Forest Fires in Portugal

João Teixeira - up200705307, Nuno Peixoto - up200804621

25-01-2015

Introduction

- Monitoring and forecasting forest fires in Portugal;
- The several variables may influence the burnt area;
- In 2003, Portugal faced the worst forest fire losing 8.6% of of the total area;
- Elevation, slope or density are some of the specifications of the data set;

Objective : Explore and predicte the data of the forest.

Exploratory analysis of the data

- Global Summary
- Main Variables
- Target Variable

Global Summary

- Number of Columns: 81.
- Number of Rows: 990.
- Number of Data: 80190.
- Target Value: 1 (TotalBurntArea) - Numeric variable.
- Number of Unknown Values: 0.

Global Summary (cont.)

Climate Variables - The climatic conditions may affect the probability of a fire to occur;

Landscape Variables - The landscape has been extensively associated with fire occurrence;

Socio-economic Variables - Human have impact in historical fire patterns;

Topographic Variables - The topographic features may influence the fire ignitions;

Main Variables

In the following table we have the **TOP5** main variables:

attr_importance	attribute
0.2037	ELEV_MAX
0.1962	bio1
0.1926	ELEV_MEAN
0.1898	bio7
0.1844	DensPop01

Main Variables (Number of outliers)

- ELEV_MAX: 8 (0.81%)
- Bio1: 21 (2.12%)
- ELEV_MEAN: 9 (0.91%)
- Bio7: 1 (0.1%)
- DensPop01: 132 (13.33%)

Main Variables (Standard Deviation)

- ELEV_MAX: 339.100654
- Bio1: 14.710837
- ELEV_MEAN: 251.9971412
- Bio7: 30.9059137
- DensPop01: 1222.3683295

Target Variable

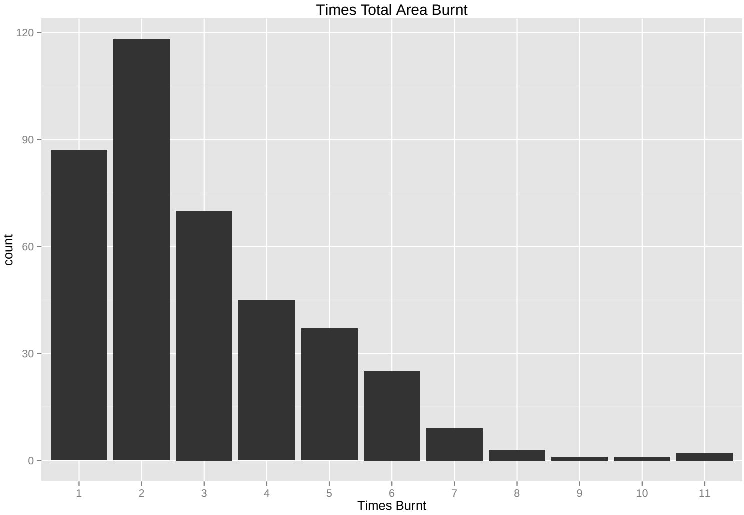
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	0	609	2550	2752	68981

Target Variable (Number of outliers)

- TotalBurntArea: 106 (10.71%)

We can see that more than 10% of the total burnt area values are considered outliers.

Target Variable (Total Area vs. Total Burnt Area)



Data Pre-Processing

- Remove None importance Variables
- Normalizing Value

Remove None importance Variables

attr_importance	attribute
NaN	TCI_STD
NaN	LPI
NaN	ED
NaN	FRAC_SD
NaN	IJI
NaN	ENN_AM
NaN	eucalipto_AREA_perc
NaN	outfolhosas_AREA_perc

- This pre-processing that does not prejudice the information

Normalizing Value

- Data normalization pre-processing we will use for the analisys;

```
##   ELEV_MAX ELEV_MEAN ELEV_STD SLOPE_MAX SLOPE_MEAN SLOPE
## 1      396  168.5080  76.7385   44.9590   18.87750  11.9
## 2      706  604.4890  42.7725   39.1152    8.99396   6.0
## 3       88   34.2032  23.7021   14.3287    2.32026   1.7
```

```
##     ELEV_MAX ELEV_MEAN   ELEV_STD  SLOPE_MAX SLOPE_MEAN
## 1 -0.2493703 -0.569479  0.2047242  0.5829093  1.0775585
## 2  0.6648126  1.160624 -0.4090371  0.2295771 -0.3401967
## 3 -1.1576551 -1.102441 -0.7536369 -1.2690826 -1.2975129
```

Predictive Models

In order to find the best regression model that can predict the target variable of the test data set with less error, we analysed the following forecasting models:

- Multiple Linear Regression
- Regression Trees
- K-Nearest Neighbors (KNNs)
- Support Vector Machines (SVMs)
- Artificial Neural Networks (ANNs)
- Random Forest (Ensembles)

Model comparison

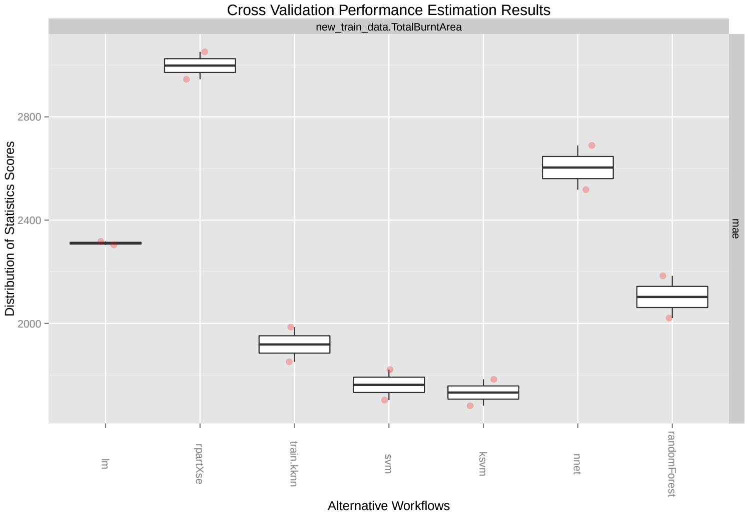
The metric evaluation to be considered for these forecasts is the MAE - Mean Absolute Error. We use a cross validation method with 2 repetitions of 3 folds.

To evaluate the models we will use the performanceEstimation package that provides a set of functions and arguments that allow us to change the values of parameters in order to check the best fit for an specific model.

Model performance

Model	MAE ERROR	Parameters
SVM (ksvm)	1732.284	**epsilon**:1e-09, **C**:1, **kernel**
SVM	1762.223	**cost**:1, **gamma**:0.01
k-Nearest Neighbors	1916.375	**scale**:TRUE, **k**:11, **distanc**
Multiple Linear Regression	2310.785	Default Parameters
ANN	2393.626	**size**:2, **maxit**:200, **decay**:0
Random Forest	2393.626	**ntree**:500, **nodesize**:5, **co**
Regression Trees	2802.628	**se**:1, **minsplit**:15

Models Plot


Cross Validation Performance Estimation Results
new_train_data.TotalBurntArea
mae
Distribution of Statistics Scores
2800
2400
2000
lm
rpartXse
train.kknn
svm
ksvm
nnet
randomForest
Alternative Workflows

Conclusion

Best performance models * **SVM** (ksvm) * MAE error: 1732.284

- **SVM**
  - MAE error: 1762.223
- **k-Nearest Neighbors**
  - MAE error: 1916.375

Clustering

The following clustering methods were used to try to find different groups of observations present in the data set:

- Clustering Large Applications (CLARA)
- Partitioning Around Medoids (PAM)
- Hierarchical Clustering
- K-Means Clustering

Clustering results

Using a R script with the help of the silhouette function we could find the best number of clusters for each used method

CLARA		PAM	
nClusters	SilhCo	nClusters	SilhCo
2	0.7144933	2	0.7336859
4	0.6361579	3	0.6187632
3	0.6243904	4	0.4812423
5	0.3102661	5	0.4318594
10	0.2789919	6	0.3595906
9	0.2749504	7	0.3234947
6	0.2645142	9	0.2750479

Clustering results (cont.)

hclust		kmeans	
nClusters	SilhCo	nClusters	SilhCo
2	0.55027091	2	0.7791398
3	0.32391734	3	0.6993844
5	0.17051099	4	0.6888222
4	0.16893921	5	0.5839630
6	0.09489308	6	0.5492905
7	0.09160570	7	0.3886676
8	0.08740959	8	0.3868419
10	0.08135908	10	0.3321750
9	0.07454503	9	0.3175380

Silhouette Plot