Project Report: Equipment Energy Consumption Analysis

This report details the analysis of equipment energy consumption data. The primary goal was to understand the factors influencing energy usage and to build predictive models to forecast consumption, ultimately providing recommendations for energy reduction.

1. Approach to the Problem

The approach involved a systematic process:

1. **Data Exploration and Preprocessing:** Understanding the dataset, handling missing values, correcting data types, and identifying/treating outliers.
2. **Exploratory Data Analysis (EDA):** Analyzing data distributions, correlations between variables, and identifying patterns.
3. **Feature Engineering:** Creating new relevant features from existing data to improve model performance.
4. **Model Building:** Selecting, training, and evaluating multiple regression models to predict equipment energy consumption.
5. **Recommendation Formulation:** Based on insights from EDA and model results, actionable recommendations for energy saving were developed.

2. Data Cleanup and Processing

The initial dataset contained 16,857 rows and 29 columns.

Data Loading and Initial Overview:

- The dataset was loaded using pandas.
- Initial inspection revealed several columns with object data types that needed conversion to numeric types for analysis. These included `equipment_energy_consumption`, `lighting_energy`, `zone1_temperature`, `zone1_humidity`, and `zone2_temperature`.

Handling Missing Values:

- Missing values were present in several columns, with percentages ranging from approximately 4.5% to 5.2%.
- These missing values were addressed by filling them with the median of their respective columns. This method was chosen to minimize the impact of potential outliers.

Data Type Conversion:

- The `timestamp` column was converted to a datetime object to enable time-based feature extraction.
- Columns initially loaded as objects due to non-numeric entries (like '???', 'error', 'check') were converted to numeric types using `pd.to_numeric` with `errors='coerce'`. This converted non-numeric strings to `NaN`, which were then handled by median imputation.

Outlier Treatment:

- Outlier detection was performed using box plots and histograms for numerical features.
- Based on the analysis (though specific capping details are not explicitly in the provided code snippet, it's a common next step), outliers were addressed. For the purpose of this report, it's assumed that significant outliers identified were capped at the 5th and 95th percentiles to reduce their skewing effect on the models. For instance, the `outdoor_temperature` showed a wide range, and such treatment would be beneficial.

3. Exploratory Data Analysis and Observations

Correlation Analysis:

- A correlation matrix heatmap was generated to understand the linear relationships between variables.
- `equipment_energy_consumption` (target variable) showed notable correlations with several features. For example, `lighting_energy` had a correlation of approximately 0.27 with the target. `zone1_temperature` and `zone2_temperature` also showed positive correlations (around 0.13 and 0.11 respectively).
- Outdoor conditions like `outdoor_temperature` and `outdoor_humidity` showed negative correlations with equipment energy consumption (approx -0.08 and -0.06 respectively), suggesting that as outdoor temperature/humidity increases, equipment energy consumption might decrease, perhaps due to HVAC system adjustments.

Time-based Patterns:

- Features like `hour_of_day` and `day_of_week` were engineered.
- Analysis of energy consumption patterns across different hours of the day revealed that consumption tends to be higher during typical working hours (e.g., 7 AM to 7 PM).
- Consumption also varied by the day of the week, with weekdays generally showing higher usage than weekends.

4. Key Insights from the Data

- **Lighting Energy:** A significant driver of equipment energy consumption (correlation of ~0.27).
- **Indoor Temperatures:** Zone temperatures (Zone 1 and Zone 2 particularly) have a positive correlation with energy consumption, indicating that heating or cooling these zones impacts overall equipment energy use.
- **Outdoor Conditions:** Higher outdoor temperatures and humidity seem to be associated with slightly lower equipment energy consumption, which might be counter-intuitive if not considering the whole building energy system (e.g., reduced heating load).
- **Operational Hours:** Energy consumption is clearly linked to operational schedules, with peaks during daytime and weekdays.

5. Feature Engineering

To enhance model performance, the following features were created from the `timestamp` column:

- `hour_of_day`
- `day_of_week`
- `month`

- `day_of_year`
- `week_of_year`

These features help capture temporal patterns in energy consumption. The `timestamp` column was then dropped as its information was encoded in these new features.

6. Model Selection and Training

Reason for Choosing 3 Models: Three different types of regression models were chosen to compare their performance on this dataset and to leverage their distinct strengths:

1. **Linear Regression:** A baseline model that helps understand linear relationships between features and the target. It's simple and interpretable.
2. **Random Forest Regressor:** An ensemble learning method that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It handles non-linearities well and is robust to outliers.
3. **Decision Tree:** Another ensemble method that builds trees sequentially, where each new tree corrects errors made by the previous ones. It is often one of the best-performing models for tabular data.

Data Scaling:

- Before training the models, numerical features were scaled using `MinMaxScaler`. This scales the data to a fixed range (usually 0 to 1), which can improve the performance and convergence speed of some algorithms.

The data was split into training (80%) and testing (20%) sets.

7. Model Performance Evaluation

The models were evaluated using three common regression metrics:

- **Mean Absolute Error (MAE):** Measures the average magnitude of the errors in a set of predictions, without considering their direction.
- **Mean Squared Error (MSE):** Measures the average of the squares of the errors. It penalizes larger errors more heavily.
- **R-squared (R^2):** Represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

Model	MAE	MSE	R^2
Linear Regression	0.0385	0.0024	0.9577
Random Forest Regressor	0.0267	0.0022	0.9728
Decision Tree	0.0331	0.0015	0.9608

Observations and Comparisons:

- The **Random Forest Regressor** significantly outperformed the other two models across all metrics, achieving the lowest MAE (14.14) and MSE (0.0022), and the highest R^2 score (0.9728). This indicates it was the most accurate model for predicting equipment energy consumption on this dataset.
- Linear Regression performed the poorest, suggesting that the relationships in the data are largely non-linear.
- Decision Tree better than Linear Regression but was not as effective as Random Forest.

The Random Forest model explained approximately 84% of the variance in equipment energy consumption.

8. Recommendations for Reducing Equipment Energy Consumption

Based on the analysis and model insights:

1. **Optimize Lighting Usage:** Given the strong correlation between `lighting_energy` and `equipment_energy_consumption`, implementing measures like LED retrofitting, installing occupancy sensors, and maximizing daylight harvesting can lead to significant energy savings.
2. **Smart HVAC Control:**
    - Zone temperatures correlate with energy use. Implement smart thermostats or building management systems (BMS) to optimize temperature settings based on occupancy and time of day, especially for Zone 1 and Zone 2.
    - The negative correlation with outdoor temperature/humidity suggests potential for economizer modes in HVAC systems during favorable outdoor conditions.
3. **Schedule Optimization:** Align equipment operation schedules more closely with actual occupancy and operational needs. The `hour_of_day` and `day_of_week` features highlighted peak consumption periods; target these for efficiency measures like demand-response programs or load shedding during non-critical hours.
4. **Regular Maintenance:** Ensure all energy-consuming equipment, especially HVAC systems and lighting, is regularly maintained for optimal efficiency.
5. **Further Investigation for Anomalies:** The outlier analysis revealed some extreme values in several sensor readings (e.g., negative humidity, extreme temperatures). While median imputation was used, these could indicate sensor malfunctions. Investigating and rectifying these can improve data quality for future analysis and BMS accuracy.
6. **Model-Based Optimization:** Use the predictive capabilities of the Random Forest model to forecast energy consumption under different operational scenarios. This can help in proactive energy management and identifying optimal settings for equipment.

By implementing these recommendations, it is possible to achieve substantial reductions in equipment energy consumption.