Machine Learning Mid-Project Report
Title: Electricity Consumption For Household Appliances

Shivansh Mittal
shivansh20128@iiitd.ac.in

Rishabh Sharma
rishabh20236@iiitd.ac.in

Tushar Suredia
tushar20254@iiitd.ac.in

Niharika Singh
niharika21545@iiitd.ac.in

1. Abstract

The contemporary world grapples with an escalating demand for energy and its consequential environmental impact, necessitating novel approaches. Conventional methods of monitoring household electricity usage lack precision, failing to elucidate appliance-specific consumption patterns. This deficiency often results in suboptimal energy utilization and elevated utility bills. Our initiative seeks to tackle these issues by employing sophisticated machine-learning methodologies. Leveraging submeter readings and robust regression models, we can accurately forecast and categorize electricity usage for individual appliances. This provides homeowners with granular insights, empowering them to optimize energy consumption, curtail costs, and play a part in fostering a more sustainable future. Armed with this knowledge, individuals can make informed choices about their energy usage, ultimately contributing to a reduction in overall energy demand and a more environmentally conscious society.

2. Introduction

In the contemporary era, the world faces an unprecedented surge in energy consumption, accompanied by pressing environmental concerns. As societies strive for heightened energy efficiency and sustainable practices, leveraging innovative solutions becomes paramount. One crucial aspect of this challenge lies in understanding and optimizing electricity usage within households. Traditional methods of monitoring electricity consumption lack the necessary granularity, making it challenging to pinpoint specific appliance usage patterns. Consequently, households face inefficient energy utilization and soaring utility bills. To address this issue, our project employs advanced machine learning techniques, using a dataset containing a vast amount of entries and key features related to electricity consumption. With a focus on predicting global active power—the primary determinant of electricity usage in a household—we aim to develop a robust machine-learning model. By harnessing the potential of submeter readings and regression models, our objective is to provide homeowners with accurate and insightful predictions. These predictions empower individuals to make informed decisions, optimize energy usage, and contribute to a sustainable future by minimizing their environmental footprint.

3. Literature Survey

This problem has already been approached in various techniques like Support Vector Machines(SVMs), Neural Networks, ACO, and Logistic Regressions:

1. Prediction of energy consumption by a passive solar building using neural networks [1] by S. A. Kalogirou and M. Bojic. Artificial neural networks (ANNs) were used to predict the energy consumption of a passive solar building. The study utilized a dynamic thermal building model based on finite volumes and time marching to evaluate the building's thermal behavior. The ANN model, trained using simulated data, demonstrated high accuracy in predicting energy consumption, with a coefficient of multiple determination (R2 value) of 0.9991 for unknown data. The ANN model proved to be faster than dynamic simulation programs, making it a valuable tool for modeling the thermal behavior of buildings.

2. This study examines the impact of economic and demographic factors on electricity consumption in New Zealand. Previous research used multiple linear regression to explore variables like GDP, electricity prices, population, customers, tourism, and climate. This research focuses on GDP, electricity prices, and population as the key variables for predicting electricity use in New Zealand, using multiple linear regression. We compare the model's forecasts to national predictions and a Logistic model. While the Logistic model accurately describes historical consumption

patterns, it tends to be conservative in its forecasts due to saturation limits in the curve.

3. Using support vector regression and ant colony optimization to predict power load [3] by D. Niu, Y. Wang and D. D. Wu. The paper proposes a system for power load forecasting using support vector machine (SVM) and ant colony optimization (ACO) techniques. ACO is employed to process large amounts of data and eliminate redundant information, reducing the training data for SVM and overcoming the disadvantage of slow processing speed. The system mines historical daily loading data with similar meteorological features to the forecasting day, improving the accuracy of the forecasting model. The paper also introduces a new feature selection mechanism based on ACO, which is applied to find optimal feature subsets in the data reduction process. The proposed method achieves greater forecasting accuracy compared to single SVM and BP neural network models.

4. DATASET

Dataset details with data preprocessing techniques. This dataset offers a valuable resource. It comprises six months of electricity consumption records for a household, spanning from January 2007 to June 2007. The dataset encompasses various metrics, including global active power, global reactive power, voltage, and global intensity, as well as sub-metering data for specific areas within the home, such as the kitchen, laundry room, and electric water heater and air conditioner. With a total of 260,640 measurements, this dataset serves as a valuable tool for comprehending and analyzing patterns in household electricity consumption.

Insight 1: The dataset chronicles electricity consumption data, recorded at one-minute intervals, commencing on January 1, 2007. It encompasses a range of metrics, including global active power, global reactive power, voltage, global intensity, and sub-metering readings.

Insight 2: An initial observation showcases a consistent trend in global active power (representing electricity consumption), where it hovers between 2.55 and 2.56 for the first 14 data points. This suggests a period of stable electricity usage at the commencement of the recorded timeline.

Insight 3: The dataset reveals a persistent and comparatively low level of global reactive power, approximately around 0.1, for the initial 14 data points. This implies that the electrical load during this timeframe does not exhibit significant reactive power components.

Insight 4: The voltage readings, as presented in the dataset, exhibit minor fluctuations, predominantly residing within the range of 241.07 to 242.88 during the initial period under scrutiny.

Insight 5: Further analysis of the dataset unveils that Sub metering 1, Sub metering 2, and Sub metering 3 report zero consumption for the initial 14 data points. This infers that the designated areas or appliances associated with these sub-metering parameters remained inactive during the specified time interval.

These insights serve as a foundational understanding of the dataset's composition and early trends concerning electricity consumption and related metrics. Subsequent research and analysis may uncover more intricate patterns and potential correlations.

Graph for visualizing Sub metering for every month. Average Voltage consumed by the corresponding appliances with respect to different months.[Figure 1]

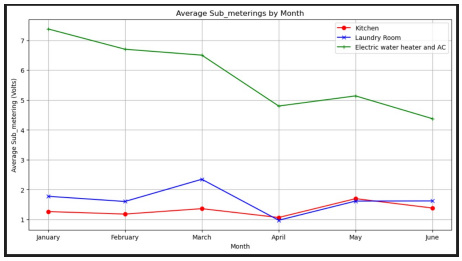

Figure 1.

Active Power Represents the actual power consumed for useful work in an electrical circuit, whereas Reactive Power reflects the power oscillating between source and load, vital for maintaining electromagnetic fields but not performing useful work. The graph below [Figure 2] represents the Power difference between the active and reactive power of our dataset.

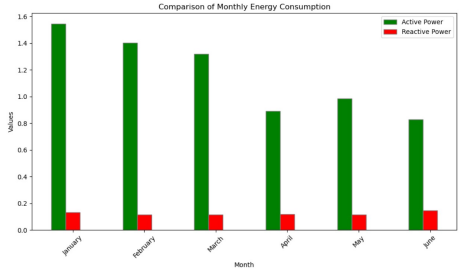

Figure 2.

A box plot is a concise data visualization that shows the median, spread, and potential outliers of a dataset, making it easy to compare distributions across different groups. Visualisation for Box plot for global active power to check for outliers.[Figure 3]

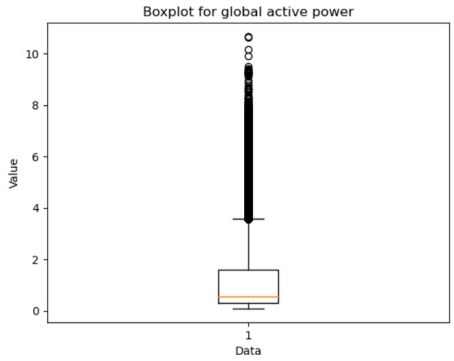

Figure 3.

5. METHODOLOGY

Our main objective is to develop a predictive model for predicting household electricity consumption by employing machine learning algorithms. In addressing the regression problem, we have explored different ML regression algorithms on the Dataset containing 2075259 measurements gathered between December 2006 and November 2010 (47 months). It features crucial attributes such as 'Date', 'Time', 'Voltage', 'Sub metering 1', 'Sub metering 2', 'Sub metering 3', 'Global reactive power', 'Global intensity', and the target variable 'Global active power'.

1. **Data Preprocessing**
   - **Label Encoding**
     Employed Label Encoding to convert the 'Time' column into a numerical format, ensuring compatibility with machine learning algorithms.
   - **Categorical Encoding**
     Transformed the 'Month' column into a categorical data type for efficient encoding.
   - **One-Hot Encoding**
     Implemented One Hot Encoding on 'Month new' to create a binary representation of months.
   - **Feature Selection**
     Carefully selected relevant features for the regression task, including 'Time', 'Voltage', 'Sub metering 1', 'Sub metering 2', 'Sub metering 3', 'Month new', 'Global reactive power', 'Global intensity'.
   - **Target Variable**
     Designated 'Global active power' as the target variable for regression.
2. **Data Splitting**
   In order to objectively assess the model's performance, the dataset was methodically partitioned into distinct training and testing subsets. The training set comprised 80 percent of the data, serving as the foundation for model development, while the testing set constituted the remaining 20 percent, providing an independent dataset for rigorous model evaluation.
3. **Models**
   These are the following models we have tried so far:
   (a) **Linear Regression**
   Linear Regression was employed for its simplicity and interpretability. It enables us to model the relationship between variables, aligning with our goal of predicting 'Global active power' based on selected features.
   (b) **SGD (Stochastic Gradient Descent) Regression**
   Stochastic Gradient Descent (SGD) was employed due to its efficiency in optimizing large-scale machine learning models given the extensive dataset with over two million measurements.
   (c) **Ridge Regression**
   Ridge regression is used to prevent overfitting in linear regression models. It adds a penalty term to the loss function that shrinks the coefficients towards zero. We used this model to improve the performance of your linear regression model.
   (d) **Bayesian Regression**
   Bayesian Regression was employed to get a confidence interval and a point estimate, with this Bayesian processing, and to get the full range of inferential solutions.
   (e) **Lasso Regression**
   Lasso Regression is similar to Ridge regression but uses L1 regularization instead of L2 regularization. Lasso was employed to find a balance between model simplicity and accuracy by adding a penalty term to the traditional linear regression model and for feature selection, as it can automatically identify and discard irrelevant or redundant variables.
   (f) **Decision Trees**
   Decision trees are a machine learning technique that comes under supervised learning algorithms. It works with the help of a tree where the nodes represent features, the edges represent the feature value, and the leaf nodes represent the final outcome of the tree model. We have tried it for our project since decision trees are efficient for regression problems.
   (g) **Random Forest Regressor**
   Random forests are a collection of several decision trees trained in similar-sized datasets made from the same original dataset using random

sampling. Random forests prevent overfitting and provide a better performance for regression problems. For our project, since our dataset is already very big, having 2.5 lakh data points, we have used only 100 decision trees.

(h) **MLP Regressor**
An MLP Regressor is a supervised machine-learning technique. It uses multilayers of neurons to make a neural network and then trains the model using the training dataset. For our case, we are using an MLP Regressor having two hidden layers each having 3 neurons/nodes. This configuration has been used after trying different combinations of values and seeing the performance metric of accuracy.

(i) **KNeighbours Regressor**
KNeighbours Regressor is a derivative of the KNN algorithm, which is generally used for classification purposes. KNN Regressor works on the principle of predicting values according to the values of the k-nearest neighbors. For our case, after trying different values and checking performance, we have run our algorithm for n=7, where n is the minimum number of nearest neighbors a point will look at.

4. **Model Evaluation**
The model's performance was rigorously evaluated using three key metrics:

(a) **Mean Squared Error (MSE)**
MSE provides an average of the squared differences between actual and predicted values.

(b) **Root Mean Squared Error (RMSE)**
RMSE offers a more interpretable error metric, particularly useful for conveying the magnitude of prediction errors.

(c) **R-squared (R2) Score**
R2 score measures the proportion of variance in the dependent variable that can be predicted from the independent variables.

6. RESULT AND ANALYSIS

After trying out all the regression models, we collected results for all of them and compared them using established metrics like MSE, RMSE, and R2 Score. The values of the metrics can be seen in Figure 10. Out of all the machine learning models we used, Linear Regression gave us the best results, which can be established using the graphs we obtained as shown and the metrics calculated by us. The performance of the models can be seen with the help of the following visualizations.

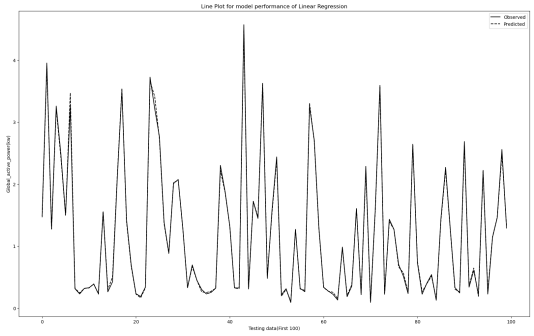

Figure 4. Linear Regression Performance

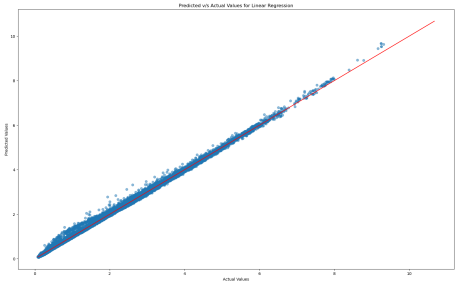

Figure 5. Predicted VS Actual values for Linear Regression

These visualizations show that the predicted values are very close to the actual values and our model is working very efficiently.

The following visualizations are for the other models we tried.

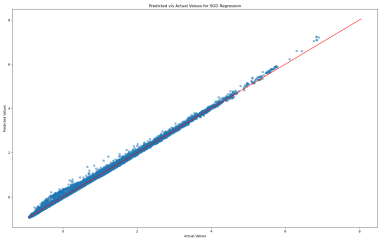

Figure 6. Predicted VS Actual values for SGD

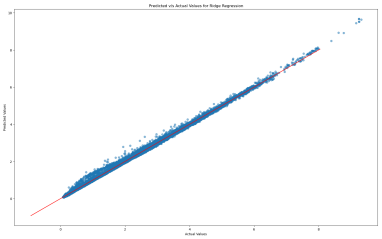

Figure 7. Predicted VS Actual values for Ridge Regression

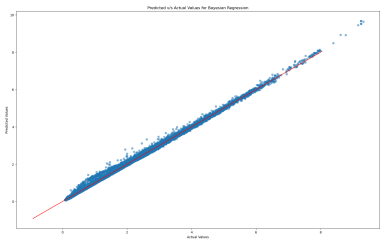

Figure 8. Predicted VS Actual values for BayesianRidge Regression

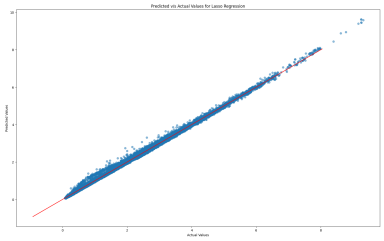

Figure 9. Predicted VS Actual values for Lasso Regression

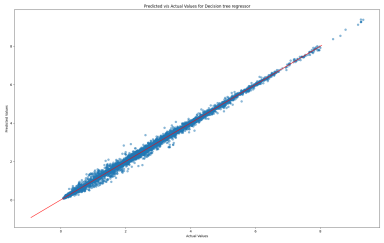

Figure 10. Predicted VS Actual values for Decision Trees Regressor

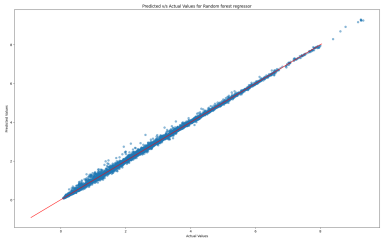

Figure 11. Predicted VS Actual values for Random Forest Regressor

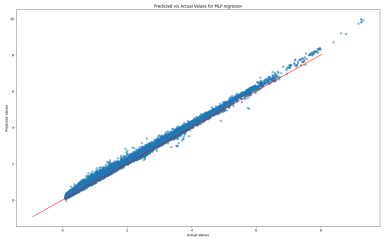

Figure 12. Predicted VS Actual values for MLP Regressor

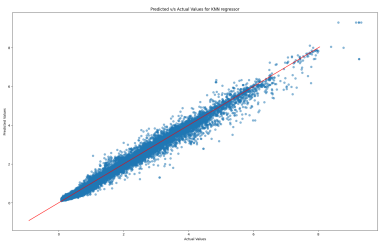

Figure 13. Predicted VS Actual values for KNN Regressor

Model	MSE	RMSE	R2 Score
Linear Regression	0.00204	0.04518	0.99854
SGD Regressor	0.00148	0.03856	0.99851
Ridge Regression	0.00204	0.04518	0.99854
Bayesian Ridge Regression	0.00204	0.04518	0.99854
Lasso Regression	0.00237	0.0487	0.9983
Decision Tree Regressor	0.00154	0.03934	0.99889
Random Forest Regressor	0.00083	0.02894	0.9994
MLP Regressor	0.02206	0.14852	0.98423
KNN Regressor	0.01626	0.12753	0.98837

Figure 14. Metrics for Regressions models

7. CONCLUSION

From our results, we can conclude that although all the regression models give good results. But since we need to choose the best model, we can go with either decision trees or random forests. The reason behind this is that there is a trade-off between the R2 score and time complexity in these models. In decision trees, the time complexity is the lowest, but its R2 score is slightly higher than that of Random forest. On the other hand, in random forests, the R2 score and the MSE value are the lowest, but the time complexity is higher as it works over 100 hundred decision trees and makes a combined result from all of them.

So, if complexity is not a problem, we should go with random forests. Otherwise, we can proceed with decision trees as the difference between their performance is not that significant. one thing to note here is that the small difference in the performances of decision trees and random forests is because our dataset is very large. So it is not overfitting as it is in a decision trees.

8. References

1 Artificial neural networks for the prediction of the energy consumption of a passive solar building by S. A. Kalogirou and M. Bojic

2 Forecasting electricity consumption in New Zealand using economic and demographic variables by Z. Mohamed and P. Bodger

3 Power load forecasting using support vector machine and ant colony optimization by D. Niu, Y. Wang and D. D. Wu