A Methodological Analysis of Consumption Patterns and Anomaly Detection in Prepaid and Postpaid Metering Systems: An N-BEATS-Driven Predictive Modeling Approach

by

Nusaba Islam
20301407
Partha Debnath
20301074
Md Rakibul Islam
20301373
Awon Bin Kamrul
20301367
Md Rishat Sheakh
20301305

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Nusaba Islam
20301407

Partha Debnath
20301074

Md Rakibul Islam
20301373

Awon Bin Kamrul
20301367

Md Rishat Sheakh
20301305

Approval

The thesis titled "A Methodological Analysis of Consumption Patterns and Anomaly Detection in Prepaid and Postpaid Metering Systems: An N-BEATS-Driven Predictive Modeling Approach" submitted by

1. Nusaba Islam(20301407)
2. Partha Debnath(20301074)
3. Md Rakibul Islam(20301373)
4. Awon Bin Kamrul(20301367)
5. Md Rishat Sheakh(20301305)

Of Summer 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on October 20, 2024.

Examining Committee:

Supervisor:
(Member)

Md. Tawhid Anwar
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Md. Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator::
(Member)

Dr. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The introduction of the prepaid metering system has caused severe dissatisfaction among the masses, claiming that there was a noticeable spike in the bills of their monthly electricity bill. This issue was addressed in this thesis paper through a rigorous investigation, comparing the existing prepaid metering system with the previously used post-paid metering system to identify the underlying causes of this discrepancy. Related datasets were collected through the Dhaka Power Distribution Company (DPDC), mainly focusing on the Paribagh, the first area introduced with this intelligent prepaid meter system. The dataset includes 22,000 individual customers' billing information, and then a robust survey was conducted to accumulate 1797 households' information on appliance usage, family size, and other relevant factors.

To predict the prepaid and postpaid consumption trends, we incorporated the deep learning model N-BEATS, which efficiently forecasts time series data. We also conducted several operations on our prepaid and postpaid data to better address the issue. Moreover, we deployed the isolation forest model to address the anomalies that would align with the underlying cause of the dissatisfaction of the masses.

To assist in validating our survey data, we deployed several data validation methods, such as the Kolmogorov-Smirnov test, and Pearson's correlation. Furthermore, Pearson's correlation technique has been used to demonstrate the correlation between prepaid and postpaid appliance usage. With N-BEATS providing a crystal difference between prepaid and postpaid data, Isolation Forest also directed us to the irregularities that may lean towards the reason for the increased billing.

To address customers' growing concerns, the authority might find our findings fruitful in solving the irregularities to ensure a seamless transition to the prepaid system. This research also illustrates actionable recommendations to ensure a satisfactory customer experience with transparent energy billing.

Keywords: Prepaid metering system, postpaid metering system, DPDC, electricity consumption, anomaly detection, time series forecasting, Prophet model, Isolation Forest, energy billing, consumption patterns, sliding window technique, energy consumption trends, billing discrepancies, machine learning, utility data analysis, seasonal trends, electricity usage prediction, residual analysis, outlier detection, predictive modeling, consumption behavior, N-BEATS model, household appliances, energy analytics, billing accuracy.

Dedication

We would like to dedicate this thesis to our loving parents, teachers, friends, as well as well-wishers.

Acknowledgement

We want to show our gratitude to the Almighty who blessed us with the strength, patience and wisdom to undertake this research journey.

We have found this research to have been a special experience, and without the support and encouragement of so many caring people, it could not have been. We greatly appreciate each and every one of them.

We would like to give a special thanks to our Supervisor, Md. Tawhid Anwar, Senior Lecturer, and our Co-supervisor, Md. Sabbir Ahmed, Lecturer. This research has been underpinned by their unwavering support and valuable guidance. We are beyond grateful for their patience and wisdom, and it has helped us big time. Working with such encouraging mentors has been a true blessing.

It is also a matter of great appreciation to A.S.M Saiful Islam Laskar, Executive Engineer, and Under ground Cable (North) Department, Mohammad Emdadul Haque, Manager (ICT), and DGM (ICT), Development Department, Dhaka Power Distribution Company Limited (DPDC). Their valuable insights, permissions and suggestions during the design of our survey form and throughout the data collection process were incredibly valuable. We sincerely want to thank them for the time and support they have given us

We would like to thank the residents of the Paribagh area, who graced us with their presence and time to participate in our survey. We would like to thank them for helping out and making this research possible through their cooperation.

Finally, and most importantly, we are humbled by our parents, brothers, and sisters, who have been our strongest supporters on the journey. Without their endless love, sacrifices, and encouragement, this research would not have been possible. This is thanks to their constant support and belief in us.

To all of you, this research is dedicated.

Table of Contents

List of Figures

Chapter 1

Overview

1.1 Introduction

Understanding energy consumption patterns and their strange behavior in the prepaid and postpaid meter systems has become crucial for optimizing resource allocation and stability. In this research work, A Methodological Analysis of Consumption Patterns and Anomaly Detection in Prepaid and Postpaid Metering Systems: An N-BEATS-Driven Predictive Modeling Approach, we will show the different behavior of the prepaid and postpaid energy consumption data through the use of several machine learning models and algorithms. The research centers on the N-BEATS (Neural Basis Expansion Analysis Time Series) model, a state-of-the-art deep learning method that analyses time series data. We collected the required dataset for this research from the Dhaka Power Distribution Company Ltd. and an in-person door-to-door survey. Using different models and algorithms, we showed different characteristics of prepaid and postpaid data; this makes the model adept at spotting complex patterns such as sudden spikes.

This research explores the fundamental trend by applying the N-BEATS model to both prepaid, postpaid, and survey data. Therefore, using this data, the model looks for more profound, hidden patterns to help inform how various consumer groups use energy. For example, prepaid users tend to watch and adjust what they consume based on their balance and real-time feedback, whereas postpaid users may behave differently as they receive their bills later. N-BEATS then facilitates highlighting these behavioral differences that are hard to see through the noise in the data. For power distribution companies, these insights are essential as they can lead to better demand management, better billing, and building programs to maximize energy efficiency. In addition, this research addresses the bigger picture of energy management. The recent development of smart meters and advanced metering interfaces (AMI) necessitates better utility tools to monitor energy consumption online, detect problems, and respond promptly. Here, the combination of N-BEATS and isolation forest models could represent a blueprint for a future energy monitoring system that is not only more accurate but more intrinsic to the ever-changing energy landscape.

One of the parts of this research is to understand consumption trends. Prepaid and postpaid metering systems can get anomalies—abrupt variations away from the norm

that might indicate a serious problem like faulty meters, billing errors, or energy theft. One of the problems associated with these issues is that traditional methods for detecting these problems, such as manual inspections or simple rule-based systems, can be prolonged and inefficient for large complex data sets usually generated by metering systems.

This is the second part of this research; we introduce the Isolation Forest algorithm. The isolation forest (IF) is an unsupervised machine-learning algorithm that finds anomalies. Unlike many other approaches in this domain, it does not require labeled data which is very difficult to obtain. Instead, it picks data points and isolates them in the binary (Liu et al., 2008). This research will show how the Isolation Forest can detect unusual energy consumption patterns. When there is a thief, when a piece of equipment fails, or when an abnormality such as a spike or drop happens this algorithm will separate them. This Isolation Forest is conducive to use on large and complex datasets because it is extremely good at finding anomalies in complex energy systems. Additionally, we use N-BEATS for trend analysis and Isolation Forest for anomaly detection to provide a holistic view of energy consumption. Isolation Forest detects unusual or suspicious behavior that might need further investigation, while N-BEATS gives detailed forecasting and reveals normal usage patterns. These models provide a robust methodology for addressing critical problems in prepaid and postpaid metering systems.

One of the main objectives of this research is to find out the users' concern of increasing their electricity bill after introducing the prepaid meter by replacing the traditional postpaid meter. This research will improve our understanding of how prepaid and postpaid meter users use different amounts of energy. For example, prepaid consumers are more likely to behave based on real-time feedback, such as watching their balance go down, while postpaid consumers are more likely to behave based on billings. The research intends to provide such utility companies with a better understanding of these differences to build better strategies such as cutting operational costs, managing demand, and improving customer satisfaction.

It is envisaged that the research findings will have far-reaching implications for those affected by this, such as utility companies and government policymakers, all the way to the consumers. More accurate forecasting and better anomaly detection may help utility companies boost operational efficiency, reduce loss from meter errors or theft, and increase customer satisfaction. Comparing prepaid and postpaid systems can be useful to policymakers trying to shape regulations or incentive programs to spur responsible energy use. However, this research has the potential to provide consumers, particularly those on prepaid meters, with greater energy use control and, therefore, more predictable costs and a clearer picture of the individual's consumption habits.

Therefore, this thesis extends the limits of energy consumption forecasting and anomaly detection through the integration of advanced machine learning models—N-BEATS for trend analysis and Isolation Forest for anomaly detection—into a comprehensive analysis of prepaid and postpaid metering systems. By shedding light on the usage patterns, this research will reveal the anomalies that will enable us to gain an actionable insight to deploy to achieve sustainability, resilience, and

economy of energy management in an ever evolving and complicated energy domain.

1.2 Research Problem

Customer concerns have also risen in the shift from postpaid to prepaid metering systems, with many customers stating an unbelievable rise in electricity bills. Since it filled this void, questions of prepaid billing systems' transparency, accuracy, and fairness have arisen. Adversely, prepaid meters were initially widely perceived by customers as a means of obtaining better control over their consumption and expenditure("Perceptions And Perceived Acceptance," 2020)[1]. However, many users now argue that the cost of electricity is high, as some argue that the prepaid meters will consume more units than expected, increasing the overall cost (Aribisala & Mohammed, 2021). [2]. As a result of this perception, more dissatisfaction and mistrust surrounding this new system has arisen when contrasted with the previous postpaid setup, where users could use more flexibility with their payments.

So the central question is whether the increased cost in these markets is attributable to fundamental changes in consumption behavior, or technical inequities in the metering systems, or other systematic error. Prepaid meters mitigate users' tendency to forget their consumption. However, this shift may lead to perceived costs growing as users start to monitor their consumption closely and have to pay upfront. Furthermore, the shift from a credit-based, post-pay accounting to a prepay system creates financial pressure on subscribers who had settled their bills at month's end, which makes the prepay experience more burdensome.

This research aims to analyze and compare consumption patterns using the latest predictive model N-BEATS applied to prepaid and postpaid datasets. Predicated on the assumption that billing discrepancies are caused by technical or other changes in user behavior, amongst other possibilities, we employ anomaly detection methods, such as the Isolation Forest, to highlight outlier instances where consumption deviates significantly from expected patterns (Journal of Digital Food, Energy & Water Systems, 2021). Additionally, the study explores factors such as year's seasons, household characteristics, and appliance usage to better understand consumption trends.

The findings will provide utility companies and policymakers with actionable insights to address customer concerns and a transparent process toward more prepaid systems. The study seeks to restore trust in prepaid metering by identifying the root causes of dissatisfaction and thus providing a fairer and more efficient energy management solution in the developing and developed contexts (Aribisala & Mohammed, 2021).[2]

1.3 Aims and Objectives

1.3.1 The Aim of the Research

The main objective of this research is to develop a framework for full analysis of electricity consumption in prepaid as well as postpaid metering systems along with capturing possible instances of fraud, technical faults, or unauthorized usage. To do this, we've used two advanced machine learning models:

- **N-BEAT (Neural Basis Expansion Analysis Time Series)**: The model described here can work with time series data, such as electricity usage patterns. It will help us to predict what energy consumption will do over time and show the primary differences between prepaid and postpaid systems.
- **Isolation Forest**: What makes this model so great is that it's capable of detecting anomalies like irregular consumption patterns, which could be cases of fraud or technical issues.

Additionally, we employed some additional techniques to improve the effectiveness of our models. To enhance the predictions of the N-BEATS model over time, we applied a sliding window technique so that it could react faster to changes in electricity usage patterns. We also applied a rolling average function to the Isolation Forest model to smooth these fluctuations and give us more accurate results when identifying unusual behavior in the data. These brought out more reliability and accuracy in our framework.

With both models, our framework can handle large amounts of data, predict typical consumption patterns, and detect any significant deviations that would warrant further investigation. But our research goes beyond the technical: we are looking to develop insights that can guide utility companies to run more efficiently, help policymakers manage electricity better, and foster more sustainable electricity practices.

In the end, this research is intended to provide a data driven solution that will not only help us to understand electricity consumption patterns better but also enhances the overall efficiency as well as reliability of both prepaid and postpaid metering systems.

1.3.2 Objective

There are several significant objectives we had while starting the study:

I. **Analyzing Consumption Patterns**: To explore and compare the electricity consumption trend between prepaid and postpaid meters with postpaid dataset from July 2021 to May 2022 and prepaid dataset from June 2022 to December 2023 we used the N-BEATS model. Now, by looking at these time periods, we would like to understand what behaviors with these two systems are different with respect to users, and how the switch to prepaid from postpaid affects energy usage. We will compare people's consumption habits with different

billing methods to see how they behave differently and get insights that can help to make electricity metering better fair and more efficient.

II. **Anomaly Detection**: To find any peculiarities in the electricity usage data we ran the Isolation Forest model. It facilitates the detection of possible problems like (meter tampering and unauthorized usage) and technical problems (that can have utilized the billing efficiency or system performance) radiation. If a utility company can detect these anomalies early, then it can improve the functioning of the system and improve its accuracy by speedy fixing of these anomalies. Not only does this approach reduce losses of revenue, but it also improves the general reliability of electricity distribution.

III. **Evaluate Prepaid Metering Effectiveness**: Another thing we looked at was how effective prepaid meters actually are at altering consumer behavior. We track changes in electricity consumption through time with prepaid systems and see if it changes us to be more mindful of our electricity usage compared to postpaid infrastructure. This analysis will determine whether prepaid meters influence energy conservation as users begin to be cognizant of their consumption. This will provide utility companies with important information about what the benefits and obstacles to the introduction of sustainable energy habits can be using prepaid systems.

IV. **Scalability of Models**: It should be ensured that deep learning models, especially the N-BEATS model being used in this research, can handle the growing amounts of data as smart meters get more predominant. At larger scales, these models must maintain their accuracy and performance as data volumes increase, and must continue to be applicable. The ability to scale like this is critical, as the broadening of smart metering systems will generate massive quantities of time series data, which deep learning models such as N-BEATS are suited to handle, as they are lateral in their capacity to understand intricate patterns and dependencies distributed across big data.

V. **Provide Recommendations for Policy Makers**: Based on the consumption patterns analysis and anomaly detection results, offer a few actionable recommendations to electricity companies and policymakers. The recommendations will include reducing energy efficiency, fighting fraud and optimizing future prepaid and postpaid metering systems.

VI. **Enhance Energy Management**: We use consumption patterns and anomaly detection to suggest strategies for electricity providers to optimize energy distribution and reduce inefficiencies. The system reliability will be improved, energy waste will be minimized, and more effective electricity will be managed.

Through these objectives, this research aims to create a data-driven framework that supports the development of more efficient, reliable, and scalable electricity metering systems, benefiting both consumers and electricity providers while informing future energy During our survey visits, many people shared a common concern: When they changed from postpaid to prepaid meters, they noticed that their electricity bills went up. This wasn't something that someone said once but it was something we heard over and over again from different households. There was frustration from

people, saying that the prepaid meters were taking them by surprise, and they were leaving them with higher bills than they should. On hearing this, we realized that it was an issue that our research needed to start addressing as well, because it seemed to matter to a lot of consumers.

1.4 Thesis Outline

Chapter 1: Overview
This chapter presents an initial discussion of consumption patterns and anomaly detection using prepaid and postpaid data. It introduces the motivation behind the research and outlines the problem statement. This section also includes research objectives, research questions, and, lastly, the overall research structure.

Chapter 2: Background Study
The background study includes an overview of the two types of prepaid and postpaid data. It also talks about two types of metering systems and their operation. We discussed the previous works in this field. Model sectiontion, methodology, and some algorithms are discussed in this chapter.

Chapter 3: Data Collection and Preprocessing
This is one of the main sections of this research, which includes a detailed description of the datasets and their collection process. It discusses the datasets' features and behavior. Since we used two distinct types of data collection methods, they are described in detail in this section. Lastly, data cleaning and preprocessing steps are included in this section.

Chapter 4: Experimental Results and Analysis
Chapter 4 dives into the experimental results by applying various machine learning models and algorithms, such as N-BEATS and Isolation Forest, to detect anomalies in consumption data. This describes the processes of training, testing, and validation methods such as Pearson correlation and KS test. Later, it focuses on the result using different algorithms and metrics, such as mean squared error (MSE) and mean absolute error (MAE) R-squared and analyzes the overall outcome of this research.

Chapter 5: Conclusion and Future Work
This final chapter recaps the overall by summarizing the key findings of this research. It shows the limitations of this research and suggests probable future work in this field, such as employing additional machine learning techniques, expanding the dataset, or exploring the impact of external factors on consumption patterns.

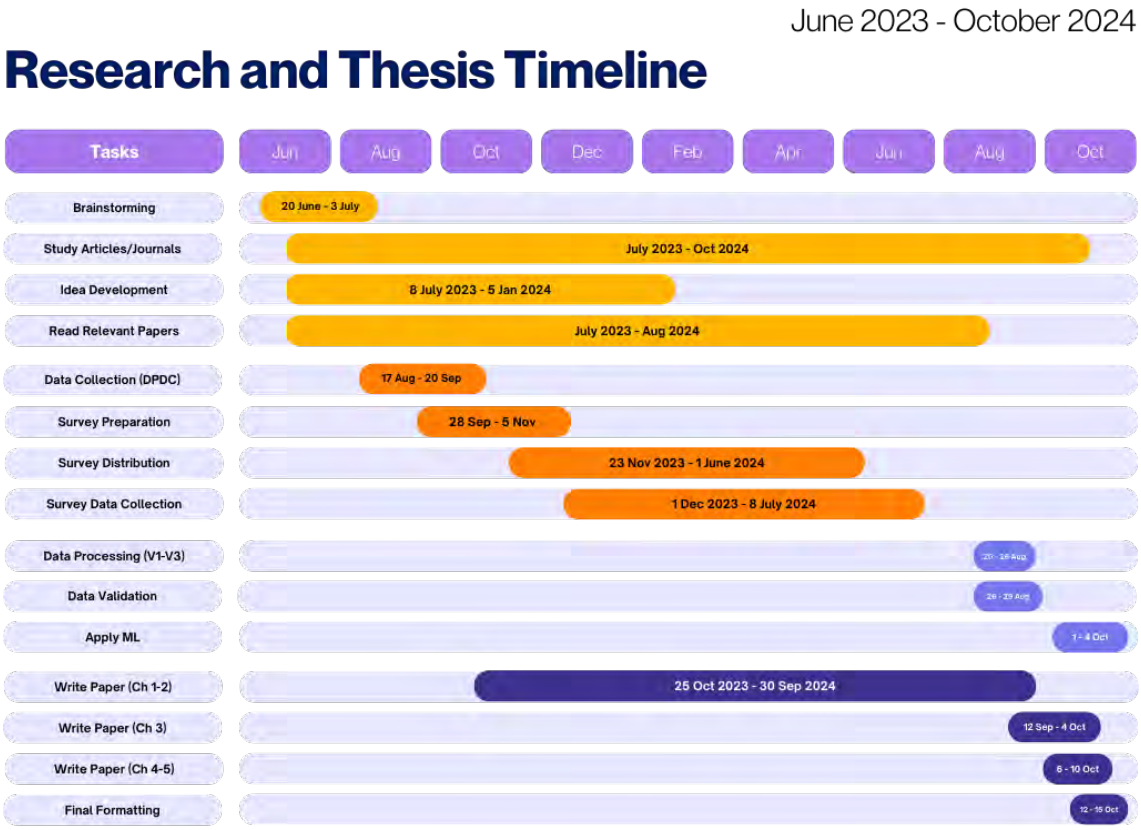

Figure 1.1: Timespan of the research

Chapter 2

Background Study

This chapter provides a detailed overview of the background study for the prepaid and postpaid metering system and a detailed discussion of the similar works in the literature review.

2.1 Literature Review

The authors of the work, Haque et al. (2021)[3], conducted a study to analyze the energy consumption of a six-story residential building. The building has six types of flooring sizes. It is located in Chittagong. For analyzing the energy consumption, they used different machine learning models including the features temperature, humidity, electric bill, etc. They used the coefficient of determination (R), mean absolute error (MAE), mean square error (MSE), root mean squared error (RMSE), and coefficient of variance (CV) for evaluating the outcome. Additionally, they used multiple linear regression (MLR), random forest (RF), support vector regression machine (SVR), and k-nearest neighbors (KNN) for analyzing the energy demand assessment. From the performance parameter table, they found that the RF model responded better than the other three models, and it calculated the total energy consumption, which was 340 MW per year.

As the title suggests, Khan et al. (2021)[4] studied the influencing factors affecting machine learning-based energy consumption prediction. The article aimed to improve forecast accuracy and analyze factors that cause significant errors using a novel error learning model. However, the data on energy consumption used in the experiment were taken from Jeju Island, South Korea, from January 2013 to 2018. The study used three machine-learning models namely Catboost, XGBoost, and Multi-Layer Perceptron sequentially to predict energy consumption, implementing the error curve learning technique. Thus, Catboost obtained the error curve, XGBoost predicted the error, and Multi-Layer Perceptron predicted the final energy consumption. After using different combinations of models for the simulation and testing phases, they came up with different statistics, which were shown in the article as graphs and tables. The average error was 2.78% on weekdays, 2.79% on weekends, and 4.28% on special days. In contrast, the article contributed to combining three machine learning models, improved forecast accuracy using the model's error as a feature and employed a genetic algorithm for optimal feature selection. Additionally, it examined the impact of weather, weekdays, weekends, and special days on energy consumption.

Demirel et al. (2022)[5] suggested a novel ML-based framework for energy distribution companies to improve their business and services to their customers. This framework uses the customer's historical consumption data to predict energy use. It includes different unsupervised and supervised methods such as isolation forest (iForest) for detecting and removing anomalies, a K-means clustering algorithm for clustering, and finally a long short-term memory (LSTM) network for energy consumption. It also includes a combination of convolutional neural networks (CNN) and LSTM (referred to as CNN-LSTM) for predicting energy consumption. It also used a time series dataset of energy consumption that was collected from 370 consumers in Portugal between 2011 and 2014. After dividing the data into three categories, namely training, validation, and test, containing 85%, 10%, and 5% respectively of the entire time series dataset, the results showed that LSTM achieved better outcomes in predicting 1-hour-ahead energy consumption and CNN-LSTM performed better for 24-hour-ahead forecasting

González-Briones et al. (2019)[6] proposed an article that reviews some of the main machine-learning models that are used in the prediction of energy consumption. As part of the DREAM-GO and EcoCASA research projects, this study assessed whether the recommendations to reduce energy consumption were effective or not. Therefore, the study evaluated the outcomes produced by models such as K-Nearest Neighbors, Linear Regression (LR), Random Forests (RF), Support Vector Regression (SVR), and Decision Tree and showed which had a higher accuracy in energy consumption. To do so, the data they used for the models was from a shoe store located in Salamanca, Spain, in the range of 05/01/2016 and 11/12/2018. After the evaluation, they found that LR and SVR had the highest accuracy of 85.7%, and RF had the lowest prediction accuracy of 79.9%.

In the study by Alzoubi (2022)[7], a novel strategy is proposed to enhance energy efficiency and cost reduction in smart homes. In order to create a distributed intelligent solution, the method uses self-contained agents with semantic interoperability and protocol interaction. These agents make decisions and communicate with other agents to optimize energy usage and production using data from sensors. The study incorporates AI techniques for energy demand forecasting, optimization, and scheduling to solve shortcomings in current smart home systems, such as intelligence, interoperability, and flexibility. The design creates a comprehensive home energy management system (HEMS-IoT) that takes advantage of AI, IoT, and cloud technologies to provide intelligent load control. It has layers for presentation, IoT services, security, and data. The paper emphasizes accurate energy consumption forecasts with a 92% prediction accuracy, particularly taking into account weather and user behavior. Scalability, security, and agent coordination are issues even though they demonstrate efficiency and cost-saving prowess. The study advances energy optimization in smart homes and offers opportunities for further investigation in energy forecasting, smart grid integration, and AI-driven optimization strategies. It also illustrates potential cost savings and environmental sustainability through AI and IoT integration within a strong HEMS framework.

In this research, Anh-Duc et al. (2020)[8] introduce an effective approach for short-term energy consumption prediction across multiple buildings, utilizing machine

learning. Using measures like MAE, MAPE, RMSE, and SI, the study compares Random Forests (RF) with M5P and Random Tree (RT) models as a predictive model. The building data genome project's year-long datasets of hourly energy usage are used by the authors to assess the performance of the RF model. Results consistently show that RF outperforms the M5P and RT models in terms of prediction accuracy. The finding is significant because it gives owners and managers a strong tool for improving a building's energy efficiency. However, restrictions such the absence of metadata and seasonality effects and preset RF parameters are accepted. Despite them, the study creates a viable framework for furthering machine learning-based energy-efficient methods. In summary, this study provides a solid basis for improving energy efficiency through precise energy consumption projections made by Random Forests.

In this research, Zekić-Sušac et al. (2021)[9] has addressed energy efficiency in public buildings within the smart city paradigm. In addition to highlighting how underutilized machine learning in this situation, it also introduces the MERIDA framework which combines Big Data, machine learning and IoT for energy management. Based on data from Croatia's energy management information system, this study uses machine learning methods to forecast energy usage, including deep neural networks, Rpart regression trees, and random forest models. The Random Forest algorithm produced the best results. With its six-stage structure, MERIDA supports decision-making by providing energy rules and retrofitting advice, encouraging energy efficiency and cost-cutting. Ontology modeling enables data collection from IoT sensors and automated control of energy devices, enabling IoT integration. The study recognizes problems such a lack of data and addresses them by working with data suppliers and investigating other techniques. Future potential include expanding models to include data on CO2 emissions, water use, and renewable energy sources. In conclusion, the article advances talks on energy efficiency by providing a thorough framework that combines machine learning and IoT for sustainable urban planning.

In their study, García-Martín et al. (2019)[10] addressed the central theme of energy estimation in machine learning within a research article. The authors explored diverse methods for quantifying the energy consumption of machine learning algorithms. The study delves into two specific use cases: data stream mining and Convolutional Neural Network (ConvNet) inference. These use cases vividly demonstrate how energy estimation techniques are applied and the subsequent impact on accuracy and energy usage. Furthermore, the research emphasizes the imperative for continued investigation to adapt energy estimation methods to evolving neural network topologies and hardware configurations. Challenges emerge from the increasing use of application-specific hardware and GPU-based energy estimation. The paper underscores the significance of integrating energy usage considerations during the design phase of machine learning systems. It highlights the limitations of existing energy estimation models and advocates for more comprehensive and adaptable approaches. Ultimately, the study concludes by recommending that future research concentrate on developing novel hardware designs and enhancing GPU-based energy estimation techniques. In summary, this research furnishes valuable insights into the integration of energy consumption considerations into the design of machine learning systems.

In their investigation, Liu et al. (2020)[11] explored the potential of implementing Deep Reinforcement Learning (DRL) methods for the prediction of building energy usage, a crucial element in managing building energy resources. While data-driven methodologies have garnered attention for their effectiveness, the integration of deep learning and reinforcement learning, referred to as DRL, has been relatively underexplored in this field. To fill this research gap, the study employed three distinct DRL models—Asynchronous Advantage Actor-Critic (A3C), Deep Deterministic Policy Gradient (DDPG), and Recurrent Deterministic Policy Gradient (RDPG)—to forecast energy consumption within an office building. Furthermore, the research conducted a comparative analysis with three supervised models. Notably, the findings underscored the superior performance of DDPG in both single-step and multi-step predictions, whereas RDPG demonstrated particular strength in multi-step scenarios. However, A3C struggles in all prediction scenarios. Notably, DRL models require more training time compared to supervised models. The study underscores DRL's potential in energy prediction and suggests its extension to other fields. This fills a research gap and contributes valuable insights to the field of building energy consumption forecasting. The paper concludes with a call for further exploration in medium and long-term predictions.

Culaba at al. (2020)[12] proposed a research that addresses energy efficiency in mixed-use buildings using a machine learning model. The goal is to predict energy consumption accurately. The study uses regression techniques, specifically supporting vector machines, to develop this predictive model. The dataset is available publicly and consists of simulated energy consumption data from 30 mixed-use buildings in Miami, Florida. The research integrates clustering with regression to improve accuracy, and the proposed model outperforms previous methods. The study's limitation is that it's specific to buildings in a tropical climate zone. Overall, it's a valuable advancement in energy consumption prediction for mixed-use buildings.

In this paper, Tsanas et al. (2012)[13] investigated building energy consumption using historical and simulated data that includes a variety of criteria such building attributes, environmental factors, weather information, and occupancy information. It summarizes the results of previous research that use regression-based predictive modelling to estimate building energy use. The report does not explicitly evaluate data accuracy but acknowledges differences in prediction accuracy across different models, features, and datasets, even though some of the datasets utilised in these experiments are publicly accessible. The paper's focus is restricted to the examined research and their methodology, which include keyword-based searches, data exploration, statistical analysis, and the use of random forests, and there are no disclosed conflicts of interest. It identifies gaps in the literature, emphasizing the need for more multivariate analysis and attention to data types, sizes, and feature selection in energy consumption prediction studies, thereby guiding future research directions.

In this research, Amasyali et al. (2018)[14] delves into data-driven methodologies for predicting building energy consumption, crucial for energy conservation. It examines variables like building types, prediction timeframes, energy categories, and data characteristics. It highlights complexities in energy prediction due to architectural, environmental, and occupant factors. The study contrasts physical models with data-

driven models, favoring the latter for its practicality despite limitations. The paper conducts a comprehensive review of machine learning algorithms like Support Vector Machines and Neural Networks. It scrutinizes building types, prediction timeframes, data types, features, and dataset sizes. This leads to discussions on limitations and future directions. In conclusion, the paper not only analyzes data-driven energy prediction but also spotlights unexplored areas. It suggests exploring big data analytics and occupant behavior understanding. The hybridization of data-driven and physics-based models is endorsed for improved accuracy. This analysis aligns to enhance energy efficiency and reduce carbon emissions.

In this research, Shapi et al. (2021)[15] aimed to develop an energy consumption prediction model for two commercial departments in an intelligent building environment.Microsoft Azure Machine Learning studio (AzureML) was used to analyze and pre-process data from June to December 2018. Three supervised machine learning methods were compared (k-Nearest Neighbor, Support Vector Machine with Radial Basis Function kernel, and Artificial Neural Network with Multilayer Perceptron model), with Support Vector Machine (SVM) yielding the most promising results in terms of lower RMSE and MAPE for specific tenants. Because AzureML lacked the necessary algorithm, cloud-based predictive model creation required R programming. Future research recommendations include adopting more potent platforms to shorten the time required for SVM training, gathering more pertinent data for increased accuracy, considering hybrid or ensemble approaches, and comparing outcomes with those of other smart buildings. The study establishes a basis for forecasting future energy usage and emphasizes SVM's potential as a successful predictive model.

In this research , Moon et al. (2018)[16] intended to power consumption forecasting models for higher education institutions, notably university campuses. The primary goal is to maintain stable and dependable electric power supply by precisely anticipating electric demand, which changes according to time, environment, and consumption patterns. Two machine learning models, Artificial Neural Network (ANN) and Support Vector Regression (SVR), are used to anticipate electricity usage. The study gathers and preprocesses data on electricity usage from four building clusters at a university, taking into account factors such as time, weather, events, work hours, and class schedules. Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are used to evaluate the models.The findings show that the ANN-based model outperforms SVR with an average error rate of 3.46-10%. Including work hours and class timetables increases accuracy, especially in educational facilities. This research offers a helpful technique to creating more detailed and precise power consumption projections, which will aid in effective power management on university campuses. Future studies might look at further improvements to the models and their applicability in different contexts, such as improved algorithms, real-time changes, and interaction with smart grid systems, to further maximize energy efficiency.

In this research , Khan et al. (2020)[17] addresses the issues faced by integrating renewable and nonrenewable energy sources into electric networks through precise energy consumption forecasts utilizing soft-computing approaches.The major goal is to create a viable forecasting model for energy consumption using historical load data

and a unique hybrid technique that combines multilayer perceptron (MLP), support vector regression (SVR), and CatBoost algorithms. The study focuses on the trends in energy consumption from renewable and nonrenewable sources. The suggested hybrid model is compared to different prediction approaches, and its performance is measured with a variety of criteria. Jeju Island is the subject of the case study, where attempts are being made to switch to renewable energy sources. The examination of the data, the hybrid forecasting model, and its comparison to other techniques are the paper's contributions. The hybrid model may need to be improved in the future, as well as real-time data integration and improving the model's capacity to react to shifting energy landscapes.

In this research, Touzani et al. (2018)[18] aimed to work on the use of complex statistical learning models, notably the Gradient Boosting Machine (GBM), to estimate correct savings in energy efficiency initiatives. The increasing availability of high-frequency interval data from commercial buildings' advanced metering infrastructure (AMI) permits the use of such models for accurate baseline energy consumption projections.The study presented and validated a GBM-based energy consumption baseline modeling approach on 410 commercial buildings. GBM beat conventional models such as Time-of-Week-and-Temperature (TOWT) and Random Forest (RF) by more than 80% in R-squared prediction accuracy and CV(RMSE).The versatility of the GBM model in dealing with a large number of input factors, convenience of incorporating additional variables, and the ability to retain accuracy with shorter training periods are highlighted. GBM's accuracy and scalability benefits make it a viable technique. Future research might look at using the GBM model to forecast near-future energy use, detect ongoing anomalies, and reduce demand-responsive loads.

The energy information system (EIS) has become vital for this purpose because it aims to optimise energy use and achieve energy efficiency with real-time data collected, analyzed and transmitted in order to improve users' ability to manage their own usage (Chou et al., 2019)[19]. A major component enables smart metering, which gives users real-time detailed consumption data they can use to adjust their usage patterns. More recently, hybrid models combining machine learning and SARIMA algorithms have been promoted as a method to increase forecasting precision in the presence of both linear and nonlinear trend consumption. Regression-based methods in anomaly detection tend to help detect anomalies from unusual consumption patterns and provide real-time alerts to reduce energy waste. These platforms also provide web-based interfaces that make energy use monitoring and responding to anomalies not only possible but also user-friendly. This leads to proactive energy saving behaviour, although user engagement remains a challenge adapting to evolving consumption patterns is challenging.

Prepaid and postpaid metering systems have significantly different energy consumption patterns; prepaid systems promote conservative use because of the payment prior to use and postpaid systems exhibit frequent patterns stemming from delayed bill cycles (Huber et al., 2018)[20]. Rules based approaches have been used in thas paper to detect anomalies such as energy theft or meter malfunction. Advanced machine learning algorithm such as X means clustering and Local Outlier Factor

(LOF) have emerged for smart meter data, as they are good at detecting irregular patterns (Huber et al., 2018)[20]. N-BEATS is a deep learning model aimed at forecasting processes of time series to work with non linear data without extensive preprocessing. It has shown to be highly applicable to capture the complex consumption patterns and detect anomalies both in prepaid and postpaid systems. The integration of N-BEATS improves energy demand forecasting and anomaly detection, and subsequently facilitates better energy management and system reliability.

Recent advances in energy information systems (EIS) have enabled better monitoring and management of energy consumption in residential buildings resulting in increased efficiency and a decrease in energy waste (Grewal et al., 2019)[21]. Real time data collection and predictive analytics are combined to provide a better understanding of consumption patterns and finding anomalies. A core of such a system is smart meters, that deliver detail bytes that enable energy forecasting and unusual consumption behaviour detection. Further advance in prediction accuracy of these systems has been achieved by integrating machine learning models like SARIMA plus optimisation techniques with increased capability to adjust consumption patterns. These insights are facilitated by web based platforms that allow homeowners to interact with these insights and influence adjustments in their energy use.

Ozkan, in his research, analysed the customers behaviour in prepaid and postpaid metering systemsand showed that load patterns are significantly different between the two systems. Prepaid systems push towards more conservative load patterns, while postpaid systems suffer from unusual load patterns caused by delayed feedback on consumption (Ozkan, 2016)[22]. Integrated smart appliances in Home Energy Management Systems (HEMS) can optimally control device operation and achieve considerable energy savings and cost reductions. Although machine learning methods have been used to detect anomalies in the energy data, he typically relied on rule-based systems due to their simplicity. Time series forecasting of energy consumption has been achieved with deep learning models like N-BEATS in analytics that can capture short- and long term trends. Since nonlinear data is being processed, it makes N-BEATS an ideal candidate for anomaly detection in metering systems, detecting consumption spikes, and also spotting fraud. N-BEATS can be used for noticing abrupt changes in usage in prepaid systems and better forecasting consumption to help manage peak demand consumption in postpaid systems.

Bridging smart metering systems with machine learning models has taken a significant step in terms of energy consumption prediction and anomalies' identification. Smart meters offer the actual use of data, which is vital for conserving energy use, and detecting such anomalous situations as energy theft or faults in the equipment (Haq et al., 2020)[23]. Employing this data, machine learning models enhance the prediction accuracy in traditional prepaid systems as well as new postpaid systems, thus enhancing the energy management and load shedding. Beside, application of clustering and classification techniques for energy consumption has proved to be very efficient due to the hybrid nature of the ML models. The study by proposed a model which includes faster k-medoids clustering, SVM and ANN for the prediction of household energy consumption with an efficiency of 99.2%. The proposed approach outperforms the baselines in terms of appliance-specific power consumption and

peak demand, which are instrumental in energy load control. Anomaly detection is very important in order to maintain the stability of the grid and to avoid energy fraud. Since the predicted consumptions are to be compared with the actual use measurements continuously in real time, machine learning like the one proposed bycan be applied. These systems are very useful for prepaid metering so that consumers can be prevented from running out of credit without any intimation and for a postpaid system, accurate billing.

Building occupiers directly impact the energy use through the control and regulation of HVAC systems, lighting, and appliances, accounting for more than one-third of the overall buildings' energy consumption (Chen et al., 2021)[24]. Occupant behavior can be divided into three main categories: use of space, people contacts, and how they perform their tasks, all of which influence energy consumption. This warrants the integration of timely behavior data in energy models for better prognosis. Different studies show that the utilization of smart control systems present many possibilities for reductions in energy consumption. For instance, occupancy-based HVAC control can save up to 28% of energy, whereas automatic lights control can save as much as 43%. In addition, machine learning algorithms, originally built for anomaly detection, can help identify irregular consumption patterns, perhaps suggesting that systems are installed incorrectly or people are not using resources efficiently. With the focus on enhancing behavioral efficiency, including occupant education, further energy saving of up to 30% can be realized. The author also stresses the need to employ finer-grained approaches to classify occupant behavior because the current models can be quite simplistic about the range of activities people can perform.

As the title suggests, Oreshkin et al. (2021)[25] explored using the N-BEATS neural network for mid-term electricity load forecasting. The N-BEATS neural network model was used because it has a great feature to resolve time-series forecasting, especially on electricity load. It was chosen for its non- linear regression function that is able to handle trend and seasonality factors in the data without prior data transformation. The specifics of the mid-term electricity load forecasting made this research one of the primary objectives important for enhancing operational efficiency in power systems and particularly in deregulated markets and energy contracts and planning. This paper argues that accurate forecasting minimises the financial risks involved and enhances the decision-making process in energy management. For future work, the research suggests exploring the use of interpretable N-BEATS models that break forecasts into understandable components like trends and seasonality. Another area of future focus is expanding the application of this model to other domains or energy systems, potentially enhancing interpretability and scalability.

In this research, El-Hadad et al. (2022)[26] have discussed Anomaly prediction in electricity consumption is done by detecting abnormally used patterns. A blend of such algorithms as the Isolation Forest is then applied in classifying the data as either normal or abnormal. After that, the so-called supervised models like Random Forest (RF) and Decision Tree (DT) are employed to predict the anomalies 30 minutes in advance by using the labelled data sequences. The models predict whether the future consumption will be abnormal, helping prevent power issues and optimise grid management. The main goal is to identify abnormal power consumption in advance,

allowing users and grid operators to act promptly and avoid unnecessary energy waste, power surges, or faults. Possible directions for future work include better handling of data imbalances, extension of the model to other datasets, and employing more powerful options such as deep learning for improved and more efficient detecting anomalies.

In this work, Mahapatra et al. (2018),[27] addressed that the adoption of prepaid electricity meters depends upon the significance of the resultant change in satisfaction because it provides more control to the users, and increases the understanding of bills. The research does this by assessing customers' perceptions of this change in Pune, India. A linear regression model was used to analyze survey data from 300 respondents. They learned that such aspects as consumption tracking, having choices to recharge, convenience of making payments online and timely billing impact it immensely. The main goal of prepaid meters is to enhance collection efficiency for utility companies, reduce operational costs, and encourage energy-saving behaviors. From the consumer side, prepaid systems allow users to buy electricity as needed, monitor usage, and avoid surprise bills, ultimately leading to better financial management. The results showed that users prefer prepaid systems over postpaid due to these benefits.

In their study, Kurniawan D. Irianto (2023)[28], the research focuses on the one hand consumer perception and satisfaction when moving from postpaid to prepaid electricity consumption. The work underscores the benefits of prepaid electricity, including metering of energy usage and costs, and singleton issues like monitoring of credit balance. The new Prepaid Smart Energy Meter Monitoring System (Pre-SEMMS) was developed out of IoT to enhance the prepaid model without changing standard electric meters. Hence, users can warrant they control their electricity usage by installing sensors including cloud platforms such as Blynk.

This is the primary rationale for developing the main functionality, aimed at unburdening consumer's control over electricity consumption and avoiding outages caused by a lack of monitoring of credit depletion. The system's feasibility is validated through household appliance testing and therefore its efficiency in measuring energy. Future work includes enhancing the system's capabilities to cover more devices, applying machine learning for energy optimisation, and expanding usage to larger buildings. It also recommends improving user interfaces for better accessibility.

2.2 Research Methodology

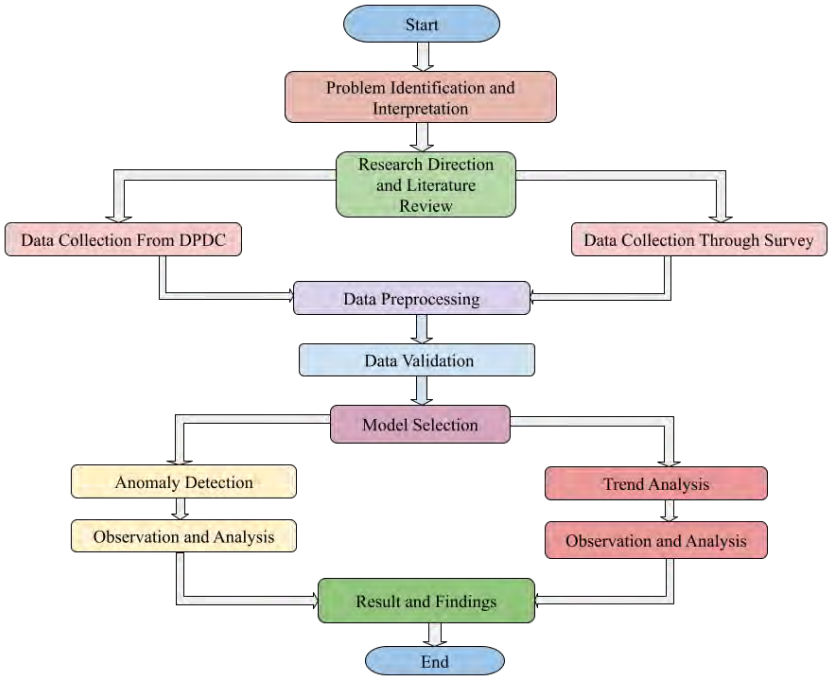

Figure 2.1: Pipeline of methodology

2.2.1 Problem Identification and Interpretation

Problem identification is about determining exactly what problem we need to solve, the gap to overcome or the challenge our research will be considering. The first step, i.e., is where we identify the main problem, either from an incognita in the existing body of data or from a gap in a specific domain or from a trend that needs additional understanding. We often found earlier studies where some questions were left unanswered, contradictions were found, or a new perspective was needed. If the problem is apparent there is a solid foundation for our research, and we know exactly what we are looking for and where to look.

Once the problem is identified, it's time to interpret it; that is, break it down so the understanding of it improves. The scope of the problem, causes, and potential impacts of the problem is problem interpretation. It helps us to narrow down specific research questions and or hypotheses. In other words, we interpret the problem, define what the problem is, and ensure it's well defined by saying which parts of the problem we are going to concentrate on by determining who or what is affected and how the problem fits in the larger context of existing knowledge. A good interpretation ensures that our problem is reasonable and brings us closer to the research goals.

2.2.2 Research Direction

The research direction is basically about setting a clear path on how we will achieve the study. It includes working out specific goals, creating research questions, and developing some hypotheses to test if necessary. It's to keep our research narrowly focused without getting too broad or vague. Our direction decides what methods to use, what type of data to collect, and what we have to do with the results, but along the way, the overall flow of our study is formed.

It is a literature review – a deep dive into the work out there that is related to our topic. It allows us to know areas to fill with our study and a way to refine our research question. Secondly, the literature review also offers a foundation for our research since it helps us link our research to existing shows and theories. It ensures that we are not rehashing stuff that's already been covered and that what we are coming up with fits not just the current mood of the times but also the broader terrain of what's already known.

2.2.3 Data Collection From Dhaka Power Distribution Company Ltd (DPDC)

Dhaka Power Distribution Company Ltd (DPDC) is a power provider which is engaged in the distribution of electricity to the capital city of Dhaka and its surrounding areas. Located under the Ministry of Power, Energy, and Mineral Resources, it has a vital part to play in ensuring that homes, industries and businesses get a regular, dependable power supply. DPDC operates prepaid and postpaid metering, where customers pay for electricity in advance (prepaid) or after the fact (postpaid). It also services billing, payments, new connections, and the upkeep of electrical infrastructure. DPDC also keeps the power grid together, substations, and transformers, to ensure that we have a stable supply through security, and this facility can solve any rapidly resolved problems. DPDC is an important part of the Bangladesh power sector by modernizing the energy distribution system and improving the efficiency of electricity management and billing in the country.

Collecting data from Dhaka Power Distribution Company Ltd (DPDC) was a tedious and time-consuming process. DPDC was initially reluctant to release the data, particularly the records from prepaid and postpaid metering systems. After many attempts, we finally got the data and managed to negotiate them only under strict conditions. As a thesis group we agreed that we had to make sure that we were fully responsible for keeping the dataset confidential. If any data from our research leaked or were shared in some way outside of our research we would be held liable. Getting access to the data was essential, this was part of our research, and without it we would not have been able to complete or continue with our work. We are able to analyze consumption patterns and detect abnormal behavior in prepaid and postpaid systems. It also provides a starting point in applying predictive models to explain why there are differences in billing between the two systems.

2.2.4 Data Collection Through Survey

For our thesis, we obtained postpaid and prepaid consumption data from DPDC . The post-paid metering system in use in Bangladesh is already functioning, however, the prepaid system is used only in some areas of Dhaka such as Paribagh and Dhanmondi. We knew clearly about prepaid and postpaid data given to respective areas of Paribagh so we targeted our survey in this area. In Dhaka, the prepaid metering system is managed by two organizations: DPDC and DESCO. About 56 percent of customers use prepaid meters in the northern and northeastern areas—Gulshan, Mirpur, and Uttara—served by DESCO.

On the other hand, about 30 percent of consumers in the western and southwestern parts of the city, such as Paribagh, Dhanmondi, Mohammadpur and Lalbagh, use pre paid meters and DPDC manages the areas of Paribagh, Dhanmondi, Mohammadpur, Lalbagh (Bangladesh Power Development Board, n.d.)[29]. Meanwhile both companies are trying to expand prepaid meter use amid government efforts to rollout smart meters countrywide. One of the goals of this initiative is to improve billing accuracy, reduce losses and give consumers more control over when they use their electricity (Dhaka Tribune, 2022)[30], (The Business Standard, 2021)[31], (BPDB, n.d.)[29]

2.2.5 Data Preprocessing

In the second part, after collecting data from DPDC and doing customer surveys, data preprocessing was one of the most critical steps needed to be taken to process the data for a correct analysis. First, we started with data cleaning of missing values, duplicates and outliers because these could skew the results. It also made sure there were no errors on the DPDC metering data, or the survey data which in turn, reduced the chance of bias. We then integrated the DPDC data with the survey results, to integrate data, so we could link appliance usage patterns with electricity consumption and explore how varying appliances impacted on energy use. Second, we did data transformation, where we standardized units of measure, e.g. where types of appliances could be given as categories. Still, we needed to standardize them for the sake of consistency. We also applied feature engineering to create additional variables, for example: total power consumption per appliance and average consumption per device, to gain a better understanding on how appliances affect the total electricity use. Finally, we performed normalization and scaling to bring all variables to a more or less similar scale so they would not dominate the analysis. It was crucial in training machine learning models such as N-BEATS to perform anomaly detection and consumption forecasting. By processing our data to the best of our ability with the entire preprocessing workflow we made sure our data was reliable, accurate, and ready for in depth analysis, which helped elevate the quality of the final results.

2.2.6 Data Validation

Data validation is one of the critical steps of the research process since it needs to be sure the data it is working with is accurate, reliable, and capable of giving meaningful parts of data. At first, the goal is to make sure that the models or

statistical methods that are used on the data, are appropriate for conducting the research. In my thesis, we used three main validation techniques: It was done using the Kolmogorov-Smirnov test, and Pearson correlation.

The Kolmogorov-Smirnov test verifies whether data agree with the given distribution so that certain statistical methods are valid. Pearson correlation tells us how strongly two variables are related, which helps us to understand the relation between electricity consumption with factors such as appliance usage. These techniques were crucial to the validity and reliability of the analysis in my thesis.

2.2.7 Model Selection

The primary goal of this research was to examine how people use electricity in prepaid versus postpaid metering systems and detect any unusual consumption patterns. For this, we applied some advanced tooling to predict energy usage and detect anomalies. Particularly, we used the N-BEATS model to predict future consumption trends from past data. We primarily wanted to see what energy consumption was like for prepaid and postpaid users and identify any outliers that might suggest examples like higher bills or unexplained usage.

2.2.8 N-BEATS

Since we wanted to predict time series such as energy consumption, we decided to use the N-BEATS model as it is known to deal with complex data very well. Different from many other models, once we have customized N-BEATS a bit, it does not need a huge amount of manual, disjointed customization; in particular, it works well with a variety of types of data without the need for a lot of distinctive feature creation (Oreshkin et al., 2019).[25]

Historical consumption data of prepaid and postpaid users are used to train the N-BEATS model. The model analyzed this past data to forecast how the energy used would appear. We used sliding window technique to improve the accuracy and prevent the model from overfitting, i.e., it can be overfit if the model is too specialized for the training data. Such an approach helped us think about the order of the data, leading us to more reliable predictions for both short-term and long-term trends. In essence, it enabled us to track the amount of energy used at different times and to predict better what that energy use will be at different times.

2.2.9 Isolation Forest

We predicted future energy use and saw strange patterns in the data that deserved further investigation. To detect these anomalies or irregularities, we applied the Isolation Forest model. This specific type of algorithm aims to identify outliers in huge datasets. It is performed by shuffling out the outliers in the data through a series of random splits in the data to allow easier identification of patterns that do not fit the norm.

Beyond Isolation Forest, we also deployed a rolling average procedure. This smoothed out the tiny fluctuations in the data, making it easier to see the more prominent trends.

The Isolation Forest could more accurately identify natural anomalies (instances in which energy use significantly diverged from established patterns). These anomalies could indicate faulty meters or even unauthorized electricity use.

2.2.10 Anomaly Detection

This research presents itself as anomaly detection, where it seeks to identify unusual patterns of energy consumption beyond what would be expected. This also aids in catching problems such as bad meters, or potential fraud. These odd data points are removed from the analysis, allowing the real consumption trends to focus on the predictions.

The Isolation Forest model is pretty good at identifying outliers in large and complex datasets. It divides the data into smaller groups using many decision trees. The number of steps for isolating an anomaly is much less than that for the regular data points. This study is a good match for the model because of its ability to handle large datasets and its flexibility— it doesn't assume the data follows a specific pattern.

The Isolation Forest model was applied to both prepaid and postpaid data and scored each data point for how likely it was to be an anomaly. This allowed us to remove any unusual data so the prediction models would run with clean data.

Observation and Analysis of Anomaly Detection

An observation and analysis phase is performed to understand the results of the anomaly detection process. In this step, we explore unusual patterns in energy use using the Isolation Forest model, such as a sudden spike or drop. We look at each anomaly to see if it represents real problems, like faulty meters or mistaken billing, or something else.

This analysis is done across both the prepaid and postpaid data and enables us to compare the behaviour between these two groups. That way, we'll be able to see how anomalies change the overall use of energy. These insights help to make the forecasting models more accurate and how energy data is managed more efficiently.

The analysis also scans patterns amongst anomalies, searching for any repeated issues or trends. For example, a few days that appear consistently anomalous could point out a repeated problem or a definite user behavior pattern. By understanding these trends, predictions can be refined, and the metering system associated with the consumption can be improved to work better with anomalous changes in consumption.

2.2.11 Trend Analysis

At the trend analysis phase, it finds patterns of energy consumption over time by looking at long-term trends and seasonal changes for both prepaid and postpaid data. This helps us anticipate future consumption and handle energy better.

The N-BEATS model is used for trend analysis because it can learn complex time series patterns directly from data. One of the best is because it handles recent

changes as well as long term trends without needing predefined features, which is good for capturing extremely unique consumption patterns.

In this study, we study past data by applying the N-BEATS model to predict future energy use. These forecasts aid growth, decline, and seasonal changes that clarify usage trends. This insight is used as a baseline for comparing accurate data, enabling us to detect unexpected changes more readily.

Observation and Analysis of Trend Analysis

In the observation and analysis phase, through analysis, we understand what the N-BEATS model foresaw. This allows us to verify that the model captures patterns in energy use over time. Once we have the predicted values, we can compare them to the actual consumption data we have and judge how accurate the forecasts are and how we can improve upon them.

In this phase, using the N-BEATS model, we examine how well the predicted trends align with actual data before comparing them to any changes or unexpected shifts from the existing data. That's why it is also meant to examine short—and long-term patterns to ensure that the prediction matches actual behavior.

Postpaid and prepaid data are both included in the analysis to offer a full picture of how energy is used across all account types. Finding seasonal peaks and sudden changes with this process will give us a nice margin for improving the model and making future predictions more accurate. This step's insights ensure that the model is sensitive to the changes in energy use.

2.2.12 Result and Findings

The final phase in the MI is Result and Findings, which combines the anomaly detection and trend analysis insights. This step aims to provide a complete picture of how the energy is used based on unusual patterns and overall trends.

Through anomaly detection from the isolation forest model, we identified unusual spikes or drops in energy use. Such problems could indicate faulty meters or possibly fraud. These irregularities tell us where to go next when we're trying to understand what's happening.

Combined, these insights allow a better understanding of what's normal and what's not in energy consumption. With this complete view, we can proceed with improved future models and make better decisions about reducing energy use.

2.3 Anomaly detection

Identifying the abnormal data is called Anomaly. It is a strange transaction that may indicate fraud in fraud detection. Sometimes, something is wrong, or an attack catches unusual behavior, which is monitored in systems. Three types are used to categorize significant expected patterns from anomaly detection. Point anomalies occur when

a single data point is compared to others. Contextual anomalies refer to data points that are normal in one context but abnormal in another (e.g., high electricity consumption during nighttime). When a group of data points, though individually standard behavior patterns, collectively indicate unusual behavior patterns that occur as Collective anomalies (Chandola et al., 2009) [32]. In the context of the workflow of anomaly detection for electricity consumption, the approach aims to identify abnormal usage before it occurs.

As outlined in our paper, the process begins with data preparation. Consumption data collection from smart meters at 30-minute intervals loaded and cleaned the data and removed outliers to ensure consistency. Next, the Isolation Forest (IF) algorithm performs anomaly detection, which assigns an anomaly score to each case. The instance is labeled as abnormal If the score is less than 0.5; otherwise, it is labeled as average. After labeling, the data is transformed into sequences, enabling a prediction model to work on it. For example, a sequence like "NAAANN" is proven to predict whether the next value will be normal or abnormal. It is easier to detect several abnormalities in consumption, which helps facilitate efficient management of energy providers and consumers.

Energy efficiency, identifying fraud, and ensuring operational safety is essential for maintaining anomaly detection in prepaid and postpaid meter systems. Both meters collect one year of data, which is used to monitor for unusual patterns. In prepaid meter systems, the consumer pays the bill in advance, and that's the reason and their consumption behavior tends to be more deliberate. In contrast, in postpaid meter systems, the bill comes at the end of each month, leading to irregular usage until the bill arrives. Therefore, anomaly detection ensures early detection of irregularities in both systems, aiding consumers and energy providers in taking timely action. The variability of electricity consumption at different times of the year requires the identification of abnormal trends. In the case of prepaid meters, unconventionality, such as some increases in electricity consumption patterns, could result from theft or interference with the meter.

Similarly, if consumption drops sharply without a corresponding behaviour change (low consumption) significantly different from the behaviour (e.g., a vacant household), it may indicate manipulation of the meter. Sometimes, the device malfunctions, which causes a lot of power consumption. Because of this, postpaid meter users sometimes face the shock of unexpected billing. For this, we can now specify a flat early, which can detect anomalies and irregularities so that users can investigate and solve problems before the end of the billing cycle. Such pre-emptive detection enables early consumer notification and checks large-scale problems such as overloaded grids. It also minimizes unnavigable financial and operational implications.

Anomaly detection is mandatory in electricity systems because of efficiency, safety, and cost control. It helps identify inconsistent power consumption caused by power theft, faulty appliances, or defective equipment, leading to grid overloads, safety hazards, and financial losses. In prepaid metering, irregularities in the use of energy, such as high consumption rates, ensure that users have enough credits. In contrast, in postpaid metering, irregularities in consumption rates guarantee their users never

face a bill shock by providing real-time feedback. This is a very effective way of preventing outages as it is straightforward for a small problem to become a big problem on a power grid if not noticed early. It also assists utilities in fighting power theft and keeps customers happy by avoiding surprises in their bills. The study by (El-Hadad et al., 2022)[26] highlights using machine learning models like Isolation Forests to predict abnormal energy usage, ensuring sustainable energy management proactively. Thus, anomaly detection avoids risks and optimizes energy consumption, which is essential for consumers and providers.

2.4 Isolation Forest for Anomaly Detection

Identifying the abnormal data is called Anomaly. The Isolation Forest model is one of the easiest ways to work with the dataset. Firstly, It is divided into smaller groups. Most data point anomalies are rare and different because they focus on the fact that anomalies are unlike other models, making them easier to separate. The model picks a random feature (like a column in a dataset), splits the data, and chooses a random point (Liu et al., 2008) [33]. Anomaly data is quite different from average data. They are separated into fewer partitions and are captured by shorter paths in the decision tree. Each observation is given an anomaly score, determining how quickly it is isolated. If the score is higher, there is a higher chance of anomaly.

In many areas, the Isolation Forest model is used for energy consumption. It also helps to increase electricity usage suddenly, solve equipment problems or improper use, and find unusual patterns. It is a strange transaction that may indicate fraud in fraud detection. Sometimes, something is wrong, or an attack catches unusual behavior, which is monitored in systems. It doesn't need labeled datasets and also works well with complex data. Many industries use it as a valuable tool for finding outliers in real time. Example: Daily electricity consumption data will be input into the model.

The model will generate segmentations from different features (hour, date, month) to distinguish each day's consumption. If a day has a much higher power consumption than an average day, the model will quickly distinguish it and mark it as an anomaly.

2.5 Prophet

In our study, we employed the Prophet model to forecast normal electricity consumption in the future accurately. We selected Prophet because it is flexible when tackling fluctuations in usage and energy use patterns by customers before and after converting them to prepaid meters. Prophet is also good at dealing with missing data and sudden changes, which are common in real-world electricity data.

2.5.1 Why Prophet Was a Good Fit:

- **Handling Seasonal Patterns and Trends:** Electricity use often increases in summer due to air conditioning and heating in winter. The Prophet can easily track such patterns and long-term shifts and is suitable for demonstrating

how energy consumption fluctuates throughout the year. This gave us a clear picture of what customers' usage is normal, allowing us to start the analysis from a familiar baseline.

- **Dealing with Irregular Data:** Our data had some gaps and inconsistencies because it was collected during the switch from postpaid to prepaid meters. The Prophet can handle these missing values and sudden spikes without losing much accuracy, allowing us to remain prescriptive when the data is not completely precise.

- **Integration with Isolation Forest for Anomaly Detection:** The Prophet model was implemented primarily to forecast normal electrical consumptions in the present study. With these predictions, we find out the actual and predicted usage (residuals). We then used the isolation forest model to check these differences. The Isolation Forest then sought to determine 'anomalies' where the actual usage was considerably higher or lower than the expected level. This combined method allowed for more accurate identification of unexpected flares or falls in energy consumption while excluding fluctuation due to time of day or day of the week.

- **Clear and Useful Results for Utility Companies:** Our process had to be simplified since we aimed to provide relevant information that would help utility companies, including DPDC. The Prophet breaks down electricity usage into different parts: daily, weekly, monthly, and yearly, occasionally relying on the period, the location and certain occasions. This breakdown made it easier to explain how we came up with our predictions. It also helped us show where the anomalies came from, making it easier for utility companies to understand and act on the information, especially in response to customer complaints about higher bills.

2.5.2 How does the Prophet Model work

Prophet is a time series forecasting model invented in 2017 by a group of scientists from Facebook (Meta). As mentioned, it can address seasonality, trends, and holidays in time series data (Taylor, et al., 2018) [34]. It was designed to be a dependable and easy-to-use means of forecasting business needs – from demand to sales and inventory. Thus, its broad adoption is due to the simplicity in implementation, versatility, and high predictive performance requiring little calibration, at least when tested by Meta Research Blog in 2017. [35]

Model Equation:

The core of the Prophet model is based on an additive decomposition of the time series into the following components (Taylor et al., 2018) [34]:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{2.1}$$

Model Equation:

The core of the Prophet model is based on an additive decomposition of the time series into the following components (Taylor et al., 2018) [34]:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{2.2}$$

where:

- $y(t)$: Observed value at time t.
- $g(t)$: Trend component that effectively measures the data series's general rate of increase or decrease.
  - *Piecewise Linear*: Linear segments with changepoints.
  - *Logistic Growth*: Illustrates an S-curve for data with upper and lower bounds.
- $s(t)$: Periodic component that models daily, weekly, and yearly seasonal patterns using a Fourier series.
- $h(t)$: Holiday component which captures the effect of special events.
- ϵ_t: Residuals that reflect random noise in the data set.

Key Features

- **Automatic Changepoint Detection:** This technique recognizes trends and makes changes in the forecast (Taylor et al., 2018)[34]
- **Flexibility with Missing Data:** Combined with other algorithms, it can handle the missing values without deeper preprocessing (Meta Research Blog, 2017)[35]
- **Robust to Outliers:** Reduces noise while keeping the integrity of the primary trend and seasonal structure (Taylor, Lu, and Wang, 2018)[34]

Prophet is one of the flexible time series models that can be applied when trends and seasonality are critical in business. Thus, the method enables users to obtain precise predictions with a low level of parameter optimization, thus serving as the primary forecasting tool for many real-world requirements (Taylor et al., 2018).[34]

2.6 Deep Learning

Deep learning uses the concept of the human brain; it uses artificial neural networks. This has multiple layers of networks; these layers of connected units are called neurons. Deep learning models can quickly identify complex patterns from pictures, text, and sounds and provide accurate predictions. It processes data inspired by the bases of the human brain and uses AI(artificial neural network) to learn from the provided data. The main structure contains three parts: the input, hidden, and output layers. The input layer takes in data, the hidden layers transform the data, and the output

layer gives results. The first step involves assembling a large amount of data, such as past electricity usage, weather conditions, and economic factors. This data is then preprocessed and formatted to train the developing neural network and deep learning models. During the training period, a deep learning model analyses patterns in the data set using predictions made and actual results produced. The resulting errors to give improved accuracy of the output is called backpropagation. Once trained, the model is used in real-world applications, such as forecasting electricity consumption; it can help manage energy use more efficiently. Another paper shows that Deep learning models are significant in short-term electricity load forecasting (STELF) due to their ability to capture complex nonlinear patterns, making them more effective than traditional methods in predicting dynamic load changes (Dong et al., 2018).[36] The use of deep learning techniques brings benefits for this case: higher accuracy of forecasts, a large amount of dataset work, and pattern recognition that the other methods cannot pick.

2.7 How does N-Beats Work

N-BEATS is an advanced architecture of deep learning used for time series forecasting. It uses a wholly connected neural network to predict past and prospective time series values. First of all, N-BEATS are several blocks, each responsible for filtering the forecast made by the previous block. Each block produces two outputs: the reconstruction of past data to help correct previous errors, which is called backast, and the future values based on the predicted past information, which is called forecast. This is the iterative process where each block focuses on the residual errors left by the previous block; mainly, it works to improve and elaborate on the model's predictions. N-BEATS will capture both long-term and seasonal (daily, monthly) fluctuations and trends in the data and improve accuracy with each block of residual learning method. The task is repeated for multiple blocks to ensure proper accuracy. N-Beats perform better in the long-term pattern than other models. Unlike other models, it does not require domain-specific features, making the work more versatile. For example, in summer, people use more air conditioners, so electricity consumption increases. Here, "season" is a domain-specific feature essential for forecasting specific cases (power consumption). It's also designed to provide interpretable outputs, decomposing the time series into components like trend and seasonality without requiring complex feature engineering. In machine learning frameworks such as PyTorch or TensorFlow, where N-beats can be easily implemented, it makes it accessible for In real-world forecasting applications. N-beats uses energy consumption, stock prices, sales data, and more because of its high accuracy and comprehensive prediction. N-BEATS reduces errors through an optimiser such as Adam (Adaptive Moment Estimation), an algorithm that generally incorporates the Mean Absolute Percentage Error (MAPE) as an objective function. The parameters are adjusted permanently to improve the predictions in subsequent iterations through the training data. Compared to most other models, n-BEATS can work with raw data without preprocessing and demonstrates high accuracy in various tasks, including energy consumption forecasting.

2.8 Prepaid postpaid meter comparison

Bangladesh's electricity supply companies are the DPDC (Dhaka Power Distribution Company) and DESCO (Dhaka Electric Supply Company). These companies use two types of electricity meters: prepaid and postpaid.

Consumers purchase electricity before use in a prepaid metering system. This means that the meter is set to enable electricity consumption when a balance is available. When the balance runs out, the meter automatically cuts off electricity, which helps people manage their budget. In real-time, the prepaid meter given to people how much they spend on their electricity usage boosts them to have more control over their spending. In contrast to the postpaid meter system, the meter records the total units consumed over a month. Customers receive a bill and pay it on a specified monthly date. When the consumer bill is not paid, the company disconnects the electricity afterward. This system gives customers more flexibility but sometimes leads to unexpected costs and unpaid bills if not managed carefully.

In Dhaka, the first prepaid electricity meters were established in Paribagh, where people served the first smart meter. DPDC helps consumers make electricity affordable through intelligent meter systems, making it easy for the masses to pay bills through prepaid meters. This meter helps people to avoid unexpected bills and reduce the chance of getting a loan. However, some people still find a postpaid meter because they pay a monthly bill at the end of the month, which is convenient for them. However, since then, the prepaid meter has become a better option in the eyes of many consumers who directly see how much they spend on their energy costs.

2.9 Sliding window experiment

In the Machine learning model, the sliding window is an experiment commonly used in time series analysis; it can handle sequential data for better signal processing. It is performed step by step over a window of a fixed size over the data set, picking up only an element of the data at a time in each location in the window. This window works on a certain amount of data and then can predict what will happen in the future to work sequentially. The sliding window used in its CSV file works efficiently in Microsoft Excel. It creates sequences (past values of rows as input and output as the next value) and divides sequences into training and test sets.

Chapter 3

Data Collection and Preprocessing

A detailed analysis and descriptions of the data collection from the Dhaka Power Distribution Company (DPDC) and through the survey, procedures of data cleaning, and statistical analysis have been presented in this section.

3.1 Recruitment and Data Collection Procedure

Two distinct data collection methods were followed to conduct this research to ensure a better understanding of energy consumption patterns and their associated anomalies. At first, the direct energy consumption data of the users of the Paribagh, Dhaka region was collected from the energy distribution provider Dhaka Power Distribution Company (DPDC). The region was selected through the analysis of the deployment of prepaid meters in Dhaka. We had several meetings with the DPDC regional officer to discuss and have the proper data that will fit our research work. This dataset includes the prepaid and postpaid data of the users from July 2021 to December 2023. In addition, we have received data from the DPDC for July 2022 daily basis from all the users. This data set includes the customer's category, ID, billing date in the DMY (Date, Month, Year) order, and the unit consumption of that particular date. After getting all the possible data from DPDC, we needed to increase the dataset and get more variety in it to have a clearer research outcome. We received 12 months of postpaid data and 21 months of prepaid data from the DPDC. Postpaid data contains the time range from July 2021 to June 2022.

It is vital to recognise residents' personal experiences with the implemented changes. Consequently, a survey was initiated. In the next phase, the data was obtained from a door-to-door survey in the same Paribag region, where the participants of this study were the residents of that area. The Department of Computer Science and Engineering, BRAC University, has approved the conduct of this survey study and its procedure. All the participants who participated here were asked to give their consent. None of their personal information was used in this research. Instead, the numerical value of their household appliance was collected through the survey form. We have gone through several previous survey samples to have questionnaires with proper standards, structure, and requirements. In addition, we approved the survey from our supervisor, and all the participants participated with their consent. In addition, we talked about our motive behind this survey and what we plan to achieve from this research work. Thus, we convinced the participants in the survey

only if they were willing to participate. We also inform the participants about the rules and regulations they should follow while answering the questionnaire. The rules were that participants may leave the survey at any time if they decide not to proceed. Additionally, they are expected to share their feelings and avoid submitting any inaccurate information honestly.

3.2 Dataset Description

3.2.1 Household and Appliance Information through the Survey

This subsection of the dataset description will discuss the data and its features collected through the door-to-door survey. This feature plays a significant role in understanding household energy consumption.

I. **Number of rooms:** The number of rooms in a household reflects the size of the living space in each household. It plays a vital role in energy consumption because larger households use more energy through many different home appliances. The larger the household, the more energy it will consume. This data will help us understand how much energy a household reasonably consumes.

II. **Number of household members:** A household's energy usage directly correlates with the number of people living there. More family members tend to consume more energy due to increased appliance use, cooking needs, and increased frequency of laundry and air conditioning activities. This data helps to show the variation of consumption patterns among different households.

III. **Major appliances:** This section discusses the various household appliances that a household typically uses. This appliance is directly connected to the energy consumption. Some appliances may consume more energy than others. The following segments delve into the specific aspects of each home appliance.

   i. **Fans:** One of the most common and widely used household appliances that consumes a decent amount of energy for each household.

   ii. **Lights:** Another essential household appliance that directly indicates basic electrical usage. Households with more rooms typically have more lights, while those that adopt energy-saving practices are likely to use energy-efficient lighting solutions.

   iii. **Freezers:** Freezers are generally energy-intensive appliances, and their constant 24/7 operation can contribute substantially to a household's overall energy consumption. Monitoring the number of freezers provides insight into the baseline energy requirements.

   iv. **Microwave ovens:** The frequent use of microwaves for heating and cooking can significantly impact overall energy consumption. Collecting this data enables a more detailed analysis of kitchen appliance usage and its impact on household energy patterns.

v. **Geysers:** A geyser is one of the household appliances that is used in the comparatively colder region. It has a good amount of energy while in d. Its inclusion in the dataset helps identify households that might experience significant seasonal energy-use variations.

vi. **Air conditioners (AC):** Air conditions are one of the significant contributors to energy consumption, particularly in warmer climates. A household's number of AC units offers valuable insights into cooling demand, often resulting in energy spikes during the hotter months.

vii. **Induction stoves:** Induction stoves are among the most used home appliances, replacing traditional gas or electric stoves. They consume a good portion of a household's total energy.

viii. **Washing machines:** Washing machines are standard in households and can contribute to irregular energy spikes depending on their usage frequency. A larger household with more laundry consumes more energy for washing and drying.

ix. **Television (TV):** Television is another most-used home appliance, and almost every family has it at their home. It consumes a decent amount of energy depending on its size and type. It is typically used for several hours daily.

x. **Personal computers (PCs):** Personal computers used for work, school, or fun are a big reason homes use so much energy. More energy is required in households with more PCs depending on their uses.

3.2.2 Energy Consumption Data (Prepaid and Postpaid)

For this research, two types of energy consumption data are used from the energy supply company Dhaka Power Distribution Company (DPDC): prepaid and postpaid data. The DPDC has provided us with 12 months of postpaid related data and 21 months of pre-paid related datasets of Paribag region, Dhaka users.

In prepaid systems, users purchase energy in advance, and their consumption is deducted from the pre-purchased balance. Each recharge transaction is logged with a timestamp and the quantity of units purchased. As users near the exhaustion of their prepaid balance, they are required to recharge to continue service. The data collected from the DPDC includes the 21-month postpaid data of 22347 users from April 2022 to December 2023. Even though the prepaid system allows for real-time energy use tracking, the data was combined into monthly consumption for comparison with postpaid systems. The monthly kilowatt-hours (kWh) consumption reflects the total energy used within a billing period. In the prepaid system, customers may recharge multiple times within a month, each contributing to their total monthly consumption.

On the other hand, postpaid energy consumption data was collected in a similar way from DPDC. However, unlike prepaid systems, postpaid users receive their bill at the end of the month after consuming energy. Each postpaid customer has an energy meter that records the total energy used over a month. We have collected 12 months of postpaid data from DPDC, whose time range is from July 2021 to June 2022. In the postpaid metering system, meters record the energy used throughout a billing

period; for example, the data we have collected is monthly energy consumption for each user in the Paribag region.

3.3 Survey Data Overview and Questionnaire

For this research, a door-to-door survey was conducted to obtain an accurate picture of each participant's energy consumption. A total of 1797 users from the Paribag region participated in the survey. This survey aimed to gather detailed demographic and household-level information that complements the energy consumption data provided by DPDC.

3.3.1 Survey Design and Objectives

This survey was conducted to gather household data that would be used along with the energy consumption information provided by DPDC. By combining this data with the consumption records, we gained more profound insights into the behaviors of prepaid and postpaid users. The survey information was taken through a survey form. The structure was as follows:

- **Customer Information:** This section provides the basic details of the participants. It includes the area name, customer number, meter number, and address. This information incorporated the survey data with energy usage data from DPDC.
- **Household information:** This is the main section of the form that the participant is required to fill out. The main focus of this part was on the household appliance count. This information is essential for analyzing energy use and detecting anomalies in prepaid and postpaid systems.

3.3.2 Questionnaire

The survey form was elementary and straightforward to understand. We explained the necessity of this survey to each participant and answered their questions before providing them with the survey form. This section highlighted the key features of the questionnaire and their significance for the study. The home appliances were fans, lights, freezers, microwave ovens, geysers, air conditioners (AC), washing machines, TVs, computers/laptops, and induction stoves. Along with these, two other pieces of information were collected from the participants: salary and profession. Among these two, the salary was an optional segment. These were collected to map the different types of users in any future cases.

3.3.3 Participants

The participants for this survey were the residents of the Paribagh area in Dhaka. The region includes a mix of both residents and renters. Paribagh is a large and diverse area. Therefore, the survey data was collected from various corners of the neighborhood to ensure a well-rounded and accurate representation of the region. This approach helped to have a broad range of energy consumption behaviors and

household characteristics. It is seen that the residents tend to have bigger houses and more household appliances than the residents. Therefore, we went to both parties to get a diverse dataset from them. The participants were mainly the senior or leading members of a family who could give accurate and detailed insights into their household's energy use.

3.3.4 Complications Faced During Survey

The final dataset combines socioeconomic data from the in-person survey with conventional invoicing measures from DPDC. Amidst all the positive aspects, there were some crucial adversities that were dealt with in times of data collection. The DPDC data collection process required adherence to certain official procedures. However, the most significant concerns raised by the public were related to their privacy. While conducting the in-person survey, it was a herculean task to make everyone ensure to keep their personal data safe and secured. Moreover, it was necessary to contact them continuously to get the data within time, and it did not benefit them directly; they were a bit indifferent to the survey questions or filling up the questionnaire, and that made the survey more difficult as well as time-consuming. It took almost 7 months to complete the survey and get the data we needed. During these extended 7 months, we encountered a variety of challenges, including political unrest, heatwaves during the intense summer heat, and finally the revolutionary quota reform movement in July 2024.

3.4 Data Cleaning and Preprocessing Steps

Data cleaning is necessary to make the data more precious and reliable for the models to have a better output (Hosseinzadeh et al., 2021)[37] . To ensure the accuracy and reliability of the dataset, we have used several data cleaning and preprocessing steps. Failing to preprocess data correctly resulted in a substantial negative impact on the research. There needed to be more consistency in our original dataset, and we have overcome that through the proper use of data preprocessing. We addressed the missing data, standardized the dataset, and removed any inconvenience that would affect the analysis. The methods below are followed to clean the dataset.

3.4.1 Data compiling

In this research procedure, different types of data were collected using different methods. In each method, there was data in different segments. For example, When DPDC provided us with the customers information, these were in different files according to its user tariff and months. Moreover, the survey was conducted over a wide range of time, so combining them together was one of the major requirements for us to run the models. So the first thing that we had to do was compile all the data into a single file so that we could use them altogether. So there were different versions of the compiled data. They are as follows:

- Version 1: In the preliminary version, we had to collect and compile all the postpaid data obtained from the DPDC. In their dataset, there were irrelevant features such as tariff rate and LT, so we had to choose and option out features

from the raw data. Then data from the multiple months were compiled in a single Excel file to have them compared.

- Version 2: The second version of data compilation was done after receiving the remaining data from the DPDC. We had to collect data from them in different sets due to their internal policy. So when they provided us with the second set of datasets, we divided the prepaid and postpaid data and compiled the postpaid data to the previous version. Some monitor feature engineering was done in this version.

- Version 3: The final version of data compilation was completed once we had all the survey data in our hands from the participants. This version of compiling took a longer amount of time since we conducted the survey for several months. Since we used hard copies of the survey form, we collected them from the participants day by day. Meantime of the collection, the compiling of the on-process in a single excel file. Once we collected all the survey forms and compiled them, we completed the final version by combining all the data altogether.

3.4.2 Margin Survey Data with DPDC Data

There were two distinct methods to collect the dataset for this research. We collected data from DPDC; another set of relevant data was collected through a door-to-door survey. First, we converted our survey data into digital data since the survey was conducted using a survey questionnaire form. We incorporated the features of the dataset provided by DPDC in a single place; previously, they were in different places, as the DPDC provided us with the data in multiple handovers. After that, we did feature engineering on the t, extracting the data in different columns based on the month name. We looked for the customer ID who contributed to the survey, and then searched for those customers' IDs in the DPDC-provided dataset. Once we found the match, we merged the features of those datasets with our survey data. This allowed us to have all the features of a single customer ID in a single sheet.

3.4.3 Handling Missing Data

Handling missing data was one of the primary steps in the data-cleaning process. It was required because incomplete or missing data can lead to inaccurate conclusions. For example, while we conducted the survey, some households did not provide complete information about their appliances or demographic details. To handle this, imputation methods were used to estimate the missing values. We used the average value of similar households for those participants who mistakenly did not prove one out of fourteen data points in the survey form. This ensured that due to no appliance count, no rows got dropped. We also removed the records of those participants who did not provide critical identifiers, such as customer IDs.

Additionally, records with significant data gaps, such as missing multiple appliance types or household characteristics, were excluded from the analysis as these records could skew the results. We had 1797 customer data points from the survey, but not

all the customer information was found in the DPDC-provided. Therefore, remove those customers' information as a row from the dataset.

3.5 Data Validation

Data validation is a process to check the accuracy of datasets. It plays a crucial role when data is collected through surveys. In our paper, we have used several data validation methods to ensure the validity of data and its consist

3.5.1 Pearson correlation

Pearson correlation is a statistical measure of the linear relationship between two continuous variables, also known as the Pearson correlation coefficient. It is explained by r, whose value ranges from -1 to +1. A Pearson correlation coefficient close to +1,0,-1. A strong positive linear relationship indicates a positive correlation that means one variable increases, the other increases. A strong negative linear relationship means -1, indicating a negative correlation where one variable increases as the other decreases. A value of 0 implies that the two variables do not have any linear relationship between them. The significance level P values vary from a height level of 0.01 or 0.05. Suppose the achieved P value is less than the cutoff level. In that case, the correlation coefficients are statistically significant, and if the P is more than the cutoff level, then the correlation coefficients are statistically insignificant (Jasjit S. Suri et al., 2018)[38]. There are a few steps to calculate the Pearson correlation coefficient. Firstly, the average of both the variables is found out. Then, for each data set, the variable is subtracted from the mean of that set of data to find the deviation of each data. These difference values are then multiplied together, and the results are summed. Before the division, this sum is divided by the product of the square root of each variable's variances. This process yields one value that measures the degree of linear relationship between two variables. In many fields, Pearson correlation works like economics, biology, and energy forecasting, as it reveals the degree to which two variables are related and can be used for predictive modelling (Bensalah &Hair,2024)[39].

3.5.2 Kolmogorov-Smirnov (KS) Test

This is a well known non parametric statistical test used to compare distributions between two data samples (Massey, 1951)[40]. The main strength of the KS test, however, is that it does not require the data to follow some particular distribution, and therefore is a useful tool in the large variety of studies (Gibbons et al., 2011)[41]. KS test allows us to establish whether two samples are likely to have come from the same underlying distribution by comparing the ECDFs of these two samples. The KS test is used by us in this study to differentiate the total electricity consumption of households from that of the consuming individual home appliances and if such appliances have an equivalently consumed sequence.

How the KS Test Works

Before learning how the KS test works, we need to begin with the theory of an empirical cumulative distribution function (ECDF). The ECDF for a given data sample plots the scale of data points which are less than or equal to a specific value (Conover, 1999) [42]. This gives a clear view of how we do the data across each value.

The KS test compares the ECDFs of two samples by calculating the maximum vertical divergence between the two ECDFs(Massey, 1951)[40]. The D-statistic is identified as this difference. More the distributions vary, the larger the value. . It then checks if this observed difference is large enough to summarize that the two samples are drawn from different distributions (Gibbonset al., 2011)[41].We calculate the ECDF for both total electricity consumption and appliance specific consumption.ative distribution function (ECDF). For any given data sample, the ECDF shows the scale of data points that are less than or equal to a specific value (Conover, 1999)[42]. This offers a clear picture of how the data is d across different values.

By calculating the maximum vertical divergence between the two ECDFs, the KS test compares the ECDFs of two samples(Massey, 1951)[40]. This difference is identified as the D-statistic. The larger the value, the more the distributions deviate. . Then the KS test evaluates if this observed difference is sufficiently large to summarize that the two samples come from different distributions (Gibbonset al., 2011) [41].

Here's a analysis of how the KS test works in practice:

I. **Calculating the ECDF:** For both total electricity consumption and appliance-specific consumption, we calculate the ECDF. The ECDF is simply a step function going up as each new piece of data is added. It is an approach to visualize how data disperses to different values (Conover, 1999)[42].

II. **Measuring the Difference (D-statistic):** Next we evaluate the largest ECDF difference between the two samples. The D statistic is how far the two distributions are. Mathematically, this is defined as:

$$D = \max_x |F_1(x) - F_2(x)|$$

F1(x) and F2(x) are the ECDFs of two samples, so the where. In other words, this value tells us where the two distributions diverge the most.

III. **Null Hypothesis:** Like so many other statistical tests, the KS test begins with a null hypothesis denoted as H0H_0H0. The null hypothesis here is that the total electricity consumption and appliance consumption are from the same distribution (Gibbons et al., 2011)[41]. The aim is to reject or accept that the test is supporting that hypothesis.Once the D statistic has been calculated we then calculate a p value to see if the difference in the distributions viewed under our distribution is statistically significant (Conover, 1999)[42]. If p-value is also lesser than some chosen threshold (normally 0.05), then the null hypothesis will be rejected which means that the samples are not coming from same distribution.electricity consumption and appliance consumption are drawn

from the same distribution. The goal of the test is to either support or reject this hypothesis.

IV. **Significance Testing (P-value):** After the calculation of the D-statistic, we compute a p-value to determine whether the viewed difference between the two distributions is statistically considerable. If the p-value is less than a chosen threshold (commonly 0.05), the null hypothesis will be rejected, signifying that the two samples do not come from the same distribution. We do embrace the possibility that the distributions are the same, since if the p value is larger, then we don't have enough evidence to say that the distributions are different.

3.6 Data Visualisation and Interpretation

Data visualization is significant to any research work since it helps people understand the importance of data through visual representation. Using only text-based systems to represent data sometimes leads the viewer to misunderstand or miss important information such as trends and patterns. Moreover, data visualization sometimes discovers important correlations between the data. There are many data visualization techniques. For our research, we will be using a Line Chart with Moving Average, a Box-and-Whisker Plot visualization, few heatmap to represent prepaid-postpaid data correlation and Scatter Plot with Trend Line visualization techniques.

3.6.1 Average Monthly Consumption with 3-Month Moving Average

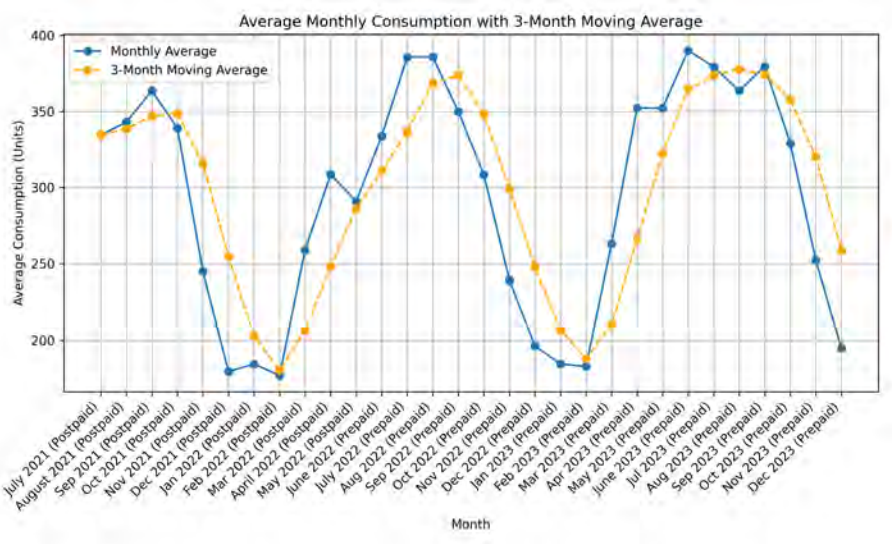

Figure 3.1: Average Monthly Consumption with 3-Month Moving Average

Figure 3.1 is a line chart diagram with a moving average trend line. It provides insight into the postpaid and prepaid monthly consumption data. The orange dotted lines represent a comparatively smooth trend over a 3-month period, while the blue

line represents the actual average consumption for each month of both of the prepaid and postpaid data from July 2021 to December 2023. We plotted months formatted as "Month Year (Type)" in the X-axis and average consumption units of the total user in the Y-axis. Upward points (peaks) indicate months with higher consumption, while downward points or troughs indicate lower consumption. In the orange dotted lines, the average was calculated using the 3 month and plotted in the middle of that. For example, the moving average for August 2021 (postpaid) is the average of the consumption in July 2021, August 2021, and September 2021. From the graph, we can see that the 3-month moving average line is less sensitive, which provides a clear picture of the overall direction of the consumption over time. If the monthly average consumption point is close to the moving average points, that indicates more stable consumption; on the other hand, larger gaps indicate more erratic consumption patterns. Therefore, by examining trends and peaks across the two series, one can assess if the shift from postpaid to prepaid (or vice versa) resulted in changes to consumption behavior.

3.6.2 Monthly Consumption Distribution (Prepaid and Postpaid)

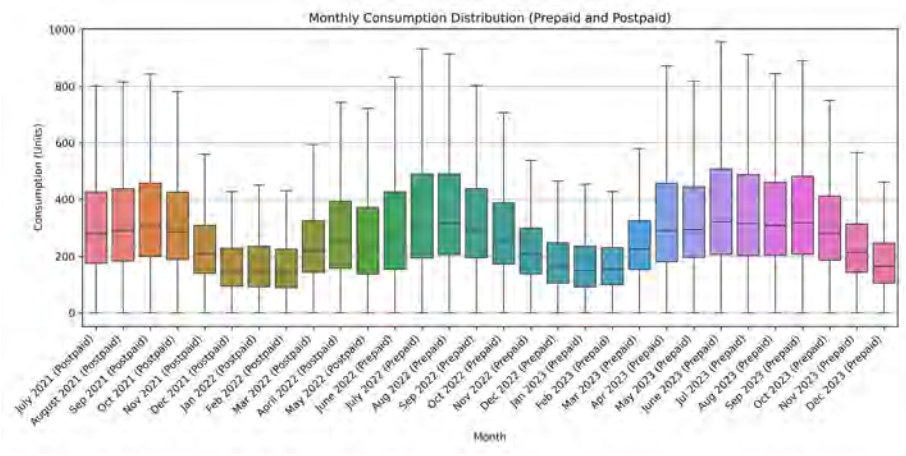

Figure 3.2: Monthly Consumption Distribution

Figure 3.2 is the "Monthly Consumption Distribution (Prepaid and Postpaid)" visualization using the box plot representation aiming to show the distribution of consumption values for each month. In the X-axis, we plotted the month, which is labeled "Month Year Type," and the Y-axis represents the range of consumption units in kilowatts. Fig 3.2 shows the characteristics of a box-plot and whisker plot diagram. The position of the median line within the box shows the central tendency of the data. It can move to the upper quartile or to the lower quartile, depending on the increase or decrease in consumption. The height of the box indicates the variability in monthly consumption. Taller boxes suggest more variability among the middle 50% of consumption values, and the shorter boxes indicate more consistent consumption. The chart allows a side-by-side visual comparison between the prepaid and postpaid consumption. The observation of the median values between the

prepaid and postpaid months reveals a significant shift in customer behavior when transitioning from postpaid to prepaid metering systems. Lower median consumption in the prepaid months compared to the postpaid months suggests a user tends to use less energy. This indicates that prepaid meters encourage greater awareness of energy use due to some reasons. One of the reasons can be the higher bill each month. These findings support the customers' concerns about a higher monthly bill each month.

3.6.3 Prepaid vs. Postpaid Consumption (Scatter Plot)

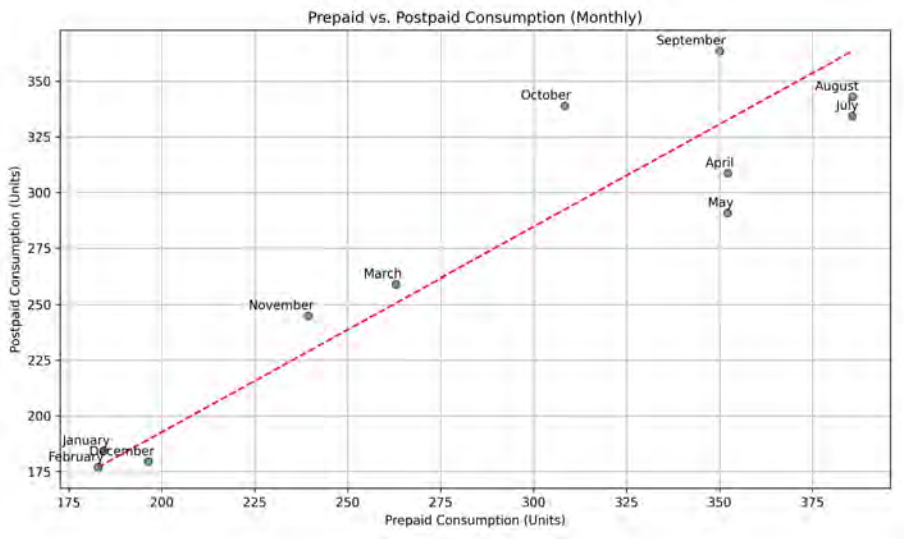

Figure 3.3: Prepaid vs. Postpaid Consumption (Scatter Plot)

This scatter plot diagram 3.3 shows the comparison of the relationship between average monthly consumption units for prepaid and postpaid meter users. In figure 3.3, x-axis represents the average monthly consumption for the prepaid user and y-axis represents average monthly unit consumption of the postpaid user. The data points inside the diagram show the prepaid and postpaid average consumption of a specific month. We see that in some months, especially in September and October, there is a tendency for postpaid users to exceed prepaid consumption. Maybe it is because postpaid users may have less immediate motivation to limit their energy use due to the delayed billing system. But the prepaid meter's user being required to pay bills in advance makes them sensible and therefore reduces the consumption. The diagram shows that the postpaid meter user tends to have higher consumption in July, August, and September. On the other hand, in April, May, and June, the prepaid meter tends to use a higher amount of energy. We can make inferences about how users manage their energy consumption under each metering system by analyzing the diagram. This is important for understanding and potentially guiding future energy management policies or utility practices.

3.6.4 Correlation between Postpaid and Prepaid Monthly Consumption

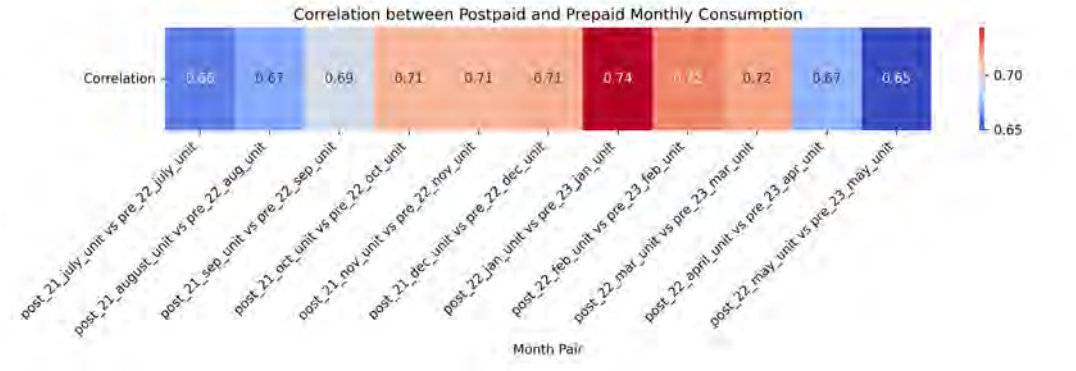

Figure 3.4: Correlation between Postpaid and Prepaid Monthly Consumption

Figure 3.4 shows the correlation heat map between postpaid and prepaid monthly consumption. We took a similar month for two different years, where in one year there was a postpaid meter system but the other year postpaid meters were replaced by prepaid meters. Therefore, the correlation between this postpaid and prepaid data shows how much deviation there is in different months after implementing the prepaid meter. To generate this correlation heat map, the average electricity consumption of all the users of a particular month was considered in both prepaid and postpaid systems. In the heat map, the blue shades represent a lower correlation between prepaid and postpaid data, and red cells represent a higher correlation. We see a lower correlation in July, August, and September (2021-2022) and April-May (2022-2023). It is because these months fall under the summer session and people tend to use cooling appliances such as ACs and refrigerators that consume more energy. From the correlation heat map, we see lower correlation in this month. This concludes that prepaid and postpaid users have different responses in the summer months. It also supports the users claim that after shifting from the postpaid to prepaid meter, there is a noticeable billing increase. On the other hand, in the months of October, November, and December (2022) and January, February, and March (2023) there is a comparably higher correlation between the prepaid and postpaid users. This show in the winter session, people consume less energy. So after shifting from postpaid to prepaid, there is a higher correlation between the prepaid and postpaid meter users due to using less deviation for less energy consumption.

3.6.5 Yearly Average Monthly Consumption Comparison (Prepaid and Postpaid)

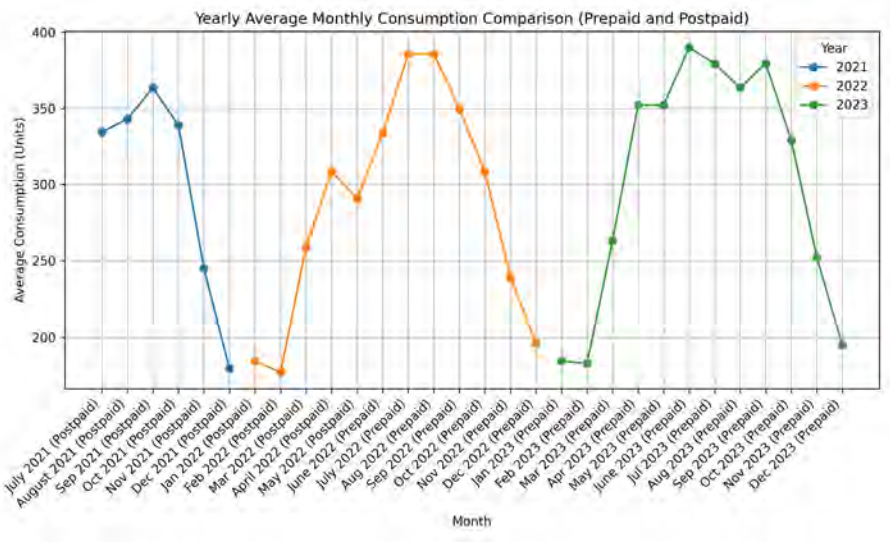

Figure 3.5: Yearly Average Monthly Consumption Comparison

Yearly Average Monthly Consumption Comparison between the prepaid and postpaid meter are shown in the line chart 3.5. It shows the yearly consumption trends in average monthly consumption for prepaid and postpaid metering systems. Different colors show different years from 2021-2023. The X-axis shows the months and meter system, and the y-axis represents the average monthly consumption. It focuses on the changes in consumption, showing how average values shift over time. The line chart shows there was only one spike above 350 units of average consumption in 2021, while the user used to use postpaid meters. But for the same group of users in 2023, there were multiple spikes above the average of 350 units in the prepaid meter system. Therefore, this trend suggests that there are gradual increases in the monthly average usage in the shifting from postpaid to prepaid meter transactions. This was also the hypothetical claim from the users that their bill has increased noticeably after introducing the prepaid meter. We have analyzed this further using the N-BEATS and Isolation Forest in the later part.

3.6.6 Average Monthly Consumption by Month and Type

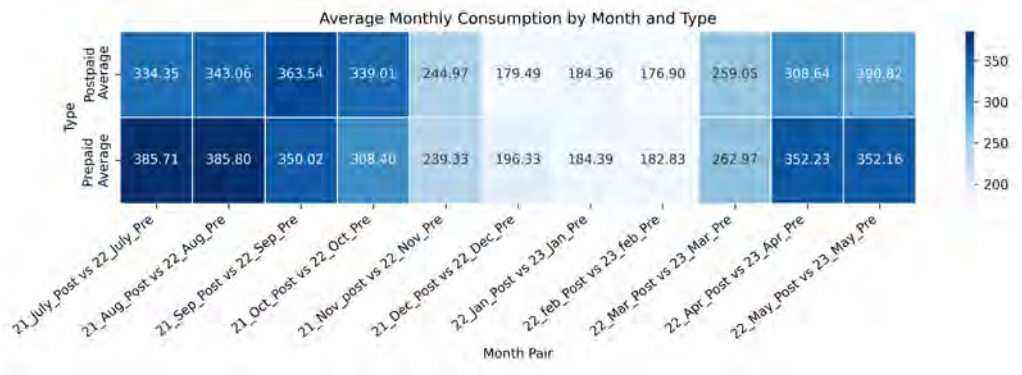

Figure 3.6: Average Monthly Consumption by Month and Type

Figure 3.6 is another heat map that shows the average monthly consumption comparison between the prepaid and postpaid meter. The x-axis represents the prepaid and postpaid comparison, and the y-axis shows the prepaid and postpaid users' monthly consumption average. The darker blue represents the higher average value, and lighter blue represents the lower consumption values. From the heatmap, we can see that the prepaid users have a higher consumption average in April 2022, July 2022, August 2022, and April 2023. On the other hand, in December, January, February, and March, prepaid users have a higher consumption average than postpaid users. In summary, in summer months such as July and August, prepaid meters demonstrate higher energy uses, and in winter months like December and January, postpaid meter users use more energy than the prepaid users. This insight aligns with our hypothesis, and in the later part of research, extensive research using N-BEATS and isolation will be conducted to find the probable anomaly.

3.6.7 Violin Plot of Monthly Consumption

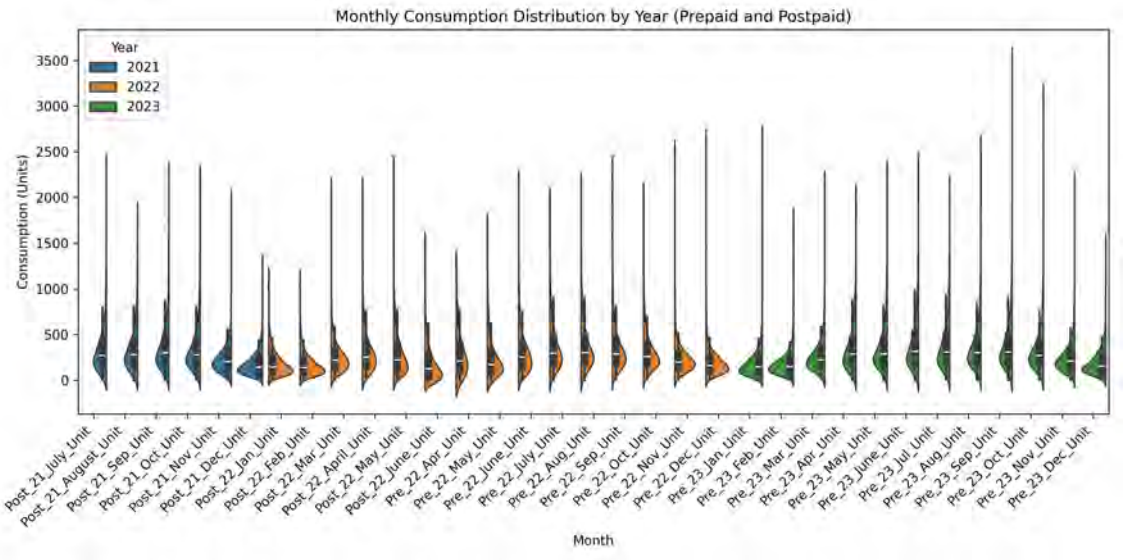

Figure 3.7: Violin Plot of Monthly Consumption

The above violin plot 3.7 shows the monthly electricity consumption distribution for prepaid and postpaid customers across three years. Here, the x-axis represents the month names for the prepaid and postpaid consumption. A violin plot mainly describes the density of data across a period. It illustrates the variety of variations there is and demonstrates broader consumption ranges.

3.6.8 Total Consumption vs Freeze

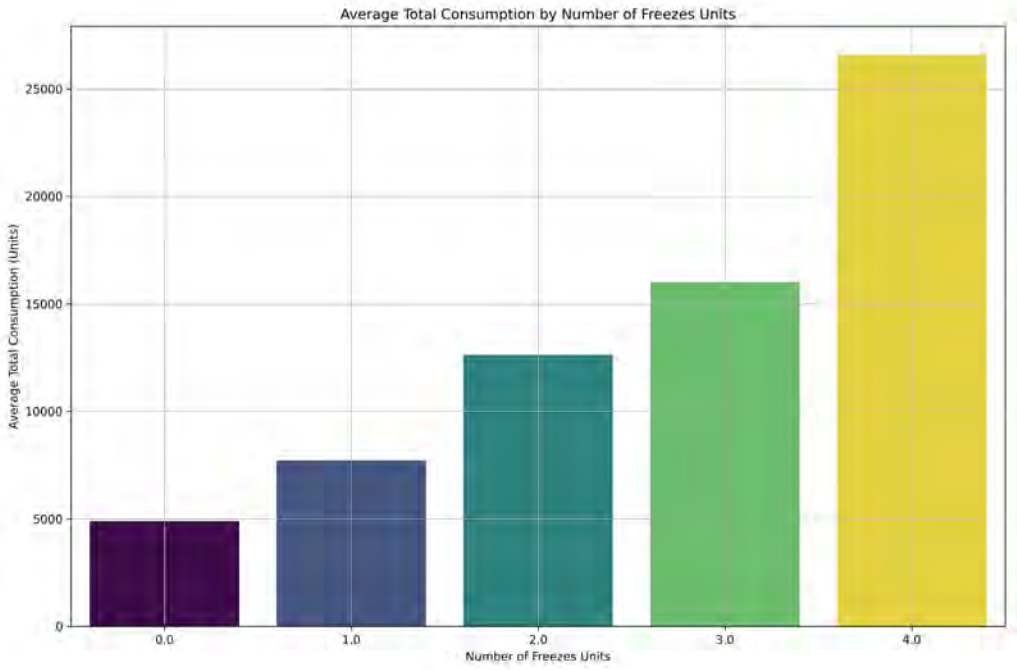

Figure 3.8: Total Consumption vs Freeze

The graph above 3.8 depicts how many freezers a household uses is related to the amount of electricity they use. The y axis tells us how much electricity a household runs averages on and the x axis shows how many freezers a household has. The trend is clear: More freezers equals more power. For instance, homes without any freezers are the lowest in electricity and homes with four freezers are the highest at almost 24,000 units. Such findings suggest that owning more freezers is related to higher electricity use, as freezers use a great deal of energy.

3.6.9 Household appliances vs Total Consumption

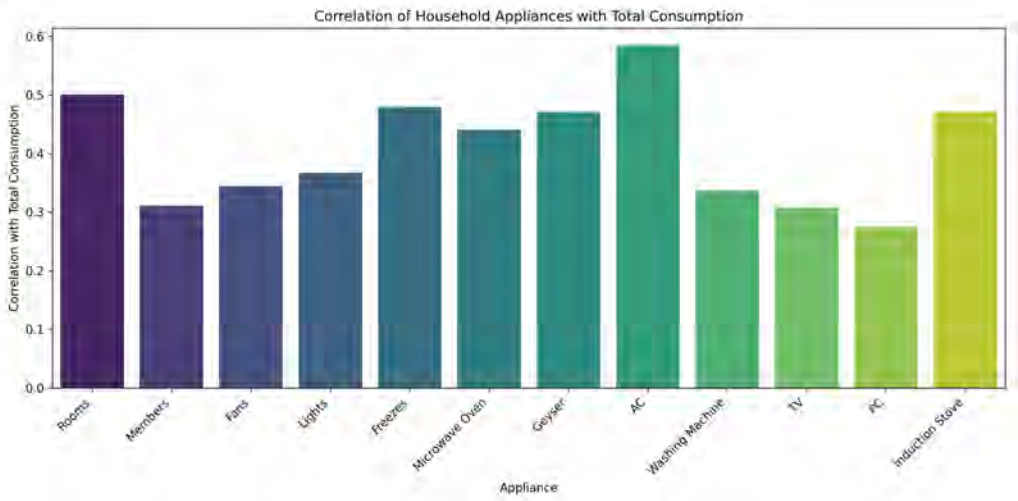

Figure 3.9: Household appliances vs Total Consumption

As this graph 3.9 statistically determines, the number of specific electrical households is proportional to the total electrical usage. On the x-axis are items such as the number of rooms, individuals in the home, or frequent utility appliances like ACs, geysers, or induction cooktops. The y-axis shows how much each factor contributes to total electricity usage.

Purposes, distances, rooms, washing machines and induction stoves should be relevant most significantly in defining the higher electricity usage. This means homes with more rooms or those large appliances that consume energy tend to consume more electricity. On the other hand, the number of household members and personal computers (PC) have a weaker connection to overall electricity use.

3.6.10 Total Consumption vs AC

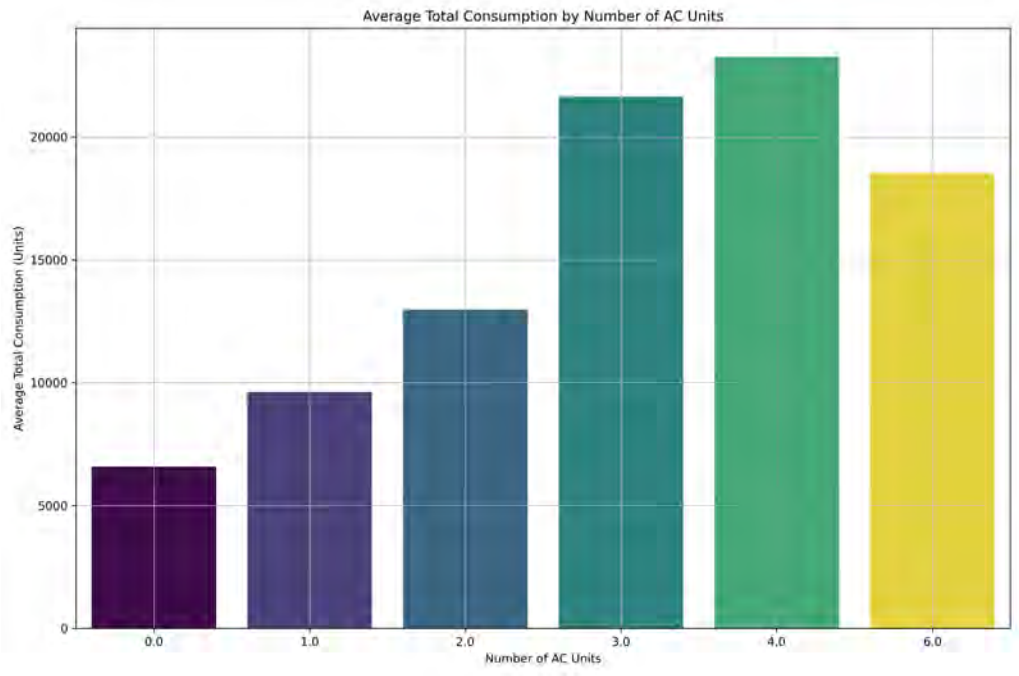

Figure 3.10: Total Consumption vs AC

This graph 3.10 represents the average electrical usage wherein the number of ACs in a house also plays a significant role. On the x-axis, the numbers represent the number of ACs in the house, and on the y-axis, the numbers represent the average electricity use in units/houses.

Residences with more ACs tend to use a lot of electricity in their homes. The consumption statistics show that homes with four ACs consume over twenty thousand units, and homes without AC consume much more power than the former. Surprisingly, houses with six ACs consume less than expected, so something other than ACs affects how much electricity is consumed.

Chapter 4

Results and Analysis

4.1 Data Validation Result

We have used several data validation methods, such as Pearson correlation, and Kolmogorov Sminov test to validate the survey data.

4.1.1 Pearson Correlation

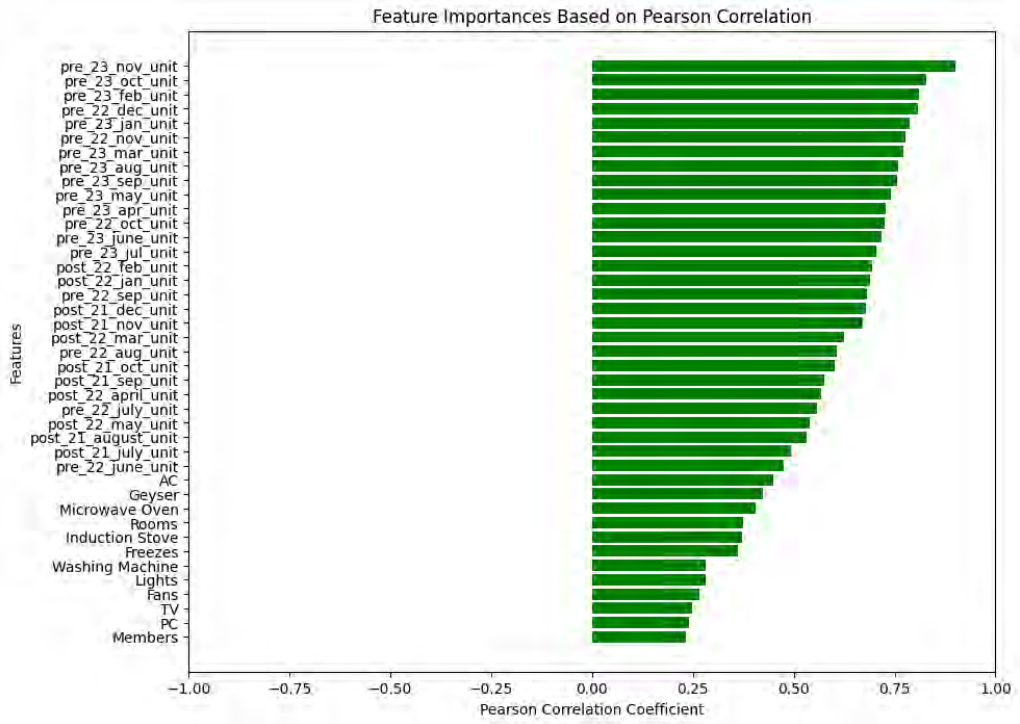

Figure 4.1: Feature Importances Based on Pearson Correlation

Here, we can visualize the most impacting factors that have been associated with the future consumption. From the graph 4.1, it can be seen that the pivotal factors that shaped the future consumption are the past behavioral data rather than the home appliances used by the customers.

But if we see closely, there is no negative impact or value present in the graph and every factor has some influence on predicting the future consumption value.

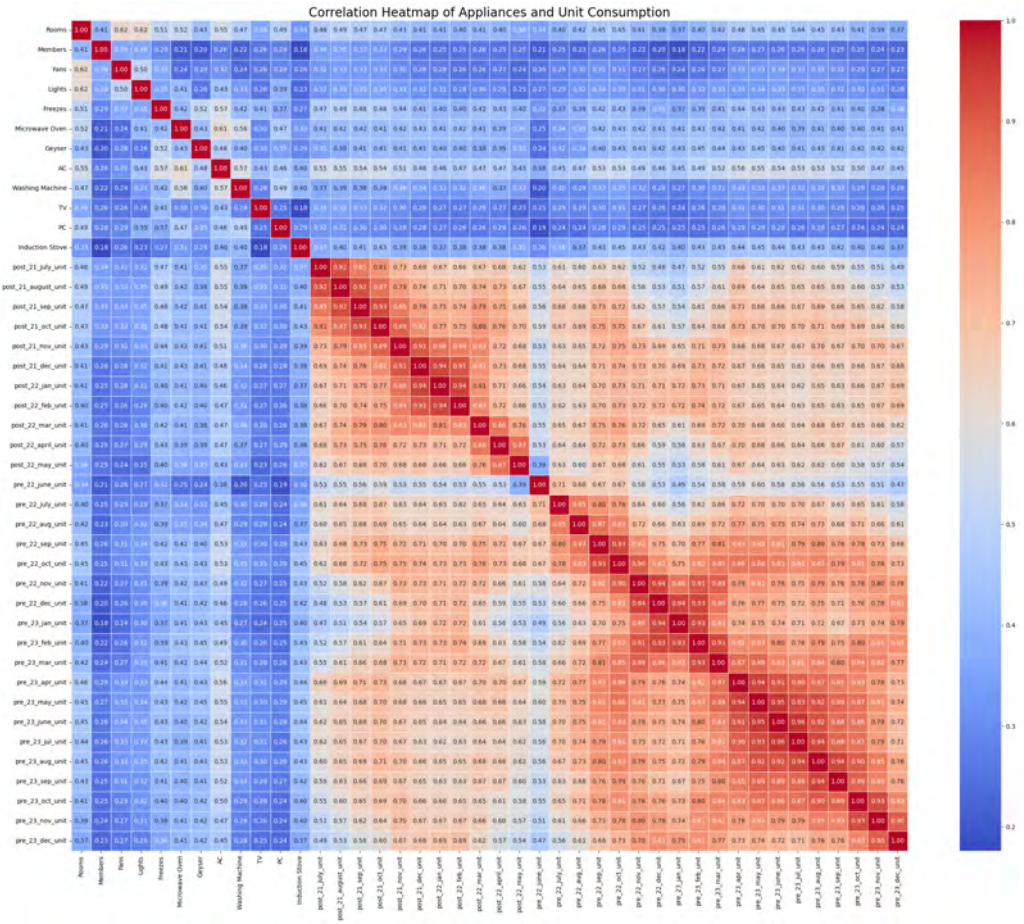

Figure 4.2: Correlation Heatmap of appliances and unit consumption

Here 4.2, the deep red color, which is shown diagonally, reflects the correlation point between features and compares among themselves, and as they are fully correlated with their own, it has come up with a value of 1.00. Moreover, Appliances features such as “TV,” “Lights,” and “Fans” are demonstrating some low correlation (numbers) with each other and it suggests that they are dependent on each other as a feature here. But the most crucial fact is, the correlation between the unit consumption and the number of rooms as well as family members are not high. This indicates that the consumption is more relying on the appliance uses but not the number of rooms as well as the family members.

Another pivotal finding would be the correlation of high energy consuming appliances and low energy consuming appliances with the unit consumption. It has been seen that the high energy consumption months have more correlation with appliances like AC, Freeze and Geyser. Whereas, in those months, lights, fans and TVs have a lower impact on unit consumption.

4.1.2 Kolmogorov Smirnov Test

Applying the KS Test in This Study

The KS test was distinctively pivotal in this research due to its flexibility in managing data that does not adhere to a specific distribution. The determination of whether the electricity consumption patterns of individual appliances were consistent with the overall household consumption or if specific appliances exhibited unique utilization behaviors was our main objective. The ECDFs were calculated for total household electricity and appliance-specific consumption (e.g., air conditioners, washing machines, microwave ovens, etc.). By comparing these ECDFs, we could illustrate differences in consumption behavior among households with different appliances.

For the trial, we classified households based on the presence or absence of specific appliances such as Fans, Lights, Freezes, Microwave Ovens, Geysers, ACs, PCs, TVs, and Washing Machines. Then the KS test was used to contrast the consumption distributions between households in each category. For instance, we analyzed whether households with ACs (air conditioners) had dramatically different total electricity consumption compared to households without ACs.

The results were illustrated through heat maps that signified the p-values from the KS test between appliance-specific consumption patterns. Appliances which possessed high p-values (close to 1) were considered to have consumption patterns analogous to the total household consumption, whereas lower p-values that are close to 0 indicated significant divergences, suggesting discrete consumption behaviors.

KS Test Results and Interpretation

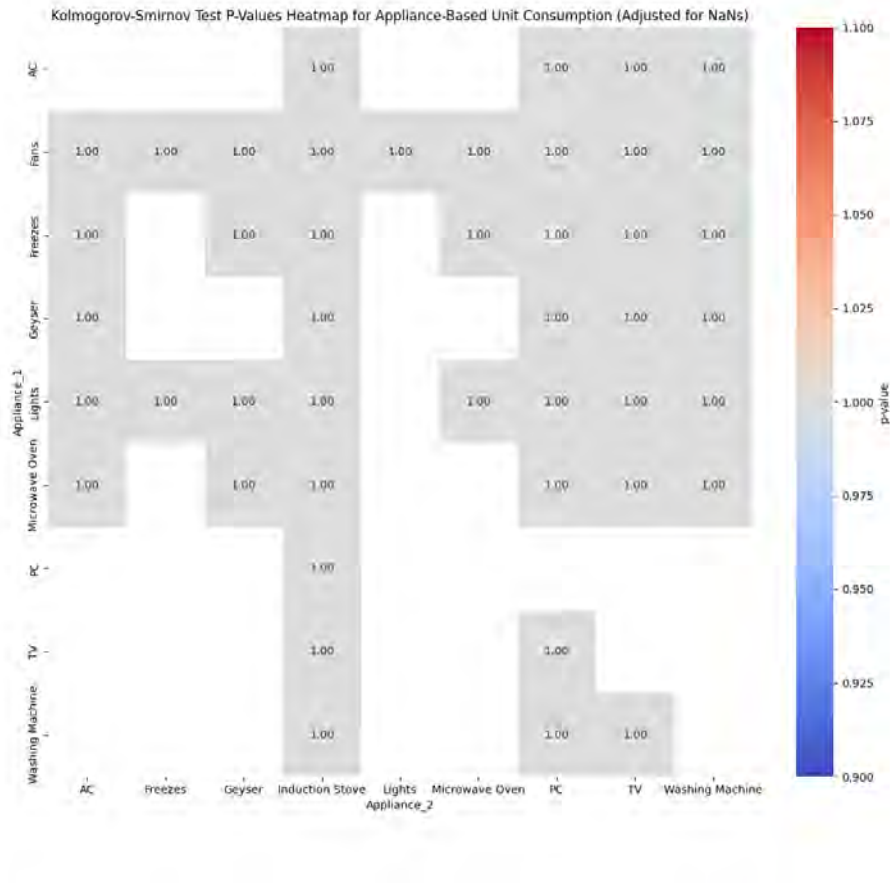

Figure 4.3: KS Test P-Values Heatmap for Appliance-Based unit Consumption

A comprehensive visualization of the p-values from the KS test was provided by the Heatmap 4.3, unveiling the statistical relevance of differences in electricity consumption between different appliances. Significant insights from the heatmap analysis include:

- **Fans and Lights:** These apparatus had a p-value of 1.00, which implies that their consumption patterns were highly similar. This indicates that these appliances are frequently used together, possibly at periods of high household activity.
- **Freezes and Geysers:** A lower p-value of 0.01 was observed between these two appliances, showing significant differences in their consumption patterns. Freezes operate continuously, while Geysers are used periodically. leading to distinct energy consumption patterns.
- **Microwave Ovens and Washing Machines:** A p-value of 0.03 showed remarkable differences in their usage patterns. Microwave ovens usually work

in short bursts, while washing machines run for longer cycles with varying energy requirements.

- **ACs vs. Fans and Other Appliances:** For the comparisons between ACs and other appliances such as fans, lights, and washing machines, significant differences were observed, and the KS test conveyed p-values close to 0. Due to their (ACs) high energy demand, especially in hot climates, have an asymmetrically large impact on total household electricity consumption.

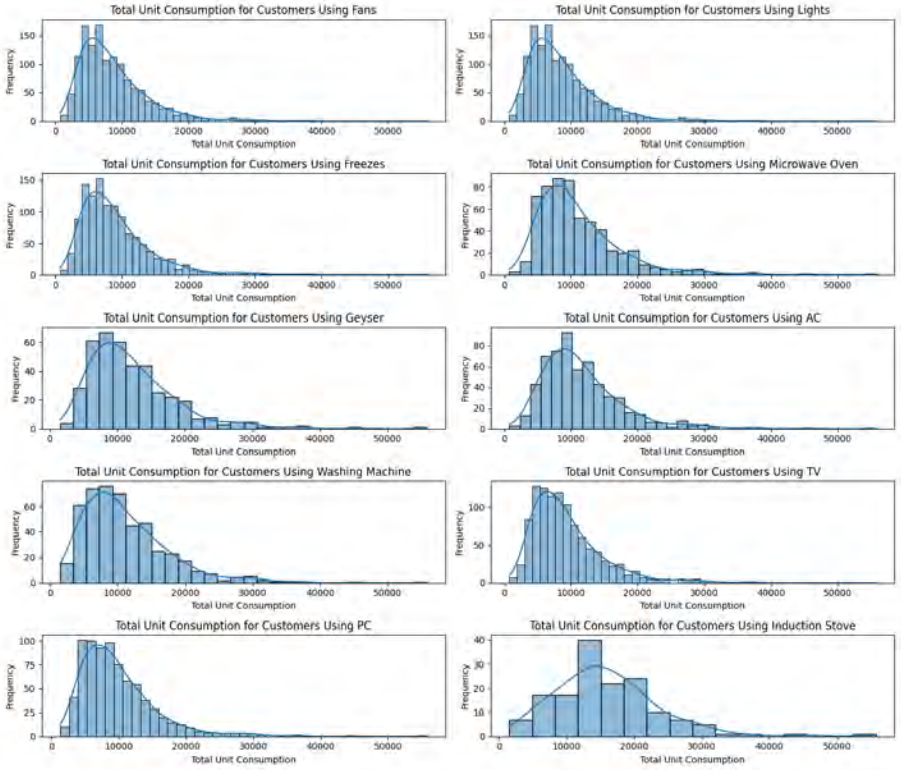

Figure 4.4: Histogram for total electricity consumption distributions for households

Along with the heatmap, histograms 4.4 were generated to illustrate the total electricity consumption distributions for households with and without air conditioners (ACs). The histograms revealed that households with ACs had a more extensive distribution, with higher peaks around 8,000 to 10,000 units and a long right tail. This indicates that these households drew considerably more power. On the other hand, households without ACs had a more focused range, peaking at around 4,000 to 6,000 units, with more moderate consumption.

The Kolmogorov-Smirnov (KS) test demonstrated the statistical significance of this difference with a p-value of 1.04×10^{-18}, highlighting the substantial influence of ACs on overall electricity consumption. These observations suggest that households with ACs consume drastically more electricity than those without, likely due to the energy-consuming nature of air conditioning.

Insights and Implications for Energy Efficiency

The KS test results provide valuable insights into how different appliances contribute to overall household energy consumption. Appliances such as ACs, Freezers, and Geysers illustrate recognizable and often higher consumption patterns, making them leading candidates for targeted energy-saving initiatives. Some criteria of encouraging the use of energy-efficient appliances, such as energy-saving ACs or freezers, and promoting behavioral modifications such as adjusting AC usage during cooler hours could significantly diminish household energy consumption.

Whereas appliances like Fans and Lights showed comparable usage patterns, suggesting that optimizing their energy use, such as switching to LED lighting or using fans instead of ACs when possible, could help reduce overall energy demand. Using these appliances in predictable patterns also indicates that minor interventions could have a broad impact across many households.

Applying the KS test in this study established clear validation of variations in electricity consumption patterns between household appliances. The heatmap of p-values showed that certain appliances, like Fans and Lights, follow similar consumption patterns, while others, such as Freezes and Microwave Ovens, display substantial differences. Moreover, the histogram analysis illustrated the major influence of AC usage on total household electricity consumption, with air-conditioned households consuming significantly more energy.

These insights offer functional advice to improve household energy efficiency significantly by focusing on high-consumption appliances such as air conditioners, freezers, and geysers. Understanding the distinct consumption behaviors of these appliances permits more effective energy-saving techniques, ultimately contributing to more sustainable household energy usage patterns.

4.2 Sliding Window Experiment

In our experiment, we applied a sliding window approach to create input sequences from the dataset, where past values are used to predict the next future value.

In our paper, we applied this technique where we considered window size as four and a total of 17,756 data. This technique inaugurates a moving subset of data where each window can contain N(where N=4) consecutive data points. We applied this sliding window technique to feed our model with plenty of short-term behavioral data. Through this, we have generated multiple overlapping windows to enrich our datasets. Finally, we get a different dataset containing observations from our original dataset.

4.3 Deep Learning Model (N-BEATS)

We ran N-Beats thrice in our dataset for a detailed and comprehensive overview.

4.3.1 N-Beats For Whole Dataset

Performance Analysis:

Here, we have got the following values after running the model:

- **MSE (3214.4075):** It is an average of the square of the difference between actual and predicted consumption values. It gives a feel of the size of prediction error, far smaller values depicting better fits of the model. In our case, an MSE of 3214.4075 implies mild volatility in some predictions when choosing the model at large.
- **MAE (38.7822):** MAE informs us that, on average, the model has deviated from the actual energy consumption by approximately 39 units. MAE is straightforward since the average error is displayed in the same units as the data.
- **RMSE (56.6957):** The root mean square error (RMSE) gives more importance to big errors. In other words, it assists in detecting significant errors or outliers. From this figure, the RMSE value of 56.6957 indicates that though most of the predictions are proximal to the actual values, other predictions are way off from the actual values.
- **R^2 (0.8312):** This coefficient of determination indicates that the model can account for 83.12 percent of the variation in the consumption data. This high R^2 indicates a good fit that effectively captures trends and patterns of prepaid and postpaid consumption to be used for our anomaly detection and consumption patterns.

Residuals of Actual vs. Predicted

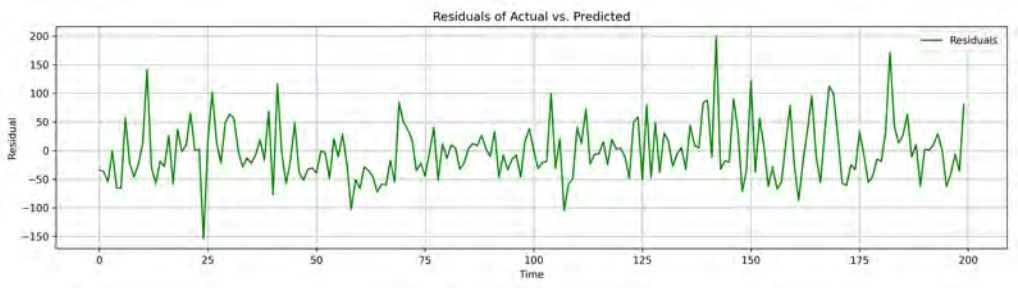

Figure 4.5: Residuals of actual vs predicted

This graph 4.5 explains the residuals (the difference between actual and predicted values) over time. The y-axis represents the residuals, and the x-axis shows time (which probably represents different time points).

Analysis:

- The model is now better at reducing error because most residuals (the difference between the actual and predicted values) appear to be close to zero.

- There are still large differences for some points, such as 25, 150, and 175, which means that the model finds it difficult to handle sudden changes or specific periods.
- Compared to the other residuals plot, few large values are observed, indicating that the new model is less erroneous and makes fewer larger errors than the shown figure.

The model performs better because the residuals are now closer to zero. However, some spikes still indicate that the model may require optimization regarding the response to the varying short-term rates.

Distribution of Residuals

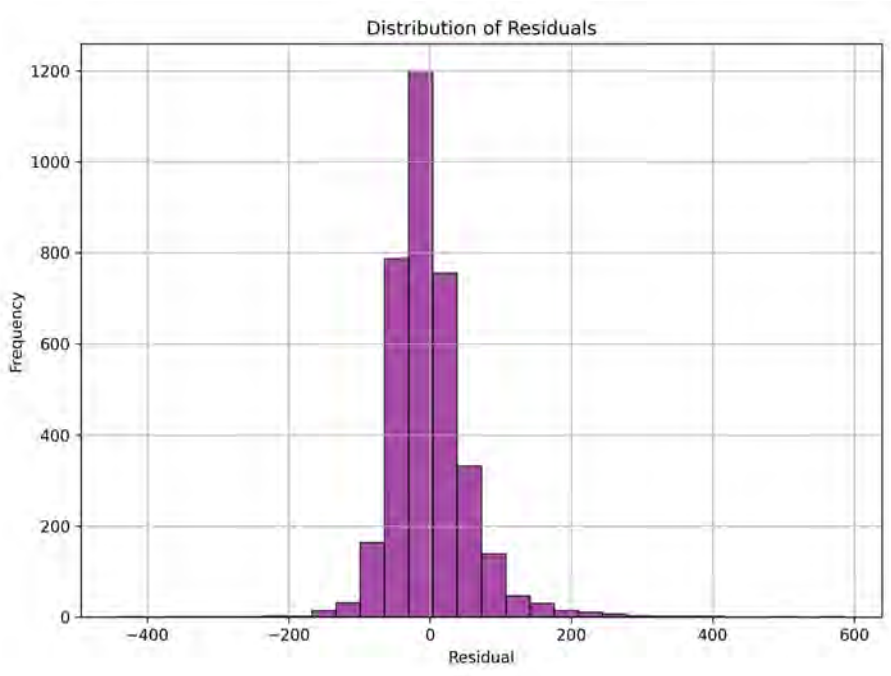

Figure 4.6: Distribution of Residuals

This histogram 4.6 shows the number of residuals on the x-axis labeled Residuals, the numbers of which indicate how close or far off the estimated values are to the actual values. The y-axis in the figure depicts the number of times those residuals were observed.

Analysis:

- More than half of the residuals, which are the differences between the actual and predicted values, are nearly zero. This is favorable since it confirms that the model's predictions are often accurate.
- The outer parts of the graph are characterized by somewhat larger residuals, which occur less frequently. There are also significantly fewer extreme values

compared to the previous case. This means the model is accurate, with fewer big yearly mistakes.

The spread of the residuals is smaller, and the averages are closer to zero, which shows that the model has minimized over- and under-prediction. Compared to previous results, there is an improvement, but there are still a few wrong classifications.

Actual vs. Predicted Energy Consumption

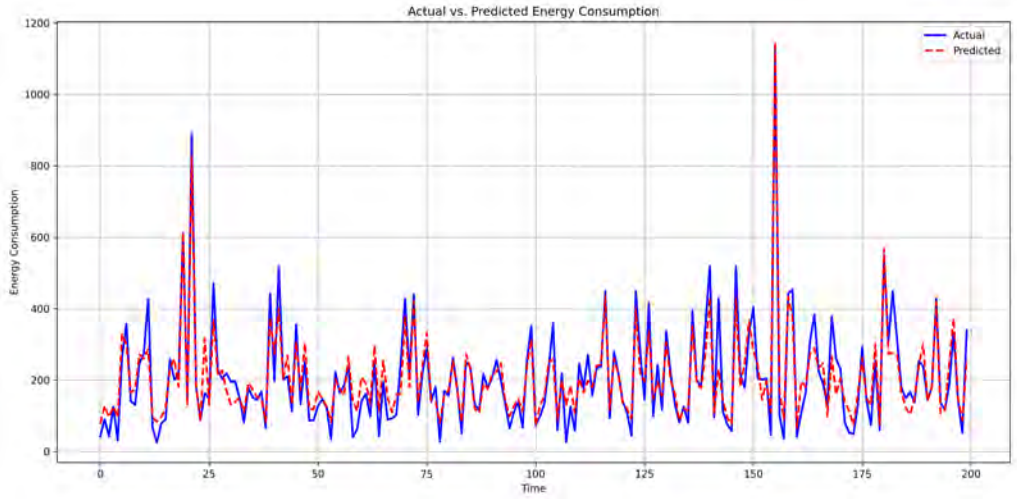

Figure 4.7: Actual vs. Predicted Energy Consumption

This graph chart 4.7 compares actual energy consumption (blue line) and predicted consumption (red dashed line) over time. The x-axis represents time, and the y-axis represents energy consumption.

Analysis:

- Comparing the results of the first and second models, it can be concluded that values have improved, and there are fewer differences than before.
- While there are still some fluctuations (such as around time steps 25, 100, and 150), the developed model is designed to get closer to the overall consumption of the products.
- The model has captured the spikes in actual consumption more accurately, though some sharper peaks still present challenges for precise prediction.

The updated graph clearly shows an improvement in the model's ability to predict energy consumption. Now, it reacts better to most overall movements and all the large fluctuations in the index.

Overall Summary: The new model seems to predict better and has fewer large errors than the previous model. The residuals are closer to zero, plus the nature of predictions is more appropriate regarding the actual consumption values. The model has improved and is much less volatile than it used to be despite the elevated level of consumption fluctuations.

4.3.2 N-Beats for Post-Paid data

Performance Analysis:

Here, we have again run the model into our postpaid datasets only, and the metrics we have got are as follows:

- **MSE (23813.0938):** Again, the mean squared error (MSE) is higher than before, which means the model is more off now. This indicates that it is even harder to predict the postpaid meter consumption, possibly because of variations in this data.
- **MAE (81.7839):** The MAE indicates that the model's variability was, on average, off by roughly 82 units. This is much higher than the whole set, indicating that accurately fitting postpaid data is more challenging.
- **RMSE (154.3149):** RMSE also emphasizes this since it responds to significant errors, as was already mentioned. This higher value shows that the actual model had a significant error. In other words, the model is used to estimate more significant changes in consumption, which may be because postpaid data has more outlying information.
- **R^2 (0.5902):** Comparing the R-squared value, the model used to fit the postpaid consumption data only accounts for 59.02% variation, implying that the model is a moderate fit. This means that the complexity of this postpaid consumption is more than the total average of the entire set, or the model is struggling to extract this information.

Residuals of Actual vs. Predicted

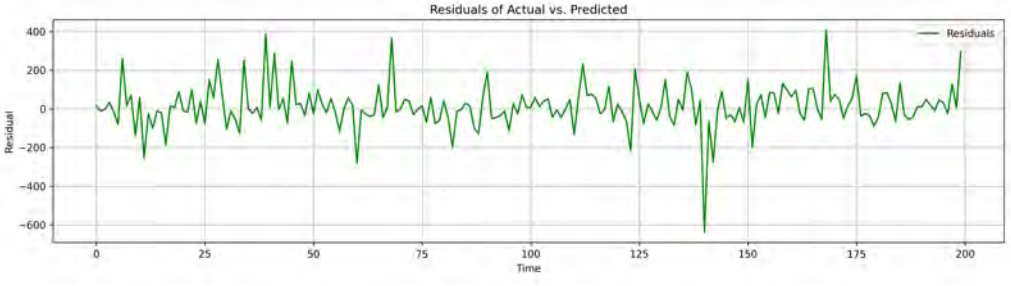

Figure 4.8: Residuals of actual vs predicted

This graph 4.8 explains the residuals (the difference between the actual and predicted values) over time. The x-axis represents time (or time intervals), while the y-axis shows the residuals.

Analysis:

- The residuals move between positive and negative values, where the errors are much more significant because of clear spikes. This shows that the model sometimes struggles to predict the actual energy consumption, and the most significant differences range from -800 to 400.
- At some step values (or time points), such as around steps 25, 150, and 175, the residuals rise above 200 or fall below -600, suggesting that the model makes rather large errors at these various points.
- The model is performing well except for these big spikes, so the residuals stay close to zero most of the time.

It can be seen that there are occasional cases indicated by arrows when the model becomes extremely sensitive to large differences while following the overall trend, implying that the model has difficulty capturing energy consumption fluctuation.

Distribution of Residuals

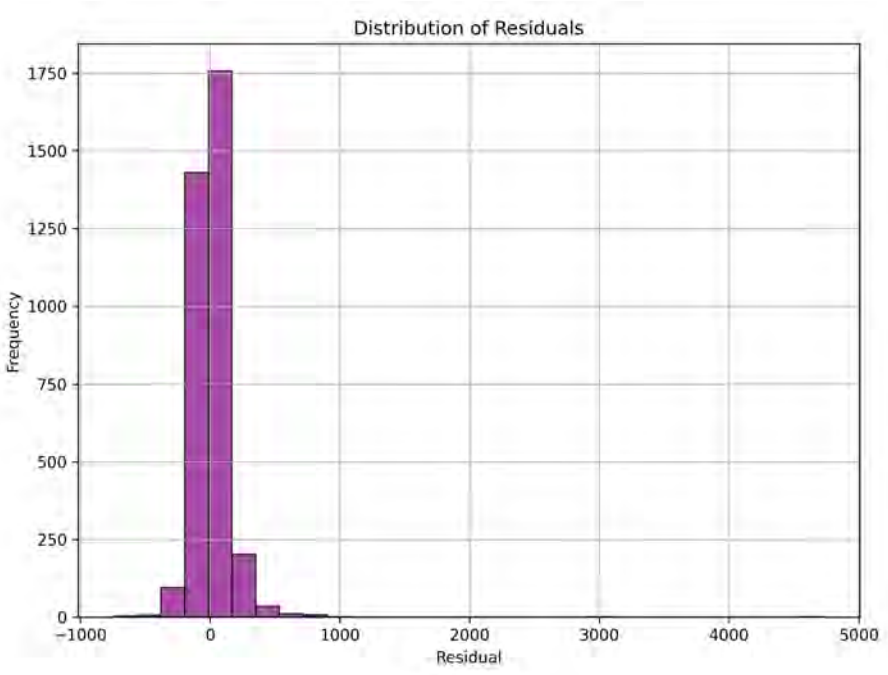

Figure 4.9: Distribution of Residuals

This 4.9 is a histogram of the residuals, showing how often the model's predictions differed from the actual values. The x-axis shows the size of the differences (residuals), and the y-axis shows how often these differences happen.

Analysis:

- The model's predictions were usually close to the actual values because most residuals are near zero.

- However, we observe that few residuals touch up to 4000 or drop down to -1000. This proves the theory that, although the model is a good one in most cases, there are situations where it can go gravely wrong. These outliers may be due to specific events or changes within the consumption levels that do not explain the model well.

Good model performance usually indicates that most residuals are centered around zero. However, the outlier observations show that the model is ineffective when given unusual or extreme cases.

Actual vs. Predicted Energy Consumption

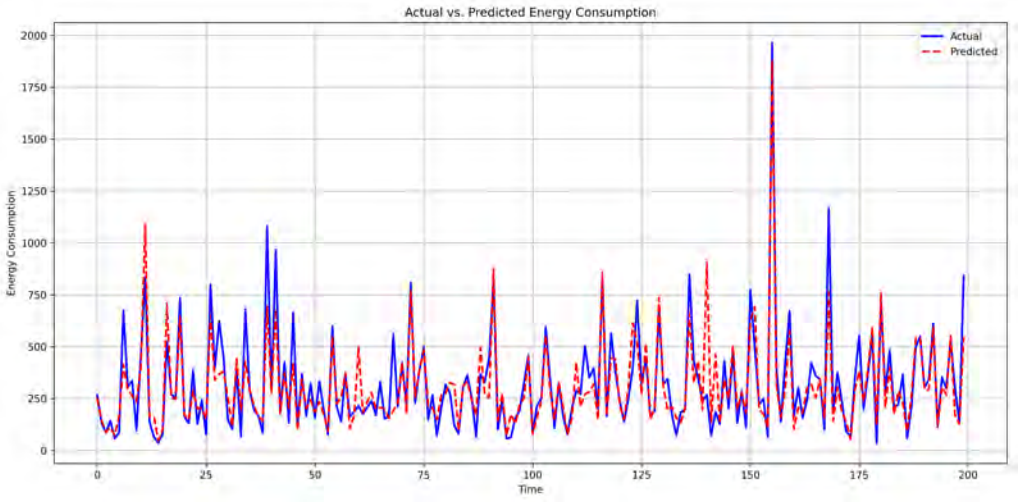

Figure 4.10: Actual vs. Predicted Energy Consumption

This graph chart 4.10 compares actual energy consumption (blue line) against predicted consumption (red dashed line) over time. The x-axis represents time, while the y-axis represents energy consumption in units.

Analysis:

- The predicted values follow the actual values closely most of the time, showing that the model captures the general trend.
- However, there are differences at any point, especially at time steps 25, 100, and 150, because the actual values rise and fluctuate randomly and the model fails to predict such changes appropriately.
- The most significant error occurs at the 150th time step, when the actual value increases significantly compared to the predicted value and the gap between them becomes very large.

The model does a good job overall but not robust for particular high-energy demand shocks. These differences correspond with the residual spikes recorded in the first graph above.

Overall Summary: The residuals also indicate that the model generally makes good predictions, but it struggles to predict large fluctuations where the errors could be large. Almost all residuals are little, near zero, but there are few large positive values showing the areas where the model is inaccurate. The model reflects the data overall trend but it does not react well to sharp changes in actual consumption rates.

4.3.3 N-Beats for Prepaid Dataset

Performance Analysis:

This time, we have run the model with the prepaid dataset and the metrics are as follows:

- **MSE (3150.1155):** Compared to the postpaid model, the MSE is lower in the prepaid model; in other words, the latter is better in predicting the consumption of prepaid meters.
- **MAE (38.4876):** The results of the experiments indicate that using the proposed approach significantly improves mean absolute error (MAE): approximately 38 units on average the model's predictions are wrong. This is just as good as how well the model did overall, meaning it did well for prepaid data all across.
- **RMSE (56.1259)** The root mean squared error (RMSE) shows how sensitive the model is to more significant errors. The RMSE here is reasonable, meaning most predictions are close to actual values, with fewer big mistakes than the postpaid data.
- **R^2 (0.8346):** The R-squared value tells us that the model explains 83.07% of the variation in prepaid consumption. This high R^2 means the model does a good job capturing the consumption patterns of prepaid customers, similar to how it fits the overall dataset.

Residuals of Actual vs. Predicted

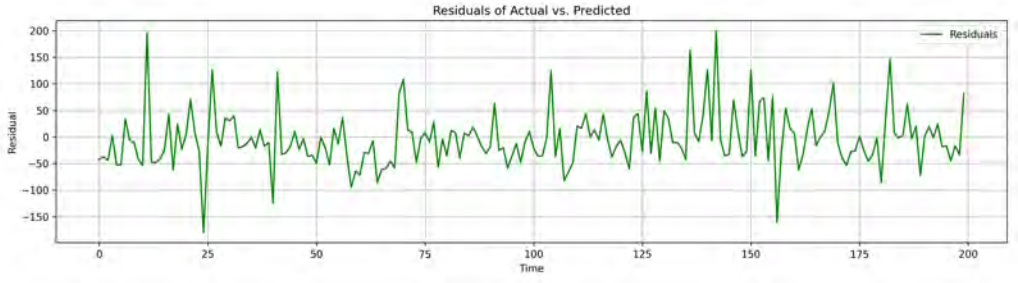

Figure 4.11: Residuals of actual vs predicted

This 4.11 shows the residuals (between actual and predicted values) over time. The x-axis represents the time intervals or steps, while the y-axis represents the residuals.

Analysis:

- This shows that the model works well for most time points, with the residuals between -100 and +100.
- However, if we look at the local level, there are relatively sharp increases, for example at step 25, or pretty sharp drops, such as at step 150, where the residuals reach approximately +200 or drop to -200. This means that during such periods, the model struggles to accurately predict the dependent variable.
- This model is more stable than previous ones, with fewer big mistakes, but some extreme changes remain

Overall, except for these occasional peaks, the residuals are generally small, which indicates that the model performs well in managing energy consumption variation.

Distribution of Residuals

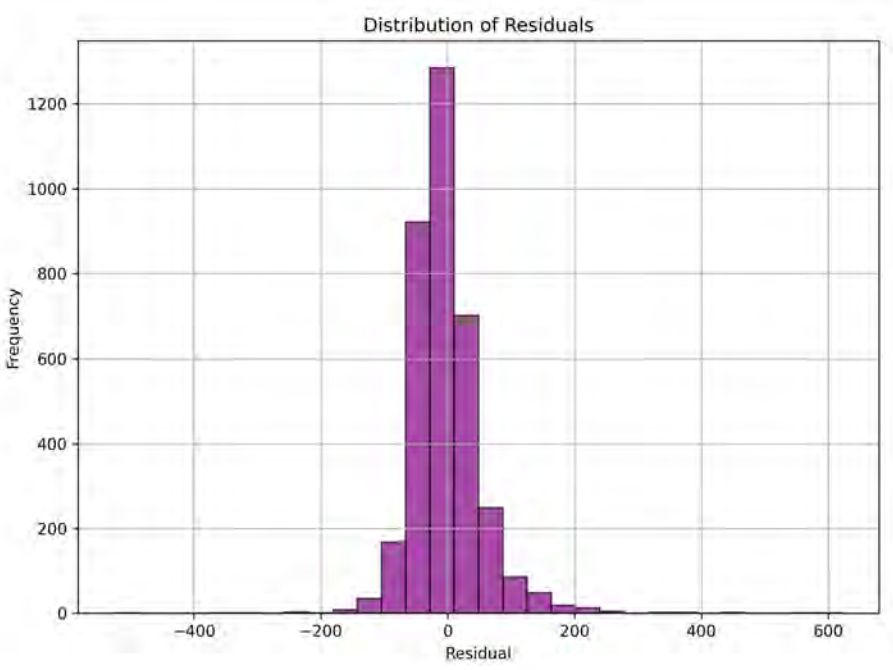

Figure 4.12: Distribution of Residuals

This histogram 4.12 shows the distribution of the residuals, with the x-axis representing the magnitude of residuals and the y-axis representing the occurrence frequency.

Analysis:

- The residuals are mostly centered around zero, which is a good sign that the model predicts well for most cases.

- The spread of residuals is less than in the previous data sets, with much less variation on the extreme values. The line touches zero in the middle, meaning the model fits most values well.
- There are only two or three outside the figure, primarily positive (+ up to 400), indicating that the model sometimes provides wrong results.

The model is good in general, with a few larger prediction errors here and there. The more closely clustered residual plots around the zero line indicate some improvement in prediction accuracy.

Actual vs. Predicted Energy Consumption

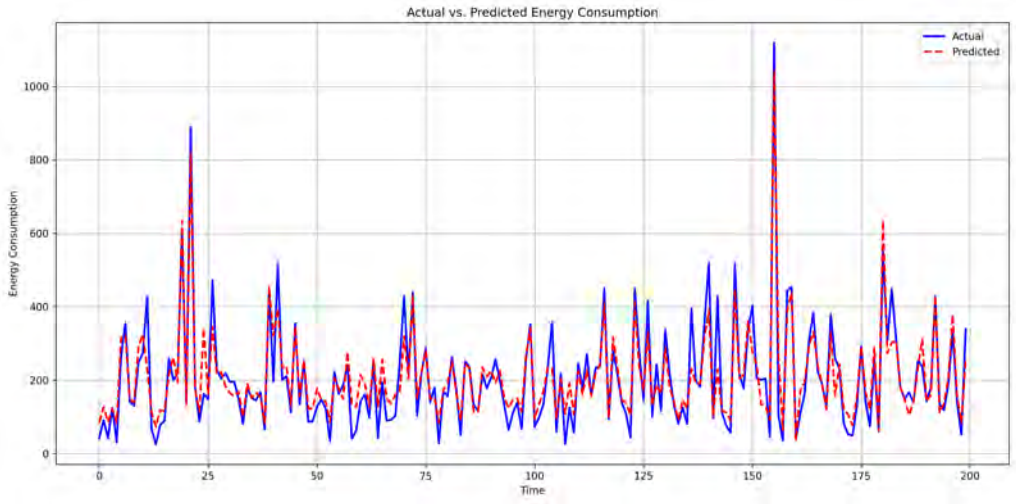

Figure 4.13: Actual vs. Predicted Energy Consumption

This graph 4.13 compares the energy consumption (blue line) with the predicted consumption (red dashed line) over time. The x-axis represents time, while the y-axis represents energy consumption.

Analysis:

- It can be observed that the predicted values presented by the model very closely fit the actual values, mainly when changes are not very large. However, there are a few sharp spike events (such as Windows 25, 100, and 150) where the predicted values are low, mainly when high consumption levels.
- Around time step 150, the actual consumption spikes to over 1000, while the model under-predicts the spike, showing a large residual in that region.
- Overall, the model seems to hold quite well in terms of general consumption patterns.

The model is good at trailing consumption as it happens, but it is not very good at areas with very high-step consumption changes. This may be because the model fails to capture sudden changes in energy consumption patterns, common during wintertime.

Overall Summary: The residual plot shows that the model works well most of the time, but there are a few spikes where the predictions are off more than usual. While the second model shows improved prediction accuracy, as most residuals are closer to zero and to the left of the x-axis, there are still some outliers. The model reproduces the general pattern of actual consumption, but more significant fluctuations in the figures imply larger prediction errors.

4.4 Compare Prepaid and Postpaid Metering Systems

After using the N-BEATS model, the difference between the prepaid and postpaid datasets was compared to understand energy consumption usage. They identified one main fact that contradicts the research question: there is no transition from prepaid to postpaid meters, which indicates some problem areas in addressing customers' complaints.

Finally, the provided model failed to estimate people's energy consumption for the postpaid data. The residual gap between actual and predicted values showed a lot of variability. For specific time steps, it was observed that the residual amount rose beyond ±200 units, indicating a model failure in those periods, such as at time steps 25, 150, and 175. This means that the post-paid meters are used randomly, so people's behavior changes since they pay their bills monthly. Performing this model on the above data was inefficient because the MAE was 92.84, and the RMSE was 162.39. This illustrates that the difference between the actual and the predicted values was wide off the mark. The R^2 value of 0.5462 indicates that postpaid data changes were only 54.62% explained, which tells how volatile this usage is to predict.

On the other hand, the prepaid data was much less volatile, so most residuals fell within the range of -100 to +100 units. However, there were again some outliers at time steps 25 and 150. This means that while the model's performance improved, it could not predict sudden changes well. The model had better predictions with prepaid data, with an MAE of 38.50 and RMSE of 56.79. The analysis shows that the model successfully accounted for 83.07% of the R^2 value, indicating a more stable consumption pattern under the prepaid system.

Many more differences between the two systems can be seen when comparing prepaid and postpaid data. Therefore, although it enhanced the overall energy usage forecast with an R^2 of 0.8372, which suggests all is well with the data fit, there were still open gaps during some periods. For instance, the model failed to predict sharp rises in energy use characterized by time discrete 25, 100, and 150, resulting in differences or residuals of about ± 400 units.

These findings suggest that switching from postpaid to prepaid meters is not as smooth as expected. The model's struggle to predict sudden changes in energy use points to irregularities or shifts in customers' behavior when they switch to prepaid meters. This difference between predicted and actual values may be one reason some prepaid customers feel their bills are higher.

To correct this, energy distribution authorities should examine such anomalies and strive to provide accurate and fair billing. Thus, the results indicate that optimizing different systems requires different approaches, as people employ energy differently depending on whether they use postpaid or prepaid systems.

This paper provides helpful suggestions for enhancing customer satisfaction based on the analysis of such energy consumption profiles and machine learning predictive models. It points out that authorities need to determine why these differences occur and ensure that billing matches energy consumption more closely. By doing this, they could close the' expectation-experience' gap on energy consumption between the prepaid and postpaid systems.

Table 4.1: Performance Metrics for Prepaid, Postpaid, and All Customers

Dataset	**Test Loss (MSE)**	**MSE**	**RMSE**	**RMSE**	**R^2**
Prepaid	0.1545	38.4876	3150.1155	56.1259	0.8346
Postpaid	0.4185	81.7839	23813.0938	154.3149	0.5902
All	0.1579	38.7822	3214.4075	56.6957	0.8312

4.5 Anomaly Detection Using Isolation Forest

In this case, through the Prophet model and Isolation Forest, we fully understood what challenges customers encounter when consuming prepaid meters. It was possible to get good quality information from data in addition to the valuable information that the energy providers can apply in improving the delivery of their services. The combined approach reflects the usage of electricity by customers as well as unusual events that might need more attention.

4.5.1 Distribution of Anomaly Scores:

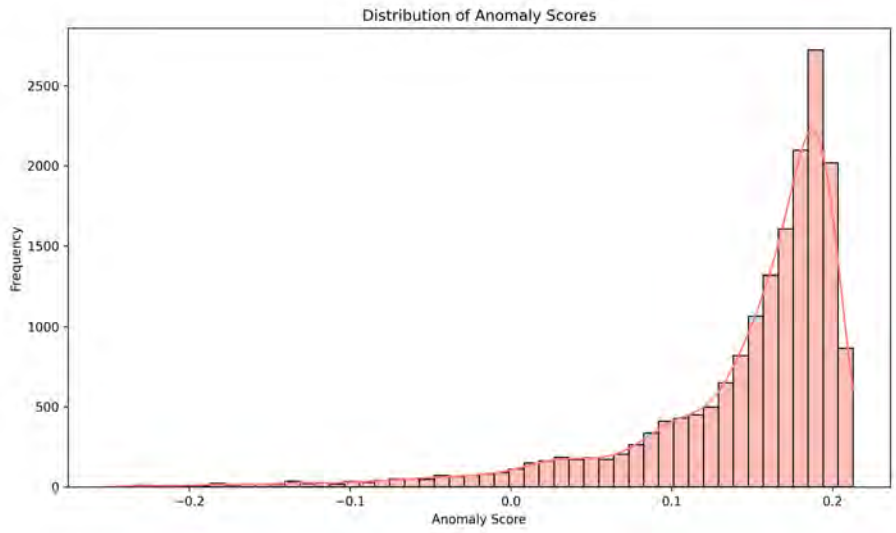

Figure 4.14: Distribution of Anomaly Scores

The graph 4.14 showing how frequently each anomaly score appears corresponds to the Isolation Forest model after Prophet analyzed the data. If the data points represent electricity consumption predicted by Prophet, this is called a residual plot or sometimes a histogram.

Interpretation:

Most anomaly scores lie in approximately 0.2, indicating that the model identified the normality of most data points. The graph is slightly skewed, which means that while most usage patterns are close to normal, a few data points stand out as significant outliers, with their scores getting smaller as they deviate more from the average.

Significance:

This graph is helpful because it shows what the model considered typical versus unusual. The skewness shows that extreme deviations (possible anomalies) are rare, but they are significant and worth investigating when they do happen.

4.5.2 Residuals for Top 5 Anomalous Customers:

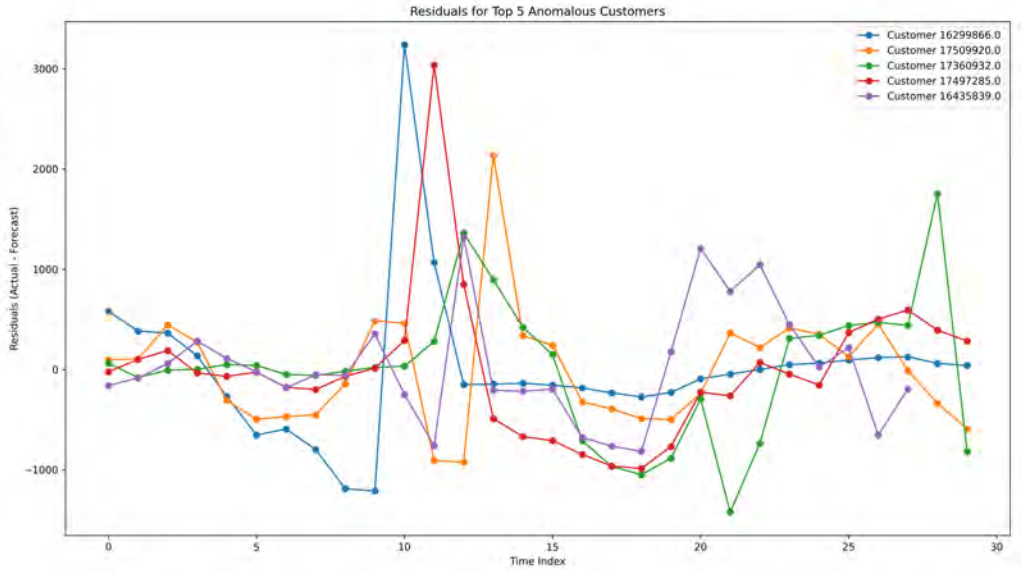

Figure 4.15: Residuals for top five anomalous customers

This graph 4.15 shows the residuals (the difference between actual and predicted electricity usage) over time for the top 5 customers with the most anomalies. Each line represents one customer, showing how their usage differed from the predicted values at different times.

Interpretation:

The residuals refer to the discrepancy between the actual electricity used by a customer and the electrical usage predicted by the Prophet model; thus, high values for the residuals indicate that at specific points in time, the electricity use was much higher or lower than was forecasted. Large upward or downward spikes suggest unexpected changes in their usage, which could point to unusual behavior or possible billing mistakes.

Significance:

Examining these residuals also enables us to see patterns of how the sales frequency of specific customers varied from the anticipated level. It proved that abrupt changes or discrepancies may have been caused by customer complaints, particularly when the conversion of postpaid to prepaid systems was being effected.

4.5.3 Count of Anomalous vs. Normal Customers:

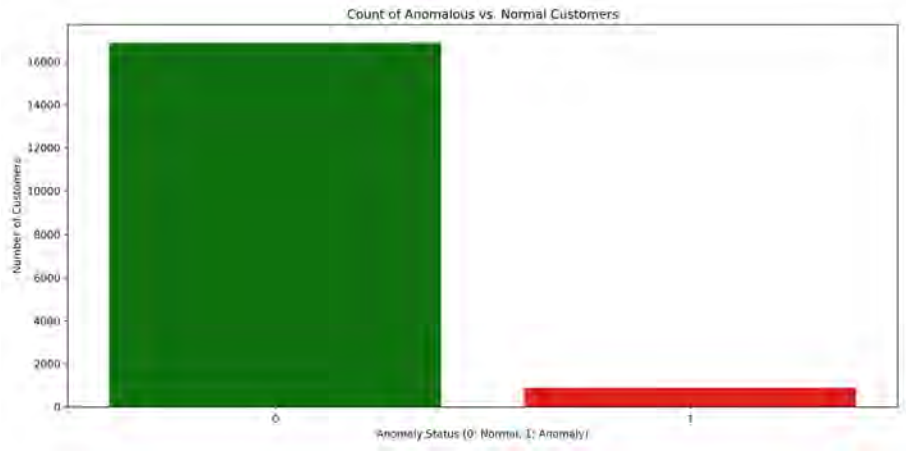

Figure 4.16: Count of anomalous vs normal customers

This bar chart 4.16 visualizes the distribution of customers classified as having anomalies versus those without. The number of regular customers is significantly higher than those with detected anomalies.

Interpretation:

It has been discovered that most customers fell under the standard load profile. Still, a subset of samples exhibited atypical behavior, and the model of the Isolation Forest recognized it. This group is qualified as having unstable rhythms.

Significance:

This means that most customers' usage profiles resemble the predicted values, but a few customers may behave differently. If these anomalies could be addressed, customers' complaints would be minimized, and the prepaid billing system would be enhanced.

4.5.4 Confusion Matrix:

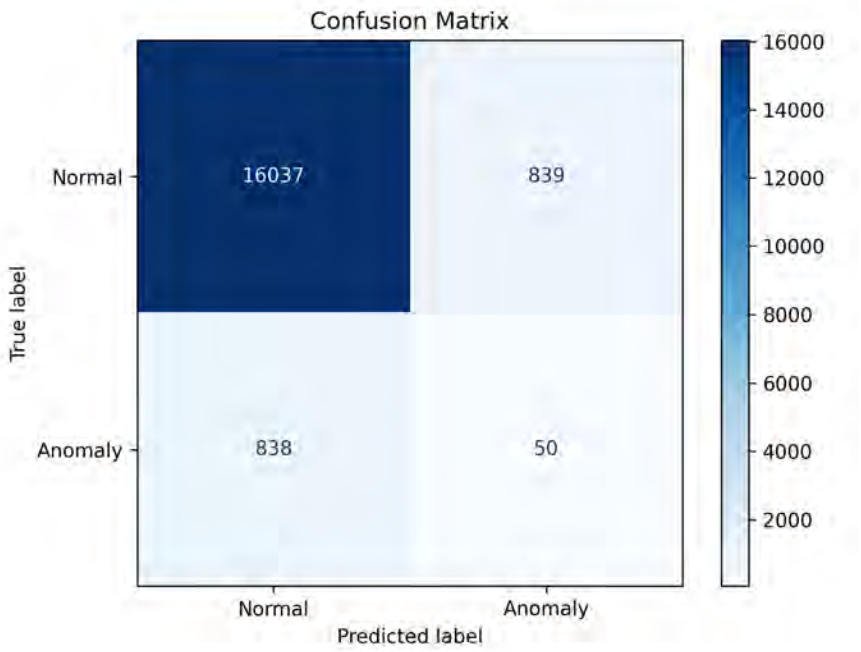

Figure 4.17: Confusion Matrix

This figure 4.17 displays the performance of the anomaly detection model by comparing accurate labels (actual anomalies vs. normal data points) against the model's predictions.

Interpretation:

- **True Positives (50):** The model correctly identified 50 natural anomalies.
- **True Negatives (16,037):** The model correctly recognized 16,037 cases of normal consumption.
- **False Positives (839):** The model wrongly flagged 839 standard cases as anomalies.
- **False Negatives (838):** The model missed 838 actual anomalies.

Significance:

Using the confusion matrix, we can accurately distinguish between the model's ability to identify authentic anomalies and misclassifications. The representation of true negatives also indicates that the model effectively identifies normal consumption behaviors. However, the similar number of false positives (wrongly flagged anomalies) and false negatives (missed anomalies) means there's room for improvement, especially in catching more real anomalies without raising too many false alarms.

Overall Insights:

We integrated the more elaborate Prophet model with the Isolation Forest to enhance the method, which brought an exciting addition to anomaly detection. First, the Prophet gave average consumption predictions, and then the Isolation Forest pointed at any anomalies. These visualizations provide a better understanding of deviation from the customers' behaviors than expected and the ability of the model to identify such deviation. This is important for solving customer complaints about billing errors when switching from postpaid to prepaid systems.

4.6 Pearson For Home Appliance Usage

We used the Pearson correlation test to get a complete overview of how different types of home appliances correlate with the customer's consumption unit.

4.6.1 Pearson Correlation based on Prepaid Meter Usage feature Importance

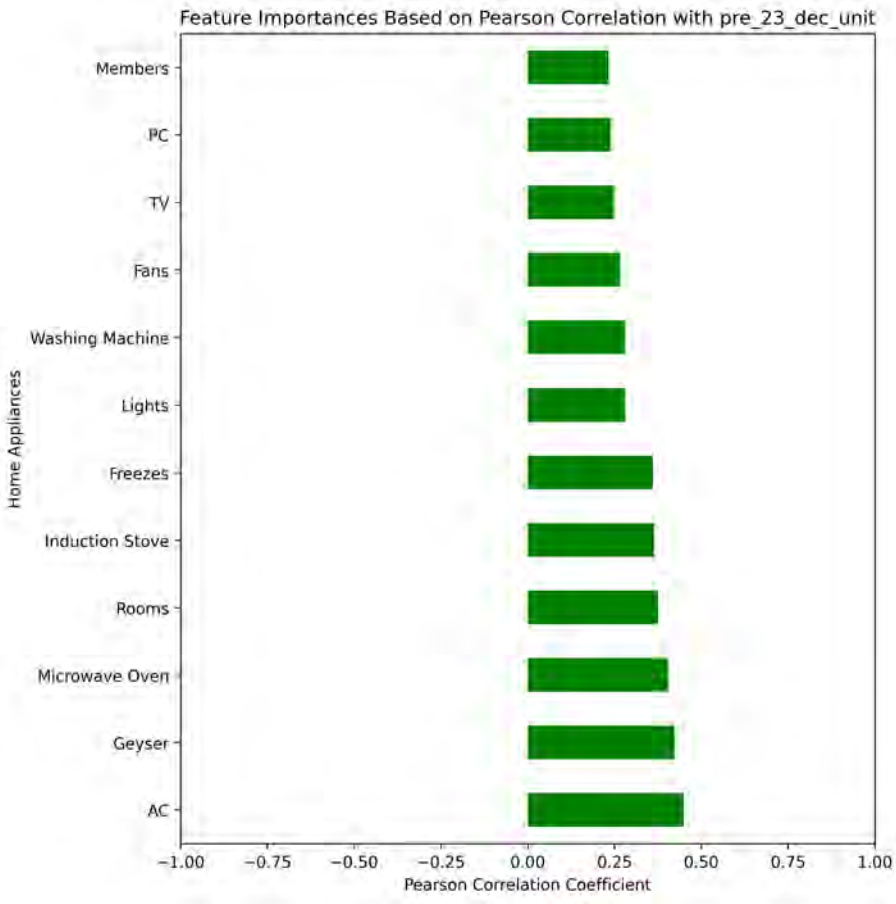

Figure 4.18: Feature Importance with Prepaid Meter Usage

This diagram 4.18 shows a Pearson correlation between different household appliances and the usage of prepaid electricity. The survey data we collected shows how many of those appliances genuinely contribute to our electricity consumption.

Prepaid electricity consumption correlates strongly with the strength of the correlation, for example, compared to air conditioners, whose correlation is about 0.77. That means that homes with air conditions consume a sizable amount of electricity. Also included are other high-energy appliances, such as geysers (0.67) and microwaves (0.58), which indicate strong positive correlations (i.e., also contribute significantly to increased electricity use). That lines up with our survey findings that showed homes with more appliances used more electricity.

Indeed, we also observed positive correlations with the number of rooms (0.45) and household members (0.42). Based on this, larger households and rooms use more electricity because there is more need for lighting, fans, and other appliances spread throughout the whole house.

Together, these findings point to how vital our survey data was overall. If we know what appliances people have in their homes, we can better analyze consumption patterns to learn why all prepaid users feel their bills are higher.

4.6.2 Pearson Correlation-Based Feature Importance with Postpaid Meter Usage

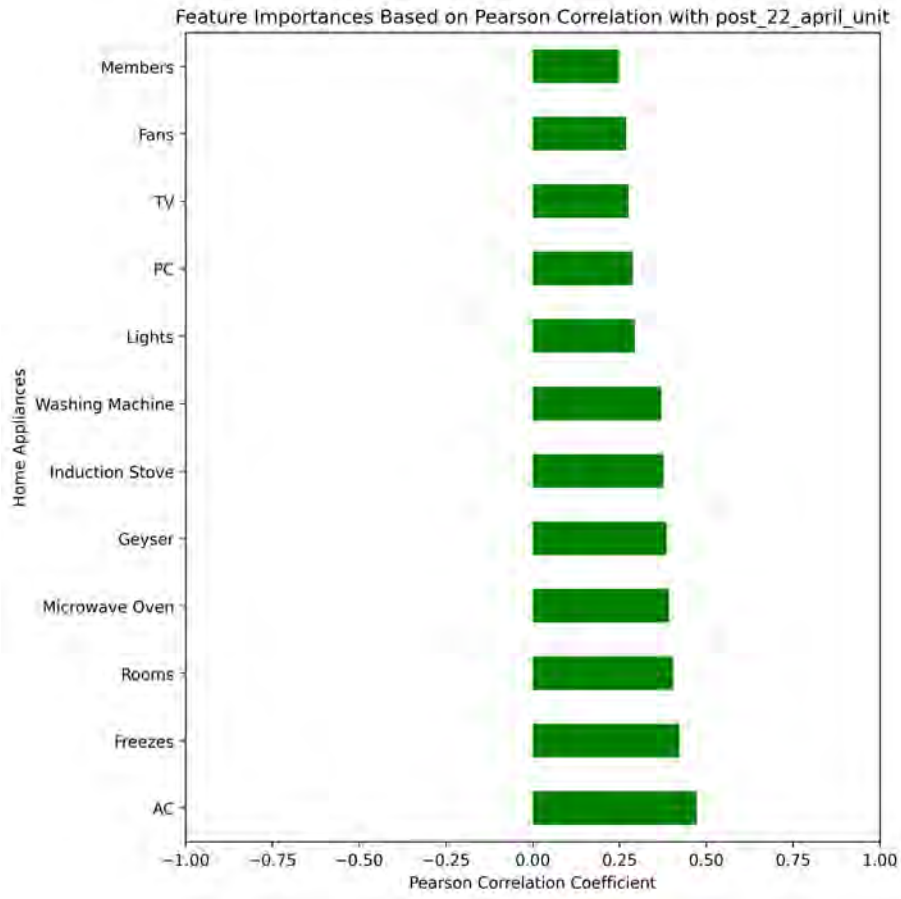

Figure 4.19: Feature Importance with Postpaid Meter Usage

This figure 4.19 shows the electricity usage correlation for postpaid meter users based on the Pearson correlation of household appliances. Prepaid data show a similar correlation with electricity consumption as air conditioners, at 0.65. It's not as high as what we saw for prepaid customers, suggesting that postpaid customers may not use their air conditioners as often due to billing perception.

We also see strong correlations in other high-energy appliances: geysers (0.55) and freezers (0.48), but with somewhat lower correlations than we found in prepaid systems. This pattern suggests that although these appliances result in higher consumption, postpaid users use them slightly less than prepaid users.

Similar correlations are seen in the number of rooms (0.41) and household members (0.38), for prepaid and postpaid meters, suggesting larger households consume more electricity whether or not that household is prepaid.

Exploiting the data from our survey, it becomes clear how each appliance influences electricity consumption. These numbers confirm our belief that understanding how

energy is used is crucial to understanding energy use by prepaid and postpaid users. They also indicate how people use electricity based on what system they pay with.

4.6.3 Home Appliances vs. Prepaid Usage Pearson Correlation Heatmap

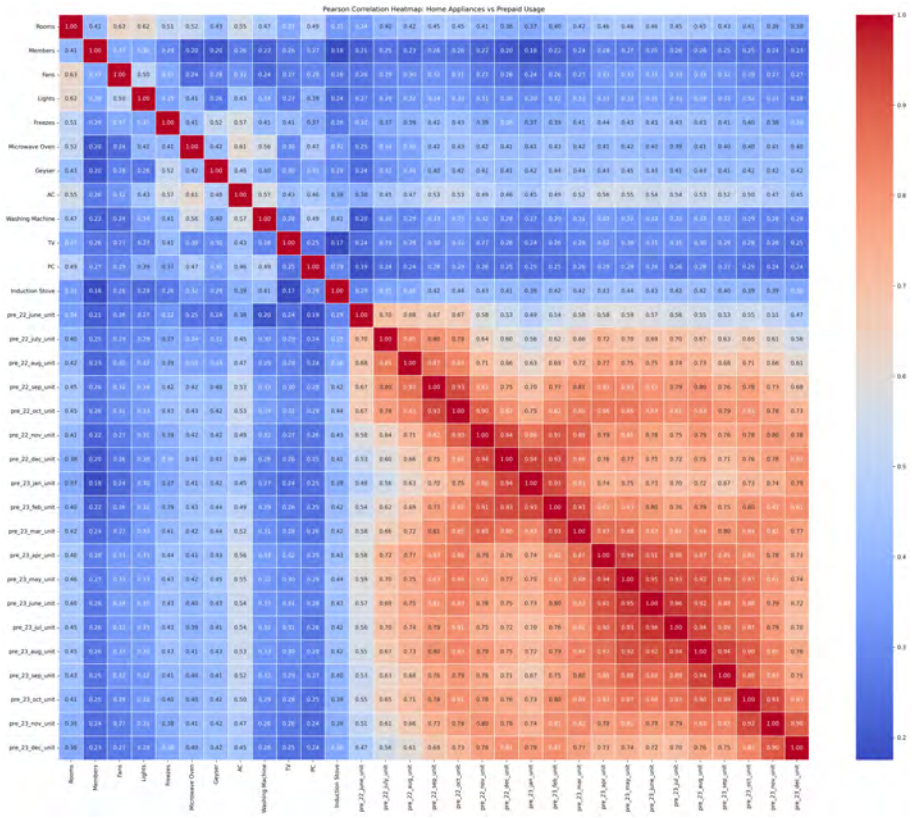

Figure 4.20: Home appliances vs Prepaid usage heatmap

Using Pearson correlation, this heat map 4.20 illustrates how household appliances affect prepaid electricity usage from Month to Month. Air conditioners have the most substantial impact on electricity consumption, and they have the most remarkable correlation of about 0.82, with peaks occurring in the summer months when they are being used. On the other hand, geysers exhibit a strong influence during most winters, commonly over 0.65. There are also low correlations with other appliances, such as freezers and microwaves (0.40 - 0.55).

The heatmap reveals that while some, like air conditioners and geysers, have a dominant seasonal impact, others constantly increase the amount of electricity used year-round. However, these variations would be hard to understand without detailed survey data on appliance usage and household characteristics.

In general, this is one way to view monthly month by month how prepaid electricity works and the nature of how it's used, notably for appliances that are in use more in

certain seasons. These correlation values tremendously validate our survey results because electricity consumption depends on the appliance used.

4.6.4 Home Appliances vs. Postpaid Usage Pearson Correlation Heatmap

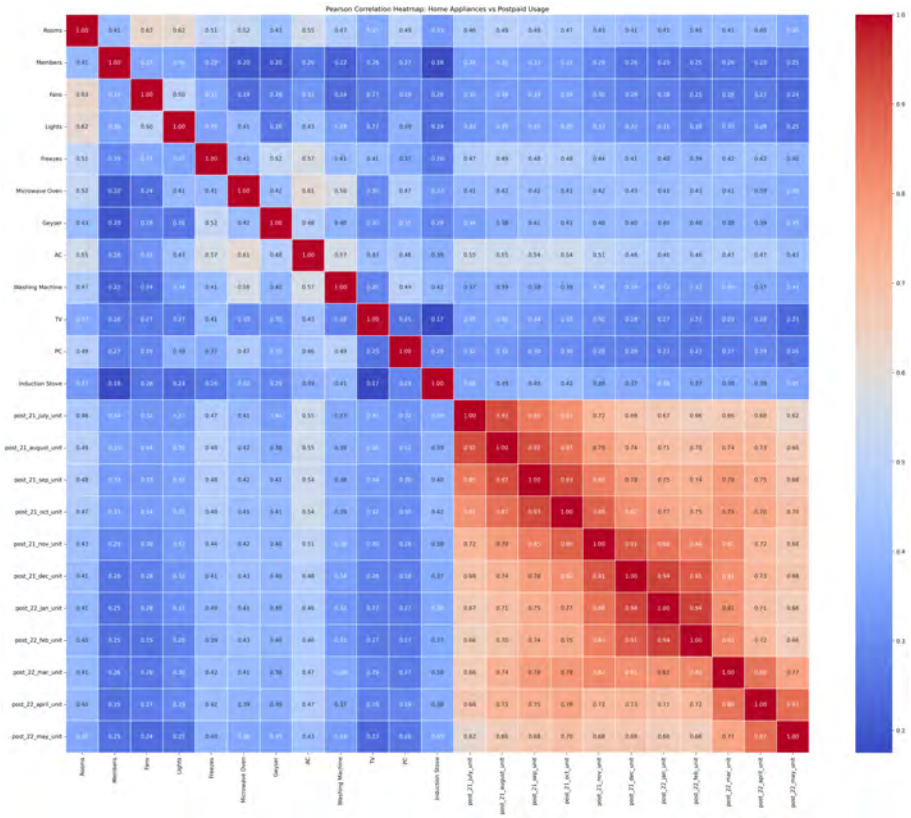

Figure 4.21: Home appliances vs Postpaid usage

This heatmap 4.21 considers how a specific appliance can affect electricity consumption for postpaid meter users, similar to prepaid users, but with less correlation overall. Air conditioners, an activity with the most decisive influence at about 0.75, are also lower for postpaid users, possibly implying that postpaid households are less focused on energy use.

The correlation to the prepaid data is similar to that of geysers and freezers, around 0.60 and 0.50. One interesting difference is that there is a slightly higher correlation for washing machines, 0.42, within postpaid households, implying that postpaid users may use these appliances more often than prepaid users.

The data collected through the survey explains the differences in access to health care. Detailed information about appliance use, household size, and room count allows us to see how these factors impact energy consumption. This correlation

analysis solidifies that the appliances listed in our survey directly affect energy usage in both prepaid and postpaid households.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This comprehensive study was pivotal in analyzing consumption patterns in the prepaid and postpaid metering systems while addressing the irregularities between them. The analytical review revealed that prepaid meters often lead to higher energy consumption, a massive concern among the masses. With the smart prepaid system, people expected to have a smooth transition and better efficiency but ended up experiencing some discrepancies. Considering the residual deviation and anomalies we have found, it is crystal clear that some potential issues exist. Last but not least, the growing concern of higher energy billings should be handled with a transparent and more efficient billing system, and through this robust research, utility companies can attain some valuable insights to improve their actions.

5.2 Limitation and Future Work

This research work has several limitations, such as due to the newly implemented prepaid metering system, there is not enough data to work on and not enough timeframe we had to consider. Moreover, the study only concentrated on a single area and it obviously does not represent the whole capital or country. As a result, we could not depict the robust consumption patterns. Hence, future research work could be expanding the field of research area and a more pervasive timeframe for deeper and more accurate understanding. In addition, future work could include exploring other time series models to verify the accuracy of predictions. Furthermore, different types of data could be added; such as customer behavior and meter calibration to assist to figure out the irregularities or anomalies.

Bibliography

[1] Z. Ismail, S. S. Baharuddin, *et al.*, "Perceptions and perceived acceptance on intention to use of prepaid meter," *European Proceedings of Social and Behavioural Sciences*, vol. 100, 2021.

[2] A. F. Aribisala and M. Mohammed, "Effect of prepaid metering system on customer satisfaction in niger state, nigeria," 2021.

[3] H. Haque, A. K. Chowdhury, M. N. R. Khan, and M. A. Razzak, "Demand analysis of energy consumption in a residential apartment using machine learning," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, IEEE, 2021, pp. 1–6.

[4] P. W. Khan, Y. Kim, Y.-C. Byun, and S.-J. Lee, "Influencing factors evaluation of machine learning-based energy consumption prediction," *Energies*, vol. 14, no. 21, p. 7167, 2021.

[5] S. Demirel, T. Alskaif, J. M. E. Pennings, M. E. Verhulst, P. Debie, and B. Tekinerdogan, "A framework for multi-stage ml-based electricity demand forecasting," in *2022 IEEE International Smart Cities Conference (ISC2)*, 2022, pp. 1–7. DOI: 10.1109/ISC255366.2022.9921933.

[6] A. González-Briones, G. Hernández, J. M. Corchado, S. Omatu, and M. S. Mohamad, "Machine learning models for electricity consumption forecasting: A review," in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, 2019, pp. 1–6. DOI: 10.1109/CAIS.2019.8769508.

[7] A. Alzoubi, "Machine learning for intelligent energy consumption in smart homes," *International Journal of Computations, Information and Manufacturing (IJCIM)*, vol. 2, no. 1, 2022.

[8] A.-D. Pham, N.-T. Ngo, T. T. H. Truong, N.-T. Huynh, and N.-S. Truong, "Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability," *Journal of Cleaner Production*, vol. 260, p. 121 082, 2020.

[9] M. Zekić-Sušac, S. Mitrović, and A. Has, "Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities," *International Journal of Information Management*, vol. 58, p. 102 074, 2021.

[10] E. García-Martín, C. F. Rodrigues, G. Riley, and H. Grahn, "Estimation of energy consumption in machine learning," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 75–88, 2019.

[11] T. Liu, Z. Tan, C. Xu, H. Chen, and Z. Li, "Study on deep reinforcement learning techniques for building energy consumption forecasting," *Energy and Buildings*, vol. 208, p. 109 675, 2020.

[12] A. B. Culaba, A. J. R. Del Rosario, A. T. Ubando, and J.-S. Chang, "Machine learning-based energy consumption clustering and forecasting for mixed-use buildings," *International Journal of Energy Research*, vol. 44, no. 12, pp. 9659–9673, 2020.

[13] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.

[14] K. Amasyali and N. M. El-Gohary, "A review of data-driven building energy consumption prediction studies," *Renewable and Sustainable Energy Reviews*, vol. 81, pp. 1192–1205, 2018.

[15] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in malaysia," *Developments in the Built Environment*, vol. 5, p. 100 037, 2021.

[16] J. Moon, J. Park, E. Hwang, and S. Jun, "Forecasting power consumption for higher educational institutions based on machine learning," *The Journal of Supercomputing*, vol. 74, pp. 3778–3800, 2018.

[17] P. W. Khan, Y.-C. Byun, S.-J. Lee, D.-H. Kang, J.-Y. Kang, and H.-S. Park, "Machine learning-based approach to predict energy consumption of renewable and nonrenewable power sources," *Energies*, vol. 13, no. 18, p. 4870, 2020.

[18] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings," *Energy and Buildings*, vol. 158, pp. 1533–1543, 2018.

[19] J.-S. Chou and N.-S. Truong, "Cloud forecasting system for monitoring and alerting of energy use by home appliances," *Applied Energy*, vol. 249, pp. 166–177, 2019, ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2019.04.063. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030626191930710X.

[20] P. Huber, M. Gerber, A. Rumsch, and A. Paice, "Prediction of domestic appliances usage based on electrical consumption," *Energy Informatics*, vol. 1, no. 1, p. 16, 2018. DOI: 10.1186/s42162-018-0035-1. [Online]. Available: https://doi.org/10.1186/s42162-018-0035-1.

[21] A. Grewal, M. Kaur, and J. H. Park, "A unified framework for behaviour monitoring and abnormality detection for smart home," *Wireless Communications and Mobile Computing*, pp. 1–16, 2019. DOI: 10.1155/2019/1734615. [Online]. Available: https://doi.org/10.1155/2019/1734615.

[22] H. A. Ã–zkan, "Appliance based control for home power management systems," *Energy*, vol. 114, pp. 693–707, 2016, ISSN: 0360-5442. DOI: https://doi.org/10.1016/j.energy.2016.08.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544216311197.

[23] E. U. Haq, X. Lyu, Y. Jia, M. Hua, and F. Ahmad, "Forecasting household electric appliances consumption and peak demand based on hybrid machine learning approach," *Energy Reports*, vol. 6, pp. 1099–1105, 2020, 2020 The 7th International Conference on Power and Energy Systems Engineering, ISSN: 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2020.11.071. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484720314967.

[24] S. Chen, G. Zhang, X. Xia, Y. Chen, S. Setunge, and L. Shi, "The impacts of occupant behavior on building energy consumption: A review," *Sustainable Energy Technologies and Assessments*, vol. 45, p. 101 212, 2021, ISSN: 2213-1388. DOI: https://doi.org/10.1016/j.seta.2021.101212. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213138821002228.

[25] B. N. Oreshkin, G. Dudek, P. Pełka, and E. Turkina, "N-beats neural network for mid-term electricity load forecasting," *Applied Energy*, vol. 293, p. 116 918, 2021.

[26] R. ELhadad, Y. F. Tan, and W. N. Tan, "Anomaly prediction in electricity consumption using a combination of machine learning techniques," *International Journal of Technology*, vol. 13, no. 6, pp. 1317–1325, 2022.

[27] S. Mahapatra and D. Golhar, "Consumers' perception and satisfaction of switching from postpaid to prepaid model in electricity consumption," *International Journal of Business, Management and Allied Sciences*, vol. 5, no. 1, pp. 239–247, 2018.

[28] K. D. Irianto, "Pre-semms: A design of prepaid smart energy meter monitoring system for household uses based on internet of things," *Journal of electronics, electromedical engineering, and medical informatics*, vol. 5, no. 2, pp. 69–74, 2023.

[29] Bangladesh Power Development Board, *Prepaid meter information*, https://bpdb.portal.gov.bd/site/page/7918c503-4c2b-43ec-96e9-cda71fbaf6b3.

[30] UNB, "Electricity: Progress in pre-paid meter installation is slow despite high hope," *Dhaka Tribune*, 2022. [Online]. Available: https://www.dhakatribune.com/bangladesh/nation/289475/electricity-progress-in-pre-paid-meter.

[31] BSS, "Over 3.85cr electricity consumers will get prepaid meters," *The Business Standard*, 2021.

[32] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, Jul. 2009. DOI: 10.1145/1541880.1541882.

[33] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[34] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018. DOI: 10.1080/00031305.2017.1380080. [Online]. Available: https://doi.org/10.1080/00031305.2017.1380080.

[35] Meta Research Blog, *Prophet: Forecasting at scale*, https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/, Facebook Research, 2017.

[36] Q. Dong *et al.*, "Short-term electricity-load forecasting by deep learning: A comprehensive survey," *arXiv preprint arXiv:2408.16202*, 2024.

[37] A. Hosseinzadeh, A. Karimpour, and R. Kluger, "Factors influencing shared micromobility services: An analysis of e-scooters and bikeshare," *Transportation Research Part D: Transport and Environment*, vol. 100, p. 103 047, 2021, ISSN: 1361-9209. DOI: https://doi.org/10.1016/j.trd.2021.103047. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361920921003448.

[38] J. S. Suri *et al.*, "Pearson correlation and statistical significance in correlation analysis," *Journal of Medical Imaging*, 2018, Explains the use of P value in correlation analysis.

[39] M. Bensalah and A. Hair, "Empowering smart cities: Sarima forecasting of power consumption in tetouan's urban grid," in *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, IEEE, 2024, pp. 1–6.

[40] F. J. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951. DOI: 10.2307/2280095.

[41] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*, 5th. CRC Press, 2011.

[42] W. J. Conover, *Practical nonparametric statistics*, 3rd. John Wiley & Sons, 1999, ISBN: 978-0-471-16068-7. [Online]. Available: https://dev.store.wiley.com/en-ca/Practical+Nonparametric+Statistics%2C+3rd+Edition-p-9780471160687.