Approach to the Problem

- **Data Cleaning & Preprocessing:**
  - First, converted object-type columns to numeric.
  - To handle the missing values, performed forward-filling technique.
  - Checked for outliers and addressed them using IQR-based filtering
  - To handle the skewness, performed log transformation.

- **Exploratory Data Analysis (EDA):**
  - Analyzed correlations to determine the main contributors to equipment energy consumption.
  - Lighting energy and Random variables appeared weakly correlated.
  - Identified highly skewed features and extreme outliers.

- **Feature Engineering:**
  - Extracted temporal features (hour, day of week, month) from timestamp variable.
  - Aggregated zone temperatures and humidities into meaningful summaries.

- **Modeling:**
  - Trained two regression models (Linear Regression & Random Forest Regressor) to predict equipment energy consumption
  - Evaluated model performance using metrics: R^2, RMSE, and MAE.

Key Insights from the Data

- **Data Quality Issues:**
  - Significant missing values and extreme outliers in several features (especially humidity).
- **Feature Importance:**
  - **Zone temperatures** (especially zones 4–6) and **outdoor conditions** (temp, humidity, pressure) are major predictors of equipment energy consumption.
- **Temporal Trends:**
  - Peak equipment energy usage during **working hours** (9 AM–6 PM).
  - Reduced usage on **weekends**, indicating operational schedule.

Model Performance Evaluation

- **Linear Regression (Ridge Regression) – Baseline Model**
  - Considered ridge regression as it assumes linear relationship between features and target and includes L2 regularization, which helps in reducing overfitting and handling multicollinearity.
  - **Performance**: Best alpha: 10.0; R^2 Score: 0.627; MAE: 0.167; RMSE: 0.220
  - Though the algorithm performed decently, its relatively lower R^2 indicates it could not capture all the variance in the data.

- ➢ The assumption of linearity likely limited its performance, as real-world relationships between sensor data (like temperature and humidity) and energy output are often non-linear and involve interactions.

❖ **Random Forest Regressor**

- ➢ Random Forest does not assume linearity, making it well-suited for complex datasets with feature interactions and non-linear relationships.
- ➢ It naturally handles missing values, outliers, and feature importance extraction.
- ➢ **Performance**: Best Params:{'max_depth': None, 'min_samples_split': 7, 'n_estimators': 100} ; R^2 Score: 0.696; MAE: 0.145; RMSE: 0.199

❖ **Comparison with Ridge:**

- ➢ R^2 increased from 0.627 (Ridge) to 0.696 (RF), and shows a significant improvement in explained variance.
- ➢ MAE decreased from 0.167 to 0.145 which indicates more accurate predictions on average.
- ➢ RMSE also improved, which suggests fewer large errors.

Analysis Through SHapley Additive exPlanations (SHAP)

SHAP values provide a unified measure of feature importance and feature impact direction for each prediction. This specific summary plot visualizes feature influence across all predictions made by the model.

Key Components of the Plot:

- **Y-axis:** Features, ranked by overall impact on the model's output.
- **X-axis:** SHAP value, representing the impact on the prediction (positive or negative).
- **Color Gradient:** Feature value (low = blue, high = red).
- **Each dot:** Represents a single observation.

Interpretations

1. energy_consumption_lag1 (Most Influential):
Impact: Strongly affects predictions both positively and negatively. High lag values (red) increase the predicted output (positive SHAP value). Low lag values (blue) decrease the prediction.
Interpretation: The previous time step's energy consumption is highly predictive of the current value, which makes sense in time-series energy modeling (autocorrelation).

2. zone6_encoded:
Impact: High variation in SHAP values across samples. Certain values (likely zone identifiers) positively or negatively influence predictions which show heterogeneous behavior across zones.

Interpretation: This zone’s encoded value has a significant relationship with energy patterns, possibly due to varying occupancy or equipment.

Some Key points:

- Autoregressive signals (lag features) are most crucial and confirm that recent consumption history is a strong predictor.
- Zone-specific behavior plays a major role and indicates strong spatial variability.
- Time-based features (like hour) add contextual understanding.
- Dimensionality reduction (PCA) contributed less (we can remove it to reduce overhead) likely due to already informative raw features.

Recommendations for Reducing Equipment Energy Consumption

1. **Fix Sensor Anomalies:**
   Replace faulty sensors reporting negative or unrealistic humidity/temperature values.

2. **Utilize Off-Peak Scheduling:**
   Reduce or shut down non-critical equipment during off-hours and weekends.

3. **Integrate Real-time Weather Feedback:**
   Link HVAC operations with real-time outdoor weather conditions.

4. **Reduce Lighting Load:**
   Lighting energy shows substantial variance and outliers, indicating potential overuse.