*Forest Fire analysis using Machine Learning Algorithms after hyperparameter tuning.*



# Abstract:

For the analysis of the Forest Fires dataset, we used Machine learning algorithms such as Random Forest, Decision Tree, ANN and Linear Regression. We used the hyper-parameter tuning methods such as *GridSearchCV()* using Cross-Validation and tuned for the model for K =10. The time taken for the model before and after the model is noted. We found that, after designing the model using hyperparameters, the Root Mean Square Error decreases from 0.737 to 0.627, but the time is taken to run the model increased from 4.004 secs for a simple model and 653.036 secs for a tuned model.

# Introduction

The Forest Fires is a common natural disaster in most of the countries. They cause huge damage to flora and fauna along with human lives. It burns the animals and plants on the way, also causes extension of rare species to extend from the world. From the past few centuries, about half a million plant and animal species already extinct from the world (Heywood & Stuart, 1992). And a million species are in danger of extinction. Forest fires affect the natural functionally, climate, eco-system, diversity and natural life of plants, animals, and humans. Though human activities play a significant role in the natural calamities, it, in turn, affects the Ecological imbalance (Malamud et al., 1998). The Forest fires generally occur during summers based on geographical location on the country. Though we cannot stop the natural disasters, if the area covered by the forest fire is detected, the damage caused due to it can be reduced. This saves the human, animal and most of the rare species algae and plants. This also helps us to find the areas where counter fires need to be focused, to stop the spread of the fire (Yang et al., 2020). This also reduces the financial and economic loss.

In this report, we analyse the Forest fires dataset from the UCI repository. This data set has 13 attributes and 517 observations. The independent's variables such as, the day and month on which the fire occurred, the parameters collected from FWI systems such as FFMC, DMC, DC and ISI, and the environmental parameters such as temperature, humidity, rainfall and the wind speed are considered. The dependent variable, the affected area due to fire is given in hectares. It is a continuous variable ranging from 0.0 to 1090.84, for additional description of the dataset, please refer to Appendix Table 2.

Our main goal here is to find the area burn due to forest fire, as the dependent variable is continuous, it is a supervised learning and Regression problem. To perform the regression analysis, we considered the Machine learning models such as Random Forest, Decision Tree, ANN, and Linear Regression. We used Label encoding and One-Hot encoding in the data preparation stage. For tuning the hyperparameter for Random Forest we used the *GridSearchCV()* method.

# AI Techniques:

**Random Forest:**

Random Forest is used our Data Analysis, considering following reasons:

1. Random Forest is based on tree based structure similar to Decision Tree, but selects the best possible tree with good performance.
2. This model is unbiased with the missing values in the data
3. The effect of outliers is minimal compared to other models (Ao et al., 2019).
4. Random Forest in build suggests the important features, which helps in optimization.

In our dataset, we have used the techniques such as Label encoding, One-Hot encoding, and Grid Search Cross-validation for the better performance of the model. As some of the input variables are categorical, the Label encoding is used to simplify the data by converting the labels to numerical values that are readable by the machines. As the machines cannot directly handle the categorical data, the One-Hot encoding helps in converting the digits into 0's and 1's format. Both Label encoding and One-Hot encoding improve the performance of the system when dealing with large datasets.

The *GridSearchCV()* gives helps in tuning the model with suggested hyperparameters. The method *best\_params()* is used to get the model specified parameter results. To cross-check the results, Cross-Validation is used along with the *GridSearchCV()* method.

# Results:

The Random Forest gave good results initially using Label and One-Hot encoding on raw data. In this regression analysis, we used MSE and RMSE to define the performance of the model. The low of MSE and RMSE represents that the model gives good results, if these values are higher, that means the model is not well fit with the training data and give a high error rate.

The results of the Random Forest model before and after performance turning is shown in below Table 2

Table 1: The performance comparison table

|  |  |  |
| --- | --- | --- |
| **Performance Metrix** | **Random Forest** | |
| **simple model** | **Tunned model** |
| **MSE** | 0.544 | 0.525 |
| **RMSE** | 0.737 | 0.626 |
| **Time is taken to train (in secs)** | 4.004 | 653.036 |

From the above table, we can observe that the simple model took approximately 4secs to train the model, whereas the tuned model took 653 secs to train the model. The cross-validation implemented to tune the parameters could be one of the reasons. The tuned model performed well with 0.62 RMSE than the simple model using the below hyperparameters shown in table 2.

Table 2: The Tuned parameters for Random Forest

|  |  |
| --- | --- |
| **Parameters** | **values** |
| max\_depth | 20 |
| max\_leaf\_nodes | 2 |
| min\_samples\_leaf | 5 |
| min\_samples\_split | 5 |

# References:

Heywood, V. H., and S. N. Stuart. "Species extinctions in tropical forests." *Tropical deforestation and species extinction* (1992): 91-117.

Malamud, Bruce D., Gleb Morein, and Donald L. Turcotte. "Forest fires: an example of self-organized critical behaviour." *Science* 281, no. 5384 (1998): 1840-1842.

Yang, Rongrong, Zhirong Wang, Juncheng Jiang, Shuoxun Shen, Peipei Sun, and Yawei Lu. "Cause analysis and prevention measures of fire and explosion caused by sulfur corrosion." *Engineering Failure Analysis* 108 (2020): 104342.

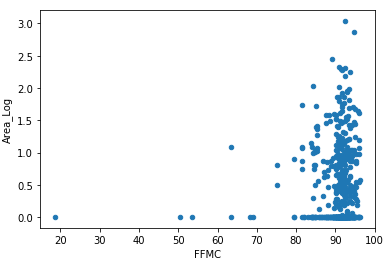
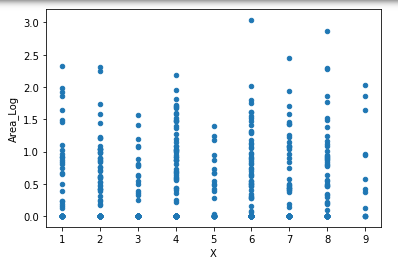
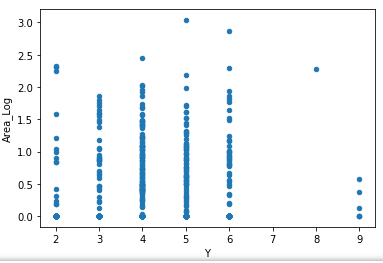
Ao, Yile, Hongqi Li, Liping Zhu, Sikandar Ali, and Zhongguo Yang. "The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling." *Journal of Petroleum Science and Engineering* 174 (2019): 776-789.

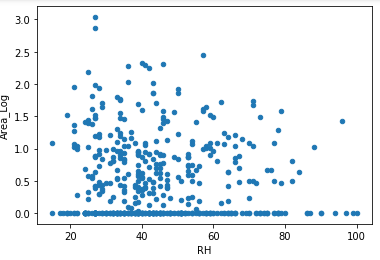
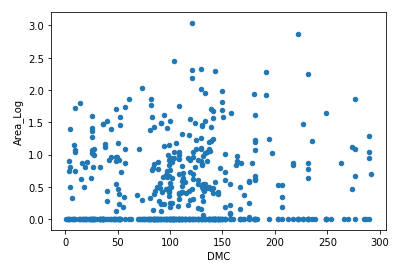
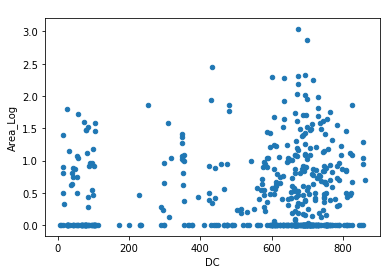
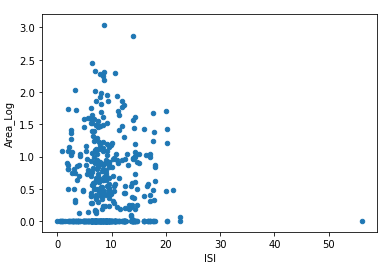
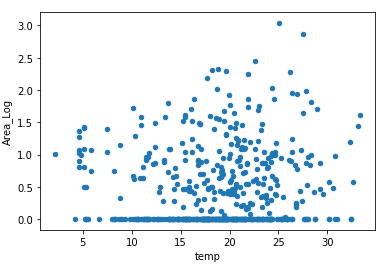
# Appendix:

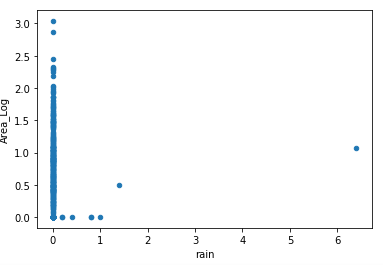
*Table 3: The dataset description of attributes*

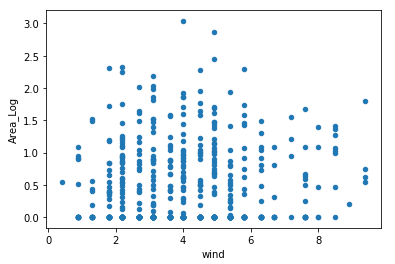
|  |  |  |  |
| --- | --- | --- | --- |
| **S.no** | **attributes** | **Description** | **range of values** |
| 1 | X | x |  |
| 2 | Y | y |  |
| 3 | month | Months in a Year | Jan to Dec |
| 4 | day | Days in a week | Mon to Sun |
| 5 | FFMC | FFMC index from the FWI system | 18.7 to 96.20 |
| 6 | DMC | 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 measured in Celsius degrees | 2.2 to 33.30 |
| 10 | RH | Relative humidity measured in % | 15.0 to 100 |
| 11 | wind | Speed of wind in km/h | 0.40 to 9.40 |
| 12 | rain | Rain in mm/m2 | 0.0 to 6.4 |
| 13 | area | Burned area of the forest (in ha) | 0.00 to 1090.84 |

The below shown images are the scatter plots between all the variables and the Area\_log

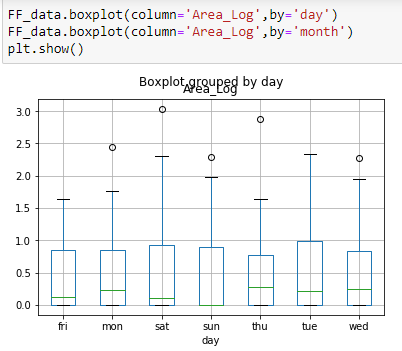
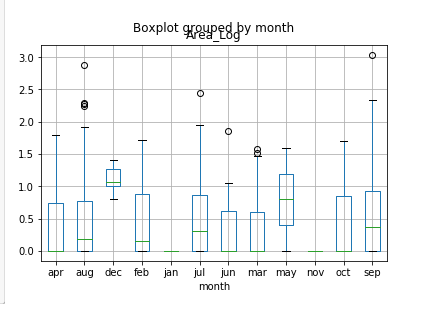








The Box plot for Area vs month and day are shown below.

The below shown is the density plot

