Amy DeMorrow

CS249 Data Science Fundamentals

3/18/2024

Regression Analysis: Energy Efficiency Dataset

ABSTRACT

A regression analysis was performed on an energy efficiency dataset obtained from Kaggle: Energy Efficiency Data Set (kaggle.com). The dataset uses multiple variables related to building design and construction to predict heating and cooling loads for the dwelling. Several models were developed and assessed, with XGBoost selected for the final model. The XGBoost model could achieve prediction R2 scores of 0.9982 for heating load and 0.8951 for cooling load.

INTRODUCTION

With climate change driving increasingly larger swings in global and regional temperatures, the energy efficiency of single-family homes is becoming a market driver in home construction. As a homeowner myself, I am interested to understand how the architectural design of my home drives my utility bills and overall energy consumption. My windows need replacing soon, and I am considering increasing the overall window sizes to gain more "natural light" inside my home. But is the aesthetic trade worth the cost (to my pocketbook and the planet)? I am most interested to see how relative compactness and window area affect heating and cooling loads.

DATA

The Kaggle dataset is comprised of 768 entries, 10 numerical variables, and no null values. All variables are either an integer or a float value (see Figure 1 for basic statistics).

	Relative_Compactness	Surface_Area	Wall_Area	Roof_Area	Overall_Height	Orientation	Glazing_Area	Glazing_Area_Distribution	Heating_Load	Cooling_Load
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000	768.000000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250	22.307201	24.587760
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096	10.090196	9.513306
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000	6.010000	10.900000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000	12.992500	15.620000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000	18.950000	22.080000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000	31.667500	33.132500
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000	43.100000	48.030000

Figure 1. The Energy Efficiency Dataset (Energy Efficiency Data Set (kaggle.com))

Running a Seaborn pairwise plot reveals a hidden classification variable – "overall height" is not continuous, it is either 3.5 or 7.0, which means we have 1 story and 2 story houses (see Figure 2).

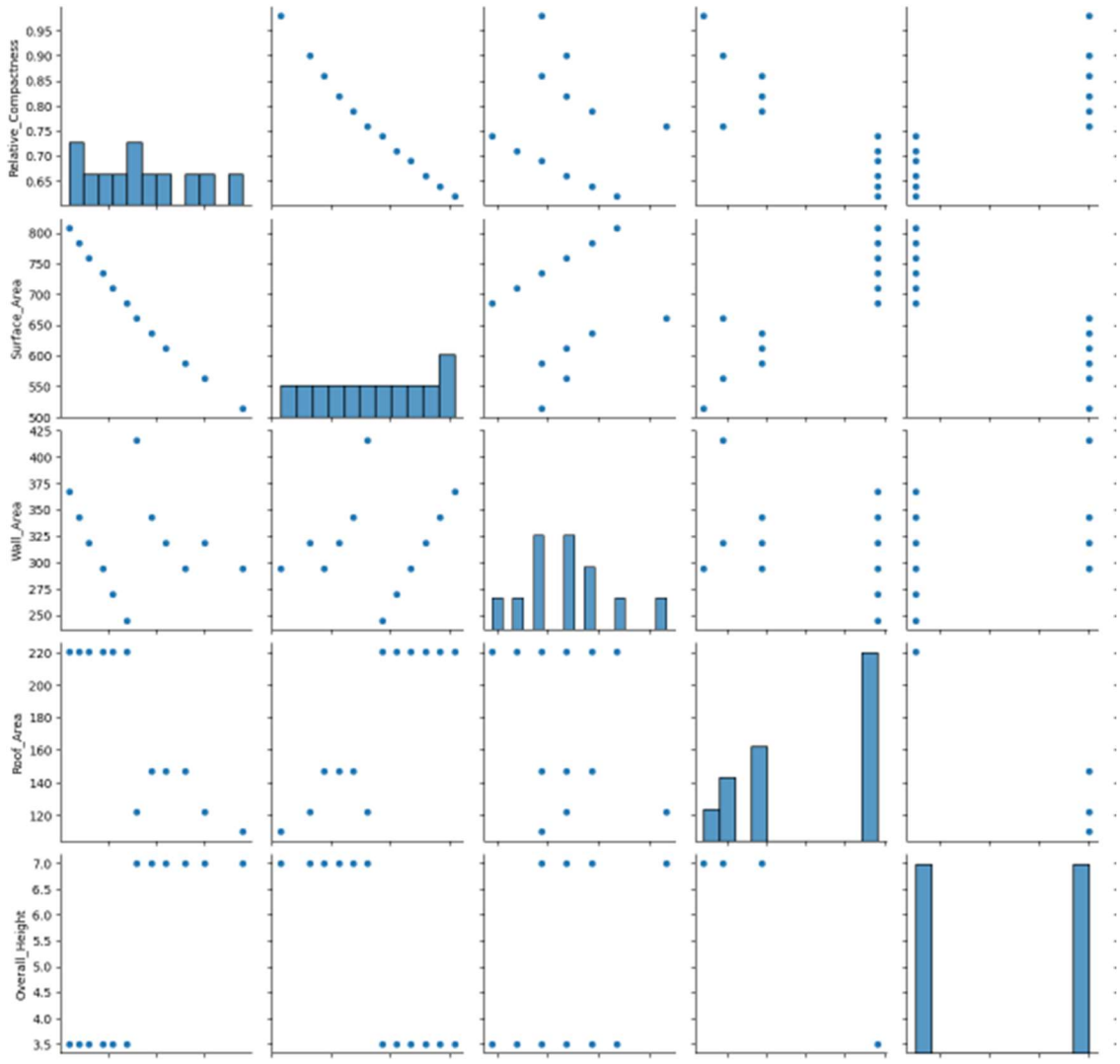

Figure 2. Pairwise plot of variable relationships revealing 1 story and 2 story houses.

Once we rerun the pairwise plot to segment on overall height, we can see how the remaining variables split between 1 story and 2 story houses. The 1 story homes have the largest roof and surface areas and are also the least compact. Wall area and window area and window distribution variables are mixed (see Figure 3):

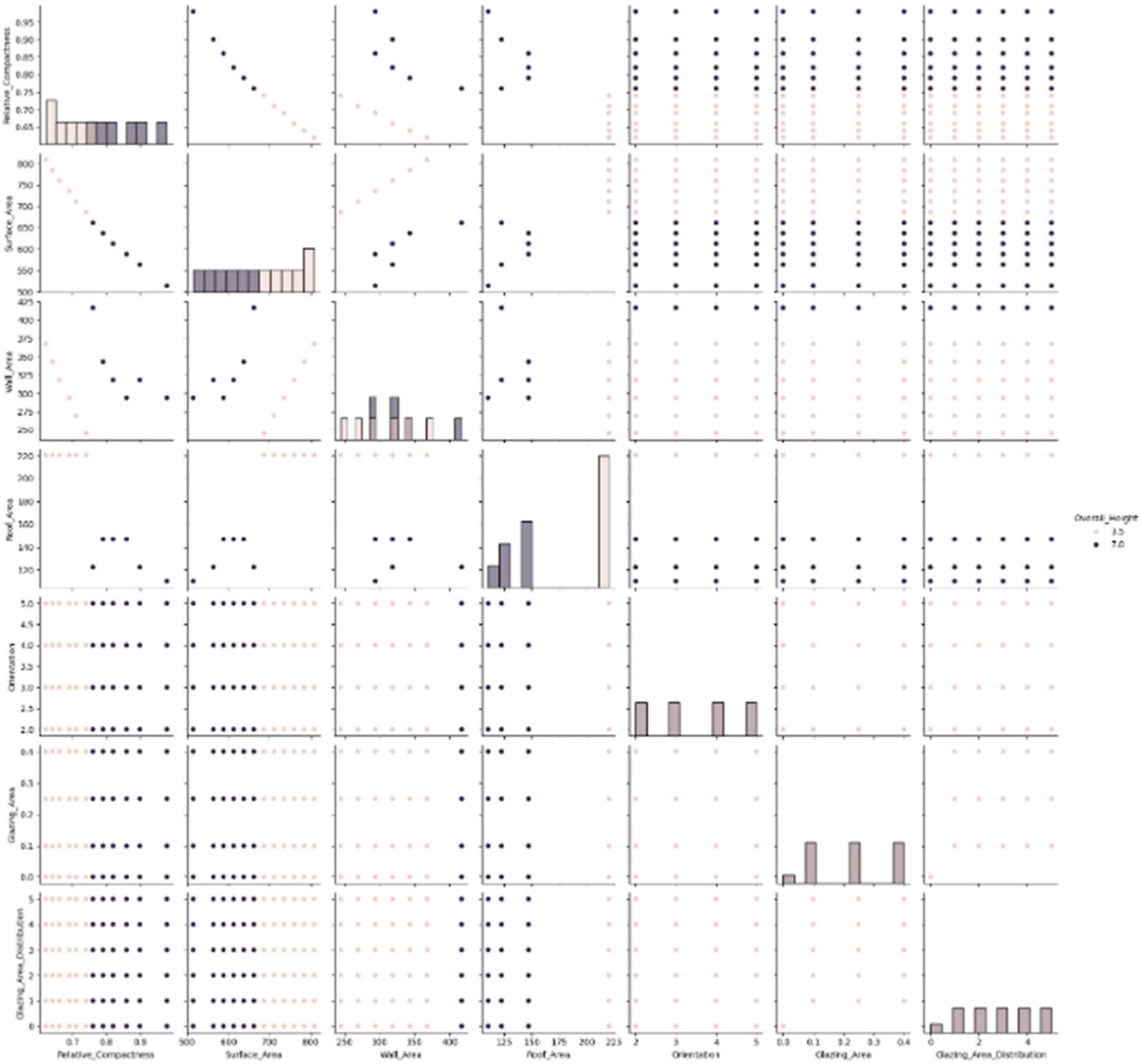

Figure 3. Variable relationships for 1 story (pink) and 2 story (purple) houses.

To confirm that heating load and cooling load differ between 1 story and 2 story houses, a hypothesis test was conducted on average heating load and average cooling load. With p values very close to zero, it was easy to reject the null hypothesis and conclude that heating load and cooling load differ for the number of stories on the home.

Several correlation matrices were created for the dataset. Collinearity was difficult to observe on the full dataset, so correlation matrices were also run on 1 story and 2 story houses. Collinearity was observed due to multiple related features (surface area, wall area, roof area) but since each feature directly corelates with energy efficiency, no features were eliminated for further analysis.

METHODS

A simple regression analysis was performed on the dataset, and heating and cooling loads were well described by the full feature dataset. OLS regression results for heating load are shown in Table 2 below:

Table 2: Heating Load OLS Regression Results

Dep. Variable:	Heating_Load	R-squared:	0.916
Model:	OLS	Adj. R-squared:	0.915
Method:	Least Squares	F-statistic:	1187.
Date:	Tue, 19 Mar 2024	Prob (F-statistic):	0.00
Time:	03:49:22	Log-Likelihood:	-1912.5
No. Observations:	768	AIC:	3841.
Df Residuals:	760	BIC:	3878.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	84.0145	19.034	4.414	0.000	46.650	121.379
Relative_Compactness	-64.7740	10.289	-6.295	0.000	-84.973	-44.575
Surface_Area	-0.0626	0.013	-4.670	0.000	-0.089	-0.036
Wall_Area	0.0361	0.004	9.386	0.000	0.029	0.044
Roof_Area	-0.0494	0.008	-6.569	0.000	-0.064	-0.035
Overall_Height	4.1699	0.338	12.337	0.000	3.506	4.833
Orientation	-0.0233	0.095	-0.246	0.805	-0.209	0.163
Glazing_Area	19.9327	0.814	24.488	0.000	18.335	21.531
Glazing_Area_Distribution	0.2038	0.070	2.914	0.004	0.067	0.341

Omnibus:	18.648	Durbin-Watson:	0.654
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37.708
Skew:	0.044	Prob(JB):	6.48e-09
Kurtosis:	4.082	Cond. No.	1.46e+16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.14e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

The cooling load OLS regression results were similar, but with R2 values around 0.88. Based on this analysis, we can see that most variables have p values below 0.05 and are statistically significant in the model. Orientation is not statistically significant for either heating or cooling load, and glazing orientation is statistically significant for heating load but not for cooling load. We are also flagged for collinearity issues in the model, but we already noted those issues in the correlation matrices.

Two OLS regressions were run on the split datasets (1 story and 2 story houses). The R2 values plunged (not a surprise), but the shift in significance of the individual variables could be assessed for each case. Some hints on relative feature importance are already evident in these simple regression models. A SHAP analysis on the OLS model shows that only roof area, surface area, and wall area are prominent features, with all other features negligible. It seems surprising that compactness and windows have essentially zero value to the model (see Figure 4):

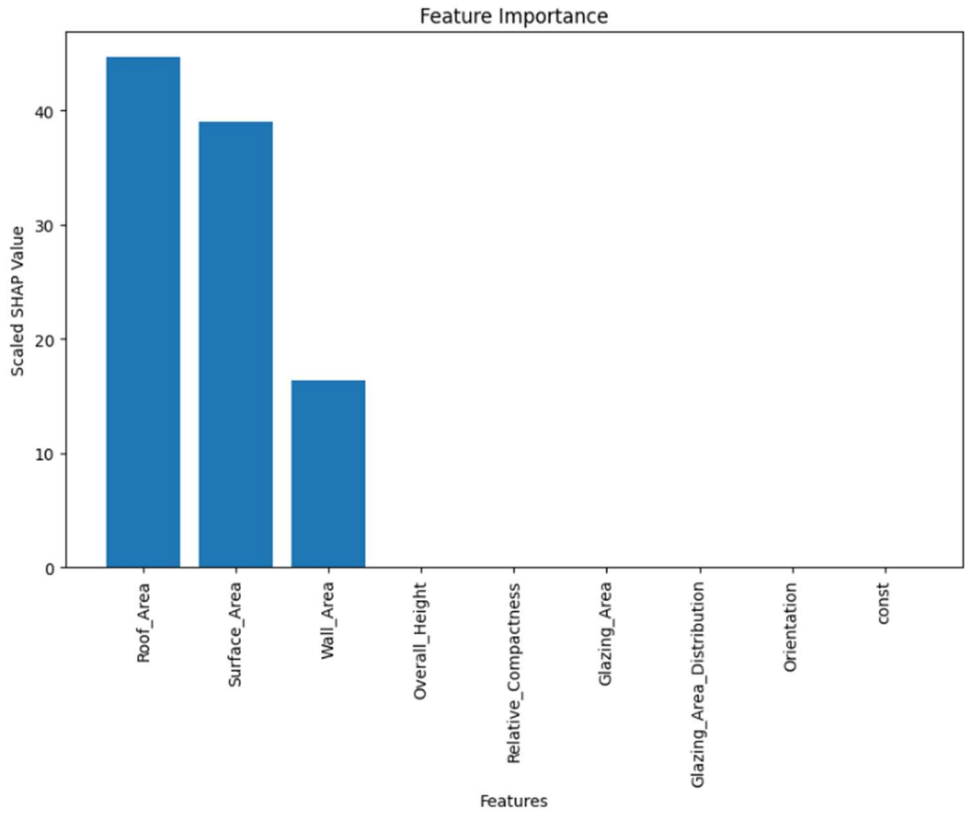

Figure 4. Heating Load Feature Important for the OLS Regression Model

The dataset was split (70/30) into training and testing sets and OLS regression was run on the training and test sets. RMSE values were generated to check for overfitting (see Table 3). The RMSE values were very low, and although overfitting is seen with heating load, it is minor. No overfitting is seen with cooling load. Regularization with a Ridge regression model was attempted to reduce overfitting, but the RMSE values didn't change. We have optimized OLS regression as much as possible given the constraints of the dataset, so other models were attempted to look for improvements in prediction power.

Table 3: RMSE Values

Model	Heating Load Train	Heating Load Test	Cooling Load Train	Cooling Load Test
OLS Regression	2.872	3.041	3.221	3.100
Ridge Regression	2.874	3.044	3.224	3.105

RESULTS

Several different models were attempted to optimize performance. The list of models and their corresponding R2 and RMSE values are shown in Tables 4 and 5. Tree models outperformed simpler linear models, with XGBoost yielding the highest R2 and the lowest RMSE values.

XGBoost was fine-tuned using grid search with cross validation (CV=5), and then trained with a learning rate of 0.3, max depths of 3 (heating) and 5 (cooling), minimum child weight of 5, number of estimators at 300, col sample by tree of 1.0, and subsample of 1.0. After making predictions on the test set, the R2 values and MAE and RMSE values are shown in Table 6.

Table 4: R2 values for Heating and Cooling Loads

Model	Heat R2 Train	Heat R2 Test	Cool R2 Train	Cool R2 Test
Linear Regression	0.917531	0.911706	0.885002	0.893025
SVR	0.930210	0.916238	0.892377	0.889505
KNN	0.963612	0.943087	0.950049	0.919558
Random Forest	0.999603	0.997164	0.995372	0.968545
Decision Tree	1.000000	0.996666	1.000000	0.934652
Bagging	0.999603	0.997481	0.993029	0.967208
AdaBoost	0.966067	0.967459	0.948174	0.943869
Gradient Boost	0.998425	0.997536	0.980189	0.975113
XGBoost	0.999981	0.998312	0.999853	0.987202

Table 5: RMSE Values for Heating and Cooling Loads

Model	Heat RMSE Train	Heat RMSE Test	Cool RMSE Train	Cool RMSE Test
Linear Regression	2.874330	3.046655	3.224343	3.106291
SVR	2.644155	2.967430	3.119238	3.156984
KNN	1.909268	2.446028	2.125048	2.693662
Random Forest	0.194368	0.551188	0.649969	1.771828
Decision Tree	0.000000	0.589996	0.000000	2.358835
Bagging	0.214217	0.565496	0.673886	1.680420
AdaBoost	1.819467	1.848830	2.344625	2.464566
Gradient Boost	0.397195	0.506423	1.338300	1.498266
XGBoost	0.043789	0.421203	0.115116	1.074424

Table 6: Final XGBoost Prediction Scores on the Test Set

Score	Heating Load	Cooling Load
R2	0.9983	0.8951
MAE	0.2844	2.5916
RMSE	0.4284	3.0761

The features were ranked and scaled between 0 and 100. Relative compactness ranked the highest, followed distantly by glazing area, wall area, and roof area. The other variables were negligible. Bar charts for feature ranks are shown in Figure 5:

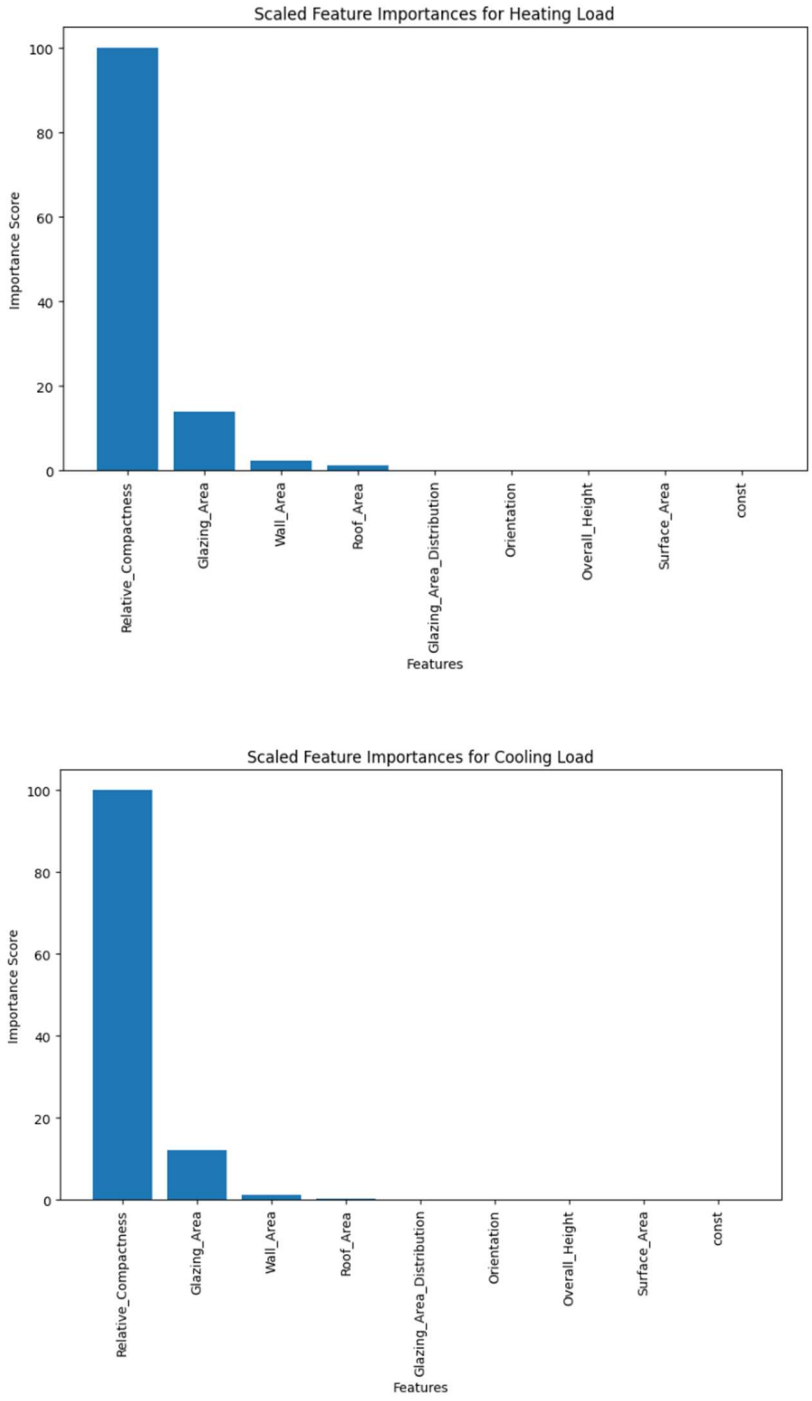

Figure 5. Scaled Feature Importances for the XGBoost Model.

CONCLUSIONS

Regression analysis was performed on an energy efficiency dataset and a model was found with extremely high prediction power for heating load, and above average performance for cooling load. I could see more overfitting with the XGBoost model on cooling load, so the lower prediciton accuracy was not a surprise. I suspect the dataset is missing an important variable related to cooling load - most likely solar radiation. Building orientation and solar radiation would interact to predict cooling load, and building orientation had no statistical significance to my models, nor any obvious correlation with the other variables.

I found it interesting that OLS regression ranked feature importance so differently from the XGBoost regression model. It became obvious that the decision tree models were superior for modeling this dataset, and I believe that the high collinaerity in this data impaired the simplier linear regression models and required the strength of a tree based model for higher accuracy.

XGBoost has been described a “Kaggle Winner” and my model performance comparisons bolster that claim. XGBoost operates by training a number of different decision trees on subsets of the data, with predictions from each tree combined to form the final prediction. The algorithm also reduces overfitting in the tree models, which I observed when comparing XGBoost performance with the Decision Tree and Random Forest models.

XGBoost also ranked the model features much differently than the OLS SHAP analysis. This model was able to look past the collinear features of roof area, wall area, and surface area and begin to make predictions from the weaker variables, which boosted accuracy. I had originally speculated that relative compactness and window area would be important features in the model, and I was shocked to find these features barely registered in the OLS regression model. OLS regression was too simplistic to effectively model this dataset; I needed a powerful tree based model for the best performance.

Just for fun, a found photos of the (retired) Ecotect software online. Seeing these photos gave me better clarification on the window distribution and orientation variables. These variables are clearly more important for multistory office and apartment buildings – with solar radiation in play. For future work I could try removing these variables from the dataset, but I doubt it would significantly alter the final model performances.