Hybrid model for predicting energy behaviour of Prosumers

MSc Research Project
Data Analytics

Sreelakshmi Chittazhi
Student ID: x22210466

School of Computing
National College of Ireland

Supervisor: Ahmed Makki

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sreelakshmi Chittazhi
Student ID:	x22210466
Programme:	Data Analytics
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Ahmed Makki
Submission Due Date:	12/08/2024
Project Title:	Hybrid model for predicting energy behaviour of Prosumers
Word Count:	8700
Page Count:	21

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

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

Signature:	Sreelakshmi Chittazhi
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Hybrid model for predicting energy behaviour of Prosumers

Sreelakshmi Chittazhi
x22210466

Abstract

The current study seeks to contribute to the solution of the problem related to the determination of imbalance in prosumer networks, as this is critical for increasing the efficiency of energy management, and therefore, for decreasing the reliance on traditional power supply. This problem is relevant because the appropriate energy management contributes to improvement of energy systems, the reduction of carbon emissions, and the availability and stability of energy supply. To address this, developed a hybrid model that integrates the weather data and market price data with advanced machine learning techniques to enhance the predictions of energy imbalance.Eventhough the base model LightGBM model(18.15) and hybrid model with meta models linear regression and ridge regression had similar MAE value (18.20) which suggest its effectiveness in managing energy imbalances in prosumer networks. The paper's contribution to the existing knowledge is in showing that using multiple data sources and implementing contemporary machine learning techniques improve accuracy of prediction. This work contributes to the existing knowledge on how the hybrid models can help manage the fluctuations of renewable energy production and prosumers' actions.The issues of the applicability of the model to other geographical areas and various patterns of prosumer activity are left unsolved. Further investigations of the presented framework are required to enhance and expand its usage for different contexts.

1 Introduction

1.1 Background

More people these days are turning their attention to the idea of producing their own electricity known as "prosumers".These individuals both generate and consume energy. The prosumers are the most essential elements of the contemporary energy market since they both produce and consume energy. There are several ways through which prosumers can generate their own power . For example, photovoltaic systems (solar panels) and wind turbines. This enables the prosumers to minimize the traditional energy sources to a big extent and help in enhancing the environmental standards. Energy demand is expected to increase by 12 % in the global scene from the year 2019 to 2030 as explained by the World Energy Outlook 2020 report (International Energy Agency, 2020)[1]. This has therefore resulted in increased demand for electricity that has negative effects on the environment hence the need to seek renewable sources of energy.

[1] https://www.iea.org/reports/world-energy-outlook-2020

However, this form of energy generation is characterized by fluctuations because it depends with the weather and this results to energy disparities. Previous studies has concentrated on individual aspects of renewable energy generation or consumption. Even though there are studies based on the energy prediction of prosumers, there but there is a lack of comprehensive studies that predict prosumer behavior to manage these imbalances effectively. Addressing this gap, this study aims to develop a hybrid model that integrates advanced machine learning algorithms to enhance the accuracy of energy imbalance predictions.

1.2 Motivation

Focusing on the sustainability theme in global energy provision, the shift towards the use of renewable energy has been deemed essential. Energy Agency's new report (2021) revealed that the possibility and probability of the renewable energy sources (RES) to attain the global capacity of the level of 55 % up to the year 2050 has been predicted[2]. This transition is meant to fasten the achievement of the goal of lower carbon emission and combat climate change. Renewable energy of the sun and wind is more accessible than fossils and causes less depletion of the environment. This transition includes improving energy efficiency as it helps the system to use less energy hence it has a minimal effect on the environment.

Both corporations and consumers play vital roles in this transition. Corporations are investing in renewable energy technologies and improving their energy efficiency to reduce operational costs and meet regulatory requirements. While customers keep seeking new supplies of energy in the form of renewable energy to use in their households and for their enterprises because of efficiency and possibly cheaper prices. However, the fact that the unpredictable energy behaviour of prosumers poses significant challenges.

This research aims at identifying the energy behavior patterns of Prosumers with a generalized hybrid model that will employ improved Machine Learning techniques and other methods to increase the efficiency of energy imbalance predictions (Benti et al.; 2023). It also helps level the demand and supply balancing problem of prosumers who are both consumers and producers of electricity to tackle unpredictability in energy management and optimization. Furthermore, the research proposal effectively addresses the major research gaps identified by (Mathumitha et al.; 2024) by incorporating additional data features (weather and market prices).

1.3 Research Question

How well a hybrid model can incorporate weather forecast data and market prices to predict energy imbalances for prosumers more effectively?

1.4 Research Objectives

Following are the key research objectives:

1) Conduct exploratory data analysis on the data.

2) Feature Extraction

[2] https://www.iea.org/reports/global-energy-and-climate-model/net-zero-emissions-by-2050-scenario-nze

3) Implement the base models (Light Gradient Boosting Regressor and Random Forest Regressor).

4) Implement the hybrid model by stacking the base models using a meta model.

1.5 Limitations

The challenges experienced in this study revolve on several elements. First, it is important to note that the paper relies on the data from Estonia and focuses on types of prosumers, thus, the conclusions can be applied only to the Estonian context and certain technologies. Also, the primary data used in the model, the weather forecasts, contain variation and uncertainty and hence less precise. Sometimes, there is no accurate recognition of fast-growing technologies and certain changing behavioral patterns of the prosumer, therefore constant model updates are required. However, these discrepancies raise issues in relation to the generalization of the findings due to variations in the regional regulations. Finally, there may be difficulties in comparing data collected from different climatic and socio-economic environments and conditions.

1.6 Structure of the Report

This document is structured as follows: Section 2 describes about the literature review, it discusses various studies conducted on predicting the energy behaviour of prosumers as well as consumers.The methodology of this research work and the specification of the proposed framework for this study are presented in Section 3 & 4. The implementation of the framework is detailed in Section 5. Finally, Section 6 and section 7 discuss about the evaluation of the model, conclusions derived from them and future work.

2 Related Work

2.1 Hybrid modelling in energy prediction

The paper aims to solve the problem of accurately predicting building energy efficiency using machine learning (ML) techniques. They specifically focused on evaluating the effectiveness of ensemble methods like bagging and boosting. A novel hybrid stacking ensemble approach combining the best-performing bagging and boosting methods was proposed (Egwim et al.; 2024). The extra trees model was the best-performing single algorithm, with an adjusted R-squared of 0.93 and an RMSE of 2.79, while the hybrid stacking ensemble achieved the highest accuracy with an adjusted R-squared of 0.9487. Limitations include the specificity of data transformation to the UK, high computational complexity of the hybrid model, and challenges in model interpretability. Despite these limitations, the study demonstrates that ensemble methods can significantly enhance predictive accuracy compared to single ML algorithms. Another similar study Chandran and Narayanan (2024) which focuses on the problem of imbalance in energy generated and consumed by prosumers in the grid threatening the stability of the grid and financial stability of electricity companies. The researchers tested the efficiency of such machine learning models as linear regression, multilayer perceptron, TabNet, XGBoost, LightGBM through a VotingRegressor method. They found out that there is a gap in the literature as regards to the development of models that simultaneously forecast energy demand, and supply. Thus, the best result was obtained with the use of the ensemble model with the MAE of 70. 16. The study also recommended a target normalization pipeline to enhance the model's performance and other important prediction variables which include lagged targets and Installed capacity.

In the field of energy usage prediction, a multi-stage energy estimating model for prosumers has been presented in (Antal et al.; 2022). The authors of this paper have designed the mentioned model to predict the domestic prosumers' energy value of each hour of the subsequent day related to past energy usage data, weather factors, and additional exogenous variables. The model is a combination of classifiers and regressors with the neural network of the multilayer perceptron and thus uses K-means clustering to classify the energy peak and valley. The model's salient features include its efficiency in handling data heterogeneity and enhancing forecast quality through the incorporation of past and future information. Moreover, the accuracy is dependent on big historical data and data preprocessing, which may be time-consuming. The accuracy metrics including mean absolute error and rmse showed a value of 0.15 and 0.22 . A research gap identified in this particular study is the model's performance is reducing over longer prediction periods, which indicates a need for further enhancements to handle extended forecasts more effectively.

Another significant contribution the establishment of an energy consumption forecasting model known as ISCOA-LSTM which stands for improved sine cosine optimization algorithm and Long short-term memory networks in (Somu et al.; 2020). The advantages of this research include high accuracy and reliability of this research, shown on different measures: MAE 0. 13, MAPE of 4. 5%, MSE of 0. 12, and RMSE of 0. 19. Thus, the drawbacks of the model are the lack of proper preprocessing and the fact that the parameters have to be adjusted based on the characteristics of real-time data which implies that it requires further tuning.The research gap here involves the preprocessing and real-time applicability, suggesting future work should focus on refining these aspects to enhance the model's practical utility.

Some works on energy load forecasting suggests that traditional ML techniques such as Support Vector Machines and Gradient Boosted Decision Trees Xg-Boost (Chen and Guestrin; 2016) and Light G-Boost (Ke et al.; 2017), particularly when there are more features that are easy to predict on.(Bagherzadeh et al.; 2021) has conducted a work predicting the power consumption of the Melbourne wastewater treatment plant. Compared to various machine learning models such as, recurrent neural network (RNN), random, forest (RF), perceptron (ANN), and the gradient boosting machine (GBM), it was found that the later one gave the best performance on the test set.

Advancements in short-term energy forecasting frameworks have been observed with the ensemble deep learning methodology identified in (Ishaq et al.; 2021) where the authors proposed a hybrid model with CNN layers for spatial features and bidirectional LSTM layers for temporal-based short-term energy forecasts. This model outperforms previous models in capturing spatial-temporal pattern motion, providing a Mean Absolute Error (MAE) of around 0.12, Root Mean Square Error (RMSE) of 0.18, and Mean Absolute Percentage Error (MAPE) of 3.8%. However, the model's complexity may lead to higher computational costs and require extensive data preprocessing, highlighting a research gap in implementing efficient algorithms that maintain accuracy while reducing computational costs. Another study outlined in (Syed et al.; 2021) and (Mathumitha et al.; 2024) presented a new deep learning network based on stacked bidirectional and unidirectional LSTM for precise energy consumption prediction at the household level. These models, incorporating LSTM, CNN, and other ensembling techniques, focus on improving forecasting capability. While the models gave good results, their complexity requires more computational resources for execution and suffers from issues related to explainability, limiting their practical application.

To a certain extent, it appears that people have not agreed on which pre diction model is better, some stated that the deep learning techniques such as ANNs and LSTMs are superior to others, whereas some individuals asserted that the more conventional ML models such as SVR and gradient boosted machine offers superior performance to ANNs. This implies that mod else are influenced by different parameters, one of which is; the type and quality of data on which the mod else are based.These studies include useful information as well as methods that is applicable to the research proposal on developing a hybrid model for predicting the energy behavior of prosumers. The stacking method discussed in one of the above papers will be useful the design and implementation of the proposed hybrid model. Also the use of advanced optimization technique such as Bayesian optimization has been implemented to get the optimal parameters which will be crucial in improving the accuracy. Based on the strengths and limitations identified in these studies, the proposed research aims to improve energy forecasting capabilities, thereby minimizing energy imbalances and improving decision-making for energy companies.

2.1.1 Renewable energy and sustainability environment

The paper by Gajdzik et al. (2023) focuses on capturing and analysing the pattern of such prosumers who are using photovoltaic panels and heat pumps in order to fulfil the increasing demand of energy and to shift towards the renewable energy sources. The data was collected from 326 Polish prosumer households and the researchers used the Computer-Assisted Web Interviewing (CAWI) method to investigate the prosumers ecological behaviours and consumption of energy. The study highlights a gap in comprehensively examining the differences between photovoltaic system users and heat pump users. Also, the measures explaining prosumers' data sharing could be limited due to the fact that the study relied on survey data, which may lead to self-bias in the data collection.

There are several studies that shows the relation between the prosumers and renewable energy. As suggested by Hu and Chuang (2023) it give more insights about the importance of transition towards renewable sources of energy among prosumers and how this can shape the future energy system. It also discusses about the challenges and opportunities posed by this transitionLeal Filho et al. (2024). In Accouche and Gangadhari (2023) the authors proposes an improvement of energy decision making for a university microgrid in Qatar through the incorporation of renewable energy sources. The energy mix model is designed using a mixed integer linear programming (MILP) with MATLAB for optimization of the proportion of solar PV systems, wind turbines, and diesel generators focusing on lowering the operating expenses as well as the emissions of greenhouse gases. The study has limitations like using an artificial market environment and using cost as the outcome measure only while neglecting the implications of the results on financial sustainability over the long term as well as policy implications.

These studies are relevant for this particular proposal in finding information on the variables and optimization techniques concerning energy behaviour prediction in prosumers as well suggest the importance of prosumers role in sustainable environment and ; the study points out areas for future research and model improvement.

2.1.2 Effect of prosumers in energy market prices

Many papers suggest that the prosumers reduce greenhouse gas emissions but augment grid vulnerability and price of electricity to non-prosumers. In response to this the Lee et al. (2023) suggests another way of subsidization . Limitations involve the application of theoretical models as well as inadequacies in the long-term operational sustainability and regional differences. There are various studies that have examined the impact of prosumers on energy market prices. For instance, Vergados et al. (2016) in their research their main objective was to minimize the total energy costs for prosumers by reducing the errors in predicting their usage of electricity and production. The authors grouped the prosumers as one unit in the energy market by using the virtual clusters which ultimately reduces the cost of energy. Furthermore, this had many advantages including an increased efficiency and cost savings.

Increase in use of renewable resources is at a very fast pace today. One of the most recent works focused on nine European countries and their current legislation and regulations regarding prosumer engagement to market prices Inês et al. (2020). This allows establishing the major threats and opportunities of a country for prosumers. Some of the European countries such as France, Germany, the Netherlands, and the United Kingdom have relatively friendly regulations in comparison to others as observed. Based on the findings of the research, the following recommendations are useful in enhancing these regulations for example, incorporating particular communities and poor families in the transformation toward renewable energy also to benefit from the change.

There have been pushes made in the existing literature to examine how prosumers influence the energy markets with recognizing trading prices, implementation, and cost. In (An et al.; 2022), genetic algorithms are utilized to identify certain optimal prices in South Korea, which are introduced with regard to greater price spans within higher electricity consumption and more number of prosumers. Looking at the specifics and applying the logic that lower technology costs coupled with higher electricity prices will onset going off-grid by 2030, it will be profitable for prosumers in the Ontario's case, it is critical to note energy policies (Kuznetsova and Anjos; 2021). In (Yu et al.; 2019) the authors unveils an efficient bidding strategy and; besides, proposing a decentralized market clearing scheme, too, which is a bang in the holy war against operational costs. Altogether, these works give information about economical approaches, policies, and optimisation regarding prosumer level.

3 Methodology

This research follows the KDD which is known as Knowledge Discover in database. This include steps such as: understanding the research problem, data gathering, data preprocessing, modelling and finally, evaluation. Through this structured approach, the author has made sure that every stage of the research was well planned and carried out, thereby producing useful insights.

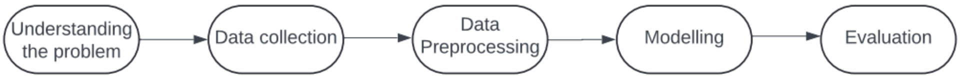

Figure 1: Research methodology

3.1 Understanding the Problem

The main goal of this research is to predict the energy behaviour of prosumers. By predicting both energy consumption and production accurately, the model can help in mitigating the energy imbalances that cause operational and financial challenges for energy companies.The study seeks to combine the weather data and market prices into a hybrid model combining machine learning algorithms to enhance prediction accuracy and provide better insights into the factors influencing prosumer behavior.

3.2 Data understanding

In this step, the dataset was obtained from an open source platform, kaggle. The dataset is basically of Estonian region. The dataset include 5 different files and those are:

- train.csv: In this dataset it contains information related to prosumers energy consumption and production. It contains the target variable that is to be predicted.
- client.csv: This contains details about the he aggregated number of consumption points (EICs - European Identifier Code) and installed solar panel capacity in kilowatts.
- gas_prices.csv: This provides historical and future natural gas prices.
- electricity_prices.csv: This data gives information about the day-ahead market prices.
- forecast_weather.csv: This data gives information about the weather information like cloud_coverage, temperature and so on.

3.3 Data Pre-processing

Data pre-processing is a significant step in research methodology, its primary purpose is to make the data free from errors, compatible for analysis and modeling.In this stage the authors will be conducting the preprocessing steps like exploratory data analysis, feature extraction, data cleaning and son on to get insights about the different datasets metioned in the above step.

3.3.1 Exploratory Data Analysis

During this stage, it helps to understand the structure and characteristics of the dataset through several ways such as visualizations, finding the null values, checking the datatype and so on. The first step is to check whether dataset is having any null or missing values as well as duplicated values. If the null observed was too less the author has drop those null values. No duplicated values was found for any of the dataset. Some of the important insights found during the analysis will be discussed in this section.

Figure 2, below represent the count plot of prosumers which indicates whether they are using it for business purpose or non-business purpose.(0 represents non-business entities and 1 represents business entities).

It is clearly visible that more data points are classified as business than those classified as non-business. Knowing the distribution of business status in the dataset can be relevant for the further analysis.

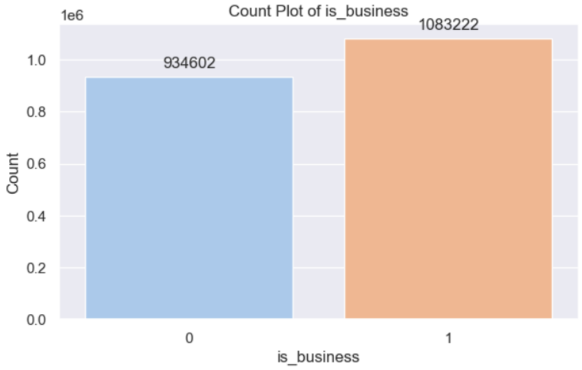

Figure 2: Bar chart of is_business

Figure 3 represents the proportion of product type that has been used by the prosumers. Each product type represents (0: "Combined", 1: "Fixed", 2: "General service", 3: "Spot")

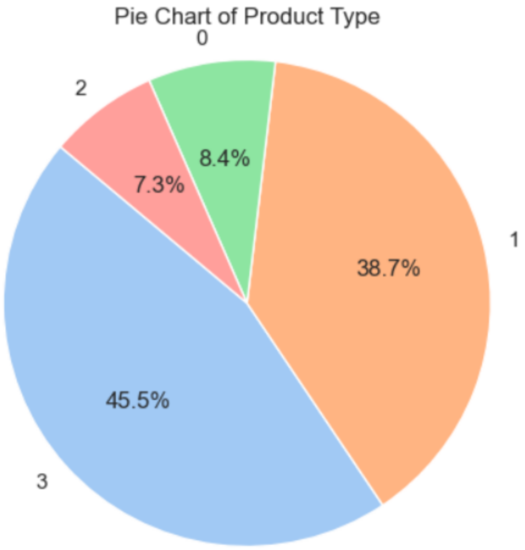

Figure 3: Pie chart of product type

The product Type 3 has the largest proportion, making up to 45.5% of the dataset. It is the most prevalent product type, while the product Type 1 is the second most commonly used constituting 38.7% of the proportion. Product Type 0 and product type 2 are the least used products as compared to the others.

From figure 4 and 5 it is noticeable that there is a pattern in energy consumption and production. During the autumn/winter months (October-February), production goes down while consumption goes up. This can be explained by people and businesses using more energy for heating, and there is less sunlight during this time. In contrast, during the spring/summer (March-September) months, the opposite happens. There is an increase in energy production due to higher temperatures (less need for heating) and more sunlight, leading to a decrease in energy consumption.

Figure 6 represents the energy for each product type. Even though it was observed that the product types 1 & 3 are the most used for counties, we see here that the types

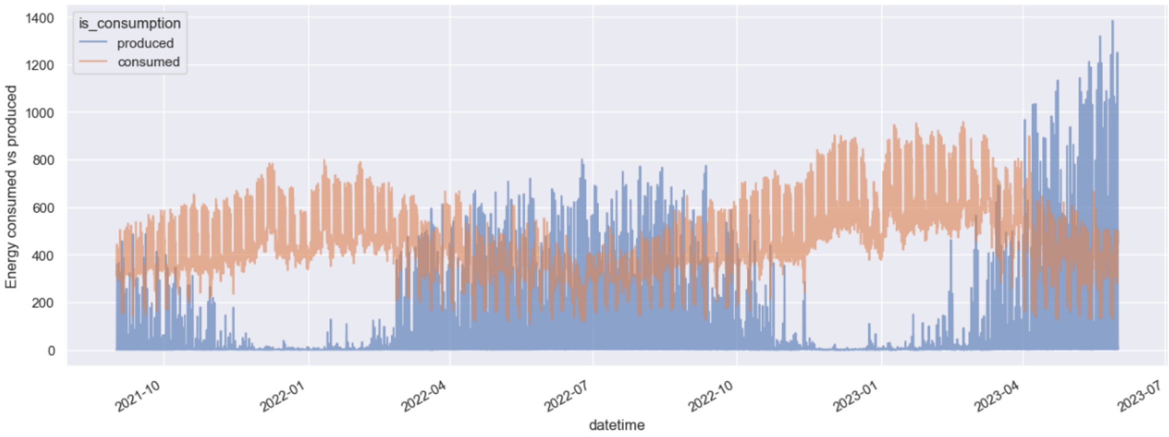

Figure 4: Visual of energy consumption and production

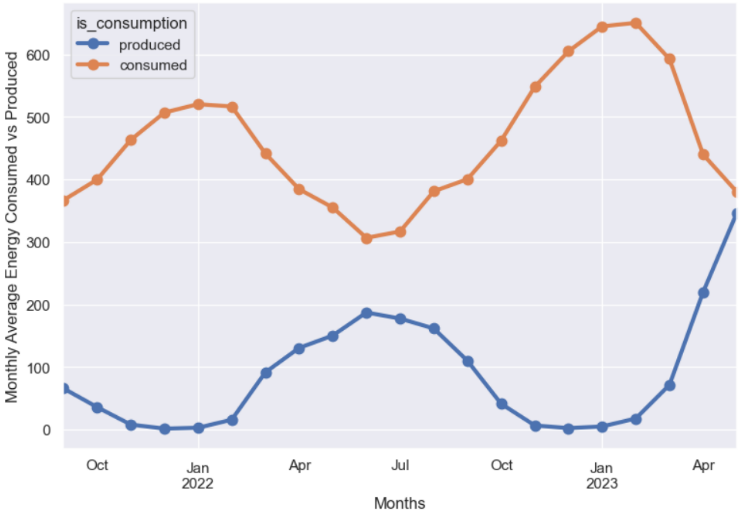

Figure 5: Monthly average energy consumption and production

0 & 3 are the one that participate more in the energy consumption/production while the the types 1 & 2 doesn't contribute as much especially type 2.

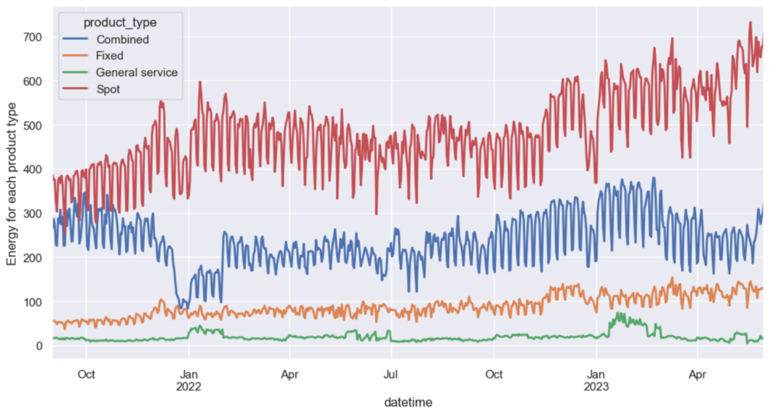

Figure 6: Energy for each product type

It is observe that the solar radiation received and temperature drops in the winter/autumn periods which explain perfectly our previous observations concerning the pattern in the energy production vs consumptions. Also the clouds are more present in that period which is normal.

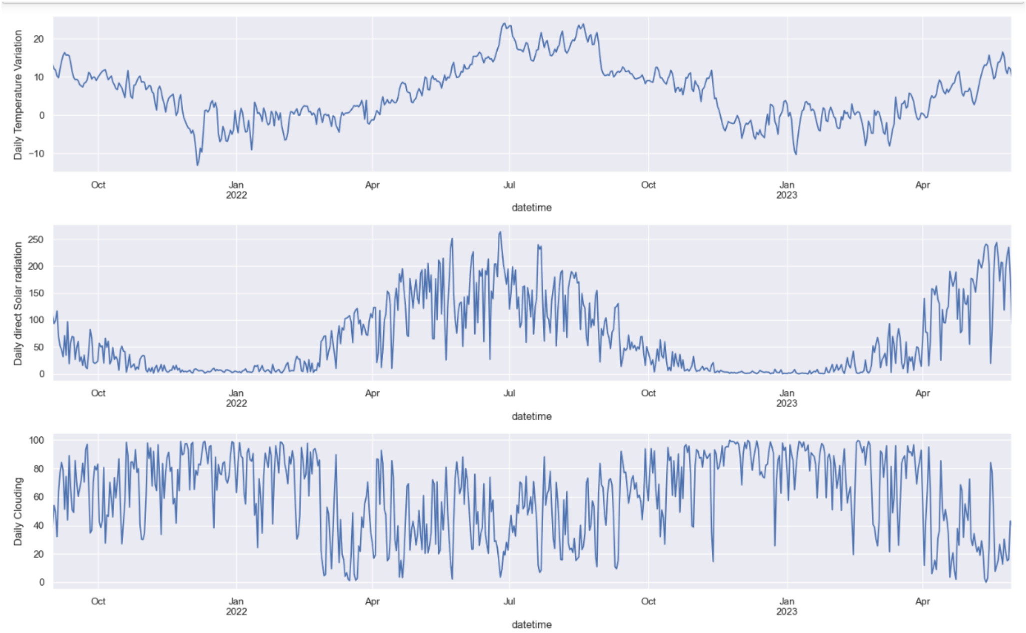

Figure 7: Pie chart of product type

3.3.2 Outlier removal

Outlier handling is one of the crucial step in data preprocessing because this can impact the the final results. Only in one dataset the outlier was observed and to remove the outlier the author has used the technique named winsorization. It is a technique which is used to trim extreme values in the data to reduce the impact of outliers.

In the elecrticity_price.csv data, the prices column was having negative values and those values has been filtered out from the data since we do not require those values. As well an outlier was found out with the help of boxplot chart.

From the above boxplot it is clear that there is a significant outlier at approximately 4000 for the prices of electricity. Moreover, there are numerous outliers between 500 to 1000 range, but it indicates it is not extreme values as that of 4000. In this context with the help of winsorization the author has removed the outlier.

The main reason to select winsorization apart from other outlier removal method is that winsorization because it does not reduce the sample size and maintains the overall distribution of data. While trimming method involves eradicating outliers altogether, and actually decreases the statistical sensitivity, Winsorization alters the extreme values to the closest percentile stated. This approach eliminates the effects of outliers without completely eliminating majorities of possible useful data points, which makes the data analysis stronger and highly resistant.It is also a better approach as compared to the Z-score or IQR techniques where some of the values may be either too much smoothened or some data values are completely removed from the dataset while Winsorization modifications rather than deleting the extreme values. Because of this, it is suited to be used

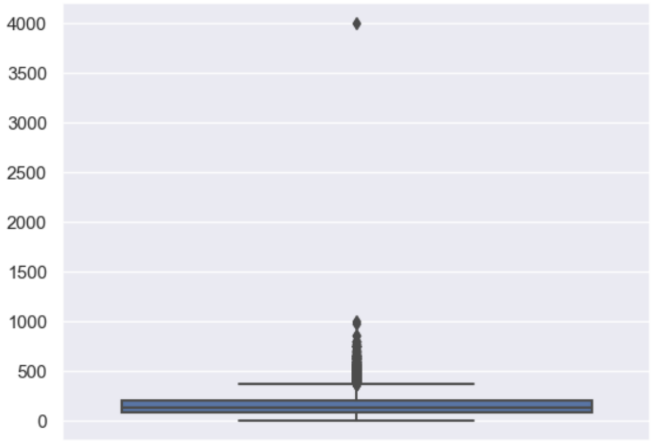

Figure 8: Pie chart of product type

when preprocessing data to guarantee that the achieved statistical analyses, and ML algorithms, become both robust and accurate.

3.3.3 Feature Engineering

Feature engineering involves adding new features from the dataset In this section certain features like hour of the day, day of the week or month of the year as well as average electricity price has been derived which can provide valuable insights for the data. Finally, the author merges the dataset for the modelling part.

3.3.4 Feature selection

The author has implemented SelectKbest method to obtain top 15 important features. This is a technique that select top k features based on a scoring function. Mutual information regression is the scoring function which the author has used with Selectkbest. It basically measure the dependency between two variables.

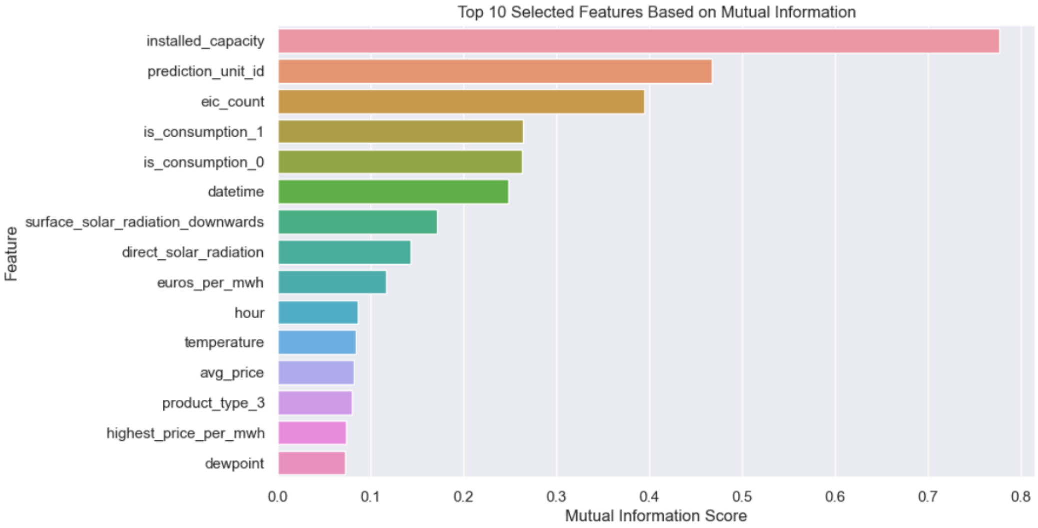

Figure 9: Feature selection

3.4 Data Modelling

This section outlines the methods used to implement the model for predicting the energy behaviour of prosumers. This particularly involves selection of best algorithms, training the models with the train data and using the test data to know how well the models perform and finally, integrating them into a hybrid model.

In this research the author is implementing a hybrid model by integrating two base models by stacking method. The two base models used for this problem are Light Gradient Boosting Machine(LightGBM) Regressor and Random Forest regressor(RF).

After the preprocessing stage, the data is cleaned and is ready for the modelling part. Initially, the data is splitted into training and testing data, 80% of the dataset is used as training data and rest 20% is used as testing data. The performance of the model will be evaluated on the testing data. Later on, with hyper parameter tuning method the best parameter for Light GBM model will be obtained and those parameter will be applied while building the two base models.

Once the two base models are trained, then with the help of test data the performance of the Light GBM and RF are evaluated. Finally, the hybrid model is implemented by stacking the predictions of these two base model to a meta model and thus giving the final predictions. The meta model used in this research is Linear Regression, but to compare the results with other meta model as well the author has used Ridge Regression and Gradient Boosting as meta model. These meta models has several advantages over other models such as linear regression gives a clear understanding about the features and the target variable, whereas the ridge regression comes with a layer of regularization, which is helpful in the case of multicollinearity and reducing the effect of much noise in many variable datasets. Gradient boosting, on the other hand, uses high levels of predictive power and has the ability to be flexible enough to capture what linear models cannot (complexity and non-linearity).

3.5 Evaluation

The evaluation section explains how the performance of the developed models was assessed. It invloves evaluating the individual base models (LightGBM Regressor and Random Forest Regressor) as well as evaluating the hybrid model. It is aimed that the models should give accurate, and robust predictions of the prosumers energy profile.

In this research the performance metrics used is Mean Absolute Error, due to its simplicity, interpretability and robustness to outliers. Absolute errors is the difference between the predicted values and the true value and Mean absolute error is the average of these absolute errors. It is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Here,

- n is the number of observations
- $\hat{y}_i$ is the predicted values
- y_i is the actual values

Unlike RMSE, which reduces large errors more than small ones, MAE does not do this. It treats all errors uniformly and thus provides a balanced view of model performance. This constant weighting of error makes MAE a robust metric . Also, MAE does not employ the condition of normality of the errors that is an advantage since the method can work for any type of distribution. These measures are then described in the same metric as the data used for their calculation, which makes it easier for the audience, including business people with no or limited statistical background, to understand the performance of the model. Due to its ability to offer an exact guide of the average prediction error, MAE aids tangible decision processes in different fields and improves the assessment of the models used within them.

4 Design Specification

Figure 10, shows the proposed design to conduct this reserach.As mentioned in the methodology part, Stage 1 is the data preparation, in which data is taken from kaggle and are subjected to data cleaning and exploratory data analysis (EDA). Here, data cleaning involves handling missing values, outliers. After this, in feature engineering the new variables are created from the raw data and in feature selection, important features are selected to improve the model performance.

Stage 2 explains how the hybrid model was implemented with the two base models. Initially the preprocessed or cleaned data is splitted into train and test data. Here the base models used are Light Gradient Boosting Machine Rgeressor and Random Forest Regressor.

Light Gradient Boosting Machine Regressor is an optimized Gradient boosting framework belonging to tree-based learning algorithms. It is designed to handle large dataset and has the benefits of faster training rate and requires lower memory space as compared to other gradient boosting algorithms such as XGBoost and thus producing accurate results Ke et al. (2017). It gets these advantages with features like histogram-based decision tree learning that accelerates the training, and leaf-wise tree growth, which optimizes for lower loss. This is especially useful when there are great demands for predictive outcomes with the capacity to work on both, numerical and categorical values.

On the other hand, Random Forest Regressor offer robustness and simplicity, it builds numerous decisions trees with a randomly selected datasets to minimize overtraining effects and improve the model accuracy Probst et al. (2019). This is because the ensemble approach is the way boosting the trees so the predictions from individual tree are averaged, this allows stability and accurate results in cases of noisy data. Thus, more complex models such as Support Vector Machines (SVM) for instance, tend to take longer to process and can pose a major hurdle to the inclusion of big data especially due to the amount of preprocessing required. Neural networks, although are very effective, sometimes fit the model extremely close to the training data which they are supposed to model and are computationally intensive and have to be fine tuned and sometimes are not very feasible for use. Decision Trees can be interpreted because of their easy and simple structure, but they have a higher chance of overfitting small data set and are not as accurate as Ensembling methods. Thus, integrating these two models, this algorithm captures LightGBM's fast speed and high accuracy with Random Forest's stability and easy interpretation, providing a better and more accurate working model.

Due to these reasons the author has finalised with these two base models. While

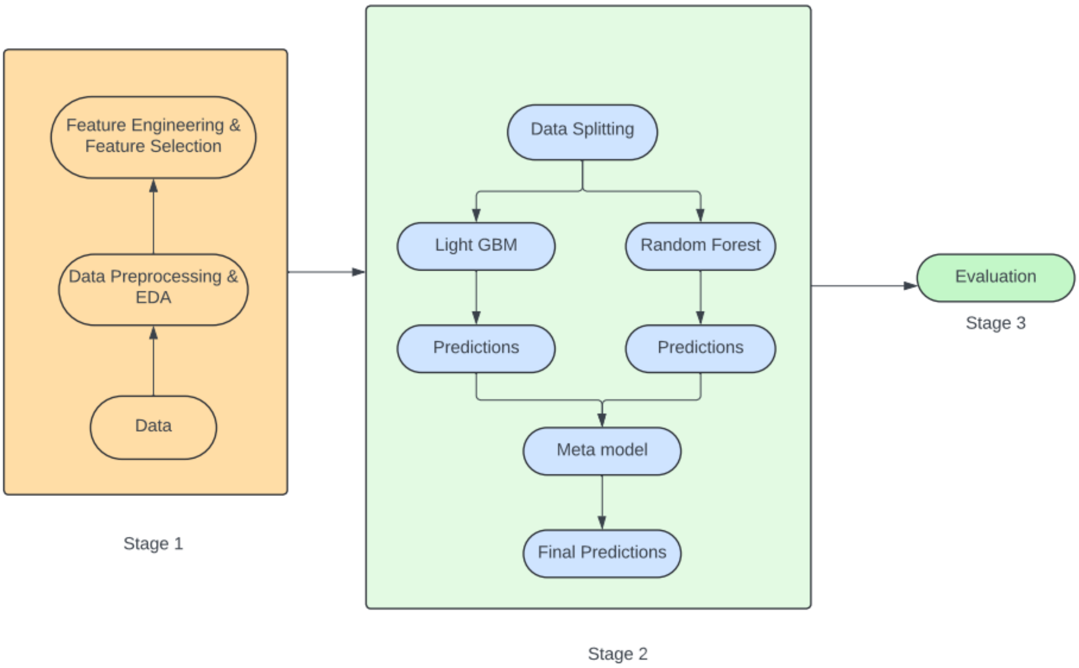

Figure 10: Design Specification

implementing the hybrid model, the author has used a technique called stacking. Once the models are trained, predictions are generated from LIGHT GBM and RF. The hybrid model is computed by combining these predictions into an array to be used as the input to the meta model. The main purpose is to aggregate the prediction of multiple models to final predictions. The meta model gets to know the proper way of combining outputs of the base models in order to enhance accuracy and reliability of the prediction. The final trained meta model which is a linear regression, is trained on this combined dataset to learn in an optimal manner how the predictions from the two base models should be used. Lastly, the meta model is applied to produce the final predictions given the LightGBM and Random Forest models' strengths to reach near-optimal performance and models stability. This makes the final result to be more accurate and reliable as compared to the final result from individual models only.

The stacking method was selected to construct a hybrid model over other ensemble methods because of its competence in utilizing the learning attributes of other different types of predictive model while simultaneously compensating their inefficiencies. Unlike bagging and boosting where new learners are generated from the same type of learner as the ones being used, stacking combines several different kinds of algorithms, each highlighting different aspects of the patterns in the data. This heterogeneity enables the meta-model to learn complex interactions and dependencies that single-model ensembles might miss. Besides, stacking help mitigate overfitting since it combines the diverse conjectures of the base models, thus improving the generality over new data. This characteristic makes it very useful in dataset which have more than one type of feature and complex interconnections.

5 Implementation

All the implementation was done in python using Jupyter notebook.

Importing libraries and data: In this research, the author initially started with loading the important libraries for data manipulation, data visualizations and data modelling such as pandas, numpy, seaborn, matplotlib.pyplot. Certain other libraries such as SelectKBest, mutual_info_regression for feature selection, optuna and lightgbm and RndomForest Regressor for hyper parameter tunning and modelling. All the data has been uploaded from csv file. As the next step for the inspection of the data first few rows of the data has been displayed to understand the structure. later on, data cleaning has been carried out by checking for duplicated rows, the shape of the dataframe is checked to find out the dimensions of the data.

Data Cleaning and preprocessing: As the next step, the null values are examined and the percentage of the nulls is then computed and the rows with nulls are removed because the percentage of nulls is very low. This step helps in cleaning the collected data to remove any unnecessary and missing values which may cause an issue in the further analysis process. Finally, details such as name of columns and their attributes including data type is obtained by info(). As a result, for accurate time-based operations, the 'datetime' is converted to a datetime object with UTC time zone, in order to align with timestamp format of the data.

Data exploration and Visualizations: Various visulations has been created from all the dataset which is already mentioned in section 3. this section is basically done to get an insight and useful information from the data.

5.1 Model Training and Building

To implement a hyrid model, initially the author has implemented two base models Light Gradient boosting Machine Regressor and Random Forest regressor in the dataset. To build Light GBM model, the hyper parameter tuning is done. The primary goal for doing hyper parameter tuning is to get optimal parameters for the two models to achieve best model performance. For this purpose the author has used Optuna framework.

5.1.1 Implementaion of Light Gradient Boosting machine Regressor

For developing Light GBM model, initially the hyper parameter tuning has done to get optimal values for the parameters. It specifies certain parameters such as including the number of estimators (n_estimators), learning rate (learning_rate), number of leaves (num_leaves), maximum depth (max_depth), and subsample ratio (subsample). An Optuna study is created to minimize the MSE, and the study.optimize method runs the optimization process over 10 trials. Every trial in the optimization process uses a different set of hyperparameters which are suggested by Optuna. Finally, the obtained best tuned hyperparameters are retrieved in order to define the LightGBM model's best hyperparameters to get the lowest MSE value for the validation set. It helps fine-tune the model on the given data set so as to improve its predictive capability on the given data.

There are 19,337 boosting iterations (n_estimators), 22 as the maximum tree depth, and 77 finalised leaves per tree, thus making it capable for capturing complex relationship in the data. It will use 80% of the data set and each tree subsample or data partition, using a learning rate of roughly 0.08 to regulate the rate of learning Section. The training process is monitored through two callbacks: one for early stopping to stop the training process should the model is not improving on the validation data for 1,000 iterations and another one to print the model evaluation on the validation data on every 1,000 iterations. The model is trained by providing the x_train ad y_train and the evaluation is done test data that is x_test and y_test. The evaluation metrics Mean absolute error has been found after making the predictions with the actual values.

5.1.2 Implementaion of Random Forest Regressor

For implementing Random Forest Regressor, hyperparameter tuning was not conducted to avoid taking up a lot of computational time. An important aspect of the Random Forest method is deciding to use bootstrapping for constructing multiple decision trees which makes it less sensitive to overfitting and usually results in good preliminary level of performance even when the parameters are left at their default values. Hence, parameters such as n_estimators, max_depth, min_samples_split and min_samples_leaf were set to meaningful values to avoid model over fitting. The values used for n_estimators = 50 and max_depth = 20 used the number of trees and maximum depth of the trees, in order to prevent overfitting the model and to reduce computation time. Moreover, when setting oob_score=True it allowed for the model assessment without the use of cross validation techniques. Random Forests are rather stable and do not require tuning. They are good to use in most cases with default parameters to compare with. This approach helped in quickly determining the baseline performance which was followed by optimizing tuning efforts on LightGBM as it is relatively more sensitive to hyperparameters because of its capability to handle the intricate aspects of data interaction.

The parameters used for implementing random forest are n_estimators=50, max_depth =20, min_sample_split=5, min_samples_leaf=2, max_features= 'sqrt', bootstrap=True, oob_score=True, n_jobs=-1, random_state=42, verbose=1, ccp_alpha=0.01.

5.1.3 Implementaion of Hybrid model

There are different ways in which hybrid model can be build. Here the author has chosen the stacking method. Once the base models are trained, their outputs/features on the training data are collected and used as features/input to a second level model called meta-model or level-1 model. For the final prediction this meta-model learns about the way, how best of the base models can be combined to produce the best results.

Hybrid model has been build with taking three different meta model. Initially, the predictions that are obtained from Light GBM model and random forest model has been combined with stacking method. Later on, these combined predictions is used as input features for the meta model and after training the meta model finally, the predictions are done. The three meta models used in this context are mainly, at first Linear regression is used as meta model and to compare the results with other meta model the author has choosen Ridge Regression and Gradient Boosting for that and Mean Absolute error has been obtained for each hybrid model and results will be explain in next section.

6 Evaluation

As mentioned above in section 3, the evaluation metrics used in this research is Mean Absolute Error(MAE).

6.1 Model Building with all the features

This section the author will discuss the results obtained for Mean Absolute Error by taking all the features for all the models. In this case study the author will be discussing the results that has obtained while using all the features form the dataset. As mentioned earlier the hybrid model implementation is done by stacking the predictions as input features to a meta model and the meta model used here is Linear regression. While taking Linear regression as meta model the mean absolute error for the hybrid model was 18.20 and for the base models Light GBM and random Forest regressor was 18.15 and 29.95, while using Ridge regression same MAE was obtained but with Gradient boosting the MAE value has increased to 114.9

6.2 Model Building with selected the features

Here for the modeling the author has taken the selected features that has obtained from SelectKbest method which has mentioned in section 3. While taking Linear regression as meta model the mean absolute error for the hybrid model in this case was 23.20 and for the base models Light GBM and random Forest regressor was 23.19 and 38.47. Similarly for ridge regression the mae was 23.20 and for gradient descent was 115.84

6.3 Discussion

The kind of experiments that have been performed in this research shed more light on how the hybrid models perform with different meta models and selection features. When all the features are utilized, the lightGBM model and random forest model had MAE value 18.15 and 29.95. While hybrid model with Linear regression and Ridge regression showed the same MAE value with 18.20 which is some what near to light GBM model. Even though there is a significant correlation between the predictions of both the models, the difference in MAE was not very significant, implying that the hybrid failed to significantly enhance the model's prediction capabilities compared to the LightGBM model.

Moreover, there is a situation where such base models as LightGBM can be better than hybrid models, for instance, when the characteristics of the dataset correspond to the base model's architecture. For example, LightGBM has a special application to work with large data and lots of features and nonlinear relationships between them. Such cases risks may arise where the added complexity of a hybrid model does not translate into improved performance, but rather hampers the performance as such models may tend to over-fit or are operationally inefficient. But when the meta-model was Gradient Boosting the MAE increased to 114. 9. This increase result indicates that, Gradient Boosting was not effective as a meta-model, probably because of over-fitting or other complications in the integration of the expectations from the base model. Hence, even though the usage of hybrid models is quite effective, the results attained working with such models heavily depend on the selected dataset and the field of application.

When selected features were used, the MAE for Light GBM model and random forest model was 23.19 and 38.47. The MAE for hybrid model with linear regression and ridge regression is .There are several reasons for model performing well with all features rather than selected features. First, it can mean that each of the features provides useful information that is provided to the model and hence, the patterns in the data can be elucidated more effectively. Some important variables such as cloud coverage has not been selected. The presence of all features might be important, which in essence means that the relations between the features could be important whereby when all features are applied in the model, the model is capable of detecting these relations and leads to better predictions.

Comparing the result, clearly states that LightGBM model has shown a significant improvement in this in this context compared to previous work. In (Chandran and Narayanan; 2024) the authors has reported MAE of 74.56, whereas the Light Gbm model in this study achieved a much lower MAE value(18.15) which clearly states its better performance. This improvement also applies to the case of the hybrid model that, as in the LightGBM model, presents a better performance with respect to the results obtained in the previous work, thus underlining the reliability of these models in this context.

7 Conclusion and Future Work

7.1 Conclusion

The research question for this study was 'How well a hybrid model can incorporate weather forecast data and market prices to predict energy imbalances for prosumers more effectively. The main objectives for this research was to conduct exploratory data analysis on the data to derive the insights for better understanding feature extraction and feature engineering and finally, implementing the base models and hybrid model by stacking the base models using meta model.

To address these objectives the author presented several experiments where the MAE values of hybrid models with different meta models such as Linear Regression,Ridge Regression and Gradient Boosting were compared to the LightGBM and Random Forest models, with all features and selected features. The presented experiments aimed to check whether the hybrid models can perfom well in predicting the energy-related performance.

The study was able to show that the 2 Hybrid models perform with equal efficiency to LightGBM wherein the MAE was established to be at 18.20 while LightGBM was 18.15 when all the features of the dataset were included. Although the performance of hybrid modelling is similar to that Light GBM model it indicates that the hybrid model is not significantly improving.

As seen, the hybrid model has the similar accuracy to the LightGBM, which was better than the Random Forest in this case. Feature Selection: The addition that all of the features improved the performance of the models showed the importance of the interaction of the subject and other variables for a good prediction. LightGBM model proposed in this study obtained significantly lower value of MAE (with the value of 18.56 as was written in the earlier studies indicating a major improvement.

This study could underpin the fact that, eventhough hybrid models can be useful in modeling, in this case it is not showing any significant improvement. The research also affirms that choice of feature is critical to model optimization for it determines the total feature collection. The study is important to the area because it provides a detailed comparison of hybrid and individual models and it demonstrates the capabilities of hybrid approaches in the framework of the complex prediction tasks.

The generalization of these findings might be slightly restricted by the nature of the data set used in this research. However, these hybrid models brought excellent results with their high parameters; especially in cases with different structures, it was seen that it could cause overfitting. Furthermore, this work was concentrated on several models and feature selection algorithms; however, others could be considered and applied to achieve better results.

7.2 Future work

It would be valuable for future studies to consider future work and analysis on these mixed models in order to expand their utility to other data types and fields. Also, a secondary study can work on the dynamic or adaptive hybrid models of analyzing the given dataset where the base models can be changed.

One of the areas that could be reviewed in the future is the extension of hybrid models concerning the integration of external real-time data feed in order to tune their performance in real-time settings, e.g. smart grids or renewable energy generation forecasting. This could include building models that upon receiving new data, the models can update their systems, hence becoming more accurate with time. Also considering with different ways of implementing hybrid modelling, in future they can try other methods such as blending and compare the results. Finally, it could be proposed to commercialize these models as the accuracy of predicting prosumer behavior is very useful in the energy sector. Potential to combine with partners in the specific industry means that appropriate tools based on these models become a reality.

References

Accouche, O. and Gangadhari, R. K. (2023). Optimizing decision-making of a smart prosumer microgrid using simulation., *Computers, Materials & Continua* **76**(1).

An, J., Hong, T. and Lee, M. (2022). Determining the optimal trading price of electricity for energy consumers and prosumers, *Renewable and Sustainable Energy Reviews* **154**: 111851.

Antal, M., Toderean, L., Cioara, T. and Anghel, I. (2022). Hybrid deep neural network model for multi-step energy prediction of prosumers, *Applied Sciences* **12**(11): 5346.

Bagherzadeh, F., Nouri, A. S., Mehrani, M.-J. and Thennadil, S. (2021). Prediction of energy consumption and evaluation of affecting factors in a full-scale wwtp using a machine learning approach, *Process Safety and Environmental Protection* **154**: 458–466.

Benti, N. E., Chaka, M. D. and Semie, A. G. (2023). Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects, *Sustainability* **15**(9): 7087.

Chandran, P. and Narayanan, A. (2024). Ensembling is all you need? evaluating machine learning models on predicting the energy imbalance of prosumers, *2024 11th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 291–296.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Egwim, C. N., Alaka, H., Egunjobi, O. O., Gomes, A. and Mporas, I. (2024). Comparison of machine learning algorithms for evaluating building energy efficiency using big data analytics, *Journal of Engineering, Design and Technology* **22**(4): 1325–1350.

Gajdzik, B., Jaciow, M., Wolniak, R., Wolny, R. and Grebski, W. W. (2023). Energy behaviors of prosumers in example of polish households, *Energies* **16**(7): 3186.

Hu, J.-L. and Chuang, M.-Y. (2023). The importance of energy prosumers for affordable and clean energy development: A review of the literature from the viewpoints of management and policy, *Energies* **16**(17): 6270.

Inês, C., Guilherme, P. L., Esther, M.-G., Swantje, G., Stephen, H. and Lars, H. (2020). Regulatory challenges and opportunities for collective renewable energy prosumers in the eu, *Energy policy* **138**: 111212.

Ishaq, M., Kwon, S. et al. (2021). Short-term energy forecasting framework using an ensemble deep learning approach, *IEEE Access* **9**: 94262–94271.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in neural information processing systems* **30**.

Kuznetsova, E. and Anjos, M. F. (2021). Prosumers and energy pricing policies: When, where, and under which conditions will prosumers emerge? a case study for ontario (canada), *Energy Policy* **149**: 111982.

Leal Filho, W., Trevisan, L. V., Salvia, A. L., Mazutti, J., Dibbern, T., de Maya, S. R., Bernal, E. F., Eustachio, J. H. P. P., Sharifi, A., Kushnir, I. et al. (2024). Prosumers and sustainable development: An international assessment in the field of renewable energy, *Sustainable Futures* **7**: 100158.

Lee, J., Hu, M., Tan, Y. R. and Wei, L. (2023). Shades of green: The hidden impact of prosumers and its mitigation, *Available at SSRN 4571381* .

Mathumitha, R., Rathika, P. and Manimala, K. (2024). Intelligent deep learning techniques for energy consumption forecasting in smart buildings: a review, *Artificial Intelligence Review* **57**(2): 35.

Probst, P., Wright, M. N. and Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest, *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* **9**(3): e1301.

Somu, N., MR, G. R. and Ramamritham, K. (2020). A hybrid model for building energy consumption forecasting using long short term memory networks, *Applied Energy* **261**: 114131.

Syed, D., Abu-Rub, H., Ghrayeb, A. and Refaat, S. S. (2021). Household-level energy forecasting in smart buildings using a novel hybrid deep learning model, *IEEE Access* **9**: 33498–33511.

Vergados, D. J., Mamounakis, I., Makris, P. and Varvarigos, E. (2016). Prosumer clustering into virtual microgrids for cost reduction in renewable energy trading markets, *Sustainable Energy, Grids and Networks* **7**: 90–103.

Yu, A., Zhang, C. and Zhang, Y.-J. A. (2019). Optimal bidding strategy of prosumers in distribution-level energy markets, *IEEE Transactions on Power Systems* **35**(3): 1695–1706.