Prediction Report

Approach to the problem

Data Preprocessing Approaches

1. **Handling Missing Values**
   Imputed missing numerical values with mean/median.
   Filled categorical missing values with the most frequent category.
   Dropped rows/columns with excessive missing data.

2. **Converting Columns to Correct Data Types**
   Converted timestamps to datetime.
   Encoded categorical variables (Label Encoding / One-Hot Encoding).

3. **Dealing with Outliers**
   outliers using box plots and statistical thresholds.
   Removed or capped extreme values.
   Applied log transformation for normalization.

4. **Feature Scaling**
   Applied StandardScaler or MinMaxScaler for normalization.
   Ensured uniform feature importance and model stability.

Feature Engineering Approaches

5. **Time-Based Feature Extraction**
   Extracted `hour`, `day of week`, and `month` from the timestamp.

6. **Correlation-Based Feature Selection**
   Identified and selected features highly correlated with the target.

7. **SHAP-Based Feature Importance (Post-modeling**)
   Used SHAP values to interpret model and rank top contributing features.

Modeling & Evaluation Approaches

8. **Grid Search for Hyperparameter Tuning**
   Tuned models like XGBoost using GridSearchCV with different combinations of parameters.

9. **Model Evaluation**
   Compared models using cross-validation scores.
   Visualized results using heatmaps and bar charts.

10. **Interpretability & Recommendations**
    Generated insights using SHAP.
    Made recommendations based on top influential features.
    Identified peak usage hours for load shifting strategies.

Key Insights from the Data:

- **Target Variable Distribution: equipment_energy_consumption** is concentrated around zero with less frequent large positive/negative values, indicating mostly low consumption with occasional spikes.
- **Feature Importance**: **energy_roll_mean_3h** is the most influential predictor, followed by **hour**, **energy_lag_1**, and **energy_lag_3**.
- **Temporal Patterns**: Energy consumption peaks around **18:00**, highlighting a period of high demand in Indore, Madhya Pradesh, India.
- **Zone Characteristics**: Average humidity is highest in **zone6_humidity**, and average temperature is highest in **zone3_temperature**.
- **Missing Values**: Significant missing data exists in temperature and humidity readings for several zones.
- **Outliers**: Outliers are present in **equipment_energy_consumption**, **lighting_energy**, **outdoor_temperature**, and **atmospheric_pressure**.
- **Correlations**: A strong positive correlation exists between equipment_energy_consumption and lighting_energy.
- **'timestamp'** needs a datetime conversion: (Implicit from the need for temporal pattern analysis).
- **random_variable1** and **random_variable2** have low predictive importance: (From previous feature importance analysis).
- **Limited Impact of Zone Environmental Factors**: (Inferred from lower feature importance).
- **Potential for Load Shifting Benefits**: (Inferred from the peak consumption at 18:00).
- **Need for Further Investigation of Outliers**: (General data quality concern).

Model Performance Evaluation Observations:

Model	RMSE	MAE	R^2
Linear Regression	28.02	18.13	0.617
Random Forest	24.18	14.27	0.715
XGBoost	24.4	14.51	0.71
LightGBM	24.19	14.49	0.714

Model	colsample_bytree	learning_rate	max_depth	n_estimators	subsample	num_leaves
XGBoost	0.8	0.1	6	100	1	N/A
LightGBM	N/A	0.1	-1	100	N/A	31

Tuned Model	RMSE	MAE	R^2
Tuned XGBoost	24.38	14.61	0.71
Tuned LightGBM	24.29	14.57	0.712

- Tree-based models outperform Linear Regression.
- The initial performance of Random Forest and LightGBM is slightly better than **XGBoost**.
- Hyperparameter tuning provides marginal improvement for LightGBM.
- R^2 values around 0.71 indicate a reasonable predictive capability.

Recommendations for Reducing Equipment Energy Consumption

1. Optimize **energy_roll_mean_3h** Control Systems: Manage operational factors influencing this key predictor to smooth energy usage.
2. Monitor Hours During Peak **Hours** (**Around 18:00**): Analyze and manage energy use during peak demand periods.
3. Maintain **energy_lag_1** Within Optimal Ranges: Control factors leading to high past energy consumption.
4. Consider Load Shifting: Reschedule non-critical energy use away from the 18:00 peak.