Matthew Landrum, Yehang Rai, Hans Holmes

DATA-151

Professor Jon Beagley

12 November 2024

Report: Part 3 of the Project

**Introduction**

With the revolution of AI and smart technology, one of the most commonly affected areas is the everyday household, which has modernized to become controllable remotely. While these provide convenience and benefit to its prospective consumers, managing the active power consumption of these homes has evolved into a monumental task as the demand for energy ever increased. In our project, we hope to remedy this issue by predicting global active power consumption, helping energy companies better prepare for and understand their energy resources and cost.

The ability to accurately forecast household energy consumption carries significant economic, environmental, and operational benefits. For individual households, such predictions enable more precise control over energy consumption patterns, which can lead to reduced costs by encouraging usage during off-peak periods. For energy providers, accurate demand forecasts allow for improved load balancing and better planning for peak demand periods. This capability is especially critical in light of the growing integration of renewable energy sources, which are inherently variable in their output. Renewable energy sources, such as solar and wind, fluctuate based on time of day and weather conditions, creating a need for demand forecasts that align with supply to ensure grid stability and prevent outages. Accurate demand forecasts support sustainability efforts by reducing reliance on fossil fuel-based energy during peak times. Ergo, we hope to, though our project, facilitate a new understanding of energy consumption that is more informed by the statistical changes and patterns inherent

**Related Studies**

Data and studies surrounding our model exist, with one of the most prominent utilizing time series analysis to forecast energy consumption. The study experiments on Korean household data. The study utilizes Weka to apply models on hourly and daily household energy publicly available, and then applies multiple models for the sake of prediction. The study found that the SVM Regression offered the best performance rates while Multilayer Perceptron and Gaussian Processes also gave good performances. The conclusions drawn in this paper not only show how important predicting energy consumption is, given the scope of their conclusions as well as the datasets utilized.

Additionally, another study exists that similarly measures energy consumption based on lifestyle data, focusing on aspects such as family size, income level, age, etc. The model utilizes an ARIMA model, an SVM model, and then utilizes a hybrid model to predict the energy consumption. Additionally, they also utilized both R and MATLAB for their coding, of which R gave a slightly more accurate result. The hybrid model is what produced the best result, with family size being the most accurate predictor of energy consumption. This paper does a good job showing how predicting energy consumption can benefit an everyday consumer, as they will know what to save on and what necessarily is driving costs up or down.

Lastly, we looked at a study that measured energy consumption in the Yangtze River Delta, China, this time looking at the effects that climate change had on energy consumption. The urbanization taking place within China has turned energy consumption into a major issue there, and the study utilizes data from the State Grid Corporation of China as well as an Econometric model. It also tests for sensitivity based on income groups.This model detects changes in peak energy consumption while also computing a damage function as well. Additionally, the study concluded that a U-shaped relationship exists between residential electricity consumption and daily temperature in Shanghai, China, while possibly not accounting for cooler or less developed areas of China since much of its data came from Shanghai. This study highlights the importance of utilizing unique variables for prediction, such as those stemming from climate change.

**Objectives**

This project centers on determining the viability of machine learning and time-series models in accurately predicting household energy consumption by capturing both technical and behavioral data patterns. We aimed to address the following research questions:

1. Can household energy consumption be accurately forecasted using time-series and machine learning models?
2. Which models and features are most effective in predicting usage patterns?
3. How do behavioral factors, such as holidays and the distinction between daylight and nighttime hours, influence the accuracy of predictions and consumption trends?

**Data and Methods**

The dataset used in this project spans multiple years of household energy consumption records, focusing primarily on Global\_active\_power, which represents the total active power consumed by the household in kilowatts and serves as the main variable we aim to predict. Other critical variables included in the dataset provide additional insights into energy consumption patterns, such as Global\_reactive\_power, which measures reactive power in kilovolt-amperes (a measure of power that alternates back and forth between the source and load). While Global\_reactive\_power does not directly contribute to energy consumption, it is essential for maintaining voltage stability and can signal shifts in energy demand. The dataset also includes Voltage, capturing the average household voltage, which fluctuates with grid conditions and appliance usage, and Global\_intensity, representing the total current intensity drawn by the household. This intensity level correlates with the power consumption, making it an important feature for predicting energy usage patterns. Three sub-metering features (Sub\_metering\_1, Sub\_metering\_2, and Sub\_metering\_3) are provided, offering data on energy consumption by specific household appliances or systems, thereby allowing the analysis of appliance-specific consumption trends.

The temporal nature of the data, captured through Date and Time columns, was enhanced by merging them into a DateTime column, facilitating time-based analysis and enabling the extraction of features that capture seasonal and daily patterns. During data preprocessing, we identified and addressed missing values and outliers to ensure a clean dataset. Missing values were primarily found in key columns like Global\_active\_power and Voltage. Since these missing values represented a small fraction of the data, we opted to remove them to maintain continuity in the time-series structure, as imputation could introduce bias. Outliers, particularly in Global\_active\_power and sub-metering columns, were detected through box plots and scatter plots and subsequently removed to prevent these extreme values from skewing the models’ predictions.

Feature engineering played a significant role in preparing the dataset for effective analysis. From the combined DateTime column, we extracted several time-based features, including Year, Month, Day, Hour, and Minute. These temporal features allowed the models to capture various cycles in energy usage, such as seasonal changes (through Month) and hourly consumption patterns (through Hour). We created behavioral indicators to capture lifestyle-related fluctuations in energy consumption. One such feature, Is\_holiday, is a binary indicator for weekends, which captures the different energy usage patterns typically observed on weekends as opposed to weekdays. Another feature, Light, is also a binary indicator that distinguishes daylight hours (6 AM to 6 PM), accounting for the impact of natural light availability on energy use, particularly for lighting and appliance usage. These engineered features added significant context to the dataset, helping the models learn from both the temporal and behavioral aspects of household energy consumption.

**Modeling**

The modeling approach in this project incorporated both traditional time-series methods and modern machine learning techniques to predict Global\_active\_power. By leveraging the rich set of features created through data preprocessing and feature engineering, we aimed to build models that could capture the inherent cycles, fluctuations, and behavioral patterns in the data. Each model was evaluated on its ability to predict future energy consumption using both the original features and engineered temporal and behavioral features.

The modeling began with linear regression as a baseline, establishing a starting point for model performance. Linear regression’s simplicity made it a useful benchmark, although it was limited in capturing non-linear relationships, which are prevalent in household energy data. Following this, we implemented a decision tree model, which improved upon the linear model by allowing for conditional feature splits based on patterns within the data. While decision trees are capable of handling non-linearity, they showed signs of overfitting, especially when used in isolation. Moving to ensemble models, the Random Forest model offered substantial improvement by combining multiple decision trees, each trained on different data subsets, to produce a robust average prediction. Random Forest was effective in handling the high-dimensional feature set, capturing interactions between variables such as Hour, Month, and Global\_intensity that influence daily and seasonal energy patterns.

XGBoost and LGBM were chosen as advanced ensemble models, with their capabilities in handling high-dimensional data and complex relationships proving particularly effective in this project. These models use gradient boosting, which sequentially refines predictions by focusing on the areas where the model struggles most, making them well-suited for high-variance data like energy consumption. XGBoost demonstrated strong feature selection abilities, frequently ranking Global\_intensity, sub-metering values, and time-based features (Hour, Day, and Month) as important predictors. LGBM’s rapid processing made it ideal for handling our large dataset, with excellent results due to its ability to weigh feature importance and reduce prediction errors iteratively.

Time-series models, specifically ARIMA and Prophet, were also applied to capture trends and seasonal cycles. ARIMA is a traditional time-series model that focuses on autoregressive and moving average components to identify correlations within the time-series data. It was effective in capturing long-term seasonal trends but struggled with short-term fluctuations due to its limited ability to handle high-frequency variations. Prophet, designed by Facebook for handling seasonality and long-term trends, performed slightly better, especially for long-term forecasting. It also had limitations with short-term predictions where frequent changes in data required higher flexibility. Together, these models provided insights into the periodicity of household energy data and acted as complementary methods alongside machine learning models.

The features selected for training each model included both the original technical variables (Global\_reactive\_power, Voltage, Global\_intensity, and sub-metering columns) and the engineered features (Year, Month, Day, Hour, Minute, Is\_holiday, and Light). These features enriched the dataset, allowing the models to capture real-world patterns and behaviors. Each model’s hyperparameters were optimized through tuning to ensure high generalizability and accuracy on unseen test data.

**Results**

Root Mean Squared Error (RMSE) was the primary metric used to evaluate model performance, selected for its sensitivity to large deviations between actual and predicted values. RMSE’s squared-error component emphasizes significant errors, making it particularly useful in energy forecasting, where the goal is to accurately capture both regular usage patterns and occasional spikes. Models with lower RMSE values were considered successful in closely aligning their predictions with actual energy consumption data, successfully capturing both recurring patterns and unexpected peaks in usage. RMSE allowed us to objectively compare models and quantify improvements across iterations and adjustments in feature engineering.

We also employed visual analysis techniques to gain qualitative insights into model accuracy. Scatter plots and time-series plots comparing actual and predicted values provided a clear view of each model’s performance.We could assess whether models accurately captured fluctuations, peaks, and troughs in energy consumption over time. XGBoost and LGBM, for instance, exhibited tight clustering around the line of expected values, confirming their high accuracy and alignment with real consumption patterns. Feature importance analysis further revealed that engineered features such as Is\_holiday and Light significantly contributed to the models’ ability to recognize behavioral patterns, validating the success of our feature engineering efforts.

Through the combination of RMSE and visual comparison, we obtained a comprehensive view of each model’s strengths and limitations. This multifaceted evaluation approach confirmed the superior performance of ensemble models, particularly XGBoost and LGBM, as they demonstrated both quantitative accuracy (through low RMSE) and qualitative accuracy (through tight clustering around actual values) in capturing household energy consumption patterns.

The performance of each model varied significantly, reflecting the complexity of household energy data and the effectiveness of different modeling approaches. Linear regression served as a baseline with moderate RMSE, confirming its limitations in handling complex, non-linear relationships. Despite its simplicity, the linear model highlighted areas where more advanced models could improve, particularly in capturing interactions between features like Hour, Month, and Global\_intensity.

The decision tree model offered an improvement by capturing non-linear relationships through hierarchical splits, though it showed a tendency to overfit, especially when not regularized. The ensemble Random Forest model marked a substantial improvement, as it averaged the predictions from multiple decision trees, reducing variance and enhancing robustness. Random Forest’s ability to handle high-dimensional data was particularly advantageous, allowing it to leverage both the temporal (Hour, Month) and behavioral features (Is\_holiday, Light) effectively.

Advanced ensemble models such as XGBoost and LGBM achieved the lowest RMSE values, approximately 0.03, indicating their proficiency in handling high-dimensional, complex data. XGBoost excelled at feature selection, consistently ranking Global\_intensity, sub-metering values, and time-based features as critical predictors. This model’s iterative refinement process enabled it to minimize errors by focusing on challenging areas in the data. LGBM similarly demonstrated strong performance, particularly due to its speed and efficiency in handling large feature sets, making it an ideal choice for our enriched dataset. Together, these models provided the most accurate predictions, capturing both seasonal and behavioral patterns with high fidelity.

Time-series models like ARIMA and Prophet contributed valuable insights into the long-term periodic trends within the data. ARIMA was effective in modeling seasonal cycles but faced challenges with short-term fluctuations, which are common in high-frequency energy data. Prophet, while slightly better suited for long-term forecasts, also struggled with frequent changes. These limitations highlighted the strengths of machine learning models in handling high-dimensional, non-linear data, particularly when enriched with behavioral indicators.

Ensemble models, specifically XGBoost and LGBM, emerged as the best-performing models, demonstrating both quantitative and qualitative accuracy. These models effectively captured the complex interactions in household energy data, leveraging engineered features to achieve high prediction accuracy and providing a practical solution for real-world energy forecasting.

**Conclusion**

This project applied machine learning and time-series forecasting techniques to predict household energy consumption, achieving high accuracy with XGBoost and LGBM models. Through extensive data preprocessing and feature engineering, the models effectively captured daily, seasonal, and behavioral patterns, supporting the project’s objective of accurately forecasting household energy demand. Our results indicate that ensemble models, particularly XGBoost, provide high predictive accuracy. Temporal and behavioral features, such as Is\_holiday and Light, significantly enhanced the model’s performance, underscoring the importance of feature engineering in capturing consumption trends.

These findings demonstrate that accurate, data-driven insights can support both cost savings and emissions reductions by enabling more efficient household energy management. This framework offers utility providers and consumers a practical tool for promoting sustainable energy practices and optimizing resource allocation.

The study raises privacy concerns, as household energy data can reveal occupant behaviors and routines. Potential biases in the dataset may limit model generalizability across different socioeconomic groups. Future research could explore anonymization techniques and validate the model across varied populations to address these concerns.

**References**

1. Rauf, S. A., & Adekoya, A. F. (2023). *Forecasting Household Energy Consumption Based on Lifestyle Data Using Hybrid Machine Learning*. Journal of Energy Systems and Technology.
2. Bilal, M., Kim, H., Fayaz, M., & Pawar, P. (2022). *Comparative Analysis of Time Series Forecasting Approaches for Household Electricity Consumption Prediction*. arXiv preprint arXiv:2207.01019.
3. Ahmad, T., et al. (2020). *Time Series Forecasting for Energy Consumption in Smart Homes*. Journal of Energy Systems.
4. Li, Yating, et al. “Climate Change and Residential Electricity Consumption in the Yangtze River Delta, China.” Proceedings of the National Academy of Sciences, vol. 116, no. 2, 24 Dec. 2018, pp. 472–477, www.pnas.org/content/116/2/472, https://doi.org/10.1073/pnas.1804667115.

Additionally,

What makes our model unique is the variables we’ve included to help predict power consumption. While studies do exist surrounding the topic, especially with the rise of smart homes, they instead focus on lifestyle decisions (often based on age and demographics). We instead opt to use more concrete data, factoring in voltage, submetering, and intensity into our model.

In order to properly predict the global active power consumption by smart homes, we utilized a couple of key regression methods. Firstly, we removed null values, which only consisted of 1.25% of the data, and performed all necessary conversions to the proper data types and and removed the date and time column; notably, we created a new column “time” that represented time in numeric format, where 0 corresponds to midnight and 1 to the end of the day, which we found worked best with our model. Next, we made a frequency histogram of the global intensity, which helped us to get an idea of the most common intensity levels. We also made a histogram of the distribution of the voltage as well as a scatterplot matrix that visualizes the relationships between global active power, global reactive power, voltage, and global intensity, while also making charts depicting the mean global active power grouped by year, quarter, month and day; because this is the variable we are attempting to predict, we felt that this was the way to attain the best understanding of its properties as possible.

For our predictive models, we utilized a regular regression model, a decision tree regression, a random forest regression, an MLP Regressor, a LightGBM Regressor, an XGBoost regressor, ADABoost Regression, Bagging Regression, and Gradient Boosting Regression. To test which one worked best for our model, we utilized RSME and found that the XGBoost regressor did the best job predicting the global active power, with an RMSE of only 0.028. As a result, we tuned the hyperparameters of the actual regressor, which gave us a slightly more accurate result.