Forest Fire Report

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

12/10/2021

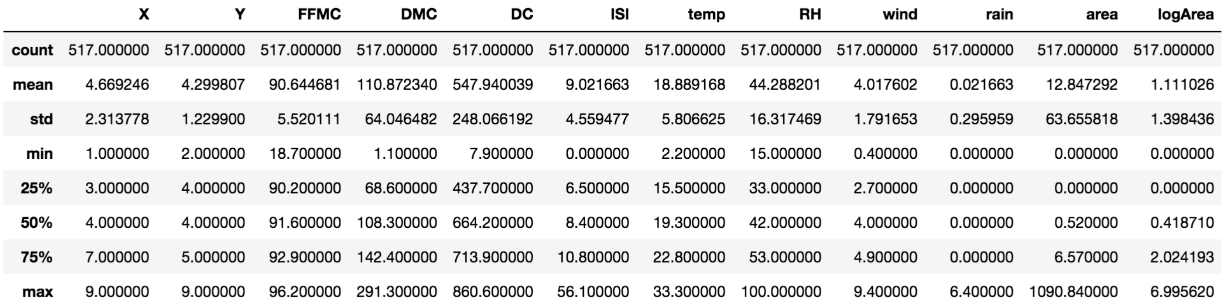
Introduction

Forest Fires are extremely damaging ecologically, destroying 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 increase.

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.

However, 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 tools is important 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 in the pre-processing we did with our predictor variables was creating dummy variables for the two categorical variables, month and day. The only other processing we did was using the StandardScaler tool in python to scale our variables for our modelling 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 ur 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 sizeable fire, and if so, predict how large that fire would end up being.

Exploratory Analysis:

The firest thing we did for explantory analysis was to look at the distribution of our independent 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.

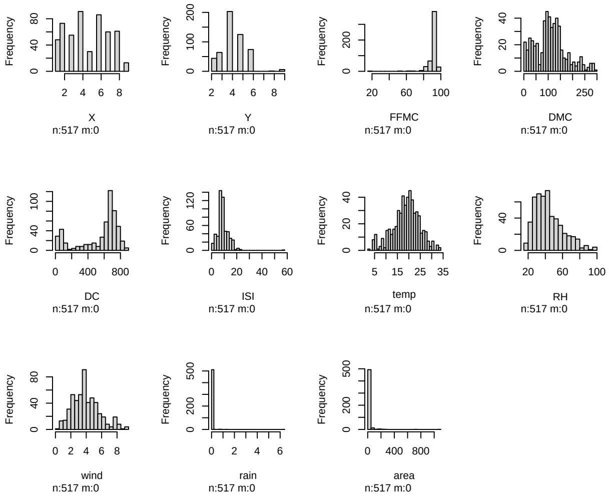
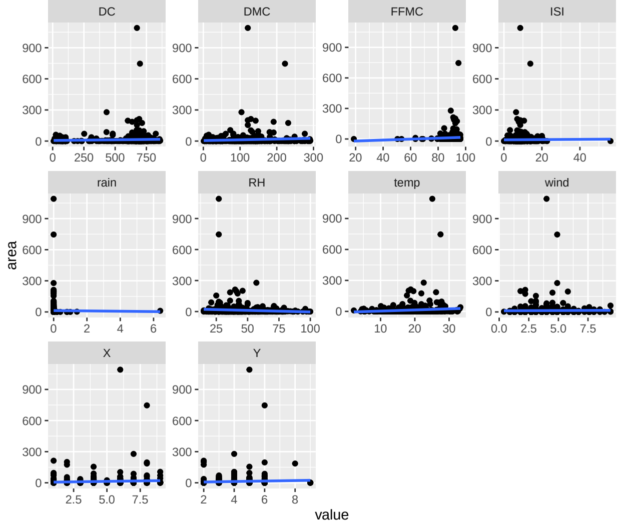
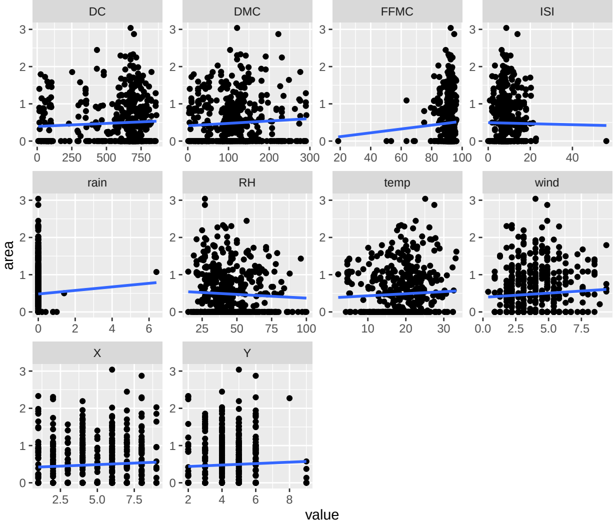


Figure 2: Histograms of Explanatory Variables

Our second piece of exploratory analysis was to look, individually, at our independent variables compared to area (our response variable), and ln(area+1), our transformed response variable.





Lastly, we examined a correlation matrix for all of our numeric variables.

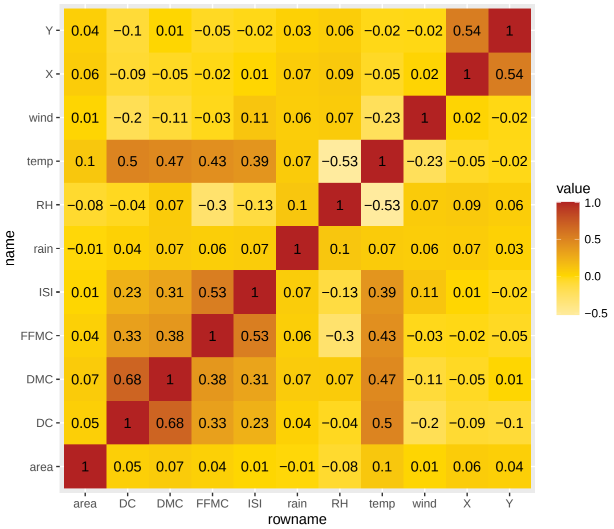


Figure 3: Correlation Heat Map

While we will not use hypthesis testing in this analysis, we did generate seveeral hypthoses based on what we saw in the data.
Hypothesis 1:
Hypothesis 2:
Hypothesis 3:

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 datasts. 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:

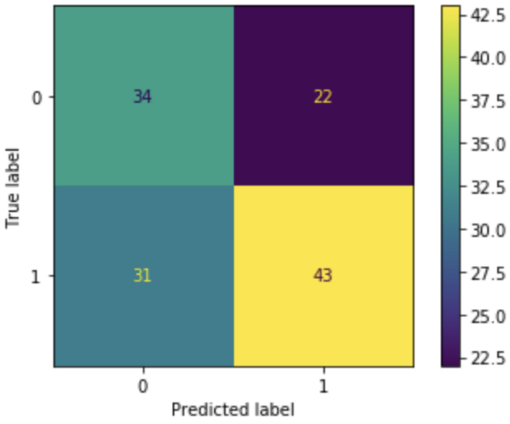


Figure 4: Random Forest Confusion Matrix

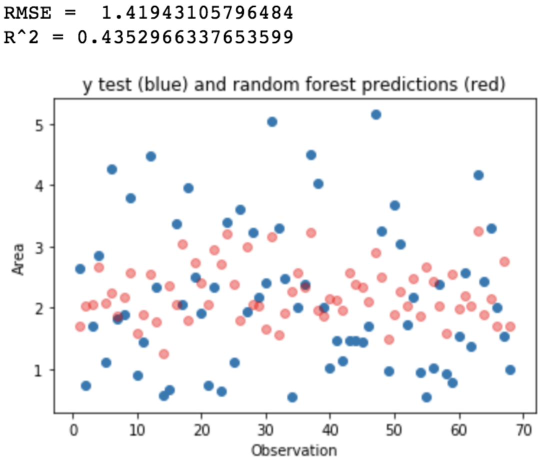
Our random forest had the following statistics:
accuracy: 0.59

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 as well. 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.

Regressions results:



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

Figure 5: Regression results

Discussion:

Our classification task

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 higher than 1.42). Our Rˆ2 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 Rˆ2. It could also be that there are important predictor variables we are missing, such as human response time or cause of fire.

Conclusion:

Acknowledgement:

Bibliography

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