BUCKINGHAMSHIRE NEW UNIVERSITY EST. 1891

BNU Research and Enterprise Conference 2024

Building Energy Usage Predictions Using Machine Learning Methods

Dr Shahadate Rezvy, Ms Tasnim Akther, Dr Tahmina Zebin
Corresponding author email: shahadate.rezvy@bnu.ac.uk

1. Introduction and Research Objectives

Buildings are major global energy consumers, responsible for substantial electricity use and CO2 emissions. Enhancing their **energy efficiency is crucial for climate change mitigation and sustainable urban development**. Traditional energy prediction methods often inadequately address the dynamics of building energy use influenced by factors like type, occupancy, and local weather. Given the context and challenges outlined, this study aims to apply the potential of machine learning to advance the predictive accuracy of building energy consumption, specifically focusing on the Site Energy Usage Intensity (Site EUI) of buildings using publicly available energy consumption datasets. The objectives of this research are:

1. **Feature Analysis and Engineering:** Identify and engineer key features from building and weather data to boost prediction accuracy.
2. **Impact of Weather and Geographical Variability:** Analyze how weather and location differences across states influence energy usage.
3. **Model Development and Optimization:** Develop multiple machine learning models, comparing their effectiveness in predicting Site EUI.
4. **Predictive Performance Evaluation:** Use metrics such as RMSE and R² to assess model accuracy.
5. **Application and Policy Implications:** Discuss the potential applications of the study's findings in policy-making and energy management.

Dataset: In this research, we are utilising a unique and publicly available dataset [1]—comprising roughly 75,757 observations of building energy usage across various U.S. states collected over seven years. This dataset includes detailed building characteristics, weather data, and historical energy consumption metrics, presenting an optimal opportunity to refine and enhance predictive models using advanced ML techniques.

2. Preprocessing and Feature Engineering

As this real-world dataset came with missing values, outliers and a few other redundant information, we have applied a thorough cleaning and feature engineering stages outlined table. These stages transform the raw data into a refined dataset optimised for modelling purposes.

Table 1: Missing value handling and feature engineering stages

Method/Technique	Features Applied	Purpose
KNN Imputation for Missing Value Handling	All features with missing values	To estimate missing values using the nearest neighbours based on a similarity metric.
One-Hot Encoding and Target Encoding	State_Factor, facility_type, building_class	Convert categorical variables into a binary representation to facilitate model understanding, facility type is also converted to broader_facility_type
Seasonal Temperature Analysis	Seasonal subsets of temperature data (winter, spring, summer, autumn)	Extract season-specific temperature statistics to better model seasonal impacts on energy usage and to capture different aspects of temperature data.
Building-Based Features	building_area, floor_energy_star_rating	Create features to represent total area and efficiency per unit area, adding context for the model.
Lag Features	site_eui, energy_star_rating, ELEVATION, temp features	Introduce historical data points to capture trends and changes over time.
Delta Features	site_eui, energy_star_rating, ELEVATION, temp features	Calculate yearly changes to understand the rate and direction of feature changes.
Group-by Transformations	State_Factor, building_class, facility_type, energy_star_rating	Aggregate stats to summarise data based on categorical groups, enhancing the model's contextual understanding.

Exploratory Feature Analysis

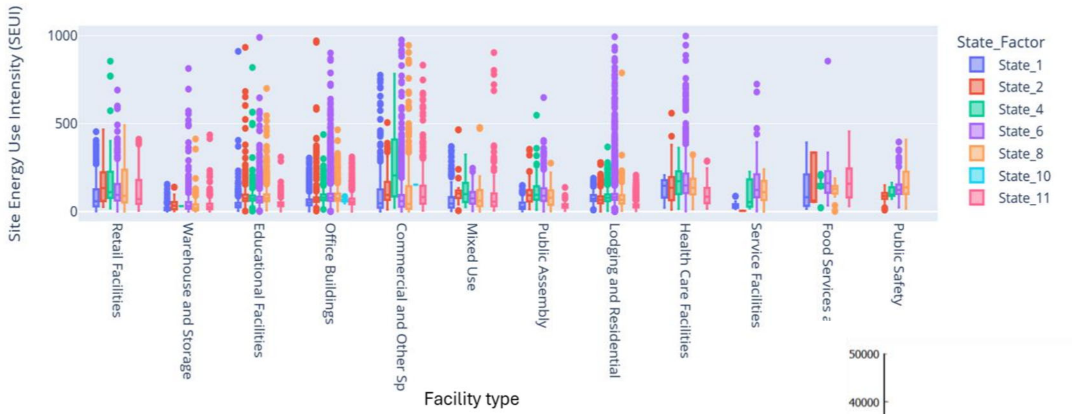

Figure 1: The plot illustrates the distribution of Site Energy Use Intensity (SEUI) across various types of facilities in different states. The y-axis measures SEUI from 0 to 1000(kBtu/ft²/year) . Each facility type displays a distribution of SEUI values represented through a scatter of dots, where each colour corresponds to a different state as indicated by the legend. A wide range of SEUI values across all facility types, suggesting significant variability in energy use intensity dependent on both facility type and geographic location.

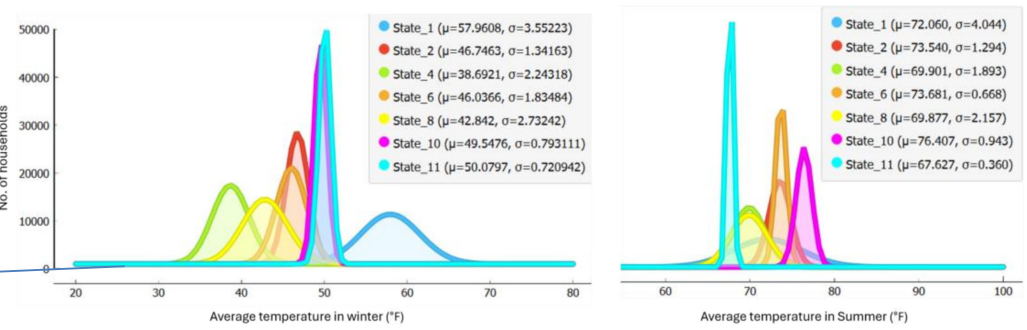

Figure 2: The distribution of average temperature in winter and summer (in °F) across different states. Both plots include labels showing the mean (μ) and standard deviation (σ) of the temperature distribution for each state. This plot also shows distinct peaks for different states, suggesting specific temperature distributions unique to each state. For instance, State_10 and State_11 show peaks around 90°F, indicating a higher frequency of these temperatures.

Model Development and Predictive Performance Evaluation

Implementation Details:
Train-Test Split: 90:10
Cross-validation: 5-fold

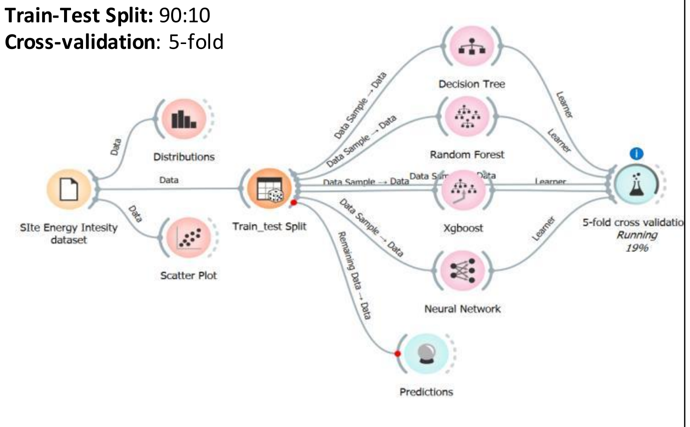

Figure 3: Workflow implemented in Orange 3, Python 3.12, Sklearn libraries.

Model Performance Comparison:

RMSE (Root Mean Square Error): Lower values are better. XGBoost has the lowest RMSE at 49.245, indicating that it has the smallest average error magnitude among the models.

MAE (Mean Absolute Error): Lower values are also better here. XGBoost again performs the best with the lowest MAE at 25.867, showing it generally makes smaller errors in predictions than the other models.

R^2 (Coefficient of Determination): Higher values are better, indicating a model explains more of the variance from the mean. XGBoost scores highest on R^2 as well, with a value of 0.286, suggesting it accounts for a larger portion of the variance in the dataset compared to the others.

Considering all these metrics together, **XGBoost** emerges as the best-performing model when compared to the Decision tree, Random forest and the Neural network Models

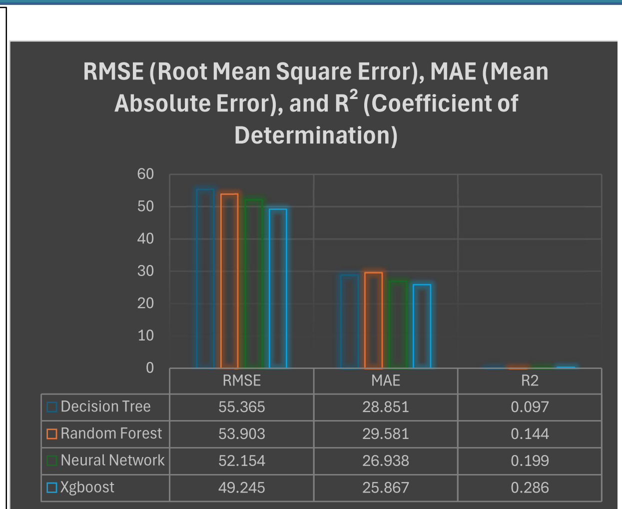

Figure 4: Comparison of the model performance

References:

1. https://www.kaggle.com/competitions/widsdatathon2022/data
2. Rolnick, D., et al. (2023). Tackling Climate Change with Machine Learning. ACM Computing Surveys, 55(2), 1-96. DOI: 10.1145/3485128
3. Qiao, Q., Yunusa-Kaltungo, A., & Edwards, R. E. (2020). Towards developing a systematic knowledge trend for building energy consumption prediction. Journal of Building Engineering, 35, 101967. DOI: 10.1016/j.jobe.2020.101967
4. Bourdeau, M., Zhai, X. Q., Nefzaoui, E., Guo, X., & Chatellier, P. (2019). Modeling and forecasting building energy consumption: A review of data-driven techniques. Sustainable Cities and Society, 48, 101533. DOI: 10.1016/j.scs.2019.101533