**1. Introduction**

**1.1. Background**

The increasing energy demand necessitates effective energy management in households. Understanding patterns in energy consumption and identifying key factors influencing it can enable optimized usage, cost savings, and environmental benefits. Leveraging machine learning and time-series analysis in big data provides precise insights and predictive capabilities to address these challenges.

**1.2. Why Conduct This Study?**

This study focuses on:

1. Developing predictive models for household appliance energy consumption.
2. Understanding the key factors affecting energy usage.
3. Informing strategies for energy optimization, scheduling, and smart home automation.

**1.3. Objectives**

* Forecast energy consumption using historical data to identify trends and patterns.
* Perform feature analysis to understand which factors most significantly influence energy usage.
* Evaluate model performance using statistical metrics and explore actionable insights.

**2. Dataset Overview**

**2.1. Description**

* **Source**: The dataset includes energy consumption data at 10-minute intervals from January 11, 2016, to May 27, 2016.
* **Size**: 19,735 rows and 29 features.
* **Variables**:

1. date time year-month-day hour:minute:second

2. Appliances, energy use in Wh

3. lights, energy use of light fixtures in the house in Wh

4. T1, Temperature in kitchen area, in Celsius

5. RH\_1, Humidity in kitchen area, in %

6. T2, Temperature in living room area, in Celsius

7. RH\_2, Humidity in living room area, in %

8. T3, Temperature in laundry room area

9. RH\_3, Humidity in laundry room area, in %

10. T4, Temperature in office room, in Celsius

11. RH\_4, Humidity in office room, in %

12. T5, Temperature in bathroom, in Celsius

13. RH\_5, Humidity in bathroom, in %

14. T6, Temperature outside the building (north side), in Celsius

15. RH\_6, Humidity outside the building (north side), in %

16. T7, Temperature in ironing room , in Celsius

17. RH\_7, Humidity in ironing room, in %

18. T8, Temperature in teenager room 2, in Celsius

19. RH\_8, Humidity in teenager room 2, in %

20. T9, Temperature in parents room, in Celsius

21. RH\_9, Humidity in parents room, in %

22. To, Temperature outside (from Chievres weather station), in Celsius

23. Pressure (from Chievres weather station), in mm Hg

24. RH\_out, Humidity outside (from Chievres weather station), in %

25. Wind speed (from Chievres weather station), in m/s

26. Visibility (from Chievres weather station), in km

27. Tdewpoint (from Chievres weather station), Â°C

28. rv1, Random variable 1, nondimensional

29. rv2, Random variable 2, nondimensional

**2.2. Key Features**

* Time-based data, enabling time-series forecasting.
* Multiple features allow multivariate analysis for deeper insights.

**2.3. Data Preprocessing**

* Removed missing values to ensure data integrity.
* Converted date strings to a DateTime format.
* Created training (70%) and testing (30%) sets using chronological order, splitting at 2016-04-16 15:20:00.

**3. Methodology**

**3.1. Tools and Libraries**

* **Python Libraries**: Pandas, NumPy, Prophet, SHAP, Gradient Boosting, ALE, and SciPy.
* **Google Colab**: Used for collaborative coding and experimentation.

**3.2. Forecasting Models**

* **Prophet**:
  + A robust model for univariate and multivariate time-series forecasting.
  + Includes trend, seasonality, and custom regressors for enhanced accuracy.

**3.3. Feature Importance Analysis**

* **Gradient Boosting Regressor**:
  + Evaluates the predictive power of individual features.
* **SHAP Values**:
  + Quantify the influence of each feature on predictions.
* **ALE (Accumulated Local Effects)**:
  + Visualize how features and their interactions affect the model.

**4. Analysis and Results**

**4.1. Univariate Time-Series Forecasting**

* Dependent variable: Appliances.
* Model correlation:

A graph with blue and black dots

AI-generated content may be incorrect.

* + Training data: **0.4336**.

A graph showing a waveform

AI-generated content may be incorrect.

* + Testing data: **0.3198**.
* Findings:
  + Captured daily energy consumption trends but limited accuracy due to single-variable input.

**4.2. Multivariate Time-Series Forecasting**

* Added regressors: lights, T1, RH\_1, T2, etc.
* Improved correlation:

A graph with blue and black lines

AI-generated content may be incorrect.

* + Training data: **0.5173**.

A graph showing a sound wave

AI-generated content may be incorrect.

* + Testing data: **0.4069**

A screenshot of a graph

AI-generated content may be incorrect.

**Interpretation:** **Interpretation:** The daily trend component shows that energy consumption begins to rise from 6:51 AM, peaks at around 7:00 PM and begins to decreases after that. The weekly trend shows a dip in energy consumption between Tuesday and Friday. In general, there is an upward trend in energy usage.

**4.3. Feature Importance Analysis**

* **Key Features**:
  + lights: Most influential factor, showing a strong positive correlation.
  + RH\_1: Significant contributor, with effects varying across its range.
  + Other key contributors: T1, RH\_out, T3, and Windspeed.
* **SHAP Insights**:

.A graph with red bars

AI-generated content may be incorrect.

**Interpretation:** The bar lengths represent the mean absolute SHAP value of each feature. This indicates how much each feature contributes, on average, to the model's predictions.Features at the top have the highest impact on the predictions. "lights" is the most influential feature, contributing a mean absolute SHAP value of 12.86, followed by "RH\_1" with 11.79.

A screen shot of a graph

AI-generated content may be incorrect.

**Interpretation:** "lights" is the most influential, where higher values (red) consistently increase predictions, while features like RH\_1 and RH\_out also significantly affect predictions but in a more mixed manner. The horizontal spread shows variability in the feature's impact, and the color gradient (blue for low values, red for high values) highlights whether high or low feature values drive predictions up or down. Less impactful features, like RH\_6 and Press\_mm\_hg,have smaller SHAP values, indicating a minimal influence on the model's output.

* **ALE PLOT**

ALE (Accumulated Local Effects) plots show how a feature or a pair of features influences the model's predictions by averaging their localized effects across the dataset while accounting for feature interactions and avoiding extrapolation bias.

A graph with a line and dots

AI-generated content may be incorrect.

**Interpretation:** This 1D ALE plot shows how the feature lights impacts the model's predictions. At lights = 0, the effect on predictions is significantly positive, indicating a strong influence. Between 10 and 50, the effect is relatively small and stable, contributing minimally to the predictions. However, when lights exceeds 50, there is a sharp increase in its effect, suggesting a strong positive relationship with the target variable at higher values. The bar plot indicates that most data points have lights = 0, making this value particularly important for the model's behavior.

A graph of a line

AI-generated content may be incorrect.

**Interpretation:** This 1D ALE plot illustrates how the feature RH\_1 influences the model's predictions. At lower values of RH\_1 (below 35), the effect on predictions is negative and stable, with a slight drop around 35. Between 35 and 50, the effect gradually increases, indicating a growing positive impact on the predictions. After RH\_1 exceeds 50, the effect sharply increases, showing a strong positive influence on the model's output for higher values of RH\_1. The black ticks along the X-axis represent the distribution of RH\_1 values in the dataset, with most data concentrated between 30 and 50, suggesting these ranges are critical for the model's behavior.

A chart of a color spectrum

AI-generated content may be incorrect.

**Interpretation:** This 2D ALE plot shows the interaction between lights and RH\_1 and their combined effect on the model's predictions. High values of both lights (above 50) and RH\_1 (above 50) result in a strong positive effect (yellow regions), increasing predictions significantly. Conversely, low values of either feature, especially when both are low, lead to a negative effect (purple regions), reducing predictions. For low lights values (below 10), the predictions are consistently negative regardless of RH\_1, while higher lights values amplify the positive influence of RH\_1, highlighting a strong non-linear interaction between these features.

* **Gradient Boosting Regressor**

A graph with red and green squares

AI-generated content may be incorrect.

**Interpretation**: Local explanation graph shows the contribution of individual features to a specific prediction made by the model, highlighting the positive (green) and negative (red) effects. At instance 519, lights > 0.00 and RH\_3 > 41.76 have the largest positive impact on increasing the prediction, with T8 <= 20.79 and T5 <= 18.28 also contributing positively but to a lesser extent. Conversely, T6 <= 3.63 and 20.79 < T3 <= 22.10 have the most significant negative impacts, reducing the prediction. Features like RH\_8 > 46.54 and Tdewpoint <= 0.90 also contribute positively, while conditions such as 750.93 < Press\_mm\_hg <= 756.10 have minor negative impacts. This graph provides a detailed breakdown of how each feature drives the prediction for a specific data point.

**5. Evaluation Metrics**

| **Metric** | **For 28 Features** | **For 20 Key features** |
| --- | --- | --- |
| **Mean Square Error (MSE)** | 7886.10 | 8236.60 |
| **Root Mean Square Error (RMSE)** | 88.80 | 90.76 |
| **R² Score** | 0.1656 | 0.1604 |

* Multivariate models yielded better accuracy compared to univariate models.
* Lower R² values highlight the complexity of modeling appliance energy consumption.

**6. Findings and Implications**

**6.1. Findings**

* Energy consumption trends reveal peak usage patterns, useful for scheduling energy-intensive activities.
* Multivariate models are effective for forecasting, capturing dependencies between features.
* Feature analysis underscores the impact of environmental and household factors on energy usage.

**6.2. Implications**

* **For Households**:
  + Insights into peak usage times can guide energy-saving strategies.
  + Identifying key factors helps optimize heating, lighting, and appliance usage.
* **For Researchers**:
  + Highlights the potential of combining time-series forecasting with feature importance analysis.
* **For Industry**:
* Aids in designing energy-efficient appliances and smart home solutions.
* The study demonstrated the efficacy of advanced models in energy consumption forecasting.
* Feature importance analysis provided actionable insights into energy optimization.
* Future work could explore deep learning models for further accuracy and real-time predictions.