Matthew Landrum, Yehang Rai, Hans Holmes

DATA-151

Professor Jon Beagley

**Predictive Modeling for Household Energy Consumption**

**Executive Summary:**

For our project, we hoped to predict global active power consumption, helping energy companies better prepare for and understand their energy resources and cost. We used a dataset that 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 aimed to predict. Additionally, we merged the Date and Time columns into a DateTime column, allowing us to better perform time series analysis. After removing outliers, which we detected through box plots, and missing data, which only accounted for a very small portion of our dataset. Additionally, we utilized feature engineering to extract several time based features, such as Year, Month, Day, Hour, and Minute, and then created behavioral indicators to capture lifestyle-related fluctuations in energy consumption, such as holidays and light considerations.

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.

From here, we incorporated traditional time-series methods and modern machine learning techniques to predict Global\_active\_power, and because of the rich set of features we attained due to our preprocessing, our model could better respond to the inherent cycles, fluctuations, and behavioral patterns in the data. As a baseline, we utilized a linear regression model, but due to its limitations we utilized a decision tree model and a Random Forest model, the latter of which showed large improvement, especially when handling high-dimensional data, due to its usage of multiple decision trees. From here, we chose advanced ensemble models, specifically XGBoost and LGBM, due to their ability to handle complex relationships and their usage of gradient boosting. 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.

Lastly, we utilized an ARIMA and a Prophet model, both of which are times series models. Both were effective in capturing long term seasonal trends, with the Prophet model performing the best out of the two, but both 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.

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. 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.

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.

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.

We found that 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.

**Background:**

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.

* 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.
* 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.
* Ahmad, T., et al. (2020). *Time Series Forecasting for Energy Consumption in Smart Homes*. Journal of Energy Systems.
* 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.

**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.

**Results (**[**Visualizations here!**](https://mybinder.org/v2/gh/Yehangraii/Data-151.git/main?labpath=Data-151.ipynb)**):**

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.

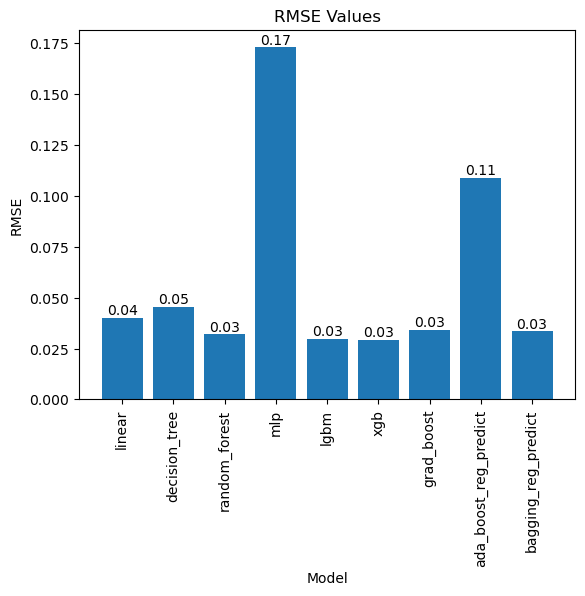
Through the combination of RMSE and visual comparison, we obtained a comprehensive view of each model’s strengths and limitations. This multifaceted evaluation approach confirmed the superior performance of ensemble models, particularly XGBoost and LGBM, as they demonstrated both quantitative accuracy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

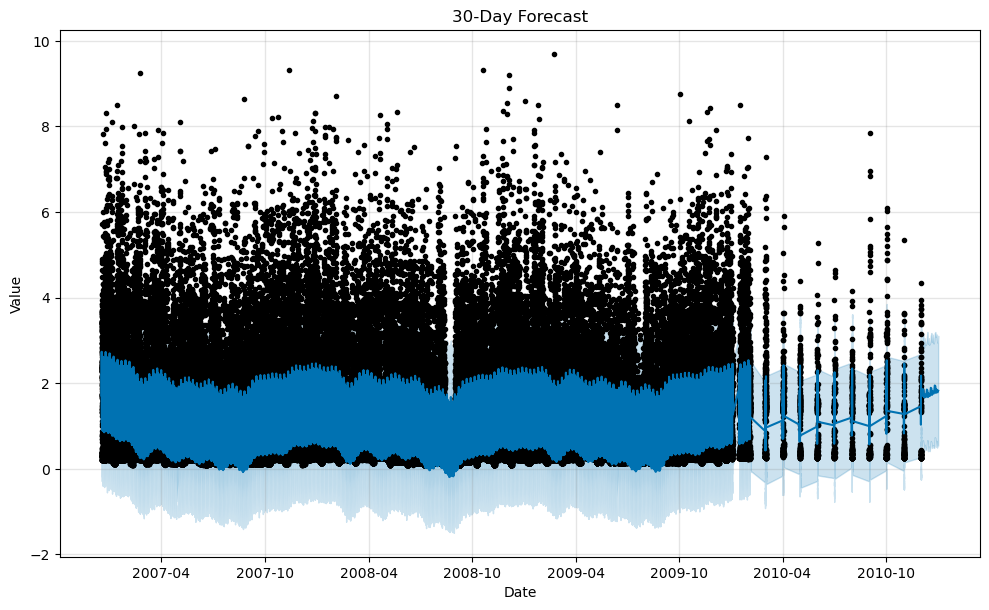
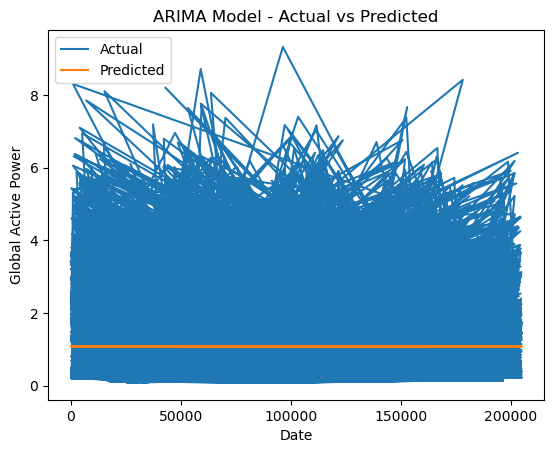
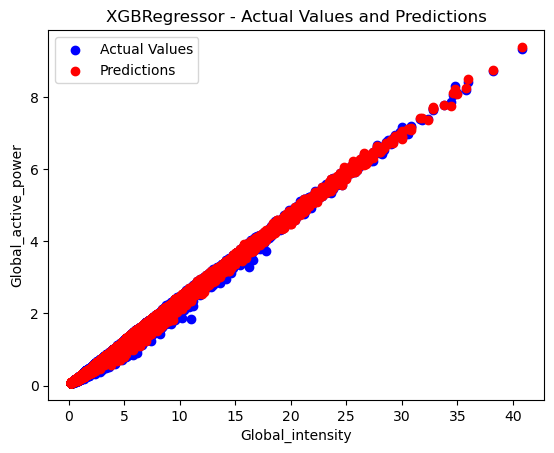
The performance of each model varied significantly, reflecting the complexity of household energy data and the effectiveness of different modeling approaches. Linear regression served as a baseline with moderate RMSE, confirming its limitations in handling complex, non-linear relationships. Despite its 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.

****