Comparison Between Regression Models to Identify Which Model Gives the Best Performance

Instructor: Dr. Erfanul Hoque

Theoretical Machine Learning DASC-5420

Thompson Rivers University

15th April 2023

Ahsan Mollani

Student ID: T00711408

1. Abstract

Forest fire is one of the major issues for our environment as it affects other important aspects of our life as well such as social, economic, etc. We know that forest fire is a natural phenomenon that can be caused by multiple factors such as wind, temperature, humidity, etc., so it is significant to know the actual cause of forest fire and also to predict the burned area due to forest fires. It is necessary to predict the burned area as it gives accurate predictions that can help in developing strategies for the future and also help in preventing such happenings and help in reducing the damage that may cause in the future.

This report is focusing on predicting the burned area due to forest fires through a machine learning algorithm specifically based on various regression models. This report will be a comparison between five regression models which are Lasso Regression, Ridge Regression, Elastic Net Regression, Random Forest Regression, and Support Vector Regression. Comparison will be done through RMSE (Root Mean Square Error) and MAD (Mean Absolute Deviation) and helps in concluding which regression model performs better for predicting the burned area of the forest fire.

Keywords: **Lasso Regression, Ridge Regression, Elastic Net Regression, Random Forest Regression, Support Vector Regression, RMSE (Root Mean Square Error), and MAD (Mean Absolute Deviation)**

2. Introduction

As we all know, forest fire has become a global concern in our contemporary world, causing a significant effect on our environment, economy, and quality of life due to which it has become a necessity now to identify causes and solutions of it which can be done by conducting the research on it. Through research, we can identify patterns and effects of forest fires which will be helpful for us to take effective measures and work on preventing this problem from happening in the future. There are multiple forest fire datasets which are available on the internet through which we can identify such trends and patterns. I will be working on one such dataset, the Forest Fire dataset obtained from the UCI machine learning repository [1].

The Forest fire dataset contains information on the multiple forest fires which happened in Portugal between 2000 and 2003. This dataset has in total 13 attributes and 517 observations. The main objective of the dataset is to predict the total burned area through forest fires. The dataset includes several predictor variables such as wind, rain, temperature, etc., and one target or response variable, area.

For doing the predictions, there are various methods available through which we can get the optimal result. In this report, we aim to use various regression models to predict the burned area of forest fires. We will compare the performance of different regression models such as Lasso regression, Ridge Regression, Elastic Net, Random Forest, and Support Vector Regression to identify the most accurate model for predicting the burned area.

3. Data

3.1. Data Gathering:

This study uses the Forest Fire dataset, which is available on the UCI machine learning repository website. The dataset includes details about factors involved in forest fires such as wind, humidity, temperature, etc. The dataset has a total of 517 observations, with 12 predictor variables in which there are 2 categorical variables, month and day, and one target variable, area [1].

3.2. Data Exploratory Analysis

The Forest fire dataset contains information on various features that can tell us the factors involved in forest fire burning, and the area which is being burned. Exploratory data analysis (EDA) was performed on this dataset to gain insights into the relationships between the different variables and the target area. The dataset consisted of 517 observations and 13 variables, of which 2 were categorical and 11 were numerical.

For the exploratory data analysis, the histogram of the area which is the response variable was made to see the skewness of the variable. As per the histogram, the response variable, Area was highly skewed and more inclined towards zero (Figure 1).

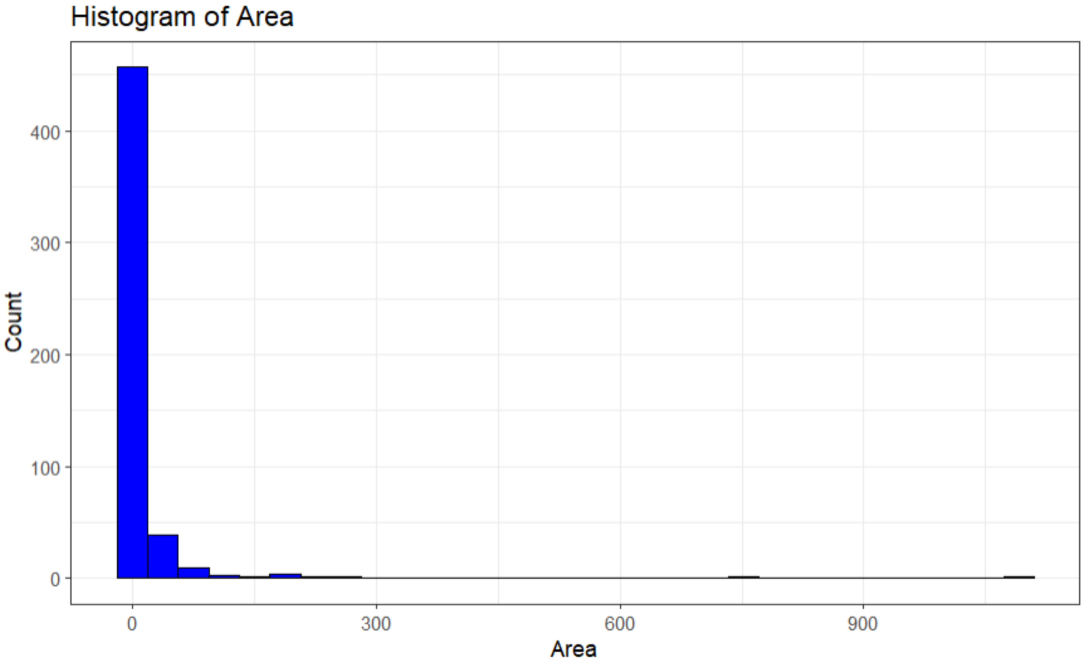

Figure 1. Histogram of response variable “Area”

Additionally, a boxplot of each of the predictor variables was made to see if there are any outliers present in the dataset and to do the comparison. As per the boxplot, variable DC has the highest range of value amongst others containing few outliers as well. Apart from that, most of the variables such as FFMC, X, Y, wind, rain etc. have a very narrow range with multiple outliers in them which tells us that we need to standardize and we also need to the outliers from the dataset.

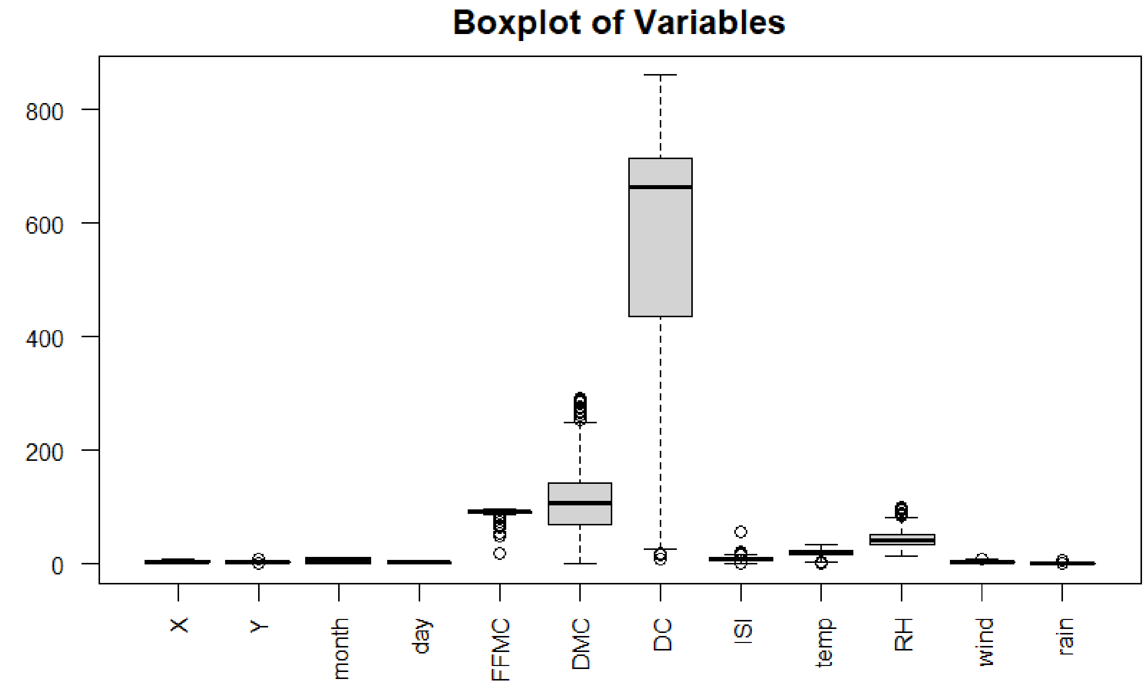

Figure 2. Boxplot of all the variables except the response variable.

Overall, the exploratory data analysis conducted provided meaningful insights into the dataset and highlighted some significant variables that could be useful to forecast the extent of the burned area resulting from forest fires.

3.3. Data Cleaning and Preprocessing

The forest fire dataset required a major part of cleaning and preprocessing as it contained a lot of outliers. The dataset contained 517 observations and 12 variables which had variables with high discrepancies among each other also due to the high skewness of the response variable we need to do the log transformation of it. Therefore, the cleaning and preprocessing step aimed to transform, standardize, and remove unnecessary variables, and outliers for accurate results.

For cleaning and preprocessing of the dataset, firstly missing values were checked and it resulted in zero missing values. Then the transformation of the response variable (area) was done to improve its values in it. The logarithm function, $y = \ln(y + 1)$ was used to improve the regression result for the rightly skewed response variable [2].

Furthermore, we also removed the outliers from the dataset by using the interquartile range (IQR) method which is known as the common approach to dealing with the outliers. After that, we removed unnecessary variables such as month and day. These two variables were the categorical variables and including them would cause the model to not perform that well that’s why these two variables were removed. At last, scaling of the predictor variable was done by making the mean of variables zero and the standard deviation one to get a better prediction.

4. Methods

There were five methods for the prediction of burned area in forest fire and to see which method is the best. The following methods are as follows:

4.1. Lasso Regression:

Lasso regression is a linear regression method used for feature selection and regularization. It involves minimizing the sum of the squared errors while constraining the sum of the absolute values of the coefficients to be less than a specified value. This formula of the lasso regression can be expressed as [3].

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \;= \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Figure 3. The formula of Lasso.

The formula of the Lasso regression includes the L1 Penalty which brings the coefficients towards zero and some of them are exactly to zero as well by shrinking the coefficients and it also includes the residual sum of squares. So through this, it combines the residual sum of square and L1 penalty [3].

The lambda is a very important parameter in the lasso regression as it determines the extent of the penalty applied to the coefficients of the model. Then cross-validation is used to determine the optimal value of lambda which results in reducing the root mean square value once the prediction is done.

4.2. Ridge Regression:

Similar to lasso regression, ridge regression is also a linear regression method for feature selection and regularization. It involves minimizing the sum of the squared errors while constraining the sum of the absolute values of the coefficients to be less than a specified value. This objective function of the ridge regression can be expressed as [3].

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

Figure 4. The formula of Ridge.

The formula of the Ridge regression includes the L2 Penalty which brings the coefficients near to zero by shrinking the coefficients and it also includes the residual sum of squares. So through this, it combines the residual sum of square and L2 penalty [3].

The lambda is a very important parameter in ridge regression as it determines the extent of the penalty applied to the coefficients of the model. Then cross-validation is used to determine the optimal value of lambda which results in reducing the root mean square value once the prediction is done.

4.3. Elastic Net Regression:

As we know that both lasso and ridge regression uses penalty terms such as L1 and L2 respectively. Elastic net uses the combination of both the L1 and L2 regularization methods. The objective function of the elastic net is as follows:

$$\sum_{i=1}^{n}(y_i-\beta_0\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\left[(1-\alpha)\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right]$$

Figure 5. The formula of Elastic Net.

As we can see in Figure 5 the elastic net uses both the lasso and ridge combination which means it not only brings the coefficients near to zero by using the L2 penalty but it also brings some of the coefficients exactly to zero through the L1 penalty.

The same procedure is also followed for elastic net where the lambda is a very important parameter and then cross-validation is used to determine the optimal value of lambda which results in reducing the root mean square value once the prediction is done.

4.4. Random Forest:

Random forest is a supervised machine learning algorithm that is used in both regression and classification models. The random forest model makes the subset of the training data points and then the features are selected to make the decision trees of each of them. Each decision tree generates an output and then it generates the average result [4].

4.5. Support Vector Regression (SVR):

Support vector regression is another supervised learning algorithm that is used for predictions. Through support vector regression we find the best-fitted line. The best-fitted line is the hyperplane which has the maximum number of points [5]. In SVR, we also find the best cost and epsilon number for which we need to assign multiple numbers to cost and epsilon, and then through the tuning function, we find the best cost and epsilon number. Once it is found, we then fit the model with that number and we perform SVR.

The Github link for the project is here.

5. Results

To evaluate the performance of the models, we used two metrics: Root Mean Square Error (RMSE) and Mean Absolute Deviation (MAD). The R-squared measures the square root of the average of the squared differences between the predicted and actual values, while the average absolute difference between the predicted and actual values. We could have also used the R-square for evaluating the performance of the model but since it’s a very difficult regression task so using R square won’t be optimal to use [1].

For Lasso, Ridge, and Elastic Net, cross-validation was used to find the optimal lambda on the training data set, after getting the optimal lambda the process of fitting the model into the respective regression is done, once this process this done the prediction of the model is done by fitting the model with test data and at the end, we find the RMSE and MAD of the respective models. Similarly, the random forest model is run on the training dataset, and then the further prediction of the model is done on test data to get the result. For performing the support vector regression, it is necessary to find the optimal cost and epsilon value and this can be done through the tuning function. Once this process is done, then the optimal value of cost and epsilon is fitted in the SVR model. After that, prediction is done on the test data to get the result.

Figure 6 and Figure 7 show the performance of five models. The Random Forest performs the best in terms of both Root Mean Squared Error (RMSE) and Mean Absolute Deviation (MAD), with an RMSE of 2.162 and a MAD of 3.482, followed by the SVR model which has an RMSE of 3.554 and a MAD of 2.28. The Elastic Net and Lasso models have similar performance, with slightly higher MSE and RMSE values than the Random Forest and SVR models. The Ridge Regression model is better than the elastic net and lasso but third place if look at the model performance.

Overall, the Random Forest model seems to be the best choice among the five algorithms because it has the smallest prediction error among others.

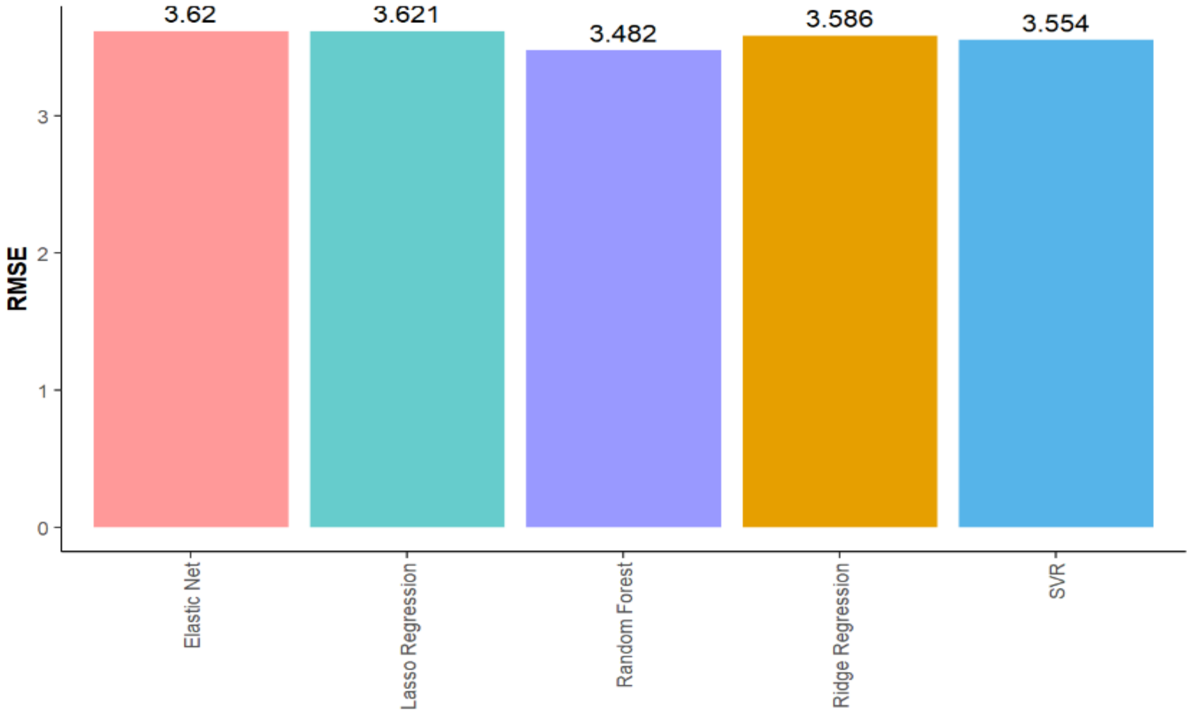

Figure 6. Comparison of models through RMSE value

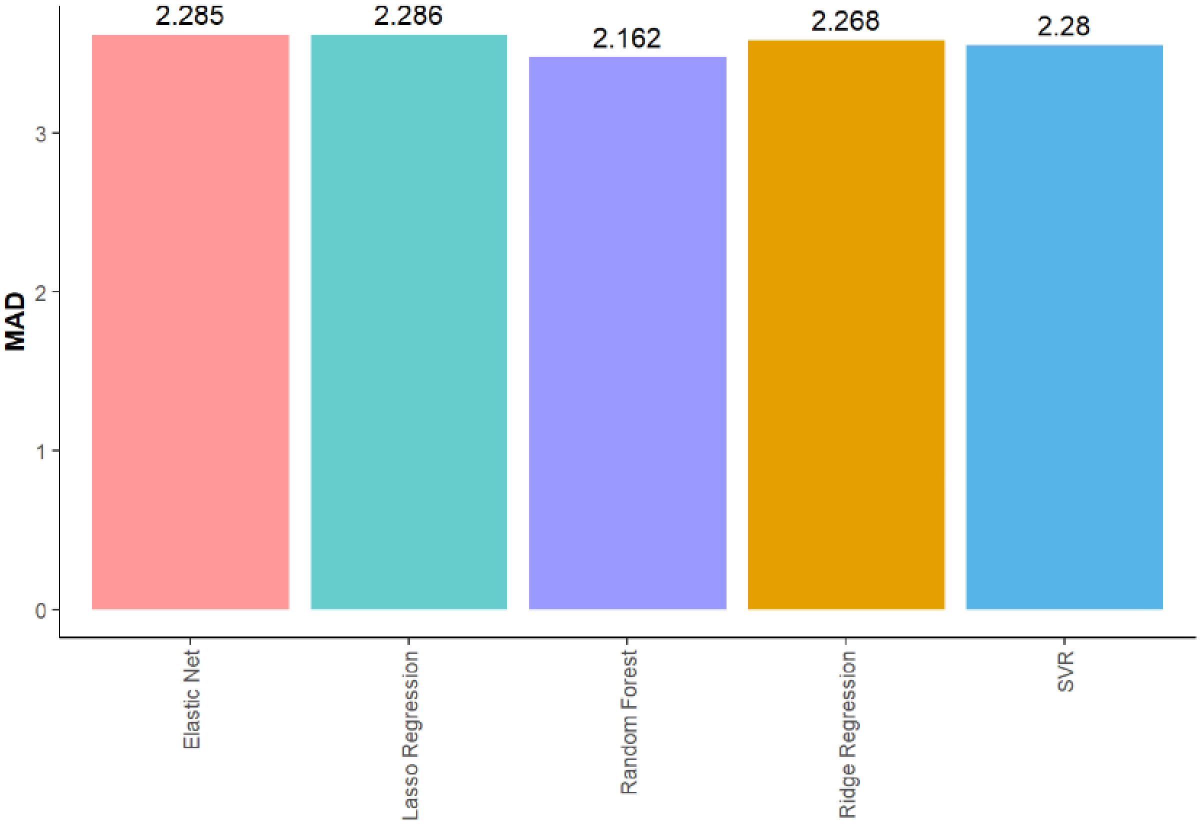

Figure 7. Comparison of models through MAD value

6. Conclusion

In conclusion, this study aimed to compare which model performs well for predicting the area burned due to forest fire. Comparison between five regression algorithms which were Lasso, Ridge, Elastic Net, Random Forest, and Support Vector Regression (SVR) on the dataset and evaluated the performance of the models using RMSE and MAD metrics. The results showed that the random forest performs better amongst the others having the lowest RMSE and MAD value followed SVR. The result provided by this comparative analysis may help people in selecting which model performs well in predicting the burned area because of the forest fire.

7. References

[1] "UCI Machine Learning Repository: Forest Fires Data Set." https://archive.ics.uci.edu/ml/datasets/forest+fires

[2] P. Cortez and A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," ResearchGate, Jan. 2007, [Online]. Available: https://www.researchgate.net/publication/238767143_A_Data_Mining_Approach_to_Predict_Forest_Fires_using_Meteorological_Data

[3] "An Introduction to Statistical Learning," An Introduction to Statistical Learning. https://www.statlearning.com/

[4] S. E. R, "Understand Random Forest Algorithms With Examples (Updated 2023)," Analytics Vidhya, Mar. 2023, [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[5] A. Raj, "Unlocking the True Power of Support Vector Regression," Medium, Dec. 16, 2021. [Online]. Available: https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.