Regression-Based Prediction of Heating and Cooling Loads in Residential Buildings

Motivation

Climate change has become a central topic of global discourse, recognized for its profound impacts on the environment, economies, and societies and the widespread discussion on climate change reflects both the growing scientific consensus about its consequences. World Economic Forum mentioned that the building value chain is responsible for 37% of total carbon emissions globally, involving multiple harder-to-abate sectors and needs to accelerate green transition from now on.

In this section, we aim to analyse the relationship through various regressions techniques between various building design parameters and subsequently help us predict and determine the energy efficiency of a building. This will be extremely useful for during the planning phase of a building construction.

Dataset

The dataset comprises 768 samples and 8 features and 2 response variables (heating and cooling load). The heating load is the amount of energy needed to heat a building, while cooling load indicates the amount of energy required to cool it. The analysis will be conducted using Residential Building Energy Efficiency dataset, available on UCI.

Variable	**Description**	**Unit**
Relative Compactness	Measure of the building's compactness relative to its volume	Dimensionless
Surface Area	Total exterior surface area of the building	m^2
Wall Area	Surface area of the building walls	m^2
Roof Area	Area of the roof	m^2
Overall Height	Overall height of the building	m
Orientation	Directional orientation of the building (encoded categorical value)	Categorical (e.g., 2-5)
Glazing Area	Total area of glazing (windows)	m^2 (or fraction)
Glazing Area Distribution	Distribution pattern of the glazing area	Categorical (e.g., 0-5)
Heating Load	Energy required to heat the building	kWh/m^2

Variable	**Description**	**Unit**
Cooling Load	Energy required to cool the building	kWh/m^2

Methodology

- Linear regression and Bayesian regression
- LASSO regression and Bayesian LASSO regression
- Ridge regression and Bayesian ridge regression

Our aim is to understand the relationship of each variable to the heating and cooling load and predict the energy efficiency in terms of heating and cooling load. The evaluation of each model will be based on the MSE of its prediction. The analysis is conducted through scikit-learn and PyMC libraries on Python. We will also split the data into 80% training set and 20% test set.

Linear regression

Linear regression is the most common statistical learning technique. The coefficients can be computed by $\beta\text{^}=(X\top X)-1X\top y$. It achieved a test MSE of 9.15 for heating load and 9.89 for cooling load. This MSE shall serve as a benchmark comparison for our models below.

Polynomial linear regression

Polynomial linear regression an extension of standard linear regression that allows us to model nonlinear relationships between the independent variables and the target variable. In this approach, we generate additional features by raising the original predictors to various powers (e.g., square, cube). We use the function scikit-learn's GridSearchCV to systematically test different polynomial degrees 1-4 to determine which transformation of the features produces the best model performance, using 5-fold cross-validation (CV) and optimizing for negative mean squared error.

The best polynomial degree for heating load is 3, which produced a MSE of 0.29. The best polynomial degree for cooling load is 2, which produced a MSE of 2.97. The MSEs for polynomial linear regression are significantly lower than the linear regression. This indicates that the variables might share a nonlinear relationship.

Bayesian Linear Regression

In this model, we assigned a normal prior to the intercept and each regression coefficient, centered at 0 with a standard deviation of 10. The noise (standard deviation of the errors) is modeled with a half-normal prior to ensure it remains positive.

This model uses MCMC to draw samples from the posterior distribution of the parameters (beta, intercept, and sigma). Predictions on the test set are computed by taking the dot product of the beta samples with the test data and adding the intercept samples and the final point prediction for each test instance is obtained by averaging over all posterior samples. We observed a Bayesian regression MSE of 9.16 and 9.89 for heating and cooling load respectively.

Next, we experimented with Bayesian polynomial regression to capture non-linear relationships between the features and heating/ cooling load. Due to computation constraint, we experimented with a degree of polynomial 3. The Bayesian prior setup is similar to above. In this case, our MSE is 1.32 for heating load and 4.96 for cooling load.

Lasso Regression

Lasso regression is a linear regression technique that adds a penalty based on the absolute values of the coefficients. This penalty helps prevent overly complex models by shrinking coefficients toward zero, effectively selecting only the most important predictors. Traditionally, Lasso minimizes an objective that combines the residual sum of squares (which measures fit) with the L1-norm of the coefficients (which measures complexity). In a Bayesian framework, Lasso is implemented by assigning Laplace (double exponential) priors to the regression coefficients, creating a similar shrinkage effect.

In our Lasso implementation, we used an alpha of 0.1. The MSEs are 9.94 for heating load and 10.76 for cooling load, indicating that Lasso regression does not outperform linear regression in this instance.

However, when polynomial features are included, the best model achieves a test MSE of 39.28. With degree-4 polynomial features, the MSE further decreases to 0.23 for heating load and 1.81 for cooling load, using a regularization parameter of 0.001 for heating load and 0.1 for cooling load.

Ridge Regression

Ridge regression is a linear regression technique that adds a penalty term to the ordinary least squares cost function. This penalty term, known as the L2 penalty, is the sum of the squares of the coefficients multiplied by a regularization parameter and it discourages large coefficient values, which helps reduce overfitting and multicollinearity. The key difference between ridge and LASSO is that ridge penalty term typically doesn’t shrink coefficients exactly to zero.

Our ridge regression produced MSEs of 9.16 and 9.90 for heating and cooling respectively. Similarly, we perform a grid search to identify the combination that minimizes the cross-validated error, and outputs the best polynomial degree and alpha along with the corresponding CV score.

When we included polynomial features of degree 3 in the ridge regression, the MSE greatly reduced to 0.39 for heating load and 2.66 for cooling load.

Summary of results

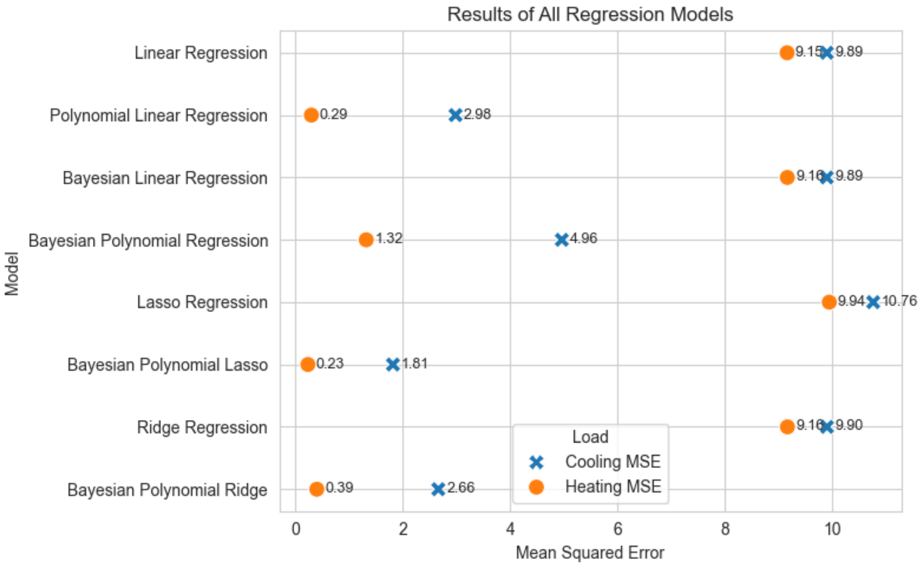

Figure 6: Summary of results

Discussion

Our analysis demonstrates that while standard linear regression provides a solid baseline for predicting building energy efficiency, incorporating polynomial features markedly improves model performance, indicating the presence of nonlinear relationships among the design parameters. The vanilla linear models yielded MSEs around 9–10 for both heating and cooling loads, polynomial expansions slashed errors to well below 1 for heating and under 3 for cooling, and the Bayesian LASSO model ultimately delivered the best performance (0.23 heating, 1.81 cooling).

The polynomial linear regression models, particularly at higher degrees, substantially reduce the mean squared error (MSE) compared to the linear approach. One possible explanation is that heating and cooling loads depend on heat transfer through walls, roofs and glazing, which follow physical laws involving areas, volumes and temperature differences. These processes aren't strictly linear in the raw features (e.g. doubling surface area doesn't simply double the load when other factors interact).

Additionally, the Bayesian regression frameworks and their regularized counterparts, such as Bayesian LASSO and Bayesian ridge regression, offer robust alternatives that balance model complexity with predictive accuracy. The results suggest that with careful feature transformation and regularization, it is possible to significantly enhance prediction accuracy for heating and cooling loads, thereby offering valuable insights for the planning phase of energy-efficient building construction.

Conclusion and Potential Improvements

In conclusion, we tested several regression models on UCI Energy dataset, with the goal of predicting the heating and cooling load of buildings. Bayesian Lasso Polynomial produced the best results. By leveraging these techniques within scikit-learn and PyMC, architects and engineers can obtain more reliable, uncertainty-aware forecasts during the design phase.

Future work could extend this framework beyond raw polynomial terms, targeted interactions guided by heat-transfer physics (e.g. wall × orientation, glazing × height), alternative prior structures, or even hierarchical and machine-learning models to push predictive accuracy and interpretability even further.

References:

Liu, M., Huo, J., Wu, Y., & Wu, J. (2021). *Stock Market Trend Analysis Using Hidden Markov Model and Long Short Term Memory*. arXiv preprint arXiv:2104.09700. https://doi.org/10.48550/arXiv.2104.09700

Aramyan, H., Ramchandani, J., & Skevofylakas, M. (2023, February 13). *Market regime detection using Statistical and ML based approaches*. Developer Portal. Retrieved from https://developers.lseg.com/en/article-catalog/article/market-regime-detection