# Prediction of Moisture content in Tall wood building envelope using Supervised Multiple Linear Regression models

Report Prepared by:

Saranya Krishnasami

# Contents

**Prediction of Moisture content in Tall wood building envelope using Supervised Multiple Linear Regression models**

# 1 Abstract

# 2 Introduction and background work

# 3 Methods

## 3.1 Description of the evaluated Multiple Linear Regression models

Model - Linear Regression: This use case is Multiple Linear Regression (MLR), which is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. The equation for multiple linear regression is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \quad (1)$$

Where:

- y is the dependent variable (target)
- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the coefficients (parameters) to be estimated and p is number of features
- $x_1, x_2, \dots, x_p$ are the independent variables (features)
- $\varepsilon$ represents the error term

The assumptions of Linear Regression are:

- A linear relationship should exist between the response and predictor variables
- Homoscedasticity is residuals having equal variance for each value of the fitted values and of the predictors

- Normality of residuals
- MLR assumes less or no multicollinearity (correlation between the independent variables) in data

Model - Lasso Regression: Lasso regression follows linear regression (equation (1)) but is efficiently used for prediction and feature selection.

Lasso (Least Absolute Shrinkage and Selection Operator) is a type of regularization technique that aims to find the values of coefficients that minimizes the sum of the squared difference between the actual and predicted values. The regularization term,

$$L1 \text{ Regularization } = \lambda * (|\beta_1| + |\beta_2| + \ldots + |\beta_p|) \quad (2)$$

Where:

$\lambda$ is the regularization parameter that controls the amount of regularization applied.

$\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients and p is number of features

- $\lambda$ denotes the amount of shrinkage
- $\lambda = 0$ implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model
- $\lambda = \infty$ implies no feature is considered i.e., as $\lambda$ closes to infinity it eliminates more and more features
- The bias increases with increase in $\lambda$, variance increases with decrease in $\lambda$

Model - Ridge Regression: The ridge regression follows L2 Regularization which also aims at reducing the sum of squared errors. Unlike lasso, ridge does not eliminate the parameters by penalizing rather reduces. In Ridge regression, denoting the cost function of a linear regression is altered by adding a penalty term equivalent to square of the magnitude of the coefficients as shown in equation (3),

$$L2 \text{ Regularization } = \lambda * \left( (\beta_1)^2 + (\beta_2)^2 + \ldots + \left(\beta_p\right)^2 \right) \quad (3)$$

$$\text{Cost Function} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{k=1}^{p} \left(\beta_j\right)^2 \quad (4)$$

Where Residual Sum of Squares,

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \Sigma_{j=1}^{p} \beta_j x_{ij}\right)^2 \quad (5)$$

Ridge regularization (L2), aims to find the values of coefficients ($\beta_1$, $\beta_2$, ..., $\beta_p$ are the coefficients) that minimizes the sum of the squared difference between the actual and predicted values. The difference between Lasso and Ridge is that the penalty term is based on the sum of the squares of the regression coefficients, rather than their absolute values

Model - Elastic Net: Lasso regressor sometimes does not perform well with highly correlated variables, and often performs worse than ridge in prediction.

To overcome these limitations, a penalty that combines the L1 and L2 constraints has been developed. An elastic net is a regularization [16][17] and variable selection procedure that makes use of the penalty (Equation 6).

$$L1 + L2 \text{ Regularization } = \lambda \left[ \frac{1}{2} (1 - \alpha) \sum_{j=1}^{P} \beta_j^2 + \alpha \sum_{j=1}^{P} |\beta_p| \right] \quad (6)$$

where,

$\lambda$ has the usual interpretation, regularization parameter

$\alpha \in [0, 1]$ is called the mixing parameter

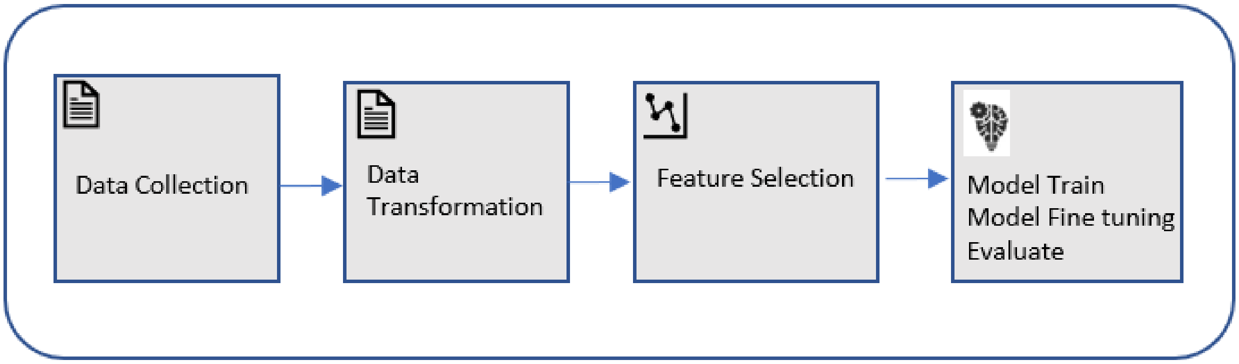
Lasso and ridge are special cases, respectively for $\alpha = 1$ and $\alpha = 0$

The mixing parameter $\alpha$ governs the extent to which the elastic net behaves as a ridge or a lasso. As $\alpha \to 0$, the ridge penalty gains more weight than the lasso; the opposite happens when $\alpha \to 1$. For example, an alpha of 0.5 would provide a 50 percent contribution of each penalty to the loss function. An alpha value of 0 gives all weight to the L2 penalty and a value of 1 gives all weight to the L1 penalty.

In sklearn, the alpha hyperparameter is set via the “l1_ratio” argument that controls the contribution of the L1 and L2 penalties [21]. The lambda ($\lambda$) hyperparameter is set via the “alpha” argument that controls the contribution of the sum of both penalties to the loss function.

## 3.2 Metamodeling steps

Metamodeling involves several key steps such as data collection and preprocessing, feature selection and engineering, data splitting, model selection, training, hyperparameter tuning, evaluation, interpretation, prediction. Balancing complexity and interpretability, addressing overfitting, and iterative refinement are crucial aspects of this process to create accurate and effective models




3.2.1 Data Collection

The data collected for tall wood building is through the Delphin simulation tool [8]. Tall Wood buildings are buildings built with Wooden structures and are 6 storey or higher buildings. The data from 12 cities with two climate time periods such as F0 (Historic) and F7 (Future) is free from incorrect and incomplete data. The data set has 31 years of climate data for each city for various time periods has 15 realizations for each time period and contains 744 data points for both train and test set.

Mean Moisture content (MMC) parameter is the response variable or the Dependent Variable (DV) predicted in this regression analysis. The Independent Variables (IV) / predictors are Moisture Index (MI), Free field Wind Driven Rain (FWDR), Orientation Wind Driven Rain (OWDR), Drying Index (DI), Global solar Radiation (GLOB_RAD), Global solar Radiation normal to wall (GLOB_RADN), Cloudiness index (CLOUD), RAIN, Wind Direction (WDIR), Wind Speed (WSPD), Relative Humidity (RH), Outdoor temperature (TEMP) and outdoor partial vapor pressure (PV).

### 3.2.2 Data Transformation

Basic exploratory data analysis is performed on the dataset and the correlation map reveals independent variables like,

- RAIN, MI, FWDR, OWDR are highly correlated with each other. While Rain and MI are correlated with 0.95 as correlation coefficient, FWDR and OWDR also have high correspondence with a coefficient of 0.91 as illustrated in Figure 1.0
- Relative Humidity and Wind speed-WSPD are highly correlated with a correlation coefficient of 0.78 as well Relative Humidity - RH Moderately correlated with FWDR (0.63) and WDIR (0.63)
- While RH and TEMP are least correlated with each other

The correlation between the independent variables is explained by the heat map shown below in Figure 1.0 generated using sklearn.

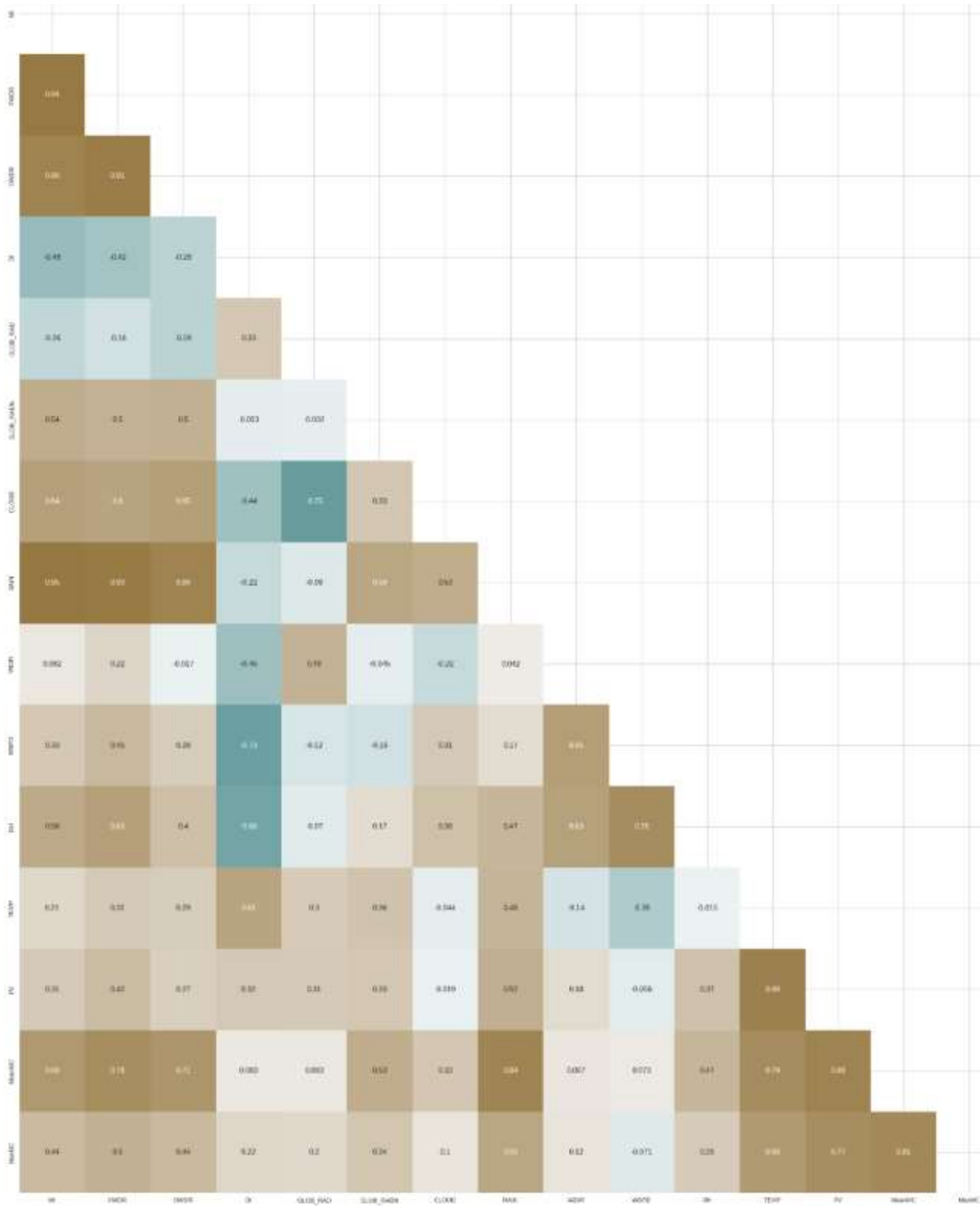


Figure 1.0 Correlation between dependent features using Correlation Heatmap

- Response Variable: Mean MC – Is highly correlated with RAIN with correlation coefficient of 0.84 and PV at 0.86. Also, to note moisture content is moderately correlated with FWDR, OWDR, TEMP and MI as shown in Figure 1.1.

Figure 1.1 Correlation of Response variable Mean MC with Independent features

Exploratory data analysis on the data set reveals the non-linearity between input variables FWDR, OWDR and RAIN against the output Mean MC. As in Figure 1.2 – 1.4, feature transformation for this experiment uses Tukey Ladder of power rule. Features RAIN and Orientation Wind Driven Rain – OWDR are square root transformed and FWDR - Free field Wind Driven Rain is log transformed. The linear relation between the features and the response variable, have significantly improved after transformation as seen from Figure 1.2 – 1.4.

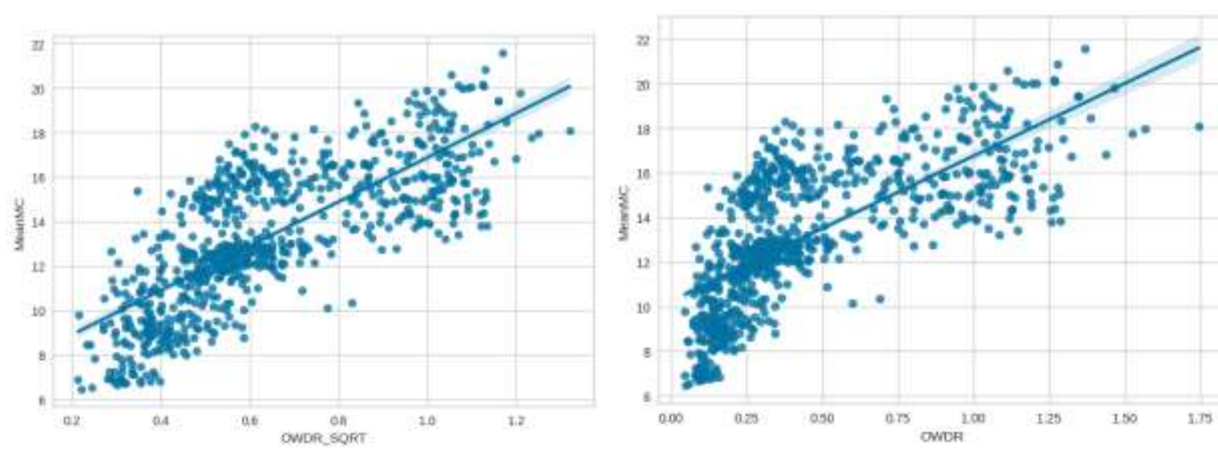



Figure 1.2 Feature OWDR after and before Square root transformation

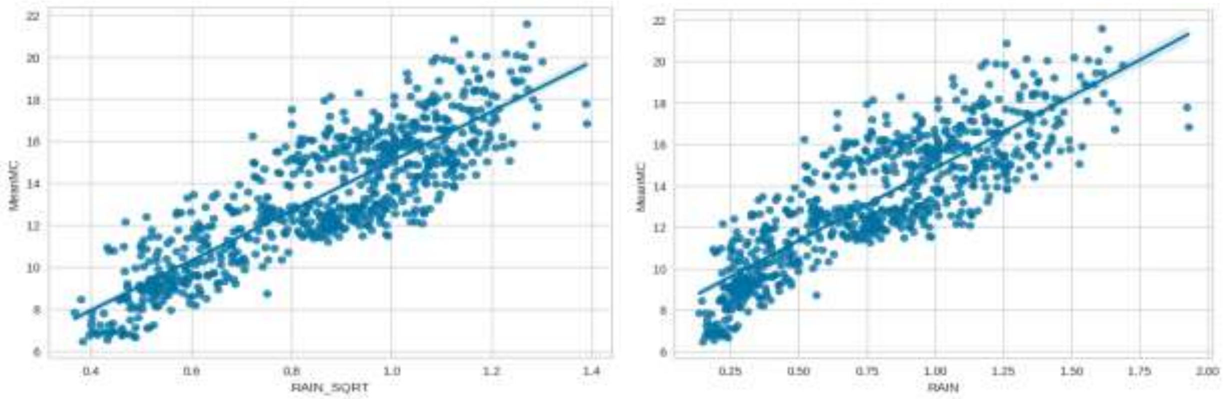



Figure 1.3 Feature RAIN after and before Square root transformation

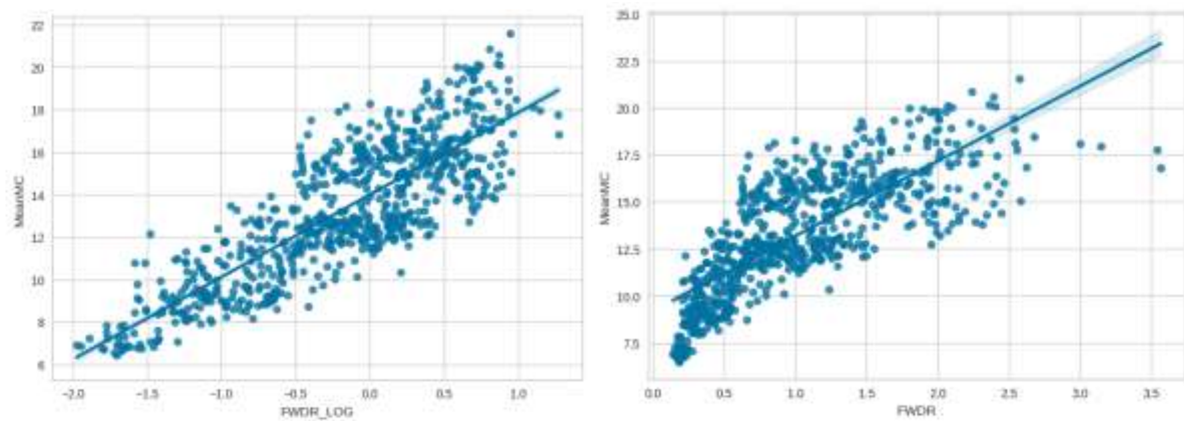
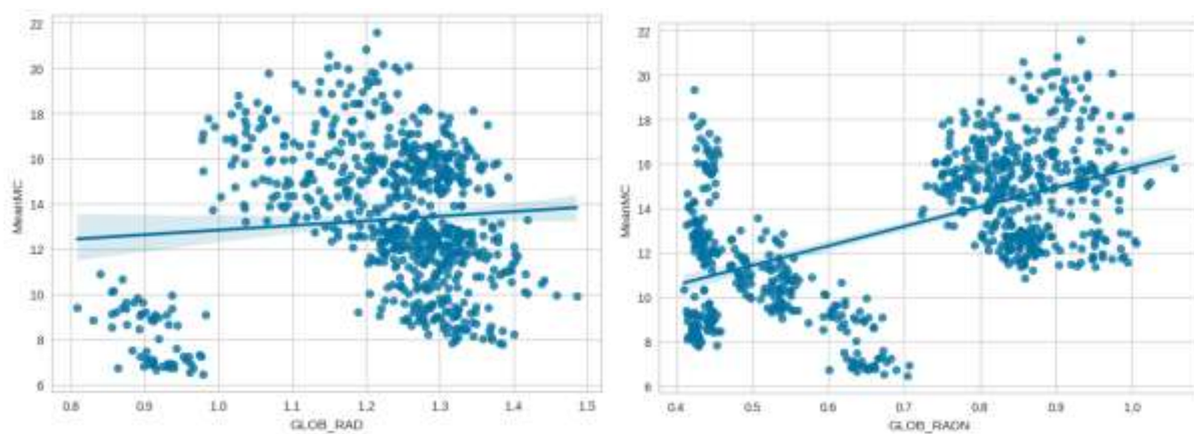



Figure 1.4 Feature FWDR after and before Log transformation

Independent features such as GLOB_RAD, GLOB_RADN and WDIR exhibit requirement to have second order term to correct the non-linearity as shown in Figure 1.5.



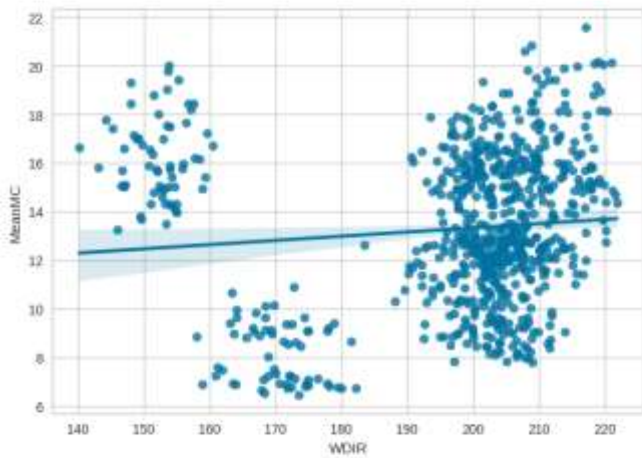



Figure 1.5  Features GLOB_RAD , GLOB_RADN and WDIR exhibiting requirement for polynomial term

After data transformation is done feature scaling is performed. Standard scaler which standardizes the features so they have zero mean and a standard deviation of 1. The distribution of features prior to data scaling subsequently after scaling is depicted in Figure 1.6.

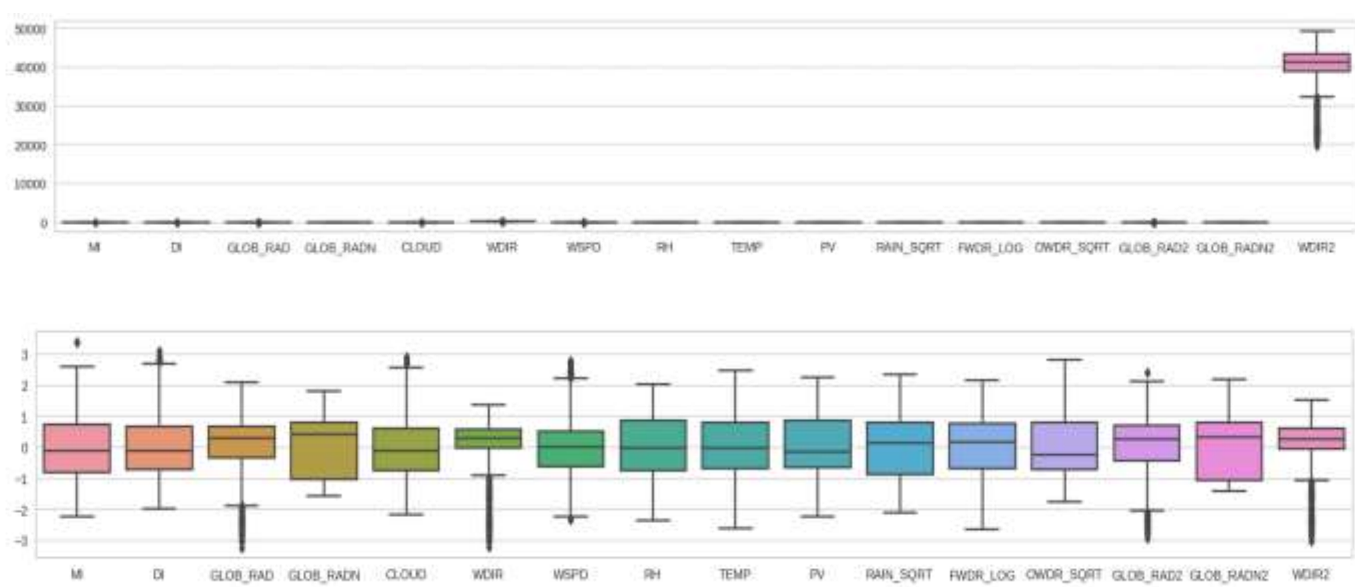



Figure 1.6 Box-and-whisker plot showing the distribution of Independent variables before and after scaling

From preliminary data analysis, the linearity of the data is corrected with the afore mentioned feature transformation techniques and all the linear algorithms are modelled with transformed and scaled data as a pre processing step.

After the data preprocessing, the input data is having 16 features such as RAIN_SQRT, PV, FWDR_LOG, TEMP, OWDR_SQRT, MI, GLOB_RADN_2, GLOB_RADN, RH, CLOUD, WDIR_2, WDIR, GLOB_RAD, DI, WSPD, GLOB_RAD_2.

### 3.2.3 Feature Selection

To analyse the effects of various feature selection techniques on the Multiple linear regression models, Wrapper, Filter and Embedded feature selection methodologies are used for the experiment [13].

The goal of feature selection methods is to select the top k significant features. The range selected for k is k = 1 to n, n being number of features in the dataset. Every feature selection method is iterated n times, n being the number of features in the dataset and for every iteration top n feature selected at every iteration. Then the models are trained with the selected features. For instance, in Univariate feature selection method came with 16 feature sets such as feature set 1 with top 1 feature selected ['RAIN_SQRT'], feature set 2 with top 2 features selected ['RAIN_SQRT', 'PV'] up till top 16 features [['RAIN_SQRT', 'PV', 'FWDR_LOG', 'TEMP', 'OWDR_SQRT', 'MI', 'GLOB_RADN_2', 'GLOB_RADN', 'RH', 'CLOUD', 'WDIR_2', 'WDIR', 'GLOB_RAD', 'DI', 'WSPD', 'GLOB_RAD_2']]. The base regression model is then modelled on the selected features and the model performance is evaluated.

Linear regression is subjected to feature selection methods such as Univariate feature selection, Recursive feature elimination and Lasso for selecting the top k features prior to modelling. F-statistic is used for univariate feature selection to select the top *k* features using SelectKBest algorithm in sklearn [18]. The k value for the selection is hyper tuned and the Linear regression model is trained with possible feature sets.

For Lasso as Feature selection methods that shrinks the coefficients of less important features to zero. The hyper parameter for Lasso is obtained by fine tuning the regularization parameter (λ). Lasso estimator uses the Akaike Information criterion (AIC), the Bayes Information criterion (BIC) to select the optimal value of the regularization parameter alpha implemented in sklearn using LassoLarcIC [19].

Alpha values are also obtained using grid search with 10-fold cross validation using the LARS algorithm [20]. The LARS algorithm starts by including the feature that has the highest correlation with the target variable. It then moves towards the next feature that is most correlated with the current residual while keeping the correlation of the selected features with the residual approximately equal. This stepwise process continues until all features are included or until a predefined threshold is reached. The methodologies used are tabulated in Table 1.0.

Table 1.0 Feature selection techniques experimented for this study

| **S. No** | **Model** | **Feature Selection** | | **Hyper Tuned Parameter** |
|---|---|---|---|---|
| | | | | |

| | | | | |
|---|---|---|---|---|
| 1 | Linear Regression | Filter | Univariate Feature selection | k Parameters |
| | | Wrapper | Recursive Feature Elimination | k Parameters |
| | | Iterative | Lasso | λ selection using AIC Criterion |
| | | | | λ selection using - BIC Criterion |
| | | | | λ selection using - Cross-validated Lasso, using the LARS algorithm |
| | | | | λ selection using - 10-fold CV using Grid Search |
| 2 | Lasso | Embedded | Lasso | λ selection using - AIC Criterion |
| | | | | λ selection using - BIC Criterion |
| | | | | λ selection using - Cross-validated Lasso, using the LARS algorithm |
| | | | | λ selection using - 10-fold CV using Grid Search |
| 3 | Ridge | Embedded | Ridge | λ selection using - 10-fold CV using Grid Search |
| 4 | Elastic Net | Iterative | Lasso and Ridge | L1 ratio and λ selection for Lasso and Ridge |

Recursive feature elimination, a wrapper feature selection technique is also experimented before performing the linear regression.

For regularized Linear models such as Lasso, Ridge and Elastic Net, embedded feature selection method is followed in this study. Lasso automatically selects important features by shrinking the coefficients (λ – regularization parameter) of less important features to zero during model training (equation (2)). Ridge regression also applies regularization to the model. While it doesn't lead to exact feature selection like Lasso, it can still help in reducing the impact of less relevant features by shrinking their coefficients. Elastic Net combines both L1 (Lasso) and L2 (Ridge) regularization. It balances between Lasso's tendency to perform feature selection and Ridge's tendency to keep all features. This can be useful when there are correlated features. The hyper parameters for the Linear models which is the regularization parameter is selected using the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC), LARS algorithm and grid search CV. The alpha value traced for the hyper parameter tuning of λ the regularization parameter is shown in Figure 1.7 for Lasso estimator with various criterions to select λ values.

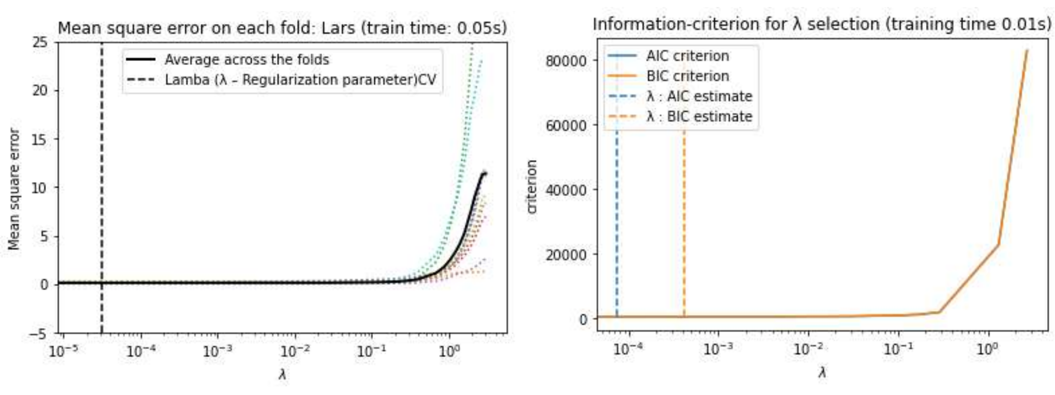



Figure 1.7 Hyper parameter tuning of λ - the regularization parameter using AIC, BIC and Lars techniques

### 3.3 Model Training, Fine Tuning and Evaluation

The transformed and scaled data set is used for feature selection for the respective models. Models are fitted on the training data and the prediction is done on the test set. The fine tuning is done at every iterative stage for the linear models to enhance the performance of prediction. Model evaluation for the linear model is using both visually using the residual plots and quantitatively using Root mean square as the evaluation metric.

*Residual plots:* Prior to evaluating the model a visual inspection is performed on the residuals. Linear regression tries to fit a line that produces the smallest difference between predicted and actual values, where these differences are unbiased as well. This difference or error is also known as residual.

$$Residual\ (e)\ =\ actual\ value\ -\ predicted\ value\ (y\ -\ \hat{y}) \tag{7}$$

*Q-Q plots* are used to check the normality of the residuals.

*Evaluation Metric* - Root Mean Squared Error (RMSE):

RMSE is the square root of the average of the squared difference of the predicted and actual value. In principle, RMSE is the root of the average of squared residuals. We know that residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals are as shown in equation (8)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{8}$$

*Evaluation Metric* - Coefficient of determination ($R^2$):

$R^2$ is a statistical metric used to measure the proportion of the variance in the dependent variable (target) that is explained by the independent variables (predictors) in a regression model. The formula to calculate $R^2$ as shown in equation (9) the Sum of Squares of Residuals (SSR) represents the sum of the squared differences between the observed values and the predicted values from the regression model and the Total Sum of Squares (SST) represents the sum of the squared differences between the observed values and the mean of the observed values. Figure 1.8 gives a representation of the terms used in calculation of $R^2$ .

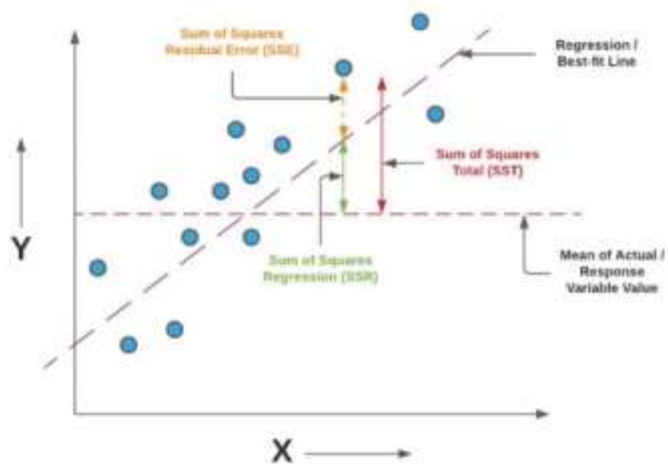



*Figure 1.8 Coefficient of determination (R2)*

$$R^2 = 1 - \frac{Sum\ of\ Squares\ of\ Residuals}{Total\ Sum\ of\ Squares} \tag{9}$$

## 4 Results and Discussion

Model - Linear Regression:

Linear regression is iteratively modeled on the feature sets selected from the feature selection methods discussed in section 2.3. A set of 36 models were analysed with feature selection techniques such as recursive feature elimination technique as a wrapper method, Filter methods like Univariate feature selection and L1 regularization. The graph below in Figure 1.9 shows the performance of Linear regression model, modelled with the features selected using filter and wrapper feature selection methods.

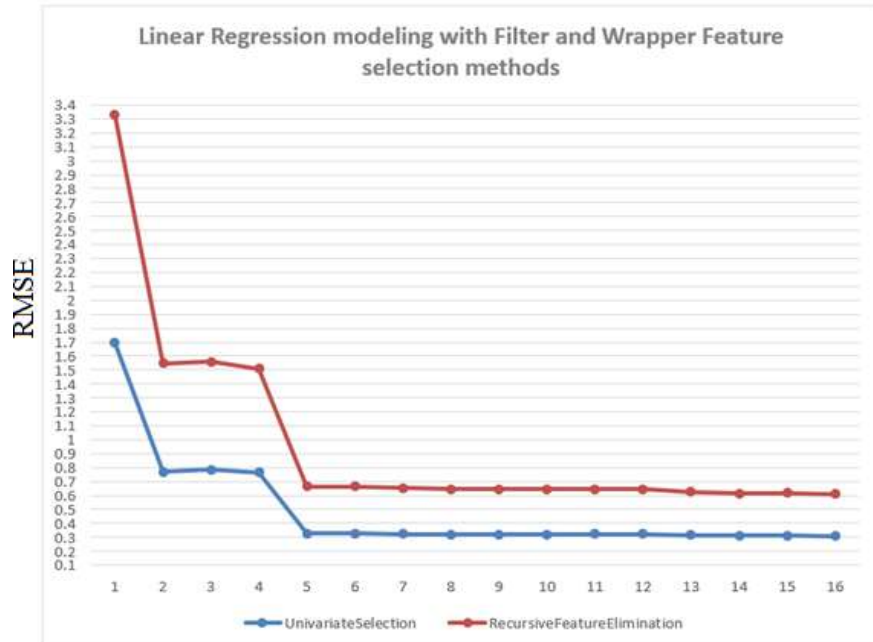



Figure 1.9 - Linear Regression performance with Filter (Univariate) and wrapper (recursive feature selection) feature selection methods. X- axis denotes Number of features and Y-axis denoting RMSE of the Linear Regression modeled with feature sets selected by respective feature selection methods.

The regularization method Lasso is also experimented to select the most important features for the linear regression. The feature selection with Lasso regularization is based on the best lambda (λ) parameter selected by the criterions such as AIC, BIC, Lars and Grid Search CV. The metrics of linear regression model on the subset of features from Lasso regularization is shown in the Table 1.1 below.

Table 1.1 Comparison of Linear model performance with Lasso Regularization.

| | Lasso (λ) - Selection Criteria | | | |
|---|---|---|---|---|
| | **AIC** | **BIC** | **Lars CV** | **Grid Search CV** |
| Linear Regression RMSE metrics | 0.3082 | 0.3082 | 0.3082 | 0.3283 |
| No of features selected | 15 | 14 | 16 | 6 |
| $R^2$ | 0.9904 | 0.9904 | 0.9904 | 0.9890 |

The best performing linear regression model was with RMSE of 0.3073 (in same scale of target - Mean moisture content) with RFE (Recursive Feature Elimination) methodology with a $R^2$ of 99.04% with 14 features selected as most significant features.

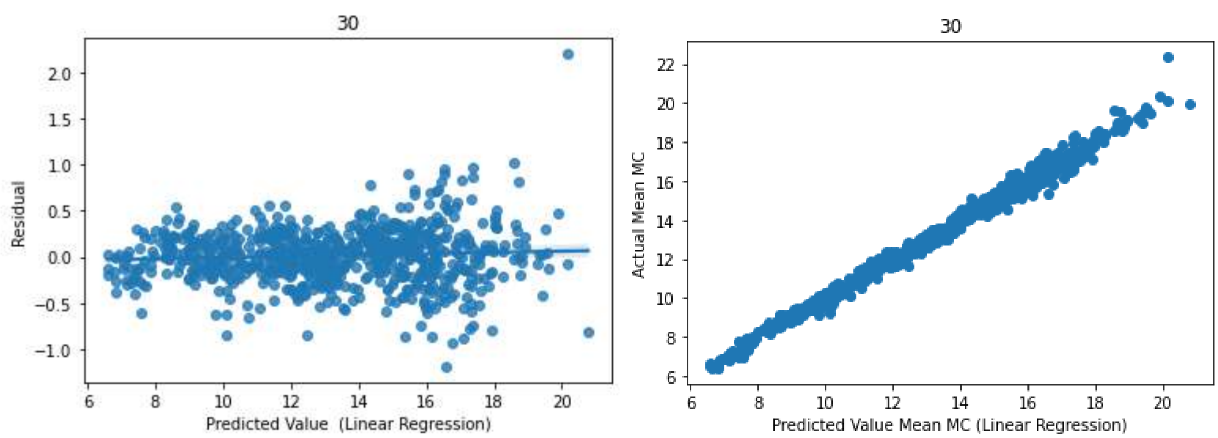



Figure 2.0 (a) - Linear Regression diagnostics: (a) Residual plot for Linear Regression Model (b) Regression plot of values predicted using Linear Regression Model versus actual Mean Moisture

content values. Though the model metrics is fair the residual plot as in Figure 2.0 (a) exhibits Heteroscedasticity that violates the basic assumption of linear regression model.

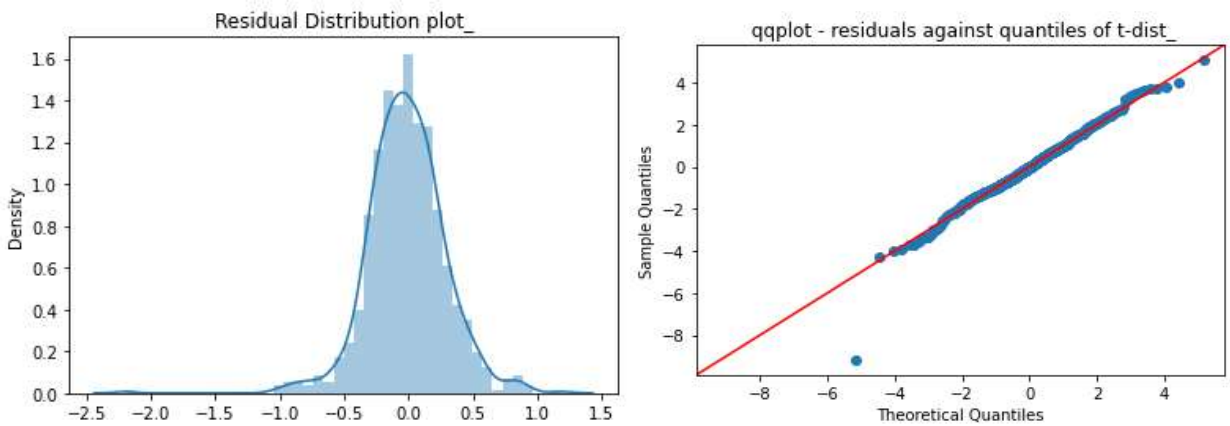



Figure 2.0 (b) - Linear Regression diagnostics: (a) Distribution plot of the residuals (b) Q-Q plot of the residuals

As well the Q-Q plot and distribution plot in Figure 2.0 (b) eliciting the absence of normality in the residuals. The statistical tests, Shapiro-wilk test and D'Agusto with p-values below 0.05 indicates that the residuals do not follow a normal distribution.

Model – Lasso L1 Regularization

The next linear model under study, the Lasso estimator is modeled with different lambda values with Lasso itself as the Feature selection method making it an Embedded Feature selection technique. The lasso estimator performance with different hyper tuned λ regularization parameter is illustrated in the below Table 1.2.

AIC finds the relative informational content of a model by considering the maximum likelihood estimate along with the count of parameters (independent variables) present within the model.

For model comparison using AIC, it's necessary to compute the AIC value for each model. If one model's AIC is at least 2 units lower than another model's, it's considered notably superior to the latter model.

Table 1.2 Comparison of Lasso model performance with Embedded Lasso Regularization. Best model with AIC and BIC. Candidate models with BIC1, AIC1 and AIC2

| | Lamba (λ – Regularization parameter)Value | | | | | | |
|---|---|---|---|---|---|---|---|
| | AIC | BIC | Lars CV | Grid Search CV | BIC1 | AIC1 | AIC2 |
| Lasso RMSE metrics | 0.308554 | 0.309680874 | 0.30847 | 0.329935197 | 0.309681 | 0.308554 | 0.308409 |
| λ Value | 7.47E-05 | 0.000418105 | 3.15E-05 | 0.0154 | 0.000418 | 7.47E-05 | 0 |
| No of features selected | 15 | 14 | 16 | 6 | 14 | 15 | 16 |
| $R^2$ | 0.990345 | 0.990287119 | 0.990345 | 0.9887762 | 0.990287 | 0.990345 | 0.990346 |

The graph in Figure 2.1 denotes the Lasso performance, hyper tuned for various λ values selected using different criterions. The AIC and BIC is used as a criterion to select the best model and the model with minimum AIC and BIC is selected as the best model. The λ that resulted in best model is selected for hyper tuning. The delta AIC measure is followed as in literature [22] to select the candidate model. The delta AIC less than 2, this indicates there is substantial evidence to support the candidate model (i.e., the candidate model is almost as good as the best model).

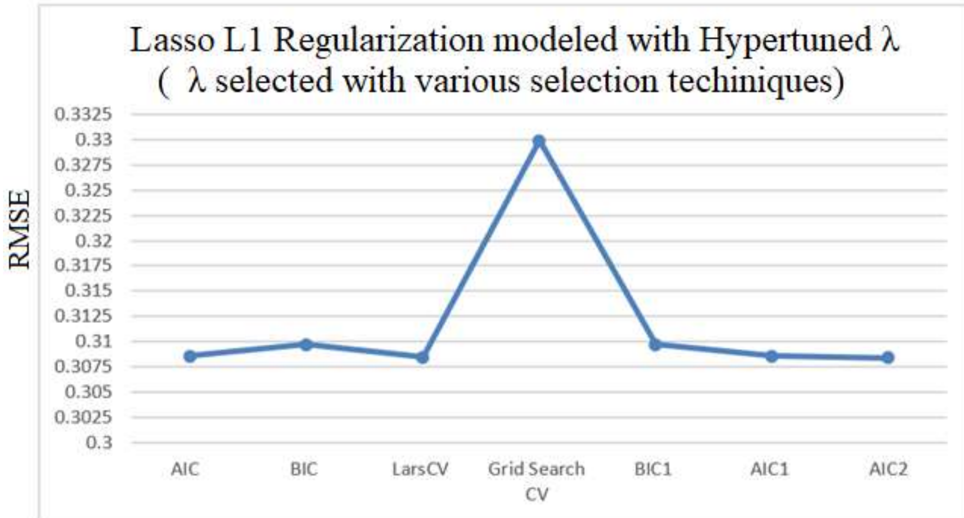



Figure 2.1 - Lasso model performance with different λ values under various λ selection criteria's

The best model with L1 Regularization had a metric RMSE of 0.308 with λ (regularization parameter) = 0 not penalizing any feature surprisingly. Figure 2.3 shows the residual plot of the lasso estimator exhibiting mild heteroscedasticity.

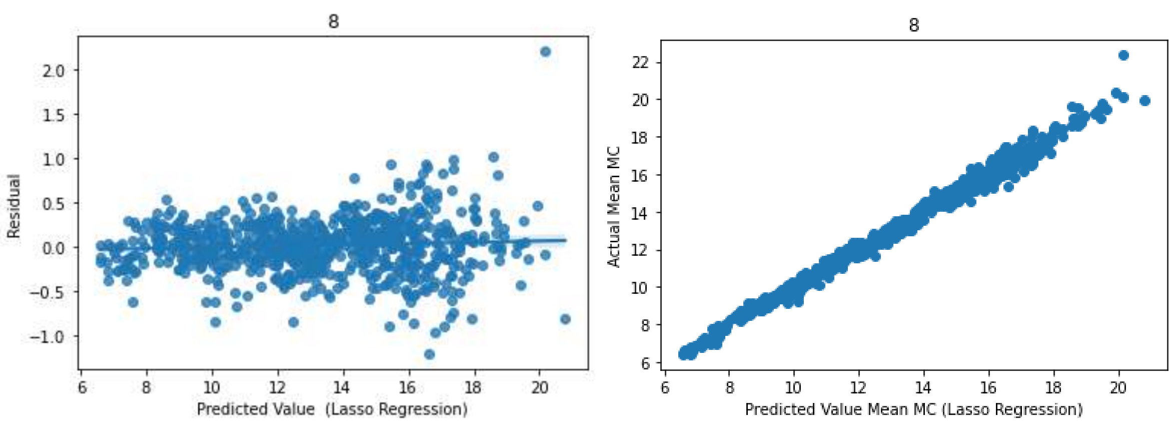



Figure 2.3 - Lasso diagnostics: (a) Residual plot for Lasso L1 Model and (b) Regression plot of values predicted using Lasso Model versus actual Mean Moisture content values.

The parsimonious Lasso model was with 6 features selected which are MI, DI, GLOB_RADN, CLOUD, PV, OWDR_SQRT with λ (regularization parameter) at 0.0154 obtained through Grid Search CV. The RMSE of the parsimonious Lasso with RMSE of 0.3299 and 98% as the coefficient of determination. The Figure 2.4 shows the residual plots with observed variability of the residuals.

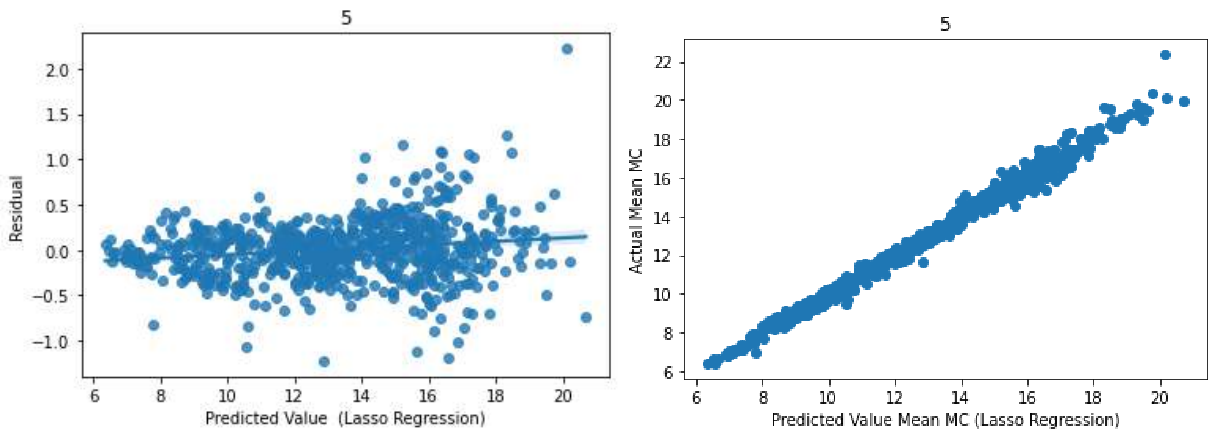



Figure 2.4 – Lasso Parsimonious model diagnostics: (a) Residual plot for Lasso Parsimonious Model and (b) Regression plot of values predicted using Lasso Parsimonious Model versus actual Mean Moisture content values.

Model – Ridge L2 Regularization:

L2 Regularization, Ridge model produced a RMSE 0.3081 with 99% $R^2$ score.

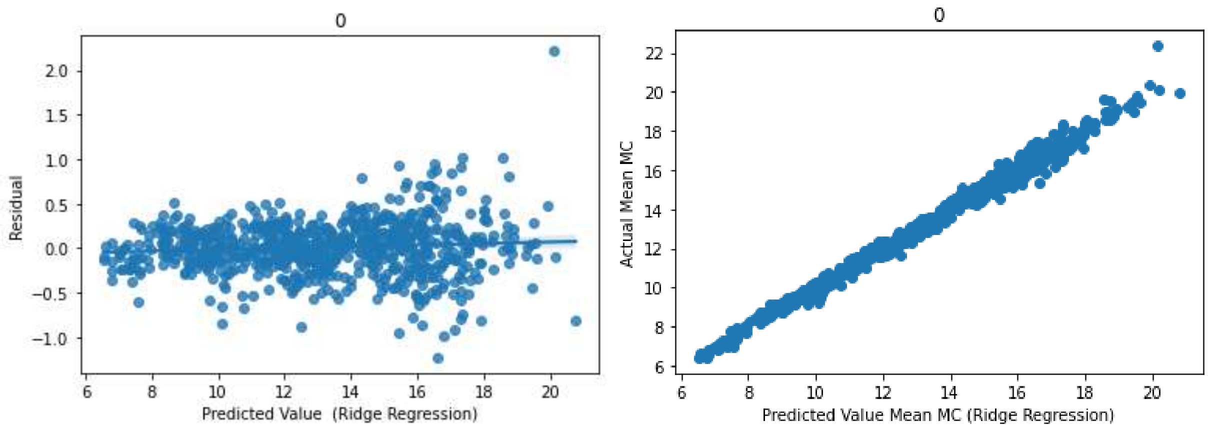



Figure 2.5 – Ridge L2 Regularization diagnostics: (a) Residual plot for Lasso L1 Model and (b) Regression plot of values predicted using Lasso Model versus actual Mean Moisture content values.

Model – Elastic Net L1 and L2 Regularization:

The Multiple regression linear model Elastic net search that utilized both L1 and L2 regularizations fine tuned for the λ (regularization parameter) and l1 ratio had a Root mean square at 0.3083 and 99.03% $R^2$ as shown in the Table 1.2. The residual plots in Figure 2.6 of the Elastic net model also exhibits Heteroscedasticity of residuals.

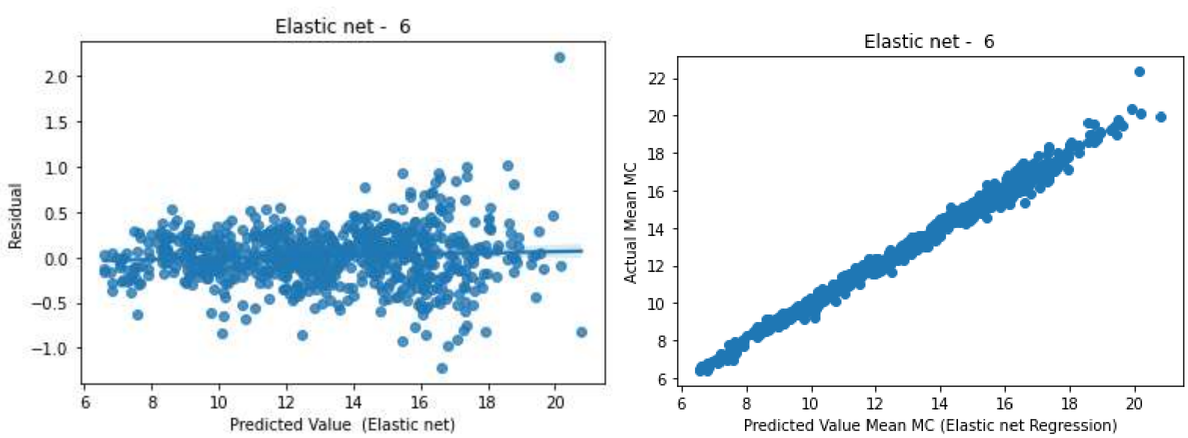


Figure 2.6 - Elastic Net model diagnostics: (a) Residual plot for Elastic net Model and (b) Regression plot of values predicted using Elastic net Model versus actual Mean Moisture content values.

Evaluation metrics shows that the all the multiple linear regression models perform well with a $R^2$ of 99% and the Baseline Linear Regression model performing well with least RMSE value. However as shown in Figures [1.9 – 2.6] the residuals of all the compared linear models exhibit Heteroscedasticity, the residuals lacking constant variance across all levels of the independent variables making the models less reliable for predictions.

Table 1.2 Comparison of Multiple Linear Regression (MLR) models using Evaluation metrics

| MLR | RMSE - 10-Fold CV | RMSE - Hold Out | Feature Selection Method | $R^2$ |
|---|---|---|---|---|
| Linear regression | 0.3073 | 0.3023 | Recursive Feature Elimination - 14 Features | 0.9904 |
| Lasso | 0.3084 | 0.3024 | Embedded Lasso penalization - 16 Features | 0.9903 |
| Ridge | 0.3081 | 0.3024 | Embedded Ridge | 0.9903 |
| Elastic Net | 0.3083 | 0.3023 | L1 and L2 Regularization | 0.9903 |
| Lasso | 0.3299 | 0.3261 | Embedded Lasso penalization - 6 Features * Parsimonious model | 0.9903 |

During my co-op term, my Problem Solving and Creativity competency improved by tackling real-world challenges in the workplace. I enhanced my adaptability by quickly learning new software tools and adjusting to changing project priorities. Additionally, I improved my teamwork by collaborating effectively with cross-functional teams on various projects.

My supervisor displayed strong leadership and decision-making abilities when guiding the team through a critical project phase. Team members exhibited exceptional problem-solving skills when brainstorming solutions to complex issues. This directly helped me to improve my critical thinking competency and learn leadership skills.

My expectations for the co-op work-term were to gain practical experience, learn from professionals in the field, and develop a better understanding of my career goals. I can confidently say that my expectations were exceeded. I had the opportunity to work on a challenging project, receive mentorship, and gain insights into the industry, which greatly contributed to the competency Intellectual Curiosity and Lifelong Learning.

My workplace also contributed to these competencies by encouraging ongoing training and development opportunities. NRC also values diversity and inclusion, which contributes to better communication and teamwork.

Overall, during my co-op experience, I have got clearer idea of the type of work I want to pursue for my full-time employment. I am interested in roles that involve data analysis and application of machine learning because of the skills and interests I've developed during this placement. Additionally, I'd like to work in a company that places a strong emphasis on mentorship and professional development to continue expanding my skill set.

## 5 Conclusion

In summary, in this case study of designing metamodels specifically Multiple Linear models for predicting moisture content brings out the below conclusions,

- The Linear models under study, Linear Regression, Lasso, Ridge and Elastic net revealed similar performance that is evident with the evaluation metrics lacking homoscedasticity
- Feature selection techniques have significantly influenced the modeling which is evident in regularization Models such as Lasso producing Parsimonious models
- Importance of Feature transformation is identified for improving the model performance

As future work we continue to study the following,

- Explore the performance of non-linear models for the dataset to handle the non-linearity of the independent variables against the target variable
- Examine and uncover the hidden patterns or non-obvious relationships within the data that linear models might overlook
- Obtain a generalized parsimonious model with minimal feature set for prediction

## 6 References

[1] IPCC, C. C. (2007). Synthesis report summary for policymakers. An Assessment of the Intergovernmental Panel on Climate Change.

[2] Lacasse, M. A., Gaur, A., & Moore, T. V. (2020). Durability and climate change—Implications for service life prediction and the maintainability of buildings. *Buildings*, *10*(3), 53.

[3] Nik, V. M., Kalagasidis, A. S., & Kjellström, E. (2012). Assessment of hygrothermal performance and mould growth risk in ventilated attics in respect to possible climate changes in Sweden. Building and Environment, 55, 96-109.

[4] Sehizadeh, A., & Ge, H. (2016). Impact of future climates on the durability of typical residential wall assemblies retrofitted to the PassiveHaus for the Eastern Canada region. Building and Environment, 97, 111-125.

[5] Nik, V. M., Kalagasidis, A. S., & Kjellström, E. (2012). Statistical methods for assessing and analyzing the building performance in respect to the future climate. Building and Environment, 53, 107-118.

[6] https://bauklimatik-dresden.de/delphin/index.php?aLa=en

[7] C. Aggarwal, M. Defo, T. Moore, M.A. Lacasse, S. Sahyoun, Validation of three methods of selecting moisture reference years for hygrothermal simulations, in: XV International Conference on Durability of Building Materials and Components, 2020, https://doi.org/10.23967/dbmc.2020.146

[8] Aggarwal, C., Ge, H., Defo, M., & Lacasse, M. A. (2022). Hygrothermal performance assessment of wood frame walls under historical and future climates using partial least squares regression. Building and Environment, 223, 109501.

[9] Astrid Tijskens, Staf Roels, Hans Janssen,Neural networks for metamodeling the hygrothermal behavior of building components, Building and Environment, Volume 162,2019,106282,ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2019.106282

[10] Taffese, W. Z., Sistonen, E., & Puttonen, J. (2015). CaPrM: Carbonation prediction model for reinforced concrete using machine learning methods. Construction and Building Materials, 100, 70-82.

[11] Wang, G. G., & Shan, S. (2006, January). Review of metamodeling techniques in support of engineering design optimization. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 4255, pp. 415-426).

[12] Simpson, T. W., Poplinski, J. D., Koch, P. N., & Allen, J. K. (2001). Metamodels for computer-based engineering design: survey and recommendations. Engineering with computers, 17, 129-150.

[13] A. Kaur, K. Guleria and N. Kumar Trivedi, "Feature Selection in Machine Learning: Methods and Comparison," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2021, pp. 789-795

[14] J. Miao and L. Niu, "A Survey on Feature Selection" Procedia Comput. Sci. Vol 91, no, itqm, pp. 919-926, 2016, doi: 10.1016/j.procs.2016.07.111

[15] Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning (p. 78).

[16] https://onlinelibrary.wiley.com/doi/full/10.1111/gcb.14447

[17] Zou Hui, and Hastie Trevor. 2005. “Regularization and variable selection via the elastic net.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2): 301–320. [Google Scholar]

[18] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[19] Lasso model selection: AIC-BIC / cross-validation — scikit-learn 1.3.0 documentation

[20] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLarsCV.html#sklearn.linear_model.LassoLarsCV

[21] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

[22] The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications. Frank J. Fabozzi, Sergio M. Focardi, Svetlozar T. Rachev and Bala G. Arshanapalli. © 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc. Model Selection Criterion: AIC and BIC, pp: 401 - 402

[23] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K. -R. Müller and G. Montavon, "Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective," in *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 40-58, July 2022, doi: 10.1109/MSP.2022.3153277.