C O V E N T R Y
U N I V E R S I T Y

Faculty of Engineering, Environment and Computing

School of Computing, Mathematics and Data Science

MSc. Data Science

7150CEM
Data Science Project

Predicting Anomalous Electricity Consumption Patterns at the Postcode Level: Machine Learning Approach

Author: Chinedu Joseph Ezenwajiaku

SID: 14227905

Supervisor: Dr Diana Hintea

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science
Academic Year: 2023/24

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Chinedu Joseph Ezenwajiaku Date: 28 July 2024

Please complete all fields.

First Name:	CHINEDU
Last Name:	EZENWAJIAKU
Student ID number	14227905
Ethics Application Number	P177443
1st Supervisor Name	Dr. Diana Hintea
2nd Supervisor Name	Dr. Peter Every

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

In an era of increasing electricity demands and environmental concerns, understanding electricity consumption patterns have become crucial for effective energy management and policy-making. Anomalies in electricity consumption can indicate various issues, from infrastructure problems to potential fraudulent activities. By identifying these anomalies, energy providers, policymakers, and consumers can take proactive measures to address inefficiencies and optimize energy supply and distribution. This study focuses on developing an advanced machine learning approach to detect and predict anomalous electricity consumption at the postcode level in the United Kingdom. The major idea is to provide a more targeted and comprehensive analysis than previous studies by examining consumption patterns at the postcode level and incorporating demographic factors and consumption metrics that are often overlooked using a combination of census data and electricity consumption records. This approach allows for a deeper understanding of how demographic factors relate with energy usage patterns, potentially revealing insights that might be missed when examining consumption data in isolation. This research explores various machine learning algorithms for both anomaly detection and predictive modelling, aiming to capture the intricate, nonlinear relationships between different variables that influence electricity consumption.

The implications of this research extend beyond mere anomaly detection. The insights gained from this study could contribute to more effective energy conservation strategies, better-informed urban planning, and the development of tailored energy policies. This also represents a significant step forward in the application of machine learning to energy consumption analysis. By providing a more nuanced understanding of anomalous electricity usage patterns, it offers valuable insights that could shape the future of energy distribution and consumption in the United Kingdom and potentially beyond.

TABLE OF CONTENT

Acknowledgements

I would like to thank my supervisor Dr. Diana Hintea, the completion of this study could not have been possible without her expertise and guidance. I would also like to thank Dr. Chinonso Ezenwajiaku for his immense support.

A debt of gratitude is owed to the entire management of the Niger Delta Development Commission, Nigeria for their sponsorship of my program and consequently this study.

Finally, I would like to appreciate the support and motivation from my parents, siblings, and friends throughout the process of this study.

CHAPTER 1: INTRODUCTION

1.1. | OVERVIEW OF ELECTRICITY CONSUMPTION PATTERNS AND THEIR IMPORTANCE

Electricity consumption patterns in modern societies can be an indicator to the electricity needs of various areas in any settlement, be it residential, commercial or industrial. To effectively manage energy, these patterns must be understood and analysed. Consumption pattern analysis can reveal much about energy usage, highlight potential inefficiencies, and point to ways that distribution and production of energy can be optimised (Gajowniczek and Ząbkowski, 2017). Electricity supply companies have to monitor and forecast electricity consumption patterns so that they are in a position to offer a reliable and cost-effective supply of energy while trying to meet the rising electricity demand precipitated by population growth and industrialisation (Kan et al., 2021).

While viewing electricity consumption at a macro level, like on a regional or national level, gives insights into the overall pattern, there is an advantage in understanding electricity consumption on a much more granular level. Postcodes which are granular geographic units will provide the opportunity to dissect consumption trends in a much more local and focused approach. This level of granular identification of anomalous electricity consumption might be very significant for utilities or energy providers in geolocating to where particular areas or even neighbourhoods are behaving irregularly (Suganthi et al., 2012). Such information would assist in targeted interventions, resource allocation, and tailored energy conservation strategies. Moreover, the anomalies that are detected at the postcode level allow potential problems like faulty equipment and illegal activities to be recognised early enough so that appropriate corrective measures could be introduced.

1.2 | PROBLEM STATEMENT AND MOTIVATION

As argued by Fried and Lagakos (2023), electricity is one of the major driving forces of the economy and development in the world today. It can also be said that electricity consumption can be directly linked to how economically developed a geographical region is. Therefore, understanding consumption patterns becomes important for energy providers and policymakers to make decisions that translate to effective electricity distribution.

However, there is an increasing complexity in electricity consumption patterns driven by various factors like lifestyle changes and technological advancements (Piao & Managi, 2023). This has made it difficult for energy providers to optimise and manage distribution effectively. Traditional statistical methods of analysing electricity consumption are mostly limited in their ability to detect and predict abnormal patterns (Schmidl et al., 2022), especially at granular levels. Such limitations can lead to inefficient management of energy, increased operational costs, and potential oversights in detecting fraudulent activities or infrastructural issues thus hindering targeted interventions and energy conservation strategies.

The problem is further compounded by the vast amount of data generated by modern smart meters and other monitoring devices, which require sophisticated analytical methods to draw meaningful insights. Using traditional methods, it is often difficult to simultaneously consider multiple important variables, such as demography, weather conditions and socioeconomic activities which can significantly influence electricity consumption (Schmidl et al., 2022). It then becomes necessary to adopt a more adaptive and robust approach to anomaly detection and prediction.

This research aims to address these challenges by developing a machine learning-based approach capable of simultaneously considering multiple variables in predicting anomalous electricity consumption at the postcode level. By leveraging advanced algorithms and using relevant and recent data sources, this study seeks to create a more accurate, adaptive, and interpretable model that can help enhance energy management practices, improve grid stability, and facilitate the transition towards more sustainable energy systems.

1.3 | OBJECTIVES

The primary objective of this research is to develop a machine learning model capable of accurately predicting patterns that could indicate unusual electricity consumption of buildings at the postcode level. By leveraging electricity consumption data from the United Kingdom's Economy 7 meters and various relevant features, such as postcodes, and human population, the proposed model aims to identify deviations from expected patterns, which can enable proactive monitoring and response. The machine learning techniques used in this research include supervised learning algorithms, such as decision trees, KNN, gradient boosting and random forest, and anomaly detection algorithms such as isolation forest and local outlier factor.

1.4 | RESEARCH QUESTIONS

In order to achieve the basic objective of this study, this research will attempt to answer the following research questions

1. How effective are different outlier detection methods (Local Outlier Factor, Isolation Forest, and Ensemble Method) in identifying anomalous electricity consumption patterns?
2. Which machine learning model performs best in predicting anomalous electricity consumption patterns and how effective are they?
3. What are the most important factors or features that contribute to predicting anomalous electricity consumption patterns?
4. How do demographic factors influence the prediction of anomalous electricity consumption?

1.5 | SCOPE AND LIMITATIONS

Scope:

- This research will focus on predicting unusual electricity use patterns at the postcode level in the UK.
- This research will use data from Economy 7 Meters in conjunction with the UK census data to draw insights on electricity consumption and develop a model for predicting abnormal usage of electricity.
- This study will look at factors like population, and meter statistics that might affect electricity use.
- This research will adopt Isolation Forest, Local Outlier Factor, Random Forest, Decision Trees, KNN, Gradient Boosting algorithms to find the best way to detect and then predict abnormal electricity usage.
- This research will try to create a model that can work irrespective of area type, from urban to rural settlements thus helping energy companies and policymakers can use to make better decisions.

Limitations:

- This study will only use data from the UK, so the models and insights drawn might not work the same way in other countries.

- This study will rely on the accuracy of the data from Economy 7 Meters and the UK census for the year 2022. If these datasets have errors, it could affect the results.
- The model in this study will be based on past data. If there are big changes in how people use electricity in the future, the model might become less accurate.
- Privacy concerns might limit how detailed the data can be, which could affect how precise the predictions are.
- This research may not be able to account for all possible unusual situations that could significantly affect electricity consumption.

1.6 | SIGNIFICANCE OF PROJECT

Energy supply companies who are the major target for this research can apply the results of this study in various ways as listed below:

1. Energy Efficiency and Conservation: Detecting anomalous electricity consumption patterns to a postal code level can help identify specific locations where there is excessive or unusual energy usage. This can help energy conservation efforts and sustainable energy practices.
2. Grid optimisation and load balancing: Identifying postcodes that show anomalous consumption patterns can help utility companies optimise load balancing and resource allocation. This information can help prevent overloads or outages by controlling supply.
3. Maintenance and fault detection: Patterns of abnormal electricity consumption at the postcode level may reflect some underlying problems with the distribution network or faulty equipment. This would facilitate early maintenance and timely repair, therefore avoiding possible outages or failures.
4. Fraud detection and revenue protection: Unusual consumption patterns at the postcode level could indicate electricity theft or meter tampering. By detecting such anomalies, utility companies can protect their revenues from loss by taking the appropriate measures.

1.7 | REPORT STRUCTURE

The subsequent sections of this report will be split into four chapters, each focusing on different facets of this research that all comes together to thoroughly explain the project. The chapters are as outlined below:

Chapter 2 covers the literature review of this study. This section gives a foundational overview of the existing works on this research topic while also identifying some gaps which will form the basis of the research carried out in this study.

Chapter 3 will focus on the theoretical framework, methodology and the experimental design adopted in the carrying out of the experiments. It also explains the rationale behind the adoption of these methods

Chapter 4 presents the results obtained from the experiments. The results will be explained while discussing the findings from the results

Chapter 5 covers the project management, the ethical considerations and the challenges faced during the process of carrying out this research

Furthermore, additional information as relating to the study will be provided in the Appendix section for further reference.

CHAPTER 2: LITERATURE REVIEW

This literature review is an overview of some methodologies and techniques used in the study of electricity consumption patterns, anomaly detection, and predictive modelling. From traditional statistical approaches to modern machine learning and artificial intelligence methods. This literature review is presented to provide a synoptic view of the existing research landscape on this topic, highlighting its strengths and limitations, along with potential avenues for further research.

2.1 | OVERVIEW OF EXISTING RESEARCH ON ELECTRICITY CONSUMPTION PATTERNS

This section is devoted to past researches aimed at understanding the characterisation of electricity consumption patterns in various settings. It covers the introduction of exogenous factors like weather conditions, occupancy patterns, and socioeconomic indicators into the models to enhance their accuracy and robustness.

Ramos et al (2022) presented a contextual reinforcement learning approach to electricity demand forecasting in buildings. While it is innovative to combine decision trees, reinforcement learning, and multiple forecasting techniques, there are several significant limitations. A single week of data was used from just one building, which cannot represent many different consumption patterns across buildings and over time sequences. Furthermore, the effectiveness of the methodology depends on how accurate the decision tree is in determining the relevant contexts. It may turn out to be hard in more complex scenarios with multiple influencing factors. There is no comprehensive comparison done against other state-of-the-art methods of forecasting, which makes it hard to assess the relative performance of the proposed approach for a reader. While this is an exciting approach, more extensive evaluation across diverse data sets, comparisons with existing methods, and analysis concerning practical applications would better assess its potential for predicting anomalous electricity consumption.

Another study by Wang, Zhang, & Chen, (2021) considered a short-term residential load forecasting model with weather features as input and uses a long-term memory recurrent neural network. This showed that an accurate short-term load forecasting for residential users can promote power system operation and facilitate demand response programs; further, it has rarely been addressed in previous works, which are highly dependent on weather factors. They used the standardized cross-correlation method to determine which correlation of different weather features is most relevant for

residential loads. While choosing the most important weather feature, apparent temperature, as additional input to the LSTM model with historical load data and holiday information, results were found to be improved in comparison to those models without weather features. However, the authors used fixed hyperparameters for all households. Tuning parameters individually for each household could improve results. Also the potential impact of different factors on residential load forecasting, like occupancy patterns and appliance usage behavior could be further explored.

The paper, in general, contributes to the understanding of how different weather factors impact short-term residential load forecasting and puts forward a practical way to integrate weather information into machine learning models. Further studies are expected in which more sophisticated methods will be considered about additional factors that could influence residential electricity consumption patterns.

Beretta et al. (2020) performed an interesting application of functional principal component analysis to model and forecast electricity consumption patterns at several spatial scales. In doing so, they used data from the Milan metropolitan area to demonstrate the efficiency of FPCA in representing fundamental physical and behavioural causes affecting these consumption patterns. They perform a detailed decomposition of patterns into principal functions while analysing their scores to infer how exogenous factors, such as weather, calendar events, and customer characteristics, may impact sales. They also provided a linear prediction model based on FPCA scores that works quite well in both short-term and long-term forecasting. Although this is a very detailed study, however, including a comparison between the FPCA-based approach and other well-known electricity load forecasting methods, for instance, time series models or neural networks would help one understand its relative strengths and limitations. Also, looking into some more advanced machine learning techniques in terms of score prediction might further increase the accuracy of such electricity load forecasting.

Support Vector Machines (SVMs) have shown promise for predicting electricity consumption patterns. Fuadi et al. (2021) applied SVM regression to forecast short-term electrical load in a laboratory setting. They used 11 weeks of historical consumption data to predict the following week's usage. The researchers optimized SVM hyperparameters using GridSearchCV, finding optimal values of C=10^6 and gamma=2.97x10^-7. Their model achieved strong predictive performance, with RMSE of 0.37, MAE of 0.21, and MSE of 0.14 when comparing predictions to actual consumption. Their findings lent strong support that SVM can effectively learn

electricity usage patterns from historical data to make accurate short-term forecasts, demonstrating its potential for anomaly detection applications.

Recent research has explored various machine learning approaches for short-term electricity consumption forecasting. Yuan et al. (2022) proposed an integrated model combining empirical mode decomposition (EMD) and long short-term memory networks (LSTM). This approach decomposes electricity consumption data into several intrinsic mode functions using EMD, then applies LSTM to predict each component independently before aggregating results. The authors found this method reduced forecasting errors by about 15% compared to direct forecasting, as it better captured the characteristics of different frequency components in the data. In view of their findings, there is a possibility that such decomposition techniques could potentially be applied to the prediction of anomalous consumption patterns by isolating irregular components in the signal.

Al Metrik and Musleh (2022) investigated electricity consumption prediction in Saudi Arabia using machine learning techniques. They employed Artificial Neural Networks (ANN) and an ensemble approach on a dataset containing monthly electricity consumption for 18 locations across the country. After preprocessing and optimizing the ANN model, they found that the ensemble outperformed the standalone ANN, achieving a correlation coefficient of 0.9116 and a mean absolute percentage error of 0.2836. Their study highlights the potential of ensemble methods in improving prediction accuracy for electricity consumption. While focused on city-level predictions, their approach could be adapted for postcode-level anomaly detection in electricity consumption patterns.

2.2 | TRADITIONAL METHODS FOR DETECTING ANOMALIES

Anything that deviates from the normal either negatively or positively can be considered as an anomaly (*OxfordLearnersDictionaries.com*, n.d.). In terms of electricity consumption, this could be a pattern which shows outliers with values deviating considerably from the other consumption values (Copiaco et al., 2023). Before the adoption of machine learning techniques, researchers used to work with traditional statistical methods and rule-based approaches for anomaly detection in electricity consumption data (Sethjiwala, 2023). This section reviews some studies on unsupervised clustering algorithms, regression analysis, and entropy-based methods

used to detect deviations from expected consumption patterns and points out their strengths and limitations within the context of evolving data landscapes.

Assouline et al. (2020) proposed a new framework for detecting anomalous patterns in electricity consumption for buildings using modern unsupervised machine learning clustering and regression methods. In this method, first, the weeks with anomalous consumption profiles are separated from the normal ones through clustering; later, a regression model is trained on the standard profiles to predict future consumption. Any deviations from the expected values are flagged as anomalies. The technique introduced here, while novel and very promising, does have certain limitations. This method depends on how precise the clustering algorithm used to find the correct contexts is. This is something which, in more complex scenarios, with multiple influencing factors, might be hard to achieve. Finally, a more reliable point-wise anomaly detection like neural networks or ensemble methods would be better than relying on advanced techniques of regression modelling (Inuwa & Das, 2024). Moreover, the real-world implications and potential applications of the method have not been discussed at length, hence limiting the understanding of its practical utility.

The entropy-based approach as proposed by Moure-Garrido et al. (2022) for detecting behavioural changes in household electricity consumption patterns was compared in terms of accuracy against two machine learning techniques—the random forest and neural network—and one statistical method: the ARIMA model. Different techniques were evaluated on a real dataset from households in the Region of Madrid, characterized by dissimilar consumption profiles. This entropy-based algorithm can detect more anomalous days than the other algorithms. The advantages are that it does not require a training period, and it can adapt dynamically to changes except in cases of long vacations. The drawback of the entropy approach is its difficulty in adapting immediately after significant changes in consumption for a long time. It would benefit the authors to explore more advanced machine learning models and ensemble techniques.

Fontugne et al. (2023) proposed a novel approach for detecting anomalous electricity consumption in large buildings using Ensemble Empirical Mode Decomposition (E-EMD). Their method decomposes non-stationary power-draw signals into different frequency modes, enabling the discovery of intrinsic inter-device correlations. By monitoring these correlations over time, they establish normal device usage patterns and detect anomalies when devices deviate from their typical behaviour. The approach successfully identified high and low power usage anomalies without prior knowledge

of expected device behaviour. While focused on building-level analysis, this method demonstrates the potential of advanced signal processing techniques in uncovering hidden patterns and relationships in electricity consumption data, which could potentially be adapted for postcode-level anomaly detection in future research.

2.3 | MACHINE LEARNING

As artificial intelligence leverages machines or computers to mimic the problem solving and the decision making capabilities of the human mind, machine learning is a subset of Artificial Intelligence that focuses more on the use of various self-learning algorithms that derive knowledge from data in order to predict outcomes.

Machine learning can either be supervised or unsupervised. In supervised learning, a labelled dataset is used to train the machine learning model which means that for each observation in the training dataset, the algorithm already knows what the correct output is. While in unsupervised learning, the algorithm is trained on unlabelled data and must find patterns or relationships in the data on its own without the guidance of a desired outcome variable.

2.3.1 | INTRODUCTION TO MACHINE LEARNING TECHNIQUES FOR PATTERN RECOGNITION

This section focuses on research that uses machine learning methods to identify unusual or anomalous patterns in electricity consumption data. The studies covered here apply both supervised and unsupervised machine learning models with hybrid approaches that combine techniques like dimensionality reduction and feature selection to achieve anomaly and pattern detection. A key advantage of these machine learning methods is their ability to effectively analyse complex, high-dimensional data sets and adapt to changing consumption patterns over time.

ELHadad et al. (2023) compared the relative performance of two anomaly detection algorithms, Enhanced Isolation Forest(E-IF) and Enhanced Local Outlier Factor(E-LOF), in labeling instances of abnormal power consumption from smart meter readings. They suggested that enhancing the Isolation Forest and Local Outlier Factor algorithms by introducing a threshold would better distinguish high from low electricity consumption anomalies. Results on ten smart meter readings with injected anomalies proved that E-IF did perform better than E-LOF: E-IF detected all the anomalies at contamination equal to 0.30 and 0.35, while E-LOF detected only an average of 68% and 78%, respectively, at respective contamination levels. While the study provided a

relative comparison between the two algorithms for power consumption data anomaly labelling, it leaves room for further discourse on an integrated approach that would enable those algorithms to be fitted in a general framework for prediction of anomalous electricity consumption patterns at a larger scale.

He et al. (2019) proposed a new method of LASSO–QRNN (Least Absolute Shrinkage and Selection Operator-Quantile Regression Neural Network) for probabilistic electricity consumption density forecasting. This suggested methodology links the Least Absolute Shrinkage and Selection Operator for variable selection and dimension reduction with the Quantile Regression Neural Network for the handling of nonlinear relationships. They showed that this method can handle high-dimensional data and exogenous variables that impact electricity consumption which produces a full probability distribution rather than a point prediction. However, their research was mainly oriented toward annual electricity consumption forecasting, which is not directly related to the detection of anomalous patterns of consumption at lower levels, such as the postcode level. Moreover, the proposed technique is dependent on past data and assumes that the exogenous variables are stationary, which may not be true for fast-changing urban environments or during exceptional events.

Soelami et al. (2021) underlined the necessity of accurate electricity consumption forecasting in energy management, budget planning, and integration of renewable energy. In their approach, they used methods for preprocessing data (the IQR clipping method) to handle anomalies, feature selection driven by mutual information, and model development using the support vector regression method driven by a radial basis function kernel.

While they provided a relatively comprehensive methodology and real-world implementation, some critical aspects could still be further be investigated. First, it would be interesting to find out how the model would perform when adding more features other than temporal data, including weather conditions, level of occupancy, and building characteristics. Since the research is focused on one building, further studies could be conducted to address how the proposed approach generalises to multiple buildings with different use patterns. Conclusively, a comparison with other machine learning techniques might project some insight into the strengths and limitations of the SVR approach they applied.

Kardi et al. (2021) proposed a novel deep learning approach for detecting anomalies in electricity consumption data one hour ahead of time. Their two-stage method

combines an LSTM network for prediction with an LSTM autoencoder for feature learning of normal consumption patterns. The study considers both global and local anomaly detection techniques and incorporates various external features, including weather variables and temporal factors. The authors found that temporal and lag features improved the model's efficiency in identifying anomalies due to seasonality in electricity consumption data. This work demonstrates the potential of machine learning techniques in predicting and detecting unusual consumption patterns, which could be adapted for postcode-level analysis.

Recent research on anomaly detection in smart grids has focused on various machine learning techniques to identify unusual electricity consumption patterns. Banik et al. (2023) provide a comprehensive review of anomaly detection methods, highlighting the importance of smart meter data in identifying abnormal usage. Machine learning approaches, including deep learning, transfer learning, and unsupervised methods, have shown promise in detecting anomalies at different levels of the grid. However, there is a need for more research on postcode-level anomaly detection. Challenges include the lack of labelled datasets, dealing with imbalanced data, and defining clear boundaries between normal and abnormal consumption. However, there should be a focus on developing robust models that can process large-scale smart meter data and provide actionable insights.

To further shed light on the ability of machine learning techniques on identifying anomalous power consumption patterns. Nayak and Jaidhar (2023) proposed a 1D Convolutional Neural Network (CNN) model for micro-moment classification to detect anomalous power consumption in buildings. Their approach achieved 96.4% accuracy and a 0.962 F1-score, outperforming previous methods. The model can identify usage types, appliance status, excessive usage, and consumption during occupant absence. This work demonstrates the potential of deep learning techniques in energy anomaly detection, which could be adapted to postcode-level analysis. However, the authors note that obtaining labeled datasets for supervised learning remains a challenge in this domain.

2.4 | RELATED WORKS

Dang et al. (2023) applied the Local Outlier Factor (LOF) algorithm to identify abnormal energy consumption patterns in a Vietnamese beverage processing factory. The unsupervised LOF method was chosen for its ability to detect multivariate and subsequence anomalies without labelled data. By analysing electricity, biomass, and

oil consumption against production levels, the researchers identified several anomalous patterns. The study highlighted periods of potential operational issues and inefficiencies. While effective, the authors noted that some points might be misclassified as anomalies and recommended cross-checking results with operational history. This approach demonstrates the potential of unsupervised machine learning techniques for detecting unusual consumption patterns at industrial sites.

ELhadad et al. (2022) proposed a method for predicting anomalous patterns of electricity consumption through a combination of machine-learning techniques. They applied the Isolation Forest algorithm to label smart meter electricity consumption readings as normal or abnormal, resulting in a data sequence with different lengths. Then, they applied supervised machine learning algorithms: the Random Forest and Decision Tree to classify the points which are normal or abnormal consumptions based on that data sequence with Random Forest giving the best accuracy of 90%. The contribution of this approach is very remarkable, as it considers the dynamic behaviour of power consumption data by turning the data into a sequence of labelled information and then feeding it into a prediction process. However, normal points were better predicted than abnormal data points of which heavy class imbalance was pointed as a possible cause. Also, there is a lack of comparison with other modern anomaly detection approaches and detailed performance evaluation for this proposed method on various datasets or the potential for scaling. Additionally, the data imbalance problem that severely affects most machine learning algorithms is not well handled.

Kesornsit and Sirisathitkul (2022) developed a hybrid machine-learning model for predicting electricity consumption in Thailand. Dimensionality reduction techniques like Principal Component Analysis were combined with feature selection algorithms like Stepwise Regression and Random Forest, followed by a Backpropagation Neural Network (BPNN) model for prediction using a comprehensive data set containing geographical, climatic, industrial, and household factors from several public sources in Thailand. They demonstrated that feature selection can dramatically improve the accuracy of predictive models. As Random Forest is integrated with BPNN, they achieve the highest accuracy in prediction when matched against any other models, thus proving that hybrid approaches are very effective. However, they could have included a discussion related to the limitations of the proposed model, such as generalizability in another region or the class imbalance and missing values implications on the data. Moreover, it could have been interesting to work on

alternative feature selection techniques or ensemble methods to really test the robustness of the approach.

The model developed gives an idea of how much electricity authorities can get help from it regarding demand planning and management strategies, especially regarding the potential of using machine learning techniques in energy-related applications.

2.5 | GAP ANALYSIS

While the studies reviewed added significant knowledge to the body, there exist some significant gaps which would be explored during this research. The gaps are as outlined below;

1. **Limited Focus on analysis on a more granular level:** most of the studies were performed on broader geographic areas or just an individual building which might not exactly capture the local trends of certain areas or risk exposure of private details for individual buildings. There is room for more research to understand how detection and prediction of anomalies will work on a geographical unit that is more fine-grained. A postcode-level analysis represents a perfect choice for this as it is more fined grained than regions or cities and broader than individual buildings. This can give a much more balanced view while capturing the local trends and maintaining some level of anonymity.

2. **Lack of Integration Between anomaly detection and prediction:** All the studies chose either current anomaly detection or future energy consumption prediction. There is room for research on models that can analyse consumption data, detect anomalies, and forecast potential irregular usage of electricity. Such models would easily be integrated into real time systems that can assist utility companies in making decisions for efficient energy management.

3. **Ensemble Method of Anomaly detection:** Most of the studies either used one algorithm for anomaly detection or simply compared the results of two algorithms. This leaves room for the possibility of achieving a better and more robust result by the combination of two or more algorithms. Ensemble

methods can take advantage of the strenghts of different algorithms and often outperform individual algorithms by capturing distinct aspects of the data (Gautam Kunapuli, 2023).

CHAPTER 3: THEORITICAL FRAMEWORK AND METHODOLOGY

3.1 | CONCEPTUAL FRAMEWORK

Because anomalies which appears as outliers are investigated in this study in relation to electricity consumption, the conceptual framework of this research is a combination of electricity consumption analysis, outlier detection, and application of modern machine learning techniques for predictions of these anomalies on a much more granular postcode level. Anomalous consumption can be seen as electricity usage patterns that deviate far from expected norms. These anomalies could be guided by several factors like the number of meters and demographic factors which will be considered in this study while control variables like weather patterns and Local events or activities will not be taken into consideration because of the contents of the dataset.

This framework proposes that these influencing factors interact in complex ways, creating unique consumption signatures for each postcode. Machine learning algorithms can be leveraged to identify and predict these patterns by analysing historical data and recognizing emerging trends.

The study will take a multi-layered approach:

- **Data collection and preprocessing:** Using postcode-level electricity usage data and census data with specific contextual data, such as demographic statistics and meter allocations. Additionally, feature engineering, normalisation, and data cleaning will be part of this layer.
- **Pattern/Anomaly Recognition:** Using recognised features and historical data, typical consumption patterns and outliers are identified within each postcode by using unsupervised learning approaches like Isolation Forest and Local Outlier Factor.
- **Predictive modelling:** this is the process of using advanced machine learning algorithms, such as KNN and decision trees, to predict future anomalies in electricity usage.

The motive behind this architecture is that all these layers, taken together, produce a robust system for making accurate predictions of abnormality in electricity consumption at the postcode level. Therefore, it is possible to employ machine learning techniques for the detection of complex, nonlinear interactions between the parameters and the patterns of consumption while the same might be difficult to achieve using traditional statistical methods.

The practical implications extend to many stakeholders, including energy providers, policymakers, and consumers. In doing so, they can leverage the ability of the models to predict abnormal consumption patterns with the potential for appropriate actions to be taken in energy distribution or consumption.

3.2 | DATA COLLECTION AND PREPROCESSING

Data collection and preprocessing are essential steps in any data analysis or machine learning project. This section details the process of collecting and preprocessing two datasets: "UK Census Postcode Estimates Table" and "Postcode Level Economy 7 Electricity 2022". These datasets will be combined using the postcode column to form a single dataset for exploratory data analysis (EDA) and predictive modelling.

The datasets have a common identifier which makes the combination possible. Combining these datasets provides a more comprehensive view of each postcode area, allowing the consideration of both electricity consumption patterns and demographic factors in the analysis. This holistic approach enriches the features and is more likely to make models that will yield more accurate and insightful predictions of anomalous electricity consumption patterns.

3.2.1 | DATA COLLECTION

The two datasets used in this project are:

- Postcode Level Economy 7 Electricity 2022: This dataset provides information on electricity consumption at the postcode level, including the number of meters, total consumption in kWh, mean consumption in kWh, and median consumption in kWh. It has 198,296 observations
- Census Postcode Estimates Table: This dataset contains demographic information at the postcode level, including the total population, the number of males and females, and the number of occupied households. This dataset has 1,308,779 observations

Both datasets are publicly accessible from the links in Appendix 1. The data is up to date as of 2022 and assumed to be accurate.

3.2.2 | DATA PREPROCESSING

Data preprocessing involves several steps to ensure the datasets are clean, consistent, and ready for analysis. The steps include:

- **Loading the Data:** The datasets are loaded into pandas dataFrames for easy manipulation and analysis.
- **Inspecting the Data:** The first few rows of each dataset are inspected to understand the structure and identify any immediate issues. This includes checking for missing values, data types, and any inconsistencies.
- **Checking and Handling Missing Values:** Missing values can significantly impact the analysis and modeling process. However, there are no missing values in the two datasets used for this research
- **Removing Duplicates:** Identifying and removing any duplicate rows to avoid redundancy.
- **Combining the Datasets:** The two datasets are combined using the postcode column. This involves:
    - ***Merging:*** Used pandas' merge function to combine the datasets on the postcode column. This creates a single dataset that includes both demographic and electricity consumption information for each postcode.
    - ***Handling Non-Matching Postcodes:*** This was to ensure that postcodes present in one dataset but not the other were handled appropriately by excluding them from the analysis.

  The resulting combined dataset has 85,846 matching observations
- **Final Inspection:** The final combined dataset was inspected to ensure all preprocessing steps have been applied correctly. This included checking for any remaining missing values, verifying data types, and ensuring the dataset is ready for EDA and predictive modelling.

3.2.3 | EXPLORATORY DATA ANALYSIS (VISUALIZATION)

This section explores the understanding of the data in this study where insights drawn here will assist in the outlier detection and predictive modelling

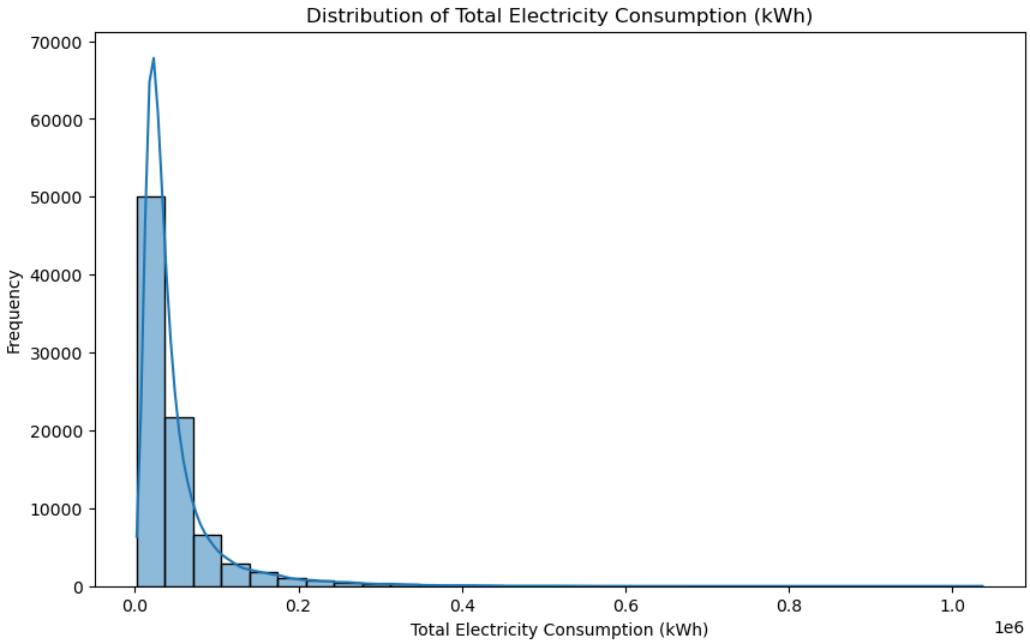


Figure 1: Distribution of Total Electricity consumption (KWh)

Figure 1 shows the distribution of electricity consumption across different postcodes, it appears to be right skewed which indicates that most postcodes have lower electricity consumption except for a few whose consumption is very high. To see how the number of occupied household affects the total electricity consumption, the scatterplot in figure 2 showed a positive correlation such that postcodes with more occupied households tend to consume more electricity.

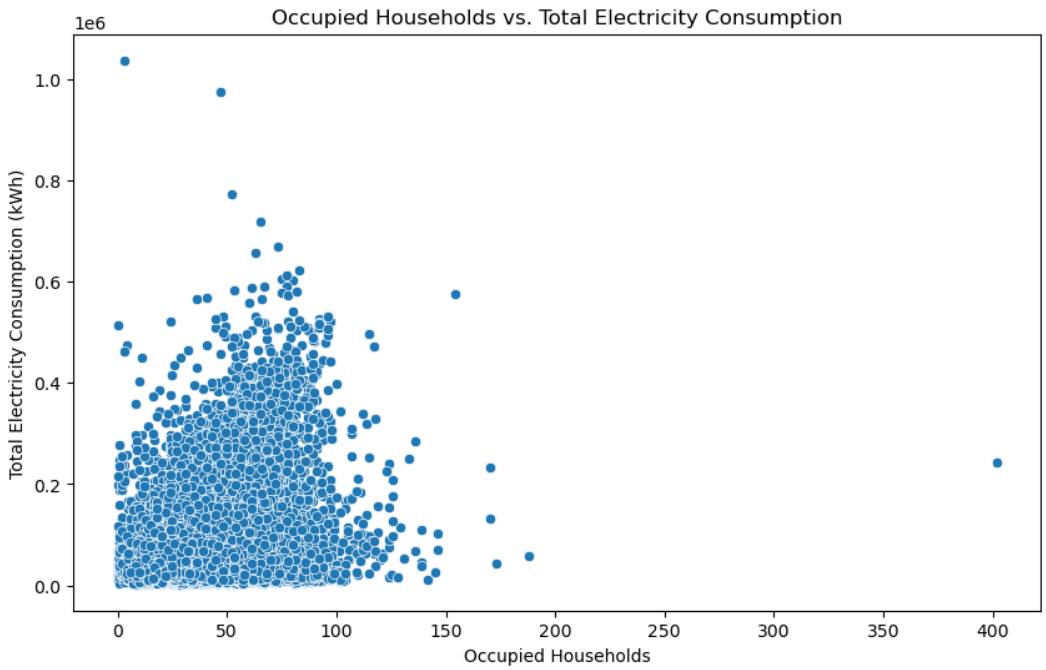


Figure 2: Scatterplot of Total Electricity Consumption against Occupied Households

Similarly, figure 3 also showed the same positive correlation between population and electricity consumption.

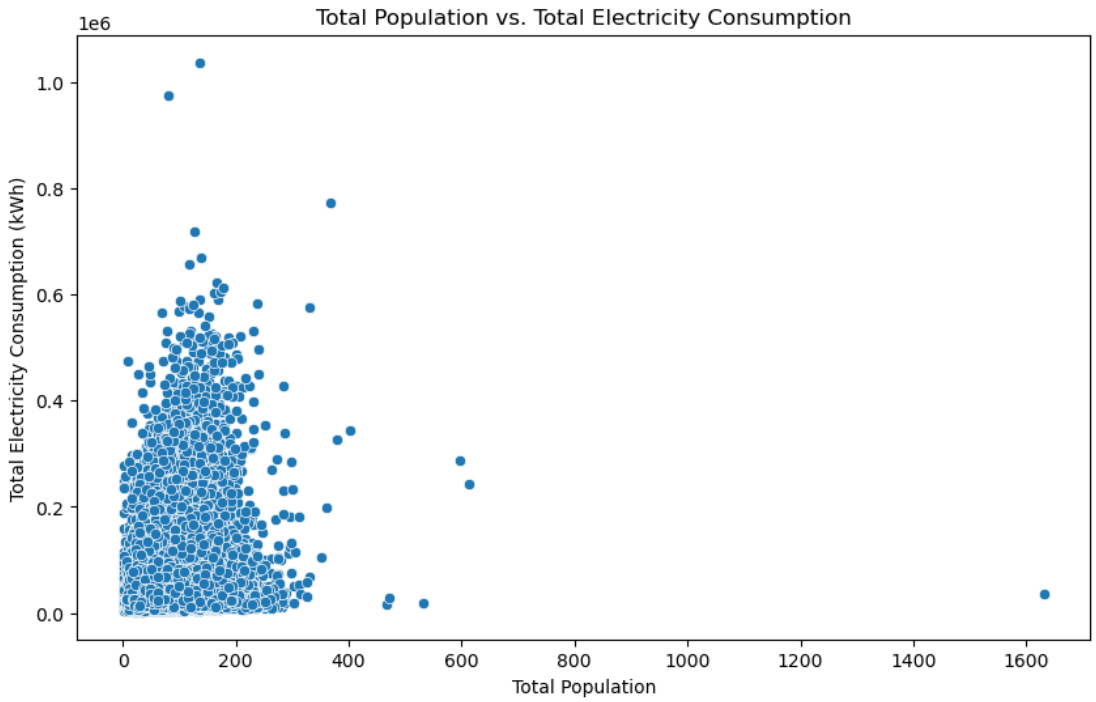


Figure 3: Scatterplot of Total Electricity Consumption against Total population

A trend analysis carried out to see the progression of electricity consumption relative to house occupancy. In figure 4, there is a general upward trend, though not

particularly linear, this indicates that more electricity is consumed as more households are getting occupied

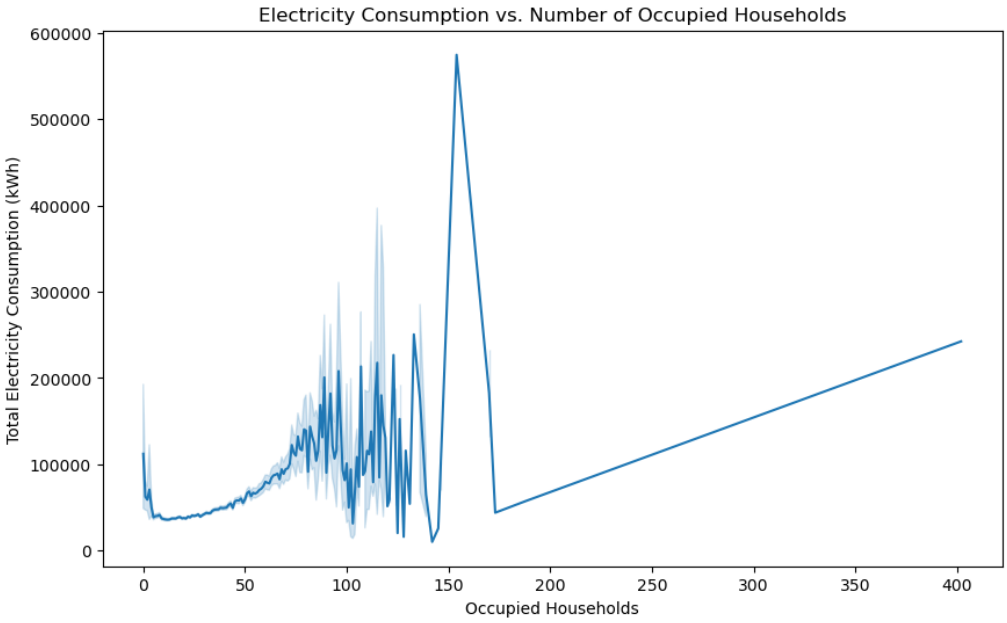


Figure 4: Scatterplot of Total Electricity Consumption against Total population

A cluster analysis using K-means clustering was also done in order to identify groups of post codes with similar patterns in demographics and electricity consumption. The data points in figure 5 are concentrated in the lower left quadrant of the plot which shows that majority of the postcodes have low population and consequently consumes low electricity.

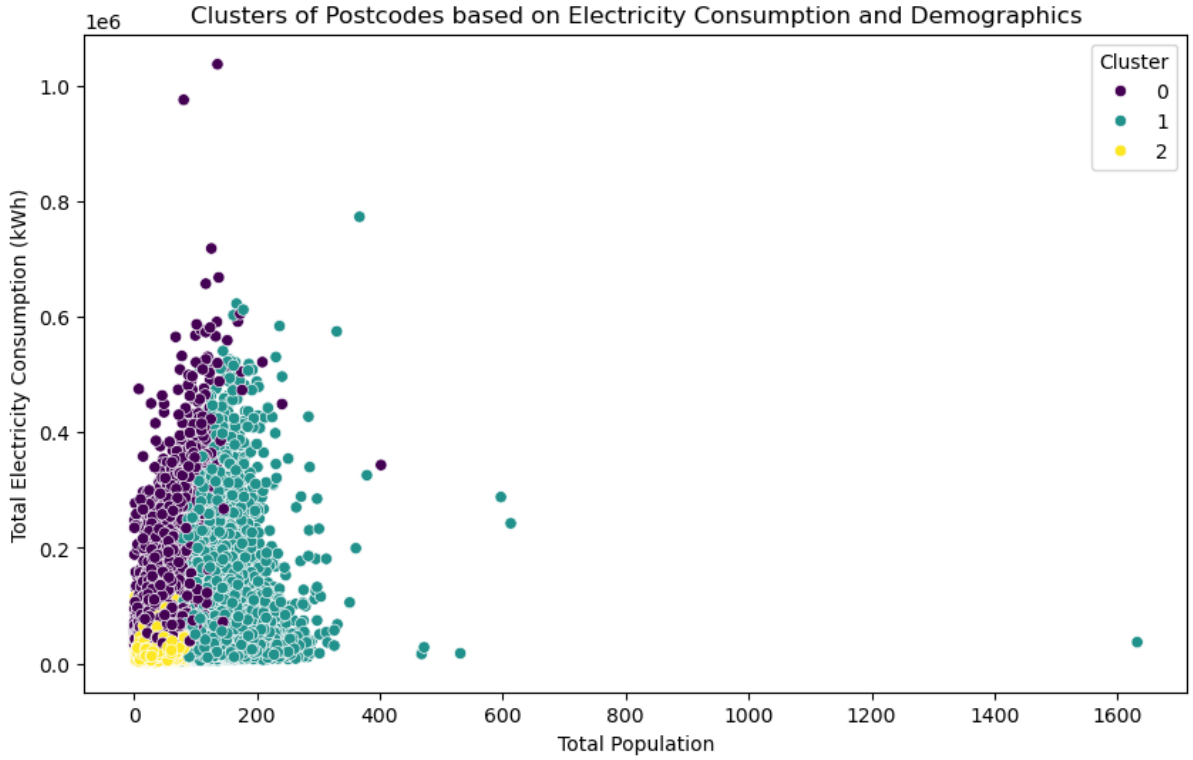


Figure 5: Clusters of Postcodes based on Electricity Consumption and Demographics

The purple clusters appear to represent postcodes with moderate to high electricity consumption relative to their population. It is spread over a wide range of population sizes which tends to have higher electricity usage.

The blue cluster represents postcodes with a wide range of populations and electricity consumption. There are outliers belonging to this cluster which indicates a very high electricity consumption unusual to such population while the yellow cluster represents postcodes with both low population and low consumption

The overlap between the purple and blue clusters is an indication of other factors influencing the clustering beyond just population and electricity consumption.

3.3 | FEATURE ENGINEERING

Feature engineering is a vital step in the data preprocessing pipeline that involves creating new features or modifying existing ones to improve the performance of machine learning models (Xie et al., 2023). It requires some domain knowledge, creativity, and a deep understanding of the data. It transforms raw data into a format

that better represents the underlying problem to the predictive models. It also helps to extract more meaningful information from the data.

3.3.1 | Importance of Feature Engineering

Feature engineering can significantly impact the performance of machine learning models. Well-engineered features can:

- Enhance model accuracy by providing more relevant information.
- Reduce model complexity by eliminating redundant or irrelevant features.
- Improve model interpretability by creating features that are more meaningful and easier to understand.

3.3.2 | Common Techniques in Feature Engineering

- **Creating New Features:** This involves generating new features from the existing data. This helps in giving more insight and understanding the data better.
- **Scaling and Normalisation:** Features with different scales can negatively impact the performance of some machine learning algorithms. Scaling and normalization ensure that all features contribute equally to the model (Bhandari, 2024).
- **Handling Categorical Data:** Converting categorical data into numerical format is essential for most machine learning algorithms. Techniques like one-hot encoding, label encoding, and target encoding are commonly used.
- **Dealing with Missing Values:** Missing values can be handled by imputation (filling in missing values with mean, median, or mode) or by creating a new feature indicating the presence of missing values.

3.3.3 | Feature Engineering in Our Dataset

For our combined dataset containing demographic and electricity consumption data, several feature engineering techniques was applied to better understand the data:

- **Electricity Consumption per Household:** This calculates the average electricity consumption per household by dividing the total consumption by the number of meters. This feature helped provide insights into the energy usage patterns of different areas.

$$\textit{Electricity Consumption per Household} = \frac{Total\ Consumption\ (kWh)}{Number\ of\ Meters}$$

- **Population Density:** This calculated the population density by dividing the total population by the number of occupied households. This feature helped in the understanding of the concentration of people in different postcodes.

$$\textit{Population Density} = \frac{Total\ Population}{Occupied\ Household}$$

- **Interaction Features:** An interaction feature was created by the division of electricity consumption per household by the population density which provided insights into how efficient or abnormal electricity is used in densely populated areas.

$$\textit{Energy Efficiency} = \frac{Electricity\ Consumption\ per\ Household}{Population\ Density}$$

- **Outlier Detection:** Outliers were identified in the electricity consumption data which was used to label the dataset as either normal or abnormal electricity consumptions. Outliers were detected using the Isolation forest and local outlier factor algorithms.
- **Label Encoding:** The categorical variables (*Outcode and Postcode)* were converted to numerical format using the label encoding method.
- **Scaling:** The continuous variables in the datasets are in varying scales. For example, the *Total_cons_Kwh* variable are in a much higher scale than *Population* and *Num_meters* variables. To ensure this disparity in scales do not affect the accuracy of the model, Normalisation was performed to put all variables on the same scale

It is important to add that the essence of feature engineering before any analysis is conducted specifically in this study is that feature engineering will help capture complex relationships within the data and also make patterns more visible to machine learning models, potentially improving their predictive performance. For example, this was evident in the research carried out by Ward et al. (2023) on energy consumption prediction on UK buildings where semantic segmentation, 3D reconstruction and

geometry extraction where used to create key building characteristics from images obtained from drive-by which subsequently helped the results of their model.

3.4 | MACHINE LEARNING MODELS

Electricity consumption data can be highly unpredictable and irregular, making it difficult to work with. For anomalies prediction, working with this type of data requires powerful machine learning models that can accurately make the required predictions from the electricity consumption patterns. In this study, three machine learning models will be applied for the predictive model- Random Forest, Decision Trees, K- Nearest Neighbors (KNN) while an Ensemble Method which will combine the Isolation forest and Local Outlier factor algorithms will be adopted for the outlier detection.

3.4.1 | MODEL ARCHITECTURE:

1. **RANDOM FOREST:** This ensemble learning algorithm uses a combination of multiple decision trees to create a more robust and accurate model (Breiman & Cutler, 2024). It uses the bagging (bootstrap aggregation) technique, where random subsets of the original dataset are created with replacement to train individual trees.

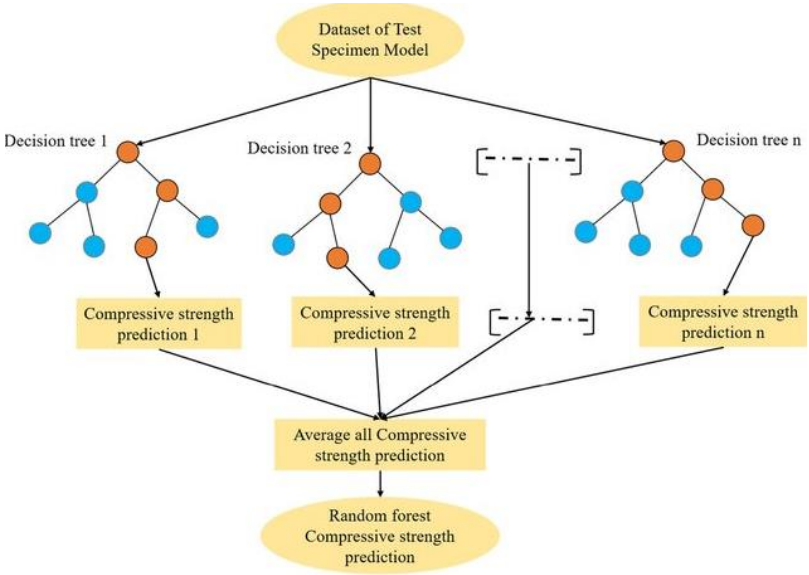


Figure 6: pictorial Representation of random forest classifier (Researchgate, n.d.)

Each tree is constructed independently, considering only a random subset of features at each node for splitting. This feature randomness adds diversity to the forest and reduces correlation between trees. For classification tasks, the final prediction is determined by majority voting from all trees, while for regression, it's the average of all tree predictions. The algorithm also utilizes out-of-bag samples for internal validation and feature importance estimation. Random Forest's architecture allows for parallelisation, efficient handling of high-dimensional data, and reduced overfitting, making it a powerful and versatile machine learning technique.

2. **DECISION TREES:** These are tree-like hierarchical models for classification and regression. They consist of a root node, internal decision nodes, branches, and leaf nodes. The tree is created by recursive partitioning data using the features in the data (Omid Chatrabgoun, 2024).

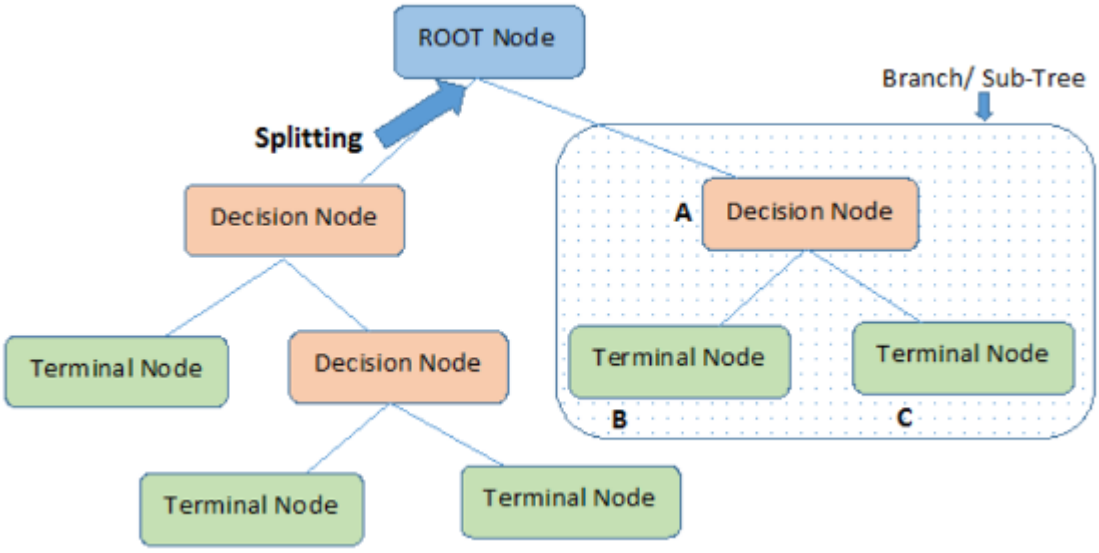


Figure 7: pictorial Representation of Decision Tree classifier (Saini, A., 2024)

On every node, the algorithm will choose the best feature to split data that will maximise information gain or minimises impurity (measured through entropy or the Gini index). This process is done repeatedly until the stopping criterion, which may be maximum depth or minimum number of samples per leaf. Prediction is achieved by traversing from the root to a leaf node according to the input features. To avoid overfitting, pruning and tuning of hyper-parameters

such as tree-depth or the minimum samples per split is necessary. Decision trees are interpretable, handle categorical and numerical data, and are basic versions of more advanced ensemble methods like Random Forests.

3. **KNN:** The K-Nearest Neighbors (KNN) algorithm is a simple yet effective machine learning technique used for classification and regression tasks. It operates by storing the entire training dataset and making predictions based on the similarity between new data points and the stored examples (Srivastava, T., 2024). When classifying a new instance, KNN calculates the distance (typically Euclidean) between the input and all training examples, identifies the K nearest neighbors, and assigns the most common class among these neighbors as the predicted label. The choice of K is crucial, as it affects the model's performance and decision boundaries. KNN is easy to interpret and implement but can be computationally expensive for large datasets.

4. **THE ISOLATION FOREST ALGORITHM:** This is an unsupervised anomaly detection algorithm which works on the principle that it is usually easier to isolate anomalies from normal data observations. The architecture is a combination of isolation trees, such that each tree is built to randomly partition the data space. The algorithm builds the trees in a recursive way by splitting the data randomly on features and split points until a prescribed tree height is attained or when the nodes hold single points. Then anomaly scores are calculated based on the average path length that would isolate a data point over the forest. Where the paths with smaller distances indicates potential anomalies (Mougan, 2022). It scales well with the number of samples and features, making it ideal for large datasets.

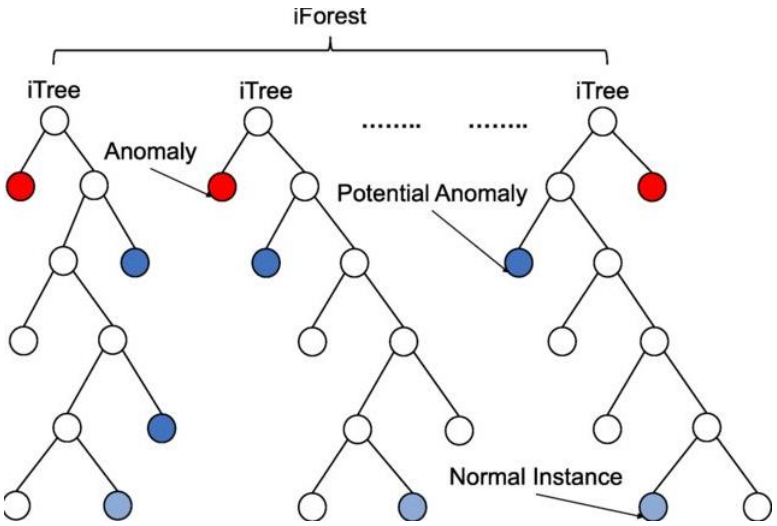


Figure 8: pictorial Representation of Isolation Forest (Researchgate, n.d.)

5. **THE LOCAL OUTLIER FACTOR (LOF) ALGORITHM:** This is an unsupervised method for detecting anomalies using the concept of local density. This is because its algorithm architecture does its comparisons of the local density of a point to the local densities of its neighbors. The k-nearest neighbours for each data point are first identified, then the reachability distance, the maximum between the actual distance and the core distance, is evaluated. In LOF, once the distances are computed, a local reachability density is computed for each point, then the LOF is determined relative to the ratio of the point's computed density to the average densities of its neighbors. Increased values of LOF show potential anomalous points.

3.4.2 | JUSTIFICATION FOR MODEL SELECTION

3.4.2.1 | CLASSIFICATION MODELS

Recent studies, such as those conducted by ELhadad et al. (2022) and Kesornsit and Sirisathitkul (2022), have demonstrated Random Forest's superior performance in anomaly prediction. Given the characteristics of our dataset which is large, has nonlinear relationships, and overlapping variables (as evident in the pair plot in the Appendix 2) - Random Forest emerges as a particularly suitable algorithm for this study.

Random Forest's advantages in this context include:

- **Handling high-dimensional data:** Our large dataset benefits from Random Forest's ability to manage numerous features effectively (Cai et al., 2023).

- **Capturing non-linear relationships:** The nonlinear patterns in our data are well-suited to Random Forest's tree-based structure, which can model complex interactions without explicit specification (Cai et al., 2023).
- **Robustness to outliers and noise:** This is crucial given the presence of outliers in our dataset as the outliers will not be removed because they are necessary in this study (ELhadad et al., 2022).
- **Feature importance ranking:** This will help in the understanding of which variables are most critical in the anomaly prediction (Cao et al., 2023).

While Random Forest will serve as the primary algorithm in this study, we will compare its performance against Decision Trees, K-Nearest Neighbors (KNN), and Gradient Boosting. This comparison is crucial as it is only through experimentation that the best model for our specific dataset will be determined.

Decision Trees: While it is simpler and more interpretable than Random Forest, they're prone to overfitting (Costa & Pedreira, 2022). However as its structure is not as complex as Random Forest, it can provide insights into the decision-making process that might be valuable for understanding anomaly patterns.

KNN: This algorithm can capture local patterns in the data that might be missed by tree-based methods. It's particularly useful if anomalies cluster in feature space. However, it may struggle with high-dimensional data and is sensitive to the choice of *"K"*.

Gradient Boosting: Like Random Forest, it's an ensemble method that can handle nonlinear data. It often provides high accuracy and can outperform Random Forest in some scenarios. However, because it is highly prone to overfitting (Sibindi et al., 2022) which could affect the accuracy of the model's result, it can only be used for comparison in this study.

By comparing these different models, the best-performing model will not only be identified but also a more comprehensive understanding of the anomaly patterns in the data will be gained. This multi-model approach allows the leveraging of the strengths of each algorithm, potentially leading to more reliable and accurate anomaly predictions.

3.4.2.2 | ENSEMBLE METHOD FOR OUTLIER DETECTION:

The strength of the Local outlier factor lies in detecting local outliers within the data while isolation Forest is more effective outlier detection in high dimensional data (Liu et al., 2024). Hence, it is important to explore the possibility of leveraging the efficiency and scalability of Isolation Forest with the local sensitivity of Local Outlier factor with the goal of reducing both false positives and false negatives hence obtaining better results. This combination seeks to leverage the strengths of both algorithms.

3.5 | EVALUATION METRICS

For notable conclusions to be drawn in this study, the performance of the models will be evaluated and compared using the following metrics:

- **Confusion Matrix:** A table that summarises the performance of a classification algorithm. It shows the true positives, true negatives, false positives, and false negatives. This helps to give insight on the errors the model is making and it is essential in the calculation of metrics like precision, recall and accuracy.

Table 1: tabular representation of a confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	**True Positive (TP)**	**False Negative (FN)**
Actual Negative	**False Positive (FP)**	**True Negative (TN)**

- **Accuracy:** This gives a general sense of model performance by showing the percentage of observations that were correctly classified. However, it can be misleading if the classes are imbalanced. According to Naidu et al. (2023), it is given by the formular:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- **Precision:** This measures the accuracy of the positive predictions. For this research, it will give a measure of how many of the predicted anomalies were actually anomalies. This is necessary in this research because a wrong anomaly prediction can be potentially costly when considered in policy or decision making. It is given as(Naidu et al., 2023):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (Sensitivity or True Positive Rate):** The proportion of true positive predictions out of all actual positives. It will measure the ability of the models to actually identify all the anomalies in this study. The formular is given below (Naidu et al., 2023):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1 Score:** This is the harmonic mean of precision and recall. It provides a balance between precision and recall (Naidu et al., 2023).

$$\text{F1 Score} = 2 * \frac{\text{Precision x Recall}}{\text{Precision+ Recall}}$$

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** This measures the ability of the model to distinguish between classes. It plots the true positive rate against the false positive rate at various threshold settings (Naidu et al., 2023).

$$\text{AUC} = \int_0^1 TPR(FPR)d(FPR)$$

3.6 | EXPERIMENTAL DESIGN

The experimental design in this study involves the following steps:

1. **Data exploration and preprocessing:**
   This will include preparing the data for analysis and then using visualisations to better understand the data.

2. **Data Splitting:**
   - Training Set: 70% of the data will be used to train the model.
   - Test Set: 15% of the data will be used to evaluate the model's performance.
   - Validation Set: 15% of the data will be used to tune hyperparameters.
     This split ensures enough data for training while reserving sufficient data for testing and validation. The validation set allows for hyperparameter tuning without compromising the integrity of the test set, which will be used for final performance evaluation (Pramoditha, 2022).
3. **Model Training and Evaluation:**
   - Training: The models will be trained on the training set using the optimal hyperparameters.
   - Testing: The models will be evaluated on the test set using the selected evaluation metrics.
   - Validation: The models' performance will be validated using the validation set, and hyperparameters will be adjusted accordingly

4. **Results and Discussion**
   The results will include:
   - **Performance Metrics:** Reporting accuracy, precision, recall, F1 score, ROC-AUC and confusion matrix.
   - **Feature Importance:** Analysis of the most important features contributing to the prediction of anomalous consumption patterns.
   - **Error Analysis:** Examination of misclassified instances to understand the model's limitations and potential areas for improvement.

5. **Hyperparameter Tuning:**

   Grid search with cross-validation will be used to find the optimal hyperparameters (e.g., number of trees, maximum depth).

6. **Tools and Equipment:**

   - **Software:** Python 3, Jupyter Notebook

Hardware: HP Envy 350, 2GHz intel core i3, 16Gb RAM running on Windows 10 Pro

The Github links to Python codes used in this experimentation can be in Appendix 1

CHAPTER 4: RESULTS, ANALYSIS AND DISCUSSION

4.1 ANOMALY DETECTION RESULTS

This section compares the results of three outlier detection methods: Local Outlier Factor (LOF), Isolation Forest, and an Ensemble Method. Table 2 shows a comparison table of how much actual outliers can be caught by each method exclusively against a substantial amount of outliers injected into the data

Table 2: Algorithm outlier detection performance against injected outliers

Method	Total injected outliers	Amount detected	Percentage detection
Local Outlier Factor	4291	466	10.90
Isolation Forest	4291	4272	99.60
Ensemble Method	4291	4291	100.00

Similar to the results obtained by ELHadad et al. (2023), Local outlier Factor had a poor performance in detecting the injected outliers and Isolation detected virtually all the outliers. However the ensemble method was able to capture all the outliers without exception.

It is necessary to note that the LOF and Isolation Forest used in this study was not the enhanced variant where the contamination levels were increased to 0.3 as ELHadad et al. (2023) employed. However, the ensemble method proved to be better at outlier detection than their proposed method. This was evident as all three methods were also applied to the data in this study without any injected outlier. The results are outlined in Table 3 below:

Table 3: Outlier detection Results

Method	Number of Outliers	Percentage of Data
Local Outlier Factor	8584	10.00%
Isolation Forest	8584	10.00%
Ensemble Method	14850	17.30%

From Table 3, it is seen that:

- LOF and Isolation Forest identified the same number of outliers: 8584, which represents 10% of the data points.
- The Ensemble Method stayed consistent in detecting more outliers: 14850, or 17.36% of the data.

To further compare the outlier detection strength of the three methods, their agreement with each other is outline in Table 4 below

Table 4: Agreement table for outlier detection methods

Agreement Between	Number of Common Outliers	Percentage of Agreement
LOF and Isolation Forest	2318	27.00%
LOF and Ensemble Method	8584	100.00%
Isolation Forest and Ensemble Method	8584	100.00%

The agreement between the three methods reveals interesting patterns:

- LOF and Isolation Forest agree on 2318 outliers, which is 27% of their total detections.
- Both LOF and Isolation Forest have 100% agreement with the Ensemble Method, meaning all outliers they detected were also flagged by the Ensemble approach.

These results suggest that the Ensemble Method is more inclusive, capturing all outliers identified by the other two methods plus additional points it considers anomalous.

The scatter plots in figure 9 provides a visual representation of the outlier detection results

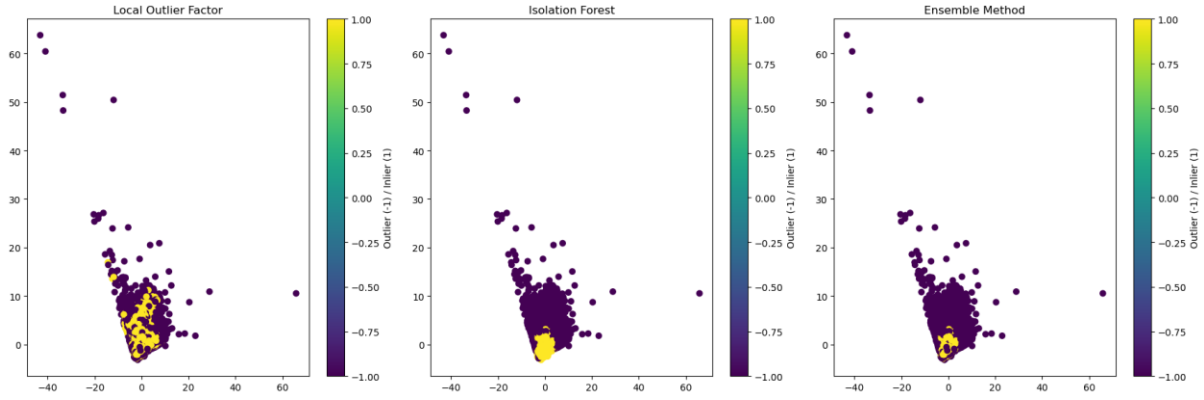


Figure 9: Scatter Plots of the outlier detection results

The following can be deduced from the plots in figure 9:

1. All three methods show a similar overall data distribution, with a dense cluster of points in the lower part of the plot and scattered points above and to the sides.

2. The colour scale indicates the outlier score, with yellow representing high outlier probability (close to 1) and purple representing low outlier probability (close to -1).

3. Local Outlier Factor:

- Shows a more nuanced distribution of outlier scores.
- The main cluster has varying shades, indicating different degrees of abnormality even within the dense region.
- Isolated points at the top and sides though not marked as outliers (yellow) show characteristics of outliers and will need further investigation.
- Some points within the main cluster are also identified as potential outliers.

4. Isolation Forest:

- Presents a more binary classification.
- The main cluster is mostly purple (non-outliers), with a sharp transition to yellow for points outside this cluster.
- Isolated points were also not distinctly marked as outliers.
- Less nuance within the main cluster compared to LOF.

5. Ensemble Method:

- Shows characteristics of both LOF and Isolation Forest.
- The main cluster has some variation in outlier scores, though less than LOF.
- Seems to flag more points within and at the edges of the main cluster as potential outliers, consistent with its higher overall outlier count.

The differences in these visualisations explain the agreement patterns seen in the table 4:

- LOF and Isolation Forest agree on obvious outliers (isolated points) but differ in their assessment of points within or at the edges of the main cluster. This explains their 27% agreement rate.
- The Ensemble Method incorporates all outliers from both LOF and Isolation Forest, leading to the 100% agreement rates. It then adds its own additional detections, resulting in the higher total outlier count.

All the results highlight the strengths and characteristics of each method and demonstrates the value of comparing multiple outlier detection techniques, as each provides a different perspective on the outliers in the data. The combination of Isolation Forest and Local Outlier factor methods in the Ensemble approach offers a more thorough detection capability than the traditional or single algorithm method used in previous study by other researchers, although there could be a potential for over-detection which would need to be balanced against the costs of missing true anomalies in the specific application context.

4.2 MODEL TRAINING AND EVALUATION

4.2.1 CLASS IMBALANCE AND SMOTE:

The initial analysis done on the dataset revealed a significant class imbalance in the dataset. This imbalance as noted by Banik et al. (2023) is a critical issue in anomaly detection tasks, because the anomalous patterns (which appear as outliers) are usually much less occurring than normal patterns. Current researches neither captured the presence of class imbalance nor handled them. Training models on this type of data could lead to unreliable and biased models that perform poorly in detecting anomalies. Hence as an oversampling method, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm was used correct this imbalance. The use of SMOTE to address class imbalance in this study is a significant methodological strength not prominently or explicitly captured in previous researches on anomaly detection as reviewed in this study. Figure 10 shows the distribution of classes before and after applying the Synthetic Minority Over-sampling Technique (SMOTE).

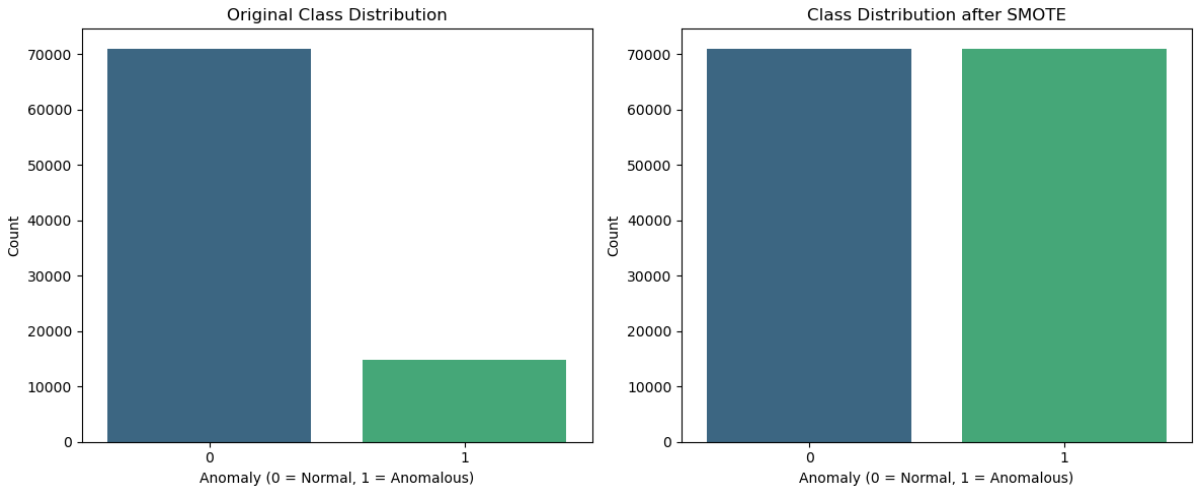


Figure 10: Class distribution before and after oversampling

Before SMOTE, as seen in figure 10 the majority class (normal consumption patterns) heavily outnumbers the minority class (anomalous patterns). Out of 85,834 recorded cases, only 14, 850 were abnormal cases with the normal cases dominating at 70,984. However, after applying the SMOTE algorithm, the class distribution became more balanced at both 70,984 for both classes, which should help in training a more robust model capable of identifying anomalous electricity consumption patterns effectively.

4.2.2 DATA SPLITTING:

The data was split into training and testing sets, with 70% used for training and 30% for testing. The test set was further split to 15% validation and 15% test sets. This split allows for model evaluation on unseen data, which is crucial for assessing the model's ability to generalize to new postcodes and detect anomalies in real-world scenarios. The distribution of the classes in each set is shown in Table 5 below:

Table 5: Class distribution in dataset split

	Total Data points	0 (normal cases)	1 (abnormal cases)
Training set	**99,373**	**49,677**	**49,696**
Test set	**21,299**	**10,505**	**10,794**
Validation set	**21,296**	**10,437**	**10,859**

4.2.3 COMPARATIVE ANALYSIS OF MODELS' RESULTS:

Table 6: Results of the model training and testing

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE	ROC-AUC
RANDOM FOREST	**0.959857**	**0.953678**	**0.966664**	**0.960127**	**0.993867**
KNN	**0.955960**	**0.942253**	**0.971453**	**0.956630**	**0.987899**
DECISION TREE	**0.916475**	**0.913173**	**0.920462**	**0.196803**	**0.916475**
GRADIENT BOOSTING	**0.864313**	**0.902229**	**0.817166**	**0.857593**	**0.939446**

1. RANDOM FOREST:

As seen in Table 6 above, the Random Forest model shows best results which is consistent with recent research. It had the highest accuracy of 0.959857 and a good balance between precision of 0.953678 and recall of 0.966664.

From the confusion matrix in Figure 11, it could be seen that out of 10,500 normal consumption cases, the model correctly identified 10150 instances while misclassifying 355. Similarly out of 10,794 abnormal consumption cases, the model correctly identified 10,294 instances while misclassifying 500 cases. The similar number of correct predictions of both classes shows that this model is well balanced and is not bias towards one class. However, there's a slight tendency towards false positives (500) over false negatives (355). This means that the model is slightly more likely to incorrectly classify a normal consumptions instance as abnormal consumption than vice versa.

Also on the trade of between precision and recall, the model has a slightly higher recall (96.67%) than precision (95.37%) for anomalous consumption cases. This indicates it's slightly more focused on capturing all abnormal cases, even if it means including some false positives.

This is a remarkable improvement to the poor prediction for abnormal data points by ELhadad et al. (2022) where precision was at 75%.

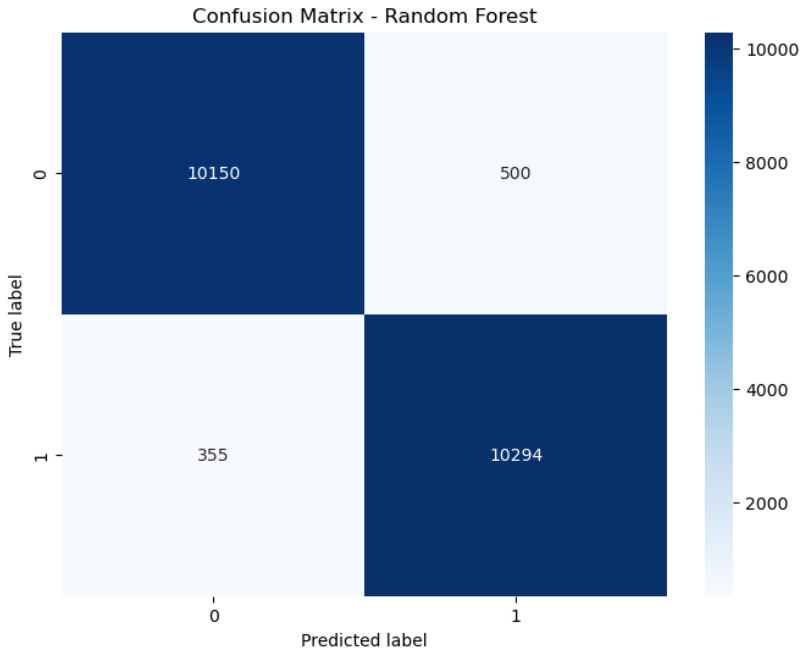


Figure 11: Confusion matrix Heat Map for the Random Forest Model

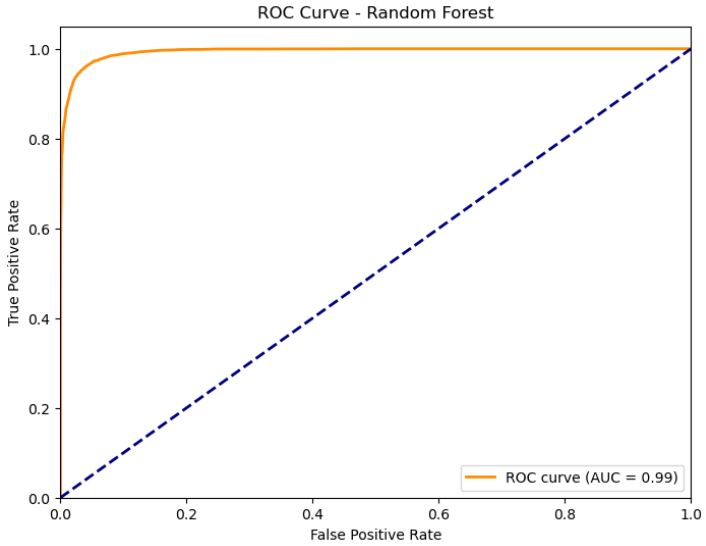
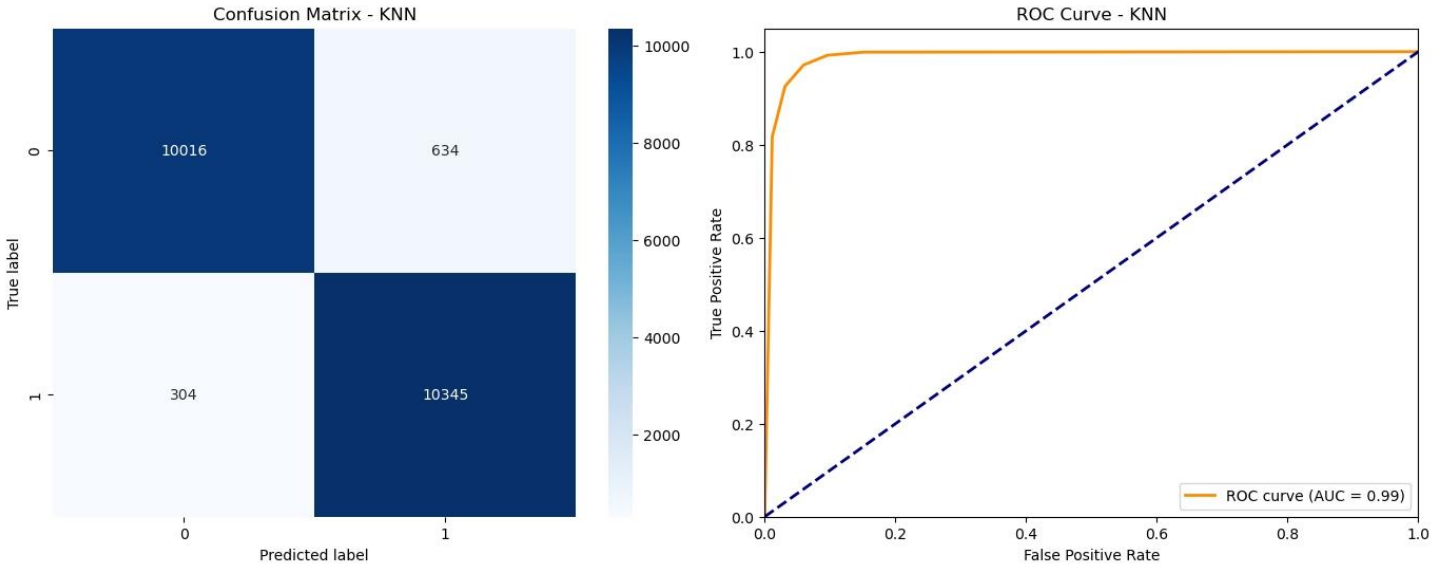


Figure 12: ROC Curve for the Random Forest Model

The ROC curve in figure 12 demonstrates excellent performance with an AUC of 0.9994, indicating the model's excellent predictive ability to distinguish between normal and anomalous electricity consumption patterns.

2. K-NEAREST NEIGHBORS (KNN):

KNN shows similar characteristic to random forest but with a slightly lower accuracy of 0.955960 and precision at 0.942253 of which both are still considered high. Interestingly the recall value is better at 0.971453 than the 0.966664 returned by



Random Forest which suggest that KNN is better at capturing positive instances (in this case anomalies) than Random Forest.

Figure 13: Confusion matrix and AUC-ROC plots for KNN Algorithm

The confusion matrix in figure 13 validates the slight superiority of KNN in capturing the anomalous consumption instances where only 304 cases were misclassified as normal unlike the 355 misclassified by Random forest. Also, the AUC-ROC value is equally 0.99 showing the same near perfect predictive ability as Random forest in distinguishing between normal and anomalous electricity consumption.

3. DECISION TREES

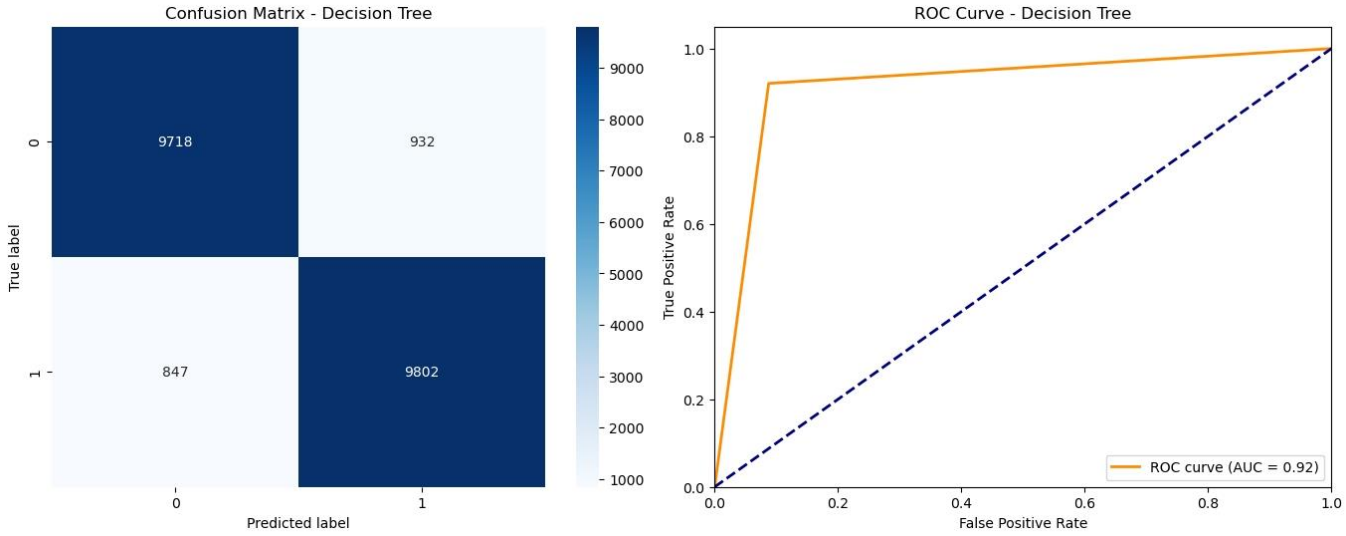


Figure 14: Confusion matrix and AUC-ROC plots for Decision Tree Algorithm

The decision tree model performed poorer than both Random forest and KNN models. However, the results are still considerably good as it was able to capture 9,718 normal cases out of 10,500 and 9,802 anomalous cases out of 10,794. But the higher number of misclassified cases of both classes (932 and 847 respectively) means that Decision trees is a less reliable model for prediction of anomalous electricity consumption. It can obviously be improved by tuning the parameter or added to other models in an ensemble training but because of the sensitivity in the classification for this study, it is preferable to go with the best performing model

4. GRADIENT BOOSTING

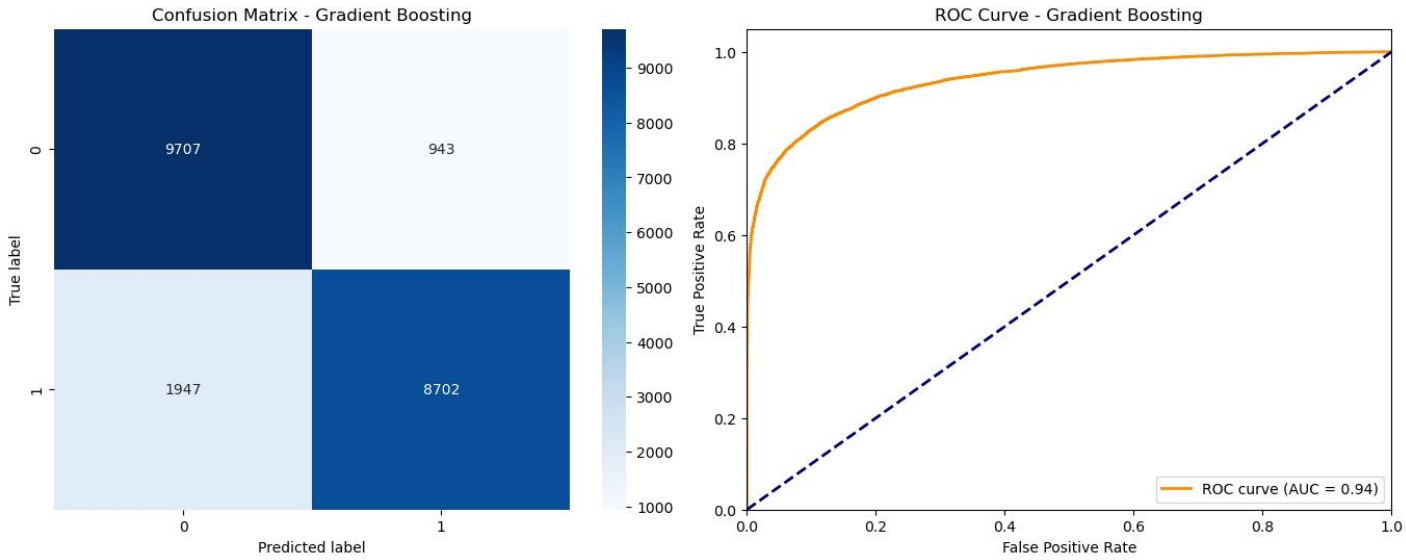


Figure 15: Confusion matrix and AUC-ROC plots for Gradient Boosting Algorithm

From the results in table 5 and the plots in figure 15 above, this model performed the poorest amongst all the models trained in this study. With a high number of misclassified data points when compared to other models, it will be a highly unreliable model to adopt in this study. For less sensitive predictive systems whose outcome consequences would not be critical, this model can be adopted as an accuracy, precision, recall and AUC-ROC of 0.864313, 0.90223, 0.817166, and 0.939446 respectively are all considerably good scores. However, since incorrect classification of electricity consumption anomalies is sensitive (for example, unfair policies made against an incorrectly classified postcode), this model cannot be relied on.

4.2.4 HYPERPARAMETER TUNING:

Grid Search hyperparameter tuning process was employed to fine-tune the Random Forest in order to obtain better efficiency. To ensure robustness in parameter selection by testing multiple configurations across different data subsets, the search used a 5-fold cross validation for each of 4 candidates totalling 20 fits. The results are given in table 6 below:

Table 6: HyperParamemeter Tuning Results

Best Parameters for Tuning				
Cross validation score	**Max Depth**	**min_samples_leaf**	**min_samples_split**	**n_estimators**
0.955322	None	1	2	200
BEST PARAMETERS OBTAINED				
Accuracy	**Recall**	**Precision**	**F1_Score**	**ROC-AUC**
0.960702	0.954259	0.96779	0.960977	0.99411

The best cross validation score achieved was approximately 0.9553 which indicates a strong performance across the validation folds such that the tuned parameters allow the model to generalise well across different subsets of the data. As a result, there is slight improvement in the performance metrics.

Other discoveries from this process is that for the prediction of anomalies, trees are allowed to grow to their full extent to achieve their best results without the risk of overfitting hence why a Max depth of None. Also 'min_samples_leaf' of 1 and 'min_samples_split' of 2 indicate that the model performs best with very granular decision-making, allowing it to capture subtle patterns in electricity consumption.

This fine-tuning process ensures that model is optimized for deployment in real world scenarios and should be capable to handling various postcode-level consumption patterns effectively, making it versatile for different geographic or demographic contexts.

4.2.5 FEATURE IMPORTANCE:

The feature importance analysis provides valuable insights into which factors most significantly contribute to detecting anomalous electricity consumption patterns. It shows the impact of the respective factors on the models predictive power.

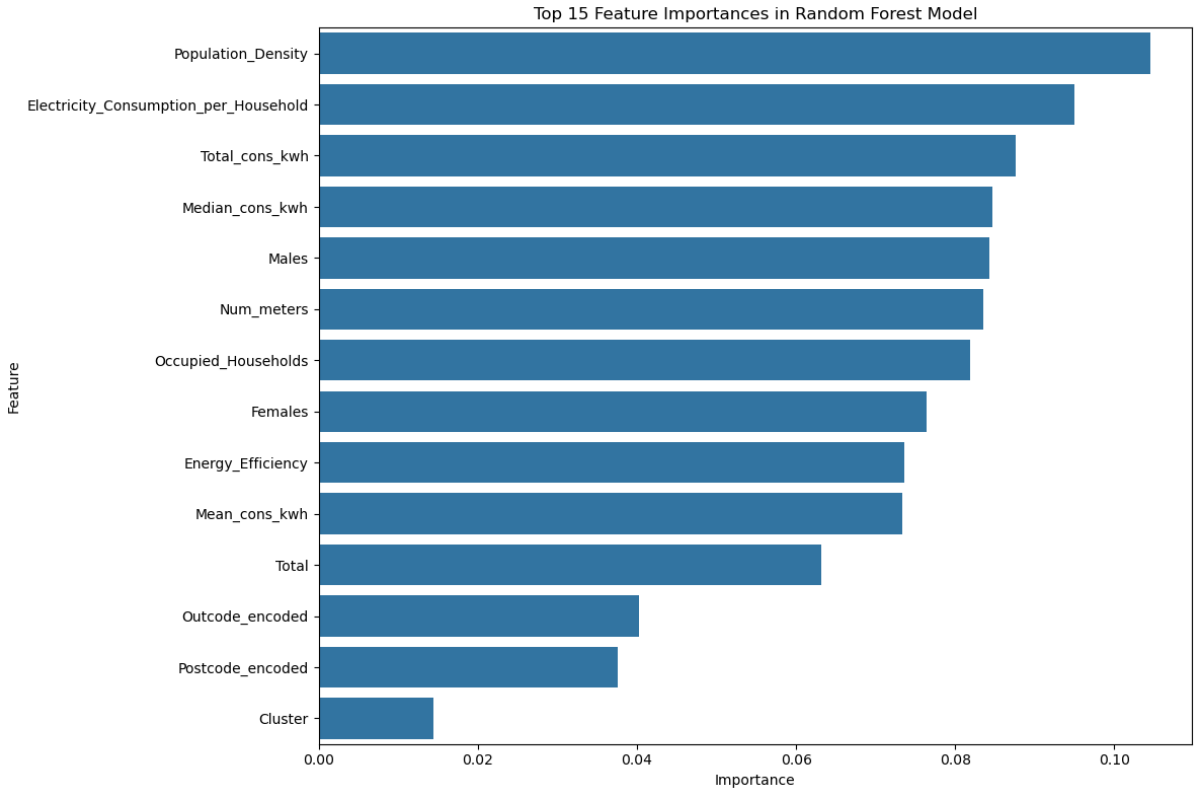


Figure 16: Top 15 Feature importance in Random Forest Model

Figure 16 shows how the variables are ranked by their impact to the trained random forest model in the prediction of anomalous electricity consumption.

Population Density emerges as the most influential feature, suggesting that the density of people in an area strongly correlates with unusual electricity usage patterns. This is closely followed by Electricity_Consumption_per_Household, this is logical because actual energy consumption by individual households should definitely play a significant role for the model to determine which consumption is abnormal. Demographic factors like the number of males and females, as well as the number of occupied households and electricity meters, all show substantial importance which aligns with the suggestion by Soelami et al. (2021) that features beyond temporal data, such as occupancy and building characteristics can influence anomalous consumption,. This implies that the model considers population characteristics and housing occupancy when detecting unusual consumption. Energy_Efficiency appears midway through the list, suggesting it plays a moderate role in predicting anomalies, possibly by setting baseline expectations for normal consumption.

Geographical factors like outcode and postcode have lower importance but still contribute to the model's predictions. This is particularly relevant for this study as it

suggest that the model can be suitable for setting up a real world postcode level anomaly prediction system.

This feature importance analysis can also guide energy providers and policymakers in focusing on the most relevant factors when monitoring and analysing electricity consumption patterns at the postcode level.

4.2.6 ERROR ANALYSIS

The correlation matrix in Figure 17 provides valuable insights into why the random forest model fails to accurately predict anomalous electricity consumption in certain cases. Each observed correlation relates to misclassification in specific ways as outlined below:

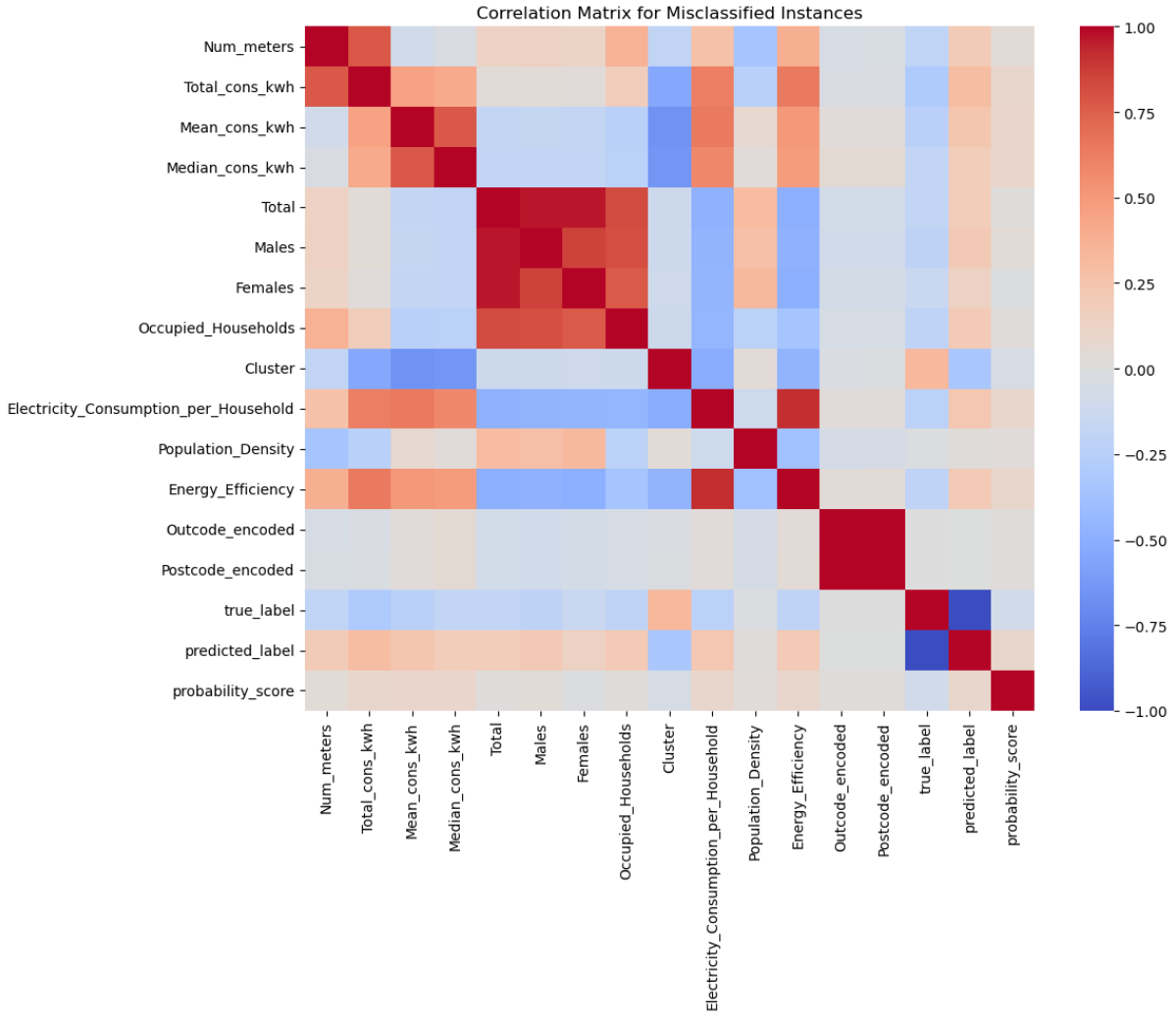


Figure 17: Correlation Matrix for misclassified instances

- The moderate positive correlation between Electricity_Consumption_per_Household and Total consumption, coupled with both of them having negative correlations with population-related features (such as total population, males, females) indicates potential misclassifications in areas where individual household consumption does not align with population trends. The model might struggle with cases where sparsely populated areas has a high consumption per household or areas that are densely populated but have an unexpectedly low consumption.
- Population_Density's negative correlations with most of the other features suggest that the model may misclassify instances in densely populated areas that don't follow typical consumption patterns. For example, urban areas with unique energy use profiles could be particularly prone to misclassification.
- Energy_Efficiency's positive correlation with consumption measures but negative correlation with population features points to potential misclassifications when energy-efficient areas show high consumption or vice versa. The model may struggle to accurately classify anomalies in scenarios where efficiency doesn't translate to expected consumption patterns.
- Minimal correlations of Outcode and Postcode with other features indicate that geographical factors may not strongly influence the model's decisions.
- The probability_score's strong positive correlation with the predicted_label and strong negative correlation with the true_label indicates that the model is confidently incorrect in its misclassifications. This suggests that the model may be overfitting to certain patterns and failing to recognise genuine anomalies that don't fit these patterns.

These observations collectively indicate that misclassifications often occur in complex scenarios where multiple factors interact in unexpected ways, or when instances deviate from the strong correlations the model has learned. The model appears to struggle with deep cases that require a more sophisticated understanding of feature interactions and local contexts.

4.2.7 DISCUSSION OF FINDINGS:

1. Model Performance:

The high performance of Random Forest models in this study and in previous studies suggests that ensemble methods are particularly effective for detecting anomalous electricity consumption patterns. At an accuracy of 0.959857 and an AUC-ROC of 0.993867, it outperformed for instance the model by ELhadad et al. (2022) who achieved their best accuracy of 90%. Both results also show that ensemble models can capture complex relationships in the data, which is crucial given the diverse nature of electricity consumption influenced by various factors such as demographics, and overall usage patterns.

2. Balancing Precision and Recall:

The high and balanced precision and recall scores across all the models indicate that the approach is effective in identifying both normal and anomalous consumption patterns. This balance is crucial in a practical setting, where both false positives (incorrectly flagging normal consumption as anomalous) and false negatives (missing actual anomalies) can have significant consequences. Banik et al. (2023) failed to achieve this balance probably as a result not handling class imbalance causing their model to be biased toward normal cases than abnormal cases.

3. Potential for Real-world Application:

The high evaluation metric values (most especially F1_score and AUC scores) suggest that the machine learning approach has strong potential for real-world application in systems that can detect and predict anomalous electricity consumption patterns. This could be invaluable for energy companies in identifying potential fraud, infrastructure issues, or unusual consumption trends that warrant investigation.

4. Scalability and Generalization:

The success achieved by the models in this study in handling postcode-level data also suggests that this approach could be scalable to larger geographic areas or even adapted to smaller units of analysis (for example, individual households), provided appropriate data is available and the right algorithm are employed.

5. Continuous Monitoring and Adaptation:

The analysis of misclassifications and feature importance also suggests that for real world deployments, it would be crucial for the models to be continuously monitored

and adapted to the current trend. This is because consumption patterns may change over time due to various factors (technological advancements, policy changes or even population), therefore regular retraining, validation and tuning of the model would ensure its continued effectiveness.

6. Integration with Domain Expertise:

Despite the strong performance of the machine learning models in this study, it will be highly beneficial to integrate the models with domain expertise for practical implementation. Irrespective of how accurate machines can be, human knowledge is still necessary. Hence, leveraging the knowledge of an expert in all the fields involved can provide more context to model predictions and help in the final decision-making process regarding identified anomalies.

In summary, the machine learning approach demonstrated in this analysis shows significant promise in predicting anomalous electricity consumption patterns at the postcode level. The high performance across multiple models, particularly Random Forest and KNN, shows an excellent and reliable anomaly detection capabilities. The insights gained from feature importance and error analysis provide valuable direction for both improving the model and understanding the underlying factors contributing to anomalous consumption patterns. This approach has the potential to significantly enhance energy management, fraud detection, and infrastructure planning in the electricity sector, ultimately leading to more efficient and reliable energy distribution systems.

CHAPTER 5: PROJECT MANAGEMENT

Project management is critical to the successful execution of complicated research projects. It gives a structured approach to project planning, execution, and control (Kaufmann & Kock, 2022). It allows for optimal resource utilisation, risk mitigation, and on-time delivery of results. Project management improves work efficiency, maintains focus on goals, and adaptation to problems, resulting in high-quality research outcomes and satisfaction. This section discusses the approaches taken in the management and carrying out of this study

5.1 | PROJECT SCHEDULE

The project schedule is outlined in the Gantt chart in Appendix 2

The project largely adhered to the original timeline. However, the error analysis phase took longer than anticipated due to the complexity of interpreting misclassifications. This slight delay was mitigated by accelerating the report preparation phase.

5.2 | RISK MANAGEMENT

Table 7: Potential Risks, impacts and mitigation plan

S/N	POTENTIALS RISKS	PROBABILITY OF OCCURRENCE	IMPACT ON THE PROJECT	MITIGATION PLAN
1	Data quality issues	Moderate	High	Implementing robust data cleaning procedures and validate data sources
2	Model performance below expectations	Moderate	High	Planning to use multiple model types and ensemble methods
3	Computational resource limitations	High	Low	Code Optimisation code and use computers available in school labs or cloud computing resources if necessary
4	Time overruns	Moderate	Low	Task prioritisation and Allowance for buffer time into the project

				schedule for tasks that pose this risk
5	Ethical concerns regarding data privacy	Low	High	Seeking ethical approval and ensuring compliance with data protection regulations and anonymize data

5.2.1 | Materialized Risks:

Two of the anticipated risks materialised during the project. Firstly, the risk on computational resource limitation which occurred during the hyperparameter tuning phase. The machine I was running the project on took almost two days to complete the grid search process. Therefore, to avoid this occurring again, subsequent grid search procedures were conducted on school provided computers at the engineering laboratories with the capacity to handle such computations.

Secondly, the risk of time overrun materialized during the error analysis phase because of the complexity of interpreting misclassifications. This was managed by seeking appropriate help and accelerating the report finalisation phase.

5.3 | QUALITY MANAGEMENT

To ensure the standard quality of the project, the following standards were adopted:

- APA referencing style
- IEEE standards for technical documentation

The following techniques were used to Review Progress:

- Bi - Weekly progress meetings with supervisor
- Supervisor Email feedback on report quality
- Code reviews using pull requests
- Peer review of analysis results

The following were used in the evaluation of results:

- Comparison of model performance against predefined benchmarks
- Validation of results through cross-validation techniques
- Supervisor’s review of final conclusions

5.4 | SOCIAL, LEGAL, ETHICAL AND PROFESSIONAL CONSIDERATIONS

To prevent any violation or misconduct, the following were considered while carrying out this research

1. **Data Protection and Privacy:** in order not to violate the data laws:
    - I ensured compliance with GDPR rules for handling personal data and made sure that all postcode-level data were in anonymity to prevent individual identification
    - Secure data storage and transmission protocols were implemented.
    - Clear data retention and deletion policies were established.
2. **Ethical Considerations:**
    - I obtained ethical approval for the use of electricity consumption data and considered potential biases in the model and their implications for fair decision-making
    - The societal implications of the research findings and recommendations were considered, ensuring that it does not disproportionately burden vulnerable populations particularly low-income households
    - Regular ethical audits were conducted throughout the research process.
3. **Professional Conduct:**
    - I maintained transparency in reporting both positive and negative results and all sources of external information were cited giving due credit to the owners of the work.
    - Established scientific methodologies and best practices in energy consumption research were adopted in this study
    - Research methods and findings were subjected to supervisor review processes

CHAPTER 6: CRITICAL APPRAISAL AND CONCLUSION

6.1 | CRITICAL APPRAISAL

This study presents a comprehensive approach to detecting and predicting anomalous electricity consumption patterns at the postcode level, addressing a significant gap which exists in most of the previous researches. While previous studies have either been on broader geographical cases or individual building analysis, the choice to adopt a postcode-level analysis strikes a balance between both, while also offering practical insights for energy management and policy-making at a community level.

The study uses a multifaceted methodology where an ensemble method combines two different anomaly detection techniques: Local Outlier Factor (LOF) and Isolation Forest whose results were used in labelling the dataset. The labelled dataset was then used in training different machine learning models. This approach shows a comprehensive understanding of contemporary machine learning process. A notable aspect of this study is its rigorous data preprocessing, particularly the use of the Synthetic Minority Over-sampling Technique (SMOTE) to solve class imbalance, which is a critical issue in anomaly detection tasks. Furthermore, the comparison of different machine learning models (Random Forest, KNN, Decision Tree, and Gradient Boosting) demonstrates a thorough approach to model selection, with the Random Forest model coming out on top with high accuracy (0.959857) and AUC-ROC (0.993867) scores.

The study's extensive feature importance analysis, which emphasises the impact of population density and per-household electricity consumption, provides useful information for energy providers and regulators. This in-depth look into the factors that influence anomalous patterns displays a thorough understanding of the complicated interplay of numerous aspects of power usage. The research further stands out for its thorough error analysis, including a correlation analysis of misclassified instances. This amount of depth in interpreting model errors and understanding the causes of misclassifications is uncommon in most researches done in this area. This helped in providing useful insights into the limitations and potential enhancements of predictive algorithms in this domain. The use of Grid Search for hyperparameter tuning of the Random Forest model demonstrates an awareness for the necessity of model optimisation, although the computational cost and efficiency of this method could be further explored.

While the research contributes significantly to the body of knowledge, there remain opportunities for improvement and future research. The lack of ground truth for validation of observed anomalies makes it difficult to determine whether the Ensemble Method's extra detections are true positives or false alarms. Though injected outliers were used to test the effectiveness of the ensemble method but there were no grounds or benchmarks to establish that the outliers it detected in the real dataset were actually outliers. Furthermore, while the use of SMOTE resolves class imbalance, the danger for overfitting associated with this technique was not adequately examined. The large performance difference between the Random Forest model and other methods, particularly Gradient Boosting, merits further investigation. Future research might concentrate on giving concrete remedies to solve reported misclassifications and investigating how the model would respond to changing consumption patterns over time, which is crucial for long-term applicability. Despite these limitations, the work displays a deep understanding of machine learning techniques used to discover anomalies in electricity usage, as well as proficiency in data preparation, developing models, and result interpretation. It adds significant knowledge to the sector, especially with its postcode-level emphasis and extensive analytical approach, opening the way for deeper and more successful energy management techniques.

6.2 | CONCLUSION

This conclusion summarizes the key achievements of this research in developing a machine learning model for electricity consumption anomaly detection and prediction. This is by reflecting on the effectiveness of the Random Forest algorithm and ensemble methods of outlier detection. Furthermore, also discussing the impact of demographic factors while acknowledging the limitations in this study. Finally, recommendations were proposed for future research in this field.

6.2.1 | ACHIEVEMENTS

With the baseline of this research centred around anomalies (outlier) detection even before any form of predictive modelling could be implemented, this study was able to establish that the ensemble method of outlier detection is more efficient and effective than single algorithms like Local Outlier Factor or Isolation Forest. This was because of the ability of Ensemble methods in taking advantage of the strengths of different algorithms.

In line with the major objective of this research, by way of extensive comparison to other models, this study was able to establish Random Forest as the best performing algorithm for predicting anomalous electricity consumption. Thus, successfully developing a model that predicts if a post-code electricity consumption data point is normal or not. This is majorly because of its strong ability to handle nonlinear data and also its robust nature in working with data that has outliers.

From the exploratory data analysis, this study was also able to highlight the direct proportionality of demographic factors to electricity consumption whether abnormal or not. The feature importance carried out on the Random Forest model validated the strong influence of demography in the predicting of anomalous consumption by showing the strong impact “population density” and “occupied households” has on the model and also showed that they are the most important influencing factors. In order words, electricity consumption is as a result of human occupancy irrespective of abnormality, which directly impacts the prediction of the model.

However, it is necessary to note that these achievements are not all encompassing but rather a key contribution to this field of research. Further research will still be needed before perfection can be achieved and consequently coming to the stage of real world deployment.

6.2.2 | FUTURE WORKS

As a result of time constraints and availability of resources, this study has some limitations which were duly recorded earlier. However, based on these limitations, this section highlights some recommended area to explore for future researches on anomaly detection and prediction.

Firstly, it will be interesting to investigate the effects of temporal factors by introducing time series data in the analysis which has the potential of revealing repetitive or seasonal anomalous consumptions. This is particularly important given the evolving nature of energy demand and consumption in the modern world. Also additional factors like weather and economic indicators which influence consumption can further be investigated

Secondly, as electricity consumption data can be large, potentially running into billions of unlabelled data points especially when collected over an extended period of time. It will be recommended to explore the application of deep learning models like neural

network. This is because deep learning models are mostly designed to work like the human brain which means they have the ability to capture patterns that traditional machine learning algorithms could not.

Lastly, as the long term goal of this research is for it to be deployed for real world application such that it can trigger immediate response by utility companies. Therefore, an investigation into the detection and predictive ability of the developed models in real time will be recommended. This could be done by designing a prototype reinforcement learning based real-time anomaly detection system that can evolve with changing consumption patterns. The system can be fed with streaming data across different locations in the process of investigation.

7.0 | STUDENT REFLECTION

According to Gibbs et al. (1998), you cannot learn from an experience if you do not take out time to sufficiently reflect on it. This project has been nothing short of a good experience while coming with its own challenges and rewards. It gave me the opportunity to see the practicability of data science in relation to electricity consumption. Coming from an electrical engineering background, I was excited to work on something related so I can have a firsthand experience on the possible roles data science can play in the advancements in electrical engineering.

I am quite satisfied with my achievements and learnings from this project as I achieved my objectives while also picking up ideas on where my further interests of research lies. Bar the time constraint, I particularly would have loved to explore the idea of seeing the performance of the models I developed here in real time. However, beyond this module I would have sufficient time to independently carry it out as a research

In this research, the first challenged faced was obtaining a suitable dataset that could fit into my plan on achieving my objectives. Given the restrictions on allowed repositories, this task proved challenging but with the guidance and suggestions of Dr Diana, I was able to obtain the two datasets with the right features that suits perfectly to this research.

Secondly, I encountered computational issues during the experimentation phase as I had the grid search hyperparameter tuning running on my machine for over two days without completion. This was a drawback as I lost some days beyond the buffer allocations on my project management. However, I was able to run these experiments more smoothly on the machines provided in the engineering laboratories and move on to other parts of my experiments

Finally, being used to the Nigerian pattern of dissertation reporting, adapting to the United Kingdom’s dissertation report structure was a big challenge. To overcome this and put together my report, I needed a lot of feedback from Dr. Diana which she graciously and timely provided

In conclusion, I was able to understand how to source for data relative to my research interest and how to write a standard academic report. Indeed, I am still in a learning process in becoming a data scientist but I am satisfied with the general knowledge I was able to gain from this research.

REFERENCES

anomaly noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com. (n.d.). https://www.oxfordlearnersdictionaries.com/definition/english/anomaly?q=anomaly

Al Metrik, M. A., & Musleh, D. A. (2022). Machine learning empowered electricity consumption prediction [Article]. Computers, Materials & Continua, 72(1), 1428. https://doi.org/10.32604/cmc.2022.025722

Assouline, D., Castello, R., Mauree, D., Zwahlen, N., Guido, M., Hamm, D., Vidulis, M., & Scartezzini, J.-L. (2020). A machine learning-assisted building electricity consumption profiling for anomaly detection. International Conference on Applied Energy 2020. https://doi.org/10.1016/j.egypro.2020.09.092

Beretta, D., Grillo, S., Pigoli, D., Bionda, E., Bossi, C., & Tornelli, C. (2020). Functional principal component analysis as a versatile technique to understand and predict the electric consumption patterns. Sustainable Energy, Grids and Networks, 21, 100308. https://doi.org/10.1016/j.segan.2020.100308

Bhandari, A. (2024, June 9). Feature Scaling: Engineering, normalization, and Standardization (Updated 2024). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/

Breiman, L., & Cutler, A. (2024). Understand random forest algorithms with examples. In Data Science Blogathon [Article]. Retrieved July 17, 2024, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Cai, Q., Li, P., & Wang, R. (2023). Electricity theft detection based on hybrid random forest and weighted support vector data description. International Journal of Electrical Power & Energy Systems, 153, 109283. https://doi.org/10.1016/j.ijepes.2023.109283

Cao, Y., Liu, G., Sun, J., Bavirisetti, D. P., & Xiao, G. (2023). PSO-Stacking improved ensemble model for campus building energy consumption forecasting based on priority feature selection. Journal of Building Engineering, 72, 106589. https://doi.org/10.1016/j.jobe.2023.106589

Copiaco, A., Himeur, Y., Amira, A., Mansoor, W., Fadli, F., Atalla, S., & Sohail, S. S. (2023). An innovative deep anomaly detection of building energy consumption using

energy time-series images. Engineering Applications of Artificial Intelligence, 119, 105775. https://doi.org/10.1016/j.engappai.2022.105775

Costa, V. G., & Pedreira, C. E. (2022). Recent advances in decision trees: an updated survey. Artificial Intelligence Review, 56(5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5

ELhadad, R., Tan, Y., & Tan, W. (2022). Anomaly prediction in electricity consumption using a combination of machine learning techniques. International Journal of Technology, 13(6), 1317. https://doi.org/10.14716/ijtech.v13i6.5931

ELHadad, R., Tan, Y., & Tan, W. (2023). Comparison of Enhanced Isolation Forest and Enhanced Local Outlier Factor in Anomalous Power Consumption Labelling. IEEE. https://doi.org/10.1109/icpea56918.2023.10093186

Fried, S., & Lagakos, D. (2023). Electricity and Firm Productivity: A General-Equilibrium Approach. American Economic Journal. Macroeconomics/American Economic Journal, Macroeconomics, 15(4), 67–103. https://doi.org/10.1257/mac.20210248

Gajowniczek, K., & Ząbkowski, T. (2017). Two-Stage electricity demand modeling using machine learning algorithms. Energies, 10(10), 1547. https://doi.org/10.3390/en10101547

Gautam Kunapuli. (2023). Ensemble Methods for Machine learning. Google Books. https://books.google.co.uk/books?id=hoK3EAAAQBAJ&lpg=PR11&ots=FcuSCs2V39&dq=ensemble%20methods%20are%20better%20that%20single%20methods&lr&pg=PR11#v=onepage&q=ensemble%20methods%20are%20better%20that%20single%20methods&f=false

Gibbs, G., Oxford Centre for Staff and Learning Development, Oxford Brookes University, Further Education Unit, Geography Discipline Network, University of Gloucestershire, Farmer, B., Eastcott, D., Sharpe, R., Alexander, J., Covell, J., Cowan, J., Habeshaw, T., Jaques, D., Kelly, T., & McGee, P. (1998). Learning by Doing, A guide to teaching and learningmethods. https://thoughtsmostlyaboutlearning.files.wordpress.com/2015/12/learning-by-doing-graham-gibbs.pdf

He, Y., Qin, Y., Wang, S., Wang, X., & Wang, C. (2019). Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network. Applied Energy, 233–234, 565–575. https://doi.org/10.1016/j.apenergy.2018.10.061

Himeur, Y., Alsalemi, A., Bensaali, F., & Abbes Amira. (2021). Smart power consumption abnormality detection in buildings using micromoments and improved K-nearest neighbors. In Department of Electrical Engineering, Qatar University, Doha, Qatar & Institute of Artificial Intelligence, De Montfort University, Leicester, UK, Int J Intell Syst (pp. 2865–2894). https://doi.org/10.1002/int.22404

Inuwa, M. M., & Das, R. (2024). A comparative analysis of various machine learning methods for anomaly detection in cyber attacks on IoT networks. Internet of Things, 101162. https://doi.org/10.1016/j.iot.2024.101162

Kan, X., Reichenberg, L., & Hedenus, F. (2021). The impacts of the electricity demand pattern on electricity system cost and the electricity supply mix: A comprehensive modeling analysis for Europe. Energy, 235, 121329. https://doi.org/10.1016/j.energy.2021.121329

Kaufmann, C., & Kock, A. (2022). Does project management matter? The relationship between project management effort, complexity, and profitability. International Journal of Project Management, 40(6), 624–633. https://doi.org/10.1016/j.ijproman.2022.05.007

Kesornsit, W., & Sirisathitkul, Y. (2022). Hybrid machine learning model for electricity consumption prediction using random forest and artificial neural networks. Applied Computational Intelligence and Soft Computing, 2022, 1–11. https://doi.org/10.1155/2022/1562942

Liu, T., Zhou, Z., & Yang, L. (2024). Layered isolation forest: A multi-level subspace algorithm for improving isolation forest. Neurocomputing, 581, 127525. https://doi.org/10.1016/j.neucom.2024.127525

Mougan, C. (2022, March 30). Isolation Forest from Scratch - Towards Data Science. Medium. https://towardsdatascience.com/isolation-forest-from-scratch-e7e5978e6f4c

Moure-Garrido, M., Campo, C., & Garcia-Rubio, C. (2022). Entropy-Based anomaly detection in household electricity consumption. Energies. https://doi.org/10.3390/en15051837

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in Machine learning Algorithms. Lecture Notes in Networks and Systems, 15–25. https://doi.org/10.1007/978-3-031-35314-7_2

Omid, Chatrabgoun, (2024, January 25). Introduction to Machine Learning, 7072CEM [pptx]. Aula. https://coventry.aula.education/

Piao, X., & Managi, S. (2023). Household energy-saving behavior, its consumption, and life satisfaction in 37 countries. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-28368-8

Pictorial-Representation-of-Random-Forest-regressor_W640.jpg (640×458). (n.d.). https://www.researchgate.net/publication/372415737/figure/fig2/AS:11431281216475279@1704771555617/Pictorial-representation-of-Random-Forest-regressor_W640.jpg

Pramoditha, R. (2022, April 16). Why do we need a validation set in addition to training and test sets? Medium. https://towardsdatascience.com/why-do-we-need-a-validation-set-in-addition-to-training-and-test-sets-5cf4a65550e0

Ramos, D., Faria, P., Gomes, L., & Vale, Z. (2022). A contextual reinforcement learning approach for electricity consumption forecasting in buildings. IEEE Access, 10, 61366–61374. https://doi.org/10.1109/access.2022.3180754

Saini, A. (2024). What is Decision Tree? [A Step-by-Step Guide]. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2024/05/decision-tree-algorithm/

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series. Proceedings of the VLDB Endowment, 15(9), 1779–1797. https://doi.org/10.14778/3538598.3538602

Sethjiwala, A. (2023, September 12). The History of Anomaly Detection - Techniques, tools and use cases. ChaosGenius - DataOps Observability Platform for Snowflake. https://www.chaosgenius.io/blog/a-brief-history-of-anomaly-detection/

Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2022). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, 5(4). https://doi.org/10.1002/eng2.12599

Silva, C., Faria, P., Vale, Z., Corchado, J. M., & The Authors. (2022). Demand response performance and uncertainty: A systematic literature review. In Energy Strategy Reviews (Vol. 41, p. 100857). https://doi.org/10.1016/j.esr.2022.100857

Soelami, F. N., Utama, P. H. K., Haq, I. N., Pradipta, J., Leksono, E., & Wasesa, M. (2021). Data driven building electricity consumption model using support vector regression. Journal of Engineering and Technological Sciences, 53(3), 210313. https://doi.org/10.5614/j.eng.technol.sci.2021.53.3.13

Srivastava, T. (2024). A Complete Guide to K-Nearest Neighbors (Updated 2024). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

Suganthi, L., Samuel, A. A., Department of Management Studies, Anna University, Chennai 600025, India, & VIT University, Vellore 632014, India. (n.d.). Energy models for demand forecasting—A review. In Renewable and Sustainable Energy Reviews (Vol. 16, pp. 1223–1240). https://doi.org/10.1016/j.rser.2011.08.014

Wang, Y., Zhang, N., & Chen, X. (2021). A Short-Term residential load forecasting model based on LSTM recurrent neural network considering weather features. Energies, 14(10), 2737. https://doi.org/10.3390/en14102737

Ward, W., Li, X., Sun, Y., Dai, M., Arbabi, H., Tingley, D. D., & Mayfield, M. (2023). Estimating energy consumption of residential buildings at scale with drive-by image capture. Building and Environment, 234, 110188. https://doi.org/10.1016/j.buildenv.2023.110188

Xie, J., Sage, M., & Zhao, F. (2023). Feature selection and feature learning in machine learning applications for gas turbines: A review. Engineering Applications of Artificial Intelligence, 117, 105591. https://doi.org/10.1016/j.engappai.2022.105591

APPENDIX 1: Links to Datasets and Experimentation Codes

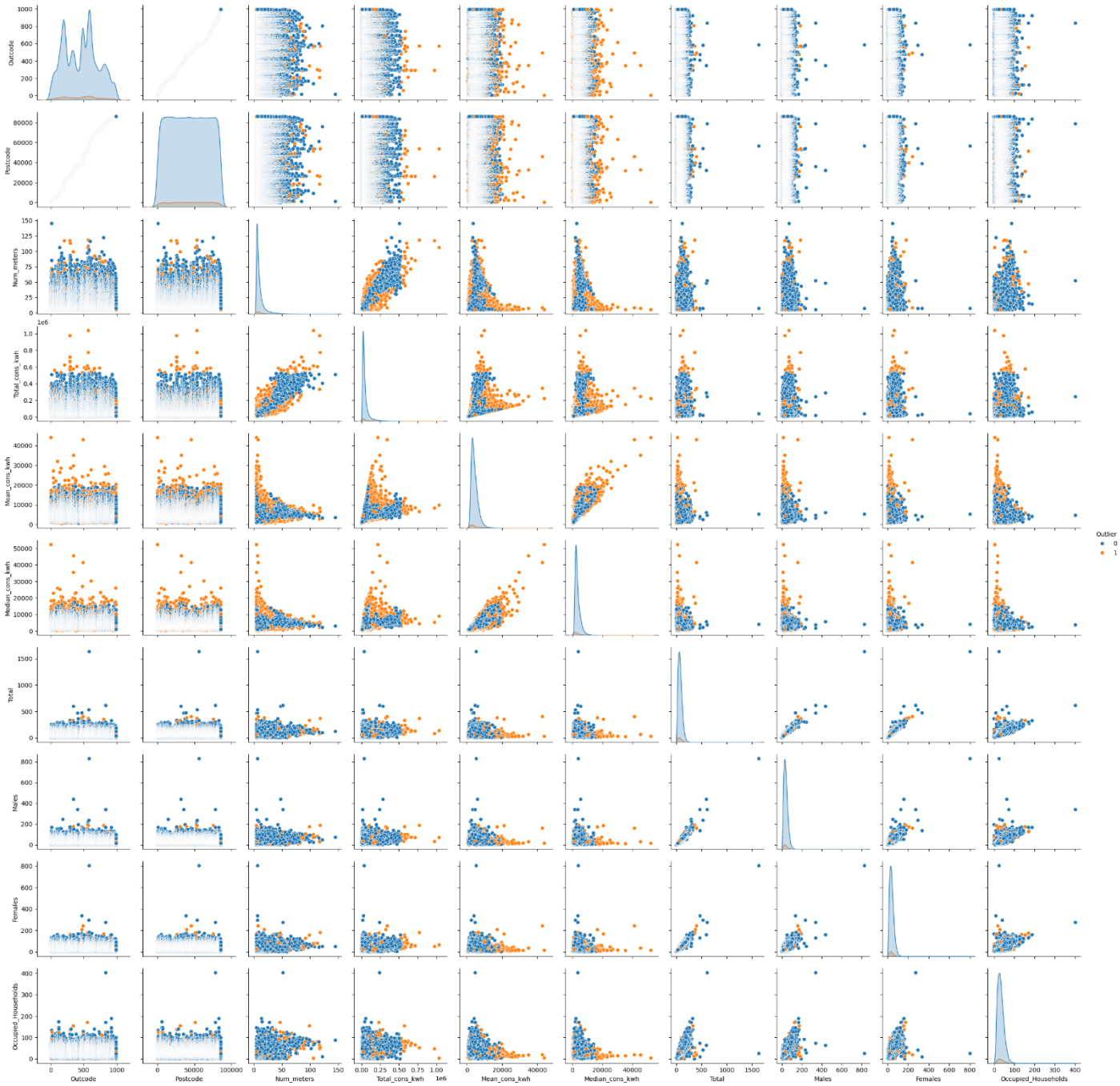
Postcode Level Economy 7 Electricity 2022 Dataset:
https://assets.publishing.service.gov.uk/media/65b0d21ef2718c0014fb1be6/Postcode_level_economy_7_electricity_2022.csv

Census Postcode Estimates Table:
https://www.nomisweb.co.uk/output/census/2011/Postcode_Estimates_Table_1.csv

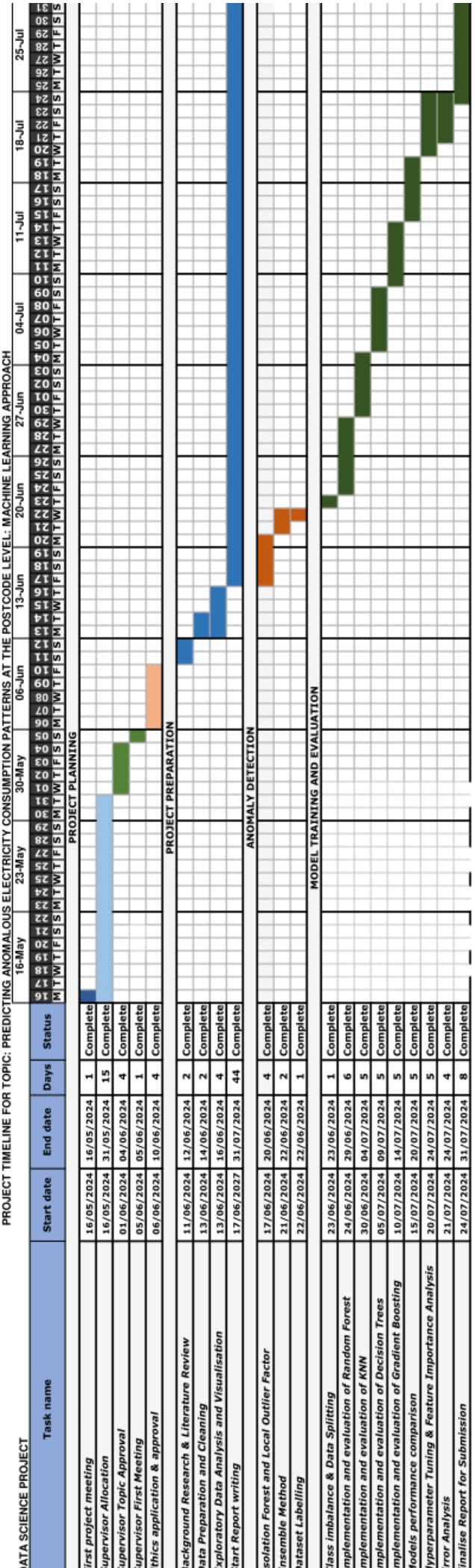
Exploratory Data Analysis: https://github.com/ezenwajiakujc/DATA-SCIENCE-PROJECT/blob/main/EDA.ipynb

Experimentation Code: https://github.com/ezenwajiakujc/DATA-SCIENCE-PROJECT/blob/main/Analysis%20and%20Results.ipynb

APPENDIX 2: Pair plot of the variables showing nonlinear relationships and overlaps



APPENDIX 3: Gantt Chart



APPENDIX 4: ETHIC APPROVAL CERTIFICATE

Certificate of Ethical Approval

Applicant:	Chinedu Ezenwajiaku
Project Title:	Predicting Anomalous Electricity Consumption Patterns at the Postcode Level: Machine Learning Approach

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	10 Jun 2024
Project Reference Number:	P177443

APPENDIX 5: ETHICS APPROVAL CHECKLISTS

Low Risk Research Ethics Approval

Project title

Predicting Anomalous Electricity Consumption Patterns at the Postcode Level: Machine Learning Approach

Record of Approval

Principal Investigator's Declaration

I request an ethics peer reviewI confirm that I have answered all relevant questions in this application honestly	X
I confirm that I will carry out the project in the ways described in this application. I will immediately suspend research and request an amendment or submit a new application if the project subsequently changes from the information I have given in this application.	X
I confirm that I, and all members of my research team (if any), have read and agree to abide by the code of research ethics issued by the relevant national learned society.	X
I confirm that I, and all members of my research team (if any), have read and agree to abide by the University's Research Ethics Policies and Processes.	X
I understand that I cannot begin my research until this application has been approved and I can download my ethics certificate.	X

Name: Chinedu Ezenwajiaku (7150CEM)

Date: 06/06/2024

Student's Supervisor (if applicable)
I have read this checklist and confirm that it covers all the ethical issues raised by this project fully and frankly.I also confirm that these issues have been discussed with the student and will continue to be reviewed in the course of supervision.

Name: Dr. Diana Hintea

Date: 10/06/2024

Reviewer (if applicable)
Date of approval by anonymous reviewer: -

Low Risk Research Ethics Approval Checklist

Project Information

Project Ref	P177443
Full name	Chinedu Ezenwajiaku
Applicant type	Taught student
Area	College of Engineering, Environment and Science
Sub Area	School of Computing, Mathematics and Data Science
Supervisor	Dr. Diana Hintea
Module Code	7150CEM
EFAAF Number	
Project title	Predicting Anomalous Electricity Consumption Patterns at the Postcode Level: Machine Learning Approach
Date(s)	16 May 2024 - 01 Aug 2024
Created	31/05/2024 16:29

Project Summary

Objective:
To develop a machine learning model to predict anomalies in electricity consumption patterns within a geographical area using the "Postcode level economy 7 electricity 2022" dataset from the UK government's national statistics.
Research Plan:
1. Data Pre-processing and Analysis: Using the dataset containing electricity consumption data to extract and derive relevant features such as geographical location, demographic information, and conditions indicative of anomalous energy consumption.
2. EDA: Identifying and analysing patterns in the data, highlighting any significant correlations or trends.
3. Model Development: Training and evaluating different machine learning algorithms to identify patterns and predict anomalies in electricity consumption.
Potential Applications: This project can be integrated into a recommendation system with significant economic benefits, including:
i. Energy demand management
ii. Identification of energy theft and irregularities
iii. Geographic targeting for renewable energy promotion
Target Audience:
Utility Companies: The model will assist utility companies in making data-driven decisions to enhance efficiency and sustainability.

Names of Co-Investigators and their organisational affiliation(place of study/employer)		
Full name	Notified	Accepted

Is this project externally funded?	No

Are you required to use a Professional Code of Ethical Practice appropriate to your discipline?	No
Have you read the Code?	No
Will this project involve international engagement or partnerships?	
Does your research fall within at least one of the 17 sensitive areas of the economy?	

APPENDIX 6: MEETING AND EMAIL RECORDS

S/N	DATE	TYPE	DISCUSSION	SUPERVISOR'S SUGGESTION	ACTION
1	04/06/2024	Email	Topic Confirmation	Obtain appropriate data for the research.	I obtained the required data
2	05/06/2024	Teams	Introductory	Guide to completion of Ethic application.	I completed and submitted the Ethics application
3	21/06/2024	Email	Dataset and Methodology	Agreed to the size of the dataset but expressed some concern with the limitation on features in the dataset. Approved chosen methodology while highlighting the importance of justifications.	I obtained the UK census data in addition to the Electricity data to add demographic factors in the research
4	01/07/2024	Teams	Chapter 1 & 2	Improve my literature review by using current research. Include a "Related works" section and Identifying Gaps	Review my literature with paper not beyond 2021 and include work related to my research while also performing a gap analysis
5	24/07/2024	Email	Chapter 1,2 & 3	Include a section in the introduction that states the structure of the report Make my research questions clear by including a "Research Question" section Include Visualisation of the data Provide justifications for the chosen Machine Learning algorithms	Included the "Report structure" section Added the research questions as a section Provided a visualization of the data and also justifications for the methodology

6	26/07/2024	Email	Chapter 4	Be more critical in comparing results with those presented in literature review Include a “Project Management” chapter	Critically compared my results with the one in the literature review
7	01/08/2024	Email	Chapter 6	Answer the research questions in the conclusion chapter Expand the social/ethical issues areas	Answered the research questions and expanded the social / ethical issues in consideration to the generic ethical practices with researches around electricity consumption