OPTIMISING HEATING SYSTEM OPERATION

QBUS2820 Individual Assignment

CONTENT

1. Executive Summary

The purpose of this report is to build an accurate predictive model for the variable *HeatingLoad*. This is done to help optimise the heating system operations. Being able to predict the total daily heating energy for a building will improve the energy efficiency in buildings, reduce costs and minimise environmental impact.

The main dataset used for analysis in this report is '*HeatingLoad training.csv*', this training dataset is used for modelling and choosing models. The final model will be evaluated based on the test dataset '*HeatingLoad test.csv*'. The dependent variable or response in the dataset is *HeatingLoad* and, and there are 7 covariates in the dataset. The description of each variable is shown in Table 1.

Predictors	Description
HeatingLoad	Total daily heating energy required (in kWh)
BuidlingAge	Age of the building (in years)
BuidlingHeight	Height of the buildings (in metres)
Insulation	Insulation quality (1 = Good, 0 = Poor)
AverageTemperature	Average daily temperature (in ^oC)
SunlightExposure	Solar energy received per unit area (in W/m^2)
WindSpeed	Wind speed at the building's location (in m/s)
OccupancyRate	Proportion of the building that is occupied (%)

Table 1. *Description of Variables*

To choose the variables in the model, forward-stepwise is used. Then by doing cross-validation, all the models will be assessed based on three metrics: RMSE, MAE and R^2. Finally, the chosen predictive model will be used to predict *HeatingLoad* on the test dataset. This is to see how the model performs on an unseen dataset, hence evaluating the model's accuracy.

2. Exploratory Data Analysis (EDA)

The dataset has 10,000 rows with no null data and duplicates. All variables data type is float except *Insulation* which is a binary variable (0 for bad insulation and 1 for good insulation). Table 2 below summarises the statistics of each variable.

	Building Age	Building Height	Insulation	Average Temperature	Sunlight Exposure	Wind Speed	Occupancy Rate	HeatingLoad
Mean	22.7655	20.7921	0.5960	18.0249	271.3616	4.4907	0.5001	260.0786
Std	12.3860	16.8827	0.4907	4.0907	229.2752	2.5312	0.2220	74.5919
Min	2.9900	3.0700	0.0000	1.6800	1.1500	0.0700	0.0100	173.6800
Max	153.8800	106.3600	1.0000	34.3400	1250.7100	18.9100	1.0000	793.9200

Table 2. *Variables statistics*

BuildingAge, BuildingHeight, SunlightExposure, WindSpeed and *HeatingLoad* have a high range compared to other variables. The maximum values of these variables are also far from the mean, which suggests that there might be outliers. Moreover, the response variable, *HeatingLoad*, has a right-skew distribution (Figure 3) with most of the value between 210 kWh and 250 kWh.

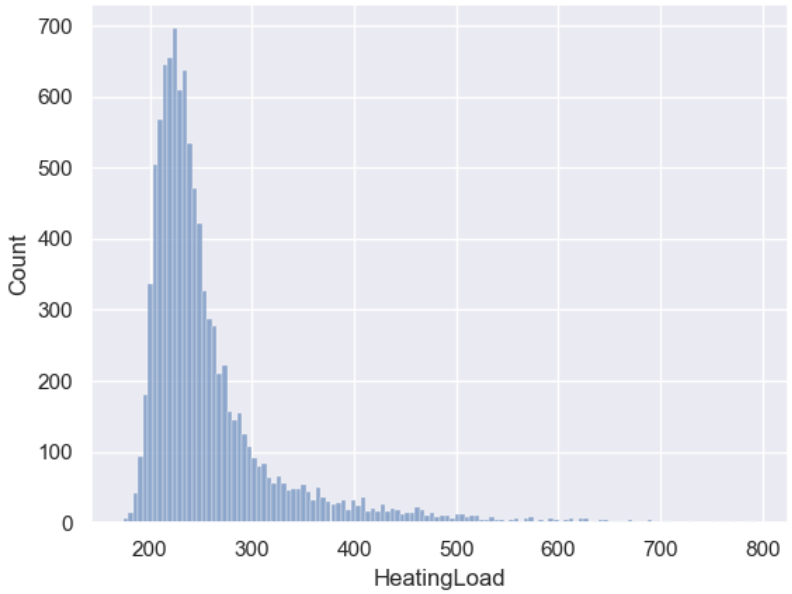

Figure 3. *HeatingLoad histogram plot*

To explore the outliers further, Figure 4 shows the boxplot of each variable and based on these graphs, it's true that the variables mentioned above have a lot of outliers. However, I won't remove or drop any of these data as the sample size will reduce significantly. These outliers are also important data points to help build the prediction model.

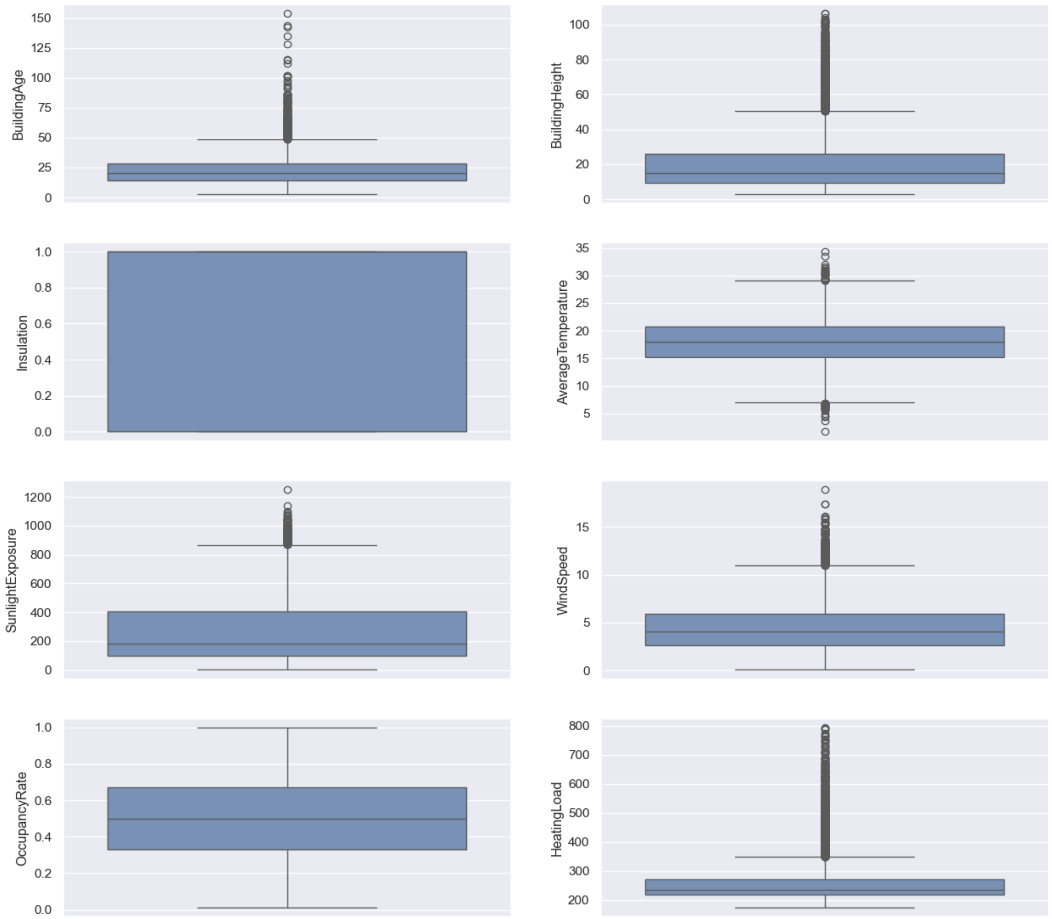

Figure 4. *Boxplot graph of variables*

Focusing on the response variable, the correlation matrix in Figure 5 shows that *HeatingLoad* and *BuildingHeight* are highly correlated (positive). *BuildingAge* and *SunlightExposure* also show a moderate positive correlation with *HeatingLoad*. For other variables, except the correlation between *SunlightExposure* and *BuildingHeight*, all other variables have a very low correlation with each other.

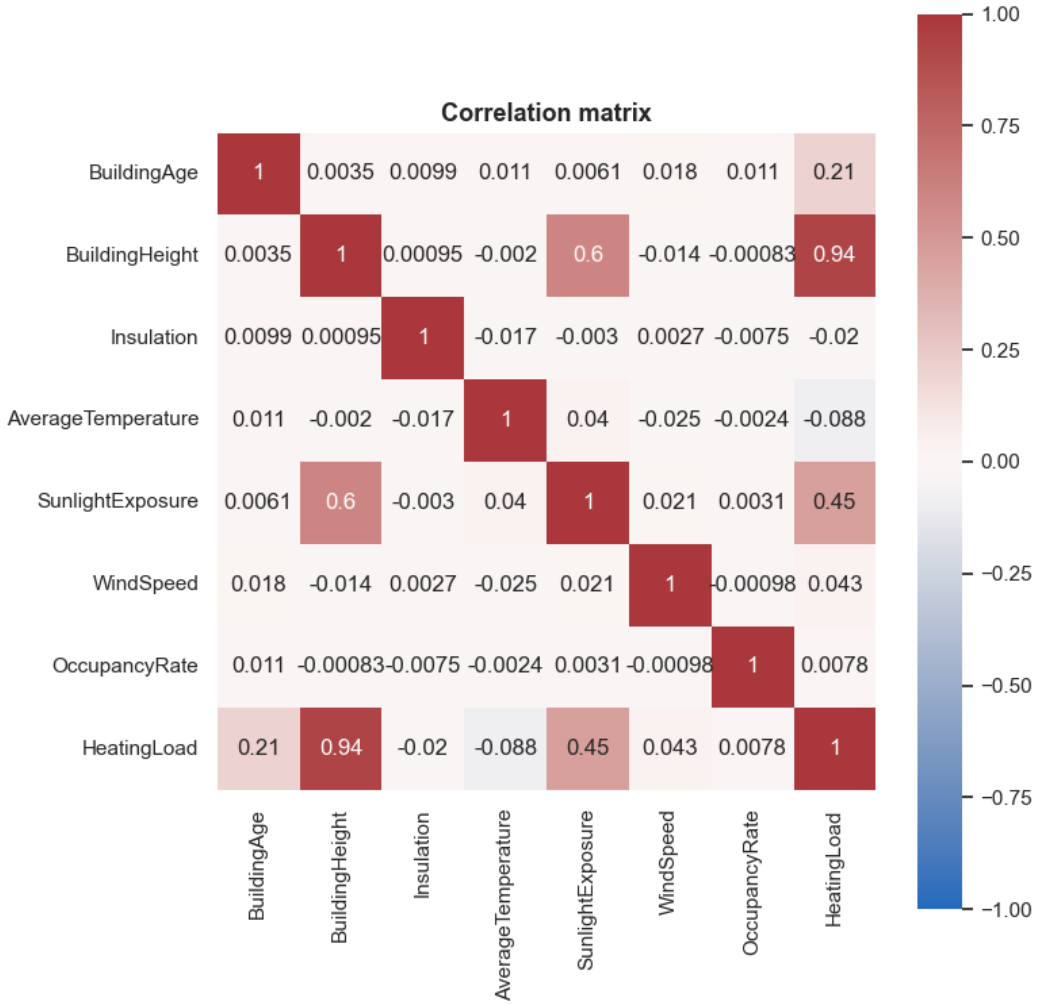

Figure 5. *Correlation matrix*

As the purpose of this report is to find the most accurate predictive model, I will further explore the relationship between *HeatingLoad* and other variables. Figure 6 shows the scatter plot and the regression line for each predictor (except *Insulation* as it's a binary variable) with *HeatingLoad*.

The plot between *BuildingHeight* and *HeatingLoad* suggests an upward curve which might indicate a polynomial relationship between the two variables. *HeatingLoad* and *BuildingAge* also seem to have a positive linear relationship. For other predictors, the regression lines have a small slope and are almost flat, suggesting a weaker relationship than *BuildingHeight* and *BuildingAge.*

For the relationship between *HeatingLoad* and *Insulation*, when splitting the *HeatingLoad* into having good insulation and having bad insulation, the mean value of *HeatingLoad* does not have a significant difference (Figure 7). The distribution of *HeatingLoad* is also similar (Figure 8). This suggests that *Insulation* might not have a significant influence on the *HeatingLoad*.

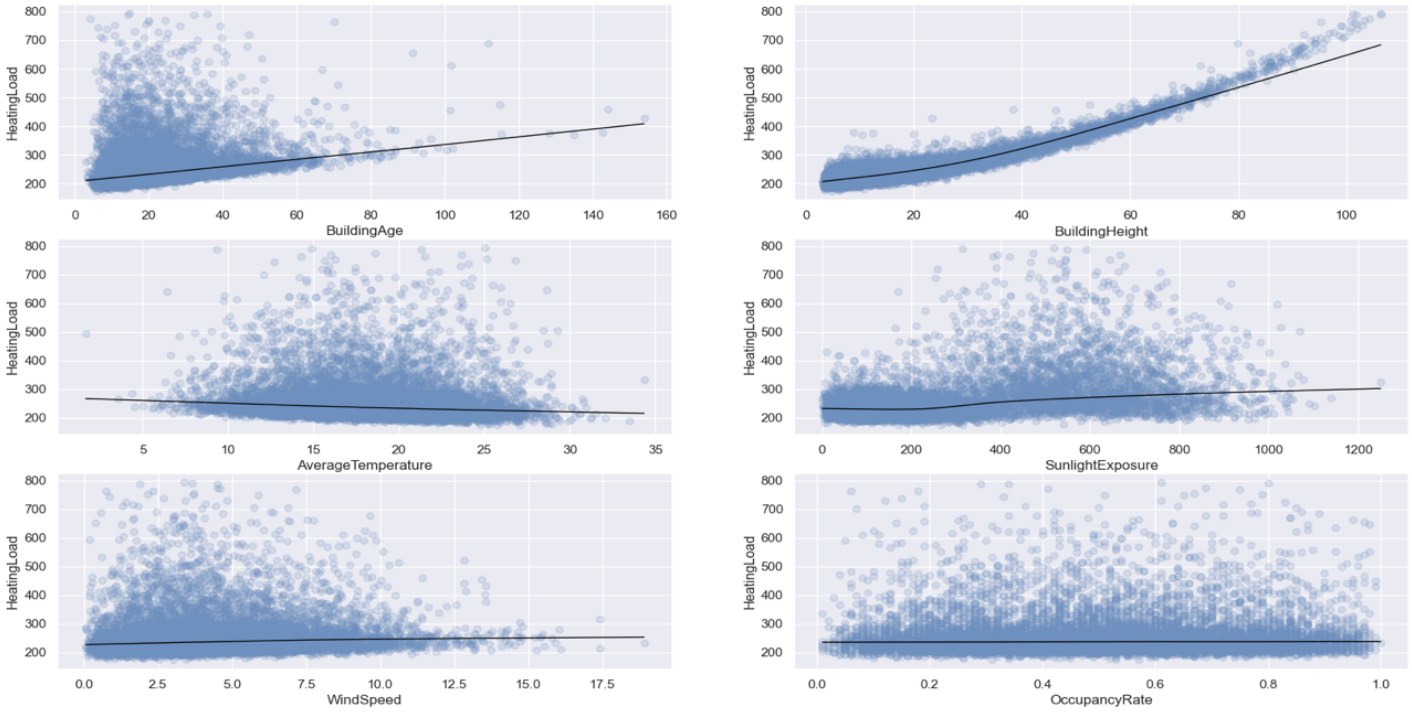

Figure 6. *Scatter plot between HeatingLoad and predictors*

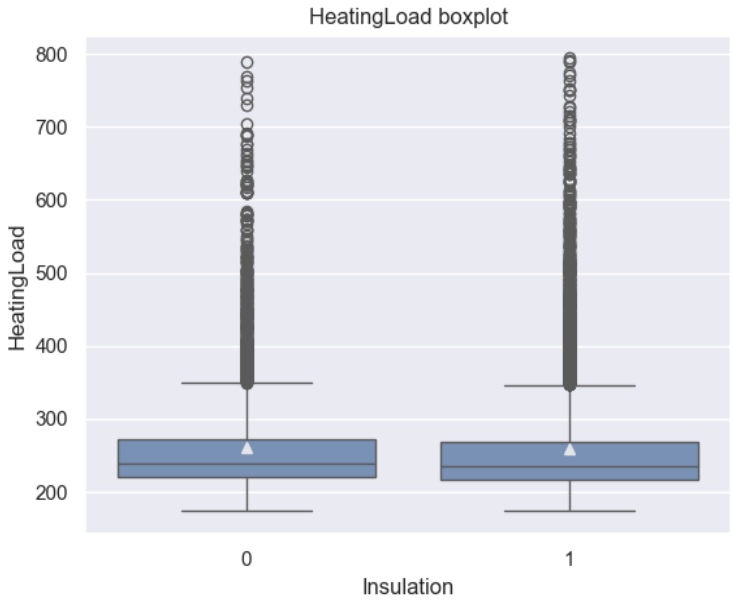

Figure 7. *HeatingLoad boxplox based on Insulation*

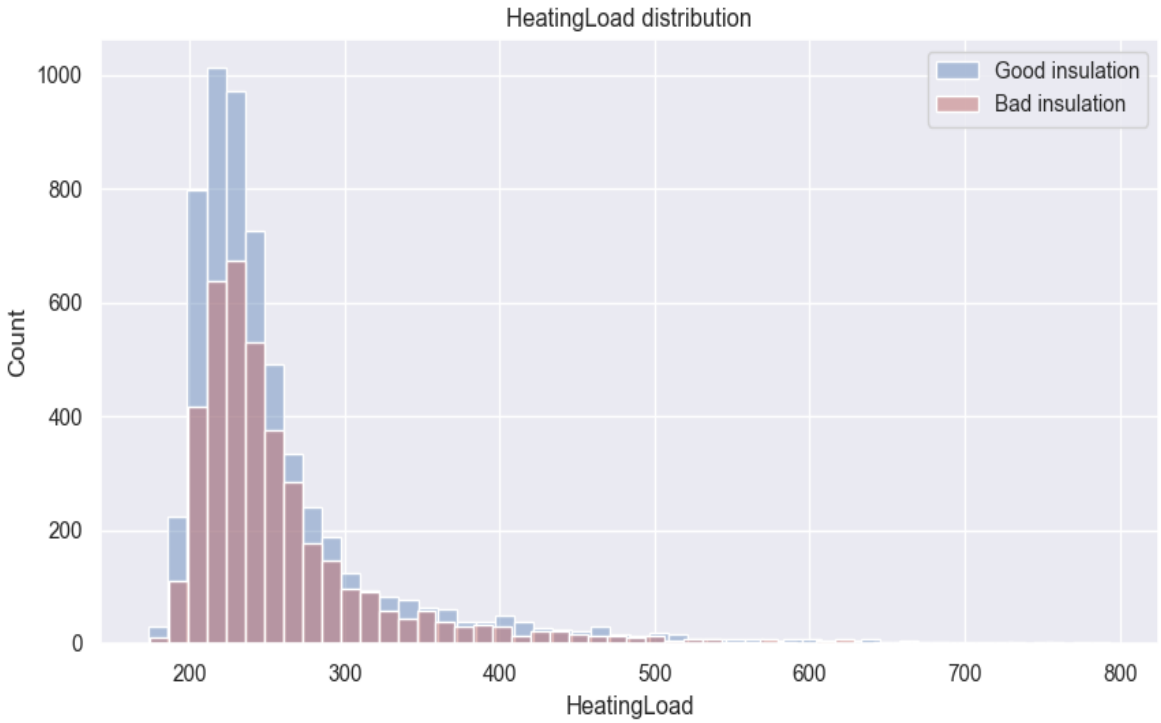

Figure 8. *HeatingLoad distribution based on Insulation*

3. Variable selection and building prediction model

To build the most accurate predictive model, I will construct multiple models and evaluate them using cross-validation, focusing on RMSE for this section.

For variable selection, I will base my decisions on multiple factors and reasoning, supported by forward-stepwise selection. This method works by iteratively building a model, adding one additional predictor at a time until the model's performance no longer improves (based on adjusted R^2). This approach will help determine the optimal set of predictors for the model (Narisetty, 2020).

To evaluate certain models, I will employ cross-validation. This method involves splitting the dataset into a number of equal-sized sets of observations (for this report, I will use 10 sets) (Brownlee, 2023). In each iteration, one set serves as the validation set while the remaining nine are used to build the model. The RMSE on the validation set is then recorded. This process is repeated for all 10 sets, and the average RMSE is calculated. This approach provides an overall assessment of the model's performance on unseen data and aids in model selection.

For this section, I will use cross-validation specifically to choose the best polynomial regression and KNN regression models. The use of cross-validation for selecting the final model will be discussed in the subsequent section on Model Selection.

a) Model 1

Using forward-stepwise selection for all 7 predictors, the result shows that Model 1 includes all the predictors. Hence, the variables in this model are: *BuildingAge, BuildingHeight, Insulation, AverageTemperature, SunlightExposure, WindSpeed* and *OccupancyRate.*

b) Model 2

Model 2 has three predictors which are *BuildingHeight, SunlightExposure* and *BuildingAge*. These predictors have the highest correlation (Figure 5) with *HeatingLoad* compared to other variables. Therefore, they might be able to give a good prediction of *BuildingHeight*.

c) Model 3

As suggested in EDA, the scatter plot between *BuildlingHeight* and *HeatingLoad* indicates that there might be a polynomial relationship. For Model 3, I will determine the optimal degree of polynomial of *BuildingHeight* in relation to *HeatingLoad*. This is done by building 10 models each with a polynomial degree from 1 to 9. Then cross-validation is used to choose the model that has the lowest RMSE from the process. Figure 9 shows the plot between the polynomial degree and CV score (RMSE) of the 9 polynomial models.

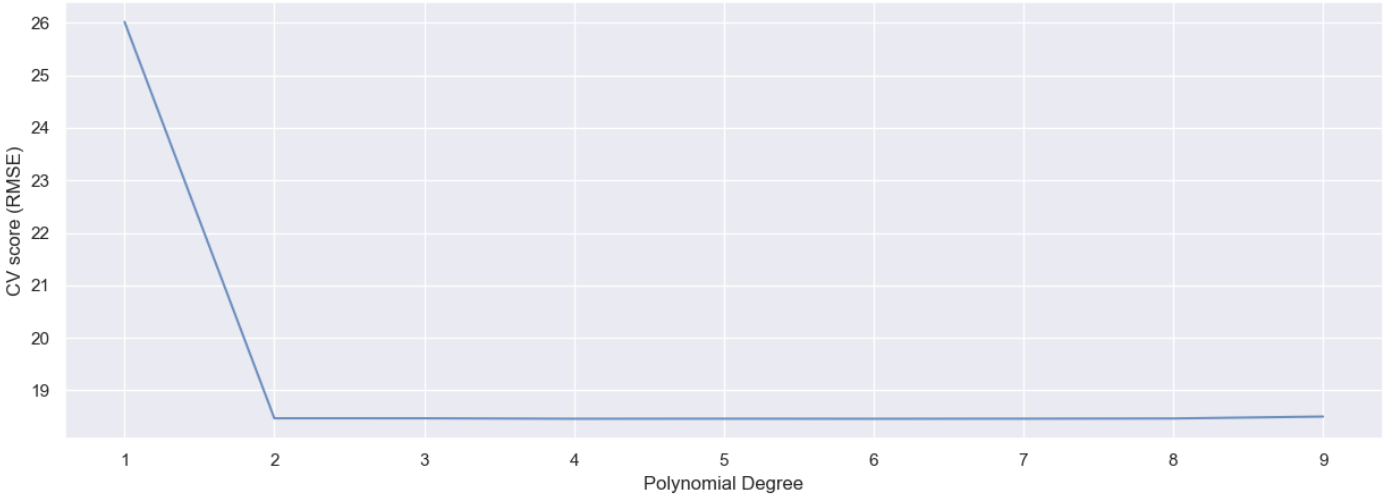

Figure 9. *BuildingHeight polynomial degree and CV score*

The polynomial degree with the lowest RMSE is 6. However, I decided to build Model 3 using the polynomial degree of 2 for *BuildingHeight*. This is because the difference between the RMSE of a model with a degree of 2 (18.4683) and a model with a degree of 6 (18.4586) is not significantly large (only 0.0526%). Therefore, by reducing the model complexity, the model will still perform similar. So Model 3 will have 2 predictors: *BuildingHeight* and *BuildingHeight_square*. Moreover, *BuildingHeight_square* will be kept as one of the predictors for the following models.

d) Model 4

Similar to Model 1, Model 4 is built based on forward-stepwise selection using all predictors and *BuildingHeight_square*. The result chooses all the given predictors: *BuildingHeight,*

BuildingHeight_square, BuildingAge, AverageTemperature, WindSpeed, SunlightExposure, Insulation, OccupancyRate.

e) Model 5

To account for the relationship between *BuildingHeight* and *SunlightExposure*, I have added an interaction term to the model (BH_SE). This decision is based on two key observations: first, as buildings increase in height, a larger area becomes exposed to sunlight due to reduced obstruction from surrounding structures. Second, Figure 5 demonstrates a high correlation between *BuildingHeight* and *SunlightExposure*.

Using forward-stepwise selection, the predictors for Model 5 are *BuildingHeight, BuildingAge, SunlightExposure, AverageTemperature, WindSpeed, Insulation, OccupancyRate, BuildingHeight_square, BH_SE*.

g) Model 6

In addition to the interaction between BuildingHeight and SunlightExposure, I have identified another important interaction: BuildingHeight and WindSpeed (BH_WS). This interaction is based on the principle that taller buildings tend to disrupt airflow more significantly, resulting in stronger wind currents. Similar to Model 5, using forward-stepwise selection, Model 6 has the following predictors: *BuildingHeight, BuildingAge, SunlightExposure, AverageTemperature, WindSpeed, Insulation, OccupancyRate, BuildingHeight_square, BH_SE, BH_WS.*

h) Model 7

Older buildings are more likely to have higher occupancy rates due to several factors. Firstly, older buildings have had more time to establish themselves in the market, allowing potential occupants to become aware of their existence. Secondly, these buildings may have developed a reputation or historical significance that attracts occupants. Therefore, I will add another interaction between *OccupancyRate* and *BuildingAge* (*OC_BA*) to reflect their relationship.

Again, with forward-stepwise selection, the predictors are *BuildingHeight, BuildingAge, SunlightExposure, AverageTemperature, WindSpeed, Insulation, OccupancyRate, BuildingHeight_square, BH_SE, BH_WS, OR_BA.* Model 7 includes all the predictors in the previous models.

i) Model 8

From Model 1 to Model 7, I have been using linear regression (parametric) to build the model. Another approach to constructing a predictive model is K-nearest Neighbors (KNN), a non-parametric method that makes no assumptions about the sample distribution. To predict the HeatingLoad, the KNN model takes the sum of k nearest points in the training data and averages the result to obtain the prediction (IBM, n.d.). To determine the optimal number of neighbors (k), I will use cross-validation to select the k value that yields the smallest RMSE, considering k values ranging from 1 to 100.

With 7 predictors, there are $2^7 - 1 = 127$ possible KNN models to compare. However, evaluating all these models would be inefficient. Instead, I will start by evaluating KNN models with only one predictor, which have the same complexity. I'll then select the best single-predictor KNN model and incrementally increase

the complexity to two predictors (changing only the second predictor). This process will be repeated until all 7 predictors are included in the model.

For single-predictor KNN models, I used Euclidean distance. However, for KNN models with multiple predictors, I employed Mahalanobis distance. This choice was made because the predictors have different scales (Table 2), which could cause some predictors to have minimal or no effect on the KNN if Euclidean distance were used (Chandavale, 2021).

The KNN model with the lowest cross-validation RMSE has the following four predictors: *BuildingHeight, BuildingAge, AverageTemperature, SunlightExposure* and 5 k-neighbours. Table 10 below shows the summary of the cross-validation RMSE for the best model for each model complexity (p) along with the optimal number of k-neighbours.

Complexity (p)	Predictors	K-neighbour	CV RMSE
1	BuildingHeight	45	18.8197
2	BuildingHeight, BuildingAge	8	11.2554
3	BuildingHeight, BuildingAge, AverageTemperature	9	9.8872
4	BuildingHeight, BuildingAge, AverageTemperature, SunlightExposure	5	9.8496
5	BuildingHeight, BuildingAge, AverageTemperature, SunlightExposure, WindSpeed	5	10.8709
6	BuildingHeight, BuildingAge, AverageTemperature, SunlightExposure, WindSpeed, Insulation	5	12.2016
7	BuildingHeight, BuildingAge, AverageTemperature, SunlightExposure, WindSpeed, Insulation, OccupancyRate	6	14.8455

Table 10. *KNN models result*

4. Model selection

From the training data, I have constructed eight different models. As previously mentioned, I will use cross-validation to select the best predictive model. To make this decision, I will assess each model's cross-validation RMSE, MAE (mean absolute error), and R^2.

Table 11 presents the cross-validation results for all the models. The optimal model should have the lowest RMSE and MAE, along with the highest R^2. However, it's crucial to note that R2 has a tendency to increase as the number of predictors increases (i.e., as the model becomes more complex). This means that R^2 does not penalize model complexity, which can introduce bias and inaccuracy when comparing models with different numbers of predictors.

Therefore, while R2 provides valuable information, it should be considered in conjunction with RMSE and MAE to ensure a balanced evaluation of model performance. This approach will help mitigate the potential bias introduced by R2 alone and lead to a more robust selection of the best predictive model.

Model	RMSE	MAE	R^2
1	16.9509	12.1025	0.9483
2	18.5615	13.4910	0.9379
3	18.4683	13.8532	0.9379
4	2.4416	1.9084	0.9989
5	2.0929	1.6731	0.9992
6	2.0893	1.6701	0.9992
7	2.0892	1.6704	0.9992
8	9.8496	6.4763	0.9824

Table 11. *Models RMSE, MAE and* R^2

From Table 11, Model 6 and Model 7 have the same R^2 value, however Model 6 has a lower MAE and Model 7 has a lower RMSE. The difference for both metrics between two models are not too significant- only 0.0001 for RMSE and 0.0003 for MAE. As the performance of the final model will be evaluated based on MSE (which is RMSE square), I will put more weight into RMSE when considering the optimal model. Hence, the optimal prediction model is Model 7.

By using all the training data, the MLR of Model 7 is

HeatingLoad =
$201.5803 + 1.4256 \times \text{BuildingHeight} + 0.0431 \times \text{BuildingHeight}^2 + 1.2555 \times \text{BuildingAge} - 0.0169 \times \text{SunlightExposure} - 1.5570 \times \text{AverageTemperature} + 1.6009 \times \text{WindSpeed} - 4.3336 \times \text{Insulation} + 1.1437 \times \text{OccupancyRate} - 0.0005 \times \text{BuildingHeight} \times \text{SunlightExposure} + 0.0031 \times \text{BuildingHeight} \times \text{WindSpeed} + 0.0154 \times \text{OccupancyRate} \times \text{BuildingAge}$

For the training dataset, this model has an adjusted R^2 of 0.999, meaning 99.9% of the variance in *HeatingLoad* can be explained by this model. Moreover, the model has a total of 11 predictors (including 3 interaction terms) with 4 predictors in the model have *BuildingHeight*. These might cause overfitting and multicollinearity, however, the result from the cross-validation indicates that this is not the case. The model still perform well on unseen data comparing to other models.

To evaluate the model further, a test dataset can be used to see how the model performs when shown a new dataset. The evaluation metric will be MSE (mean squared error). The lower the MSE, the more accurate the prediction is.

5. Conclusion

Through variable selection using forward-stepwise and model selection using cross-validation, the best predictive model has been built with 11 predictors.

The model has an average RMSE (cross-validation) of 2.0892 (Table 11) which the smallest value among the eight models. With a low error, the model will be able to predict the *HeatingLoad* accurate, provide valuable information to optimise the heating system operation in buildings. The model also used all the provided variables with interaction terms between them, this indicates that multiple factors contribute to the *HeatingLoad*. Finally, even though with multiple interaction term, it's difficult to interpret the model and assess individual predictor relationship with *HeatingLoad*, the purpose of the report is purely on accurate prediction. Therefore, this won't be a problem

References

Brownlee, J. (2023, October 4). *A Gentle Introduction to k-fold Cross-Validation - MachineLearningMastery.com*. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/

Chandavale, A. (2021, July 18). *KNN Distance Metrics and how they work Mathematically*. Medium. https://medium.com/@adityachandavale/knn-distance-metrics-and-how-they-work-mathematically-d58f65f0bd06

IBM. (n.d.). *What is the k-nearest neighbors algorithm?* IBM. https://www.ibm.com/topics/knn

Narisetty, N. N. (2020). Bayesian model selection for high-dimensional data. In *Handbook of Statistics* (Vol. 43, pp. 207-248). Elsevier. https://www.sciencedirect.com/science/article/abs/pii/S0169716119300380