Detecting Fire Area

Walid Ashraf

Abstract

Wildfire hazarads have been discussed in many research. This work utilized the Forest Fires Data Set [1] that contains fire weather indices (FWI) and weather observations. The research aimed to find months with most fire, find which parameters affect area size, and predict fire area. The author utilized graphical and regression techniques to discover answers to his goals. Results show that august and September holds the most fires, the best prediction method is Random Forest tree with five features.

Motivation

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. In this work, we explore a Data Mining (DM) approach to predict the burned area of forest fires.

The research utilizes different parameters to detect the burned area in order to be able to plan the needed firefighting forces in advance.

Dataset(s)

This research utilizes the firefighting dataset available in [1].

Attribute Information:

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 (skewed to zero)

Data Preparation and Cleaning

The area result data was very skewed to the zero, because many days passes without fires, and most fires are small.

In the research we used the logarea = log(area+1), the (+1) is to avoid zero values.

In the preprocessing phase, the work required feature minimalization because (X, Y, day) was not relevant attributes and all utilized values was normalized using standard normalizer.

Finally, at the regression phase, research **showed** that the Temprature, Relative Humedity, DMC, Wind, and DC has the most effect on the resulat value.

Research Question(s)

1 - What months have the most fires?

2 - What are the most correlated parameters with Fire Area?

3 - What is the best regression method for fire area detection and what are its parameters?

Methods

The **1st question** was answered using a histogram between the month and fire occurances.

In detecting the the burned area (**2nd question**), I have used regression algorithms. Examples are:

1. Linear Regression
2. Support Vector Machine
3. Random Forest Regression

The **3rd question** was found using GridSearch method is sklean that tries the same algorithm with different parameters and recommends the best one.

Findings

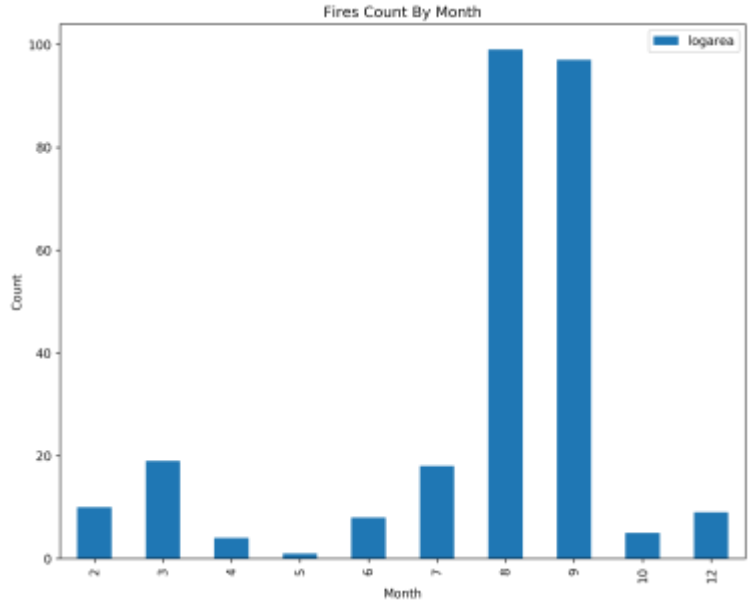
<Feel free to replicate this slide to show multiple findings>

Present your findings. Include at least one visualization in your presentation (feel free to include more). The visualization should be honest, accessible, and elegant for a general audience.

You need not come to a definitive conclusion, but you need to say how your findings relate back to your research question.

Fire By Month

- Figure Shows that August and September are have the most incidents of Fires.
- The Graph is missing 1, and 11 (Jan. and Nov.) because they are zero.

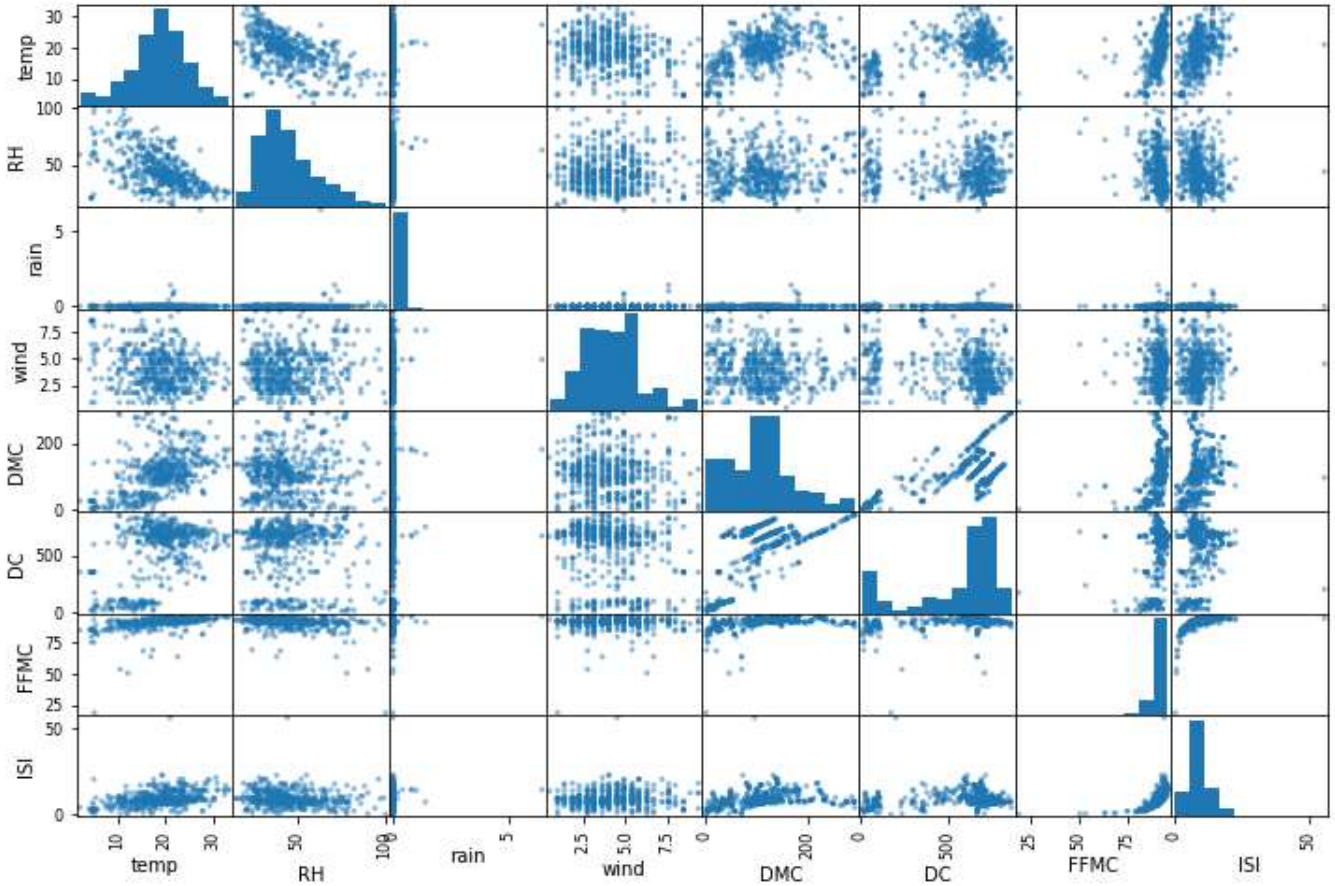


Correlation Value with Log(area+1)

- As shown in the table below, the most correlated with burned area are DMC, Wind, DC, Relative Humidity and temp.

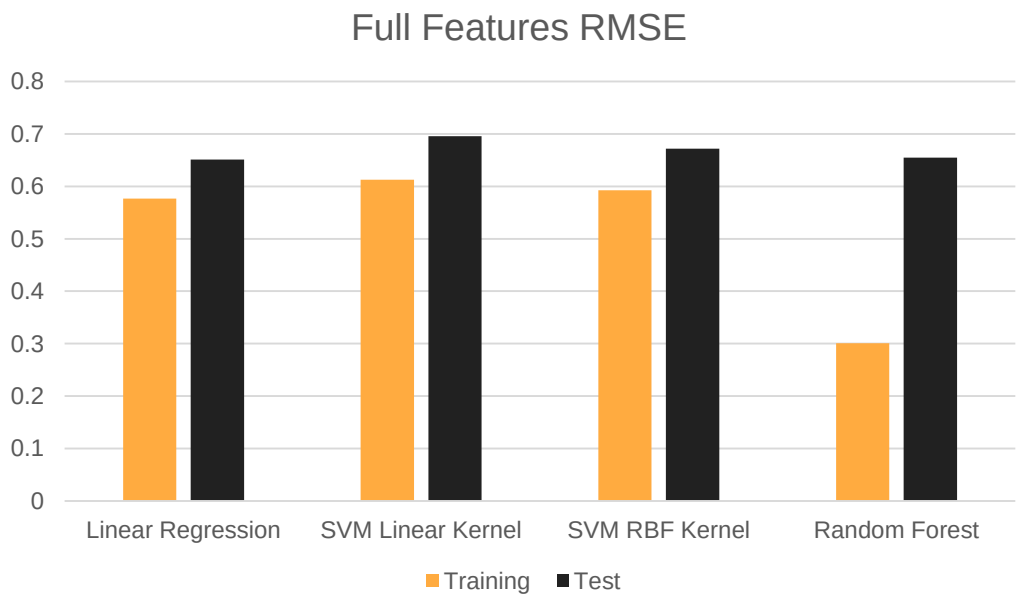
Parameter	Absolute Correlation
DMC	0.067153
Wind	0.066973
DC	0.066360
RH	0.053662
temp	0.053487
FFMC	0.046799
rain	0.023311
ISI	0.010347

Scatter Matrix



- Show Correlation between different parameters.
- Self Correlation (diagonal is subsisted with histogram)

Testing With Full Features



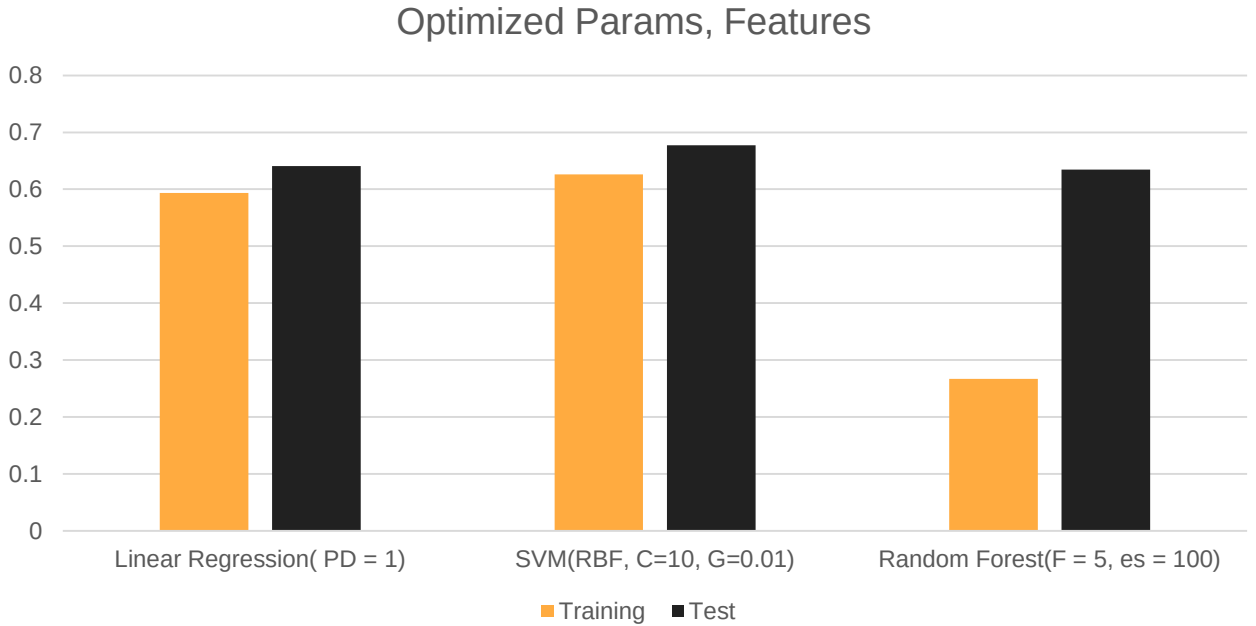
Optimizing Features

- Utilizing Random Forest Algorithm with different options using Sklearn Grid Search we are able to find the best number of features (n=5) and their effects.
- The Months (Categorical Values) was converted using Sklearn OneHotEncoder due to an assumed importance of the month, which resulted to be not.
- Finally, Other algorithms parameters are optimized using these features and Grid search for RMSE minimization

Feature	Importance
temp	0.202854
RH	0.161055
DMC	0.129180
wind	0.117419
DC	0.112222
ISI	0.105068
FFMC	0.102111
1	0.013301
2	0.010915
11	0.010380
4	0.006987
7	0.006847
6	0.004919
12	0.004685
8	0.004603
10	0.003325
rain	0.003144
9	0.000825
5	0.000143
3	0.000017

Final Results (Optimized Parameters and Features)

- With Feature Minimization all Linear Regression and SVM performs less in the training set and better in Test Set while Random Forest Performs better in both.



Limitations

The work contained two main limitations.

1. The readings was very low on some months (ex. November has 1 reading, may had two).
2. The best algorithm have larger than expected RMSE

Conclusions

The results showed that August and September holds the most fires. The results showed that DMC, wind, DC, RH and temp are the most correlated ones with the fire size. Moreover, the best test RMSE for a prediction algorithm was achieved by Random Forest tree with value 0.63. The algorithm was trained with the best scoring features to be Temprature, Relative Humdity,DMC,Wind, and DC. Nevertheless, all algorithms predicted predicted better withe this attibuted minimation scheme.

Acknowledgements

I would like to thank my friends who have helped me greatly while learning this course, and thank Aurélien Géron the author of [2] (the best technical machine learning book ever).

References

[1] https://archive.ics.uci.edu/ml/datasets/Forest+Fires

[2] http://shop.oreilly.com/product/0636920052289.do