Final Project
CM50268 Bayesian Machine Learning 2023-2024

Task 1: Exploratory analysis

The dataset consists of 10 variables, including an intercept (bias) and various features related to building characteristics. Both the training and test datasets consist of 384 samples each.

1. **Training Dataset:** Summary statistics for each variable show that values vary across variables, with heating load values ranging from 6.4 to 43.1 in the training set.
2. **Test Dataset:** Similar ranges and distributions are observed in the test set.
3. **Missing Values:** Both datasets are free of missing values across all variables.

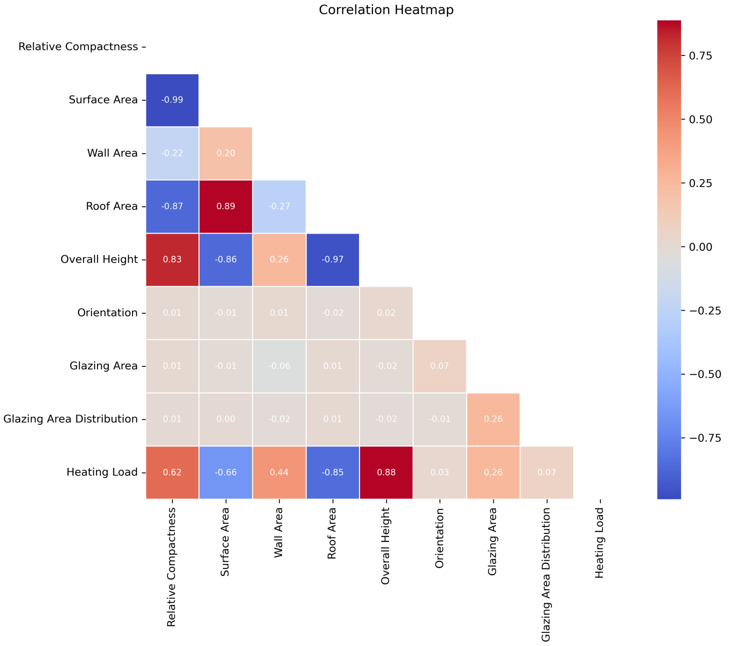

Figure 1: Correlation Heatmap

2. Correlation Analysis:

- **Relevant Variables:**
  - **Roof Area**: Strong negative correlation (around 0.85).
  - **Overall Height**: Strong positive correlation (around 0.66).
  - **Surface Area**: Negative correlation (around 0.66).
  - **Relative Compactness**: Positive correlation (around 0.62).
  - **Wall Area**: Positive correlation (around 0.44).
- **Irrelevant Variables:**
  - **Orientation**: Negligible correlation (around 0.03).
  - **Glazing Area**: Weak correlation (around 0.26).
  - **Glazing Area Distribution**: Weak correlation (around 0.07).

3. Training and Evaluation:

The model was evaluated using Mean Absolute Error (MAE) on both training and test datasets.

- **Training MAE:** 2.1609
- **Test MAE:** 2.0305

4. Initial Observations on Task Difficulty and Linearity:

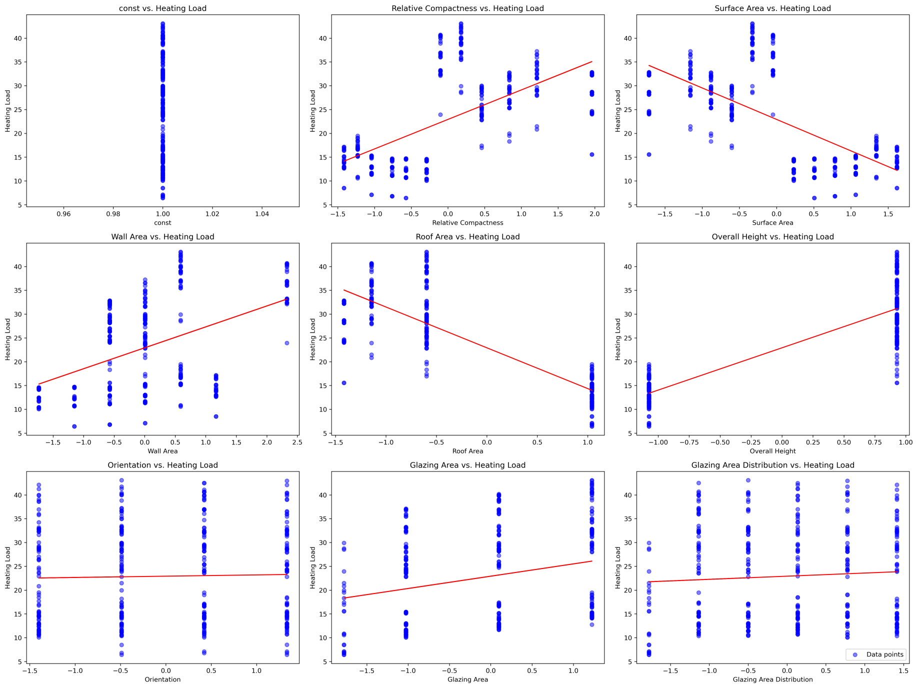

Figure 2: Correlation line plot

Preliminary analysis of the dataset suggests that at least some features in this dataset might linearly correlate with heating loads. With the case of Roof Area, Surface Area, Relative Compactness, Wall Area and Overall Height, the relation is strong. Thus, the linearity assumption is rather strong. However, correlations with Orientation and Glazing Area Distribution are slight and close to zero. That points to having little influence under linear modeling assumptions. Instead, moderate task difficulty seems to be the case and the need for a careful feature selection procedure and possibly non-linear models to achieve better predictive accuracy.

Task 2: Bayesian Linear Regression

In this task we are going to work on Bayesian Linear Regression with Gaussian priors on weights and noise. Here, we're concerned with finding the most likely values of hyperparameters governing priors by computing the log-marginal likelihood, considering all weight vectors being in its space.

1. Assumptions:

- w has a Gaussian prior $N(0, \sigma_w^2)$ where $\alpha = 1/\sigma_w^2$ is the prior precision.
- The model incorporates Gaussian noise $N(0, \sigma_\epsilon^2)$ where $\beta = 1/\sigma_\epsilon^2$ is the noise precision.

2. Covariance Matrix and Log Marginal

are defined as:

$$\text{covar} = \frac{1}{\beta} I + \frac{1}{\alpha} X X^T$$

$$\log p(y|\alpha, \beta) = -0.5\, y^T K^{-1} y - 0.5 \log |K| - 0.5 N \log 2\pi$$

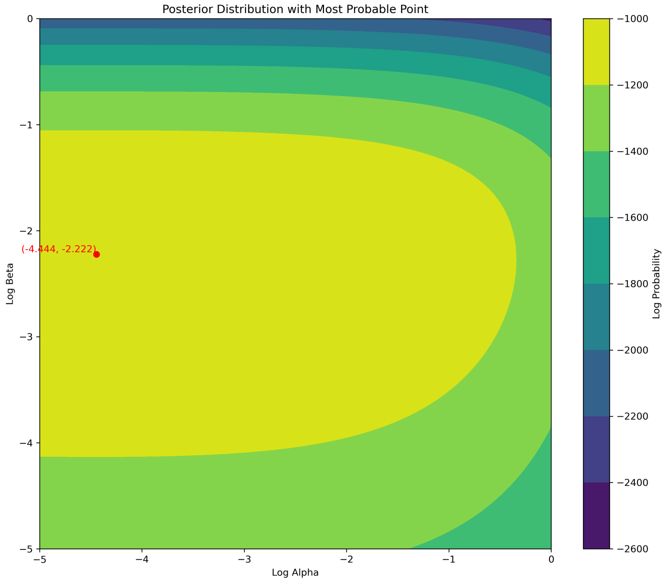

Figure 3: Posterior Distribution with Most Probable Point

3. Most Probable Values:

The posterior distribution's contour plot identifies the following optimal values:

- Alpha: $\alpha = 0.0117$ (log-alpha: $\log \alpha = -4.4444$)
- Beta: $\beta = 0.1083$ (log-beta: $\log \beta = -2.2222$)

These values correspond to the model's precision parameters, reflecting the level of uncertainty in the prior distribution over the regression weights and the noise in the data. The log-marginal likelihood at these values is -1001.4508, indicating the model's likelihood for given data and hyperparameters.

4. Posterior Calculation and Model Performance:

The model's posterior distribution is calculated based on the given dataset, finding a posterior mean (μ) and covariance (Σ) for the regression weights:

Posterior Mean:

$$\mu = \left(X^T X + \frac{\alpha}{\beta} I \right)^{-1} \left(X^T y \right)$$

Posterior Covariance:

$$\Sigma = \frac{1}{\beta} \left(X^T X + \frac{\alpha}{\beta} I \right)^{-1}$$

5. Model Predictions:

The posterior mean (μ) is used to make predictions:

- Training Set: The predictions for the training set are computed by multiplying X_{train} by μ_n.
- Test Set: Similarly, test set predictions are derived using X_{test} and μ_n.

6. Evaluation of the Model:

The model's performance is evaluated by calculating the Mean Absolute Error (MAE) on both sets:

- Training Set MAE: 2.1302
- Test Set MAE: 2.0668

Task 3: Verifying HMC on a Standard 2D Gaussian Example

1. HMC Sampling:

Parameters Used:

- R: 10,000 samples
- L: 25 leapfrog steps per sample
- **eps:** 0.36.
- **Burn-in:** 1,000 burn-in samples.

2. Mathematical Formula of the 2D Gaussian:

The standard 2D Gaussian distribution is described by the following formula:

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

3. Potential Energy Function :

The energy function calculates the negative log probability of x under the Gaussian distribution:

$$\text{Energy}(x) = -\ln p(x)$$

4. Energy Gradient Function:

The gradient of the negative log probability is given by:

$$\nabla\text{Energy}(x) = \Sigma^{-1} \cdot x$$

5. Verification of HMC's Functionality:

- **Gradient Check:** Gradient Check Performed a gradient check to ensure all was consistent The difference between analytical and numerical gradients varies between each refinement.

Parameter	Calc.	Numeric	Delta
S1	-3.01349	-3.01349	-1.06e-10
S2	2.9126	2.9126	-1.37e-10

Table 1: Gradient Check Values of HMC on a Standard 2D Gaussian

- **Acceptance Rate:** The acceptance rate for the HMC sampling was 90.1%, indicating effective coverage and stability.
- **Contour Plot:** This shows the distribution's density with sampled points overlayed:

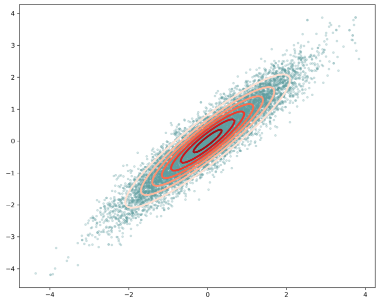

Figure 4: Contour plot and scatter points for HMC 2D Gaussian distribution

- **Weights Distribution:** This plot shows the distribution of weights after HMC sampling:

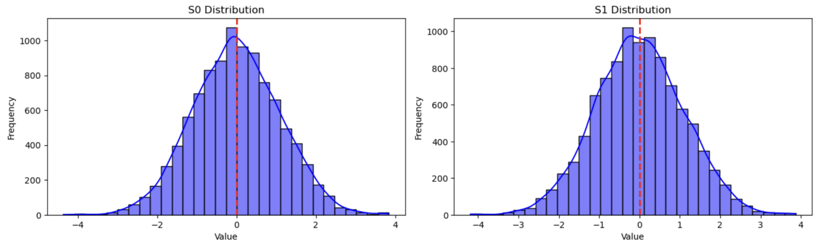

Figure 5: Weights distribution for HMC sampling

6. Optimal Values of the Parameters:

After sampling, the mean values of the model's parameters were computed as follows:

Parameter	Mean Value
S1	0.0011
S2	-0.0015

Table 2: Optimal Parameters on a Standard 2D Gaussian

Task 4: Bayesian Linear Regression with HMC

In this task, we apply Hamiltonian Monte Carlo (HMC) to perform Bayesian linear regression on a dataset. The goal is to sample from the posterior distribution of the model's parameters, evaluate the model's performance, and compare the results to other methods.

1. HMC Sampling Parameters:

- **R:** 10,000 samples.
- **L:** 20 leapfrog steps.
- **eps:** 0.08.
- **Burn-in:** 1,000 burn-in samples.

2. Distribution of Prior and Likelihood:

- **Prior:** The weights w are assumed to have a Gaussian prior $w \sim N(0, \sigma_w^2)$ where $\sigma_w^2 = 1/\alpha$. The negative log prior is:

$$\text{neg\_log\_prior} = 0.5 \times \alpha \times \sum w^2 + 0.5 \times \text{len}(w) \times \ln(2\pi/\alpha)$$

- **Likelihood:** The model assumes Gaussian noise with variance $\sigma_\epsilon^2 = 1/\beta$. The negative log likelihood for the data y given the weights w is:

$$\text{neg\_log\_likelihood} = 0.5 \times \beta \times \sum((y - y_{\text{pred}})^2) + \text{len}(y) \times 0.5 \times \ln(2\pi/\beta)$$

  Where $y_{\text{pred}} = Xw$ are the model predictions.

3. Posterior Distribution (Potential Energy Function):

The posterior distribution combines the prior and likelihood, resulting in the negative log posterior:

$$\text{neg\_log\_posterior} = \text{neg\_log\_prior} + \text{neg\_log\_likelihood}$$

This posterior distribution represents the probability distribution of the model's parameters $\theta = (\alpha, \beta, w)$ given the observation data D.

4. Gradient Function for Bayesian Linear Regression:

The gradient function calculates the partial derivatives of the energy function (negative log posterior) with respect to each parameter θ:

- **Gradient of α:**

$$\text{grad\_alpha} = (0.5 \times (\sum w^2 - \frac{\text{len}(w)}{\alpha})) \times \alpha$$

- **Gradient of β:**

$$\text{grad\_beta} = (0.5 \times (\sum((y - y_{\text{pred}})^2) - \frac{\text{len}(y)}{\beta})) \times \beta$$

- **Gradient of w:**

$$\text{grad\_w} = \alpha \times w - \beta \times X^T \times (y - y_{\text{pred}})$$

5. Verification of HMC's Functionality:

- **Gradient Check:** Gradient Check Performed a gradient check to ensure all was consistent The difference between analytical and numerical gradients varies between each refinement.

Parameter	Calc.	Numeric	Delta
Alpha	1.57649	1.5765	-3.72e-09
Beta	43.8348	43.835	1.11e08
W1	-0.6565	-0.6565	1.12e07
W2	-2.4974	-2.4974	-5.92e-09
W3	2.5414	2.5414	6.62e-08
W4	-1.1908	-1.1908	-4.45e-08
W5	3.0598	3.0598	-1.01e-08
W6	-3.0084	-3.0084	-1.31e-07
W7	1.3979	1.3979	-5.20e-08
W8	1.4223	1.4223	-1.18e-08
W9	0.2096	0.2096	-7.59e-08

Table 3: Gradient Check Values of HMC

- **Acceptance Rate:** The acceptance rate for the HMC sampling was 87%, indicating effective coverage and stability.

- **Weights Distribution:** This plot shows the distribution of weights after HMC sampling:

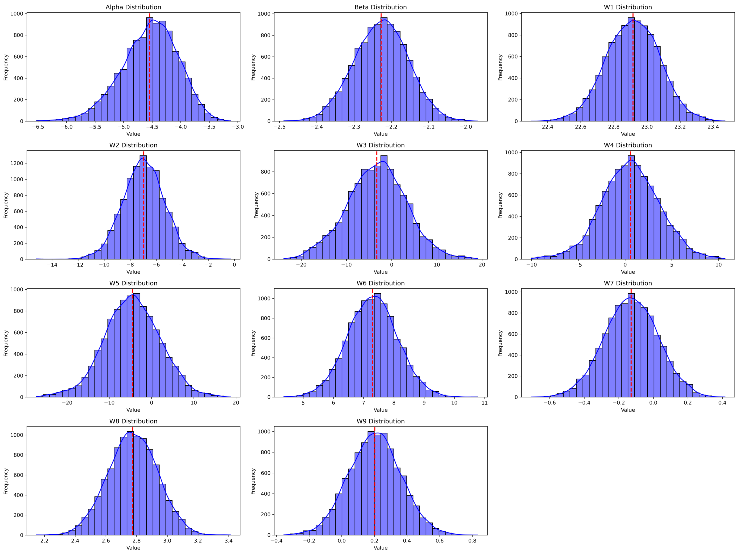

Figure 6: Posterior distribution of variables

6. Optimal Values of the Parameters:

After sampling, the mean values of the model's parameters were computed as follows:

Parameter	**Mean Value**
log(Alpha)	-4.5430
log(Beta)	-2.2373
W1	22.9154
W2	-6.9623
W3	-3.2892
W4	0.5743
W5	-4.5415
W6	7.376
W7	-0.1300
W8	2.7753
W9	0.2032

Table 4: Optimal Values of Parameters

7. Evaluation of the Model:

The model's performance was evaluated using the Mean Absolute Error (MAE) metric:

- **Training Set MAE:** 2.1292
- **Test Set MAE:** 2.0661

8. Comparative Insight:

Both the Bayesian linear regression models in both tasks turn out to have the same MAE values indicating that they are stable. Task 2 uses Type-II maximum likelihood in finding the best hyperparameters in contrast to Task 4, that employs HMC in sampling the posterior distribution. Both methods are capable of capturing the input features relationship with the target variable, which is depicted to give valid and reliable predictions.

Task 5: Apply HMC as a Classifier

In this task, we formulate the regression problem as a binary classification problem. problem using Hamiltonian Monte Carlo (HMC). Given a path needed to forecast a: "high" heating load, labeling cases with a heating load greater than 23.0 one-sided positive model.

1. HMC Sampling Parameters:

- **R:** 1,000 samples.
- **L:** 20 leapfrog steps.
- **eps:** 0.17.
- **Burn-in:** 100 burn-in samples.

2. Distribution of Prior and Likelihood:

- **Prior:** The weights w are assumed to have a Gaussian prior $w \sim N(0, \sigma_w^2)$ where $\sigma_w^2 = 1/\alpha$. The negative log prior is:

$$\text{neg\_log\_prior} = 0.5 \times \alpha \times \sum w^2 + 0.5 \times \text{len}(w) \times \ln(2\pi/\alpha)$$

- **Likelihood:** The model assumes a Bernoulli likelihood with a sigmoid link function. The negative log likelihood for the data y given the weights w is:

$$\text{neg\_log\_likelihood} = -\sum (y \cdot \ln(y_{\text{pred}}) + (1-y) \cdot \ln(1-y_{\text{pred}}))$$

3. Posterior Distribution (Potential Energy Function):

The posterior distribution combines the prior and likelihood, resulting in the negative log posterior:

$$\text{neg\_log\_posterior} = \text{neg\_log\_prior} + \text{neg\_log\_likelihood}$$

This posterior distribution represents the probability distribution of the model's parameters $\theta = (\alpha, w)$ given the observation data D.

4. Gradient Function :

The gradient function calculates the partial derivatives of the energy function (negative log posterior) with respect to each parameter θ:

- **Gradient of α:**

$$\text{grad\_alpha} = (0.5 \times (\sum w^2 - \frac{\text{len}(w)}{\alpha})) \times \alpha$$

- **Gradient of w:**

$$\text{grad\_w} = \alpha \times w + X^T \times (y_{\text{pred}} - y)$$

- **Combined Gradient:**

$$\text{grad\_posterior} = \text{np.concatenate}(([grad\_alpha], grad_w))$$

5. Verification of HMC's Functionality:

- **Gradient Check:** A gradient check was performed, ensuring consistency between the analytical and numerical gradients.

Parameter	**Calc.**	**Numeric**	**Delta**
Alpha	25.0326	25.0326	5.24e-09
W1	-0.1560	-0.1560	4.44e-09
W2	0.6321	0.6321	1.68e-08
W3	-8.2316	-8.2316	1.95e-08
W4	3.0584	3.0584	-1.47e-08
W5	-13.7917	-13.7917	5.08e-08
W6	9.6650	9.6650	-8.72e-09
W7	-14.1682	-14.1682	-2.17e-08
W8	-15.3266	-15.3266	3.24e-09
W9	-18.0012	-18.0012	-1.38e-08

Table 5: Gradient Check Value of HMC

- **Acceptance Rate:** The acceptance rate for the HMC sampling was 90.4%, indicating effective coverage and stability.

- **Weights Distribution:** This plot shows the distribution of weights after HMC sampling:

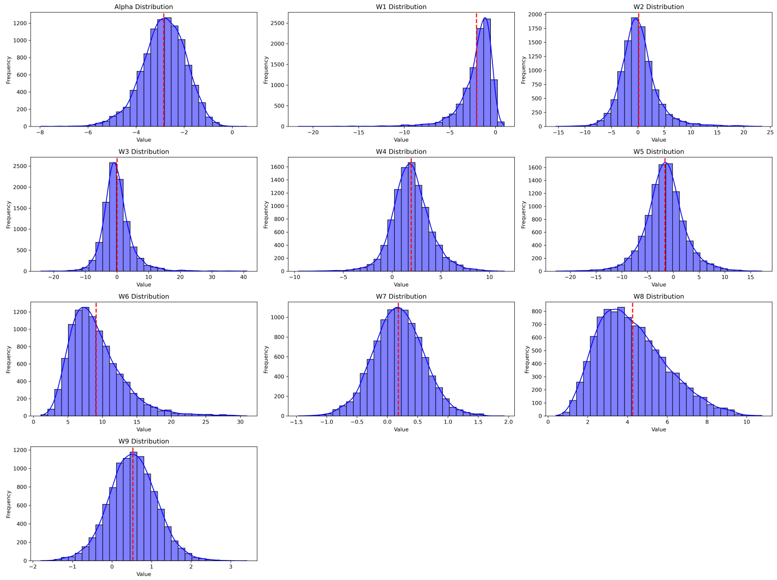

Figure 7: Posterior distribution of variables

6. Optimal Values of the Parameters:

After sampling, the mean values of the model's parameters were computed as follows:

Parameter	**Mean Value**
log(Alpha)	-2.8448
W1	-2.0561
W2	0.1128
W3	-0.0166
W4	1.9631
W5	-1.6404
W6	9.0641
W7	0.1876
W8	4.2644
W9	0.5290

Table 6: Optimal Values of Parameters

7. Evaluation of the Model:

The model's performance was evaluated using accuracy.

- **Training Accuracy:** 98.70
- **Test Accuracy:** 99.22

Task 6: Variational Inference

This task implements Variational Inference with a mean-field factorization to estimate the "most probable" values of the hyperparameters θ within a Bayesian linear regression model.

1. Assumption:

- **Precision Parameters:** α and β are modeled as Gamma distributions:

$$\alpha \propto \Gamma(\alpha|a, b)$$

$$\beta \propto \Gamma(\beta|c, d)$$

- **Weights Posterior:** The posterior distribution for the weights w is modeled as a multivariate Gaussian distribution:

$$w \propto N(w|\mu_n, \Sigma_n)$$

2. Mathematical Derivations:

- **Posterior Covariance:** The posterior covariance matrix Σ_n of the weights is derived as follows:

$$\Sigma_n = \left(X_{\text{train}}^T X_{\text{train}} \cdot \beta + \alpha \cdot I\right)^{-1}$$

  where:

  - X_{train} is the design matrix for the training set.
  - β is the precision of the noise modeled by a Gamma distribution.
  - α is the precision of the weights modeled by a Gamma distribution.
  - I is the identity matrix.

- **Posterior Mean:** The posterior mean μ_n of the weights is given by:

$$\mu_n = \Sigma_n \left(X_{\text{train}}^T Y_{\text{train}}\right) \beta$$

  where Y_{train} is the vector of target values for the training set.

- **Gamma Hyperparameters:** The hyperparameters of the Gamma distributions are iteratively updated as follows:
  - $an = a0 + D/2$
  - $bn = b0 + (0.5 \times (\mu_n^T \cdot \mu_n + \text{trace}(\Sigma_n)))$
  - $cn = c0 + N/2$
  - $dn = d0 + 0.5 \times \sum(Y_{\text{train}} - X_{\text{train}} \cdot \mu_n)^2$

  These updates refine the shape and scale parameters of the Gamma distributions, representing the precision parameters α and β.

- **Iterative Process:**

  The VI approach iteratively refines its estimates for the posterior distribution and the Gamma-distributed precision parameters over 1000 iterations

3. Results:

- **Hyperparameter Expectations:** After 1000 iterations, the expected values of the hyperparameters α and β are computed as follows:

$$\alpha_{\text{expected}} = an/bn$$

$$\beta_{\text{expected}} = cn/dn$$

  For this task, the final values are:
  - Expected α : 0.0119
  - Expected β : 0.1103

- **Model Performance:** The model's performance is evaluated using the Mean Absolute Error (MAE) metric:
  - Training Set MAE: 2.1292
  - Test Set MAE: 2.0660

Report Summary

The data analysis, that catalogs various building characteristics, underscored the efficacy of the linear models in the prediction of heating load. Critical factors like total Roof Area, Surface Area, Relative Compactness, Wall Area and Overall Height were the main predictive factors for the outcome prediction. These factors had strong linear relationship with the heating load, and thus, were the best candidates for linear predictive modeling.

In the project, there were a number of statistical methods used, including Hamiltonian Monte Carlo (HMC), Variational Inference (VI), and Type-II Maximum Likelihood Estimation. Each of these methods gave unique outputs, and the same was robust in modeling. HMC did provide a detailed posterior distribution of parameters. VI improved the estimates of hyperparameters effectively. Type-II Maximum Likelihood was useful in straightforward estimation of the parameters. All the different approaches based on their distinct methodologies showed congruent and compatible results, and the conclusions were very reliable for linear models in this context.