Energy Consumption Analysis Report

1 Introduction

This report analyzes equipment energy consumption data from a multi-zone building to identify factors influencing energy usage and develop predictive models. The dataset includes environmental measurements (temperature, humidity) across nine zones, outdoor conditions, and temporal features. The goal is to provide insights into energy consumption patterns and recommend strategies for optimization.

2 Approach

2.1 Data Preprocessing

- Loaded dataset with 16,857 entries and 29 features, including timestamp, energy consumption, zone-specific temperature/humidity, and outdoor conditions.
- Converted timestamp to datetime, extracting hour, day, and weekday features.
- Identified and handled non-numeric values in temperature and humidity columns by converting to numeric (coercing invalid entries to NaN).
- Addressed missing values (approximately 5% per feature) using median imputation for model training.

2.2 Feature Engineering

- Created temporal features: hour, day, weekday, and weekend indicator.
- Calculated zone averages for temperature and humidity.
- Computed differences between indoor (zone averages) and outdoor temperature/humidity.
- Generated interaction terms (e.g., temperature-humidity, wind-temperature) and zone-specific differences to capture environmental impacts.

2.3 Modeling

- Split data into 80% training and 20% testing sets.
- Evaluated three models: Linear Regression, Decision Tree, and Random Forest.
- Used R^2 and RMSE metrics to assess performance.
- Performed cross-validation and learning curve analysis to evaluate model stability and generalization.

3 Key Insights

3.1 Data Exploration

- Energy consumption varies by hour, peaking during daytime, suggesting operational schedules influence usage.
- Zone-specific temperature and humidity show moderate variability, but some zones (e.g., zone 6) exhibit extreme outliers, indicating potential sensor errors.
- Outdoor conditions (temperature, humidity) have weak correlations with energy consumption, suggesting indoor factors dominate.
- Missing values (up to 888 entries for some features) and non-numeric entries (e.g., 75 in zone1_temperature) indicate data quality issues.

3.2 Correlations and Feature Importance

- Top correlations with energy consumption: hour (0.121), zone humidity differences (0.03–0.05), and average zone temperature (0.041).
- Random Forest feature importance highlights zone-specific humidity (e.g., zone 5: 0.031) and temperature differences (e.g., zone 6: 0.032) as key predictors.
- No features show suspiciously high correlations (>0.8), ruling out obvious data leakage.

4 Model Performance Evaluation

- **Linear Regression**: $R^2 = 0.004$, RMSE = 163.559. Poor performance indicates non-linear relationships.
- **Decision Tree**: $R^2 = -1.357$, RMSE = 251.554. Severe overfitting, as it performs worse than a mean predictor.
- **Random Forest**: $R^2 = 0.059$, RMSE = 158.923. Best performer but still low explanatory power.

- Cross-validation for Random Forest shows high variance (mean R^2 = -0.106, std = 0.112), indicating model instability across data subsets.
- Learning curve suggests Random Forest overfits on training data, with poor generalization to test data.

4.2 Interpretation

The low R^2 scores across models suggest that the current features do not adequately capture the drivers of equipment energy consumption. Potential reasons include:

- Complex non-linear relationships not fully captured by the models.
- Data quality issues (outliers, missing values, non-numeric entries).
- Missing critical features, such as equipment type, operational status, or occupancy data.

5 Recommendations

5.1 Energy Reduction Strategies

- **Optimize Scheduling**: Since energy consumption peaks during daytime hours, implement off-peak operation schedules or automated shut-off during low-occupancy periods.
- **Environmental Control**: Maintain consistent zone temperatures and humidities, as differences (e.g., zone 6) correlate with higher consumption. Upgrade HVAC systems for better regulation.
- **Equipment Maintenance**: Investigate zones with extreme temperature/humidity readings (e.g., zone 6: -42.99°C to 55.93°C) for faulty sensors or inefficient equipment.

5.2 Data and Modeling Improvements

- **Data Quality**: Implement stricter data validation to eliminate non-numeric entries and reduce missing values. Cross-validate sensor readings for consistency.
- **Additional Features**: Collect data on equipment specifications, occupancy patterns, and maintenance logs to improve model explanatory power.
- **Advanced Modeling**: Explore gradient boosting (e.g., XGBoost) or neural networks to capture complex relationships. Perform hyperparameter tuning for Random Forest.

6 Conclusion

The analysis reveals that equipment energy consumption is influenced by temporal patterns and zone-specific environmental conditions, but current models struggle to predict usage accurately due to data limitations and complex relationships. By optimizing operational schedules, improving environmental controls, and enhancing data quality, significant energy savings are achievable. Future work should focus on collecting richer datasets and experimenting with advanced modeling techniques.