1. Approach

This project aimed to develop a machine learning model to predict equipment energy consumption in a manufacturing facility and provide actionable insights for reducing energy usage. The approach involved the following steps:

1. **Data Loading and Exploration:** The provided sensor data was loaded into a pandas DataFrame. Descriptive statistics, data types, missing values, and distributions of key variables were analyzed. Correlations between variables, including equipment_energy_consumption, temperature, humidity, and random variables, were investigated.
2. **Data Cleaning:** Missing values in the equipment_energy_consumption column were filled using the median. Outliers were handled using the Interquartile Range (IQR) method. Duplicate rows were removed.
3. **Data Preparation and Feature Engineering:** Time-based features (hour of day, day of week, month of year) were extracted from the timestamp column. Interaction terms between relevant environmental factors and energy consumption were engineered to capture potential dependencies. The inclusion of random variables was evaluated based on their correlations and potential impact on the model.
4. **Data Splitting:** The data was split into training, validation, and testing sets using a temporal split to ensure realistic model evaluation.
5. **Model Training and Optimization:** Multiple regression models (RandomForestRegressor, GradientBoostingRegressor, LinearRegression) were trained on the training data. Hyperparameter optimization was performed using the validation set, and model performance was evaluated using RMSE, MAE, and R-squared.
6. **Model Evaluation and Feature Importance:** The best performing model was evaluated on the held-out test set using the same metrics. Feature importance analysis was conducted to identify the most influential factors in predicting energy consumption.
7. **Data Visualization:** Visualizations were created to illustrate model predictions vs. actuals, relationships between key features and energy consumption, and model performance metrics.

2. Key Insights from the Data

- **Weak Correlation with Random Variables:** random_variable1 and random_variable2 exhibited negligible linear relationships with equipment_energy_consumption, justifying their exclusion from the model.
- **Interaction Terms are Key Predictors:** Interaction terms between zone temperatures (zone1, zone3) and equipment energy consumption emerged as the most important features in the Random Forest model.
- **Temporal Patterns:** Time-based features, such as hour of day, day of week, and month of year, could potentially influence energy consumption patterns.

3. Model Performance Evaluation

The optimized Random Forest model achieved the best performance on both the validation and test sets:

Validation Set:

- RMSE: 0.85
- MAE: 0.14
- R-squared: 0.9995

Test Set:

- RMSE: 2.03
- MAE: 0.65
- R-squared: 0.997

4. Recommendations for Reducing Equipment Energy Consumption

Based on the analysis, the following recommendations are proposed to reduce energy consumption:

1. **Optimize Zone 1 and Zone 3 Temperatures:** Prioritize temperature control in Zone 1 and Zone 3, particularly during periods of high equipment activity. Implement dynamic temperature setpoint adjustments or zone-specific cooling/heating systems.
2. **Improve Insulation:** Enhance insulation in Zone 1 and Zone 3 to minimize heat transfer and reduce energy required for temperature maintenance.
3. **Optimize Equipment Scheduling:** If possible, schedule energy-intensive processes during off-peak hours or periods with lower temperature requirements.
4. **Investigate Non-linear Relationships:** Further exploring potential non-linear relationships or interactions involving random variables or other features can help identify additional optimization opportunities.

By implementing these recommendations, SmartManufacture Inc. can assist their client in achieving significant energy savings and cost reductions while maintaining operational efficiency. I hope this report provides valuable insights for your decision-making. Please let me know if you have any further questions or require additional analysis.