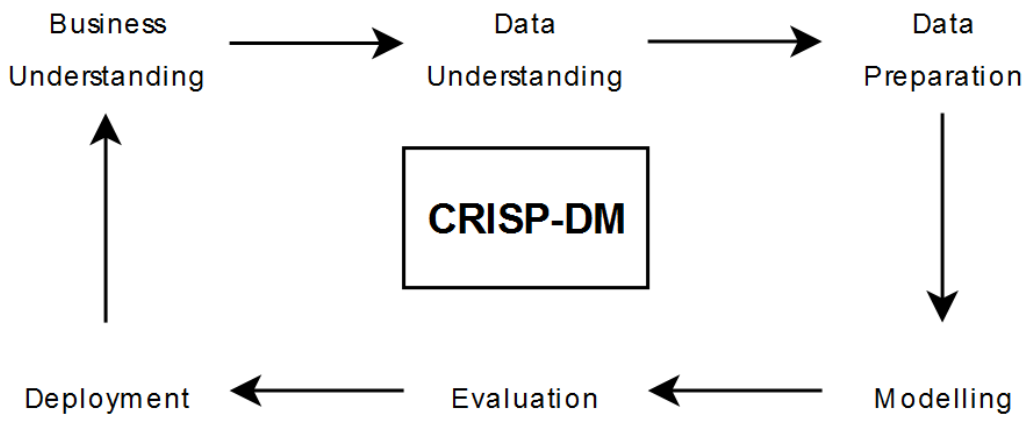


Figure 2.2: CRISP-DM Methodology

## 2.3 Missing Data - Imputation

Missing data is present in all real world datasets. There is very little that can be done to avoid it in most situations. It arises from cases where manual entry is left blank, entered incorrectly, or is poorly ingested by a data engineering pipeline. Whatever the reason, it is imperative to find a way to deal with this kind of data gap issue.

One such method of dealing with data gaps is imputation. This can be as simple as filling the missing values in a dataset with the mean, median or mode of that column, depending on the datatype. Lin et al [16] performed a review of all imputation techniques across literature between 2006 and 2017. They concluded that random forest imputation and kNN imputation were the superior techniques to use in extreme missing data cases ($< 30\%$ for kNN and $> 50\%$ for missing forest).

### 2.3.1 Missing Forest Imputation

Stekhoven and Buhlmann [17] created the algorithm of missing forest in 2011. Their goal was to overcome the limitations of kNN imputation, which is very sensitive to the curse of dimensionality as well as feature scaling. The missing forest algorithm is quite straightforward. The data is taken in its raw form, without any preprocessing of categorical variables into numerical variables required. The algorithm is tuned to both classification and regression, depending on where the data is missing. It starts off by imputing each column with its respective mean/mode. It then uses all the other features in the dataset to build a random forest model to predict the values of the imputed columns. It loops through the dataset until a threshold value is reached, which is set by the size of the missing data.

## 2.4 Data Encoding

Machine learning algorithms generally do not handle categorical variables very well. Categorical variables generally can either be ordinal, in that they have a clear hierarchy, or simply just labels with no ordinality. In all cases, we must discern the value of each label in order for it to be useful in a machine learning setting. There are a variety of ways this can be done depending on the type of categorical data one has.

Label encoding is one such technique. One simply maps each unique category to a numerical value. If we had 3 countries; Ireland, USA and Japan, this would convert to 1,2,3 for example. The downfall of label encoding is that we now have a situation where a model could interpret an order to the labels. It may capture that Ireland > USA for example, which is not true in this setting.

One hot encoding is a technique that looks to overcome the limitations of label encoding by creating dummy variables in the data. It creates additional dummy features, one for each category in the variable. It then labels with a 1 if a row of data is in that category and 0 elsewhere. Dummy encoding is a popular technique but it has limitations. For one, the new dummy variables are extremely correlated with one another, which can lead to collinearity issues. This can be overcome through analysis of the data, and dropping certain dummy columns. The other major downside of one hot encoding is that it can dramatically inflate a dataset if the number of categories to encode is large. For example, if we wanted to encode years ranging from 1900 to 2000, this is an addition of 100 columns. In big datasets, this can cause memory issues very quickly, as well as decrease model performance and increase training time.

The last method in scope for this thesis is known as category boosting encoding, or CatBoost Encoding. Prokhorenkova et al. [18] developed the algorithm in 2018 in order to combat increasing problems in research related to prediction

shift in production grade models over time. It works by first randomly permuting the dataset. This is done to ensure that the dataset is not ordered based on the target variable. Next, the categorical variables are converted to labels, like with label encoding. The following formula is then used to convert the numerical labels to a floating point numerical interpretation:

$$target = \frac{countInClass + prior}{totalCount + 1} \tag{2.1}$$

Where $countInClass$ is how many times the label value was equal to 1 for objects with the current categorical feature value. In a multi-class problem, this is extended out to beyond 0 and 1 labels. *prior* is the starting value for the label during label encoding. It is determined by the starting parameters in the dataset. $totalCount$ is the total number of datapoints (up to the current one) that have a categorical feature value that matches the current one [19]. This method has a big advantage in that it eliminates both the limitations of label encoding, by going beyond ordering and one hot encoding, keeping the same number of columns in the dataset as there were at the outset.

## 2.5 Sampling

Imbalanced datasets are abundant, and a big issue in data science. They occur when we have one class that dominates, and another that perhaps only covers 1% of what the majority class does. This leads to poor model performance, as a model would discriminate on the minority class, and generally choose the majority class. Therefore, we need a technique to combat this problem.

Oversampling is a technique that seeks to duplicate samples in the minority class, so that they are inflated to match the size of the majority class, negating the imbalance. This works very well in practice, and is fast to compute, even on larger datasets. It does not add any new information to the data, but this is not always needed.

Undersampling on the other hand seeks to do the opposite, by removing samples from the majority class at random, until it matches the minority class. The disadvantage of this technique is that we are losing valuable information in our data.

Synthetic Minority Oversampling Technique (SMOTE) was introduced in 2002 by Nitesh Chawla, et al. [20]. This algorithm seeks to oversample the minority class by oversampling. Instead of simply creating exact duplicates of the minority class, SMOTE seeks to create new samples from the minority class. It does this by selecting samples close together, finding the line of best fit between the samples, and drawing a brand new synthetic sample. This is very similar to how kNN works, in that SMOTE uses a k-neighbours parameter to choose how many points to use in the creation of a new one. It will repeat this process until the minority class/es are balanced with the majority class. The new samples are very similar to the original ones, but the slight difference allows more information to be gained from our dataset for modelling.

Tomek developed a technique to deal with undersampling in his 1976 paper

on Condensed Nearest Neighbors [21]. This method of undersampling tries to acquire the datapoints from the dataset that limit the loss in information that a model can use to predict. It does this by taking every sample from the minority class in the data, and selecting only points in the majority class that cannot be classified correctly by the model, due to them being very similar to another class. However, this technique suffers in practice as it selects data at random, which leads to unwanted data being left in the training data. The adaption used to overcome this limitation is called TomekLinks. It states that two points form a Tomek link if they are from separate classes, and each of them is the nearest neighbor of one another. This tells us that both of these points are near a decision boundary that determines class separation in the data. It also tells us one of these data points could be an outlier. Therefore, we want to remove all of these identified TomekLink pairs, in order to make the decision boundaries more distinct between classes. This will make our model much more accurate in its predictions.

Batista et al. [22] combined the methods of SMOTE and TomekLinks together to form SMOTETomekLinks (STL). The technique simply performs SMOTE as before, to oversample the data with synthetic datapoints, and is then undersampled by finding the TomekLinks in the data to make a clear decision boundary between classes. This technique is utilised in this report.

## 2.6 Supervised Learning

Supervised learning is a sub-category of machine learning. It involves the use of labelled data to train a computational model. In training the model, the machine is shown the correct outcome for each training example. The basis for this technique is that we can predict what will happen in similar future scenarios, given what has already happened in the past [23]. This assumes that the factors that initially led to these outcomes have not changed, which may not always be the case, depending on the problem. Supervised learning can be split into classification and regression tasks. Classification deals with categorical, discrete and Boolean predictions. Regression deals with the estimation of continuous values. The focus in this thesis will be on classification, as the goal is to predict BER, which is a multi-class (15 classes) categorical variable.

### 2.6.1 Classification

#### 2.6.1.1 K-Nearest Neighbours (kNN)

kNN works under a simple premise; any datapoints that have similar characteristics to other datapoints in a dataset will have similar outcomes [24]. This is known as feature similarity. In kNN classification, the algorithm starts by calculating the distance of the new datapoint to each training datapoint. This is done either by calculating the Euclidian, Manhattan or Minkowski distance.

The Euclidian distance is calculated by taking the square root of the sum of the difference of squares between the new datapoint and training datapoint [Eq. 2.2].

The Manhattan distance is calculated by taking the shortest distance between the two vectors that represent the new data and training data, and getting the sum of their absolute difference [Eq. 2.3].

The Minkowski distance is a generalisation of the Euclidian and Manhattan distance formulae. The formula relies on the constant $p$. If $p = 1$, then we have the Manhattan formula. If $p = 2$, we have the Euclidian formula [Eq. 2.4].

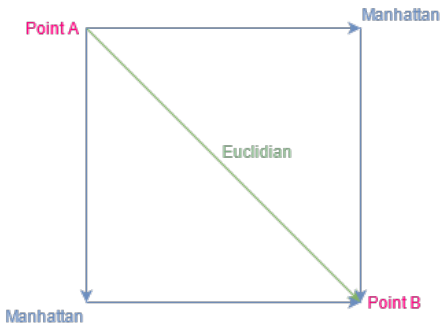


Figure 2.3: Distance Formulae

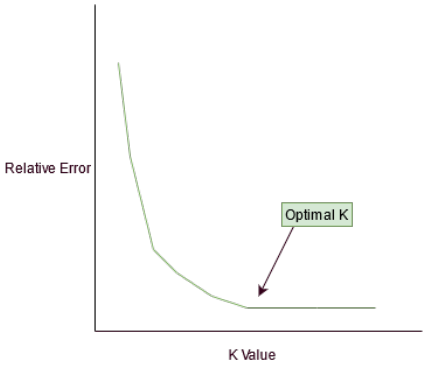


Figure 2.4: Example Elbow Plot

If a new datapoint is close to 2 or more training datapoints, the mode of the training datapoints close by are taken to be the value of the new datapoint. This is subjective relative to the value chosen for k. The optimal k value can be found by modelling for a range of values of k, and plotting the relative error of calculations against the k values. This is known as an elbow plot (Fig 2.4). A rule of thumb used is $k = \sqrt{N}$, where N is the number of training datapoints. So if we had 400 training datapoints, $k = \sqrt{400} = 20$. An elbow plot may prove more useful than this rule however.

kNN is a useful modelling approach as it is simple, intuitive, and does not assume anything about the data it analyses. However, it does suffer from the curse of dimensionality, wherein the more independent variables that are used to predict a dependent variable, the number of dimensions increases [25]. This causes confusion in N-dimensional space for the modelling in measuring distance. This can be mitigated using PCA, which is a dimensionality reduction technique

that can bring the number of dimensions back to 2.

$$D_{Euclidian} = \sqrt{\sum_{j=1}^{k}(x_j - y_j)^2} \tag{2.2}$$

$$D_{Manhattan} = \sum_{j=1}^{k}|x_j - y_j| \tag{2.3}$$

$$D_{Minkowski} = \left[\sum_{j=1}^{k}(|x_j - y_j|)^p\right]^{\frac{1}{p}} \tag{2.4}$$

#### 2.6.1.2 Decision Trees

Decision trees are useful models for classification analysis. Classification trees are developed generally using the CART methodology, which takes the independent variables and uses them to split nodes up, with the goal of predicting a dependent variable. A decision node is one that splits into another decision node and leaf nodes, or just leaf nodes. Fig. 2.5 shows a very basic example of a classification tree workflow to illustrate this. We have two classes here: Malignant or not malignant. Each node in the decision tree is decided using a feature from the data gathered. In this case, 3 features are used, and split until we reach our final

classes. This tree has also been pruned to have a maximum depth of 4 so as to not overfit the data, so it will be generally good on new unseen data.

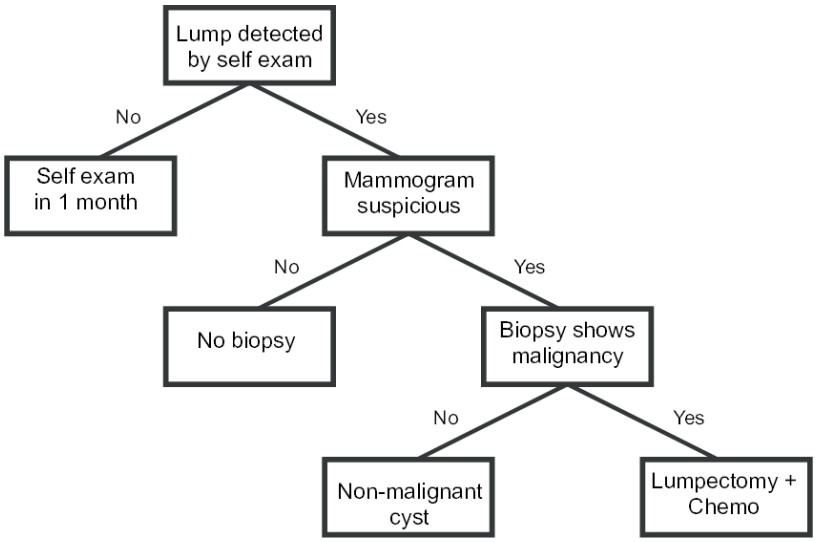


Figure 2.5: Classification Tree Flow Example (Taken from [1])

The splits of decision nodes use loss functions that will determine the best split to make, based on a metric called purity. This can be determined in classification using 2 methods.

One such method is called Gini Impurity [26]. This measures the likelihood that our tree would misclassify a test example. It can also be described as a quantification of the variance across our classes in the dataset.

$$G = \sum_{i=1}^{j} (p_i(1 - p_i)) \tag{2.5}$$

Where $p_i$ is the probability of picking a point from class $i$.

Another method is known as entropy calculation. Entropy is a measure of the proportion of class spread we have within a given node in our tree.

$$E = -\sum_{i=1}^{N}(p_i log_2 p_i) \tag{2.6}$$

Where $p_i$ is the same as the Gini calculation. Entropy is calculated for splits, and a split is only performed if the entropy of the child node is lower than that of the parent node. We can think of this in terms of information gain.

$$IG = E_p - E_c \tag{2.7}$$

The more entropy removed from the parent node in splitting into the child node/s, the more information we can say this feature gives us about the target class.

We want a child node that generates the least variance when splitting a parent node. This calculation is done for the feature variables each time the parent nodes need to be split, and chooses the child node with least variance to proceed with. The tree will look at each feature variable, and check to see which results in the lowest variance when split and choose this variable to split on, and so forth until the tree ends. Classification trees need to be pre-pruned generally, which means we set a max depth for the tree so it doesn't overfit on the training data. Another solution would be to increase the number of trees used and use the majority voting of these trees to make the best model. This is known as a random forest [27].

#### 2.6.1.3 Random Forests

Random forests make use of an ensemble method called bagging to use many decision tree learners in order to enhance a models performance. Random forests allow an individual tree in the forest to grow very large without needing to be

pruned, as we are not as concerned anymore about the high variance of a single tree. A method known as bootstrapping [28] is used to pick random samples from the dataset to train each tree with. This negates overfitting. While a single tree may overfit, as a collective the forest will not be biased to any specific training data.

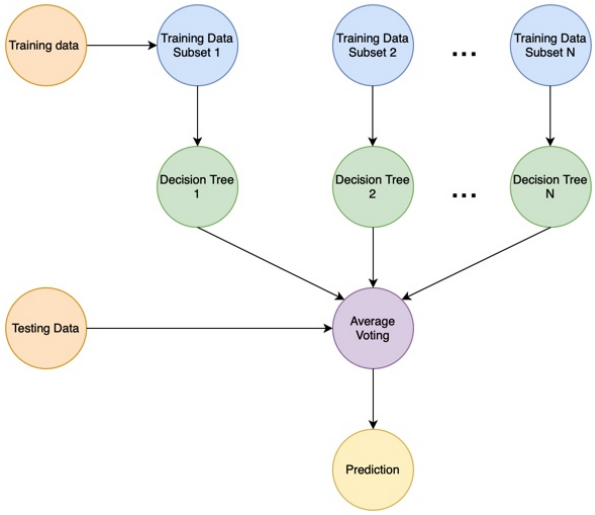


Figure 2.6: Random Forest Flow

When constructing each tree, the variables for each split are chosen randomly, as opposed to using reduction of variance. This again is to ensure that each tree is unique and they do not correlate with one another to lead to a large bias. Hence, we have low variance.

Finally, we repeat this process for $n$ trees. The average prediction of the trees is taken as the prediction for a specific test datapoint.

## 2.7 Neural Networks

Neural Networks are algorithms that are modelled on the structure of the human brain, and try to replicate the process of information retrieval and deduction. They consist of nodes called neurons, which are points that data flows through and is processed. In it's most basic form, a neural network takes input data, puts it through neurons that perform mathematical processes on the input data, and finally return this new processed data as output data [29]. Neural networks have become very popular in recent years, due to the every increasing availability of large quantities of data ,and increased computational power. Outlined below is a form of neural network that has been selected as a possible framework for this thesis.

### 2.7.1 Artificial Neural Networks

ANNs are not a new concept in machine learning. A multitude of machine learning methods can be represented as neural networks. For example, simple linear regression can be viewed as 2 input neurons that are multiplied by a weight and bias (slope and y-intercept here) and added together to reach a single output layer as shown in Fig 2.7.

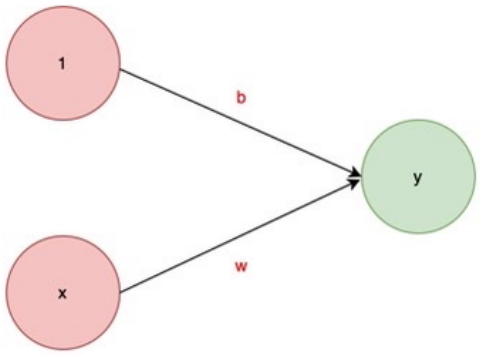


Figure 2.7: Linear Regression Network: $y = wx + b$

In more complex machine learning problems, we need hidden layers that perform processes on the input data, before being output as a prediction. The hidden layers would make use of a linear or ReLU activation function [30] by convention, but any activation function can be used. Finally when the neural network outputs a generation from the output layer, this can be used to make a prediction. Metrics such as MSE or categorical-crossentropy can be used to measure the error of the prediction, depending on the type of problem. This can be fed back to the network for backpropogation, using gradient descent, to update the weights and biases in order to improve the predictions [31; 32]. This is illustrated in Fig 2.8.

A ReLU activation function is more desirable than a linear one, as we can get derivatives of the ReLU function in order to backpropagate the network to improve the results.

$$f(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } x \geqslant 0 \end{cases} \tag{2.8}$$

The ReLU function above has derivatives 0 and 1. Any negative value will end up with an output of 0. The ReLU function is used as it diminishes the occurrence of vanishing gradients during model training. Vanishing gradients occur when the gradients of the activation function get smaller as they reach a global minimum. This can get to the point where the gradients become exponentially small, and hence the gradient descent algorithm will never reach the global minimum. ReLU solves this issue, but can also lead to some of the neurons giving an output of 0.

The use of a neural network as opposed to one of the earlier classification methods outlined could be chosen to avoid manual feature extraction that perhaps could be too complex, while also providing detailed modelling of the dataset that may not be as refined in a simple classification algorithm.

It's important to note that we can use more than one hidden layer. In fact, in more complex scenarios, this is preferred as it will transform input data away from linear format the more layers that are used. One of the benefits of these deep neural networks is that they allow us to build more complex function approximations than simple methods like linear or logistic regression regression.

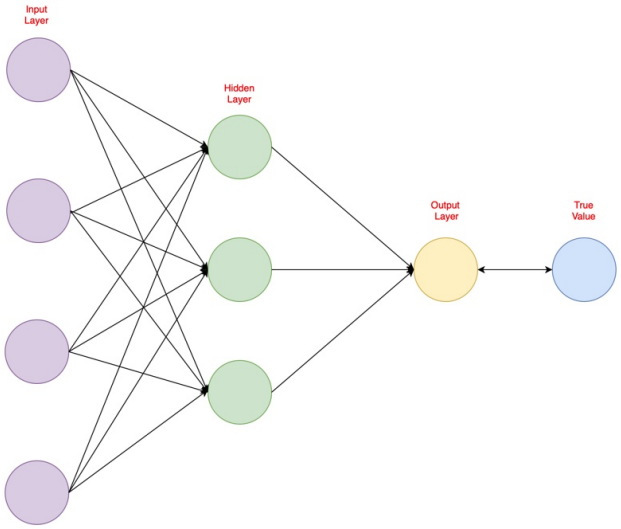


Figure 2.8: Feed Forward Neural Network

# Chapter 3

# Literature Review

## 3.1 Research Method

The search was conducted from Scopus within the NUIG Library system, IEEE Xplore and ScienceDirect, which comprised terms such as "Energy Cost Saving Artificial Intelligence" and "Home Energy Saving". Based on these searches, the most relevant titles were found and scanned for relevant abstracts. If an article was deemed relevant to this thesis' hypothesis, the papers cited by the article were examined to get even more background.

The second set of searches revolved around similar terms, input to Google Scholar. Older papers (pre-2010) were found and scanned for relevance. These papers lay the foundation for machine learning and AI techniques in the field of energy analysis and optimisation. The papers found provide an excellent level of insight that helped guide this review.

Overall, the search proved a success and provided the basis for 3 research questions for this thesis:

$RQ_1$ - What are the data requirements for modelling energy usage for households?

$RQ_2$ - How should machine learning be applied to best predict household energy efficiency?

$RQ_3$ - How can users effectively interact with and gain insight about the household base from this model?

These research questions serve as a foundation for the literature review.

**Notes**

The references included are based on the quality of publication, all of which are peer reviewed, as well as the quality of the abstract. Books were also analysed for relevance and new techniques (both books cited are new editions published in 2021). The older papers were given preference to get a baseline in the work done in the field of energy analytics and henceforth, move on to some of the more recent work, again based on title relevance and abstract quality.

The references provide a varied sample set to work with in terms of methodology and approaches. They are then scanned and analysed if they were deemed to have fallen in line with one or more of the research questions outlined. The references are in IEEE style.

## 3.2 Literary Review

Energy efficiency has been at the forefront of many industries across the world, particularly in the last 50 years, due to the oil crisis of the 1970s. Both in the private and public sectors, officials are looking for ways to incorporate methods of maximising efficient energy for many reasons, but all avenues lead to one end result; a sustainable future for the human race by reducing our carbon footprint.

Narciso et al. [33] presented an overview of 42 of the most reputable papers in the area of energy efficiency over the last 20 years, outlining the models chosen, along with input variables, pre-processing techniques etc. It highlights that despite thorough research, there are still areas that are lacking in their real world application.

The field of energy efficiency is vast and, as such, not all of it is in scope for this review. The focus will be on the technical approaches, such as data processing and model choice. The literature review is broken down by the research questions that needed to be addressed. From there, the research is collated and examined as a whole to gauge what is most relevant to the scope of this project.

### $RQ_1$ : What are the data requirements for modelling energy usage for households?

#### Machine Learning in Buildings

Before devising any model plan, data needs to be the forefront of analysis. There may be a lot of data available to avail of, but is it all relevant? Edwards et al. [34] attempted to solve the issue of needing many input variables in order to make energy models viable in residential buildings. Their data was collected from sensors attached to houses, with 140 measurements taken every 15 minutes. The sensors collected data on temperature and time, as well as previous energy consumption readings.

Ambrose et al. [35] explored all aspects of their data for energy efficiency, and how useful each one was for modelling. For example, energy billing data, can make it hard to determine daily energy patterns if it is calculated on a quarterly basis. Household characteristics like location, BER, size, as well as income and number of residents can be used as strong input variables for a model. Attributes like indoor temperature, smart sensors readings, and what kind of lighting appliances

are used for example are impossible to determine without survey on site, and therefore are beyond the scope of this analysis.

Mason et al. [36] took the approach of using monthly data readings to predict the next month's energy usage, which is more applicable to the data that will be explored in this thesis. The analysis of Satre-Meloy [37] found that household electricity use is best described through socio-demographic and physical dwelling variables, like size of the home, and ownership of an electric vehicle, for example. Finally, it was found that the conversion of research from commercial to residential buildings is not straight forward, as the usage patterns can vary quite dramatically.

### $RQ_2$ : How should machine learning be applied to best predict household energy efficiency?

The data needs to be processed before any modelling can be done effectively. Xiao et al. [38] examined the effect of splitting data according to days. For example, Monday energy usage data was used only to predict the following Monday's usage, taking into account holidays where electricity usage would be sporadic. Interestingly, they also used forecasting to predict historical data, in order to prove its validity.

Zekić-Sušac et al. [39] made use of variable reduction using $\chi^2$ tests of independence for the factor variables and correlations for the numeric variables. Based on the initial 47 variables chosen in the sample, this process determined 10 relevant input variables. The factor variables were mapped to binary categories prior to modelling. Zhu et al. [40] used the method of PLS to find the most relevant input variables, while in a lot of cases other simpler methods like normalisation and outlier removal were utilised [41; 42]. Zhang et al. [43] made use of PCA alongside a neural network in order to encapsulate the most important

inputs in the most concise and efficient form possible.

### Machine Learning in Financial Aid

Machine learning is only useful if it has a clear purpose. In the context of household energy efficiency and sustainability, this purpose needs to be clearly defined. The home owners will more than likely need a loan in order to carry out upgrades, which BERP can tell them the benefits of in relation to how much they stand to save, and how long it will take to recoup on investment. Using this as a basis, the purpose of this analysis is to help customers in their life journey, while also making it as efficient and sustainable as possible. Are they buying a new home? A financial institution like credit union or bank can offer them a green specific loan that give them a good rate of interest.

Models need a platform in which they can be utilised for their purpose. Home upgrades are not a simple matter most of the time. Therefore it is about finding the best way to use the information gathered from the model thereafter, in a real world scenario.

Liang et al. [44] explored the effect of sustainable action on financial institution cost efficiency, and found that financial institutions who embrace sustainable actions like adoption of green products, and reducing their carbon footprint, outperform those that do not embrace these actions. From a business standpoint, it is therefore in their best interests to aid home owners to adapt with the times.

Following from this, Taneja et al. [45] explored customer sentiment towards financial institutions taking up sustainable products. By being transparent and showing their intent for action to aid green initiatives, financial institutions gain approval from their customers. By creating products like the ones mentioned previously, and marketing them in a personalised manner using machine learning models to determine which customers would be most open to such a product, this

creates a perfect integration with existing processes, such as prediction of credit scores, default predictions, and so on. It also helps businesses integrate with individuals in both striving to achieve a common goal while also being realistic about the money driven world we live in. This will lead to a clear definition of energy efficiency in the context of this thesis in a commercial lens as well as residential.

### Feature Engineering and Extraction

Not all data is relevant for modelling the dependent variable, and must be narrowed down. Chou et al. [46] proposed a framework in which only 4 input variables (outdoor temperature in $^{\circ}C$, day of the week, hour of the day and the times at which these values were recorded) were taken to create a number of prediction models, using smart sensors in residential housing. Their results indicated that a hybrid model significantly outperforms single models (by 64%), such as support vector machines and linear regression, when using limited data.

Zekić-Sušac et al. [39] in contrast proposed a model structure with 47 input variables (34 continuous and 13 categorical). They devised a preprocessing technique and gathered a sample of $\approx$ 1500 buildings to train a model, created using a random forest integrated with the Boruta algorithm for variable reduction. They concluded that even with the best model found after cross validation, the accuracy was too low using this small a data source, and required a big data framework in order to get promising results. They determined that the most important input variables for their successful model were: occupational, heating and time data (tying into $RQ_1$).

Zhao et al. [47] used the correlation coefficient between each variable in their data with the dependent variable and chose which were relevant. They also performed RGS, which was developed by Navot et al. [48] for use in study the

study of brain neural activities.

### Model Selection

Model selection is crucial for success, and is unique to the dataset at hand. Edwards et al. [34] found, using environmental sensors, a LSSVM model performed best on residential properties. McLoughlain et al. [49] instead used time series forecasting, specifically Fourier transforms and Gaussian processes, and found both performed very well with the variable nature of electricity usage. Mason et al. [36] explored the usage of ANNs, as they assume less about a dataset than other models do. This combined with the monthly data format proved very effective in comparison to state-of-the-art models. Similarly, CNNs, ELMs, FLNNs and GAs were all applied in use cases for commercial properties, but could be altered to suit new data [50; 51; 52; 43].

## $RQ_3$ : **How can users effectively interact with and gain insight about the household base from this model?**

### Visualisation

Sacha et al. [53] explored integrating data visualisation techniques with machine learning models. They designed a framework in which human interactions were interwoven with machine learning models, so that results could be made more coherent, as well as allowing the changing of parameters to explore what happens on a dynamic basis.

Hadley Wickham's book, Mastering Shiny [54] provides a comprehensive insight into how to produce interactive dashboards, and how to deploy them for external usage, which would be relevant to any model generated information produced by BERP. The ability for someone to easily interact with data that is directly linked to them is beneficial in the adoption and comprehension of findings

from the analysis [55].

Ruff et al. [56] explored the use of a shiny dashboard (R language) to visualise the results of a CNN, to make the model accessible to colleagues with less experience in the field of modelling.

**Data Processing: What methods can be used to convert spend data to energy data?**

Shibano et al. [57] devised a model to convert household income into energy usage data, by linking income to the number of electrical appliances owned and hence, the demand of these appliances, which they deduced to have followed a gamma distribution for simplification. Greer [58] provides case studies in energy cost modelling, that mimic real world modelling scenarios. Geo-spacial data could prove useful when mixed with cost data in determining energy usage by regional means, to give a larger view than a single household. Both can be used in conjunction to provide dual perspectives for home owners [59].

**Note**

The review work done by Narciso et al. [33] is the perfect foundation for any paper on this topic, and will serve as a guide for a lot of the work to come in this thesis.

This thesis will make a unique contribution to sustainability in that (to the best of my knowledge), it will be the first of its kind to investigate BER in relation to home owner accessible information, as well as governing bodies who seek to drive sustainability initiatives, through the use of artificial intelligence. The work of Shibano et al. [57] also should be noted as a key area of conversion for feature engineering.

## 3.3 Review Conclusions

Overall, this review concludes that there is a plethora of evidence to support my hypothesis that energy usage for residential households across ROI can be modelled using generally accessible user household data. At the outset of this topic, 4 questions were set to be answered:

1. What are the strengths, weaknesses, threats and opportunities of AI in the field of energy efficiency?
2. What are the challenges with energy efficiency modelling?
3. What value or benefits can be achieved by AI in energy efficiency modelling?
4. Can we model household energy rating generally available household information?

This research is important as AI can be used to model energy efficiency, whether it be for residential or commercial properties. The caveats however are that usage patterns are a big factor, and therefore the same model cannot be used for both types of properties, unless something is done to mitigate this. AI presents us with the opportunity to allow home owners to see what their current BER is (for free), how it could impact their cost savings if it were elevated, and compare this to averages for households similar to their own, or on a regional basis. In addition, this has the potential to showcase the minimum energy a household requires/needs for a given BER. $500,000$ households are expected to be brought up to a minimum BER B2 in the Republic of Ireland by 2030 [60], and this model could be used to find the minimum improvements needed to reach such a rating, whether the household is already B2 or not (in the eyes of the governing bodies rolling out the changes) and finally the impact to energy spend.

The challenges surrounding the collection and processing of data for modelling are numerous, but, using techniques outlined in this review, can be minimised.

For example, the data collected will be household data, which will be in the form of annual review, and therefore sensor measurements, while providing a good level of background on model input variables and techniques, will be of no benefit to my analysis.

AI can automate tasks that would take an incalculable amount of time for governing staff to complete. The model and dashboard it hypothetically could feed, could be maintained and updated for new data incoming with little effort. The dashboard would promote helping save the environment as well as saving the home owner money in an approachable, simple and user friendly way.

In conclusion, household data can be used to create a model (using AI) and dashboard (using Shiny [54]). This model can predict energy efficiency (BER), show this rating to the relevant bodies (home owner or government sustainability department), what efficiency is versus other households similar to their own, and finally tips on how to reduce energy usage, to reduce cost and improve BER but also benefit the environment. There are no studies in the literature researched that do this.

# Chapter 4

# Data Processing & Analysis

## 4.1 Data Processing

The dataset chosen for this analysis was taken from the Sustainable Energy Authority Of Ireland (SEAI) with permission [61]. This dataset contains information on over one million private dwelling households across ROI, with 211 features recorded for each household. The data does not identify any particular individual and complies with GDPR regulations. The data [61] also contains a user guide, which is a data dictionary that was used to determine the common features that would be easily accessible by home owners without an on-site inspection. The target variable is the BER of the household. There are 15 different classes and therefore makes this a multi-classification problem, ranging from A1 to G (best to worst).

13 features out of 211 were deemed acceptable for accessible usage for model interaction, as the other ones not selected had over 80% null values, or was not easily discernible by the household occupants without outside assistance.

The data was cleaned by evaluating each feature, and determining, through graphical and statistical analysis, if the values made sense practically. For ex-

ample, any households that were built after the year 2022 (the time of writing this report) were dropped, as this is nonsensical data. Datatypes were changed to reflect how they would be used in a modelling scenario (categorical to numerical). Duplicate rows were dropped, and the first one was kept. Distributions and correlations between features were plotted and subsequently analyzed, to determine whether data outliers or multi-collinearity would be an issue.

A general gap analysis was conducted on the cleaned data, which showed a clear issue with 7 of the 13 features. 3 separate techniques were used to handle this using imputation, and compared. Firstly, the null values were dropped, leaving $300,000$ data points. However this left a huge class imbalance for the higher rated households (only 4 A1 rated homes left in dataset) and was therefore not deemed fit for use. The second method was mean/mode imputation. This proved more successful, but again was not sufficient for usage, as the gaps were still too large, and the data did not make real world sense as a result. The final and chosen method was using a random forest to impute the missing values. This was achieved using both a random forest classifier for categorical datatypes and a random forest regressor for numerical datatypes. The effectiveness of this imputer was tested on a sample of the dataset where values were replaced with nulls, the imputer was used and the imputed values were tested against the original. This was 98% accurate in the test and therefore deemed the best approach for use on the entire dataset. kNN imputation was also considered, but after reading [16], it was clear this approach is best suited to datasets with $< 30\%$ null values.

This left a complete dataset with no missing or nonsensical values, without having to drop the majority of our dataset and lose valuable information. In order to model on the dataset, further preparation steps were required.

Feature engineering was conducted on the cleaned, imputed data. This allowed the extraction of the cost of energy for each household per year, based on

the total energy used and the average costs of electricity and gas being applied where appropriate [62]. The categorical features were converted to numerical, using CatBoost encoding as explored in 2. One hot encoding was attempted, but caused data bloating that severely diminished model performance. The features, now all in numeric format, were scaled using a min-max scaler, so that no one feature would dominate the others due to scale. This meant every feature had values between 0 and 1.

The final preprocessing step of the data was sampling. Due to the A1 rated homes being by far a minority class (in ratio of 130:1 to the majority class C3) in the dataset, STL was tested to balance the dataset. This balanced the dataset and led to clearer model decision boundaries.

## 4.2 Analysis

As discussed in Chapter 2, the methods deemed most appropriate for this analysis through background research are:

- K-Nearest Neighbours (kNN) Classification
- Decision Tree Classification
- Random Forest Classification
- Artificial Neural Network Classification

These methods were chosen prior to seeing the data. Hence there is more evidence available in the analysis phase as to why these algorithms would still be deemed suitable to work or not.

kNN was chosen because it is simple to understand and easy to interpret. However upon testing the model, it performed very poorly. This is in large part to do with the scale and span of the dataset. kNN does not do well with high numbers of dimensions. Even without using one hot encoding, 13 dimensions

means the distance between points becomes unclear for the algorithm to discern, and hence is fails.

Decision trees were chosen for interpretability and ease of implementation. When tested, a single tree performed very well, but was unable to model to full scope of data effectively.

Hence, random forests became the next point of focus to create the optimal model. This model performed the best out of the 4 choices above and was chosen as the final model. Random forests excel with large datasets and high dimensions of data, as it subsets the data into trees as discussed in Chapter 2. It is also robust to the outliers in the data, which were due to the A1 household values and therefore were too valuable to remove.

Finally, an ANN was constructed and tested on the data. The ANN was constructed using 6 dense layers, with 1024 neurons in the input later. These layers ended in a 15 neuron, softmax activation function layer, which was used to determine which class each belonged to. The model was constructed to be very deep in order to account for the large number of classes and data points, but ultimately could not perform as well as the random forest. It also requires a huge amount of computational power to train, which gives the random forest model a clear advantage.

All models were tested using their base variants in sklearn using python, with a random seed set to make the analysis replicable. They were each then trained using grid search over a parameter grid, containing permutations of criteria for each model to test, and to come back with the best arrangement of arguments. This was also combined with cross validation, which shuffled the dataset into different test folds, to ensure average performance across the folds was consistent with a single test set. Each optimised model was trained and metrics such as precision, recall, accuracy and AUC were utilised to determine the best model.

# Chapter 5

# Experimental Setup

## 5.1 Data Preparation

The complete experiment is setup as shown in Fig. 5.1. The data from SEAI [61] is ingested into a data platform. A random seed is set, to ensure the results are replicable. The data is then cleaned through removal of nonsensical values, correction of datatypes and movement of target variable (BER) to the end of the dataframe.

Missing Forest imputation [17] is performed on the entire dataset, with 7 of the 13 features containing up to 55% null values. Note that the target variable contained 0 null values prior to imputation phase in this experiment.

The resulting dataset is tested for multi-collinearity, and new features are developed from previous ones to improve coverage. Total energy delivered to a household is used to get an approximation of the energy spend for that household in the last year [62].

The dataset is then split into a training and test split of 80:20 ratio, separating the independent variables (X) and dependent variable (y). This is to ensure the model is valid in a real world scenario on unseen data.

## 5.2 ML Pipeline

The machine learning pipeline is developed as shown in the green cylinder in Fig. 5.1. It is important to note the difference between how the pipeline treats the training data as opposed to the testing data. On the left, the target variable is label encoded to make it suitable with CatBoost Encoding. All of the training data is then put through the CatBoost Encoder, which transforms any categorical features into a numerical representation, and leaves the already numerical features alone. The data is scaled so all values lie between 0 and 1, to ensure no feature dominance occurs and negate the influence of outliers. The training data is oversampled using SMOTE to balance the minority class with the majority, and then undersampled concurrently using TomekLinks to remove any noise/border points in the data. This is done concurrently in one function using SMOTETomekLinks (STL). Finally, this training data is then used to fit a random forest model.

The testing data is put through a similar pipeline process, except it is never fitted, but rather transformed, based on what each encoder/scaler learned from the training data. The testing data is not sampled at all, as this would lead to over-inflation of our real world scenario. The fitted model is used to predict the test set values. Different models are put through the pipeline, and the best base pipeline is taken for grid search evaluation.

A grid search evaluation is run across a parameter grid, and cross validated to get the optimal pipeline arguments. When this process is complete, the optimal model is saved, and can be fed as a back-end into a user friendly dashboard, where household owners can input the features and be told their BER.

The model can be retrained with new batch data ingested, and put through the same cleaning, imputation, engineering and ML pipeline as before.

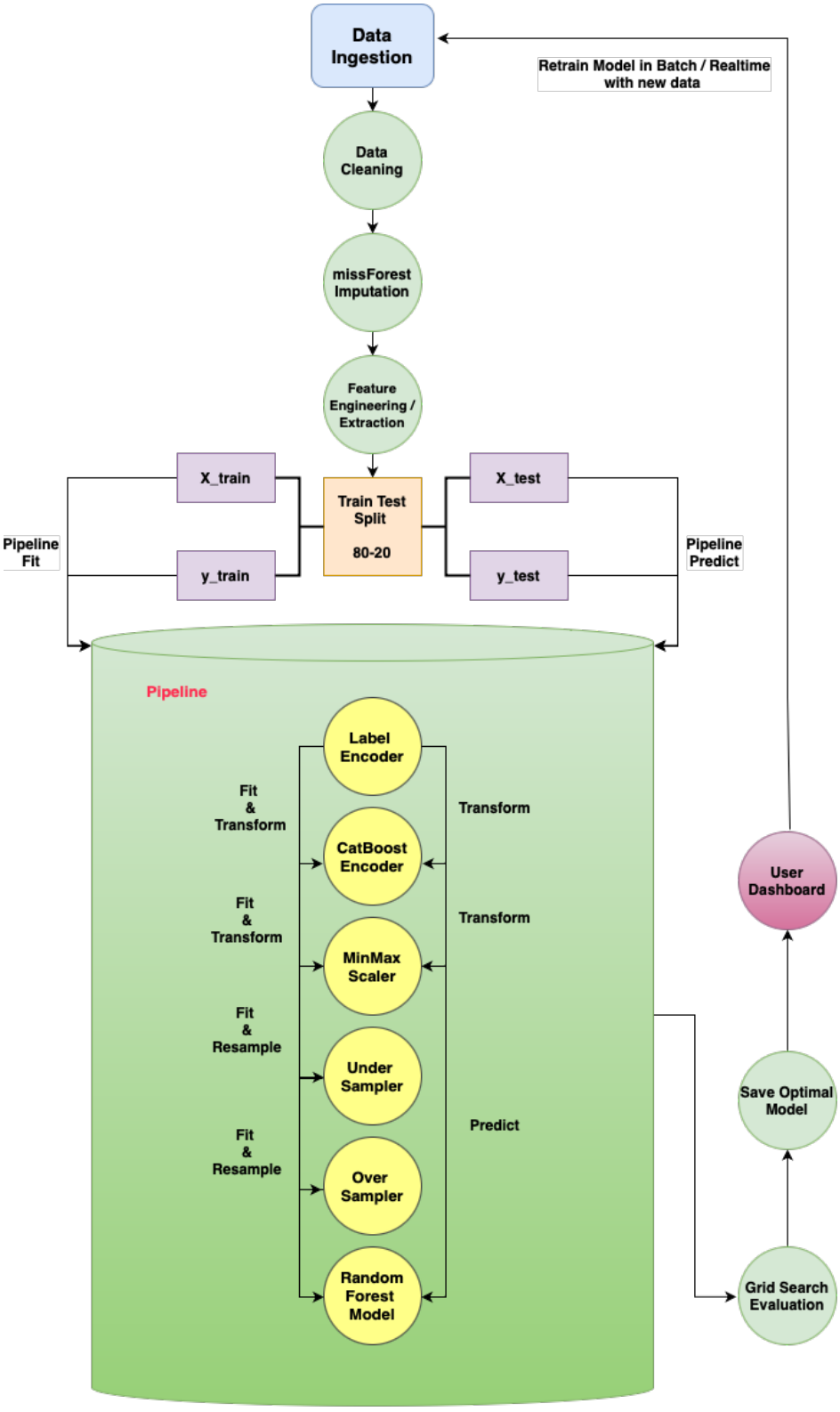


Figure 5.1: BERP Pipeline

# Chapter 6

# Results

## 6.1 Model Evaluation

The optimal model trained seeks to determine the BER of a given household.

Through the experimental method outlined in Chapter 5, the optimal model was determined to be a Random Forest Classifier. The model was able to overcome the dimensionality limitations where kNN fails. The highest optimised kNN model was only able to reach a general F1 score of 0.30, which is not adequate for use. A single decision tree performed well, but naturally over-fitted on the training data, which the Random Forest mitigated. Finally, the ANN performed similarly to the Random Forest model, but struggled on the test data. It also took an order of magnitude longer to train and optimise over the Random Forest model, which is why it was not chosen as the final model. The final model, Building Energy Rating Predictor (BERP), is evaluated in the following sections.

### 6.1.1 Classification Evaluation

In Table 6.1, the classification evaluation metrics chosen for the model are shown. The model is balanced across the classes after STL, and performs well. It struggles to classify and correctly predict A1 rated homes. The overall model has a balanced accuracy score of 0.81, and a macro-F1 score of 0.78, with a very high AUC of 0.99. The model excels at correctly classifying actual classes, and precise in being accurate in this prediction. The predictions are ranked very well according to the AUC scores, regardless of any thresholds within the model.

| *BER* | *Accuracy* | *Precision* | *Recall* | *F1 Score* | *AUC* |
|---|---|---|---|---|---|
| A1 | 0.49 | 0.50 | 0.49 | 0.50 | 0.98 |
| A2 | 0.88 | 0.89 | 0.88 | 0.89 | 1.0 |
| A3 | 0.87 | 0.88 | 0.87 | 0.88 | 1.0 |
| B1 | 0.74 | 0.65 | 0.74 | 0.69 | 0.99 |
| B2 | 0.78 | 0.75 | 0.78 | 0.77 | 0.99 |
| B3 | 0.84 | 0.82 | 0.84 | 0.83 | 0.98 |
| C1 | 0.83 | 0.85 | 0.83 | 0.84 | 0.98 |
| C2 | 0.80 | 0.84 | 0.80 | 0.82 | 0.97 |
| C3 | 0.78 | 0.81 | 0.79 | 0.79 | 0.97 |
| D1 | 0.77 | 0.80 | 0.77 | 0.79 | 0.97 |
| D2 | 0.78 | 0.79 | 0.78 | 0.79 | 0.98 |
| E1 | 0.77 | 0.68 | 0.78 | 0.72 | 0.98 |
| E2 | 0.77 | 0.71 | 0.76 | 0.74 | 0.99 |
| F | 0.83 | 0.77 | 0.83 | 0.80 | 0.99 |
| G | 0.93 | 0.94 | 0.93 | 0.94 | 1.0 |
| **Total Weighted** | **0.81** | **0.78** | **0.79** | **0.78** | **0.98** |
| **Total Macro** | **0.81** | **0.81** | **0.81** | **0.81** | **0.99** |

Table 6.1: Model Results

### 6.1.2 Confusion Matrix

The confusion matrix shown in Fig 6.1 shows the percentage of accuracy across each class. For example, A1 rated homes are being misclassified as A2 and A3 rated homes quite a lot of the time. In all other cases, BERP is generally correct in it's classifications. The misclassifications centre around a rating one up or down from the class. For example, D2 rated homes are being misclassified as being D1 or E1 rated homes, due to the feature space not being able to differentiate them as clearly. Overall, the model has an acceptable level of misclassification, with the exception of A1 rated homes.

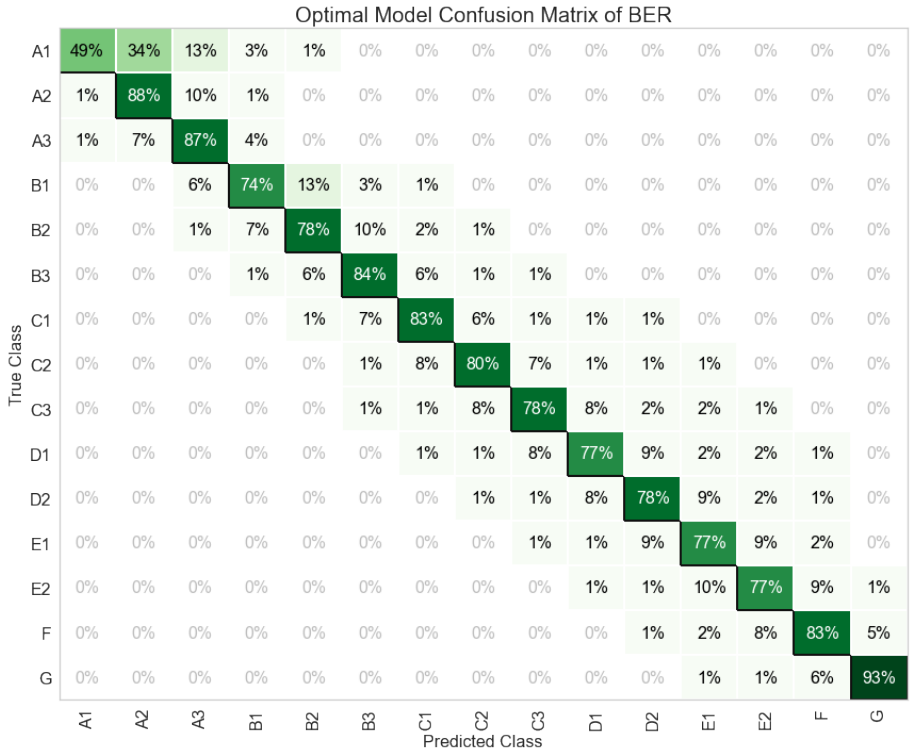


Figure 6.1: Confusion Matrix

### 6.1.3 Area Under the Receiver Operating Characteristics (AUROC)

The Receiver Operating Curve (ROC) is a probabilistic graph, that plots the true positive rate vs. the false positive rate of the BER predictions. It maps our confusion matrix into one plot, that can evaluate the model performance. It uses different probability thresholds for conversion of probabilities BERP generates to class labels, in order to plot the curve. The area under the curve, AUC, is a measure of our models ability to correctly separate classes. As shown in Fig 6.2, the AUC is 0.99 for the macro, and 0.98 for the micro averages of the class space, indicating that the model can almost perfectly separate the classes from one other.

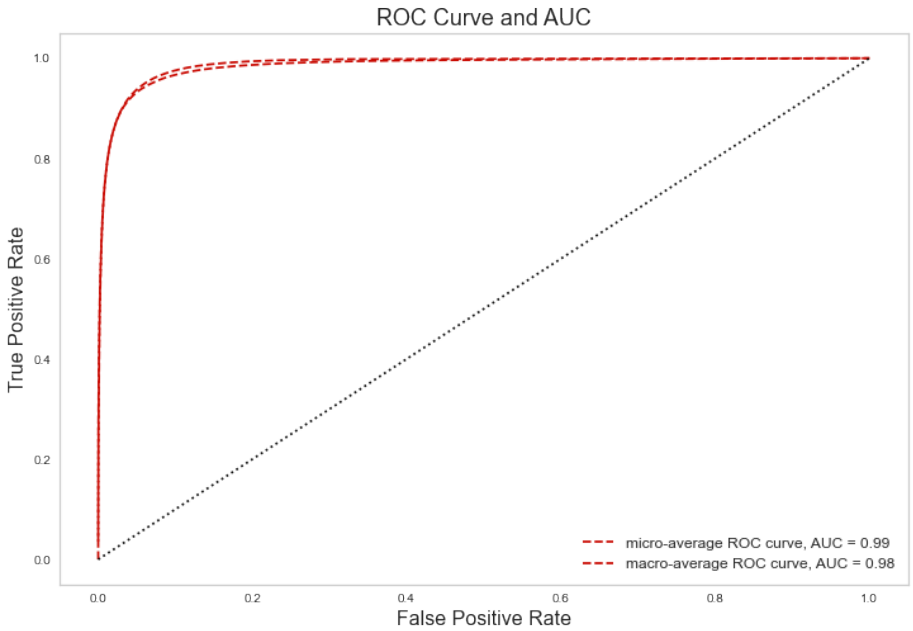


Figure 6.2: ROC Curve

### 6.1.4 Feature Importance

Fig 6.3 shows which features of the model are most useful in determining the BER. The highest importance lies with the ground floor area of a household, contributing 25% of the importance alone. 4 features fall above 10% prediction contribution, while the rest fall under 6%. However, it is important to note that when the features under 6% are removed, and the model is retrained on the remaining features, the A1 scores suffer dramatically, and therefore they are required for that class alone to be acceptable. All chosen features in the model are generally available information for household owners, and can be chosen from a drop-down in our ideal scenario dashboard.

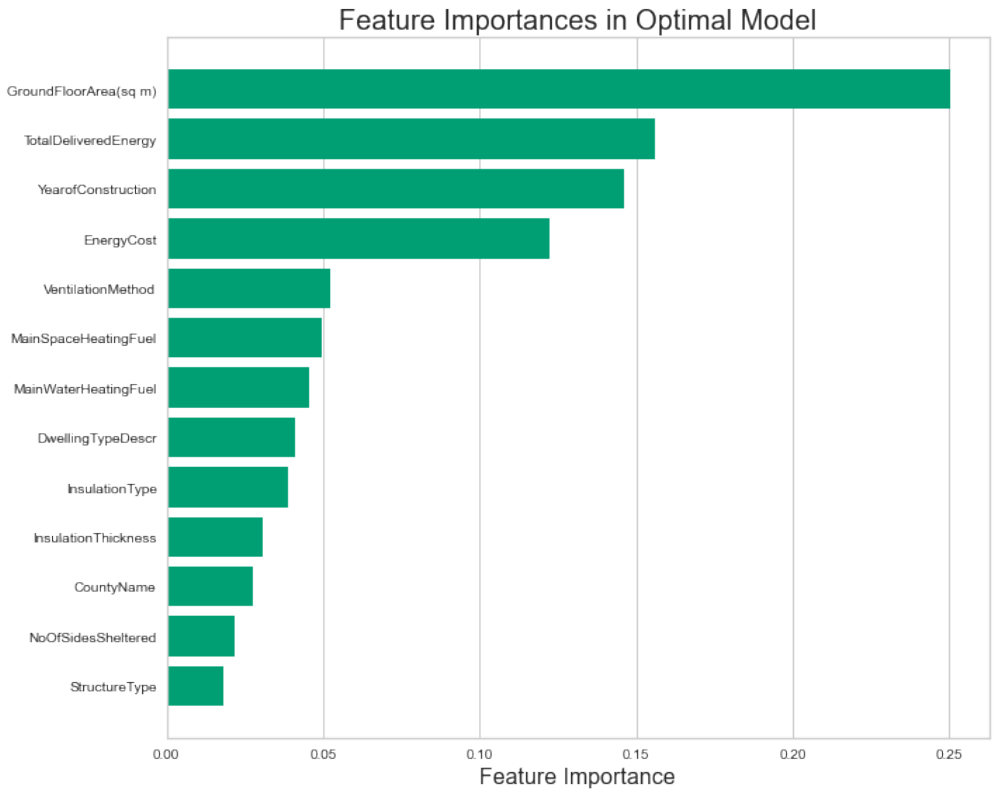


Figure 6.3: Feature Importance

### 6.1.5 Precision Recall Evaluation

The Precision Recall (PR) curve shows a tradeoff between the two metrics BERP, as well as the average precision across the classes. According to Saito et al. [63], PR curves are more descriptive when evaluating imbalanced datasets after sampling. In Fig 6.4, we wish to find how precision affects recall. We seek to optimise recall in our model, as it is more important in this example to find samples as opposed to extreme precision. The gradual curvature of the plot shows that the model is an excellent classifier. The curve shows at different probability thresholds, how precision and recall are affected. In BERP, it is clear the tradeoff lies around 0.81 for both metrics, according to Table 6.1.

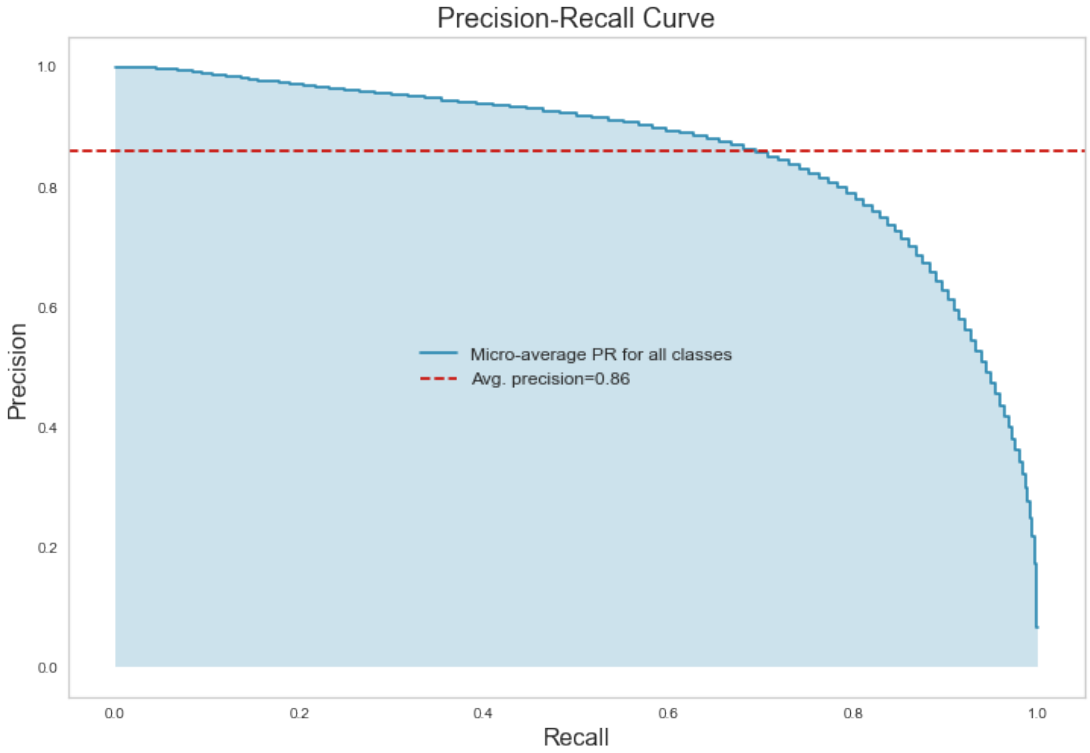


Figure 6.4: Precision-Recall Curve

# Chapter 7

# Discussion

Overall I feel BERP was a success in proving the hypothesis set for this thesis. The previous section dealt with the recording of results. However without context they mean little, and therefore this section aims to give said context.

## 7.1 Model Usage

The fundamental question of any model is: where will it be used and what is important for it to be able to capture? If a home owner is using BERP, the F1 metric is a good baseline to use, as we want to be able to correctly identify their BER, while also making sure that it is precise, so they are informed on what actions they would need to take to increase it to an acceptable level. If it is a governing sustainability body who seek to determine the BER spread across the country and where their focus must lie, then we care most about recall, as we want to be able to identify as many households as possible for each class. However, F1 would still be acceptable here, as we also need them to be precise to ensure that we are targeting the right areas. Therefore, the general F1 score of 0.78 macro and 0.81 weighted across the classes proves and exceeds the hypothesis in both

cases.

### 7.1.1 Interpretation of Confusion Matrix Importances

While A1 is the weakest class in our dataset due to class imbalance, it is not necessarily the most important one to classify in this use case. If an A1 household is miss-classified as an A2 for example, it does little to tell either a home owner, or a governing body, anything different about the property. Home owners with an A rated home will not likely seek to modify it any further, as they most likely already have, or their house was only recently built. The difference in cost will be negligible for the home owner, and the governing body will not be concerned with the difference A1 between A2 households. They are only interested in the general A case, so that they do not need to target these areas for improvement. Therefore, the most important classification for the governing model use case would be the B3 classification. This is the prior rating to the B2 threshold desired. If a house is miss-classified as a B2 (or higher) instead of B3, it means it has been overlooked for improvement. With 6% (with 1% of this coming from C1 being misclassed as B2) of the values being misclassified according the Fig. 6.1, this is an acceptable margin of error, but one to optimise in future runs. This is also true of B2 ratings being misclassed under B2, with 13% of B2 values being misclassed in this fashion. This is not as crucial however, as these households are already at the minimum threshold for BER.

### 7.1.2 Expectations

At the outset of this thesis, I was unsure of how well a model could capture the information used, given the limitations I had set upon it to include only generally accessible information to home owners, to promote usage. The class imbalances also subverted any expectations of excellent model results at the outset during

data preparation. However, through iteration and transformation, the data was able to get to a state where it was perfect for a modelling scenario, without actually changing information in the original data. The model evaluation solely lies on the real world performance of BERP on the test dataset. This dataset was transformed, but only using the transforms learned by fitting the training dataset. The imputation results were also extremely positive, and is something I will be bringing into any future work.

I had expected STL to be a substantial help in the improvement of class imbalance scores in BERP, which to my expectations overcame the limits of simple random over/under sampling techniques. The artificial datapoints for the A1 rated homes helped add information to the model. It was able to improve the recall of the model overall, in particular approximately doubling the recall of A1 from 0.24 to 0.49.

The ANN approach fell short of expectation. The network was setup to be very deep, including normalisation's and dropouts to ensure no 0 weight occurrences. However, the model found it very difficult to discern between classes, even after optimisation. I feel this due to poor model optimisation and could be improved upon given more specified research into DNNs.

The shortcomings of the kNN model lie primarily in the dimensionality and density of the data. There are certain outliers in the dataset that are important to keep in cases of class imbalance, that clearly have affected the distance calculations. The 13 dimensions also make it very hard for distance metrics to be calculated correctly, and lead to substantial model degradation. The decision tree is easier to unravel, in that it performed well, but it was not capable and capturing the full extent of the relationships in a single tree, even when optimised.

The higher classification metrics of lower BER households is a very positive result, as these are the most important to identify, and have the most to gain

from home improvements, in terms of energy savings and campaign results by governing bodies.

The year of construction feature, while not being useful for home owners in that they cannot improve it, allows generalisation. Households in a certain estate/area can generally be assumed to have similar characteristics if built by the same construction companies, and therefore one datapoint could inform 10, 20, or even 100. This is granted no development has been done on any of the houses in the area, but it gives a clear additional geo-spatial data source for BERP in construction agencies. Any future work in this area would excel from inclusion of this kind of data.

The importance of household area over energy usage is surprising. I would have expected the opposite to be the case, but both contribute highly to model predictions. Likewise, I am surprised that the insulation type or thickness is not a higher importance feature. It would impractical for households to decrease ground floor area naturally, but again this characteristic can link households of similar size together when other data is unavailable.

# Chapter 8

# Conclusion

In conclusion, this work has been successful in proving and exceeding the hypothesis set. It has contributed valuable insight into the field of energy research at a residential level, by incorporation of valuable imputation, encoding and sampling techniques from research in AI, that are generally not used together. The A1 rated households are at an acceptable prediction level, but the worst of the model. Even with STL providing a clear uplift in the model, particularly in the A1 class, it still does not provide enough information for BERP to be able to discern between other A rated homes. As discussed previously, this is the ideal place for the model to fail, as this is the least important class for us to identify.

Where things have gone better than expected is in the optimisation and speed of the model pipeline. In experimentation, from ingestion to optimisation, the dataset takes about 90 minutes to complete a training run. This is on a baseline Macbook Air, so in a production environment, on a cluster of high end CPUs, this would be greatly reduced. Looking back to the research questions included at the outset of this thesis, we can examine where they have been answered and where one could direct their focus in future work.

The data requirements for determining household energy efficiency are not

as specific as thought at the outset of this work. Previously, research has been dominated by modelling using sensor measurements and survey data. Using common information about a household shows that a model can be very successful without costly measures. Machine learning is best applied in a tree based decisioning approach, in terms of household energy efficiency. BERP clearly excels at differentiating classes from one another, especially at a lower threshold of poor energy efficiency, due to the majority of households in the training data having a poor efficiency rating. Distance based approaches do not work, due to the high dimensionality of the data, as well as the scale and span.

An ANN could work better in production, but as discussed previously, sufficient resources and expertise were unavailable to fully optimise and explore a DNN approach. While being outside of the scope of this work in creating a dashboard, the effective user interaction with BERP would be best suited to a visual domain. Simple statistical tables and bar charts can be created to show, for the energy spend one household currently consumes, how it would change if the BER was improved, as well as financial aid resources for this kind of upgrade to a home.

Likewise, a separate interface could be used by governing bodies to access BERP to find poorly rated homes in certain areas, certain household sizings or energy usage points. This could add a hierarchy of importance to choose which $500,000$ households would be best suited to target for a BER improvement to B2 by 2030 across ROI [60].

This work impacts the area of residential energy modelling, and makes it more accessible to users by inclusion of simple input features. The techniques used in this thesis to fill data gaps and encode data are also combined in a pipeline, that can be simply downloaded and altered from a repository (with permission and citation) [64].

Testing the model on my childhood home and seeing the correct result, as given by a BER site inspector, is certainly an evaluation metric hard to quantify in words, the sense of accomplishment that comes with it.

In future work in this area, I will be developing said dashboard, trying to incorporate more abundant data sources from construction companies, or public datasets, and finally create a business property variant for commercial use.

# References


[1] J. Cruz and D. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, pp. 59–77, 02 2007. v, 16

[2] A. Kaya and M. Keyes, "Energy management technology in pulp, paper and allied industries," *IFAC Proceedings Volumes*, vol. 13, no. 4, pp. 609–622, 1980, 4th IFAC Conference on Instrumentation and Automation in the Paper, Rubber, Plastics and Polymerisation Industries, Ghent, Belgium, 3-5 June 1980. 1

[3] H. Kang, M. Lee, T. Hong, and J.-K. Choi, "Determining the optimal occupancy density for reducing the energy consumption of public office buildings: A statistical approach," *Building and Environment*, vol. 127, pp. 173–186, 2018. 1

[4] A. Roslizar, M. A. Alghoul, B. Bakhtyar, N. Asim, and K. Sopian, "Annual energy usage reduction and cost savings of a school: End-use energy analysis," *The Scientific World Journal*, vol. 2014, pp. 310–539, Nov 2014. 1

[5] J. A. Rodger, "A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public

buildings," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 1813–1829, 2014. 1

[6] M. Santamouris, G. Mihalakakou, P. Patargias, N. Gaitani, K. Sfakianaki, M. Papaglastra, C. Pavlou, P. Doukas, E. Primikiri, V. Geros, M. Assimakopoulos, R. Mitoula, and S. Zerefos, "Using intelligent clustering techniques to classify the energy performance of school buildings," *Energy and Buildings*, vol. 39, no. 1, pp. 45–51, 2007. 1

[7] D. Torregrossa, U. Leopold, F. Hernández-Sancho, and J. Hansen, "Machine learning for energy cost modelling in wastewater treatment plants," *Journal of Environmental Management*, vol. 223, pp. 1061–1067, 2018. 1

[8] C. Department of the Environment and Communications, "2050 Net-Zero Act," 08 2021. [Online]. Available: https://www.gov.ie/en/press-release/9336b-irelands-ambitious-climate-act-signed-into-law/ 2

[9] T. W. Edgar and D. O. Manz, "Chapter 6 - machine learning," in *Research Methods for Cyber Security*, T. W. Edgar and D. O. Manz, Eds. Syngress, 2017, pp. 153–173. 4

[10] V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 631–641, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION , 2019 November 11-12, 2019. 5

[11] S. Lee, Y. Kim, H. Kahng, S.-K. Lee, S. Chung, T. Cheong, K. Shin, J. Park, and S. B. Kim, "Intelligent traffic control for autonomous vehicle systems based on machine learning," *Expert systems with applications*, vol. 144, p. 113074, 2020. 5

[12] I. P. Adegun and H. B. Vadapalli, "Facial micro-expression recognition: A machine learning approach," *Scientific African*, vol. 8, p. e00465, 2020. 5

[13] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021, cENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020. 6

[14] R. Wirth and J. Hipp, "Crisp-dm: towards a standard process modell for data mining," 2000. 6

[15] H. Wickham, "Tidy data," *The American Statistician*, vol. 14, 09 2014. 6

[16] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1487–1509, Feb 2020. 8, 33

[17] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 10 2011. 8, 36

[18] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018. 9

[19] Yandex, "Catboost Yandex," 06 2022. [Online]. Available: https://catboost.ai/en/docs/concepts/algorithm-main-stages_cat-to-numberic 10

[20] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *CoRR*, vol. abs/1106.1813, 2011. 11

[21] I. Tomek, "Two modifications of cnn," 1976. 12

[22] G. E. A. P. A. Batista, A. L. C. Bazzan, and M. C. Monard, "Balancing training data for automated annotation of keywords: a case study," in *WOB*, 2003. 12

[23] M. Alloghani, D. Al-Jumeily Obe, J. Mustafina, A. Hussain, and A. Aljaaf, *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*, 01 2020, pp. 3–21. 13

[24] J. S. Richman, "Chapter thirteen - multivariate neighborhood sample entropy: A method for data reduction and prediction of complex data," in *Computer Methods, Part C*, ser. Methods in Enzymology, M. L. Johnson and L. Brand, Eds. Academic Press, 2011, vol. 487, pp. 397–408. 13

[25] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966. 14

[26] W. L. Dunn and J. K. Shultis, "5 - variance reduction techniques," in *Exploring Monte Carlo Methods*, W. L. Dunn and J. K. Shultis, Eds. Amsterdam: Elsevier, 2012, pp. 97–132. 16

[27] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, pp. 14 – 23, 01 2011. 17

[28] X. Chen, Z. Y. Dong, K. Meng, Y. Xu, K. P. Wong, and H. W. Ngan, "Electricity price forecasting with extreme learning machine and bootstrapping,"

*IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 2055–2062, 2012. 18

[29] A. Prieto, B. Prieto, E. M. Ortigosa, E. Ros, F. Pelayo, J. Ortega, and I. Rojas, "Neural networks: An overview of early research, current frameworks and new challenges," *Neurocomputing*, vol. 214, pp. 242–268, 2016. 19

[30] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018. 20

[31] E. Guresen and G. Kayakutlu, "Definition of artificial neural networks with comparison to other networks," *Procedia Computer Science*, vol. 3, pp. 426–433, 2011, world Conference on Information Technology. 20

[32] S. Ruder, "An overview of gradient descent optimization algorithms," 2017. 20

[33] D. A. Narciso and F. Martins, "Application of machine learning tools for energy efficiency in industry: A review," *Energy Reports*, vol. 6, pp. 1181–1199, 2020. 24, 29

[34] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy and Buildings*, vol. 49, pp. 591–603, 2012. 24, 28

[35] M. Ambrose and M. James, "Dealing with energy efficiency data," *Energy Procedia*, vol. 121, pp. 158–165, 2017, improving Residential Energy Efficiency International Conference, IREE 2017. 24

[36] K. Mason, J. Duggan, and E. Howley, "Forecasting energy demand, wind generation and carbon dioxide emissions in ireland using evolutionary neural networks," *Energy*, vol. 155, pp. 705–720, 2018. 25, 28

[37] A. Satre-Meloy, "Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models," *Energy*, vol. 174, pp. 148–168, 2019. 25

[38] L. Xiao, J. Wang, X. Yang, and L. Xiao, "A hybrid model based on data preprocessing for electrical power forecasting," *International Journal of Electrical Power  Energy Systems*, vol. 64, pp. 311–327, 2015. 25

[39] M. Zekić-Sušac, A. Has, and M. Knežević, "Predicting energy cost of public buildings by artificial neural networks, cart, and random forest," *Neurocomputing*, vol. 439, pp. 223–233, 2021. 25, 27

[40] L. Zhu and J. Chen, "Energy efficiency evaluation and prediction of large-scale chemical plants using partial least squares analysis integrated with gaussian process models," *Energy Conversion and Management*, vol. 195, pp. 690–700, 2019. 25

[41] Y. Han, C. Fan, M. Xu, Z. Geng, and Y. Zhong, "Production capacity analysis and energy saving of complex chemical processes using lstm based on attention mechanism," *Applied Thermal Engineering*, vol. 160, p. 114072, 2019. 25

[42] B. Beisheim, K. Rahimi-Adli, S. Krämer, and S. Engell, "Energy performance analysis of continuous processes using surrogate models," *Energy*, vol. 183, pp. 776–787, 2019. 25

[43] X.-H. Zhang, Q.-X. Zhu, Y.-L. He, and Y. Xu, "Energy modeling using an

effective latent variable based functional link learning machine," *Energy*, vol. 162, pp. 883–891, 2018. 25, 28

[44] L.-W. Liang, H.-Y. Chang, and H.-L. Shao, "Does sustainability make banks more cost efficient?" *Global Finance Journal*, vol. 38, pp. 13–23, 2018, special Issue on Corporate Social Responsibility and Ethics in Financial Markets. 26

[45] S. Taneja and L. Ali, "Determinants of customers' intentions towards environmentally sustainable banking: Testing the structural model," *Journal of Retailing and Consumer Services*, vol. 59, p. 102418, 2021. 26

[46] J.-S. Chou and D.-S. Tran, "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders," *Energy*, vol. 165, pp. 709–726, 2018. 27

[47] H. Zhao and F. Magoulès, "Feature selection for predicting building energy consumption based on statistical learning method," *Journal of Algorithms and Computational Technology*, vol. 6, pp. 59 – 77, 2012. 27

[48] A. Navot, L. Shpigelman, N. Tishby, and E. Vaadia, "Nearest neighbor based feature selection for regression and its application to neural activity," vol. 18, 01 2005. 27

[49] F. McLoughlin, A. Duffy, and M. Conlon, "Evaluation of time series techniques to characterise domestic electricity demand," *Energy*, vol. 50, pp. 120–130, 2013. 28

[50] Z. Geng, Y. Zhang, C. Li, Y. Han, Y. Cui, and B. Yu, "Energy optimization and prediction modeling of petrochemical industries: An improved convolutional neural network based on cross-feature," *Energy*, vol. 194, p. 116851, 2020. 28

[51] Y.-L. He, P.-J. Wang, M.-Q. Zhang, Q.-X. Zhu, and Y. Xu, "A novel and effective nonlinear interpolation virtual sample generation method for enhancing energy prediction and analysis on small data problem: A case study of ethylene industry," *Energy*, vol. 147, pp. 418–427, 2018. 28

[52] M. Kovačič and B. Šarler, "Genetic programming prediction of the natural gas consumption in a steel plant," *Energy*, vol. 66, pp. 273–284, 2014. 28

[53] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "What you see is what you can change: Human-centered machine learning by interactive visualization," *Neurocomputing*, vol. 268, pp. 164–175, 2017. 28

[54] H. Wickham, *Mastering Shiny.* O'Reilly, 2021. 28, 31

[55] Y.-S. Kim, K. Reinecke, and J. Hullman, "Explaining the gap: Visualizing one's predictions improves recall and comprehension of data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1375–1386. 29

[56] Z. J. Ruff, D. B. Lesmeister, C. L. Appel, and C. M. Sullivan, "Workflow and convolutional neural network for automated identification of animal sounds," *Ecological Indicators*, vol. 124, p. 107419, 2021. 29

[57] K. Shibano and G. Mogi, "Electricity consumption forecast model using household income: Case study in tanzania," *Energies*, vol. 13, no. 10, 2020. 29

[58] M. Greer, *Electricity Cost Modelling Calculations - 2nd Edition.* Academic Press, 2021. 29

[59] M. V. Rocco, E. Fumagalli, C. Vigone, A. Miserocchi, and E. Colombo, "Enhancing energy models with geo-spatial data for the analysis of future electrification pathways: The case of tanzania," *Energy Strategy Reviews*, vol. 34, p. 100614, 2021. 29

[60] E. G. on Future Skills Needs, "2030 BER B2 Target," 11 2021. [Online]. Available: https://enterprise.gov.ie/en/publications/skills-for-zero-carbon.html 30, 50

[61] SEAI, "SEAI Data," 06 2022. [Online]. Available: https://ndber.seai.ie/BERResearchTool/ber/search.aspx 32, 36

[62] Sustainable Energy Authority Of Ireland (SEAI), "Average Electricity & Gas prices Ireland 2022," 04 2022. [Online]. Available: https://www.seai.ie/data-and-insights/seai-statistics/key-statistics/prices/ 34, 36

[63] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, pp. 1–21, 03 2015. 44

[64] O. Brannock$^{TM}$, "Building Energy Rating Predictor (BERP) Repository," 08 2022. [Online]. Available: https://github.com/OisinB-2814/masters_thesis_ob2814 50