Burned Area Prediction Project

# Introduction / about the data

Wild fires destroy millions of hectares of forest every year. Forest files are costly and dangerous. Fast detection and response is the key to mitigating their destruction. In this project we use weather and meteorological data from the northeast region of Portugal to predict the burned area (or size) of forest fires.

## About the data

1. Source: Forest Fire data set in Montesinho natural park, originally from  [A Data Mining Approach to Predict Forest Fires using Meteorological Data](http://www3.dsi.uminho.pt/pcortez/fires.pdf)
2. Time period: January 2000 ~ December 2003
3. Attributes: Spatial (2), Temporal (2), FWI index (4), Weather conditions (4), Area (1)

# Methods & variable selection

## eda & feature engineering

We began an Exploratory Data Analysis with the fire data set. We created several small hypotheses about relationships between the burned area and some other features to try to find out what features are significantly impact on the burned area. We did that for Coordinates, Temp, Wind, Relative Humidity, Rain, and FFMC, but were not able to find any clear evidence of correlation between the burned area and those features in the data. Next, we created two new features, Is\_summer and CatWind, based on features of Month and Wind. Upon visual inspection there appear to be a relationship between the burned area and two new features, then we decided to keep them in the data. Therefore, in total, there were 15 variables before using pre-processing techniques to prepare the data set for modelling.

## Pre-proessing

In the pre-processing, first we binarized all the categorical features in the data set which are Month, Day, CatWind, X, Y, and Is\_summer. After binarizing, there were 49 variables left. Utilizing caret package, we also centered and scaled our data to make sure all numerical features on same scales. Additionally, for training data, we also removed 13 Near Zero-Varicance predictors to improve the stability of the model fit. Lastly, in order to reduce concerns of multicollinearity or correlated predictors, we removed 3 variables that had a correlation over 0.75. In total, after pre-processing, there were 33 variables left in the data set.

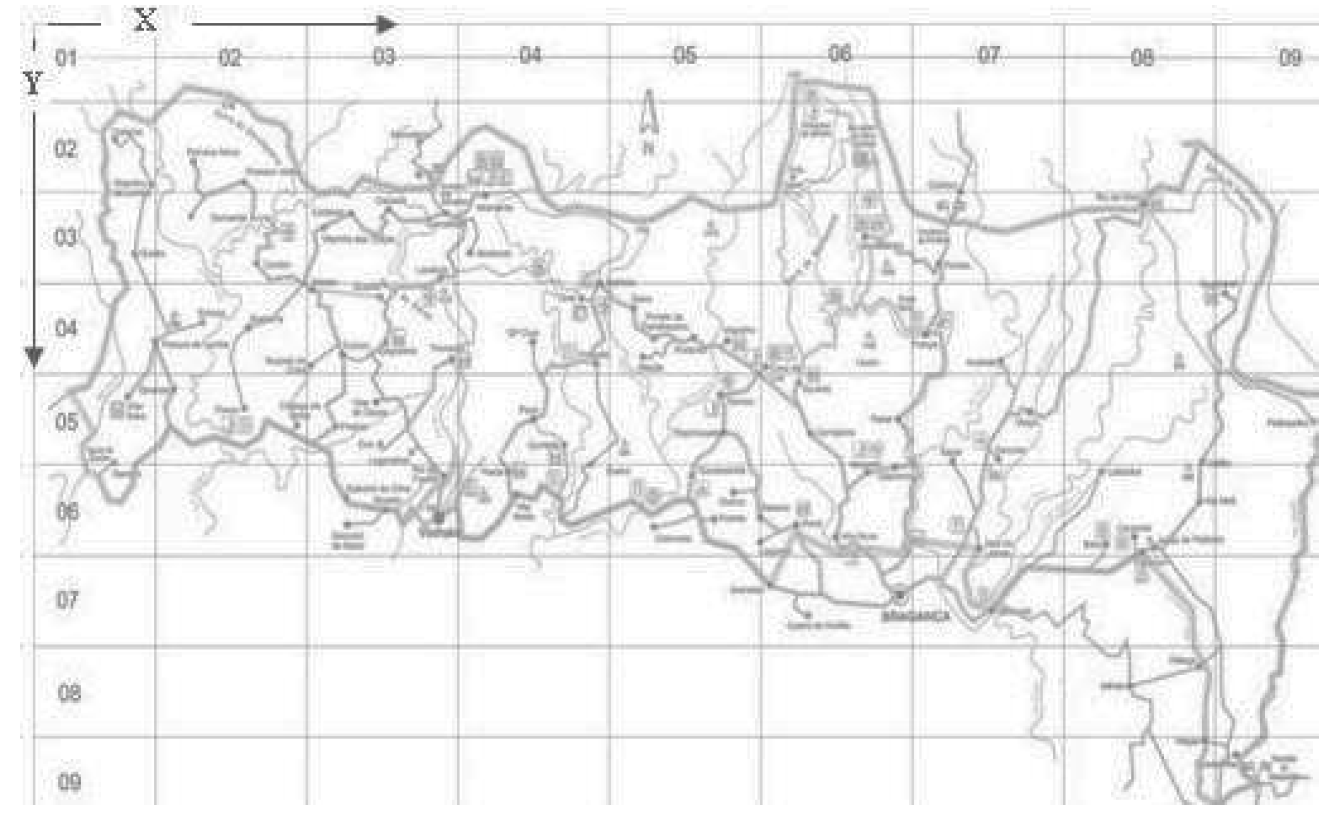
# Variable and model selection

Given the number of variables (i.e. 33 variables) available to us and limited knowledge about what specifically drives burned area, we employed a number of best subset variable selection algorithms to inform our selection of model features. The selection metric of measurement used for our analysis was the root-mean-squared-error (RMSE). The result output and most important variables can be seen represented in graphs within the appendix. 5 models that we created in this work were full model, forward model, random forest model, ridge model, and lasso model. In our case forward selection has the best average RMSE so we could select it for the final model.

Upon running these algorithms, we find that a number of the most important features are consistent across methods, though the forward selection produces the lowest RMSE on average. The most significant predictors appear to be *temp*, *X3* and *DMC.* Finally, it is important to notice that despite our selection of variables being set to a max of 20, the feature importance drops rapidly after the top 14. Therefore, we will only use the top selected features in our final model.

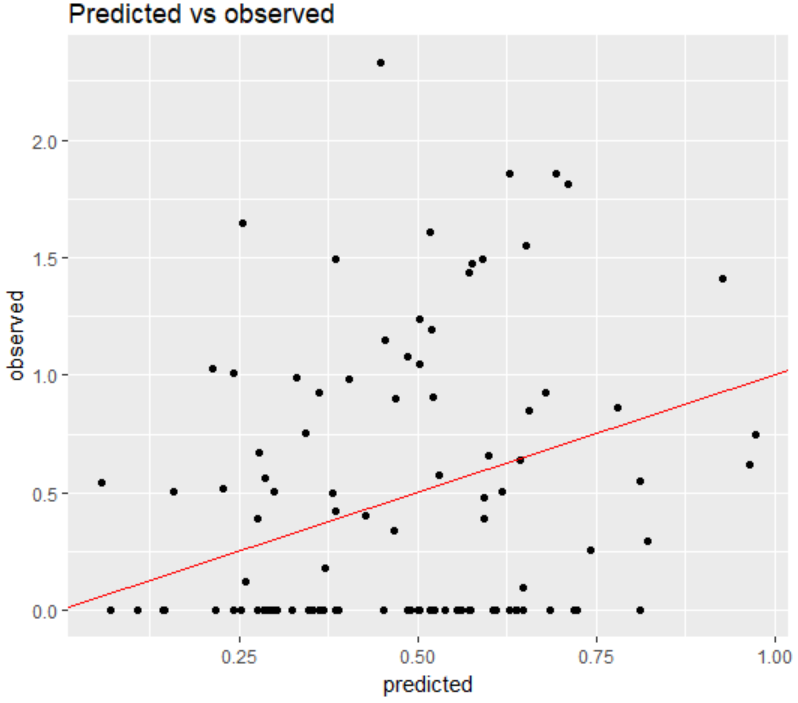
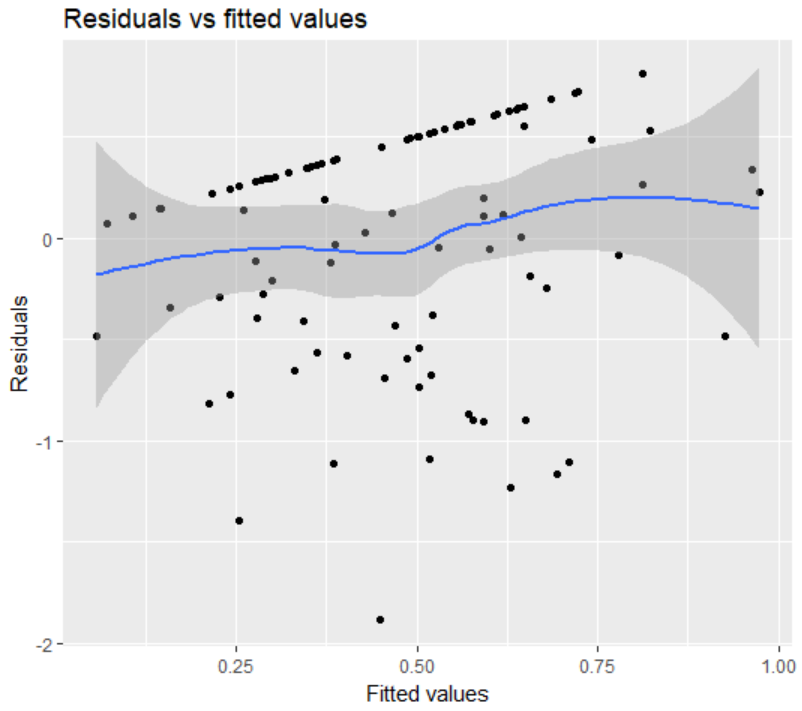
# Conclusion & recommendations

Overall, temperature, spatial location at X=3, and moisture content of shallow organic layers are the largest three indicators of the burned area of forest fire. In order to build proactive response, accurate and timely weather forecast should get more use for predicting forest fire danger. We should also pay close attention to the spatial location at X=3 in the map. In this area, even small fires should be extinguished in time so as not to affect other areas and develop into larege fires in the whole park. Additionally, upon our model, compared with the moisture content of surface litter (FFMC) and deep layers (DC) and other FWI index, the moisture content of shallow layers (DMC) is a more important predictor for the burned area. Therefore, we should closely monitor this indicator and improve the measurement accuracy of sensors.

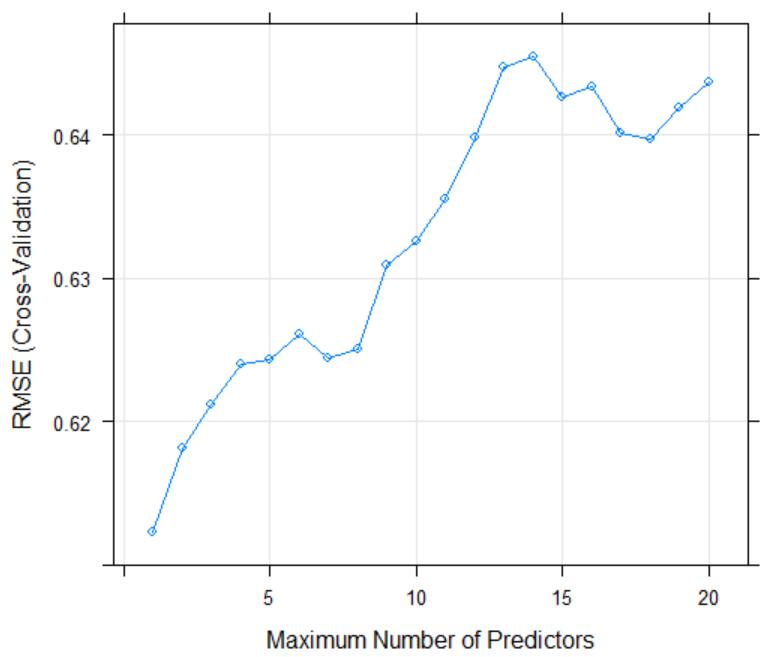
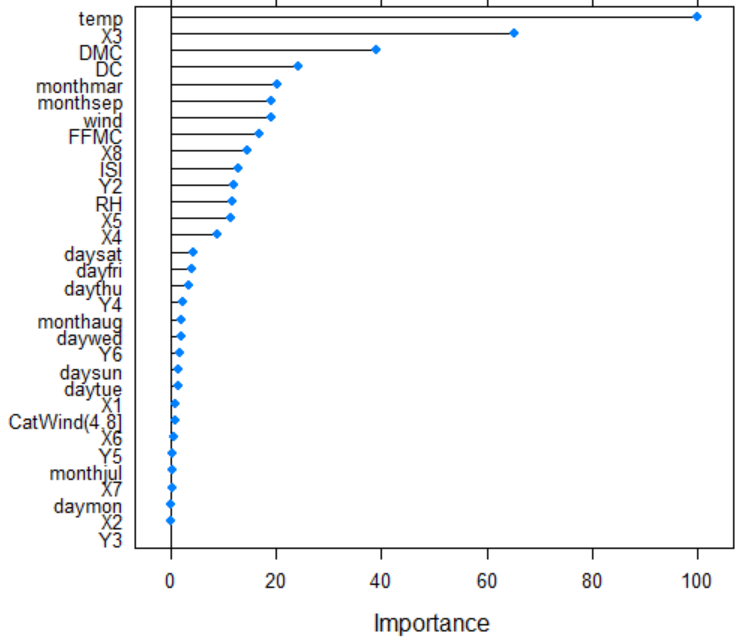


# Appendix

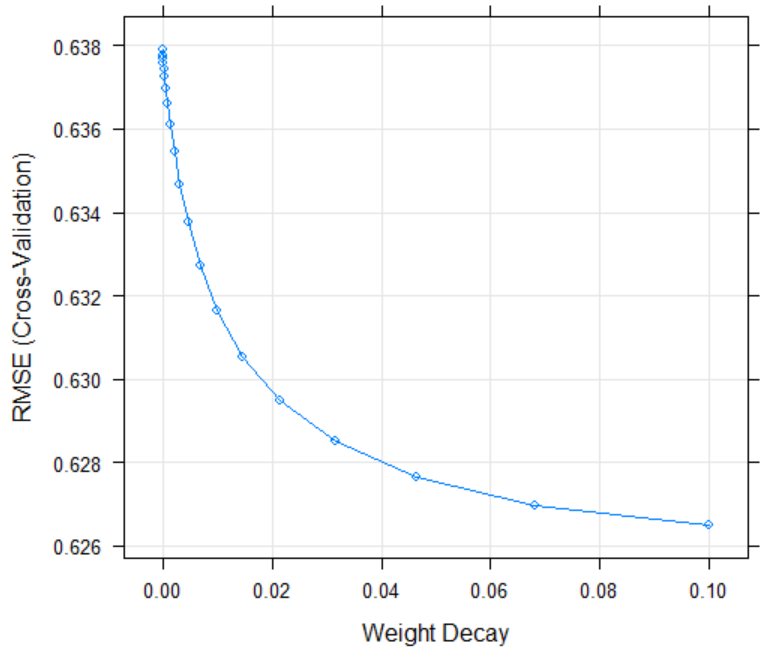
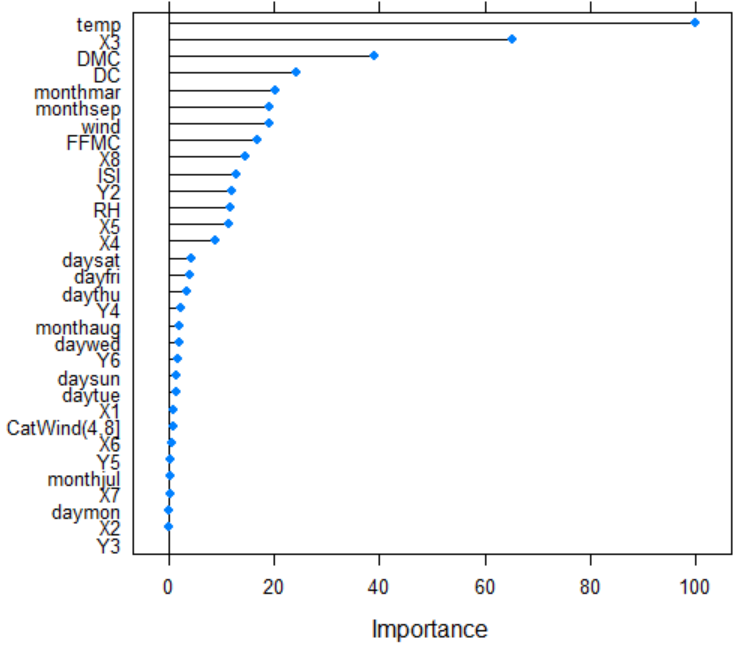
## full model results

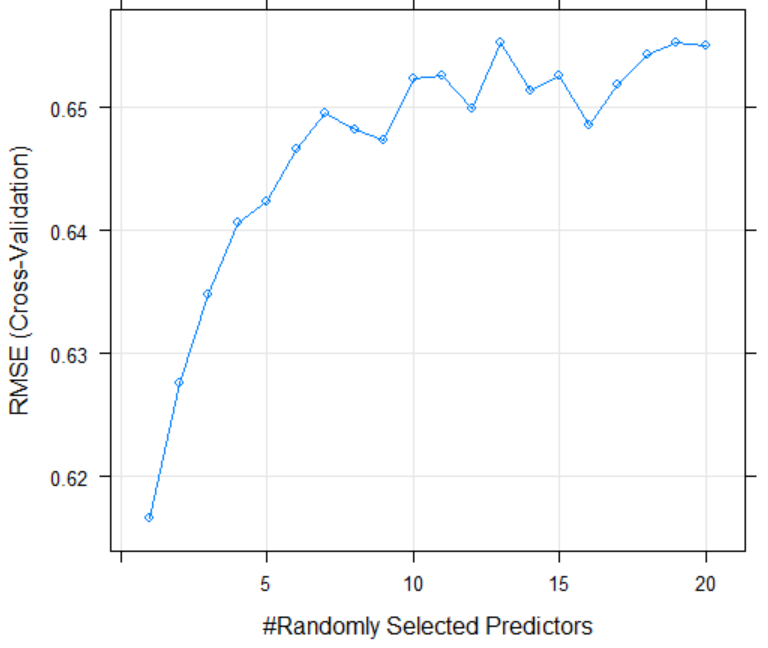
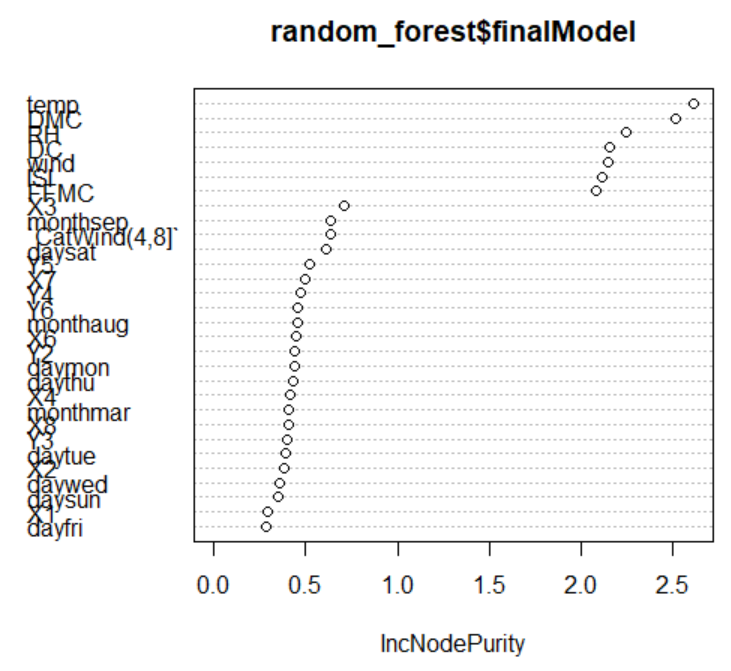
## forward model: cross-validated(10); varmax(20)

## ridge model: cross-validated(10); tune length(20)

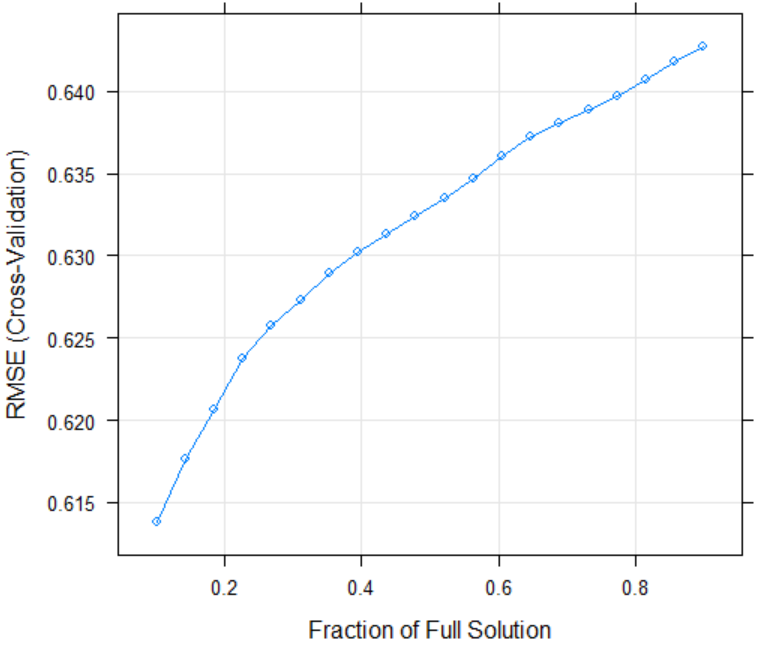
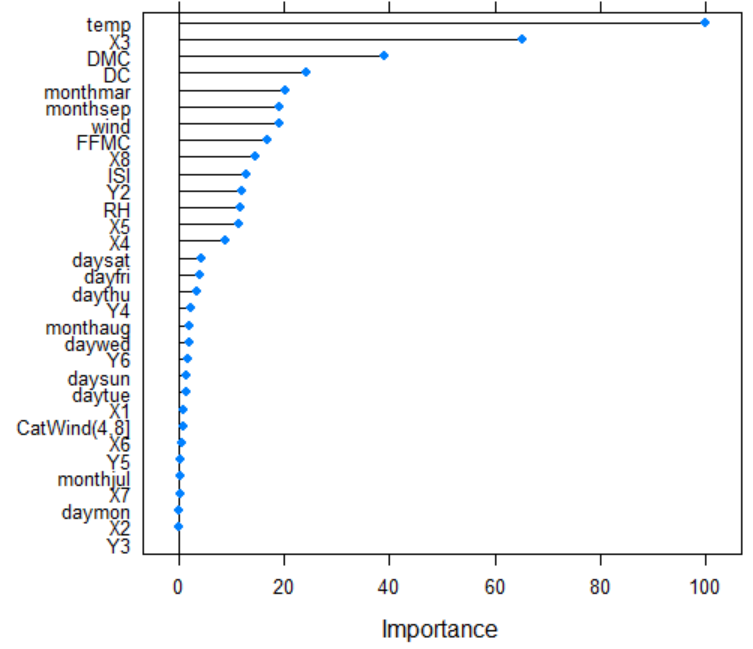
 

## random forest model: cross\_validated(10; randomly selected predictors(20)

# IncNodePurity: A specific parameter to measure the importance of the attribute for random forest model. Higher the value, higher the attribute importance. Source: <https://discuss.analyticsvidhya.com/t/how-to-extract-important-variables-from-random-forest-model-using-varimpplot-in-r/1325>

## lasso model: cross-validated(10); tune length(20)

## model metric comparison

