STAT 6950 Final Project

Fernando Anorve Lopez
Christina Sousa

April 20, 2019

1 Introduction

Forest fires are broadly define as "a fire which tends to spread freely and can burn all living and natural components within the forest" [2]. They can be costly to communities and ecological systems, for instance, "physical, chemical, mineralogical, and biological soil properties can be affected by forest fires [4]. In Portugal, the region from which the data set originated, there are 15,000 to 25,000 forest per year on average. These burn anywhere from about 150,000 hectares (ha) to 250,000 ha annually. [3]

In this report we will investigate data are from $n = 517$ forest fires occuring in the Montesinho National Park in northeast Portugal between January 2000 and December 2003. They come to us by way of Paulo Cortez and Anibal Morais at the University of Minho, in Guimaraes, Portugal [1]. The main purpose of this data is to predict the behavior of the burn area of a fire using weather-related predictors. Predicting the area of a forest fire might be useful to optimize resources to extinguish a fire.

1.1 Hypotheses

Our goal is to use 12 predictors to model the total burned area of large forest fires within the park. This could be done using a multiple linear regression model. Additionally, we would like to identify which values of the predictors yield fires with burn area less than $100m^2$, and which ones lead to larger fires. This could be done using logistic regression. We hypothesize that these approaches will yield better predictive power than the naive model.

2 Data

At each fire occurrence, a fire inspector recorded the date and and time of the fire, as well as the spatial location of the fire within the park boundaries. They also recorded the type of vegetation, weather conditions, total burn area, and six components of the Fire Weather Index (FWI), a Canadian system for rating fire danger. These included Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI).

Although the individual elements are not measured on the same scale, higher values indicate greater danger of fire in all cases.

The researchers did not include BUI in the data, because it was strongly collinear with the other predictors.

2.1 Predictor Variables

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**: index from the FWI system: 18.7 to 96.20
6. **DMC**: index from the FWI system: 1.1 to 291.3
7. **DC**: index from the FWI system: 7.9 to 860.6
8. **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

2.2 Response Variable

area: the burned area of the forest (in ha): 0.00 to 1090.84

2.3 Limitations

The burn area of forest fires is a highly variable response that depends on a myriad of factors, from the type of vegetation in the area, to meteorological conditions, to type of terrain and the methods of containment. For this reason, this problem is viewed in existing literature as a challenging prediction problem, and is often used to put various modern machine learning methods to the test. [**1**][**2**][**5**] Hence we should view any models arising from this analysis as mere steps towards improving the prospects of using on-line meteorological data to assist in fire preparedness. It is unreasonable to expect an exceedingly accurate predictive model at this time, given the complex nature of the problem and the yet-developing computational techniques for analyzing this kind of data.

Fires that burn less than $100m^2$ of area are all catalogued as zero. This poses a missing data problem, in which we do not have a continuous response for observations with burn areas less than $100m^2$.

In their paper ”Estimation of the Burned Area in Forest Fires Using Computational Intelligence Techniques,” Ozbayoglu and Bozer suggest a five-category burn area scale for classifying fires [**2**]:

- *Very Small*: 0-1.32 ha
- *Small*: 1.32-63.86 ha
- *Medium*: 63.86-273.05 ha
- *Big*: 273.05-773.17 ha
- *Very Big*: 773.17-5661 ha

We found that in the Montesinho Forest Fires training data set, even splitting on 63.86ha alone left only 13 records in the ”¿63.86ha” category. Hence most of the fires in this data set would be classified as ”Small”, and there is not much information left to try to predict ”Large” (¿63.86ha) versus ”Small” (¡63.86ha). For this reason we decided to split the data on ”Zero” ($< 100m^2$) versus ”Nonzero” ($> 100m^2$) burn area, since this yielded a more even split of the data.

3 Exploratory Data Analysis

Now that we are familiar with the data and limitations, let’s have a look at some visualizations to get an idea of the relationships between the various predictors and **area**.

First, we view the continuous predictors. Note the discretization which has occurred in the Larea plots for the zero burn area records. Also note that there do not appear to be any prominent predictors of Larea here.

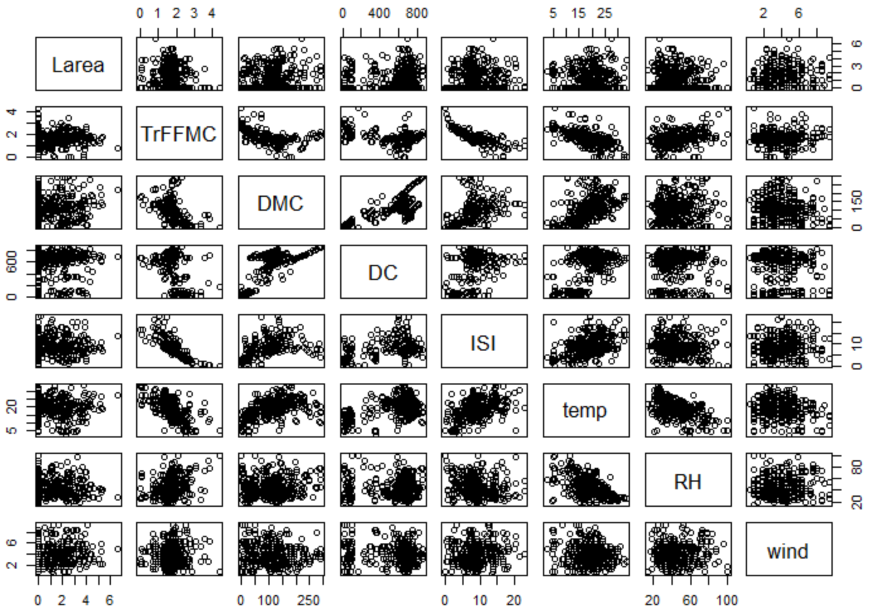

Figure 1: Scatterplot Matrix of the Transformed Continous Predictor Variables versus Log(Area+1)

Let's have a look at the categorical predictors. we note that December and May appear to stand out as having higher log(area) values, as well as **no rain**.

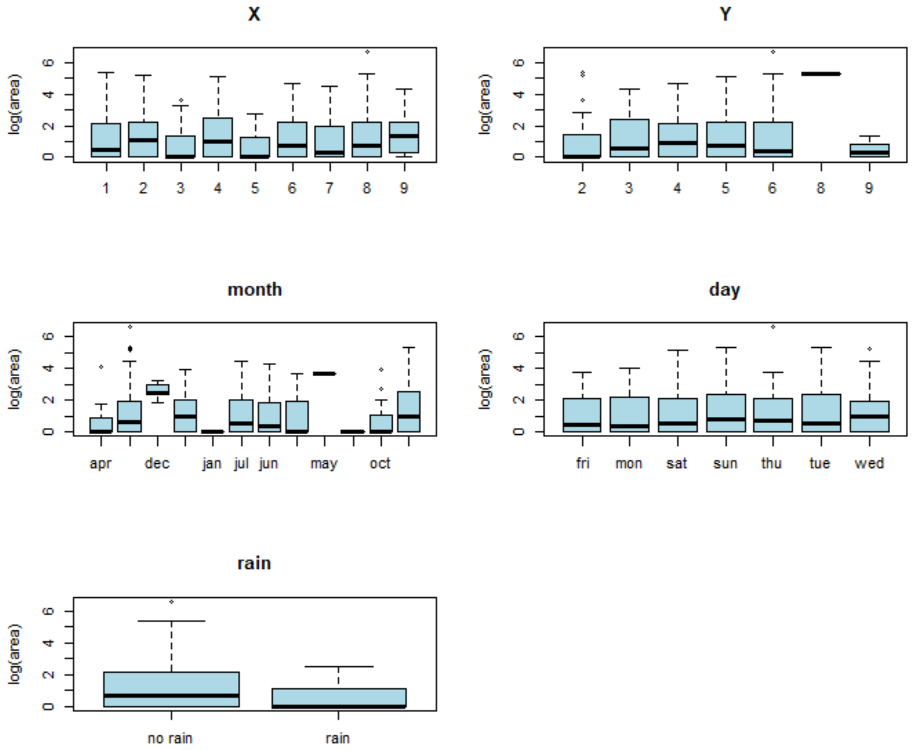

Figure 2: Categorical Predictors versus Log(Area+1

Since we are interested in classifying zero-burn area fires versus nonzero burn-area fires, we plot histograms for each of the continuous predictors, conditioned on zero or nonzero burn, to see if there is any separation in the distributions for these.

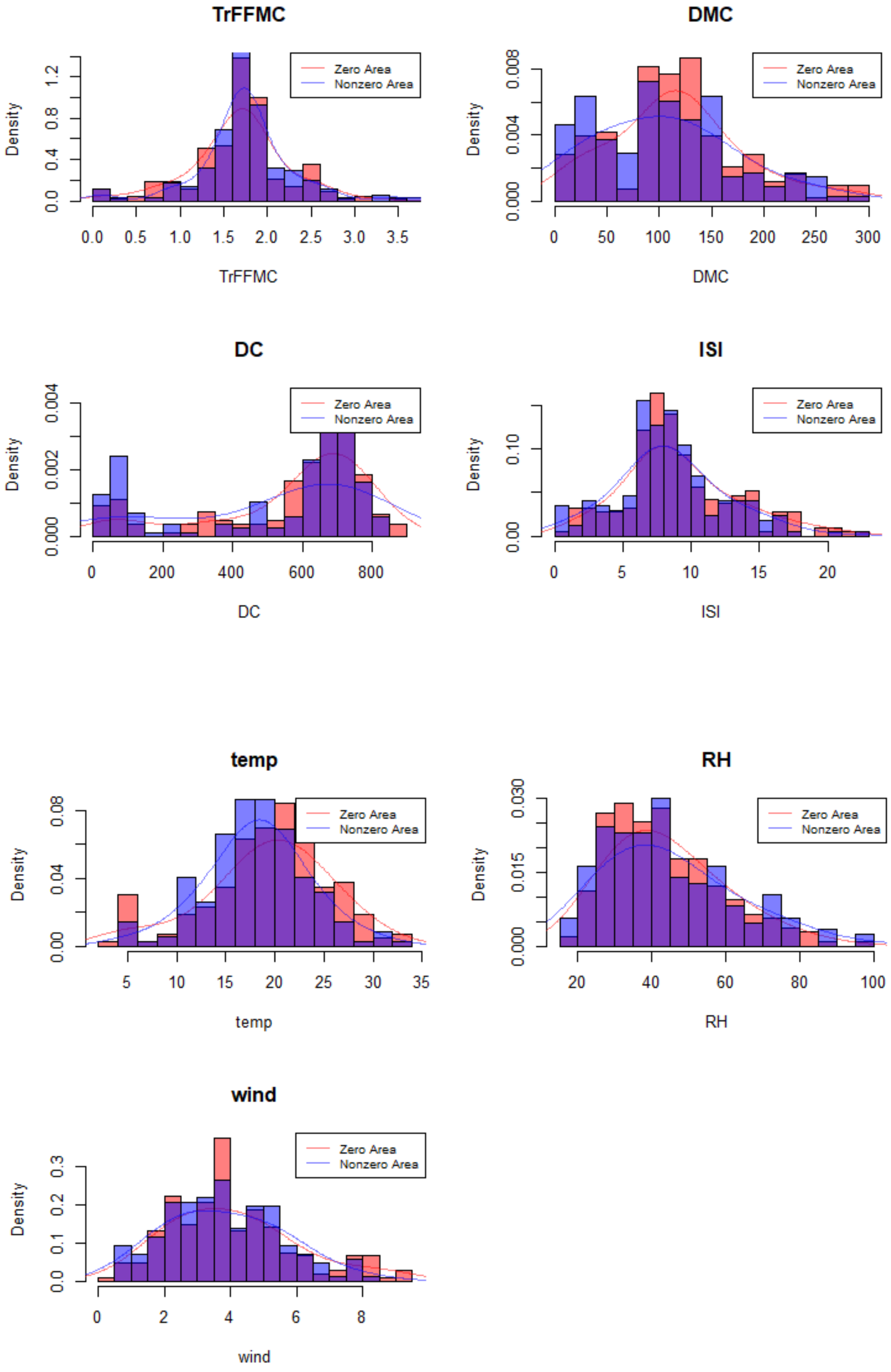

Figure 3: Comparisons of Distributions of Continuous Predictors Conditioned on Zero burn area and Nonzero Burn Area

4 Methods

We randomly split the 517 available observations into a training and testing set, using a ratio of approximately 75/25. The testing set was held out from the model selection process so as not to introduce bias.

4.1 Naive Model

Since we do not find any potentially strong predictors for total burned area within this data set, we will set a benchmark using the Naive model, which simply predicts the total burned area to be the sample mean of the **area** data obtained in the data set.

4.2 Logistic Regression

4.3 Diagnostics

4.4 LASSO Penalized Regression

4.5 Diagnostics

4.6 CART

4.7 Diagnostics

4.8 Random Forest

4.9 Diagnostics

5 Discussion

5.1 Conclusion

5.2 Future Work

6 Appendix

6.1 Data Transformations and Cleaning

6.1.1 Training and Testing Split

Data was split roughly 75/25 training to testing data. Testing data was held back and not used for model fitting.

6.1.2 Area

Since there were many observations with zero area values, we modeled the response in most cases as the transformation

$$log(area + 1)$$

6.1.3 rain

There are only 7 nonzero **rain** values: Considered **rainbin**

6.1.4 FFMC

[Explain transformation]

6.2 Justifications

6.2.1 Method and Theory for Part 4.2 and 4.3

In this section we developed a procedure for fitting confidence bands around our predicted densities as well as the entire model itself. Following are the methods and justification used for both pointwise confidence intervals as well as simultaneous confidence intervals and prediction bands.

Let x_0 be a specified value of gain.

Under the conditional normal linear model, a random observation of density from the population is given by:

$$Y \sim \mathcal{N}(\beta_0 + \beta_1 \log(x_0), \sigma^2).$$

6.2.2 Inference with respect to the mean population density for a specified x_0

6.2.3 Inference for true observed density for a specified x_0 (i.e., prediction)

6.2.4 Simultaneous inference for finite specified values of gain (Bonferroni Bands)

6.2.5 Simultaneous inference for infinitely many specified values of gain (Scheffé Bands)

6.2.6 Logistic Regression

References

[1] **P. Cortez and A. Morais**, "A Data Mining Approach to Predict Forest Fires using Meteorological Data." In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

[2] **Ozbayoglu, Murat Bozer, Recep.**, "Estimation of the Burned Area in Forest Fires Using Computational Intelligence Techniques." Procedia Computer Science. 12. 282–287. 10.1016/j.procs.2012.09.070. (2012).

[3] **Mateus, Paulo Fernandes, Paulo.**, "Forest Fires in Portugal: Dynamics, Causes and Policies." Evolution of Forest Cover in Portugal: From the Miocene to the Present. 10.1007/978-3-319-08455-8_4. (2014)

[4] **Certini, G.** Oecologia 143: 1. https://doi.org/10.1007/s00442-004-1788-8 (2005)

[5] **Guruh Fajar Shidik and Khabib Mustofa** "Predicting Size of Forest Fire Using Hybrid Model." IFIP International Federation for Information Processing (Eds.): ICT-EurAsia 2014, LNCS 8407, pp. 316–327, 2014.

[6] **Casella, George and Roger L. Berger**, *Statistical Inference: Second Edition*, Duxbury

[7] **Wickham, Hadley**, "Tidy Data." *Journal of Statistical Software*, 59.10 `https://www.jstatsoft.org/article/view/v059i10`, 2014.