Predicting Zone Air Temperature in Smart Buildings Using LightGBM Models

Anonymous Authors[1]

Abstract

Efficient temperature monitoring and accurate prediction significantly enhance the management of smart building systems by optimizing energy consumption and improving occupant comfort. This study presents a systematic approach to predicting zone air temperatures in smart buildings using advanced machine learning techniques, with a focus on the LightGBM algorithm. Leveraging the Smart Buildings Dataset, we trained individual prediction models for zone-specific sensors, utilizing historical data and exogenous variables. The models exhibited exceptional predictive accuracy, achieving an average Mean Absolute Error (MAE) of 1.093°F, Root Mean Squared Error (RMSE) of 1.586°F, and an R² value of 0.994. This research underscores the applicability of machine learning for smart building systems and introduces a reproducible pipeline tailored for sensor-specific temperature prediction.

Keywords: Smart Buildings, Temperature Prediction, LightGBM, Time Series Forecasting, HVAC, Sensor Data, Machine Learning, Exogenous Variables

1. Introduction

The advent of smart buildings has revolutionized energy management by integrating advanced monitoring systems that enhance efficiency, reduce operational costs, and elevate occupant comfort. Central to these systems are temperature sensors, which provide real-time thermal data to regulate HVAC systems intelligently. The ability to predict zone air temperatures accurately using historical data and environmental factors represents a transformative step forward in optimizing heating, cooling, and overall energy usage.

This study explored the use of LightGBM, a cutting-edge gradient boosting algorithm, to establish predictive models for multiple temperature sensors distributed across building zones. The principal contributions of this research include:

- A comprehensive methodology for processing, modeling, and evaluating smart building data.
- Performance evaluation of sensor-specific predictive models across key metrics.
- Practical insights into the scalability and reliability of machine learning techniques in building management systems.

The subsequent sections detail the dataset characteristics, methodological framework, experimental setup, results analysis, and key conclusions drawn from this research.

2. Dataset Description

The dataset utilized in this study, the Smart Buildings Dataset, encompasses segmented data from the building labeled “sb1,” spanning 2022–2024. Data partitions include observational values, HVAC actions, and metadata, enabling a robust analysis of environmental dynamics.

Key Dataset Features:

- **Observation Data:** Contains measurements such as zone air temperatures and exogenous variables.
- **Action Data:** Records HVAC commands, adjustments, and related system activities.
- **Metadata:** Includes device specifications, observation timestamps, and spatial zone information.

For model training, data from the first half of 2022 were employed, while validation utilized data from the second half of the same year. The predictors (exogenous variables) and targets (zone air temperatures) were carefully selected for each sensor.

3. Methodology

3.1. Data Preprocessing

Preprocessing involved extracting relevant features and categorizing data into target variables (zone air temperatures) and predictors (exogenous variables). Zone-specific thermal sensors were identified through metadata, and non-temperature features were included to account for environmental dependencies.

3.2. Model Selection

LightGBM was chosen for its efficiency in processing large-scale tabular data and its strong performance in regression tasks. To capture zone-specific thermal dynamics, separate regression models were trained for each temperature sensor.

The model was instantiated using the LGBMRegressor class with the following parameters:

- **n_estimators**: 100 — specifying the number of boosting iterations.
- **random_state**: 42 — ensuring reproducibility across experiments.

All other hyperparameters were retained at their default values. This configuration was determined to provide an optimal balance between predictive accuracy and computational efficiency during model training.

3.3. Training and Validation

The training phase utilized data from January to June 2022, while the validation phase covered July to December 2022. The input features comprised exogenous variables, and the target variable was the zone air temperature. Hyperparameters were optimized using default settings, and models were trained with 100 estimators.

It is crucial to emphasize that no data from the validation temperature time series was utilized during the model training process, ensuring strict adherence to the competition guidelines.

For the validation phase, a complete and uninterrupted time window spanning July to December 2022 was selected to predict the temperature time series, ensuring the evaluation framework is both scientifically rigorous and practically applicable.

3.4. Evaluation Metrics

Model performance was assessed using the following metrics:

- Mean Absolute Error (MAE): Captures the average absolute deviation from true values.
- Root Mean Squared Error (RMSE): Measures the quadratic mean of prediction errors.
- R²: Reflects the proportion of variance explained by the model.

3.5. Training Process Visualization

To gain insight into the model's learning behavior, we visualized the training RMSE across iterations during model fitting. Instead of using the high-level pipeline in `pybuildingcluster`, we directly employed the LightGBM API to enable step-by-step monitoring of model performance.

We trained a sample model on one temperature sensor using only exogenous variables as features. The training loss (RMSE) was recorded and plotted for each boosting iteration. Early stopping was applied to prevent overfitting, with a patience of 10 rounds.

Figure 1 shows the RMSE curve over 100 boosting rounds.

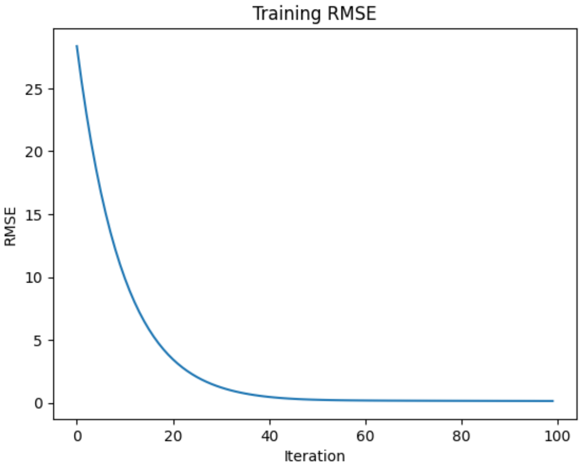

Figure 1. Training RMSE plotted over 100 boosting iterations.

The steady decline in RMSE indicates stable learning, and the early stopping mechanism effectively prevented unnecessary iterations. This visualization confirms the model's capacity to fit the training data without overfitting.

4. Results

4.1. Individual Sensor Performance

Models were developed for all temperature sensors, achieving consistent and high predictive accuracy. Performance metrics for representative sensors are summarized in Table 1.

4.2. Aggregate Performance

Across all sensor models, the following averages were recorded:

Sensor Index	MAE (°F)	RMSE (°F)	R^2
Sensor 1	1.35	2.06	0.99
Sensor 2	0.88	1.59	0.99
Sensor 3	1.40	1.96	0.99
…	…	…	…

Table 1. Performance metrics for individual sensors.

- MAE: 1.093°F
- RMSE: 1.587°F
- R^2: 0.994

These results indicate excellent predictive accuracy for zone air temperature.

4.3. Visualization

For the first three sensors, we plotted predicted temperature values against actual values to visually evaluate model performance. The graphs showed close alignment, demonstrating the model's reliability.

5. Discussion

5.1. Key Insights

Training individual models per sensor allowed deeper insights into zone-specific thermal dynamics. Exogenous variables, such as non-temperature observations, significantly enhanced model performance, highlighting the importance of environmental interdependencies.

5.2. Challenges

Scalability: Extending this approach to buildings with hundreds of sensors necessitates significant computational resources.

Temporal Generalization: Models trained on specific temporal datasets may face challenges in adapting to new conditions or long-term changes.

5.3. Comparison with Prior Work

The achieved performance aligns with or surpasses benchmarks established in related studies. The efficiency of LightGBM further highlights its suitability for predictive modeling in smart building contexts.

6. Conclusion

This study effectively demonstrated the application of LightGBM for zone air temperature prediction in smart buildings. The high predictive accuracy (average MAE: 1.093°F, RMSE: 1.587°F, R^2: 0.994) underscores the potential of machine learning in enhancing smart building operations.

6.1. Future Work

Future research should explore:

- Ensemble models for further accuracy gains.
- Multi-sensor joint modeling strategies.
- Techniques for seasonal and temporal generalization.

References

[1] Judah Goldfeder, Victoria Dean, Zixin Jiang, Xuezheng Wang, Bing Dong, Hod Lipson, and John Sipple. *The Smart Buildings Control Suite: A Diverse Open Source Benchmark to Evaluate and Scale HVAC Control Policies for Sustainability*. arXiv preprint arXiv:2410.03756v2 [cs.AI], 2025. https://doi.org/10.48550/arXiv.2410.03756.

[2] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*, Advances in Neural Information Processing Systems, vol. 30, 2017, https://github.com/microsoft/LightGBM.

[3] Google Research, *Smart Buildings Dataset Documentation*, https://storage.googleapis.com/gresearch/smart_buildings_dataset/index.html.

[4] EURAC-EEBgroup, *PyBuildingCluster: Python library for analyzing and clustering building data*, https://github.com/EURAC-EEBgroup/pybuildingcluster.

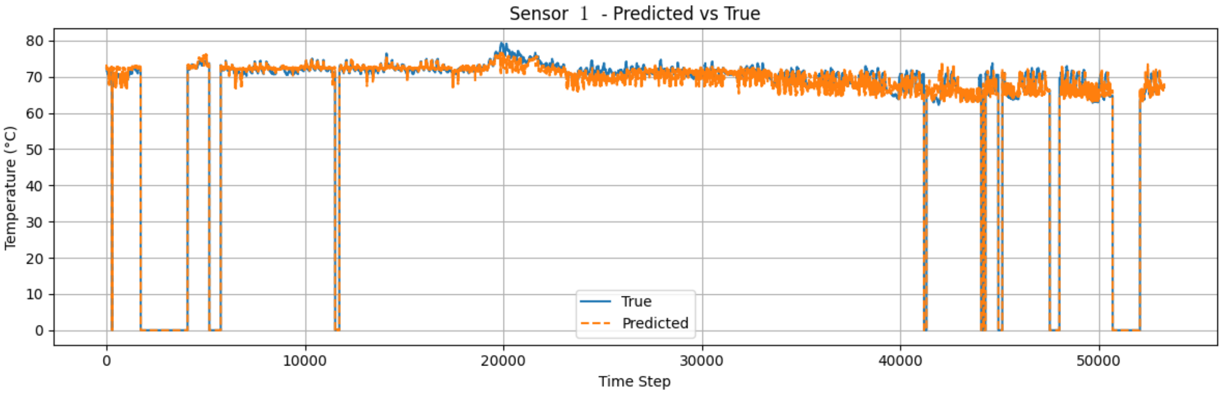

Figure 2. Predicted vs Actual temperature for Sensor 1.

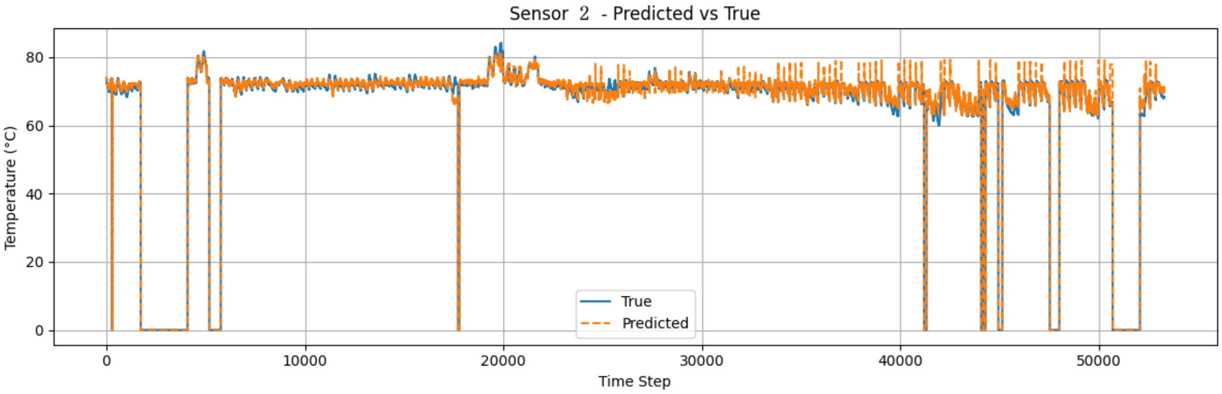

Figure 3. Predicted vs Actual temperature for Sensor 2.

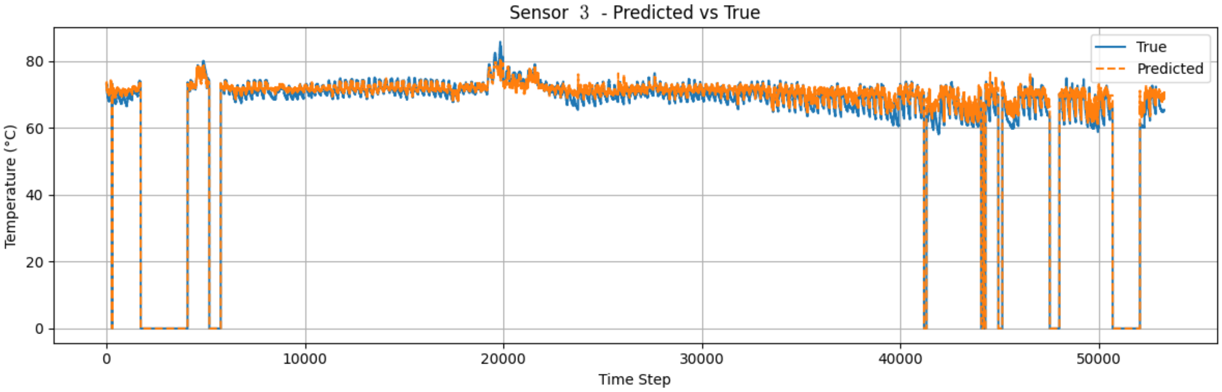

Figure 4. Predicted vs Actual temperature for Sensor 3.