Devdutt & Steven | 12/2/2025 | Group 6

Early Prediction of Residential Energy Consumption

Abstract

This presentation introduces a **Machine Learning (ML) model** designed to accurately **predict future energy consumption** across multiple categories, including electricity, natural gas, and other utilities. The model utilizes historical usage data, seasonal trends, and external factors (e.g., weather and operational schedules) to forecast demand with high precision.

Introduction - Energy Consumption Prediction

Residential buildings account for a large share of total energy use and emissions.[13]
Top determinants influencing national energy include using electricity for space and water heating [13]

Utilities and policymakers need **accurate demand predictions** to plan infrastructure and design efficiency programs.

Household-level energy use is influenced by many interacting factors:

- Building size and age
- Climate and weather
- Heating/cooling equipment and fuels
- Occupant behavior [13]

Literature Review

Traditional Statistical Models:

- Conventional methods such as time series analysis, regression, and exponential smoothing have been widely used for their ability to capture seasonal patterns and trends.
- However, these engineering-based methods often fail to capture the complex, nonlinear relationships inherent in modern energy systems.[2]

Machine Learning (ML) Approaches:

- Recent studies suggest ML techniques offer superior adaptability and prediction accuracy compared to traditional methods.[6] [7]
- Commonly employed algorithms include Support Vector Machines (SVM), Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), Random Forest (RF), and Gradient Boosting. [6]
- Tree-based methods like Decision Trees and Random Forest are valued for their interpretability and ability to handle nonlinear relationships.[6] [3]

Literature Review - Previous Work

Tested Algorithms:

- Random Forest[1][3][6]
- Feed Forward Neural Networks[3]
- Support Vector Regression[3]
- Long Short-Term Memory[3]
- Gaussian Process Regression[3]
- Linear regression[8]
- Convoluted Neural Networks[2]
- Deep Neural Network[4]

Common limitations:

- One-hot encoding of many categorical variables which equals high dimensionality.
- Limited attention to **feature engineering** (climate/occupancy interactions).

Our Approach:

CatBoost

Hypothesis:

- H1: *CatBoost*, with its native handling of categorical variables and ordered boosting, will yield better results than XGBoost and Random Forest on this dataset.
- H2: Feature engineering (climate-degree-day features, occupant density, equipment counts, etc.) will significantly improve prediction accuracy compared with using raw features only

Dataset

Residential Energy Consumption Survey 2020

- The **Residential Energy Consumption Survey (RECS)**, administered by the U.S. Energy Information Administration (EIA), is the premier source of data on energy usage in American homes. [5]

Target Feature:

TOTAL BTU = *total annual site energy use* for the home, measured in BTU, summing all fuels (electricity, natural gas, propane, fuel oil, wood, other). It's the best single "how much energy was used" number, independent of fuel type [5]

18,496

Total Samples

A robust dataset representing millions of U.S. households.

798

Total Features

Includes **29 Categorical** variables covering diverse attributes.

Methodologies - Software

- Visual Studio
- Jupyter Notebooks
- Python

Methodologies - Libraries

- Scikit Learn
- CatBoost
- xgboost
- Pandas
- Numpy
- Matplot
- seaborn

Methodologies - Random forest and XGboost

Random forest was implemented with Scikit-Learn (90/10 train test split).

XGboost was implemented using XGBRegressor (90/10 train test Split).

Reasons for picking these two:

- Easy to implement
- Industry Standards
- Would work well on our dataset
- Provides interpretable feature importance metrics.

Methodologies - CatBoost

(90/10 train test split)

CatBoost is a machine learning library for gradient boosting on decision trees, developed by Yandex. It's designed to work especially well with tabular data that includes **categorical features**, which it can handle directly without heavy preprocessing. Like XGBoost and LightGBM, it's used for tasks such as classification and regression, often giving strong performance with minimal tuning.[15]

- **Pros:**

  Handles categorical features natively
  Often strong performance with sensible default parameters
  Built-in handling of missing values
  Provides useful tools like feature importance and efficient evaluation

- **Cons:**

  Training can still be slower than simpler models (e.g., linear models, small trees)
  Tuning for best performance can be time-consuming
  Less ubiquitous ecosystem and documentation compared to libraries like XGBoost/LightGBM Model interpretability is lower than linear or simple tree-based models

Results - Test Accuracy CatBoost

Performance: R^2 = **0.9932**, MAE ≈ **1,327 BTU** → most accurate model.

Prediction behavior: Predicted vs actual points lie very close to the 45° “perfect prediction” line, even for high energy values.

Residuals: Small, tightly clustered around zero with no strong pattern → low bias and variance.

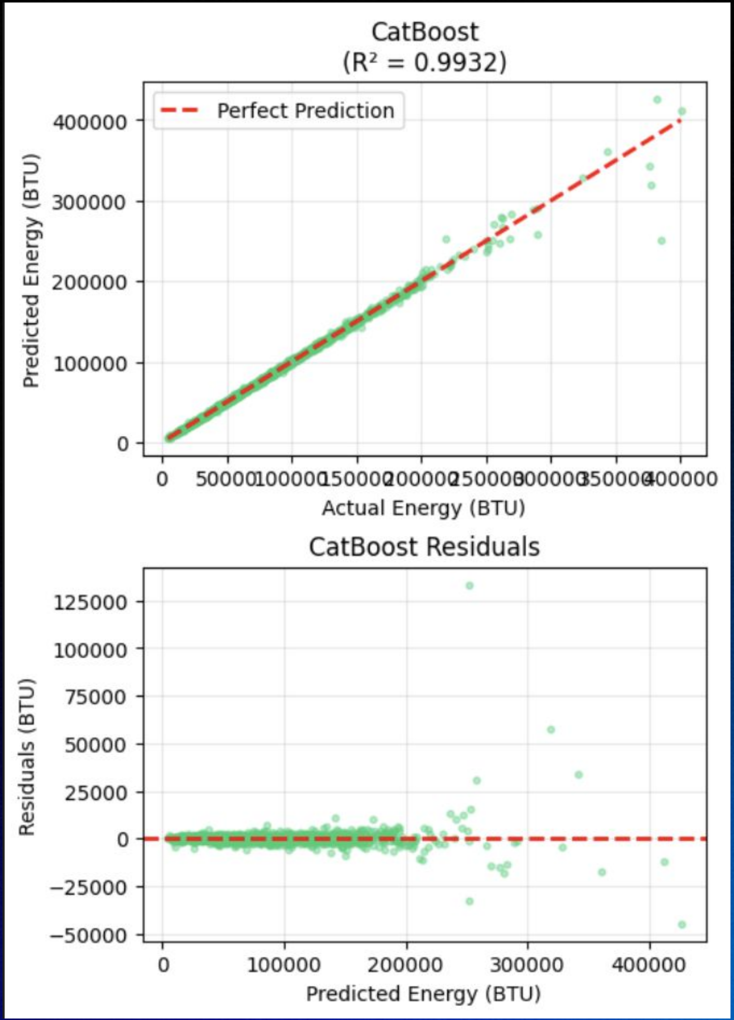


Figure 1.Catboost stats

Results - Test Accuracy XGboost

- **Performance:** R^2 = **0.9890**, MAE ≈ **1,753 BTU** → slightly worse than CatBoost but still very strong.
- **Prediction behavior:** Points follow the perfect-prediction line, with a bit more scatter, especially at higher BTU values.
- **Residuals:** Mostly centered around zero but with higher spread than CatBoost → slightly higher error and variance.

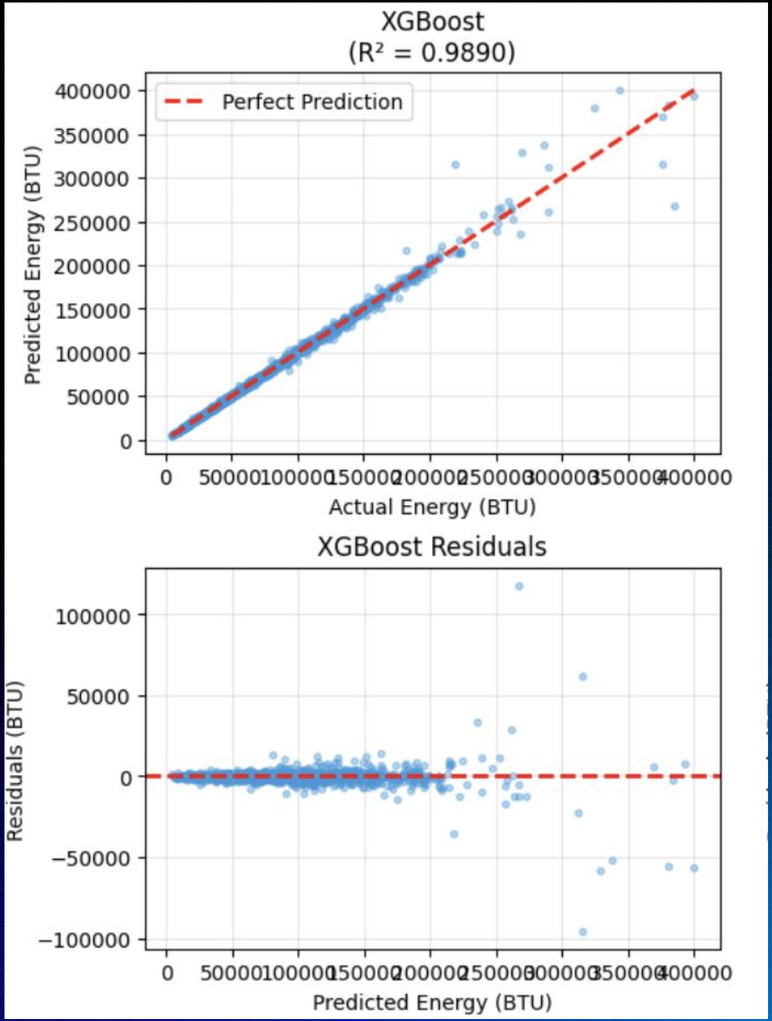


Figure 2.XGboost stats

Results - Test Accuracy Random Forest

- **Performance:** R^2 = **0.9511**, MAE ≈ **6,261 BTU** → clearly less accurate than CatBoost and XGBoost.
- **Prediction behavior:** Larger spread around the perfect-prediction line, particularly at higher energy levels.
- **Residuals:** Wide, more dispersed residuals with noticeable structure → higher systematic error.

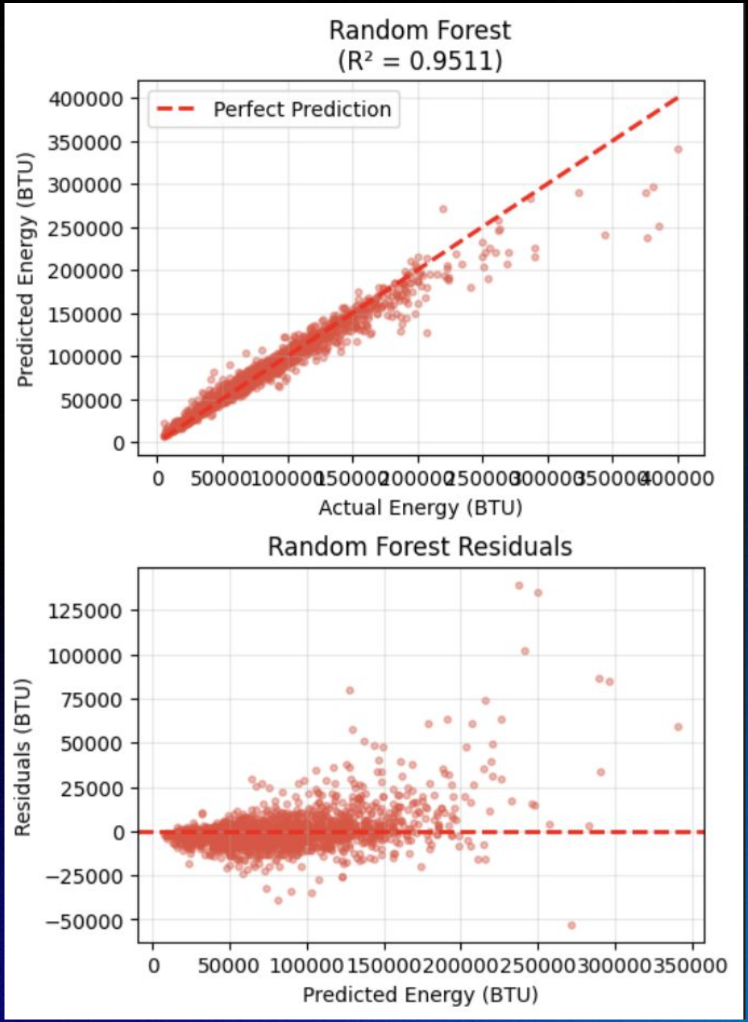


Figure 3.Random Forest stats

Results - Model Comparison

Accuracy: CatBoost > XGBoost >> Random Forest

- Best R^2 and lowest MAE from CatBoost.

Error patterns:

- CatBoost has the tightest residuals and least bias.
- XGBoost is close behind with slightly more residual spread.
- Random Forest shows the largest and most structured residuals.

Training time: Random Forest (fastest) < XGBoost < CatBoost (slowest).

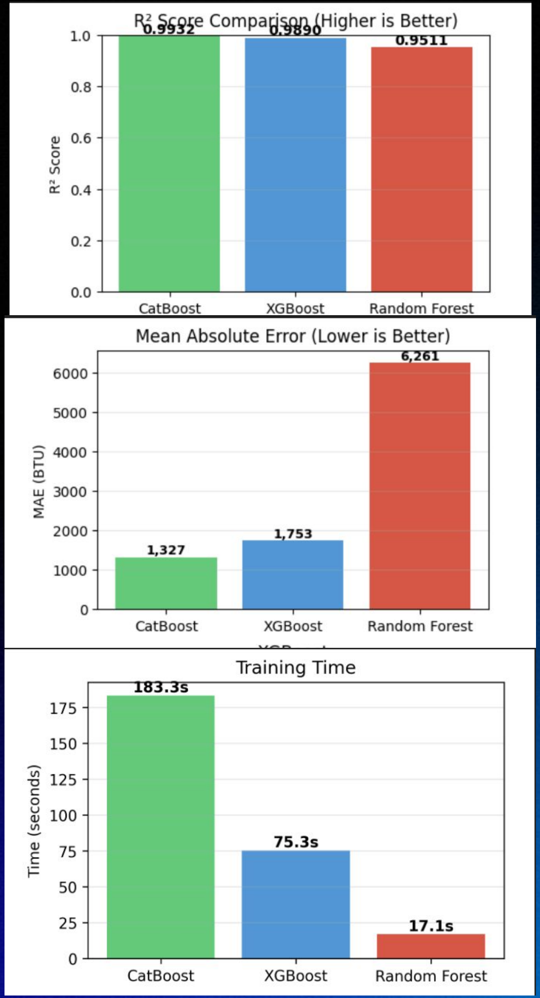


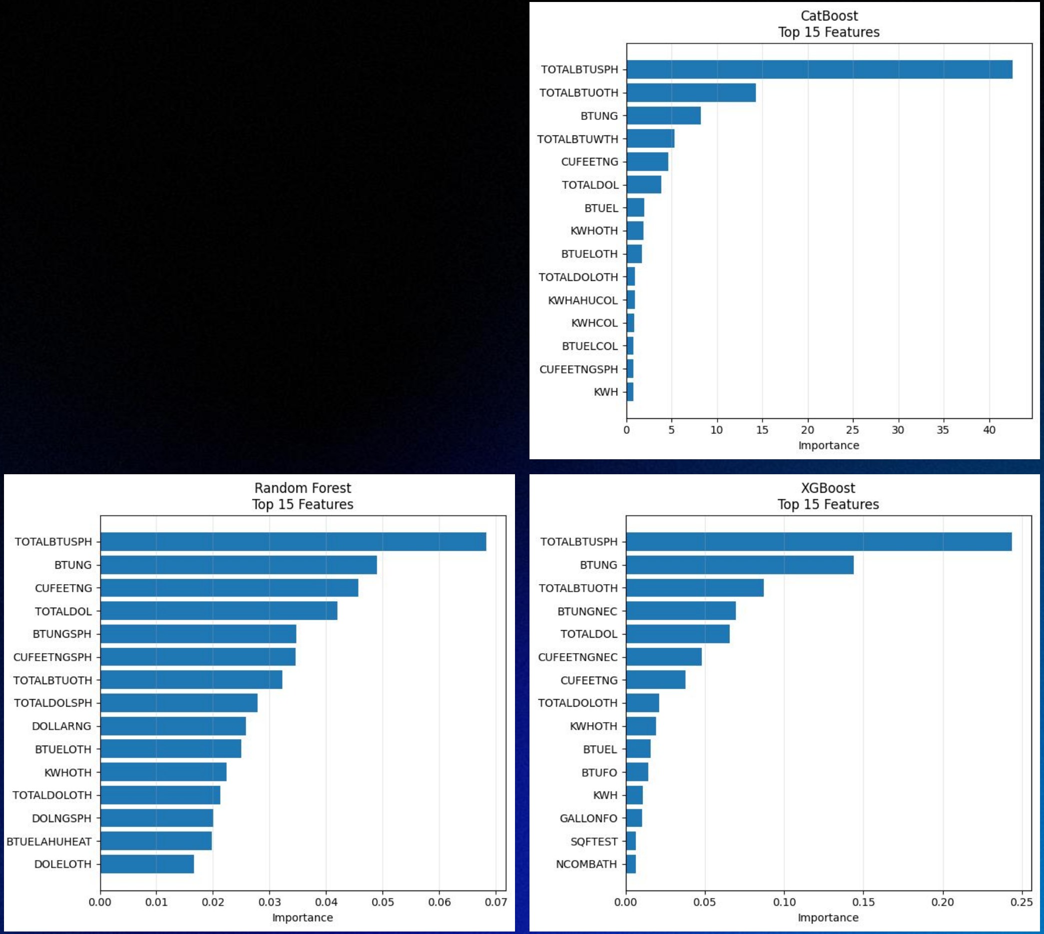
Figure 4. Comparing stats

Results - Feature Importance

Top drivers from our feature-importance chart (consensus across CatBoost, XGBoost, RF):

1. **Space heating energy (TOTALBTUSPH)** is the #1 driver.
2. **Natural gas usage** (e.g., BTUNG, CUFEETNG) has a strong influence on total energy.
3. **Water heating energy (TOTALBTUWTH)** contributes significantly.
4. **Other loads (TOTALBTUOTH)** and **electricity** add meaningful share.
5. **Cost variables** (e.g., TOTALDOL) appear as **proxies** for overall consumption.

Figure 5.Feature Importance

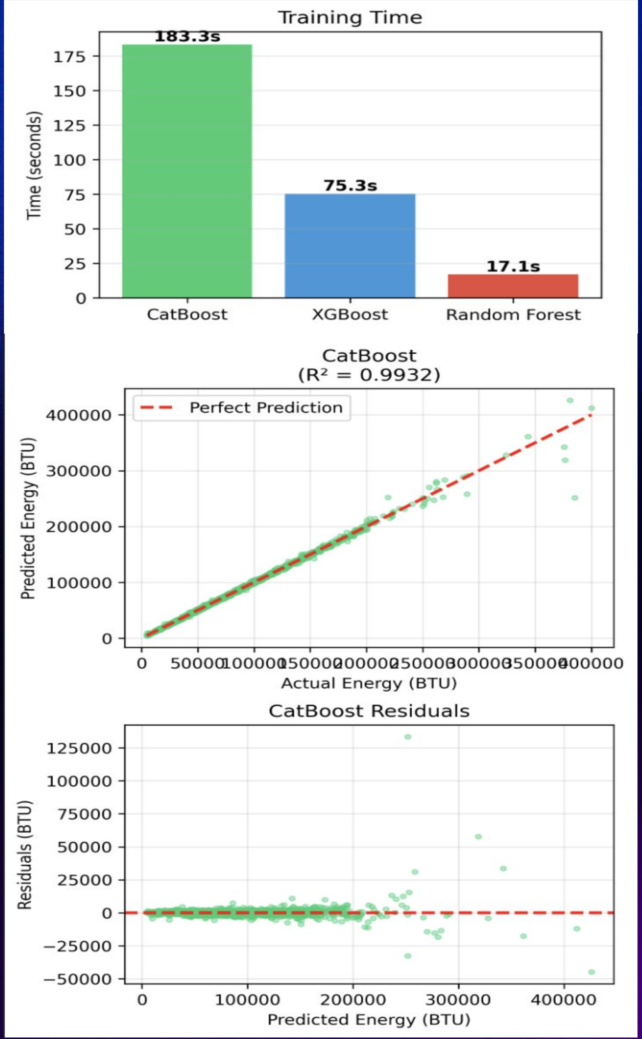


Contributions

Contributions

Demonstrated the efficacy of **CatBoost** for residential energy prediction, minimizing the need for extensive data preprocessing

- *Previous Models:* Focused on traditional ML (ANN, SVM, DT) and generic GB.
- *Benefit:* Successfully modeled complex, heterogeneous data without the high dimensionality issues often seen in One-Hot Encoding.
- Time was the only constraint compared to other models.



Discussion - Conclusion

We Choose to keep Moving Forward with Catboost.

- Research Gap- not previously explored by any studies and papers we found
- Practical Advantages-Intuitive workflow, Minimal preprocessing, Well-maintained library with extensive examples
- Easier to understand compared to Physics-Informed Neural Networks,Graph Neural Networks etc
- Worked better against random forest in our preliminary results.

Discussion - Limitations Of Catboost

Training Time

- Trees must be built one after another
- Not fully parallelizable
- Scales poorly with data size

Memory Consumption

- High RAM requirements
- Categorical feature overhead: Maintains statistics for all categorical combinations
- Gradient computation: Requires storing gradients for all samples

Feature Complexity

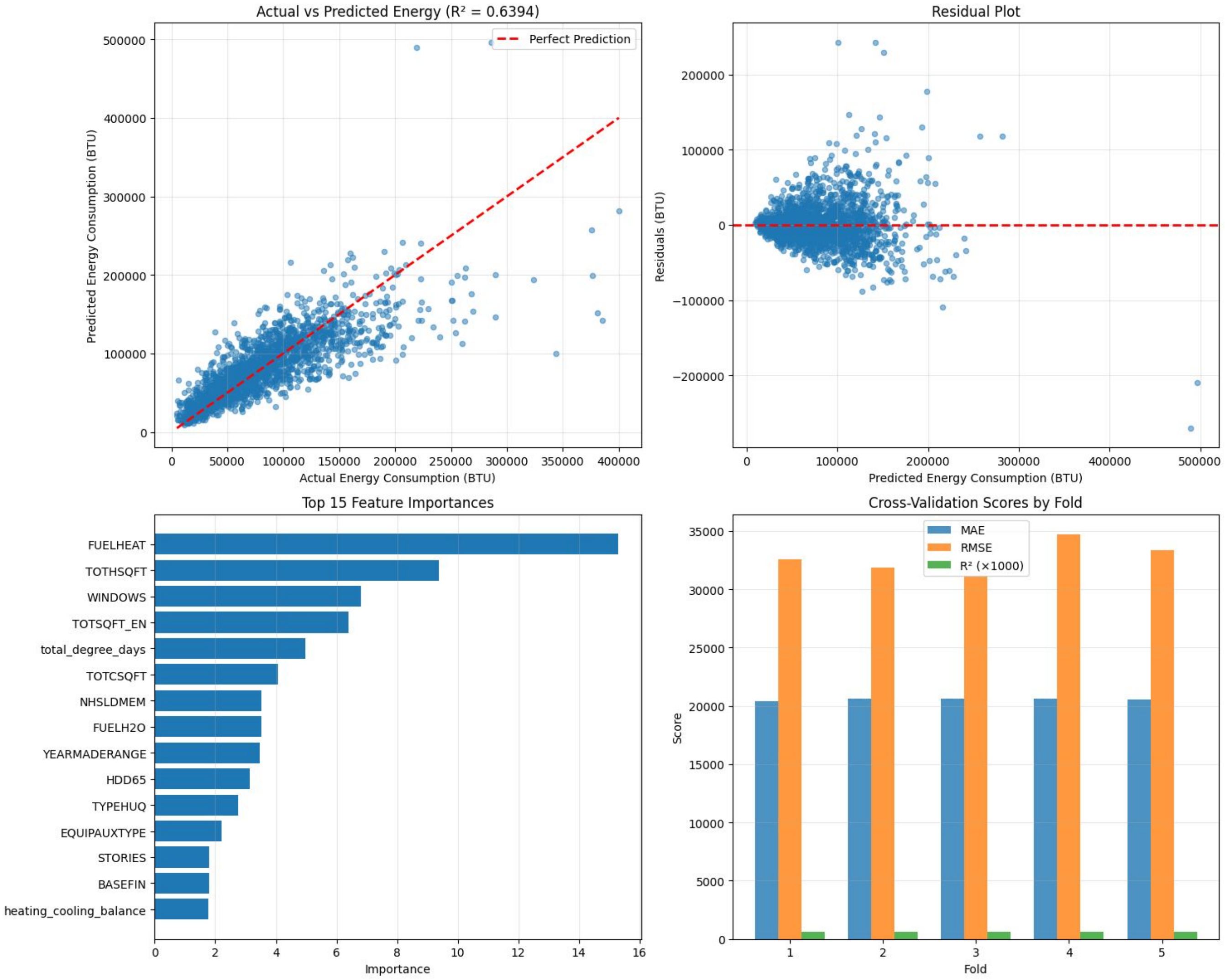
- High Performance degrades with many unique categories

Discussion - CatBoost with Stratified K-Fold Cross-Validation + Hold-Out Test Set

Tried this method to check on overfitting.

- **1. Initial Data Partition (90/10 Split)**
  - **Full Dataset:** 18,496 samples.
  - **Training Set (90%): 16,646 samples** are Used for all subsequent tuning and model fitting.
  - **Test Set (10%): 1,850 samples** are Held out until the very end.
- **2. Stratified 5-Fold Cross-Validation (CV)**
  - Applied to the **Training Set (16,646 samples)**
  - Each fold involved **80% Training** and **20% Validation**.
  - The outcome was the **average performance metric** across the 5 validation folds.
- **3. Final Model Training & Evaluation**
  - The final **CatBoost** model was trained on the **ENTIRE Training Set (16,646 samples)**.
  - Evaluation was performed **only once** on the **untouched Test Set (1,850 samples)**.
  - This result provides the **final, true performance score** on unseen data.

Results:


Actual vs Predicted Energy (R² = 0.6394)
Perfect Prediction
Predicted Energy Consumption (BTU)
Actual Energy Consumption (BTU)
Residual Plot
Residuals (BTU)
Predicted Energy Consumption (BTU)
Top 15 Feature Importances
FUELHEAT
TOTHSQFT
WINDOWS
TOTSQFT_EN
total_degree_days
TOTCSQFT
NHSLDMEM
FUELH2O
YEARMADERANGE
HDD65
TYPEHUQ
EQUIPAUXTYPE
STORIES
BASEFIN
heating_cooling_balance
Importance
Cross-Validation Scores by Fold
MAE
RMSE
R² (×1000)
Score
Fold

Discussion - Future Work

Apply CatBoost algorithm to the Residential Energy Consumption Survey dataset to validate model performance on a different, comprehensive residential building dataset

- **Check if it's overfitting:** Apply catBoost with K-Fold it on RECS 2021 and look for results.
- **Data Cleaning & Preprocessing**
- **Train-Test Split Strategy(80,10,10)**
- **Model Development & Comparison to DNN, ANN, RF, etc.**

Question?

Reference List

[1] M. Fellah, S. Ouhaibi, N. Belouaggadia, and K. Mansouri, “Energy Consumption Forecasting and Thermal Insulator Selection with Random Forest Regression,” Scientific African, p. e02870, Jul. 2025, doi: https://doi.org/10.1016/j.sciaf.2025.e02870.

[2] Mehdi Neshat, Menasha Thilakaratne, M. El-Abd, Seyedali Mirjalili, A. H. Gandomi, and J. Boland, “Smart buildings energy consumption forecasting using adaptive evolutionary bagging extra tree learning models,” Energy, vol. 333, pp. 137130–137130, Jul. 2025, doi: https://doi.org/10.1016/j.energy.2025.137130.

[3] O. Gulaydin and M. Mourshed, “Machine learning for subnational residential electricity demand forecasting to 2050 under shared socioeconomic pathways: Comparing tree-based, neural and kernel methods,” Energy, vol. 336, p. 138195, Sep. 2025, doi: https://doi.org/10.1016/j.energy.2025.138195.

[4] W. Xu, J. Tu, N. Xu, and Z. Liu, “Predicting Daily Heating Energy Consumption in Residential Buildings through Integration of Random Forest Model and Meta-Heuristic Algorithms,” Energy, pp. 131726–131726, May 2024, doi: https://doi.org/10.1016/j.energy.2024.131726.

Reference List

[5] “U.S. Energy Information Administration - EIA - Independent Statistics and Analysis,” www.eia.gov. https://www.eia.gov/consumption/residential/data/2020/index.php?view=microdata

[6] P. B. Asamoah and E. Shittu, “Evaluating the performance of machine learning models for energy load prediction in residential HVAC systems,” Energy and Buildings, vol. 334, p. 115517, Feb. 2025, doi: https://doi.org/10.1016/j.enbuild.2025.115517.

[7] K. H. Baesmat, E. E. Regentova, and Y. Baghzouz, “A Hybrid machine learning–statistical based method for short-term energy consumption prediction in residential buildings,” Energy and AI, vol. 21, p. 100552, Jul. 2025, doi: https://doi.org/10.1016/j.egyai.2025.100552.

[8] G. SRIRAM, “Energy Consumption Dataset - Linear Regression,” Kaggle.com, 2024. https://www.kaggle.com/datasets/govindaramsriram/energy-consumption-dataset-linear-regression/data (accessed Oct. 22, 2025).

Reference List

[9] R. B. Ayoola et al., “Data-driven optimisation of residential air-to-water heat pump performance using IoT and machine learning,” Energy and Buildings, vol. 348, p. 116352, Aug. 2025, doi: https://doi.org/10.1016/j.enbuild.2025.116352.

[10] M. Bozorgi, S. H. Tasnim, and S. Mahmud, “Machine learning-driven hybrid cooling system for enhanced energy efficiency in multi-unit residential buildings,” Energy and Buildings, vol. 336, p. 115613, Mar. 2025, doi: https://doi.org/10.1016/j.enbuild.2025.115613.

[11] Y. Li, H. Zhang, X. Shen, and K. Qu, “Interpretable machine learning for predicting and optimizing residential building performance in cold regions,” Energy and Buildings, vol. 347, p. 116321, Aug. 2025, doi: https://doi.org/10.1016/j.enbuild.2025.116321.

[12] L. Sun, Z. Hu, M. Mae, and T. Imaizumi, “Deep transfer learning strategy based on TimesBlock-CDAN for predicting thermal environment and air conditioner energy consumption in residential buildings,” Applied Energy, vol. 381, pp. 125188–125188, Dec. 2024, doi: https://doi.org/10.1016/j.apenergy.2024.125188.

Reference List

[13] S. S. Korsavi, R. Azari, L. D. Iulo, and M. Mahdavi, “Determinants of U.S. residential energy consumption at national and state levels: Policy implications,” Energy Policy, vol. 202, p. 114594, Jul. 2025, doi: https://doi.org/10.1016/j.enpol.2025.114594.

[14] I. A. Kachalla, C. Ghiaus, and M. Baseer, “Comparative analysis of machine learning models for prediction and forecasting of electric water boilers energy consumption,” Applied Thermal Engineering, vol. 267, p. 125799, Feb. 2025, doi: https://doi.org/10.1016/j.applthermaleng.2025.125799.

[15] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost Classifier with other Machine Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 738–748, 2020, doi: 10.14569/IJACSA.2020.0111190.