STAT 6950 Final Project

Fernando Anorve Lopez
Christina Sousa

April 22, 2019

Contents

1 Introduction

In this report we will investigate data from $n = 517$ forest fires occuring in the Montesinho National Park in northeast Portugal between January 2000 and December 2003. The data comes to us by way of the UCI Machine Learning Repository, and was collected by Paulo Cortez and Anibal Morais at the University of Minho, in Guimaraes, Portugal [**1**][**6**]. 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, and meteorological data is readily available and constantly updated. Any insight this type of data can offer would be helpful in mitigating these disasters.

Forest fires are broadly defined 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 wildfires per year on average. These burn anywhere from about 150,000 hectares (ha) to 250,000 ha annually. [**3**]

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 (MLR) 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 Variables

1. **X**: x-axis spatial coordinate within the Montesinho park map: 1 to 9 (See A.1.1)
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
13. **area**: (Response) the burned area of the forest (in ha): 0.00 to 1090.84

To provide a better understanding of the spatial information of the fires in this dataset, Figure 1 illustrates the locations and magnitudes of each of the fires within the park.

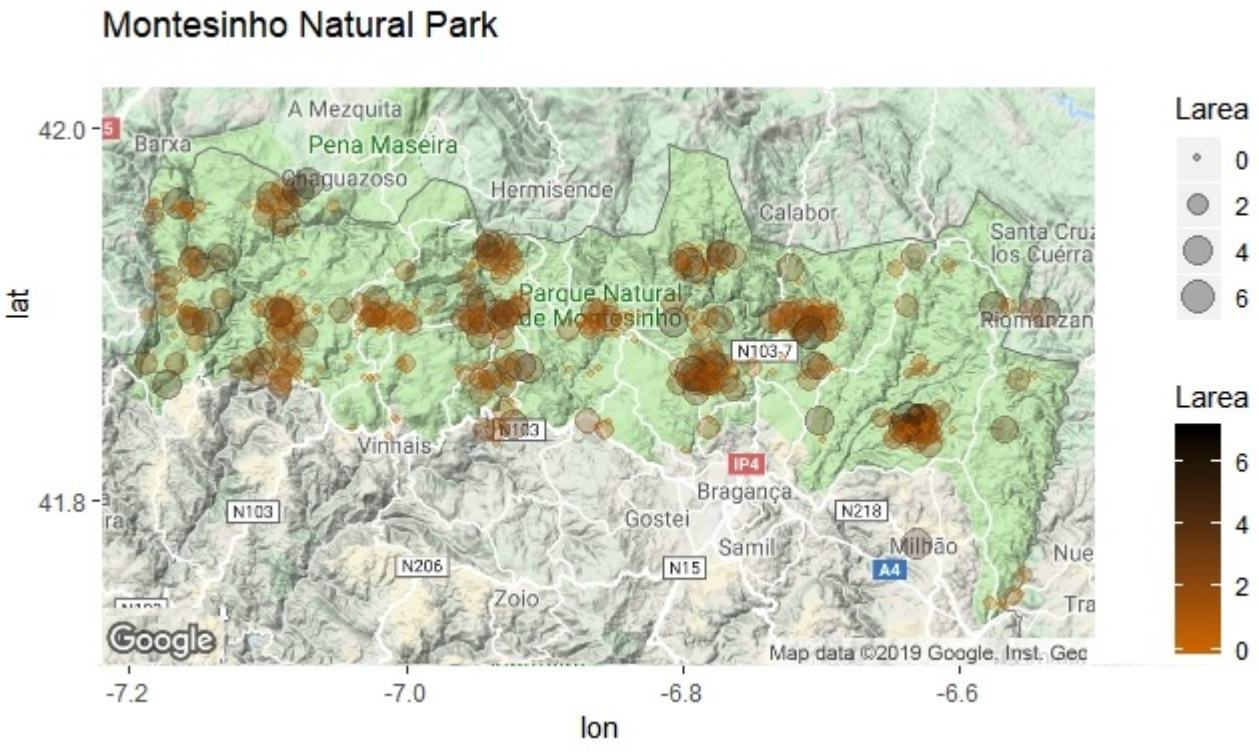


Figure 1: $\log(\mathbf{Area} + 1)$ with respect to the spatial coordinates. Burn area was plotted using the transformation $\log(\mathbf{Area} + 1)$ This particular transformation is useful because small fires appear with a 0 value, while the large fires area appear in a logarithmic scales. Notice that the data set has discrete coordinates $\mathbf{X}$ and $\mathbf{Y}$. Hence a randomly jittered version of the data is shown to avoid overplotting.

2.2 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**]. We should view any models arising from this analysis as a step 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 given the complex nature of the problem and the still-developing computational techniques for analyzing this kind of data.

Complicating the analysis of this particular data set is the fact that weather conditions were not reported on days without fires. It is reasonable that, for example, lack of rain for weeks prior to a fire event may result in a larger burn area, so this poses a major limitation. The data also was not recorded sequentially because it is the combination of two data sets, one which was entered in

order of fire size, and the other which was ordered chronologically. This hinders the ability to use a time-series approach (See A.2.5 for more on this).

Furthermore, fires that burned less than $100m^2$ of area were 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 $100m^2$.

Finally, we note that 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, indicating that most of the fires in this data set would be classified as "Small", and there is not much information with which to predict "Large" ($>$ 63.86ha) versus "Small" ($<$ 63.86ha) fires. 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 its limitations, let us have a look at some visualizations to get an idea of the relationships between the various predictors and **area**. When modeling the transformation log(**area**+1), we will refer to this variable as **Larea**.

First, we view the continuous predictors (Figure 2). Note the discretization which has occurred in the **Larea** plots for the zero burn area records. Within these continuous predictors, none are highly linearly correlated with **Larea**. **DMC** and **wind** are the most highly correlated with **Larea**, but at only 8% and 10% respectively. To make matters worse, many of the predictors are highly correlated with one another. For more details see A.1.2.

Considering now the categorical predictors (Figure 3), we note that December and May appear to stand out as having higher log(area) values, as well as **no rain**. To determine whether there may be some useful categories for prediction here, ran ANOVA on each of **month**, **day**, and **rain**. For **month**, **dec** indeed emerged as having significantly higher mean **Larea** than other months (p-value=0.0136). Also **May** had somewhat significantly higher mean **Larea** (p-value=0.0565). The ANOVA test for **day** and **rain** did not reveal any groups with significantly higher mean **Larea**, so these do not appear to be useful.

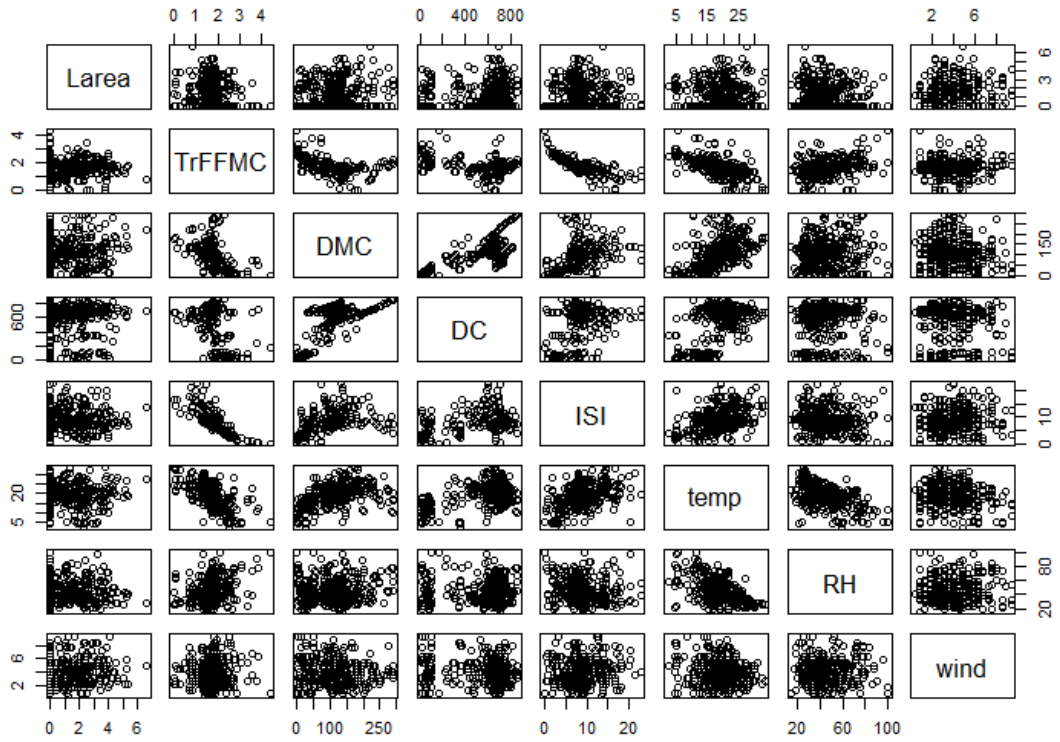


Figure 2: Scatterplot Matrix of the Transformed Continous Predictor Variables versus Log(Area+1)

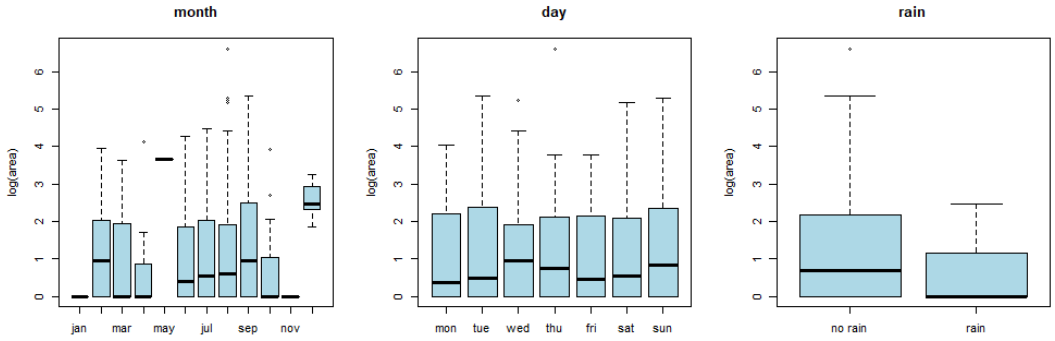


Figure 3: Categorical Predictors versus Larea

Since we are interested in classifying zero-burn area fires versus nonzero burn-area fires, we plotted 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. Those that had significantly different means are shown in Figure 4; the rest can be found in A.1.4. **TrFFMC**, **DMC**, and **temp** all significantly differed at the $\alpha = .10$ significance level (but not at $\alpha = 0.05$) according to two-sample Welch's t-tests. Since the distribution of **DC** is highly non-normal, we used a Wilcoxan Rank Sum test, which also rejected with a p-value of 0.07461. These are promising results, as they indicate that there are at least detectable mean shifts in zero-area fires versus nonzero-area fires for these predictors.

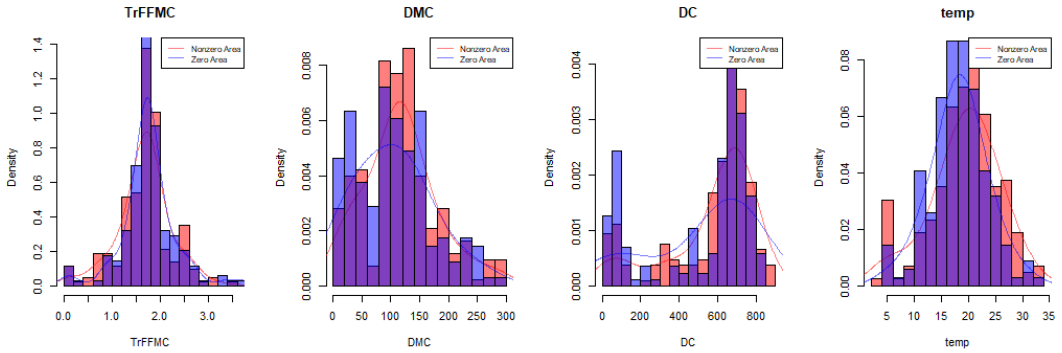


Figure 4: Histograms for continuous predictors conditioned on zero and nonzero burn area (significant mean shifts).

For additional plots investigating separation in the distributions of continuous predictors conditioned on zero burn area versus nonzero burn area, see Appendix A.1.3.

Next, we fit simple linear regressions with each of the continuous predictors separately. The slopes are not striking, however again **TrFFMC** and **DMC** appear to emerge, as was the case in the histogram comparisons. Also, **wind** looks like it may be slightly helpful in predicting **Larea**. We note that the slope for **temp** is not as striking as one might expect intuitively.

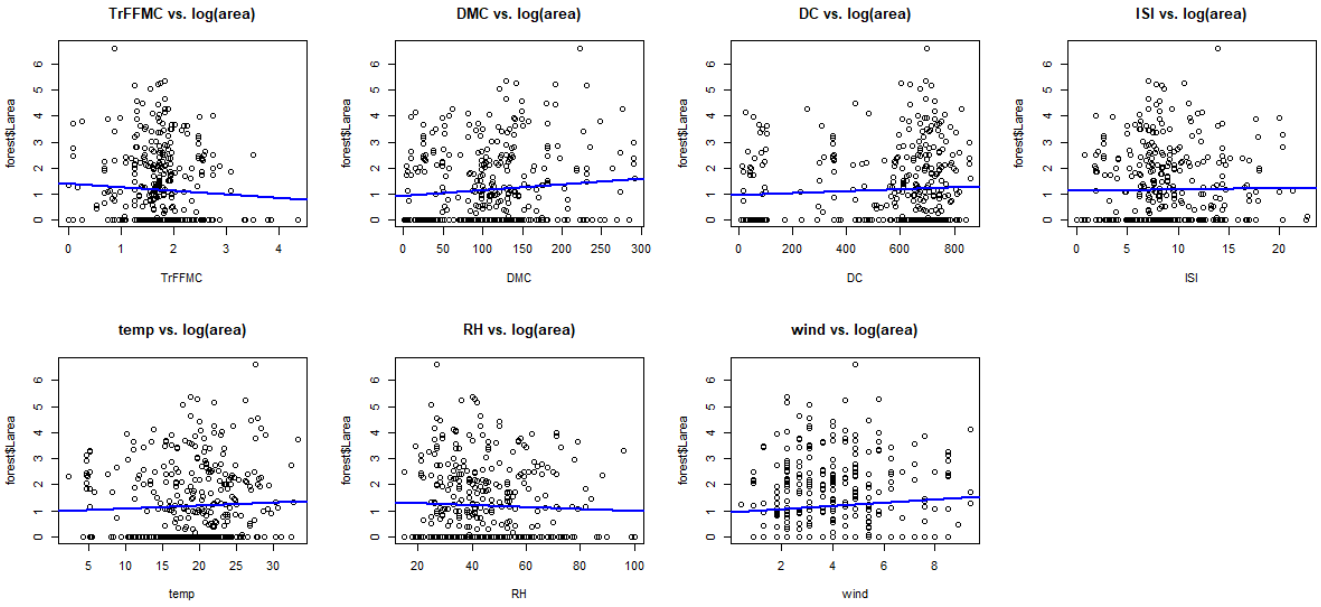


Figure 5: Simple linear regressions of continuous predictors on Larea

We also investigated the two-way interactions between these predictors. After all, weather systems are very dynamic and it is reasonable that there are interactions between wind, humidity, temperature, and the fire indexes. Making use of intermediate multiple linear regressions to estimate linear coefficients, we plotted estimated two way interactions. The plots revealed evidence of strong interactions between the predictors shown in Figure 6. For example, with respect to **temp** and **wind**, we observe that as **temp** increases, the effect of **wind** on **Larea** appears to change from

negative to positive. Because such a pronounced change could mask effects of individual predictors, these interactions must be considered in the modeling process. For all two-way interaction plots, see A.1.5.

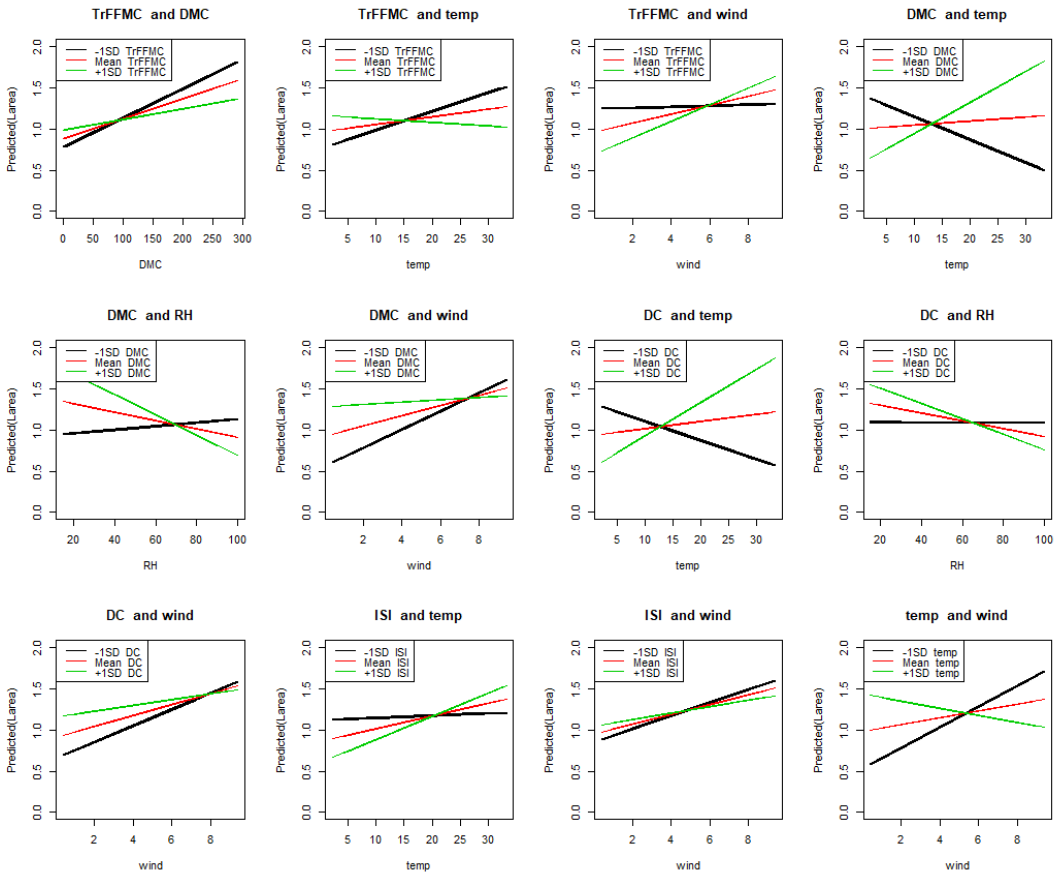


Figure 6: Two-way interactions of continuous predictors with respect to Larea (selected plots; for all plots see A.1.5). These lower-dimensional plots show how the effect of the first predictor on **Larea** changes as the second predictor is varied through three discrete levels: the mean minus one standard deviation, mean, and mean plus one standard deviation.

We conclude this Exploratory Data Analysis section by reiterating that this is not a typical regression or machine learning problem, where suspected predictors emerge somewhat immediately through plotting and remain consistent when viewed from different perspectives. Here we have that **wind** and **DMC** appeared to be candidates in terms of linear correlation, **month=dec** appeared as a candidate in consideration of categorical predictors, while in seeking separation of zero and nonzero burn areas, **TrFFMC**, **DMC**, and **temp** appeared to be promising candidates. **TrFFMC**, **DMC**, and **wind** also emerged as promising candidates in the simple linear regressions. The presence of two-way interactions will also pose challenges in our analysis. Hopefully logistic regression and multiple linear regression methods can help detect what the human eye cannot, to beat out a Naive model and advance the case for predicting fire burn areas with on-line meteorological data.

4 Methods

4.1 Naive Model

Since we did 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

Our first approach to model the data was to apply a logistic regression using the weather predictors mentioned in Section 2.1 (as we wished to focus exclusively on the meteorological predictors) to predict the likelihood that a given fire has a small or a large burn area.

4.2.1 Model selection

As the variable selection technique all of the possible models using weather predictors were compared using the deviance as a reference of fit. Recall that the weather related variables are **FFMC**, **DMC**, **DC**, **ISI**, **temp**, **RH**, **wind**, and **rainbin**.

Figure 7 shows how deviance is reduced as more predictors are added to the model. For each number of predictors, the two best models in terms of the deviance. Its should be noticed that the deviance of the naive model is approximately 535, while the deviance of the complete model is 520 (i.e. there is not a very significant change).

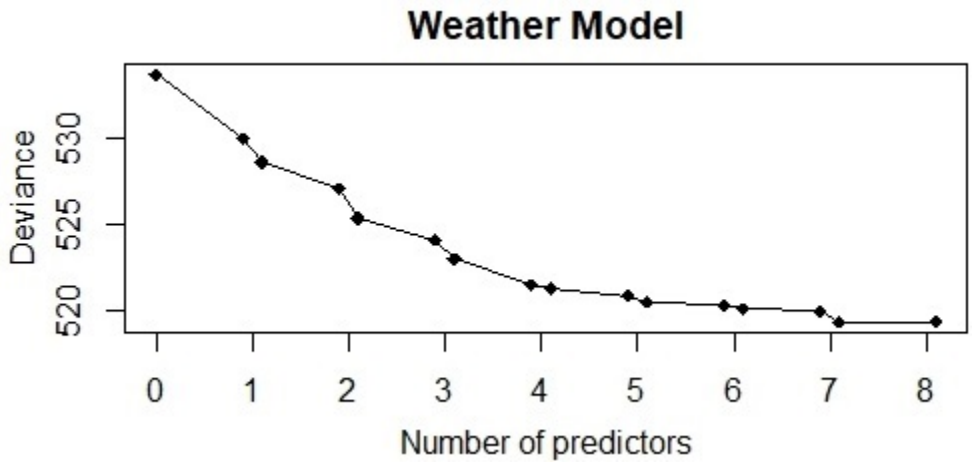


Figure 7: Deviances vs. Number of Predictors: Exhaustive Comparison

Again from Figure 7 we observe that using more than four predictors does not decrease the deviance of our model considerably. Hence we can say that using four predictors is appropriate. Restricted to four predictors, the best model uses the variables **DC**, **temp**, **wind**, and **rainbin**, and its summary is shown in A.4. Results show that only **wind** appears to be significant at a 0.05 level, while **DC** seems to be significant at a $\alpha = 0.10$ level but not at $\alpha = 0.05$.

4.2.2 Diagnostics

As in this case there is only one observation per covariate, an analysis based on the residuals plot or a qq-plot can be misleading. Instead we show the Residuals vs. Leverage plot in Figure 8.

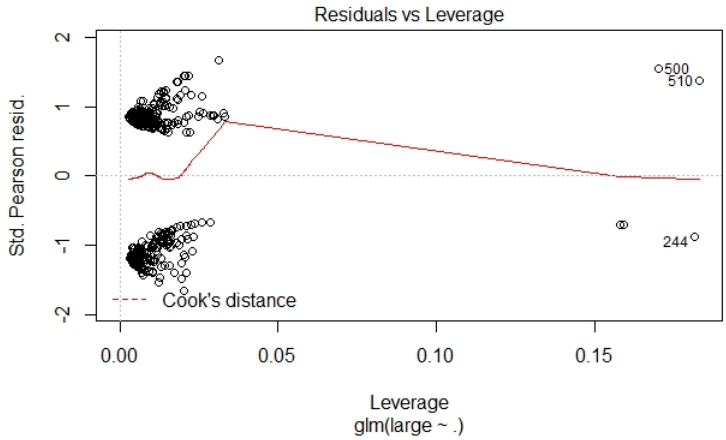


Figure 8: Residuals vs. Leverage Plot - Logistic Model

Note in Figure 8 that even though five points have higher leverage than the rest of the observations, Cook's distance lines do not appear in the graph, implying that they still have low Cook's distance scores.

4.2.3 Testing Results

Predictions on the testing set using the logistic model provided useful diagnostics. We can get predicted values by calculating the predicted probability

$$\hat{p}_{X^*} = \hat{p}(Y|X_1^*, X_2^*, X_3^*, X_4^*),$$

and next assigning the value

$$\hat{Y} = \begin{cases} 1 & \text{if} \quad p_{X^*} \geq 0.5 \\ 0 & \text{if} \quad p_{X^*} < 0.5 \end{cases}$$

We obtained the following results:

	Small Fires	Large Fires
Predicted	27	102
Actual values	64	65

	Frequency
Accurate Predictions	74
Wrong Predictions	55
% of Accurate Predictions	57%
False Small Fires	9
False Large Fires	46

The logistic regression model correctly predicted 57% of the actual outcomes of the testing set. This is a rather poor result considering that the naive model, which assigns $\hat{Y} = 1$ regardless of the predictors, would have 50.38% of accuracy. Interestingly, this model can be considered conservative in the sense that the predicted values indicated that most of the fires were likely to be large given the regressors.

4.3 CART

An alternative model is a classification tree. Our goal is to find a model for predicting small versus large fires from our weather data. Our set of predictors was the same as the logistic regression: **FFMC**, **DMC**, **DC**, **ISI**, **temp**, **RH**, **wind**, and **rainbin**.

Figure 9 shows that, in terms of the Cross-Validated Error, the tree does not seem to improve considerably as complexity parameter varies. Hence a pruned version of the tree with CP = 0.014 was considered, as this was roughly the point where the error was lowest (Figure 9).

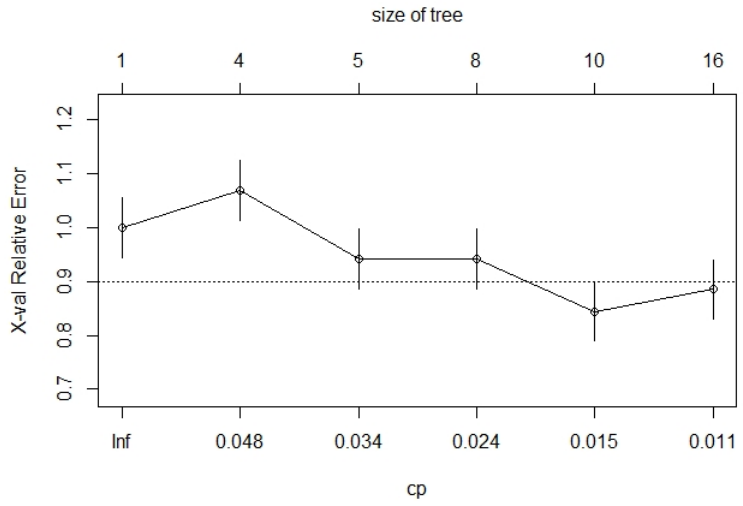


Figure 9: Cross-Validated Error Versus Complexity Parameter

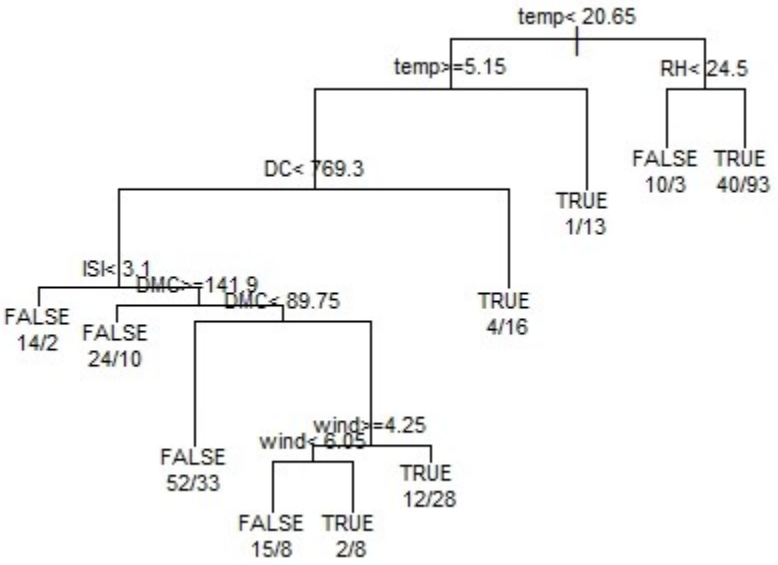


Figure 10: Classification tree pruned with CP = 0.014

4.3.1 Testing Results

As we did in the logistic model we can get the predicted probability

$$\hat{p}_{X^*} = \hat{p}(Y|X_1^*, X_2^*, X_3^*, X_4^*),$$

and assign the value

$$\hat{Y} = \begin{cases} 1 & \text{if} \quad p_{X^*} \geq 0.5 \\ 0 & \text{if} \quad p_{X^*} < 0.5 \end{cases}$$

We obtained the following results:

	Small Fires	Large Fires
Predicted	49	80
Actual values	64	65

	Frequency
Accurate Predictions	96
Wrong Predictions	33
% of Accurate Predictions	74.4%
False Small Fires	9
False Large Fires	24

The classification tree correctly predicted the likelihood of 74.4% of the actual outcomes of the testing set. Interestingly, this is a better result than the obtained with the logistic model. Similar to the logistic model, this classification tree can be considered conservative in the sense that the predicted values indicated that, given the regressors, it is more likely to predict large fires than small ones.

4.4 Multiple Linear Regression

4.4.1 Model Selection

We tried four different sets of predictors for the Multiple Linear Regression model with **Larea** as a response. We did not consider **month** and **day** here, as we wished to focus exclusively on the meteorological predictors. The first set of predictors were those suggested by our exploratory data analysis (EDA), along with their respective two-way interaction terms. Second, we considered the set of predictors suggested by full enumeration of the model space, including all two-way interactions where we had evidence of those, using the **leaps** package. Since this approach resulted in a model that was not structured well hierarchically, i.e., it included interaction terms without including the lower-order single predictors, we also used the **add1()** method starting from the mean model and **drop1()** method starting from a full model, conditional on maintaining a hierarchical structure. Finally, we used the sets of predictors suggested by forward selection, backward elimination, and stepwise regression, where forward selection and stepwise regression suggested the same predictors, while backward elimination suggested the mean model.

Set	**Predictors**
EDA	TrFFMC, DMC, wind, temp, TrFFMC:DMC, TrFFMC:temp, TrFFMC:wind, DMC:temp, DMC:wind, temp:wind
Leaps	TrFFMC, DMC, ISI, RH, wind, TrFFMC:temp, TrFFMC:wind, DMC:RH, DMC:wind, ISI:temp, ISI:wind, temp:wind, RH:wind
Add1	(mean model)
Drop1	TrFFMC, DMC, DC, ISI, temp, RH, wind, TrFFMC:DMC, TrFFMC:temp, TrFFMC:wind, DMC:temp, DMC:RH, DMC:wind, DC:temp, DC:RH, DC:wind, ISI:temp, ISI:wind, RH:wind
Forward/Stepwise	DMC, wind
Backward	(mean model)

4.4.2 Diagnostics

We use BIC to compare models (see Appendix A.2.6 for further discussion of why other metrics were not preferred). It is interesting that the statistical intuition of our EDA actually beat the **leaps** package here in terms of BIC reduction. However, since forward selection beats them both, as well as the Drop1 set, with fewer predictors, we would opt for that model from the ones considered in this section. The mean model does achieve lower BIC, but perhaps at the expense of bias.

Predictors	df	BIC
EDA	12	1413.093
Leaps	16	1431.508
Drop1	21	1465.298
Forward/Stepwise	4	1376.473
Mean Model	2	1372.353

Unfortunately, it appears that the error distribution for this model is rather skewed, so that assumption of normal errors has been violated here. This is likely due to the distcretization of the burn areas less than $100m^2$.

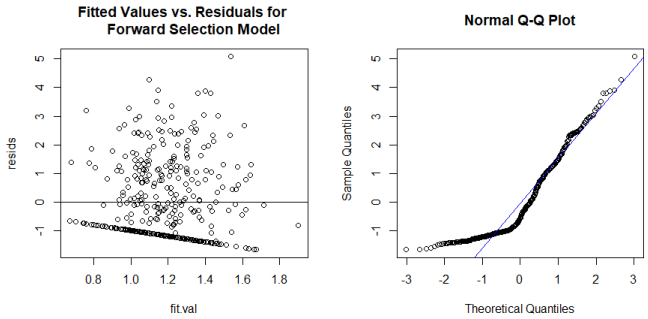


Figure 11: Residual Plot and QQ Plot from Forward Selection Model

Indeed, if we condition this model on nonzero burn area, we observe somewhat better conformance to assumptions, although the residual distribution still appears to be slightly skewed. The Durbin-Watson test does not detect any auto-correlation among the errors (p-value=0.702), and the Cook's Distances for this model are acceptable. However the K-S test for normality of errors returns a p-value of 0.05214. This demonstrates some evidence against normality of the errors. The Shapiro-Wilk test, on the other hand, returns a p-value of 2.164e-05, providing strong evidence against normality. Therefore we cannot assume that the model assumptions have been met in this case.

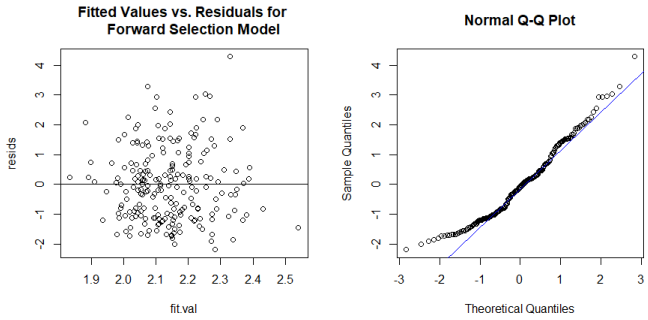


Figure 12: Residual Plot and QQ Plot from Forward Selection Model conditioned on Nonzero Burn Area

4.4.3 Testing Results

The performance of the four MLR models and mean model on training and testing sets is summarized below. We observe that the **Leaps** set of predictors yielded the smallest error on both the training and testing sets. This model has therefore been stated fully with further diagnostics in Appendix A.5.

Predictors	In-Sample MSE	Prediction MSE
EDA	1.875301	1.962941
Leaps	1.849228	1.931399
Drop1	1.868427	1.986195
Forward/Stepwise	1.929605	2.043869
Mean Model	1.968872	2.038497

To in interpret the size of this error, recall that the response has been transformed using $g(y) = log(y + 1)$. Hence the average prediction error for the **Leaps** model in hectares is given by

$$e^{\sqrt{PMSE}} - 1 = 3.01ha,$$

or about a half of a football field. While this may not be incredibly precise given that the median area of fires in this data set is 0.9 hectares, this model could still be useful for gauging how large of fires could be expected given current weather conditions. Note, however, that the model only achieves an adjusted R^2 value of 0.02542. Therefore even though this model appears to offer come predictive power, it is important to remember that there are other variables not included in this model which would explain more of the variation in burn area, namely method of containment and response times.

5 Discussion

5.1 Conclusion

In the task of predicting $> 100m^2$ fires, the logistic model produced rather mediocre results in both in-sample and testing diagnostics. The residual deviance of the model was not significantly different from the null deviance, and its accuracy of prediction using the sample set was barely higher than the naive model. Our second approach, the classification tree, yielded better results at predicting $> 100m^2$ fires using the test set. While the CART model seems to work very well at predicting these fires, these models are often difficult to interpret and generalize. Nevertheless, it is worth mentioning that this model took into account the predictors **DMC**, **DC**, **ISI**, **temp**, **RH**, and **wind**, thus suggesting that these variables could be useful for future research.

In the task of predicting the area of forest fires, the MLR model including the **Leaps** predictors succeeded in beating the mean model in terms of prediction error by about .15ha, or about $150,000m^2$. Still, it seems that many factors influencing the response have not been accounted for in this data set, resulting in a very low adjusted R^2 of 0.02542. This means that these predictors explain only a slight amount of the variability in the response. However, this model could still provide valuable information for the park managers by giving a better-than-average sense of how large fires might be given current weather conditions.

5.2 Future Work

In future research regarding weather conditions to predict wildfires or other similar events, it would be useful to order the data set chronologically to make possible a time series analysis. It could also be convenient to provide more detailed data of small fires so that a continuous response is available for all types of fires, and even to record weather conditions of the days when no fires were reported.

Future works based on these predictors could focus on **DC**, **temp**, and **wind**, as these variables consistently appeared in our models. Since some evidence showing interaction was found, it could also be useful to consider at least second order interaction terms. We found that these second order interactions improved the predictive performance of our MLR models, so it is possible that they will show even better results when used in more advanced machine learning algorithms.

A Appendix

A.1 Supplemental Figures

A.1.1 Map of Montesinho Natural Park

The **X** and **Y** spatial variables that appear in this data set divide the map of the park into 81 rectangles that are shown in Figure 13.

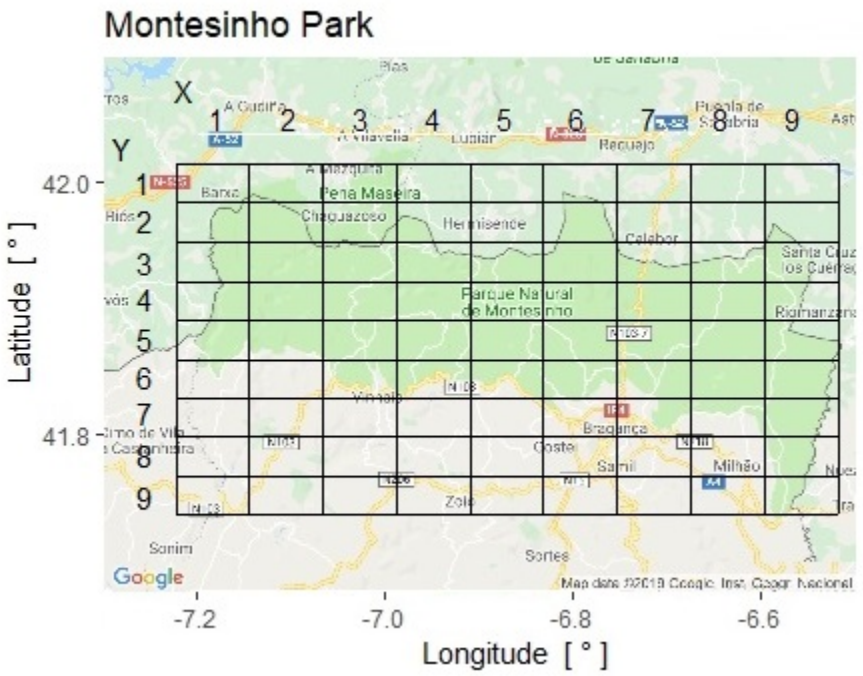


Figure 13: Map with spatial coordinates **X** and **Y** of the Montesinho Natural Park

A.1.2 Correlation Matrix and VIF

Following is the correlation matrix for the continuous predictors in the data set. Note that while many of the predictors are correlated with one another, none exhibit a high degree of correlation with the response **Larea**.

It appears that **TrFFMC** and **ISI** are highly correlated (-0.7665), as well as **DC** and **DMC** (0.6825), **temp** and **TrFFMC** (-0.6047). Furthermore, **temp** and **DMC** have somewhat high linear correlation (0.4923),along with **temp** and **RH** (-0.4866), **temp** and **DC**(0.4886) and **TrFFMC** and **DMC** (-0.4847). Finally, **temp** and **ISI** appear to be somewhat linearly correlated (0.4417), along with **TrFFMC** and **DC** (0.4886).

	Larea	TrFFMC	DMC	DC	ISI	Temp	RH	wind
Larea	1.0000	-0.0550	0.1032	0.0667	0.0114	0.0472	-0.0421	0.0802
TrFFMC	-0.0550	1.0000	-0.4847	-0.3910	-0.7665	-0.6047	0.3011	0.0708
DMC	0.1033	-0.4847	1.0000	0.6825	0.3597	0.4923	0.0849	-0.1459
DC	0.0667	-0.3910	0.6825	1.0000	0.2970	0.4886	0.01256	-0.1859
ISI	0.0114	-0.7665	0.3597	0.2970	1.0000	0.4417	-0.1366	0.0839
Temp	0.0472	-0.6047	0.4923	0.4886	0.4417	1.0000	-0.4866	-0.2342
RH	-0.0421	0.3011	0.0849	0.0126	-0.1366	-0.4866	1.0000	0.0576
wind	0.0802	0.0708	-0.1459	-0.1859	0.0839	-0.2342	0.0576	1.0000

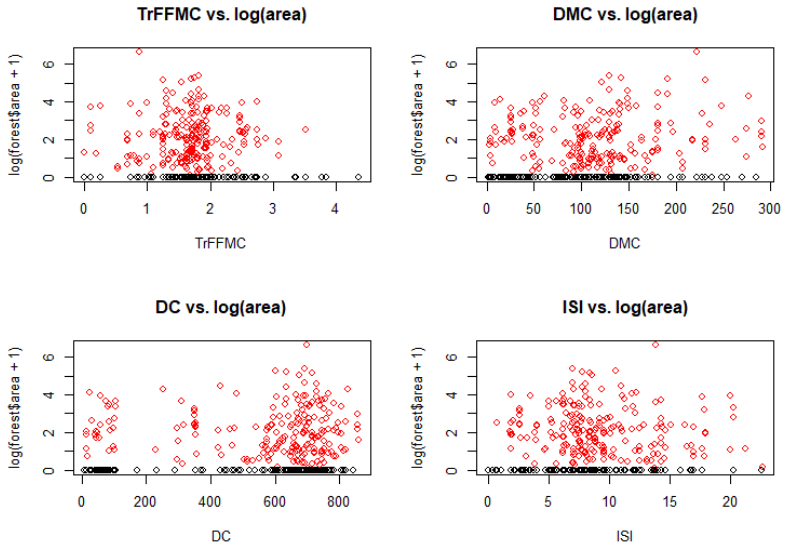
The VIF_k scores for this dataset follow:

Larea	TrFFMC	DMC	DC	ISI	Temp	RH
3.404234	2.385843	2.021183	2.633326	2.619569	1.712791	1.135073

These do not seem to be large enough to cause concern, despite the high marginal linear correlations.

A.1.3 Scatterplots Colored by Burn Area

Here we see the scatterplots of the continuous predictors, colored by zero versus nonzero burn area. What we are looking for is some separation on the x-scale between black and red dots. Since this is just another way of looking at the histograms in Section 3, we similarly observe that we do not find the separation we hope to see. It only appears that the frequency of the zero burn area points corresponds to the frequencies of the nonzero burn areas along the x-scale, indicating that as we move along the range of each of these predictors, we do not find even a fuzzy transition from zero burn area to nonzero burn area. Both zero burn areas and nonzero burn areas appear to occur with proportional frequencies throughout the ranges of these predictors, emphasizing the challenge of singal detection in this problem.



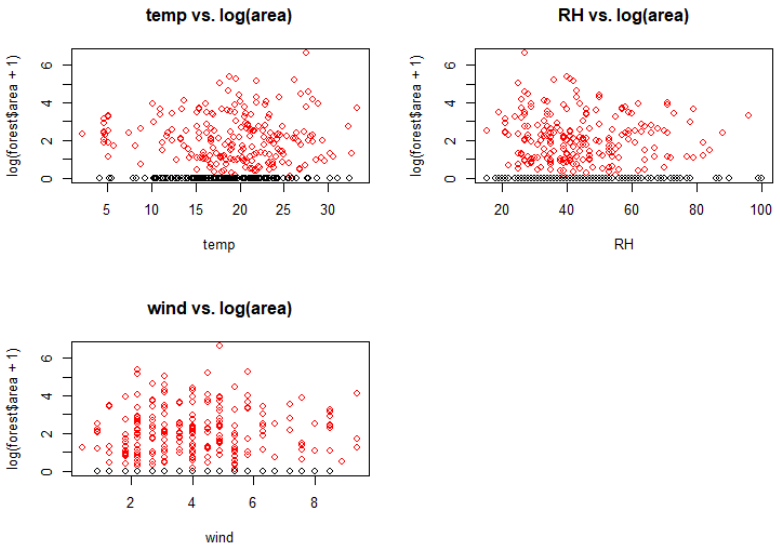


Figure 14: Comparisons of Distributions of Continuous Predictors Conditioned on Zero burn area and Nonzero Burn Area

Figure 15 shows the scatter plot matrix using different colors to differentiate zero values from positive ones. The lack of separation between both groups prevents us from identifying a clear predictor for the logistic regression.

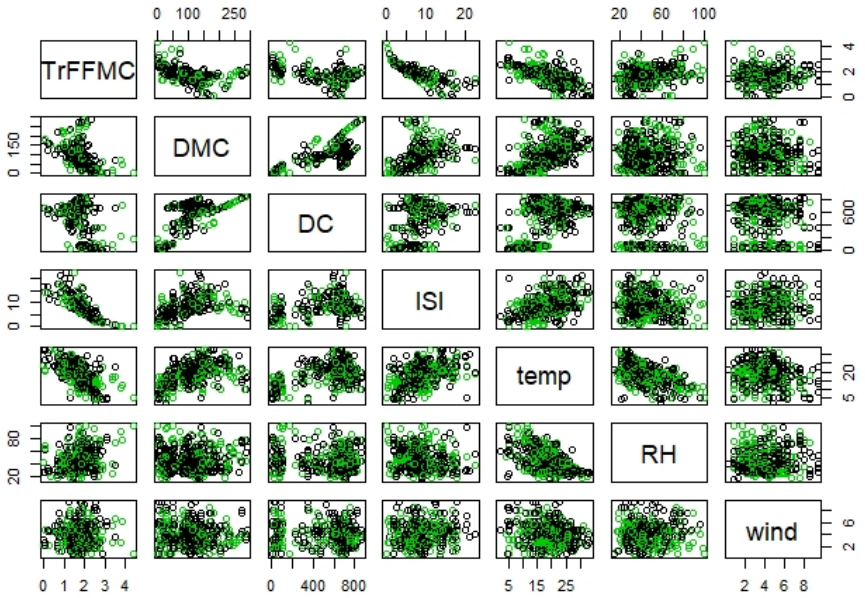


Figure 15: Scatterplot Matrix of the Transformed Continous Predictor Variables with Different Colors for Zero (green) or Nonzero Burn (black). Note the lack of separation here.

A.1.4 Histograms for Continuous Predictors Conditioned on Zero and Nonzero Burn Area

Since we are interested in classifying zero-burn area fires versus nonzero burn-area fires, we plotted 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. The modes of the distributions shown in 16 do not appear to be separated, except in the case of **TrFFMC** (slight), **DMC**, and **temp**. However because there is not any clear distinct visual separation, we ran two-sample t-tests on each of the continuous predictors to determine whether any of these means significantly differed from one another. **TrFFMC**, **DMC**, and **temp** all significantly differed at the $\alpha = .10$ significance level (but not at $\alpha = 0.05$). Since the distribution of **DC** is highly non-normal, we instead used a Wilcoxan Rank Sum test, which also rejected with a p-value of 0.07461.

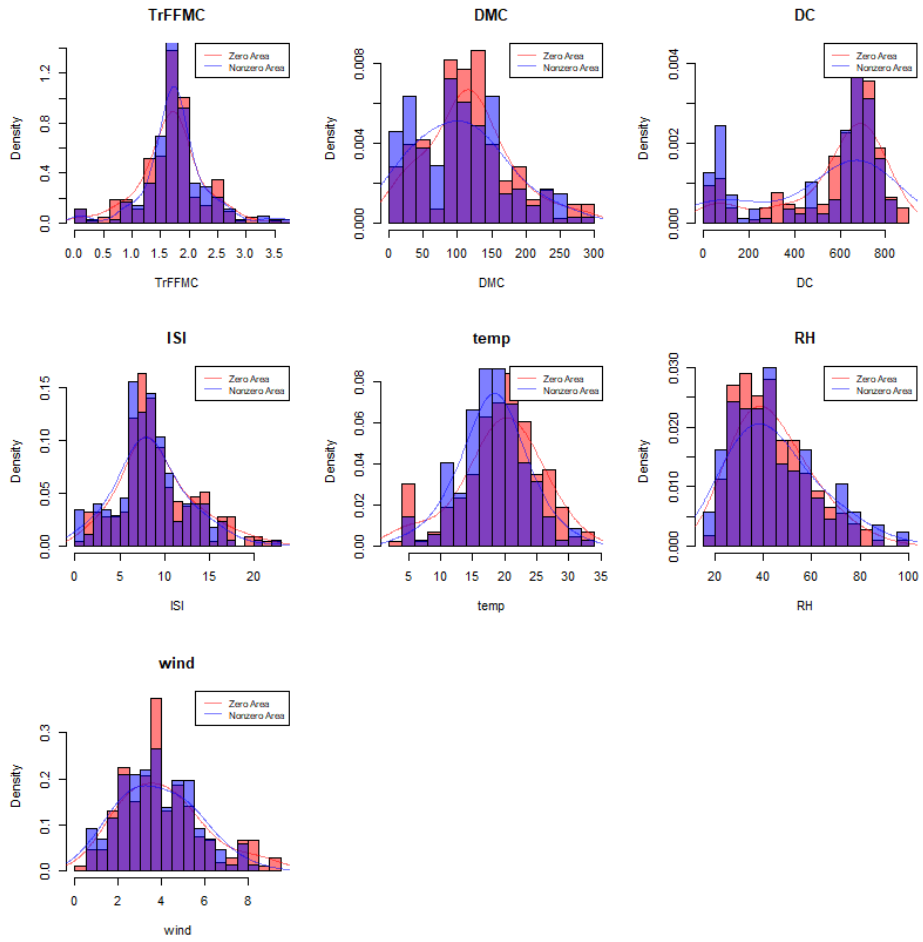
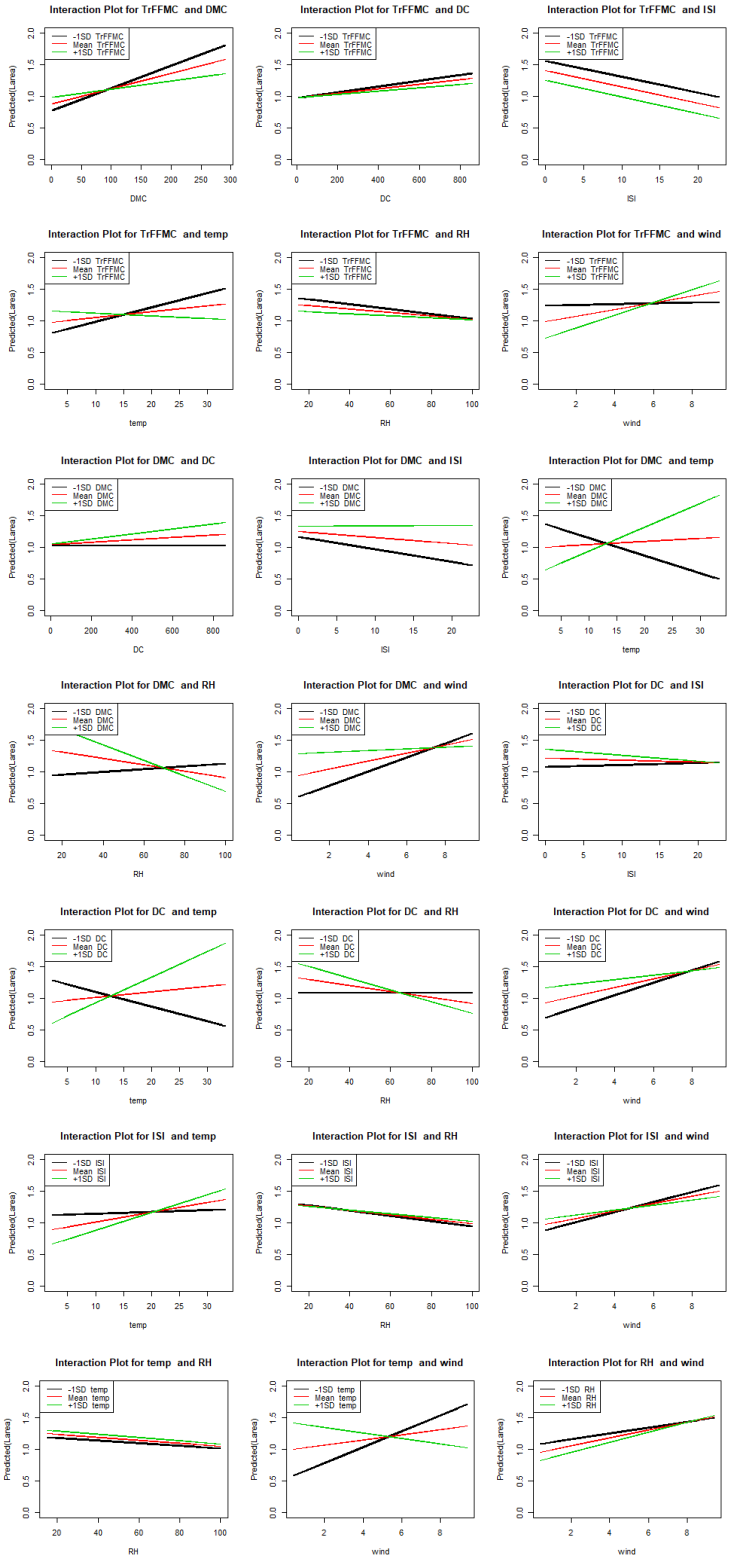


Figure 16: Comparisons of Distributions of Continuous Predictors Conditioned on Zero burn area and Nonzero Burn Area

A.1.5 Two-Way Interactions



A.2 Methods for Generating Interaction Plots

To generate the interaction plots given in the EDA section, we created 21 separate multiple linear regression models for the $\binom{7}{2}$ possible combinations of two continuous predictors. These were each fit with both predictors as well as the interaction term. We then conditioned on the values of the first predictor by predicting from the fitted model, while fixing predictor 1 at levels one standard deviation below the mean, the mean, and one standard deviation above the mean, obtaining fitted prediction lines for all three levels. This allowed us to see how the prediction of Larea using the second variable would change, given varying values of the first variable. A change in slope between these three prediction lines (i.e., non-parallelness) indicates that as the first variable changes, the rate of change of the second variable is changing with respect to the response, and hence interaction exists. For further details please reference our code which is available at `https://github.com/Catalysta/ForestFires`.

Because the coefficients of these individual MLR models are only estimates and do not reflect the true parameters, there is some variability associated with these plots which has not been quantified here. For this reason we only considered the most extreme interactions which were discovered through this process.

A.3 Data Transformations and Cleaning

A.3.1 Training and Testing Split

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.

A.3.2 Area

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

$$\log(\mathbf{area} + 1)$$

This transformation has the property that all zero values remain unchanged while the positive observations are presented in a logarithmic scale.

A.3.3 rain

Because there were only 7 nonzero **rain** values, we instead considered a binary predictor **rainbin** which was categorized as "1" for **rain** and "0" for **no rain**.

A.3.4 FFMC

As observed in Figure 17, the original version of the predictor FFMC was negatively skewed. Since log transformations are generally useful for positive skewness, we tried to normalize this predic-

tor by reflecting it, adding an appropriate value to make it positive, and finally applying a log transformation.

$$f(\mathbf{FFMC}) = \log(-\mathbf{FFMC} + \max(\mathbf{FFMC}) + 1)$$

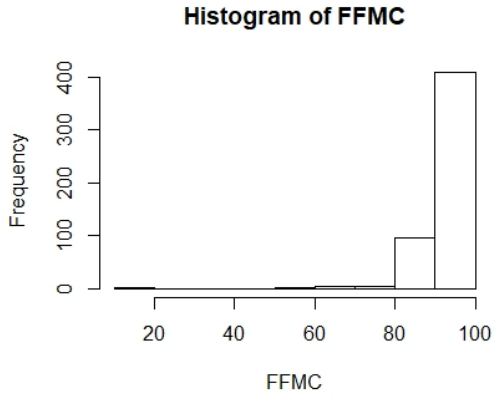


Figure 17: Histogram of FFMC Variable

Figure 18 shows the final distribution of this transformed variable.

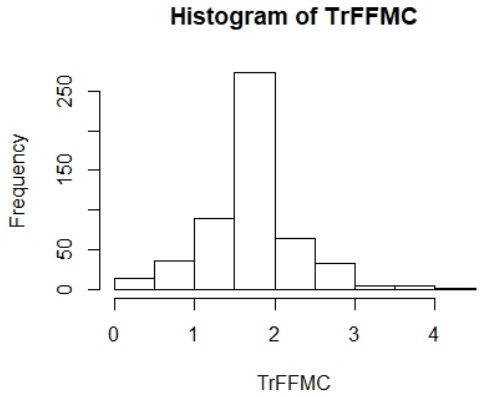
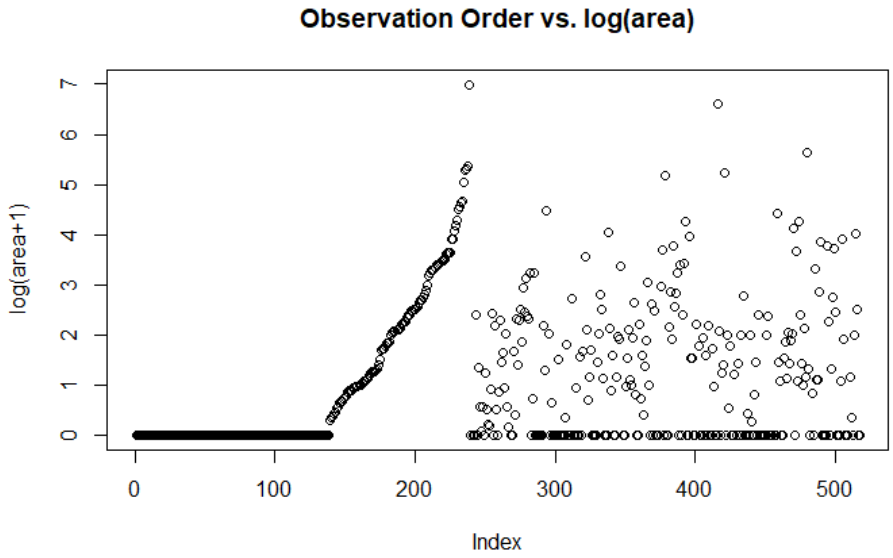


Figure 18: Histogram of FFMC Variable

A.3.5 Observation Order

We also noticed a trend in observation order with respect to **Larea**: this is likely due to the fact that this data came from two different data sets which were combined manually. [**1**] When combining the two data sets, it seems that the first set was sorted by **area**, while the second set was ordered chronologically. This prevented us from using a time series approach on the data set in its entirety.



A.3.6 Selection Criteria

The Mallow's CP criterion does not apply in the MLR scenario because we cannot assume that the full model exactly describes behavior of burn area. Also, since the available predictors do not explain much of the variability in the response, adjusted R^2 is being influenced more by the number of variables in the model than the amount of variability explained, in some cases turning negative. This is why we chose to compare models based on the BIC criterion.

A.4 Results of the Logistic Model

Following is the computer output from the Logistic regression model. We note only mediocre improvement from the Null deviance and a still-high AIC, indicating remaining uncertainty in the model fit.

```
Call:
glm(formula = large ~ ., family = "binomial", data = train.x)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6192  -1.2464   0.9231   1.0619   1.6245

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.2585155  0.4956536  -2.539   0.0111 *
DC           0.0008903  0.0004845   1.838   0.0661 .
temp         0.0259064  0.0209739   1.235   0.2168
wind         0.1301154  0.0597702   2.177   0.0295 *
rainbinTRUE -1.3880929  0.8897375  -1.560   0.1187
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 533.75  on 387  degrees of freedom
Residual deviance: 521.49  on 383  degrees of freedom
AIC: 531.49

Number of Fisher Scoring iterations: 4
```

Figure 19: Summary of the Logistic Regression Model Using **DC**, **temp**, **wind**, and **rainbin** as Regressors

A.5 Results of the 'Leaps' MLR Model

This model is fully stated as

$$E[Larea|X_1, X_2, X_3, X_4, X_5, X_6] = 1.582 - 0.8891X_1 + 0.008758X_2$$
$$-0.2024X_3 + 0.02296X_4 + 0.01400X_5 + 0.01562X_6 + 0.01599X_1X_6$$
$$+0.08724X_1X_5 - 0.0001629X_2X_4 + 0.0006927X_2X_5 + 0.003829X_3X_6$$
$$+0.02260X_3X_5 - 0.02182X_5X_6 - 0.002197X_4X_5,$$

where

$$X_1 = TrFFMC, X_2 = DMC, X_3 = ISI, X_4 = RH, X_5 = wind, X_6 = temp.$$

To further check diagnostics on this model, we note that the residual plots again look slightly skewed:

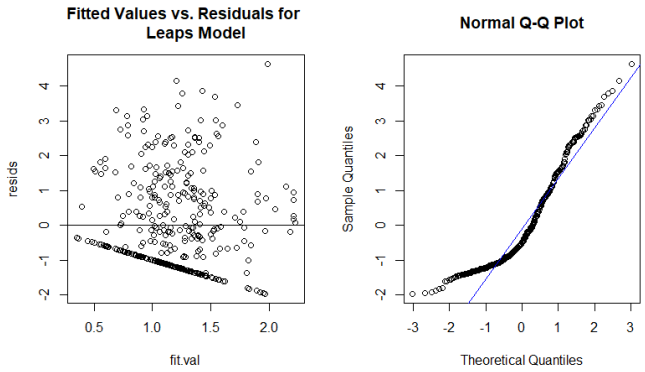


Figure 20: Residual Plot and QQ Plot from Leaps Model

If we, as in the case of the forward-selection model, condition on nonzero burn area, we see a slight improvement. However, different from the forward-selection model, we observe a bit of clumping on the right of the residuals which is a little suspect.

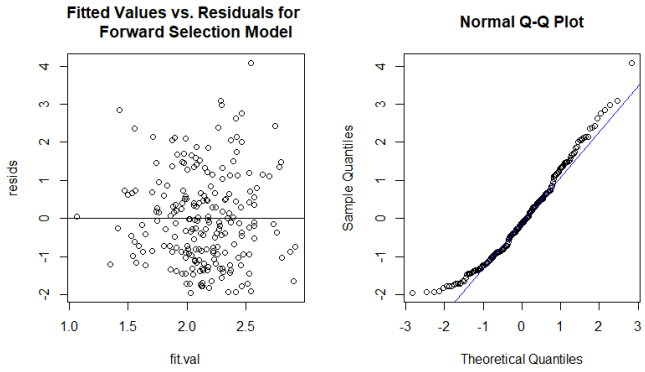


Figure 21: Residual Plot and QQ Plot from Leaps Model

The Durbin-Watson test does not detect autocorrelation, and the K-S test fails to reject with a p-value of 0.01299, so we cannot rule out normality of the residuals.

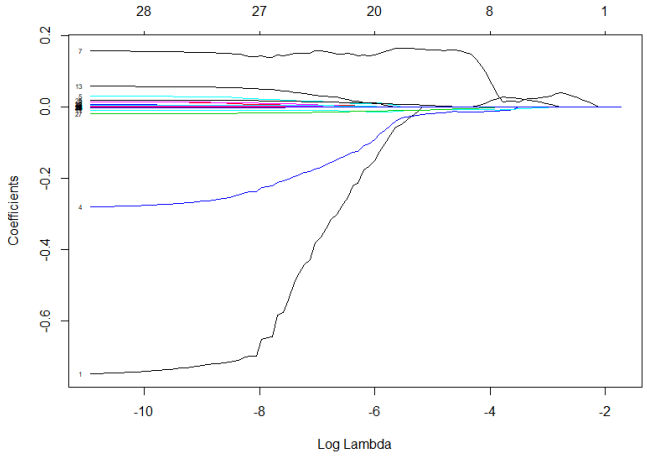
Despite the slight pattern in the residual plot, this model does not seem to severely violate the assumptions of independence of errors, homoscedasticity, and normality. However, the model only achieves an adjusted R^2 value of 0.02542. Therefore even though this model appears to offer improved predictive power over the mean model, it is important to remember that there are other variables not included in this model which probably explain more of the variation in burn area, namely method of containment and response times.

A.6 LASSO Model

We also fit a LASSO model using all predictors and two-way interactions. This model identified **wind** and **DMC:temp** as important predictors, and it shrunk their coefficients almost to zero using the "lambda min" rule. This model achieved an in-sample MSE of **1.946107** and PMSE of **2.028175**. Since this did not beat the **Leaps** MLE model that we selected, we did not include it in the main part of our analysis.

The LASSO coefficients, path, and CV error plots appear below.

```
29 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept)   1.053919e+00
TrFFMC        .
DMC           .
DC            .
ISI           .
temp          .
RH            .
wind          7.172880e-03
`TrFFMC*DMC`  .
`TrFFMC*DC`   .
`TrFFMC*ISI`  .
`TrFFMC*temp` .
`TrFFMC*RH`   .
`TrFFMC*wind` .
`DMC*DC`      .
`DMC*ISI`     .
`DMC*temp`    4.432168e-05
`DMC*RH`      .
`DMC*wind`    .
`DC*ISI`      .
`DC*temp`     .
`DC*RH`       .
`DC*wind`     .
`ISI*temp`    .
`ISI*RH`      .
`ISI*wind`    .
`temp*RH`     .
`temp*wind`   .
`RH*wind`     .
```



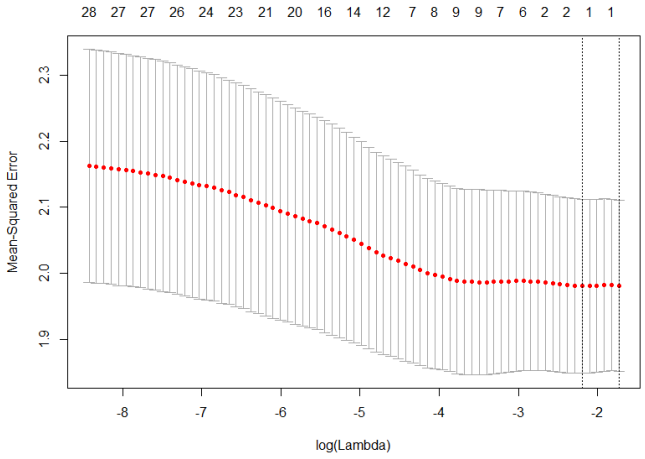


Figure 22: LASSO Coefficients, Path, and CV Error Plot.

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] **Dua, D. and Graff, C.**, *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science. [http://archive.ics.uci.edu/ml]. (2019)