Project 2: Energy-Efficiency-Statistical-Analysis

Christina Mourad, Victor Um, Joe De Leon, Martin Ha

2024-12-03

Abstract

Utilizing UC Irvine's "Energy Efficiency" dataset, we developed multiple statistical machine learning models in order to create predictive models that will depict the most efficient heating and cooling loads based on the building characteristics of residential buildings, aiming to provide insights that can inform sustainable design practices. We analyzed 8 input variables (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area) and its relationship with two output variables, heating load (HL) and cooling load (CL), to optimize model performance by experimenting with different regression techniques and feature selection methods. We initially explored the dataset discovering the data distribution, correlation strength of each feature/target variable with each other, and the feature's significance with each target variable. Afterwards, we constructed multiple linear regression models to discover any additional associations, building upon our initial discoveries and fine-tuning our model, as well as comparing it against a more robust machine learning model, in this case decision trees, when predicting HL and CL. Lastly, to provide more actionable, real-world insights, we transformed the original regression tasks into a classification problem by implementing a multi-class logistic regression model. This approach, alongside the use of multiple linear regression models and decision trees for discovering associations, allowed us to not only predict outcomes but also offer more concise, clear, practical recommendations.

Introduction

With the growing concerns over the rise of artificial intelligence comes the increased demand for energy, particularly from data centers and computational resources. Moreover, these issues are only further compounded by the already pressing issue of global warming, worsened by years of wasteful energy use, over-consumption, and a lack of energy-saving guidelines.[1] Yet very little attention is brought to individuals and families living in residential homes - an often overlooked group that bear the brunt of increasing energy costs. It is estimated that the global contribution from buildings towards energy consumption is heading towards 20% and 40% in developed countries[2]. Therefore, a primary method of combating the energy crisis is on focusing on energy consumption in buildings. For the sake of clarity, we will focus on residential buildings, as our data is limited to this sector.

With this in mind, the main objective of this study is to analyze the dataset in terms of the variables in order to develop a practical predictive model that will describe the building characteristics most aligned with efficient heating and cooling loads in order to provide insights that can inform sustainable design practices while optimizing model performance by experimenting with different regression and feature selection methods.

[1]Celina Filippin, Energy consumption profile of public housing for lower-mid income families in a fast growing city of Argentina, Habitat International, Volume 23, Issue 1, 1999, Pages 125-134, ISSN 0197-3975, https://doi.org/10.1016/S0197-3975(98)00025-3.

[2]Luis Pérez-Lombard, José Ortiz, Christine Pout, A review on buildings energy consumption information, Energy and Buildings, Volume 40, Issue 3, 2008, Pages 394-398, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2007.03.007. (https://www.sciencedirect.com/science/article/pii/S0378778807001016)

In this study, we evaluate 8 features designed to determine the target variables heating (HL, energy required for heading) and cooling load (CL, energy required for cooling) requirements of residential buildings: Relative Compactness (A measure of the building's shape efficiency), Surface Area (The total surface area of the building), Wall Area (The area of the walls, contributing to heat transfer), Roof Area (The area of the roof, affecting thermal insulation), Overall Height (Building height, impacting air flow and heat transfer), Orientation (Cardinal direction of the buildings facade), Glazing Area (Total window area, influencing natural light and insulation), and Glazing Area Distribution (Spread of window area on each facade). The dataset amounts to 768 observations with no missing data using 12 simulated different building shapes in Ecotect[3].

Methodology

1. Initial Data Exploration

Following best practices, the first step is to explore the variables in order to uncover any significant statistical properties. For instance, understanding the distribution of the dataset is key to determining which statistical tools and machine learning models we decide to utilize. One way in which we can uncover the dataset's distribution is through histograms. While seemingly basic, histograms are unbiased in that they do not assume distribution, such as linearity, and are relatively easy to create. This step is essential in that it reveals whether the dataset follows a Gaussian or normal distribution.

Upon inspection, the histograms reveal that the dataset is Non-Gaussian. As such, Spearman rank correlation coefficient was used as to take a first look at the variables association with each other, including the target variable. To elaborate, Spearman rank correlation coefficient was used because it does not assume a normal distribution and is excellent at measuring monotonic relationships. Additionally, p-values were then used to evaluate the significance of each variable with both target values as a means to understand what we can expect going forward with the analysis.

2. Multiple Linear Regression

Results from the Spearman rank correlation indicated that the dataset had a high chance of multicollinearity. As such, we opted to use Multiple Linear Regression since it can help reveal the presence of multicollinearity. Using multiple linear regression, we split the dataset into 3 instances: train, validation, and testing, where 60% of the data was used as training and the remaining 40% was split between validation and testing. Furthermore, having no prior model or processes to compare our model's performance, we opted to create a baseline model whose predictions were only the mean of the target variable as the standard deviation of the target value would equal the RMSE of the baseline model[4].

After creating the multiple linear regression model, we plotted the summary of the model, giving us 4 graphs: Residuals vs Fitted, Normal Q-Q Plot, Scale-Location, and Residuals vs Leverage. A brief look at the plot reveals possible issues with normality, homoscedasticity, and linearity. Afterwards, we utilized Variance Inflation Factor (VIF) in order to determine multicollinearity. Unexpectedly, initially VIF did not perform as intended as it failed to work with the model. However, it was revealed that since one of the features could be perfectly predicted by the other features, VIF would ultimately fail due to how its calculations were made behind-the-scenes. Using alias on the multiple linear regression model revealed the main cause for multicollinearity to be X4, or Roof Area. Eliminating X4 from the model allowed us to utilize VIF properly, revealing several predictors with extremely multicollinearity.

On the other hand, while already surpassing the baseline model, it was believed that the multiple linear regression model could still be improved in terms of its predictive capabilities. As such, we decided to

[3]Athanasios Tsanas, Angeliki Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy and Buildings, Volume 49, 2012, Pages 560-567, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2012.03.003.

[4]Vahid Asghari, Mohammad Hossein Kazemi, Mohammadsadegh Shahrokhishahraki, Pingbo Tang, Amin Alvanchi, Shu-Chien Hsu, Process-oriented guidelines for systematic improvement of supervised learning research in construction engineering, Advanced Engineering Informatics, Volume 58, 2023, 102215, ISSN 1474-0346, https://doi.org/10.1016/j.aei.2023.102215.

apply Lasso and forward, backward, and bidirectional stepwise to the model, expecting Lasso regularization technique to perform better then the stepwise techniques[5].

3. Decision Tree

Unlike the Multiple Linear Regression model, Decision trees are much more robust in that they do not have strong assumptions of the dataset such as linearity, independence of errors, homoscedasticity, etc, making them another good choice as a predictive model. While a simple notion, decision trees works by successively splitting the input feature space into smaller and smaller sub-regions, growing until it isn't possible or until it meets a certain condition[6]. To add to this, we applied the Grid Search Cross-Validation (GridSearchCV) technique to fine-tune the model, choosing the complexity parameter (cp or cost-complexity pruning parameter) as the hyperparameter to fine-tune. Essentially, GridSearchCV allows us to pass in a range of values for the hyperparameter of our choice and tests all combinations on the model, selecting the best choice to keep. Also, cp determines the minimum reduction in cost function required to make a further split at the node. To put it differently, if the improvement given a split is less than the cp, tree will stop splitting and prune the 'branch'. As a result, not only the the Decision Tree perform better, but its interpretability may also improve.

4. MultiNomial Logistic Regression

While previous models were used for understanding the data and used within the context of regressions tasks, the previous models did not meet our goals for a practical predictive model. To elaborate, while both multiple linear regression and decision tree demonstrated good predictive capabilities and data insight, the fact that they can only predict the heating or cooling load as a continuous value limits the ability of the average non-statistical user to make decisions without additional thresholds or context. As such, we decided to turn the regression task into a classification by using a MultiNomial Logistic Regression model in addition to the data and feature insights gained from the multiple linear regression model and decision tree.

To put it differently, we separated the target values into 3 groups based on quantile range: low, medium, and high. Furthermore, using the insight gained from the previous models, we removed already-identified poor features, increasing the accuracy of the multinomial logistic regression model. Additionally, we opted to use accuracy to evaluate the model due to little to no class imbalance when splitting the data into 3 groups. However, we still created a confusion matrix in order to understand the finder details of the model.

[5]SUDHEER KUMAR, S.D. ATTRI, & K.K. SINGH. (2019). Comparison of Lasso and stepwise regression technique for wheat yield prediction. Journal of Agrometeorology, 21(2), 188–192. https://doi.org/10.54386/jam.v21i2.231

[6]Athanasios Tsanas, Angeliki Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy and Buildings, Volume 49, 2012, Pages 560-567, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2012.03.003.

Data Analysis

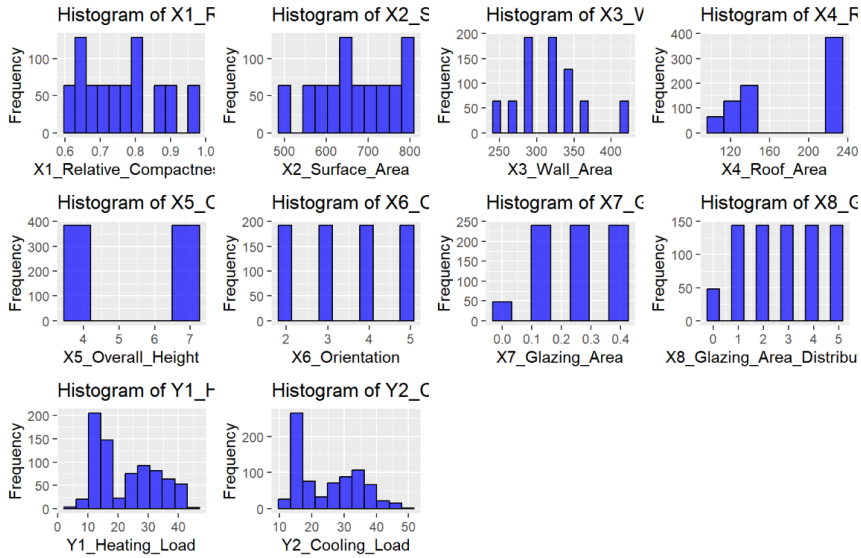

Fig. 1 showcases the distribution of the data points for each variable, showcasing that all of the variables are Non-Gaussian.

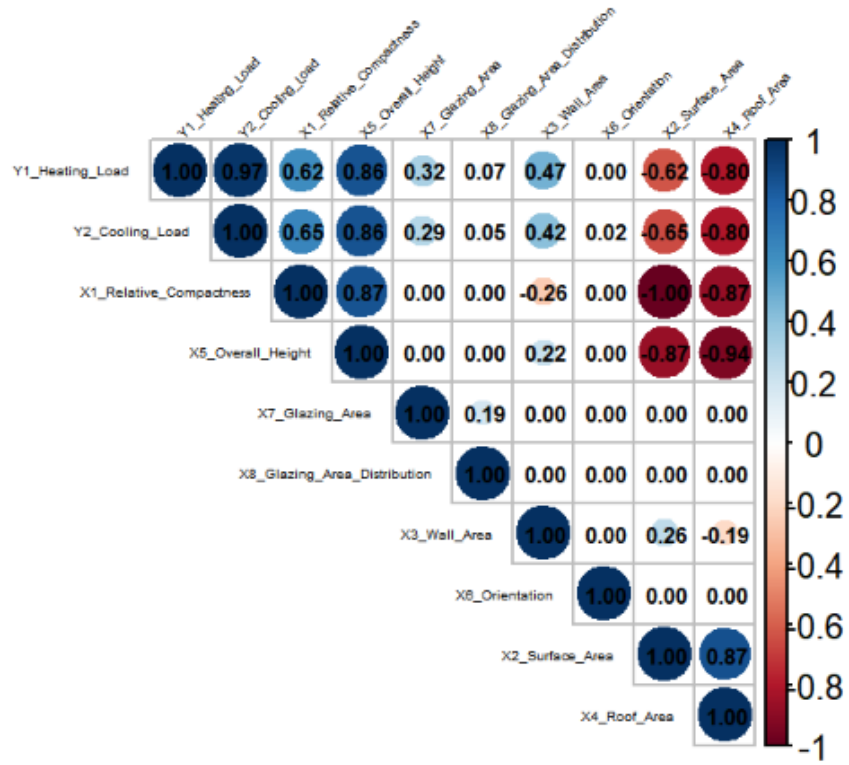

Fig. 2 showcases the Spearman rank correlation coefficient, showcasing signs of multicollinearity.

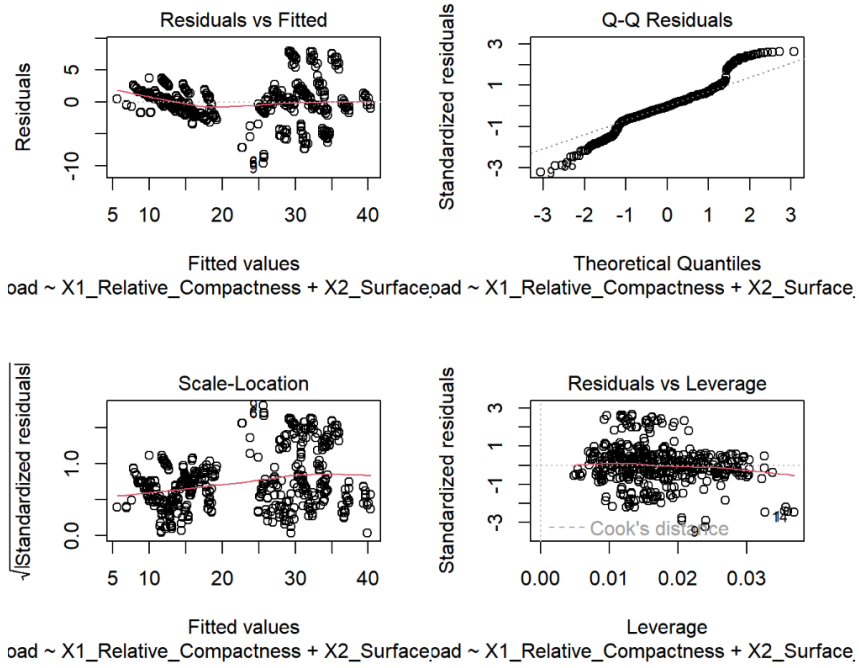

Fig. 3: Residuals vs. Fitted (Top Left): The residuals do not appear to be evenly scattered around the horizontal line (y=0), which suggests potential non-linearity or model misspecification. Normal Q-Q Plot (Top Right): The points deviate from the diagonal line at the tails, indicating the residuals may not follow a normal distribution. Scale-Location (Bottom Left): There appears to be a pattern in the spread of residuals, suggesting heteroscedasticity. Residuals vs. Leverage (Bottom Right): Most points lie within the Cook's distance boundaries, but there may be some moderately influential points that warrant further investigation.

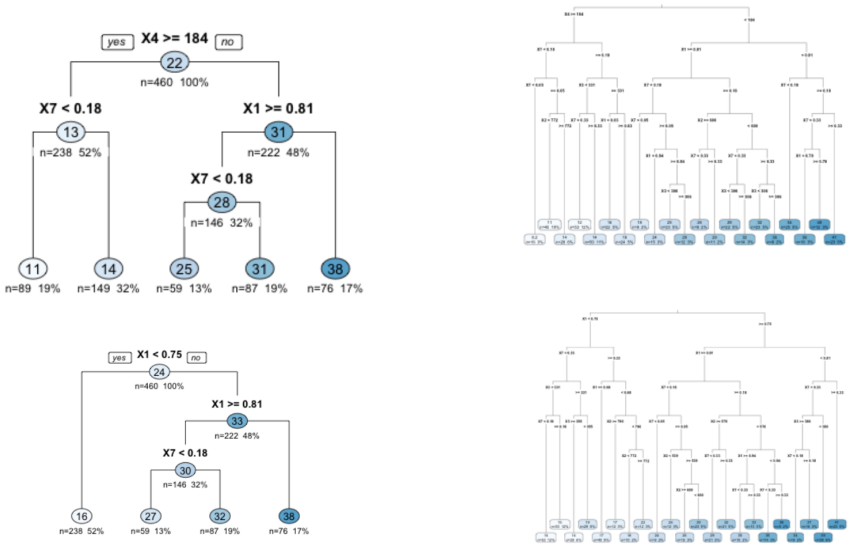

Fig. 4: Top Left: Decision Tree for Heating Load Top Right: Decision Tree For Heating Load after GridSearchCV Bottom Left: Decision Tree for Cooling Load Bottom Right: Decision Tree For Cooling Load after GridSearchCV

Furthermore, when comparing Lasso multiple linear regression models' performance against GridSearchCV Decision Tree for Heating Load and Cooling Load, GridSearchCV Decisoin Tree outperforms Lasso multiple linear regression on both fronts:

Lasso heating: “RMSE on Validation Data: 2.91816828116011” Lasso cooling: “RMSE on Validation Data: 3.17224400910463”

gridsearch heating: “Validation Root Mean Squared Error: 1.52752918017046” grisearch cooling: “Validation Root Mean Squared Error: 2.13034746196226”

As such, in terms of regression tasks, Decision Tree applied with GridSearchCV is the better model predictive model. Given that decision tree with GridSearchCV is better, it demonstrates that the features most important to both Y1 and Y2 will be the features utilized by the GridSearchCV tree. Those being X7, X3, and X1.

```
predicted_classes2 <- predict(model2, newdata = df, type = "class")
conf_matrix2 <- confusionMatrix(predicted_classes2, df$Y1_class)
print(conf_matrix2)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Low Medium High
##     Low    211     33    0
##     Medium  43    188   20
##     High     0     32  241
##
## Overall Statistics
##
##                Accuracy : 0.8333
##                  95% CI : (0.8051, 0.859)
##     No Information Rate : 0.3398
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7499
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Low Class: Medium Class: High
## Sensitivity              0.8307        0.7431      0.9234
## Specificity              0.9358        0.8777      0.9369
## Pos Pred Value           0.8648        0.7490      0.8828
## Neg Pred Value           0.9179        0.8743      0.9596
## Prevalence               0.3307        0.3294      0.3398
## Detection Rate           0.2747        0.2448      0.3138
## Detection Prevalence     0.3177        0.3268      0.3555
## Balanced Accuracy        0.8833        0.8104      0.9301
```

Fig. 5 Multinomial Logistic Regression Results

Conclusion

Overall, the combination of regression and classification models allowed us to develop a comprehensive approach to predicting energy needs in residential buildings. These results can inform future sustainable building designs by highlighting the most important factors influencing energy consumption, such as glazing area, roof area, and orientation. Future work could explore further feature engineering or more advanced machine learning techniques to refine the predictions and make them even more actionable.

Appendix