UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Academic Year	Module	Assessment Number	Assessment Type

Regression Report

Student Id : 2407743

Student Name : Utkarsha – Aryal

Section : L5CG7

Module Leader : Siman Giri

Tutor : Siman Giri

Submitted on : 11-02-2025

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Table of Contents

Abstract

Purpose : To predict heating load in building using a buildings structure to optimize energy efficiency.

Approach: The energy efficiency dataset which has 8 features and 768 rows was analyzed using EDA , 1st linear regression from scratch was made and after that Random Forest and Ridge Regression done by using scikit learn and finally a fined tuned version of random forest was made again. A column X6 was dropped it was a categorical non ordinal value which did not show correlation with any other data.
Hyperparameter optimization (GridSearchCV/RandomizedSearchCV) and feature selection were implemented.

Key Results: Random Forest performed better than ridge but random forest R^2 came out to be 1 which could suggest that there had been a data leak. Ridge regression also performed well.

Conclusion: Random Forest did better than the other models , but potential overfitting was noted. Key features driving heating load were observed to be Relative Compactness(X1) and Surface Area (X2).

Introduction

Problem Statement

Predicting heating load using a buildings features to increase energy efficiency
Dataset
Source: Kaggle
Description: Contains 8 features (X1,X2,X3,X4,X5,X6,X7,X8) describing building parameters (Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, Glazing Area Distribution) and 2 targets (Y1=heating load, Y2=cooling load) ,but only Y1= heating load is being used
Link to UNSDG: Aligns with **Goal 7 (Affordable and Clean Energy)**

Objective

To develop a regression model to predict heating load and identify critical feature affecting energy efficiency.

2.Methodology

2.1 Data Preprocessing

Handling Missing Values: No missing values.
Scaling: Features standardized using StandardScaler for model consistency.
Train-Test Split: 80/20 split.

2.2 Exploratory Data Analysis (EDA)

Correlation Heatmap: Revealed strong relationships between features (e.g., X1 (Relative Compactness) and X2 (Surface Area) were negatively correlated).
X6 is a categorical data which did not have any correlation with other data so it was dropped.
Key Insight: Compact designs (higher X1) correlate with lower heating loads, suggesting energy efficiency benefits.

2.3 Model Building

Linear Regression (From Scratch): Implemented gradient descent (MSE=9.02,R^2=0.91).
Random Forest (RF): Default model achieved MSE=0.25, R^2=1.00
Ridge Regression: MSE=9.21, R^2=0.91.
Final model with Random Forest (RF) :

2.4 Model Evaluation

Model	MSE	R^2
Linear Regression (Scratch)	9.02	0.91
Random Forest	0.25	1.00
Ridge Regression	9.21	0.91

2.5 Hyperparameter Optimization

Random Forest: GridSearchCV optimized max_depth=20, n_estimators=200, improving robustness.
Ridge Regression: RandomizedSearchCV selected alpha=0.1 to balance bias-variance tradeoff.

2.6 Feature Selection

Random Forest: Top 5 features: X1 (Relative Compactness), X2 (Surface Area), X3 (Wall Area), X5 (Height), X6 (Orientation).
Ridge Regression: RFE selected X1, X2, X4 (Roof Area), X7 (Glazing Area), X8 Glazing Distribution).

3.Conclusion

Key Findings

Random Forest showed near perfect performance likely due to overfitting from correlated features

Ridge Regression provided reliable predictions ($R^2=0.91$) and better generalizability.

Critical features: Compactness (X1) and Surface Area (X2) most strongly influenced heating load.

Final Model

Random Forest is optimal for prediction (MSE=0.25), but Ridge Regression is recommended for interpretability and avoiding overfitting.

Challenges

Overfitting risk in Random Forest due to feature correlations.
Limited dataset size (768 samples) restricted model generalization.

Future work

Collect more data to validate Random Forest's robustness.
Investigate interactions between features (e.g., X1 and X2).