**STAT515 FINAL PROJECT REPORT**

Regression analysis of forest fires

PRESENTED BY: Jaswanth Erusu | Thanmayee Akkineni | Monali Sheetal Kathepaga

SUBJECT KNOWLEDGE: STATISTICS 515

PROGRAMMING LANGUAGE: R STUDIO

DATASET: QUANTITATIVE CSV FILE

**ABSTRACT:**

The goal of this challenging regression challenge is to forecast the burned area of forest fires in Portugal's northeast region using meteorological and other data. Forest fires are a serious environmental problem because they harm the economy, the ecology, and put people's lives in danger. Rapid detection is essential for managing this problem. Using automated tools based on nearby sensors, such as those offered by meteorological stations, is one option for achieving this. In fact, various fire indices, such the forest Fire Weather Index (FWI), employ meteorological factors (such as temperature and wind) to impact forest fires. It is possible to foretell the burned area of smaller, more frequent fires. The management of firefighting resources, such as the prioritization of objectives for air tankers and ground troops, can be improved with the help of such knowledge.

**Introduction:**

Forest fires, often known as wildfires, are a significant environmental issue because they threaten the preservation of forests, harm the environment and economy, and hurt people. Millions of hectares (ha) of forest are lost each year throughout the world because of this phenomenon, which has a variety of causes (including human error and lightning). Particularly, forest fires have a significant negative impact on Portugal. Over 2.7 million acres of forest were lost between 1980 and 2005, which is the same amount of land as Albania. With 4.6% and 3.1% of the territory affected by fire in 2003 and 2005, respectively, there were 21 and 18 fatalities.

For a firefighting operation to be effective, quick detection is essential. The development of automatic solutions has received a lot of attention because conventional human surveillance is expensive and influenced by subjective factors. These can be categorized into three main groups [1]:

local sensors, satellite-based, and infrared/smoke scanners. Satellites are expensive to acquire, have delays in localization, and their resolution is not always sufficient.

**Dataset:**

The dataset of forest fires contains 13 rows and 518 columns.: Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), latitude, longitude, month, day, temperature, RH, Wind, rain, area. While the DMC and DC stand for the moisture content of shallow and deep organic layers, respectively, which affects fire intensity, the FFMC represents the moisture content of surface litter and effects ignition and fire spread. The FWI index, which measures fire intensity. High numbers indicate more severe burning conditions, even though distinct scales are utilized for each of the FWI aspects.

**A screenshot of a computer

Description automatically generated with medium confidence**

1.X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2.Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9  
3. month - month of the year: 'jan' to 'dec'  
4. day - day of the week: 'mon' to 'sun'  
5. FFMC - FFMC index from the FWI system: 18.7 to 96.20  
6. DMC - DMC index from the FWI system: 1.1 to 291.3  
7. DC - DC index from the FWI system: 7.9 to 860.6  
8. ISI - ISI index from the FWI system: 0.0 to 56.10  
9. temp - temperature in Celsius degrees: 2.2 to 33.30  
10. RH - relative humidity in %: 15.0 to 100  
11. wind - wind speed in km/h: 0.40 to 9.40  
12. rain - outside rain in mm/m2: 0.0 to 6.4  
13. area - the burned area of the forest (in ha): 0.00 to 1090.84.

**EXPLORATORY DATA ANALYSIS**

The exploratory analysis of data is performed to integrate, analyze, and research data to offer summaries learned from the dataset's features. This is best communicated through data visualization models. By using EDA, we may learn about the dataset's structure, quality, variability, and co-variability. We can obtain crucial data features by conducting exploratory data analysis.

Chart, box and whisker chart

Description automatically generated

BOX PLOT FOR MONTH VS TEMPERATURE

**Chart, box and whisker chart

Description automatically generated**

BOX PLOT FOR MONTH VS WIND

**Table

Description automatically generated with low confidence**

POINT PLOT FOR MONTH VS RAIN

**Regression Analysis Models:**

For this project regression analysis, we divided our dataset into another part which we call as “mm” which has the 6 variables of dataset which has only numerical values in it to perform the following regressions we used in this project:

* Linear regression
* Multiple regression
* Ridge regression
* Lasso regression
* Elastic net regression
* Ridge vs Linear vs Lasso vs Elastic net

In the end we compared all the regression models to get a result of which model is the best fit.

**Linear regression:**

A regression model that depicts the association between variables using a straight line. By looking for the value of the regression coefficient(s) that minimizes the overall model error, it locates the line that fits your data the best. The distribution of the dependent variable must be normal for each value of the independent variable. For all possible independent variable values, the variance of the distribution of the dependent variable should remain constant. All observations should be independent, and there should be a linear relationship between the dependent variable and each independent variable.

Graphical user interface, diagram

Description automatically generated

**For data set rr**

Chart, scatter chart

Description automatically generated

**For data set mm**

Chart, scatter chart

Description automatically generated

Correlation between the variables

**Multiple Regression:**

Multiple regression is a statistical method that can be used to assess a single dependent variable and several independent variables. Multiple regression analysis makes use of independent variables whose values are known to predict the value of the single dependent value. To predict the value of one variable based on the information of another, we can apply the simple linear regression function. A linear regression can only use two continuous variables: an independent variable and a dependent variable. The independent variable is the element that is used to calculate the dependent variable or outcome. The explanatory variables included in a multivariate regression model are numerous.

A picture containing diagram

Description automatically generated

**Ridge Regression:**

It reduces coefficients to non-zero values while keeping all variables to prevent overfitting. It determines the ideal value of lambda, which causes coefficients to decrease as lambda grows, and tries to accomplish this while keeping all the model's variables in place. We can comprehend fractioned deviation if we utilize the x-variable as the dev. We can determine the order of most significant to least significant variables by plotting the variables' relative relevance.

Chart

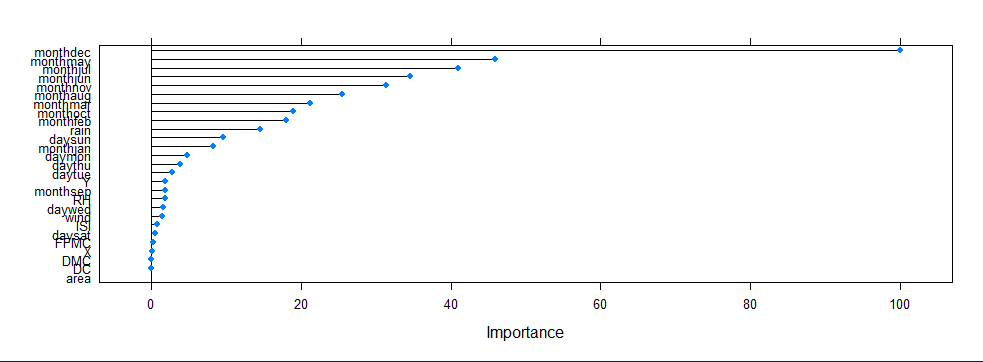
Description automatically generated

Log Lambda vs coefficients

Chart, diagram

Description automatically generated

Fraction Deviance vs coefficients



Importance vs variables

Chart, scatter chart

Description automatically generated

Ridge vs LinearModel

**Lasso regression:**

Reduces regression coefficients, some of which are reduced to zero, which aids in feature selection. Lasso regression chooses one variable and disregards others if there is a collection of closely linked variables that are generating multicollinearity. When you wish to automate specific steps in the model selection process, such as variable selection and parameter elimination, this sort of regression is ideally suited.

Chart, line chart

Description automatically generatedChart, line chart

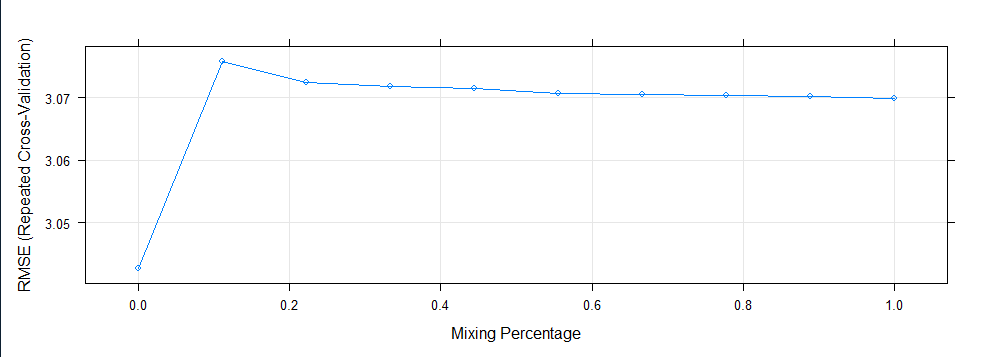
Description automatically generated

**Text

Description automatically generated**

**Elastic Net Regression:**

Regression models are regularized by elastic net linear regression using the penalties from the lasso and ridge procedures. By considering their drawbacks to enhance the regularization of statistical models, the strategy combines the lasso and ridge regression methods. There are 204 variables and 0 coefficients when log lambda equals 8. When the x-variable is specified as "DEV," a fraction deviance explained plot is generated. Root means square error (RMSE) is the standard deviation of the residuals (prediction errors).

Chart

Description automatically generatedChart, diagram

Description automatically generated

We can adjust the alpha parameter in elastic nets so that alpha = 0 represents ridge regression and alpha = 1 represents lasso regression. Like this, the penalty function reduces to the L1 (ridge) regularization at alpha = 0 and to the L2 (lasso) regularization at alpha = 1. To optimize the Elastic Net, we can select an alpha value between 0 and 1, which will cause some coefficients to shrink and others to be set to 0 for sparse selection. The lambda hyper-parameter in elastic net regression substantially and primarily depends on the alpha hyper-parameter.

**RIDGE VS LASSO VS ELASTIC NET VS LINEAR MODEL:**

The minimum, first quartile, median, third quartile, and maximum mean absolute errors give us a summary for each of our models. The mean RMSE, which is 3.042648, is lowest for the elastic net, ridge, and lasso. Look at r squared, which quantifies the proportion of the response variable's fluctuation that can be accounted for by the model's independent variables. To determine which model is significantly better than others, we made a box plot.

A picture containing diagram

Description automatically generatedA picture containing calendar

Description automatically generated

**FUTURE WORK:**

What we have done in this Final Term Project is a very basic analysis and model predictions. Since this data is huge, a deep cleanup and filtering out processes needs to be implemented to retrieve only important characteristics. That would help better prediction and model fitting.

**CONCLUSION**

Finally, we conclude that all four models we attempted to match our data against- Linear regression, Multiple regression, Ridge regression, Lasso regression, and Elastic net regression- The Lasso Regression model may be the best match for our training data. As we all know, once a model has been trained using the training set, it must be tested by making predictions against the test set. Lasso Regression has an accuracy of 88% in this case.

References:

1. B. Arrue, A. Ollero, and J. Matinez de Dios. An Intelligent System for False Alarm Reduction in Infrared Forest-Fire Detection. IEEE Intelligent Systems, 15(3):64–73, 2000.

2. J. Bi and K. Bennett. Regression Error Characteristic curves. In Proceedings of 20th International Conference on Machine Learning (ICML), pages 43–50, Washington DC, USA, 2003.

3. L. Breiman. Random Forests. Machine Learning, 45(1):5–32, 2001.

4. L. Breiman, J. Friedman, R. Ohlsen, and C. Stone. Classification and Regression Trees. Wadsworth, Monterey, CA, 1984.

5. R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Quebec,

For dataset:

<https://archive.ics.uci.edu/ml/datasets/Forest+Fires>