Forest Fire Report

Joel Cepeda, Jonathan Richards, Rohan Shah, Matthew Bradley, Kaushik Sivakumar

12/10/2021

Introduction

Forest Fires are extremely damaging ecologically, destroying significant amounts of wildlife habitat, as well as contributing to greenhouse gas emissions. Forest fires also cause severe damage to humans, destroying homes, and leading to economic burdens. And there is mounting evidence to show that as climate change worsens, forest fires will only increase in frequency (Weaver et al. 2004).

For these reasons, it is important that humans work towards finding effective ways to mitigate forest fire risks. An important aspect of this is learning to identify when an area is at larger risk for forest fires, and taking preventative measures. However, these predictions are very difficult to make. Two useful tools in predicting fires are weather patterns and systems such as the fire weather index system, which tracks indexes that keep track of soil moisture, wind speed, and other factors to examine fire risk in an area.

It is very difficult to create a reliable predictive model for forest fires. Trends between fire size and weather patterns are complex, and may not tell the whole story (human response time etc... may also be important factors). Still, analysis of available metrics is important and can provide meaningful insight to forest fire patterns, and potentially lessen the ecological and economic impact that fires will have in the future. For this reason, we will attempt, through classification and regression, to build a predictive model that takes a variety of input and predicts forest fire area burned.

Data

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area	logArea
count	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000	517.000000
mean	4.669246	4.299807	90.644681	110.872340	547.940039	9.021663	18.889168	44.288201	4.017602	0.021663	12.847292	1.111026
std	2.313778	1.229900	5.520111	64.046482	248.066192	4.559477	5.806625	16.317469	1.791653	0.295959	63.655818	1.398436
min	1.000000	2.000000	18.700000	1.100000	7.900000	0.000000	2.200000	15.000000	0.400000	0.000000	0.000000	0.000000
25%	3.000000	4.000000	90.200000	68.600000	437.700000	6.500000	15.500000	33.000000	2.700000	0.000000	0.000000	0.000000
50%	4.000000	4.000000	91.600000	108.300000	664.200000	8.400000	19.300000	42.000000	4.000000	0.000000	0.520000	0.418710
75%	7.000000	5.000000	92.900000	142.400000	713.900000	10.800000	22.800000	53.000000	4.900000	0.000000	6.570000	2.024193
max	9.000000	9.000000	96.200000	291.300000	860.600000	56.100000	33.300000	100.000000	9.400000	6.400000	1090.840000	6.995620

Figure 1: Data

Our fire data comes from measurements taken in 2007 in Montesinho park is Portugal. The data consists of 12 predictor variables that include coordinates of fire location, month, day, temperature, relative humidity (RH), wind, rain, and four fire weather indexes; Fine Fuel Moisture Code, Duff moisture Code, Drought Code, and Initial Spread Index. The first step we did in the pre-processing the data was to create dummy variables for our two categorical predictor variables, month and day. The only other processing we did on our predictor variables was using the StandardScaler tool in python to scale our variables for our modeling analysis.

Our response variable is area burned in hectares. However, if a fire was less than 1/100 hectares large, it was marked as 0. This caused several difficulties for our analysis. First, our area variable was extremely skewed

due to so many 0 entries (247 of our 517 entries had an area of 0). Secondly, it was very hard to build a regression analysis that could handle so many 0s, along with some large outliers. We handled these problems in two ways. First, to deal with the skew of the area variable, we did a log transformation ln(area + 1), so that 0 values were still 0, but our range of values was not so right skewed. The second thing we did was build a classifier in order to predict whether a fire would have size 0 or not, and then we did a regression analysis on the fires with size greater than 0. For this we had to create a new variable that was 0 if a fire had a size of 0, and 1 if the fire had a size greater than 0. This allowed us to create a tool that could be used to predict if there would be a sizable fire, and if so, predict how large that fire would end up being.

Exploratory Analysis:

The first thing we did for explanatory analysis was to look at the distribution of our variables. We did not end up doing any transformations here, but it was useful to see which variables were skewed and which were more normal. We see in our predictor variables that rain (especially) and ISI are both right skewed. We also see FFMC is left skewed. Viewing the histograms of the data allowed us to have an idea of which variables may be affecting the regression in different ways than we expect (for example, if we are not happy with our analysis results, we could explore the effects of th few rainy days in our dataset, and use these to help with classification etc. . .).

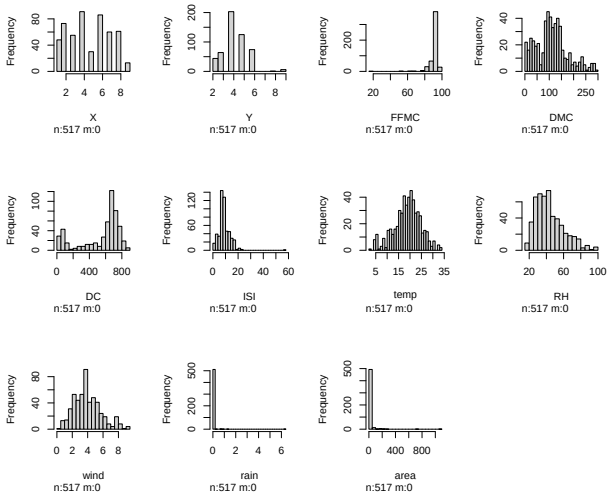
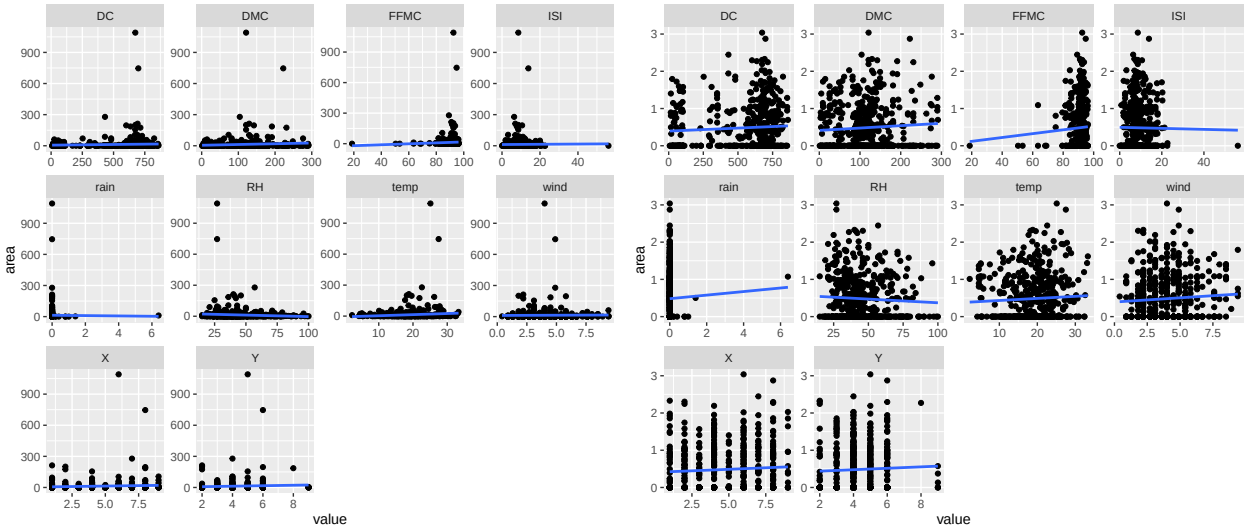


Figure 2: Histograms of Explanatory Variables

Our second piece of exploratory analysis was to look, individually, at our independent variables compared to area (our response variable)(left), and ln(area+1), our transformed response variable (right). This plot shows the effectiveness of our log transformation, it beocmes much easier to see the general trends of our scatterplots because outliers do not have such a large impact on the plots. Still, there are not clear linear relationships between independent variables and our log transformed dependent variable, so it appears that we may need a more complex machine learning algorithm for our regression, rather than linear regression.



Lastly, we examined a correlation matrix for all of our numeric variables. We see that there are no variables that are very highly correlated with area on their own (the most significant relationships with area are with RH, which has a correlation of -0.08, and DMC, which has a correlation of 0.07). But there are some interesting trends in the heatmap: first, the FWI variables are all relatively highly correlated with each other. Second, all of the FWI variables are relatively highly correlated with temperature, which tells us temperature is probably an important factor in these indexes.

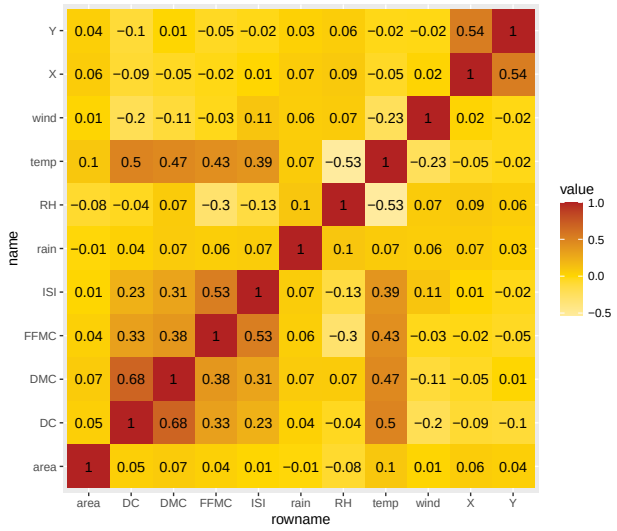


Figure 3: Correlation Heat Map

While we will not use hypothesis testing in this analysis, we did generate several hypotheses based on what we saw in the data.
Hypothesis 1: There is a significant relationship between FWI variables and forest fire size.
Hypothesis 2: Relative Humidity is significantly correlated with area affected by forest fire.
Hypothesis 3: Temperature is significantly correlated the FWI indexes.

Modelling: Classification and Regression

For our classification task, to decide whether a fire would have size 0, or size greater than 0, we tried three different classifiers, and created train and test datasets. We examined accuracy, recall and precision of the

three models we tried. We tried an svc classifier, a knn classifier, and a random forest classifier. The random forest classifier had the highest accuracy, highest precision, and was nearly equal with the knn classifier for highest recall. A confusion matrix for our random forest is shown below:

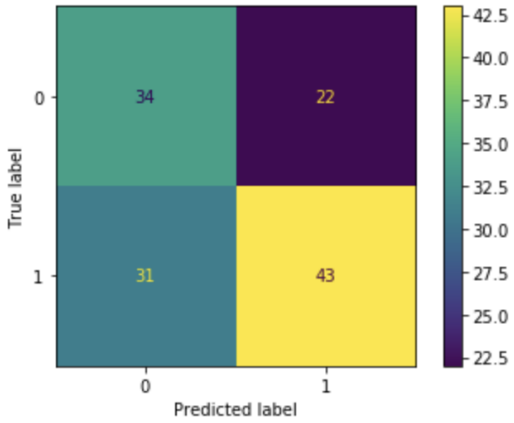


Figure 4: Random Forest Confusion Matrix

Our random forest had the following statistics:
accuracy: 0.59
Precision: 0.66
Recall: 0.58

Satisfied with our classification, we moved on to the regression analysis.

For our regression, we also used a random forest algorithm (and remember, our regression is only on values greater than 0). Our data has nonlinear relationships and the response variable is continuous, so neither logistic or linear regression make sense. Out of the machine learning algorithms we learned in class, random forest regression appears to be a robust option. We used train and test datasets to examine the performance of our regression model. For our results, we used cross validation, and used root mean square error, because it is easily interpretable (it returns the same units as the original measurements). We also used cross validation to measure the R^2 value of th regression.

Regressions results:

Discussion:

Our classification task was around 0.6 for accuracy and recall (true positives/(true positives + false negatives)). So, if the classifier should predict a significant sized fire, it predicts it correctly about 60% of the time (58% exactly). The precision (true positive/(true positive + false positive)) was at 66%, meaning that if our classifier predicts a significant fire, it is correct about 66% of the time. These results indicate that our model may be helpful in predicting whether there will be a significant fire, but not something to completely rely upon.

Our regression had a root mean square error 1.42, which is in the same units as our response variable (but remember, our response variable was log transformed, so the actual average error in our area prediction is ((e^1.42) - 1) = 3.137 hectares). Our R^2 value was 0.435, meaning we can explain around 43.5% of the variation in the data using our model. This leaves us with the question of whether we could improve our regression analysis somehow, or is our data naturally noisy, and thus hard to predict with a high value of R^2. It could also be that there are important predictor variables we are missing, such as human response time or cause of fire. Further analysis could help us determine which of these factors are most important in keeping our findings from being more accurate. Lastly, our regression model struggled to predict extremely high or low values (this can be seen in the plot of actual areas (blue point) versus predicted areas (red points)).

```
RMSE =  1.41943105796484
R^2 = 0.4352966337653599
```

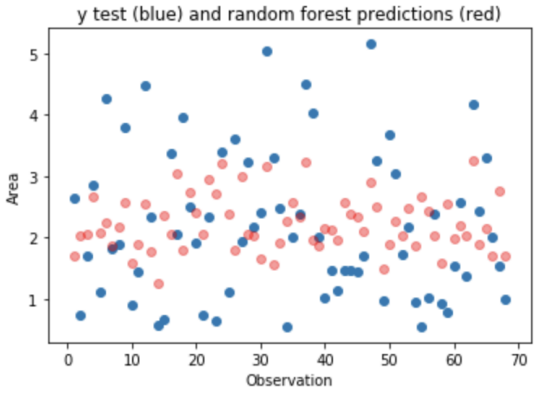


Figure 5: Regression results. Plot shows actual test data points (blue), and regression predictions for these points (red).

The main limitations of our analysis are that the data comes from a single year at a single park. Thus, the results may not be useful in other parts of the world, or even in other years at the same park. This severely limits the applicability of our model. Another limitation of our analysis is that we did not check smaller subsets of our independent variables. For example, maybe running a classifier only using FWI indexes would be more accurate than including all variables, so this is something that would be important to explore in future work.

Conclusion:

Our main findings were that, given our data, we were able to build a model that could help predict forest fire size in Montesinho Park. Our classfier correctly predicted significantly sized fires around 58% of the time, which could be a useful tool for people trying to mitigate fire risk before fires happen. Our regression could help predict how large a fire would be, but only explained around 43.5% of the variability in the data, so there are possible improvements to be made to make the model more useful for predicting fire size.

Acknowledgement:

Matthew Bradley: Coded parts of classification and regression tasks in python. Coded parts of exploratory data analysis code in R. Made associated slides on powerpoint. Wrote associated sections of PDF report, along with producing and inserting images for report, and helping with introduction and discussion sections.

Bibliography

https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004GL020876