Regression Prediction

Brian Wisniewski

Introduction

The purpose of this report is to predict the burned area of forest fires, in the northeast region of Portugal. Specifically, the goal is to fit multiple regression trees to predict the area burned by a forest fire and minimize the predictive error of our model.

Exploratory Analysis

To inspect the variables and assess which seem to be highly skewed, and which seem to have explanatory power with fire area, I used a combination of histograms and scatter plots (Figure 2 & Figure 3). Figure 1 is included to gain a sense of the overall fire activity in the area. Figure 2 demonstrates frequency, which is primarily useful for assessing skew, and I determined that there was significant skew in the distribution of the burned variable. To account for this skew, that variable was log-transformed.There also appeared to be some skew present in other variables, such as RH and DC, though nowhere near as drastic as the burned variable and, in my opinion, not enough to warrant log transformation. Figure 3 demonstrates explanatory power between variables and the log-transformed fire area. I created these charts with the log transformed data, versus the original fire area data, primarily to make the basic associative relationships more apparent in the graphs. I tried graphing with the original data and the associations were not as easily discernible, because of extreme values skewing the y-axis. Based on my subjective interpretation of Figure 3, it appeared that there was potentially explanatory power between DMC, temp, and DC. In this report, I omitted charts of other variables because they seemed superfluous. For example, graphs related to the explanatory power of rain were omitted because fires almost always occur when it is not raining. Essentially, I used the contextual information in the data set documentation in conjunction with some basic analysis, to determine which variables I would include in this report.

```
## Loading required package: leaps
```

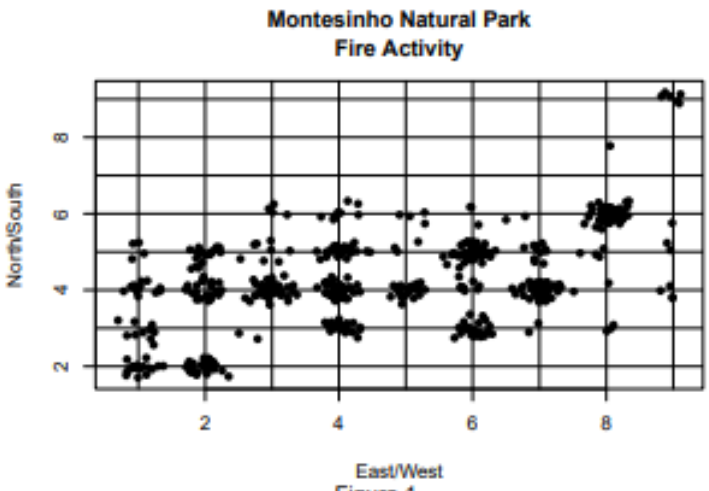

Figure 1

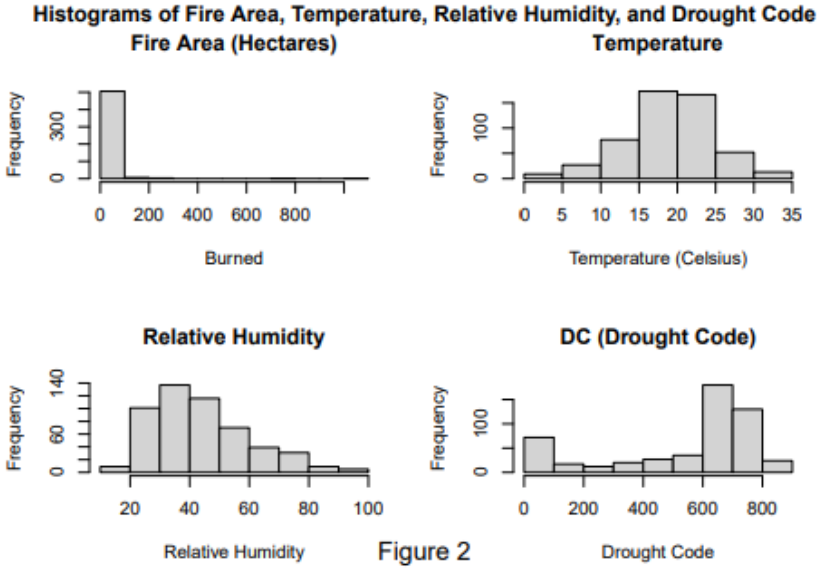

Figure 2

```
# log-transform the response variable
forestfires$area_xform <- log(forestfires$burned+1)
```

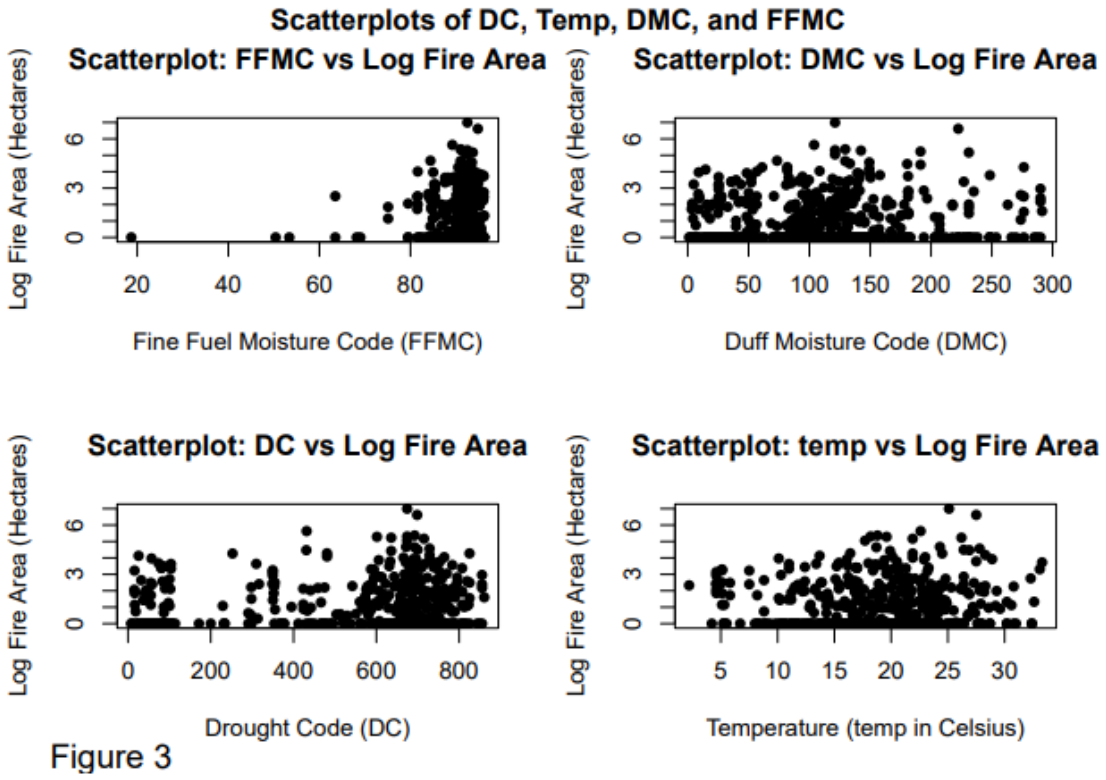

Figure 3

Splitting Training and Testing Data

This code creates a single random subset of nTrain = 467 observations. Then, it sets aside the remaining nTest = 50 observations. I set a random number generator seed to ensure reproducibility. Then I generated a sample size of 50 observations that is utilized in creating a set of test observations. I excluded those test observations from the original data set in order to create the training values

```
# generate random training and testing sets
set.seed(1212)
testvalues <- sample(1:517, size=50, replace=FALSE)
trainvalues <- (1:517)[-testvalues]
```

Model Building

This code, which unfortunately has been cut off and lost in the screenshot, fits and plots five CART regression models to predict fire area, with different variables and tuning criteria. The code also trains each tree on the training observations, and then makes predictions on the forest 3 fire area in both the log-hectare and hectare scale using the testing observations. The variables I focused on, comprising the trees in Figure 4, were based on assessment of explanatory power from Problem 1. I experimented with the tuning criteria, to maximize readability and practical utility. Temp, DMC, and DC were variables that I focused on. It appears

that temp and DMC dominated the computation in the tree formation and are probably two important factors to consider when predicting fires.

```
# build a single tree. You can and should fit more
# trees with greater care than I've taken here.
# Training the tree models
t1 <- tree(burned ~ DMC, data = MontesinhoFires[trainvalues,], split = "deviance")
t2 <- tree(burned ~ RH + DC  + FFMC , data = MontesinhoFires[trainvalues,], split = "deviance", mincut
t3 <- tree(burned ~ FFMC + temp, data = MontesinhoFires[trainvalues,], split = "deviance", minsize = 5)
t4 <- tree(burned ~ DMC + temp, data = MontesinhoFires[trainvalues,], split = "deviance", mincut = 2, m
t5 <- tree(burned ~ FFMC + DMC + DC + temp, data = MontesinhoFires[trainvalues,], split = "deviance", m
```

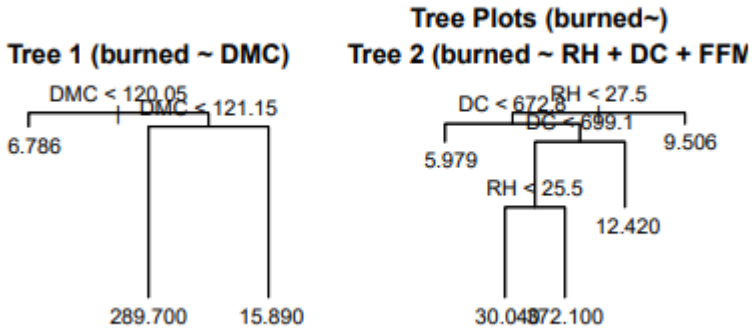

Figure 4

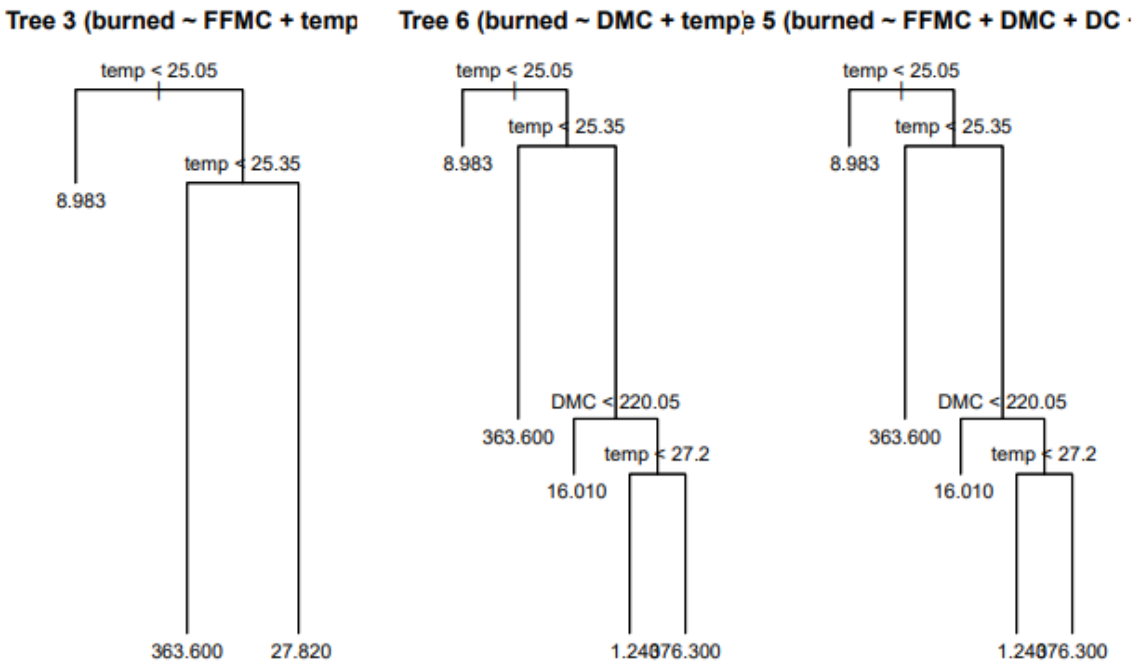

Figure 4 cont

```r
# Making predictions on the testing data (these are log transformed ones)
p1 <- predict(t1, newdata = MontesinhoFires[testvalues,])
p2 <- predict(t2, newdata = MontesinhoFires[testvalues,])
p3 <- predict(t3, newdata = MontesinhoFires[testvalues,])
p4 <- predict(t4, newdata = MontesinhoFires[testvalues,])
p5 <- predict(t5, newdata = MontesinhoFires[testvalues,])

# Convert predictions to hectare scale
p1_hectare <- exp(p1) - 1
p2_hectare <- exp(p2) - 1
p3_hectare <- exp(p3) - 1
p4_hectare <- exp(p4) - 1
p5_hectare <- exp(p5) - 1
```

Validation

This code computes the root mean squared prediction error for each tree, with error reported in both log-hectares and hectares in Table 1.

```r
# Compute RMSPE for each tree model
rmspe1 <- sqrt(mean((p1_hectare - MontesinhoFires$burned[testvalues])^2))
```

```r
rmspe2 <- sqrt(mean((p2_hectare - MontesinhoFires$burned[testvalues])^2))
rmspe3 <- sqrt(mean((p3_hectare - MontesinhoFires$burned[testvalues])^2))
rmspe4 <- sqrt(mean((p4_hectare - MontesinhoFires$burned[testvalues])^2))
rmspe5 <- sqrt(mean((p5_hectare - MontesinhoFires$burned[testvalues])^2))

# Compute RMSPE for log-transformed predictions
rmspe1_log <- sqrt(mean((p1 - MontesinhoFires$burned[testvalues])^2))
rmspe2_log <- sqrt(mean((p2 - MontesinhoFires$burned[testvalues])^2))
rmspe3_log <- sqrt(mean((p3 - MontesinhoFires$burned[testvalues])^2))
rmspe4_log <- sqrt(mean((p4 - MontesinhoFires$burned[testvalues])^2))
rmspe5_log <- sqrt(mean((p5 - MontesinhoFires$burned[testvalues])^2))
```

Table 1: RMSPE for Each Tree Model (Table 1)

Tree	RMSPE_Hectare	RMSPE_Log_Hectare
Tree 1	5.046318e+06	17.14984
Tree 2	1.570353e+12	16.98600
Tree 3	Inf	53.46405
Tree 4	Inf	53.02368
Tree 5	Inf	53.02368

Averaging Predictions

This code takes the same predictions for each of the five tree and computes the average of the five predictions I got for each of the 50 testing values. It uses that average as the global prediction for each testing observation. Then, it computes the RMSPE based on this averaged predictive value. That RMSPE using the log transformed hectares, was ~34.6. That RMSPE using

the other vales was infinity. The RMSPE of infinity, suggests there are cases where the predicted values differ significantly from the actual values. The RMSPE of ~34.6 indicates the predicted versus actual values are much closer. This may indicate that the log-transformed values may be a better representation of the data in this context. It may also indicate that certain models may be extremely inaccurate, and we should adjust them or remove them. Once we did that, the RMSPE for the original data may not be infinity.

```
# Predictions for each tree model on the testing data
predictions <- data.frame(
  Tree1 = p1,
  Tree2 = p2,
  Tree3 = p3,
  Tree4 = p4,
  Tree5 = p5
)

# Compute average of the predictions for each testing obs
averaged_predictions <- rowMeans(predictions)

# Compute RMSPE based on the averaged predictive values
```

```
rmspe_averaged_log <- sqrt(mean((averaged_predictions - MontesinhoFires$burned[testvalues])^2))
rmspe_averaged_log
```

```
## [1] 34.60796
```

```
# Calculate predictions for each tree on testing data
predictions <- data.frame(
  Tree1 = p1_hectare,
  Tree2 = p2_hectare,
  Tree3 = p3_hectare,
  Tree4 = p4_hectare,
  Tree5 = p5_hectare
)

# Compute average of predictions for each testing obs
averaged_predictions <- rowMeans(predictions)

# Computes the RMSPE based on averaged predictive vals
rmspe_averaged <- sqrt(mean((averaged_predictions - MontesinhoFires$burned[testvalues])^2))
rmspe_averaged
```

```
## [1] Inf
```

Conclusion

The process went well overall, but there were aspects that could've gone better. Some of the tree models that I chose produced infinite RMSPE, which caused the average RMSPE, for the original data, to be infinite. I think my method for choosing variables could be more thorough- It would be great to speak to an expert in the field, to understand the context and how the variables function in the context. I think choosing 5 models is a reasonable number of

trees to use for this data set, so that choice went well. I think it allowed me to explore combinations of variables and fine tuning of the models enough to be informative, but not superfluous. Overall the structure of the process went well, but I think there is room for improvement in the preliminary analysis of the variables and laying the groundwork for my modeling. In summary, I would spend more time thinking about what to include in the models, before creating the models. As far as whether the average prediction outperformed any of the individual trees' RMSPE: Only the average predicting using log transformed data did. It outperformed trees 3,4,5, which checks out with my intuition because trees 1,2 were the ones that I put most preliminary thought into. Interestingly, changing the seed causes drastic changes in the results. Changing the seed from 1212 to 1210, causes the relationship I described in the last paragraph, to invert. Additionally, none of the RMSPE values for log transformed or original data were infinity. In fact, the models that had RMSPE of infinity with seed 1212, performed extremely well with seed 1210. I think to gauge the power and utility of our models, or make real decisions based on our analysis, we may want to experiment with many seeds and observe those results in context with one another.