Feature Selection & Regression Modelling for Climate Analysis Using Dutch Climate Data

Tuğba Nur Işık

M.Sc. Student, Ozyegin University, İstanbul, Turkey
Email: nur.isik@ozu.edu.tr

ABSTRACT

Climate prediction and detection of essential parameters that affect building energy consumption is a challenging process due to the complex relationship between climatic parameters. It is essential to identify the parameters affecting yearly temperature and determine their importance, especially for the climate resilience. Many tools successfully predict the climatic conditions, hence the space heating and cooling demands. However, these tools are complex for non-expert users and they may need data that may not be available. This study explores the potential of Linear Regression (LR) as an alternative method to uncover dynamics between climatic variables and predict Heating Degree Days (HDD) using historical data of Dutch Province De Bilt spanning 1800-2014. The research question addresses whether LR can effectively model the complex relationships in given climate data and predict the HDD. The study identifies multicollinearity and overfitting as key issues in initial models, mitigated through feature selection techniques, including Variance Inflation Factor (VIF) analysis and Backward Elimination. The final model's performance metrics show a significant improvement in balance between training and test results as R-Score for training set is 0.99 and the test set is 0.91 and Mean Squared Error (MSE) for training set is 15.77 and for test set is 38.17, justifying the application of LR, which predicts the HDD with a high R-Score. Limitations and opportunities for future research, such as incorporating Cooling Degree Days (CDD) and forecasting climate change impacts, are discussed.

Keywords: climate data, multiple linear regression, heating degree days, feature selection, data science

1. INTRODUCTION

Research Problem

The nature of climatic normal is very complex as it depends various factors such as yearly temperature, seasonal weather events, precipitation rate, evaporation and humidity. In the context of climate change, countries with a colder climate like the Netherlands are experiencing warming trends, with rising annual average temperatures and shifts in seasonal temperature patterns. As a result, the demand for heating energy may decrease over time whereas the demand for cooling energy may increase, leading to changes in energy consumption patterns. Understanding these shifts is essential for evaluating the implications of climate change on energy systems and for developing building energy retrofit strategies. *("The Future of Cooling," 2018)*

Heating Degree Days (HDD) is a metric widely used to estimate the energy required to heat buildings. It is calculated based on the difference between the outdoor temperature and a baseline indoor temperature, typically set at 18℃. While other variables, such as annual average temperature or precipitation, provide valuable insights into climate trends, HDD directly links these trends to building energy use and occupant activity. It also reflects the influence of climate on energy demand, making it a critical indicator in assessing building energy needs and consumption. *(Li et al., 2012)*

Research Gap

The impacts of climate change on building energy consumption patterns create a reliance in data-driven models and machine learning algorithms. Modeling climate change effects, in this case Heating Degree Days, requires a deep understanding of the underlying climatic variables and their interdependencies. In this regard, data analysis and machine learning algorithms offer powerful tools for identifying relationships between variables and hidden patterns that may not be immediately apparent through traditional methods. Machine learning methods are particularly useful for handling large data sets, revealing non-linear relationships between the variables, and accurate prediction. While advanced machine learning techniques methods are proven to be powerful for such tasks, they may hinder accessibility for beginner users. *(Ciulla & D'Amico, 2019)*

Research Question

This study focuses on Multiple Linear Regression (MLR) to demonstrate the potential of simpler yet effective models and offer an alternative approach for users with limited knowledge in machine learning methods. This study explores whether Multiple Linear Regression (MLR) can serve as an effective and accessible method to find which variables are most effective on Heating Degree Days variable.

Research Objective

The aim of this study is to analyze the patterns and the relationship between the variables, examine the variables that are most relevant to HDD and finally to evaluate the effectiveness of MLR in modelling HDD. This study contributes to the literature by showing the potential of simpler machine learning algorithms in predicting change in climatic conditions and discover the hidden relationships between the climate variables. Ultimately, presenting a feature-based climatic data prediction tool to be used in decision-making processes in building retrofitting.

1.1. Dataset

The Climate Data De Bilt data set, provided by the KNMI (Royal Netherlands Meteorological Institute), offers a historical climatic observation, making it a valuable resource for this study. Climatic data includes observations of variables including year-based, temperature-based and atmospheric conditions, collected between the years of 1800-2014. This dataset is open to public access through the Dutch Government's open data platform *(Climate Data De Bilt; Temperature, Precipitation, Sunshine 1800-2014, n.d.).* Historical data, such as provided in this dataset, is essential for examining change in climatic conditions over time and for designing predictive models.

1.2. Description of the Variables

In the following section, the variables in the dataset are briefly introduced, categorized and sorted based on their relevance to climatic context and their prevalence in the common climate of the Netherlands.

- **Year:** The year of observation, ranging from 1800 to 2014.

Temperature-based variables

- **Yearly Average Temperature:** The average temperature recorded across the entire year.
- **Winter Average Temperature:** The mean temperature during the winter season.
- **Winter Average Minimum Temperature:** The average of the lowest daily temperatures during the winter months.
- **Summer Average Temperature:** The mean temperature during the summer season.
- **Summer Average Maximum Temperature:** The average of the highest daily temperatures during the summer months.

Day-based variables

- **Heating Degree Days (HDD):** The sum of temperature differences below a specific threshold (e.g. 18℃).
- **Sunless Days:** The total number of days with no measurable sunlight.
- **Days with Fog:** The total number of days where fog was observed.
- **Days with Precipitation:** The total number of days with measurable rainfall and/or snowfall.
- **Ice Days:** Days where the maximum temperature remains below 0℃.
- **Frost Days:** Days where the minimum temperature dropped below ℃.
- **Snow Days:** Days where snowfall was observed.
- **Summery Days:** Days where sunlight was measured. Usually when maximum temperature exceeds 25℃.
- **Tropical Days:** Days where extreme hots were observed. Usually when maximum temperature exceeds 30℃.
- **Dry Days:** Days without precipitation.

Other variables

These variables relate to overall climatic conditions or measurements.

- **Hours of Sunshine:** The total number of sunlight hours recorded throughout the year.
- **Quantity of Precipitation:** The total amount of precipitation recorded. Presumably in millimeters since the unit was not shown
- **Relative Humidity:** Indicates how much water vapor the air contains and provides an estimate of the likelihood of precipitation. It is measured in percentages.
- **Evaporation:** The total amount of water lost through evaporation, typically measured in millimeters. It is inversely proportional to humidity and precipitation. If the air is humid, the evaporation rate decreases because the air is already saturated.

1.3. Focus Variable: Heating Degree Days

Heating Degree Days (HDD) is influenced by a range of climatic variables. For instance, variables such as yearly average temperature, winter average temperature and winter average minimum temperature are expected to have a *strong inverse* relationship with HDD, as the warmer outdoor temperatures reduce heating demand. Similarly, the variables like weather events such as fog and snow may also influence HDD by affecting heating requirements. Therefore, it is reasonable to make assumptions regarding the dynamics between the variables and their significance in determining HDD. However, the true extent of their influence can only be examined through statistical analysis, which will be covered in following sections.

In this study, the methodology includes correlation analysis and feature selection, followed with multicollinearity detection and feature elimination to identify the most critical predictors for HDD.

2. METHODOLOGY

In the Data Preprocessing part, attributes of the data set and the relationship between the variables are analyzed and the dataset's integrity was ensured by addressing missing values and inconsistencies. For the data engineering and coding, Jupyter Notebook platform and various libraries for Python are used.

2.1. Data Preprocessing

Data Cleaning: In this step, following steps were yearly data with missing values are removed and features with no significance to data set are removed from the dataset to ensure consistency.
Data Rearrangement: Names and the order of the features are re-arranged for enhanced data visualization.

Ultimately, the features are between the years of 1981-2014.

2.2. Data Analysis

Data Visualization: Initial descriptive analysis included correlation matrix using visualization tools (e.g. heatmap) highlighted potential multicollinearity. Figure (1) shows the heatmap that represents correlations between the variables.

Correlation Analysis: To interpret the correlations between the features, the Pearson Correlation Coefficient (r) is used. The Pearson correlation coefficient measures the linear correlation between variables and takes a value between -1 and +1. Table (1) summarizes the interpretation of corresponding values for the correlation rate. Intervals for the correlation rates are adjusted to the data set. *(Jaadi, 2021)* Table (2) shows the estimated correlations between HDD and other features.

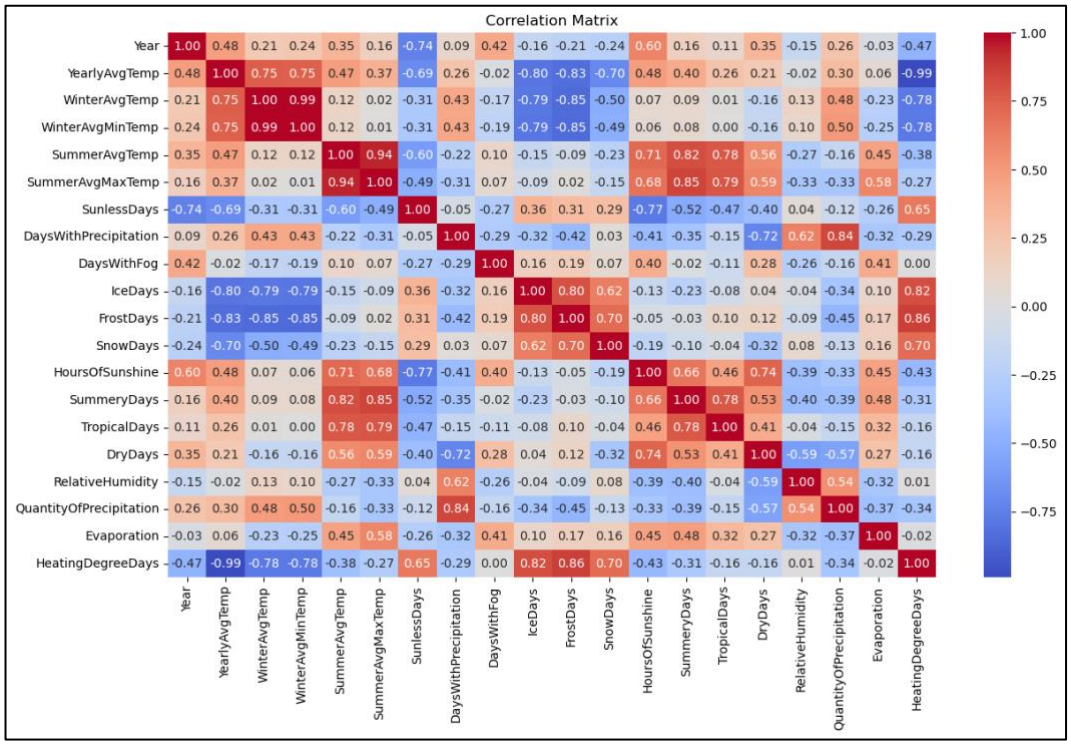

Figure 1. Correlation matrix of the variables

Table 1. Correlation coefficient

Correlation rate	Explanation
$0.7 \leq r < 1$	Strong positive relationship
$0.3 \leq r < 0.7$	Moderate positive relationship
$0 \leq r < 0.3$	Weak positive relationship
$-0.3 \leq r < 0$	Weak negative relationship
$-0.7 \leq r < -0.3$	Moderate negative relationship
$-1 \leq r < -0.7$	Strong negative relationship

Table 2. Correlation between HDD and other variables

Correlation range	Features	Correlation rate (r)
$0.7 \leq r < 1$	Frost Days	0.86
	Ice Days	0.82
	Snow Days	0.70
$0.3 \leq r < 0.7$	Sunless Days	0.65
$0 \leq r < 0.3$	Relative Humidity	0.01
	Days with Fog	< 0.00
$-0.3 \leq r < 0$	Evaporation	-0.02
	Tropical Days	-0.16
	Dry Days	-0.16
	Summer Average Maximum Temperature	-0.27
	Days with Precipitation	-0.29
$-0.7 \leq r < -0.3$	Summery Days	-0.31
	Quantity of Precipitation	-0.34
	Summer Average Temperature	-0.38
	Hours of Sunshine	-0.43
	Year	-0.47
$-1 \leq r < -0.7$	Winter Average Minimum Temperature	-0.78
	Winter Average Temperature	-0.78
	Yearly Average Temperature	-0.99

2.3. Multiple Linear Regression Model

Multiple Linear Regression is used when predicting a dependent variable that is a function of more than one predictor variable. The dependent variable is a linear function of random variables (x^t). The regression equation can be written as Equation (1) (Parsons, 2010b)

$$y^t = ax^t + b$$

$$E\ (r^t - y^t) = \frac{1}{N}\sum_{t=1}^{N}[r^t - (ax^t + b)]^2$$

$$a_i = \frac{\sum_{t=1}^{N}\left(x_i^t - \bar{x}_i\right)\left(r_i^t - \bar{r}_i\right)}{\sum_{t=1}^{N}\left(\left(x_i^t - \bar{x}_i\right)\right)} \quad \textbf{(1)}$$

$$b = \bar{r} - \sum_{i=1}^{F}\sum_{t=1}^{N} a_i\bar{x}_i$$

$$X = \{x^t, r^t\}_{t=1}^{N}$$

$$a = \{a_i\}_{i=1}^{F}$$

2.3.1. Error estimation metrics

Coefficient of determination (R^2): Coefficient of Determination (R^2) measures the model's predictability and evaluates how well the model represents the data set. The R^2 takes a value between 0 and 1. It is closer to **1,** so the developed model can predict the output variable. It is calculated using Equation (2) (Parsons, 2010b)

$$R^2 = 1 - \frac{\sum\left(r_i^t - y_i^t\right)^2}{\sum\left(r_i^t - \bar{r}\right)^2} \quad \textbf{(2)}$$

Where, y_i is the i-th expected output; r is the i-th predicted output and $\bar{r}$ is the average of the expected output.

Mean Square Error (MSE): MSE is used to calculate the variance between the target and the predicted values. It is calculated using Equation (3) (Parsons, 2010b)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(r_i - y_i)^2 \quad \textbf{(3)}$$

2.4. Initial Model Training and Evaluation

Data Splitting: Random and year-based Train/Test splits were applied to evaluate temporal dependencies. For the year-based split, data from 1981 and 2004 was used for training and 2004-2014 for testing. In Model 1, high test MSE suggested overfitting where in Model 2 temporal dependency issues were more evident. Comparison of model performance metrics are shown in Table (3).

Table 3. Model performances of Model 1 and Model 2

Model		Model 1		Model 2	
Description		Random Train-Test Split		Year-Based Split	
Set		Train Set	Test Set	Train Set	Test Set
Metrics	**R-score**	0.99	0.91	0.99	0.55
	MSE	248.76	1457.38	124.10	26552.16

These results point to two major issues:

1- Multicollinearity
2- Overfitting

Multicollinearity arises when independent variables in a model are highly correlated. This might lead to problems such as noise in the model where highly correlated predictors make it difficult for the model to train properly and confusion in interpretation which happens when useless information complicates the analysis and identification of significant variables. Overfitting arises when the model learned patterns and also the noise that it mimics the specific relationships between variables rather than the generalized ones. (Parsons, 2010b)

2.5. Multicollinearity Analysis and Feature Elimination

Variance Inflation Factor (VIF) is a tool used to identify multicollinearity. It measures how much the variance of the regression coefficient deviated because of multicollinearity. Variables that have high VIF values indicate that there is strong correlation with other variables. A general rule suggests that variables with VIF values above 10 are considered to be a problem for the model's learning hence needs to be eliminated. *(Shrestha, 2020)* VIF detection revealed high multicollinearity among several features, as shown in Table (4). These features were iteratively removed. Backward Elimination identified key features for HDD prediction. Selected features and their respective VIF rates are shown in Table (5).

Table 4. Variance Inflation Factor (VIF) values for variables

Feature	**VIF**	**Feature**	**VIF**
Year	15540.22	Days with Precipitation	300.47
Summer Average Max Temperature	8790.85	Evaporation	288.47
Relative Humidity	4829.96	Dry Days	107.07
Summer Average Temperature	3029.11	Frost Days	70.44
Yearly Average Temperature	1848.88	Quantity of Precipitation	32.25
Hours of Sunshine	752.75	Summery Days	20.17
Winter Average Temperature	499.01		

Table 5. VIF values of selected features

Selected Features	**VIF**
Tropical Days	1.98
Winter Average Minimum Temperature	2.597
Snow Days	6.85
Days with Fog	7.27
Ice Days	7.41
Sunless Days	8.99

2.6. Final Model Training and Evaluation

After the elimination of features, the final model achieved a balance between training and testing performance, addressing overfitting and multicollinearity. Table (6) shows the comparison between the performance metrics of Model 1 and Final Model. Figure (2) shows the best-fit line which shows the Final Model's prediction performance.

Table 6. Comparison of initial and final models

Model		Model 1		Final Model	
Description		Random Train-Test Split		Random Train-Test Split Feature Elimination	
Set		Train Set	Test Set	Train Set	Test Set
Metrics	**R-score**	0.99	0.91	0.99	0.91
	MSE	248.76	1457.38	15.77	38.17

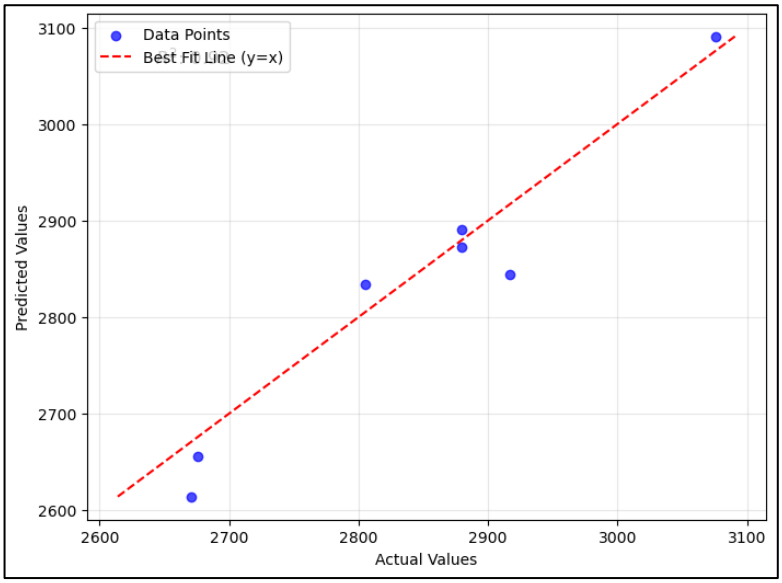

Figure 2.Actual vs. Predicted HDD Values for the Final Model

3. CONCLUSION

This study demonstrates that MLR can effectively model HDD when multicollinearity is addressed. The results of the final model's balanced performance highlight the potential of feature selection techniques in improving interpretability and reliability. The evaluation of three modeling approaches showed that in Model 1, where train and test data was split randomly, a high training and test R-scores, 0.99 and 0.91 were achieved, respectively. However, the MSE for train set was 248.76, whereas MSE for test set was 1457.38, considerably higher than the train set. These imbalanced and high scores of MSE values suggested a potential overfitting in learning test data patterns.

Therefore, in Model 2, an alternative approach for train and test sets was applied. Considering that the HDD variable changes with respect to years, the year-based split was employed, where the 70% of the data represented the train set, the data recorded between the years of 1981-2004 and the 30% of the data represented the test set which is the data recorded between the years of 2004-2014. Results of Model 2 shown a strong performance on training set, R-score of 0.99 and achieving a lower MSE of 124.10 but its test performance was significantly worse with a lower R-score of 0.55 and a much higher MSE of 26552.16.

These results were pointing out a possibility of multicollinearity between the variables, hence the model was struggling with confusions and noises. To prevent the confusion, feature selection was applied using Backward Elimination method. Features with high VIF scores were eliminated from the dataset. In Final Model, where the remaining features had a relatively low VIF values, a high performance for both training and test sets were achieved. R-scores of test and training sets were 0.99 and 0.91 and MSE for test and training sets were 15.77 and 38.17, respectively.

4. CONTRIBUTION AND FUTURE WORK

This study shows the importance of tailoring feature selection and the multicollinearity problem to enhance model performance. The results of this study justified the use of The Multiple Linear Regression as a powerful tool, considering its ability to simplify the complex interactions between the input parameters and its simplicity. Moreover, this study revealed the hidden dynamics between climate variables.

Limitations of this study consist of lack of information of the dataset: such as absence of Cooling Degree Days (CDD) variable, which is also an important feature for climate change impact estimation and limited historical data. For the future work, this study could be used to forecasting future HDD data. Furthermore, this research could be used as a basis for developing a CDD prediction model.

5. REFERENCES

- The future of cooling. (2018). In OECD eBooks. https://doi.org/10.1787/9789264301993-en
- Li, D. H., Yang, L., & Lam, J. C. (2012). Impact of climate change on energy use in the built environment in different climate zones – A review. Energy, 42(1), 103–112. https://doi.org/10.1016/j.energy.2012.03.044
- Ciulla, G., & D'Amico, A. (2019). Building energy performance forecasting: A multiple linear regression approach. *Applied Energy*, *253*, 113500. https://doi.org/10.1016/j.apenergy.2019.113500
- *Climate data De Bilt; temperature, precipitation, sunshine 1800-2014*. (n.d.). Data Overheid. https://data.overheid.nl/en/dataset/4818-climate-data-de-bilt--temperature--precipitation--sunshine-1800-2014#panel-description
- Jaadi, Z. (2021, December 12). Everything you need to know about interpreting correlations. *Medium*. https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8
- Parsons, S. (2010b). Introduction to Machine Learning, Second Editon by Ethem Alpaydin, MIT Press, 584 pp., $55.00. ISBN 978-0-262-01243-0. *The Knowledge Engineering Review*, *25*(3), 353. https://doi.org/10.1017/s0269888910000056
- Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39–42. https://doi.org/10.12691/ajams-8-2-1