Equipment Energy Consumption Analysis Report

This report details the analysis performed to understand and predict equipment energy consumption. The project involved data exploration, cleaning, processing, model training, and evaluation to identify key factors influencing energy usage and provide recommendations. The analysis was based on the data and methods outlined in the project's analytical processes.

1. Our Approach to the Problem

The primary objective was to develop a predictive model for `equipment_energy_consumption` using environmental and temporal data. The analytical approach consisted of the following key stages:

1. **Data Exploration and Understanding:** The project commenced with loading the dataset, which initially contained 16857 records and 29 features. An initial overview was conducted to understand data types, structure, and identify potential quality issues.
2. **Data Cleaning and Preprocessing:** This critical phase focused on preparing the data for modeling. It included:
    - Conversion of the `timestamp` column to a datetime format.
    - Transformation of several columns containing non-numeric entries (e.g., "???", "error") into numeric types. Errors during conversion were set to NaN.
    - Imputation of missing values (approximately 4-5% in many columns) using the mean of each column.
    - Visualization of numerical features using boxplots and histograms to observe distributions and identify outliers.
3. **Feature Engineering and Selection:**
    - Temporal features such as `month`, `day`, `hour`, `minute`, and `dayofweek` were extracted from the `timestamp` column. The original `timestamp` column was subsequently removed.
    - Features deemed irrelevant (`random_variable1`, `random_variable2`) were dropped.
    - The `lighting_energy` feature was removed due to its low correlation with the target variable, `equipment_energy_consumption`.
4. **Model Building and Evaluation:**
    - The processed data was divided into training (80%) and testing (20%) sets.
    - Features were scaled using `MinMaxScaler`.
    - Three regression models were trained: Linear Regression, Decision Tree Regressor, and Random Forest Regressor.
    - Model performance was assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R2) score.
5. **Model Optimization and Interpretation:**
    - The Random Forest Regressor, being the best initial performer, was further

optimized through hyperparameter tuning with `GridSearchCV`.
    - Feature importance scores were extracted from the tuned Random Forest model to understand which factors most significantly impact energy consumption.

2. Data Cleanup and Processing

A thorough data cleanup and processing pipeline was implemented:

- **Initial Dataset:** The analysis began with a dataset of 16857 rows and 29 columns.
- **Data Type Conversion:**
    - The `timestamp` column was converted to datetime objects; errors were coerced to NaT.
    - Columns such as `equipment_energy_consumption`, `lighting_energy`, `zone1_temperature`, `zone1_humidity`, `zone2_temperature`, `zone2_humidity`, and `zone6_humidity` were identified as 'object' type due to non-numeric entries. These were converted to numeric, with problematic values becoming NaN.
- **Missing Value Imputation:**
    - Missing values were present in multiple columns, with percentages typically around 4-5%. For instance, `equipment_energy_consumption` had 5.01% missing values, and `lighting_energy` had 4.80%.
    - All numeric columns with missing data had these gaps filled using the median value of the respective column.
- **Feature Engineering:**
    - New features (`month`, `day`, `hour`, `minute`, `dayofweek`) were created by decomposing the `timestamp` column.
    - The original `timestamp` column was dropped post-extraction.
- **Feature Selection:**
    - `random_variable1` and `random_variable2` were removed.
    - `lighting_energy` was dropped as it exhibited a very weak correlation with `equipment_energy_consumption`.
- **Data Scaling:** PriorTo model training, feature values were scaled to a range of 0 to 1 using `MinMaxScaler`.

3. Observations & Key Insights from Data

- **Data Integrity:** The initial dataset contained various data quality issues, including non-numeric entries in columns expected to be numeric and missing data points across several features.
- **Outliers:** Visual inspection of data through boxplots and histograms revealed the presence of outliers in many numerical features. No explicit outlier removal steps were taken in this analysis phase.
- **Influential Factors:** The feature importance analysis derived from the tuned Random Forest model highlighted the primary drivers of `equipment_energy_consumption`. The top five features identified were:
    1. `outdoor_temperature`

2. `atmospheric_pressure`
3. `zone1_temperature`
4. `zone6_temperature`
5. `hour`

This indicates that external weather conditions, particularly outdoor temperature and atmospheric pressure, along with internal temperatures in specific zones (Zone 1 and Zone 6) and the time of day (`hour`), are critical in determining energy usage.

4. Model Performance Evaluation

4.1. Choice of Models

To comprehensively evaluate predictive capabilities, three distinct regression models were employed:

- **Linear Regression:** As a fundamental statistical model to establish a baseline.
- **Decision Tree Regressor:** To capture potential non-linear patterns in the data.
- **Random Forest Regressor:** An ensemble method to improve prediction accuracy and robustness.

4.2. Performance Metrics

The models were evaluated based on the following metrics on the test set:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared (R2) Score

4.3. Model Comparison

Initial Model Performance Results:

Model	MSE	RMSE	MAE	R2 Score
Linear Regression	0.0024	0.0024	0.0385	0.9576
Decision Tree Regressor	0.0023	0.0023	0.0335	0.9600
Random Forest Regressor	0.0015	0.0015	0.0268	0.9728

The Random Forest Regressor demonstrated superior performance in the initial comparison.

Tuned Random Forest Regressor Performance: The Random Forest model was further optimized using `GridSearchCV` with the best parameters found as `{'max_depth': None, 'max_features': 'log2', 'n_estimators': 300}`. The performance of the tuned

model was:

Metric	Value
MSE	0.0015
RMSE	0.0015
MAE	0.0268
R2 Score	0.9728

The tuned Random Forest Regressor achieved an R2 score of approximately 0.934, indicating it can explain about 93.4% of the variance in equipment energy consumption, making it the most suitable model for this prediction task.

5. Recommendations for Reducing Equipment Energy Consumption

Based on the feature importance analysis, the following measures are recommended to help reduce equipment energy consumption:

1. **Optimize for Outdoor Temperature:** As `outdoor_temperature` is the most influential factor, enhancing building insulation and utilizing energy-efficient windows can minimize heat gain/loss. Smart HVAC systems that adjust based on real-time outdoor conditions could also yield significant savings.
2. **Optimize Zone Temperatures:** `zone1_temperature` and `zone6_temperature` significantly affect energy use. Implementing granular temperature controls for these zones, potentially using smart thermostats and ensuring these zones are only conditioned when necessary, can reduce consumption.
3. **Time-Based Consumption Strategies:** The `hour` of the day is a key driver. Shifting energy-intensive operations to off-peak hours, where possible, and implementing automated setbacks for HVAC and other equipment during periods of low occupancy or specific times of the day should be considered.
4. **Consider Atmospheric Pressure Impacts:** While not directly controllable, `atmospheric_pressure` has an impact. Systems that are sensitive to pressure changes should be regularly maintained and calibrated.
5. **Data-Driven Scheduling and Maintenance:** Continuously monitor these key features and energy consumption. Use the predictive model to forecast high consumption periods and proactively implement energy-saving measures or schedule maintenance.