Forest Fires Prediction

# Introduction

Forest fires cause great environmental damage while endangering human lives. Human surveillance to prevent fires can be costly and ineffective. Automatic meteorological stations are most often available and provide valuable data that can be used to create predictive models. The goal of this project is to forecast forest fires using predictive models based on the available information. All the programming is done in R.

# Literature Review

As human surveillance is costly and unreliable method to prevent forest fires, some automated fire detection solutions have developed over time: satellite scanners, infrared smoke scanners and meteorological sensors. While the first two have high costs, the local meteorological sensors can provide real-time data at low costs. Many different data mining techniques are used to predict forest fires and area captured by it. This work makes use of a publicly available data set containing information on the forest fires in Portugal’s Montesinho Natural Park collected over the period of 3 years. The main goal of this study is to build multiple regression model(s) to predict area at risk of forest fire and determine the key contributing features to forest fires. The goodness of the fit for the linear regression was measured using adjusted R-squared and the accuracy of a model was evaluated using the RMSE (Root Mean Squared Error) and MAD (Mean Absolute Deviation). The last two metrics have the useful property of being in the same units as the target variable and lower values indicate a better fit. Both MAD and RMSE represent a good measure of how accurately the model predicts the response, which is the most important criterion for fit when the main purpose of the model is prediction.

Due to the small size of the data set used (517 observations), a 10 K-Fold cross validation was implemented to validate predictive models [2]. Cross-validation has been suggested as an effective technique [3] [4] when the data at hand is too limited to perform split validation. The variable selection method Stepwise regression [5] was used to find the best linear model with optimal number of variables. Multicollinearity between predictor variables was assessed by means of the Variance Inflator Factor [6]. The interactions effect between predictors [7] was explored with the goal of improving the performance of the linear regression model. Two independent variables interact if the effect of one of the variables differs depending on the level of the other variable. Allowing for interactions turned out to be very helpful in improving the accuracy of the fitted model (see Table 2). In addition to Multiple Linear Regression, other data mining algorithms such as Regression Trees [8],[9] and Random Forests [10] were also implemented. As it will be shown, these two techniques didn’t prove effective at predicting the area burned based on the data at hand. As a separate task, a logistic regression [11] model was built using the same data set to predict whether the fire area would fall into one of two different categories: small or large fire. For this task the dependent variable area of fire was categorized into small and large and recoded as a binary target (1 or 0). The Logistic regression model assigns a probability to each observation based on the main predictors. The performance of the model was assessed by means of the area under ROC curve [12].

All data manipulation, processing, visualization and modeling was done using the open source software R version 3.3.2

# Data set

This work focuses on the forest fire data set publicly available at the following link: http[://archive.ics.uci.edu/ml/datasets/Forest+Fires](https://archive.ics.uci.edu/ml/datasets/Forest+Fires)

The data set contains information about fires in Montesinho Natural Park in Portugal that has been collected over the period of 3 years from January 2000 to December 2003. The data was compiled using two sources: 1. data collected by the local fire inspector containing the area burned by the fire, along with time, date, spatial location, vegetation involved and four component of the Fire Weather Index (FWI) and 2. weather data collected by the local meteorological station on the day of the forest fire. The two sources were merged into one data set containing 517 observations and 13 variables. The vegetation information was missing for the 80% of the information and it was omitted. For the temporal values it was decided to keep only month and day of the week. The variables contained in the data set can be divided into 5 groups:

1.Fire location annotated within 9 X 9 grid of the park coordinates:

**X**- values 1 to 9

**Y**- values 1 to 9

2.Month and day of the week

**month** - values jan to dec

**day**- values mon to fri

3.Components that make up the FWI

**FFMC -** Fine Fuel Moisture Code - a numeric rating of the moisture content of litter and other cured fine fuels

**DMC**  - Duff Moisture Code – a numeric rating of the average moisture content of loosely compacted organic of moderate depth

**DC** - Drought Code - a numeric rating of the average moisture content of deep compact organic layer

**ISI**  - Initial Spread Index - a numeric rating of the expected rate of fire spread;

4.Weather features

**temp** – temperature in Celsius scale

**RH** - relative humidity in %

**wind** - wind in km/h

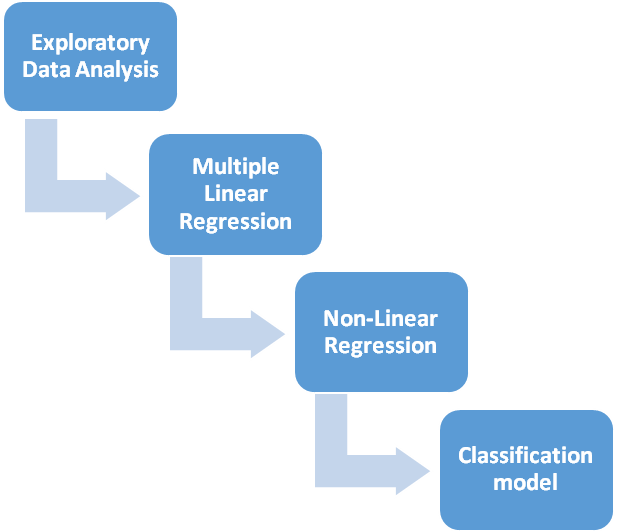
**rain** - rain in mm/m2

5.Area burned – response variable

**area** – area burned by fire in squared meters (m2).

It is worth mentioning that the metadata associated with the dataset indicates that if the area burned was less than 1 ha (100 m2) it was recorded as zero. Also, out of the four weather variables, temperature, RH and wind are captured by the station sensors when the fire was detected, and rain variable denotes accumulated precipitation within the previous 30 minutes. The data set contains no missing values. A good portion of the data contains small fires (smaller than 100 m2 as mentioned above) which are all recorded as area equal to zero.

# Approach



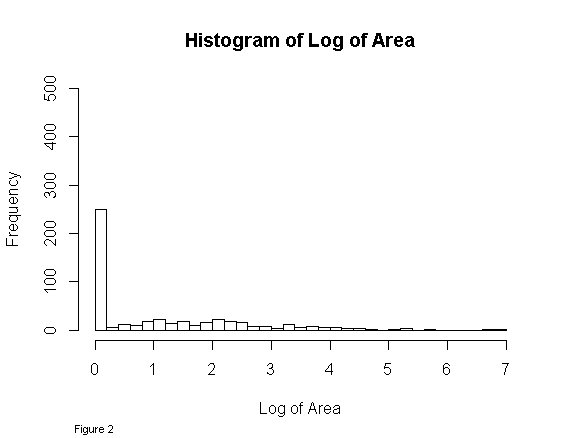
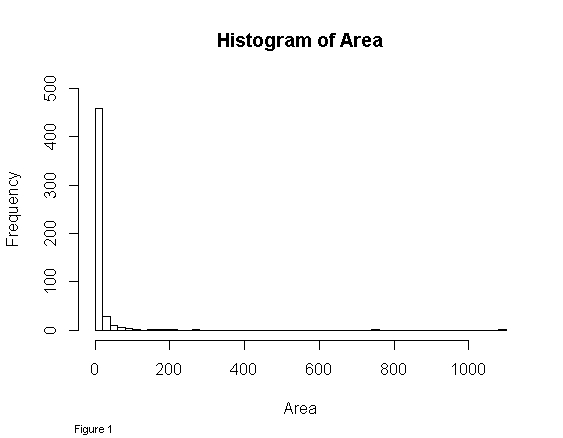
# Step 1: Exploratory Data Analysis

The data set has 517 observations and 13 variables. There are no missing values in any of the variables.

## Data Challenges

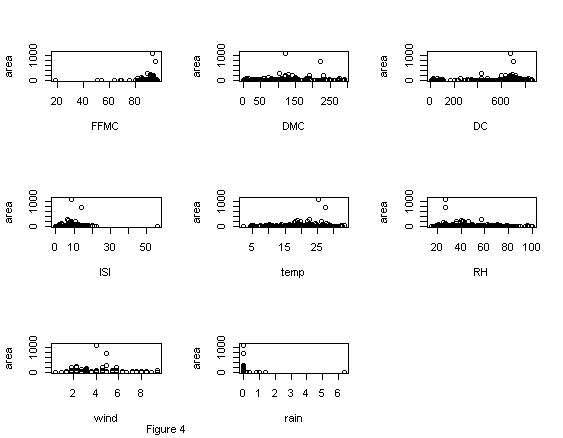
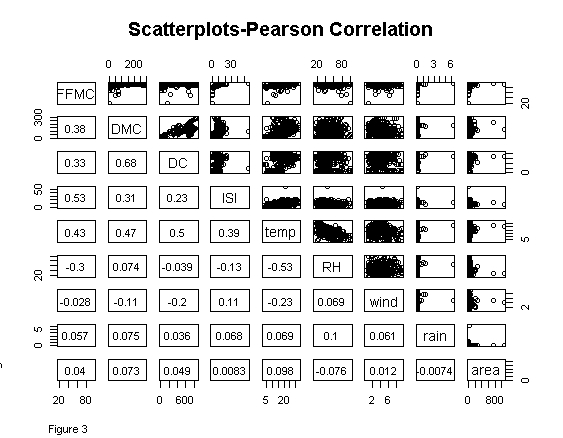
**Skewness**

Approximately 48% of the observations in the data set contain a value of zero value for the response variable (area), which is a continuous variable. Those values of Area=0 are valid data as explained in the previous section. The distribution of the dependent variable, area, is highly skewed to the right as per Figure 1. To reduce skewness, a log transformation log(area+1) was applied. The effect of this transformation can be seen on Figure 2, where the shape of the histogram for log(area+1) looks closer to a normal distribution than the non-transformed variable, area.



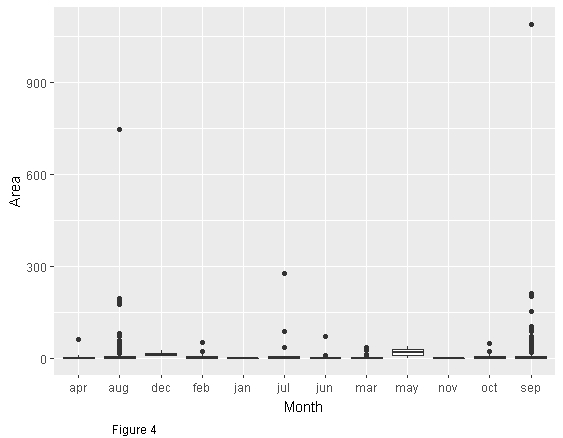
**Correlation among continuous predictors**

As shown in Figure 3, the Pearson correlation between log of area and each independent variable is extremely weak (less than 0.07 in all cases), which poses a great challenge for building a predictive model. As per Figure 3 (the scatter plot of Figure 3 can be zoomed in when executing ‘Data Analysis.R’ code script from R studio), the wind and DMC are the predictors with the highest Pearson correlation with log of are (0.067) which is very low. Moreover, non-linear patterns are not obvious to identify from the scatter plots. The correlation matrix in Figure 3 also reveals that there may be some collinearity among some predictors. For example, DMC and DC show a Pearson correlation of 0.68 and Temperature and Relative Humidity have a correlation of 0.53. For this reason I have assessed the multicollinearity among the continuous predictors using the Variance Inflation Factor, which is reported in the next section.



## Outliers and missing values

There are no missing values for any of the variables. The box plot Figure 4 shows two outliers however they are valid values and have been kept for the analysis.



## Categorical predictors

I have grouped the nominal variable month into four seasons to reduce the number of dummy variables and reduce the number of parameters in the model overall. For the same reason the variable day was collapsed into two levels: weekday and weekend. The park map X and Y coordinates are to be used jointly to uniquely identify the geographical location of a fire location. Also with the goal of reducing the number of dummy coding, I have joined the X-Y coordinated and simplified the map grid from 9X9 to a 3X3 grid.

## Multicollinearity

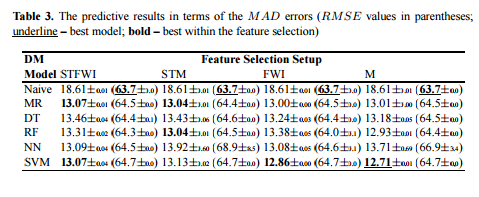
Multicollinearity between the numeric independent variables could be a problem when performing linear regression. From the scatter plots (Figure 3), between independent variables it appears that DMC and DC might raise a concern as the Pearson correlation between the two variables is 0.68. Other variables with higher person correlation index are ISI and FFWI (0.53) and temperature correlation with RH (-0.53), DMC (0.47) and FFMC(0.43). To identify collinear predictors, the Variance Inflator Factor was used by means of the R “usdm” package to verify if the correlations identified above are of any concern. As shown on Table 1, the results of the VIF for all predictor variables are less than 5, which is considered a threshold for moderate collinearity. With this information we can conclude that there is no need to remove any inputs on grounds of multicollinearity. The data set at this state is ready for regression analysis.

|  |  |
| --- | --- |
| **Variables** | **VIF** |
| FFMC | 1.695255289 |
| DMC | 2.330688362 |
| DC | 2.078204513 |
| ISI | 1.578257622 |
| temp | 2.661897308 |
| RH | 1.899988696 |
| wind | 1.140609509 |
| rain | 1.044800721 |

Table 1

# Step 2: Linear Regression

For all my regression models I will be using the results from Cortez and Morai’s [1] paper below as a benchmark.



The Cortez and Morai’s paper created four combinations of predictor variables:

1.STFWI - – using spatial, temporal and the four FWI components (groups 1,2 and 3 from the data set description)

2. STM –using spatial, temporal and four weather variables components (groups 1,2 and 4 from the data set description)

3. FWI – using only the four FWI components (group 3 from the data set description)

4. M – with the four weather conditions (group 4 from the data set description)

For each combination of predictors a model was validated with the RMSE and MAD measures. Their best linear model was using STM predictors with RMSE of 64.4 and MAD of 13.04 which I will use as my benchmark for the linear models.

## 1. Linear model using numeric predictors

My approach is not to use the same groups Cortez and Morai’s but instead to build the best possible linear model using any set of predictors. To begin with, the first linear model was built using all numeric predictors (FFWI and weather variables) to see how that would compare to the authors linear models. This model produced a very low R-squared (0.01988) and adjusted R-squared (0.004446), which was expected considering the very low correlation between the predictors and the area. The RMSE and MAD values were close to the paper (RMSE 64.44362), (MAD 12.93822). Applying the stepwise variable selection to the above model, the best predictors were found to be RH, FFMC and wind with the R-squared of 0.0142 and slightly improved adjusted R-squared of 0.00848. The results for the accuracy metrics: RMSE (64.45021) and MAD (12.9581) did not show any improvement over the model that is using all fire and weather predictors.

## 2.Linear model using classified categorical predictors

In this step I have added the spatial and temporal categorized predictors: park coordinates, season and week day. As in the previous step I did the linear model using all predictors and then performed the stepwise variable selection thereafter. The model using all the predictors improved over the previous one with R-squared of 0.05259 and adjusted R-squared of 0.01835. Performing the stepwise regression it was suggested not to use the RH and FFMC variables and to keep the wind variable. Other than the wind variable, the temperature, DMC and season were selected as significant variables. The best predictors as per this model are DMC, wind, temp and season. Adjusted R-squared for the model has improved over the previous model to 0.02638. Testing the model on the entire data set the RMSE value was 64.39361 and MAD 12.89218 which is better than the best linear models reported in the reference paper.

## 3.Polynomial regression

In an attempt to improve the fit to the data and to account for non-linear relationships between the predictors and the outcome variable, area burned, I have introduced polynomials into the regression model. For each numeric predictor I have included terms with progressively higher degrees (quadratic, cubic, etc) and tested their statistical significance. This way I found that temperature, rain and wind lead to a better adjusted R-squared when polynomials of second and third degree are present in the model. Based on this information, I have modified the best model from the previous step by having terms for wind, temperature and rain up to the 3rd power. This model resulted in an adjusted R-squared of 0.03834 with lower values for RMSE (64.34795) and MAD (12.78248). I have tried adding polynomials for other variables and only DC made an improvement. With the DC predictor in the model the adjusted R-squared was 0.04215, and the test results were a little bit better for the accuracy of the model(RMSE: 64.30408 and 12.75965).

## 4.Linear model with the predictor variables interaction effect

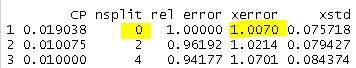
A further effort improve the linear model consisted in considering interactions between the predictor variables interaction effect. To this effect, I have tested the interaction effect between all predictors (second order interactions) for the best model obtained in the second step (table 2; model 2a - DMC, wind, temp and season) and found that it improved over the original model having adjusted R-squared of 0.03552, RMSE of 64.35336 and MAD of 12.79324. This model is suggests that there is an interaction of DMC with temperature and DMC with season. In the next model I have included all the predictor’s interactions and stepwise elimination got the best results so far. The values R-square was 0.1533 and adjusted R-squared 0.0724. The RMSE and MAD values: 64.06728 and 12.4756. The drawback with this model is that includes too many interactions, that model interpretation becomes overly complicated.

# Step 3: Recursive Partitioning: Tree-based Modeling

## Regression Trees:

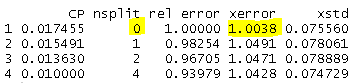
Decision Trees is a known classification technique used to predict an outcome for a categorical response variable based on the set of predictors. The response variable can have minimum of 2 (yes and no or binary 0 and 1) or more classes. Regression Trees is a regression modeling technique used to predict numeric continuous response variable. Both methods form a tree structure by breaking down a dataset into smaller and smaller subsets with similar values while at the same time decision or regression tree is incrementally developed. The fully grown tree has decision nodes and leaf nodes. Regression trees start by finding data splits for each independent variable. The best predictor variable is selected as the as the topmost decision node. At each split point, the error between the predicted and the actual values are calculated as Sum of Squared Errors. The errors between the variables are compared and the variable with the lowest SSE is chosen as the root node. The process is continued with the full grown tree having terminal leafs. The predicted value in a terminal leaf is the average value of all the observations that belong to that leaf. As in the linear model section, the log of the area is used to create regression trees and transformed back to its original values to calculate the RMSE and MAD in the cross validation script.

The first tree that I have created was using the weather variables (wind, relative humidity and temperature) to predict area burned. The weather predictor model had the RMSE value of 64.49935 and MAD 13.15522 which is in line with the reference paper. This modeling was obtained using the rpart package with minbuckets set to 8, indicating the minimum number of observations per leaf node. Changing the number of buckets did not affect the model accuracy significantly. The complexity parameter table for a model indicates the cross-validated (xerror column) error for each number of splits. For a full grown tree, the lowest error for each number of splits indicates the complexity parameter at which the tree should be pruned. Pruning is a technique used in tree modeling to avoid overfitting and to ensure that the tree-based predictions generalize well to new data. The table below shows that the cross-validation error, xerror, never reaches a minimum, in fact it keeps increasing with the number of splits instead of decreasing. This indicates that the optimal cross-validated pruned tree model is a tree with no leafs. In other words the regression tree that we obtained is one that does not generalize well to new data and is not useful for prediction. In fact the best prediction in this case would be the average value of outcome variable, log(area+1). I have tried different values for minbucket and the result didn’t change.



Just to demonstrate how pruning is done in R, I have pruned the tree to the he suggested size in my R code, but did not validate it or used it for any reference.

The second tree I created was using all the predictor variables. Its performance increased slightly over the weather tree model having RMSE of 64.391 and MAD of 13.08937. For this tree it is also suggested to prune the tree to zero nodes as per results below.



I have tried many different trees using different number of predictors and different minbucket values and for all of them it was suggested to use the number of splits equal to zero. It is worth mentioning that rpart R package would check automatically for interactions between variables automatically. My conclusion, based on the suggested number of splits equal to zero is that regression trees obtained for this data set do not perform well on unseen data and therefore I will not use this technique as a predictive modeling technique for this data set. This is most likely related to the very low association between the predictors and the outcome variable as mentioned in the data analysis step.

## Random Forest:

Random Forest as regression data mining technique is an ensemble learning method. Ensemble learning generates multiple algorithms and puts them together to make a prediction. Decision/regression tree use the entire training data set to create a model, Random Forrest selects randomly a number K of data points (subset) from the training data set and builds a Tree-based model. This process is repeated N number of times and each time a new tree is created using random K of data points (with replacement) and possible different combination set of predictors. To make a prediction for a new data point, all the Decision Tree models are used. Each model will predict a value and the average of all the predicted values will be used as the final predicted value for that data point. Individual models in the Random Forest will be less accurate considering they are based on subset of training data compared to Decision Trees where the entire training data is used. Having a large number of these models and getting their average when predicting is the advantage of Random Forests and is expected to outperform predictions done by a single regression tree.

Using the randomForest R package, I built a random forest model for the weather predictors using wind, rain, temperature and relative humidity and obtained an RMSE value 64.43 and MAD 12.93 which is in line with the paper. Full results from the random forest are presented in the results section.

# Step 4: Classification model

## Logistic Regression:

In this section my task is to create a Logistic regression classifier model to predict if the area captured by the fire is going to be large or small. Logistic Regression is a data mining technique used to predict a categorical response variable. If the response is binomial (there are two possible outcome values: 0 and 1) than it is a binomial Logistic Regression. Multinomial logistic regression is used when the response variable has three or more possible classes. To perform logistic regression on the forest fire data set, I have converted the response variable area to a categorical binary variable by splitting the data into 2 classes: class 1 includes large fires that cover area over 0 ha, class 0 includes small fires that cover area of 0 ha (as mentioned in the data set description if the area burned was less than 100m2 it was recorded as a zero value). The goal is to predict if a fire is going to be large or small based on the values for the independent variables. To evaluate performance of logistic models I have used the (AUC), Area Under the Receiver Operator Curve (ROC). ROC is a way to visualize performance of a binary classifier by plotting the True Positive rate on the Y axis and False Positive rate on the X axis for every possible classification threshold which range from 0 to 1. AUC values range from .5 meaning the classifier is not any better than a random guess, to 1 meaning the classifier’s accuracy is 100%

To build the first classifier model I have ran the stepwise logistic regression starting with all the predictors. The suggested predictor variables for this the selected model were DC and wind. To evaluate the model I have calculated AUC value using 10 fold cross validation with 30 repeats. The model’s AUC value was 56% which generally is not a considered a good classifier. In the next step I have cross-validated model using all predictor variables excluding rain only. The accuracy of this model was improved over the previous one to AUC value of 63%. In the next two steps I have introduced interactions and quadratic terms which proved to improve my linear models. The best model had AUC value of 70% which is considered pretty good considering the low correlation of the predictor variables and area.

# Results

## Linear regression:

Table 2 contains the summary of the linear regression models. It shows gradual improvement in the Adjusted R-squared columns as different features were selected (models 1 to 2a). The trend had continued after the introduction of polynomial terms (models 3 and 3a) and variable interactions (4 to 5). The model with the highest R adjusted-value of 0.0744 is a mix of interactions and one polynomial term. Testing each model using 10 fold cross validation where the RMSE and MAD values were following the adjusted R-squared, meaning as the adjusted r-squared increases for a model the values for RMSE and MAD would decrease.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Model description** | **Model name** | **R-Squared** | **Adjusted R-squared** | **RMSE** | **MAD** |
| 1 | Numeric variables (weather and fire index) | lm.wf | 0.01988 | 0.004446 | 64.44124 | 12.93817 |
| 1a | Numeric stepwise (selected significant variables: RH, wind and RH) | stepB | 0.01425 | 0.008485 | 64.44783 | 12.95806 |
| 2 | All variables including factor variables: season, wday and coordinates | lm.all | 0.05259 | 0.01835 | 64.33444 | 12.85998 |
| 2a | Stepwise of the previous model (selected significant variables: DMC, wind, temperature and season) | stepC | 0.0377 | 0.02638 | 64.39118 | 12.89213 |
| 3 | Polynomial terms applied on the model above plus rain^2 | lm.poly | 0.06256 | 0.03834 | 64.34547 | 12.78232 |
| 3a | Model above plus DC^2 | lm.polyDC | 0.06999 | 0.04215 | 64.30157 | 12.75959 |
| 4 | Interactions on the model 2a | stepD | 0.05421 | 0.03552 | 64.35089 | 12.79311 |
| 4a | Interactions with all variables | stepE | 0.1533 | 0.07245 | 64.06452 | 12.47535 |

Table 2

## Recursive Partitioning: Tree-based Modeling:

## Decision Tree Modeling:

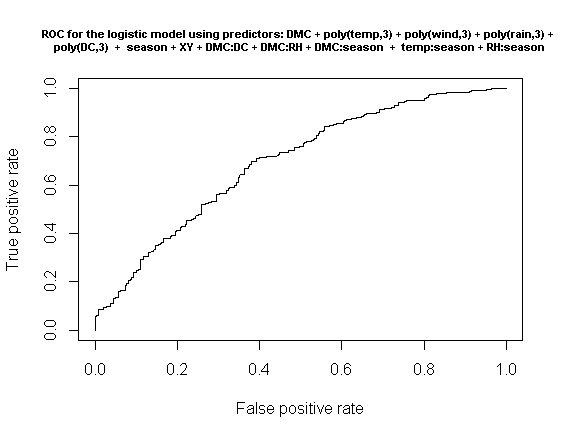
As described in the previous section, the decision tree models obtained for this dataset were such that did not generalize well to unseen data. Indeed, using different sets of inputs and settings lead in all cases to cross-validated tree with no leaves. I concluded that the very weak association between our predictors and the outcome variable, area, turns out to be too big of a challenge for this algorithm.

## Random Forest Modeling:

Thanks to the ensemble nature of random forest, better results were obtained with this algorithm than with decision trees. As described in the previous section, my random forest results for the weather predictors, temperature, rain, relative humidity and wind reproduce the results reported by Cortez and Morai’s [1]. Moreover, when I used all the variables the Random Forest predictive model improved in terms of the RMSE, which was 64.39, compared to 64.43 for the weather variables, with MAD equal to 13.16, slightly greater than 12.93, the value obtained for the weather variables. With the aim of increasing performance I increased the number of trees in the forest to 500, which increased considerably the CPU runtime, however, the results did not show noticeable improvement. There are other settings that we could “play with” to try to improve performance, such as the number of inputs selected at each split, however this was not explored due to time limitations.

## Logistic Regression:

As for the logistic regression, the best model in the logistic regression was using the mix of polynomials and interactions. The model had an AUC value of 70% as per ROC plot below



# Conclusions

Predicting forest fires and area burned would be very beneficial to human society. Predicting and preventing forest fires would save human, animal and plant lives and reduce greenhouse effect which is negatively affecting our lives in all aspects. The goal of this work was to create effective predictive model(s) for Portugal’s Montesinho Natural Park that can also be applied to other parks globally as the weather and Fire index measures are becoming widely available at cheaper costs. Building a highly performing predictive model proved to be very challenging. On the one hand, almost half of the observations are reported in the dataset as having the response variable, area burned, equal to zero. On the other hand, all the predictors showed very weak association with the response variable and the Pearson correlation was close to zero overall. Not surprisingly, the attempted linear models with only linear terms on the predictors yielded close to zero adjusted R squared. This led me to allow for higher complexity in the model. This way, I succeeded to improve the adjusted R-squared significantly by introducing polynomials and variable interactions. With polynomial terms and interactions at play I was able to obtain a cross-validated linear model with higher accuracy (smaller RMSE and MAD) than any of the linear models reported by Cortez’s and Morai [1].

The tree-based models (regression tree and random forest) have not improved over the linear models and in the case of regression trees did not result in predictive models that would generalize well to new, unseen data.

The Logistic regression classification model proved to be a very good model to predict whether the fire would be small or large. The 10-fold cross-validated value (repeated 30 times) of AUC was 70%. This model can be useful in a way that would allow park and government authorities to better plan the resources in advance depending on whether the fire has been predicted to be large or small.

Thanks to the challenges inherent to the dataset that I used, this project resulted in a great learning experience for me as it gave me an opportunity to apply different data mining techniques and to assess their performance, all while using such a powerful tool as the open source software R.

References:

[1] [Forest Fires Data Mining](http://www3.dsi.uminho.pt/pcortez/fires.pdf)

[2] [CVTools Package in R](https://cran.r-project.org/web/packages/cvTools/cvTools.pdf)

[3] [Cross Validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics))

[4] [10-Fold Cross Validation](https://www.openml.org/a/estimation-procedures/1)

[5] [Stepwise Variable Selection](http://www.statmethods.net/stats/regression.html)

[6] [Multicollinearity and Variance Inflator Factor](https://onlinecourses.science.psu.edu/stat501/node/347)

[7] [Interpreting Interactions in Regression](http://www.theanalysisfactor.com/interpreting-interactions-in-regression/)

[8] [Regression Tree Analysis using Rpart package](http://www.statmethods.net/advstats/cart.html)

[9] [Xerror in CP table](http://stats.stackexchange.com/questions/13471/how-to-choose-the-number-of-splits-in-rpart)

[10] [Random Forest Regression in R](https://cran.r-project.org/web/packages/randomForest/randomForest.pdf)

[11] [Logistic Regression](https://onlinecourses.science.psu.edu/stat504/node/149)

[12] [ROC Curve](http://thestatsgeek.com/2014/05/05/area-under-the-roc-curve-assessing-discrimination-in-logistic-regression/)