CM50268 Bayesian Machine Learning: Final Project

Exploratory Analysis

This section provides an exploratory analysis of the training dataset used in predicting heating load. We aim to understand the characteristics of the dataset, identifying relevant variables for heating load prediction, and assess their relationships with the target variable.

The dataset contains 384 observations, each described by nine features, this includes a base unit “x0”, and eight over variables (x1 to x8). The target variable is the heating load of a building which we aim to predict from the given features. As part of the preprocessing, these eight feature variables were normalised to have a mean of zero, this is crucial for removing scale effects that could be affect the values of performance of predictive models.

Firstly, we plotted histograms of each feature, using the seaborn python library, to examine their distributions. These are shown in Figure 1. These show the range and distribution of each feature, giving insight into their variability.

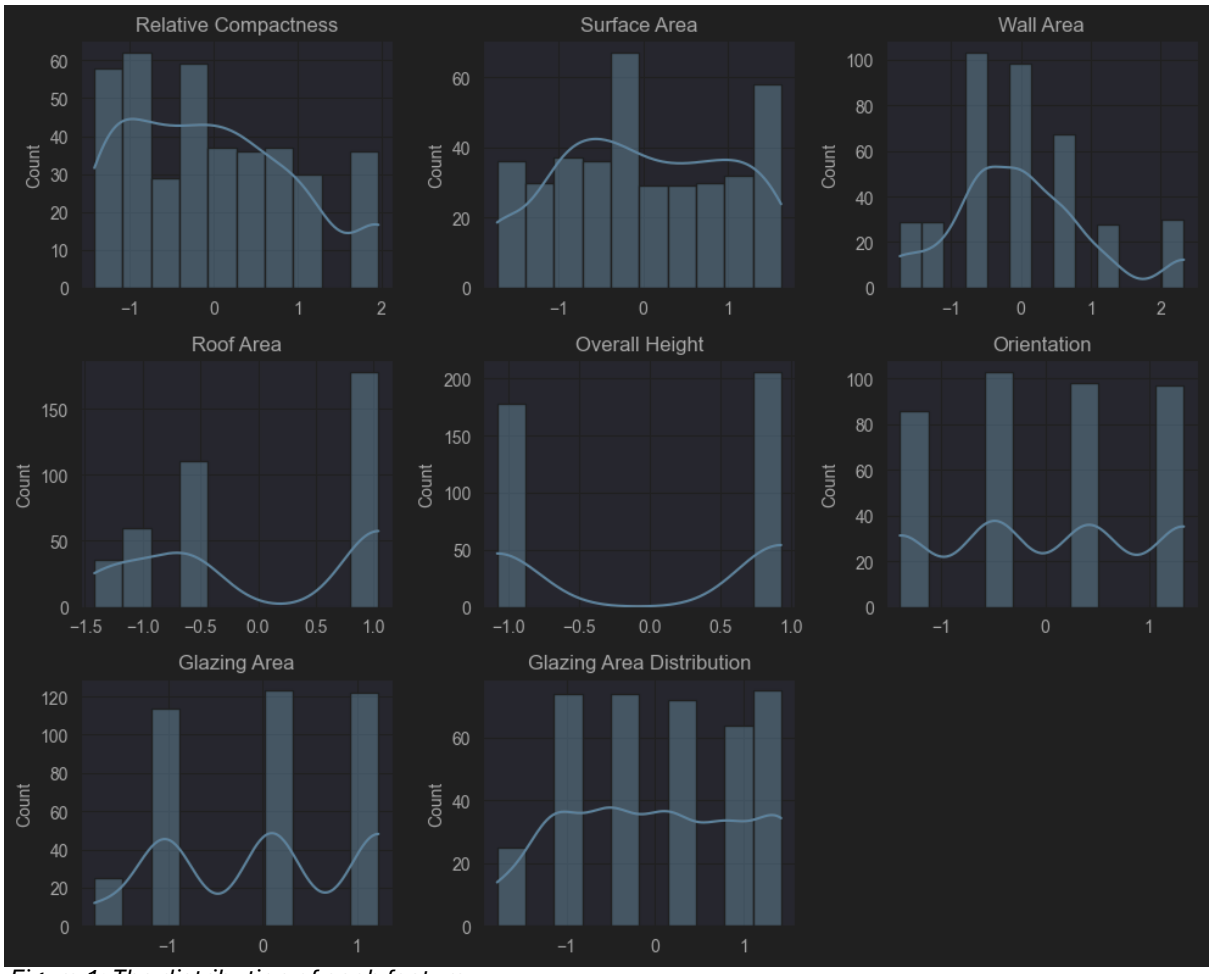

Figure 1: The distribution of each feature

To evaluate the relationship between ach feature and the heating load, scatter plots were used to highlight the patterns and trends, indicating potential linear or non-linear relationships. This is essential for determining if

linear regression could be a suitable modelling approach or if more complex models are needed. These are shown in Figure 2.

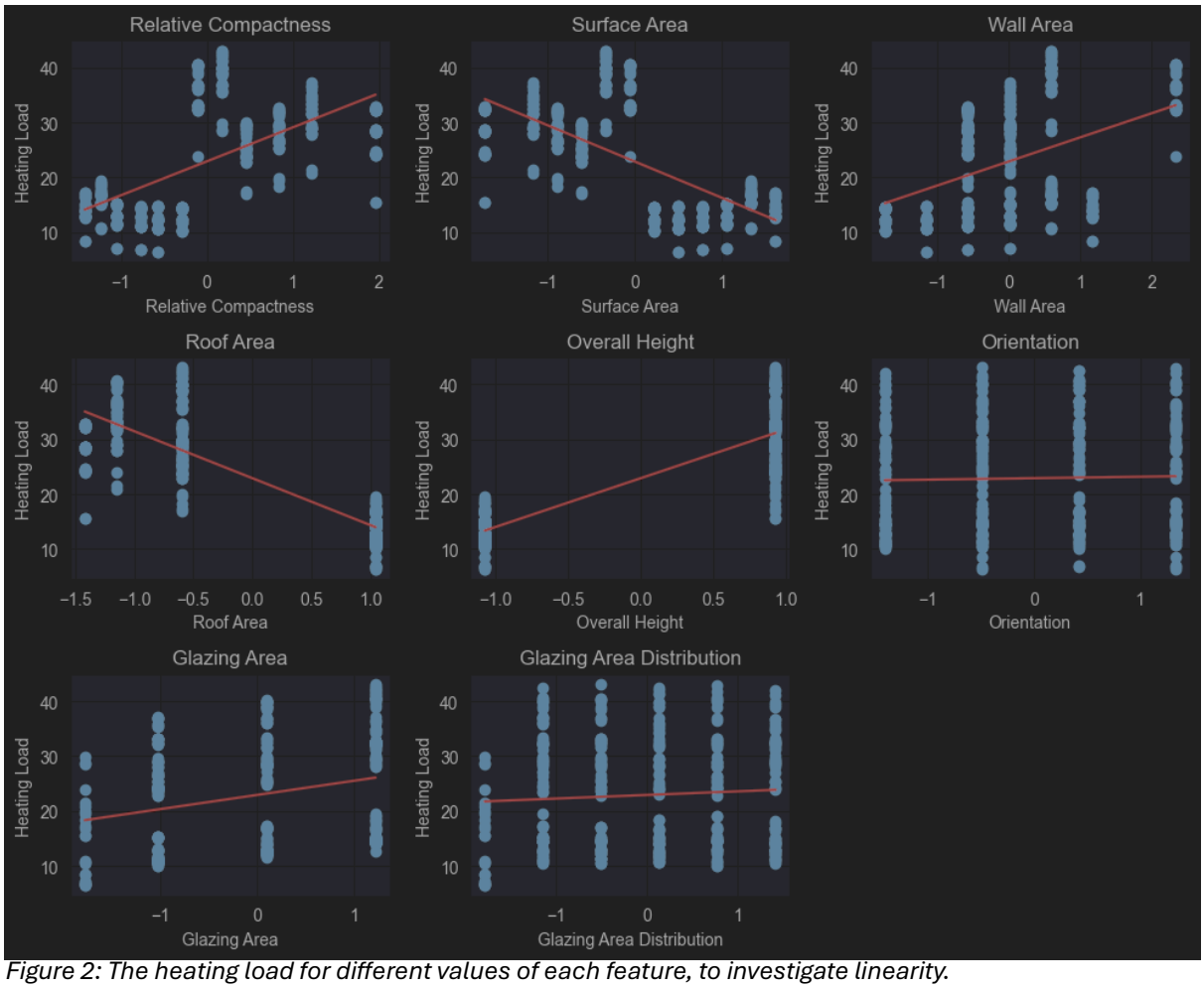

Figure 2: The heating load for different values of each feature, to investigate linearity.

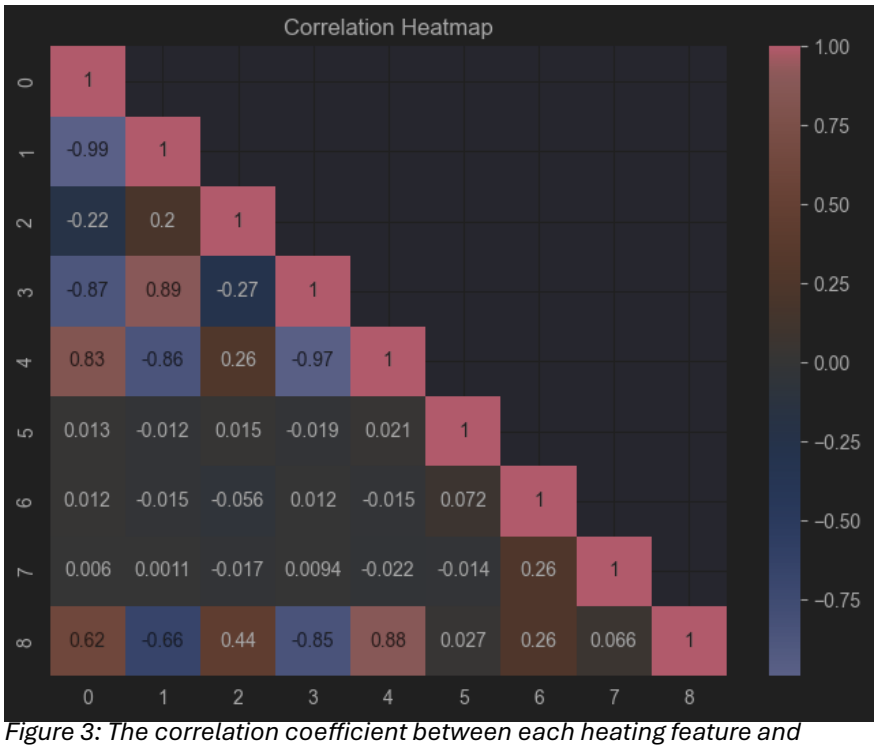

Figure 3: The correlation coefficient between each heating feature and heating load, and between each other.

Across the 8 variables we can clearly see a rough idea of what causes a higher heating load, and the variables with almost no relationship, such as Orientation and Glazing area distribution. We can also very clearly see that the Relative Compactness, Wall Area, Overall, Height and Glazing Area all have a positive linear effect on heating load. Contrarily, Surface Area and Roof Area both have a negative linear effect on the heating load. To give a more visual and quantitative measure of how each feature is related to the heating load, and to each other, we produced a correlation heatmap as shown in Figure 3.

The exploratory analysis suggests varying degrees of linearity in the relationships between features and the heating load, with some features potentially more relevant than others. The normalisation and preliminary visualisations support the initial understanding of the dataset which will guide the subsequent modelling efforts. To finalise the exploratory analysis, a predictive baseline was established by fitting a linear model to the training set using the least-squares method. The model was trained on the training set and evaluated on both sets. The MAE for the training set was 2.13, with a value of 2.07 for the test set, suggesting that the model generalised well to unseen data.

Bayesian Linear Regression

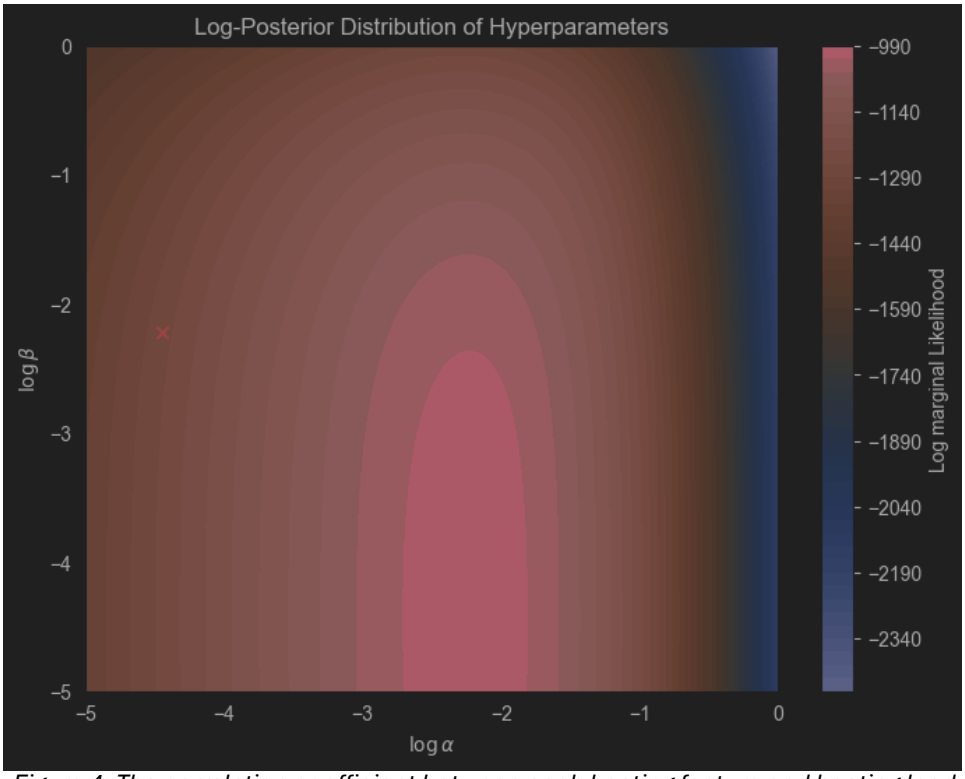

Figure 4: The correlation coefficient between each heating feature and heating load, and between each other.

Having analysed the dataset, we can now delve into the application of Bayesian linear regression for predicting a heating load. We employ Type-II maximum likelihood methodology to estimate the most probably values for the hyper-parameters. A visualisation of the posterior distribution is shown in Figure 4. This aids in understanding the distribution of the parameters of interest, with the most probably value marked.

The most probable values of the hyper-parameters are estimated to be α = 0.012, β = 0.108, and the log-likelihood corresponding to these values of -1001. The performance of the Bayesian linear regression model is evaluated based on the Root Mean Squared Error (RMSE) of predictions on both the training and test sets. The training RMSE is 3.012, with the test RMSE at 2.843.

HMC verified on a standard 2D Gaussian.

In this section, we verify the effectiveness of Hamiltonian Monte Carlo (HMC) by applying it to a standard 2D Gaussian example. The purpose is to validate the functionality of out HMC implementation, visualising its performance. To apply HMC, energy_func and energy_grad are defined, relating to the negative log probability of the 2D variables, and the gradient of the energy function, respectively. The standard 2D Gaussian probability density function is given by:

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(1 - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \quad (1)$$

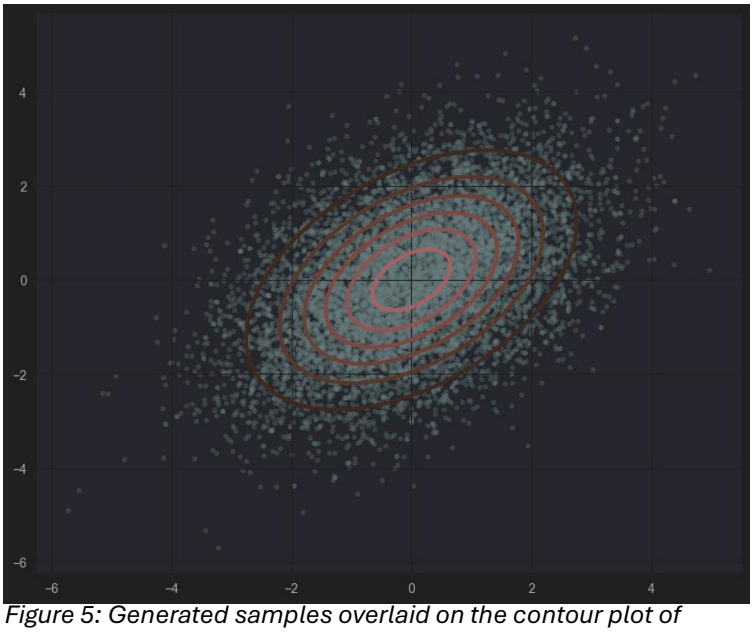

Figure 5: Generated samples overlaid on the contour plot of the PDF to visualise their distribution and coverage.

We verify the implementation of HMC by visualising the performance of the sampling process, including a contour plot of the Probability Density Function (PDF), with the generated samples overlaid as shown in Figure 5. The parameters we use are:

- Number of Samples, R: 10,000
- Leapfrog steps, L: 25
- Step Size, ε: 1.2

The acceptance rate finalises at 90.5%, around the expected value.

Applying HMC to the Linear Regression Model

Now, we were able to apply the HMC to obtain samples from the joint posterior over linear regression coefficients and hyperparameter sets for the linear regression model using the energy efficiency data. We defined the energy function its gradient again, tailored to the linear regression model. The energy function represents the negative log probability of the parameters under the Gaussian likelihood and Gaussian priors, while the energy gradient computes the partial derivatives of the same function, with respect to the parameters.

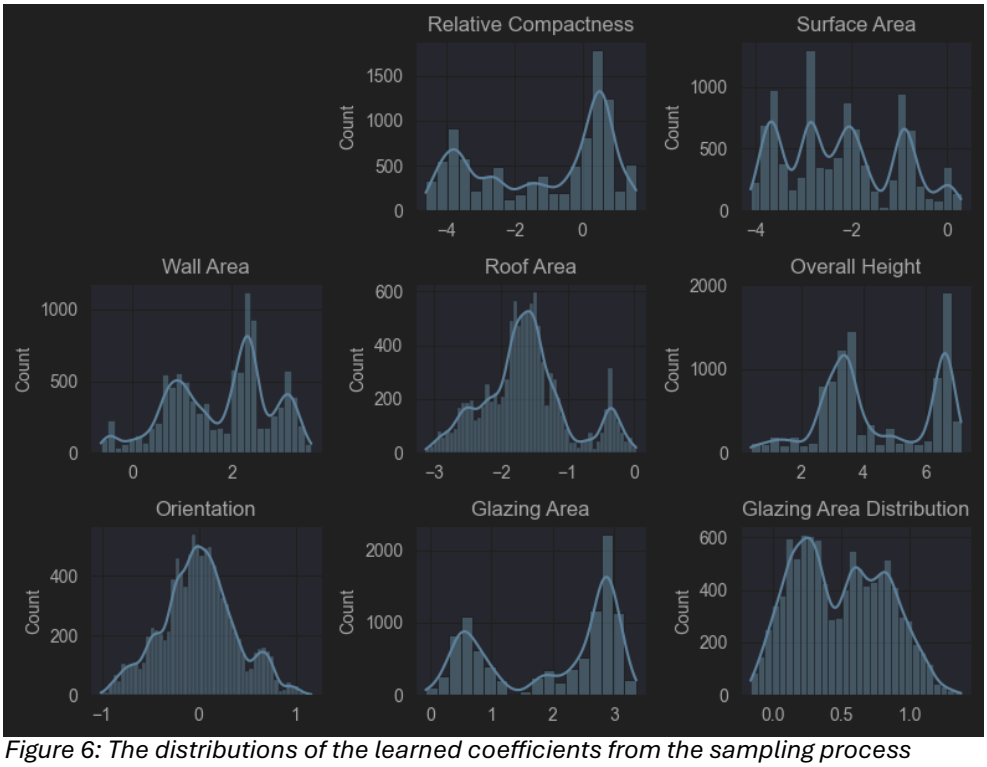

Figure 6: The distributions of the learned coefficients from the sampling process

Initialising the sampling process with random initial states and specified hyperparameters and a burn in parameter:

- Number of Samples, R: 10,000
- Leapfrog steps, L: 20
- Step Size, ε: 0.001
- Burn in period: 1,000

This gave an acceptance rate of 88.6%. We were also able to visualise the distributions of the learned coefficients obtained from the sampling process in histograms, as shown in Figure 6. This gave a Training and Test RMSE of 7.96 and 7.62 respectively, and a

Training and Test MAE of 7.04 and 6.84 respectively. These obtained results indicate that the HMC method effectively explored the parameter space. These error values are much larger than the ones mentioned earlier in the report due to randomness of the sampling process and assumptions of the model.

Apply HMC as a Classifier

Here, our aim was to use the HMC algorithm as a classifier to predict a binary class indicator based on the heating data, predicting whether the heating load is high or not. We first label all cases with a heating load greater than 23.0 as positive. By modifying the energy function and its gradient to use the Bernoulli likelihood with a sigmoid link function, we have effectively converted the problem into a logistic regression model. The initial state is defined again with specified hyperparameters:

- Number of Samples, R: 10,000
- Leapfrog steps, L: 20
- Step Size, ε: 0.2
- Burn in period: 1,000

This gives an acceptance rate of 83.8%, roughly what is expected still. We also compute the misclassification rate on the test set using the obtained samples, representing the proportion of misclassified instances compared to the true labels. We achieve a value of 0.78%, which is impressive as it suggests a very small number of cases are misclassified, showing the model's ability to effectively predict whether the heating load is high or not based on the given features.

Variational Inference

Finally, we apply Variational Inference with simple Mean-Field Theory factorisation to estimate the most probably values for the hyperparameters of our linear regression model. We define a function "VI" that implements the VI algorithm with mean-field approximation. This function updates the hyperparameters and posterior parameters to approximate the true posterior distribution. The expectations of alpha and beta were then produced, representing the most probably values for the hyperparameters. We then calculated the RMSE for both the training and test sets using the obtained posterior mean.

- Expectation of alpha: 0.012
- Expectation of beta: 0.108
- RMSE of the training set: 3.012
- RMSE of the test set: 2.843

Summary

Across our exploration of heating load prediction, we completed an analysis of the dataset aiming to find relationships between different features and the heating load. Through early histograms and plots we were able to identify which variables showed linear and non-linear correlations with the heating load. Throughout the project, we applied various machine learning techniques. Bayesian linear regression provided a principled approach to modelling uncertainty, while the Hamiltonian Monte Carlo methods showed their effectiveness in exploring complex parameter spaces. Additionally, variational inference offered an efficient alternative, using mean-field theory to approximate the posterior distribution. By reporting the performance of each method, we are able to gain insight into their strengths and limitations in this task.