Forest Fires in Portugal

João Teixeira - up200705307, Nuno Peixoto - up200804621

Contents

Introdution

The data set of the fires in the Portugal forests allow monitoring and forecasting the fire outbreaks taking into account a number of variables that may influence the amount of burned area (Torgo 2010).

In 2003, Portugal faced the worst forest fire season ever, during which the burned area greatly exceeded the average for the last few decades. The year 2003 was marked by the loss of 8.6% (423,949 hectares) of the total area of Portuguese forests, representing a value four times the annual average of the 90's (Leite et al. 2013).

The catastrophic forest fires involving human and infrastructures loses are becoming increasingly common in different parts of the world, particularly in bioclimatic regions of the Mediterranean due to dry and hot seasons. Besides leaving a long trail of burnt areas, the catastrophic fires have caused a significant number of human victims.

Each data set entry contains a set of specifications (variables) into an forest area, such as elevation, slope, the percentage area of each existing type, density and **the total burnt area**.

The purpose of this work consists int the data analysis and exploration of predictive models in order to predict the burned forest area and thereby create prevention actions to fighting forest fires.

Exploratory analysis of the data

Global Summary

The train dataframe has the following main attributes:

- Number of Columns: 81.
- Number of Rows: 990.
- Number of Data: 80190.
- Target Value: 1 (TotalBurntArea) - The variable is a numeric value so the predict model will be regression.
- Number of Unknown Values: 0.

The data are organized into four categories that affect somehow the area of burned forest:

Climate Variables - The climatic conditions may affect fuel accumulation and moisture having an effect on the probability of a fire to occur;

Landscape Variables - The landscape features of the Earth's surface has been extensively associated with fire occurrence;

Socio-economic Variables - Human factors have been used in predictive modeling of historical fire patterns;

Topographic Variables - The topographic features and compositions may influence the fire ignitions and the accessibility limitations to reach the fire occurrences;

Main Variables Summary

In order to know and summary only the variables that have more impact in the target value we can calculate the gain ratio between each variable with the TotalBurntArea.

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

Table 1: Variables and the importance for the targe value.

There's a short description of the main variables:

- **ELEV_MAX** - Maximum altitude;
- **bio1** - Annual Mean Temperature;
- **ELEV_MEAN** - Mean altitude;
- **bio7** - Temperature Annual Range (between max temperature of warmest month and min of coldest month);
- **DensPop01** - Population density in 2001;

1. Summary

The summarization of the data allows to get an overview of the fundamental properties of the data submitted for analysis, with the aim to describe the properties of the data between the observations measures.

With *summary* command is possible to view the following statistics:

ELEV_MAX	bio1	ELEV_MEAN	bio7	DensPop01
Min. : 14.0	Min. : 90.04	Min. : 3.081	Min. :124.3	Min. : 1.40
1st Qu.: 211.2	1st Qu.:136.00	1st Qu.: 100.619	1st Qu.:200.6	1st Qu.: 30.45
Median : 414.0	Median :143.25	Median : 244.958	Median :225.9	Median : 102.85
Mean : 480.6	Mean :143.38	Mean : 312.015	Mean :224.6	Mean : 400.66
3rd Qu.: 714.5	3rd Qu.:153.52	3rd Qu.: 476.824	3rd Qu.:250.3	3rd Qu.: 265.30
Max. :1990.0	Max. :175.20	Max. :1355.150	Max. :286.7	Max. :15597.00

Table 2: Summarization of top5 variables

2. Number of Outliers

An outlier is an observation that is outside the overall pattern of a distribution. Usually, the presence of an outlier indicates some sort of problem.

It is importante check the number and the rate of existing outliers for the most importante variables of the data that may influence in the predictive models.

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

It can be concluded that the main variables (except **DensPop01**) have a low outlier rate which does not significantly affect the forecast values.

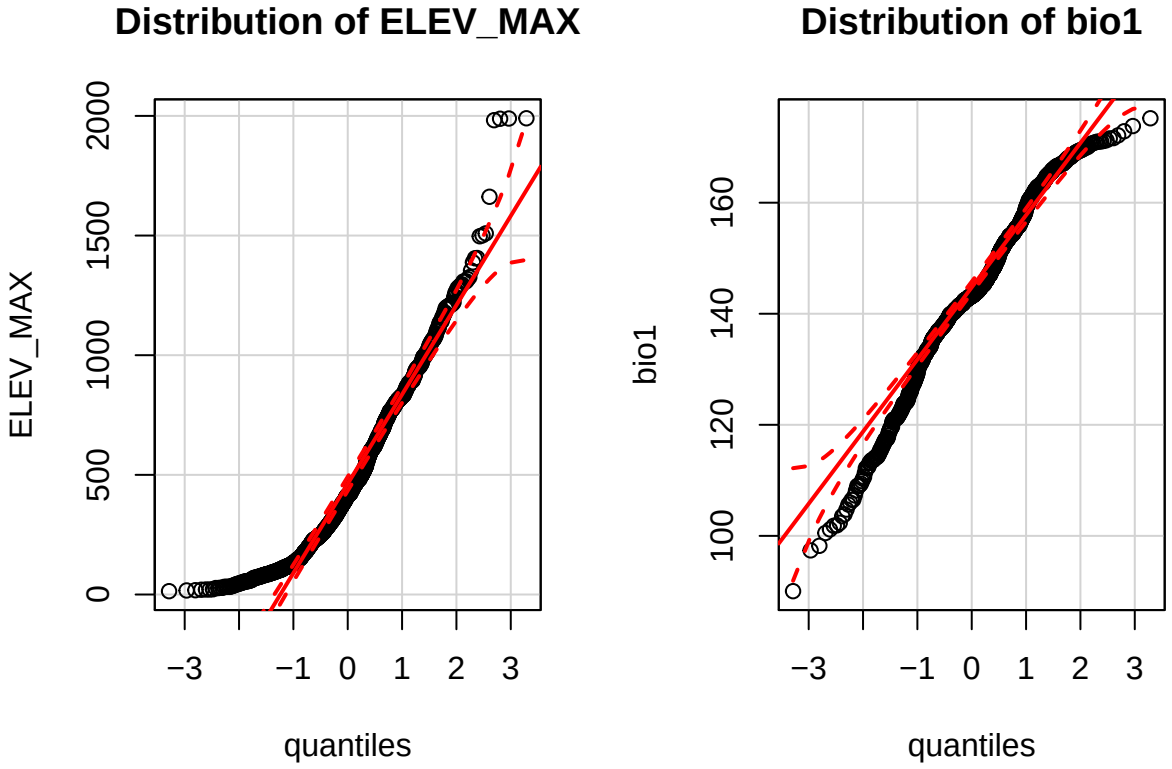
With the variable DensPop01 it may have to do some sort of analysis or pre-processing once has a higher value of outliers.

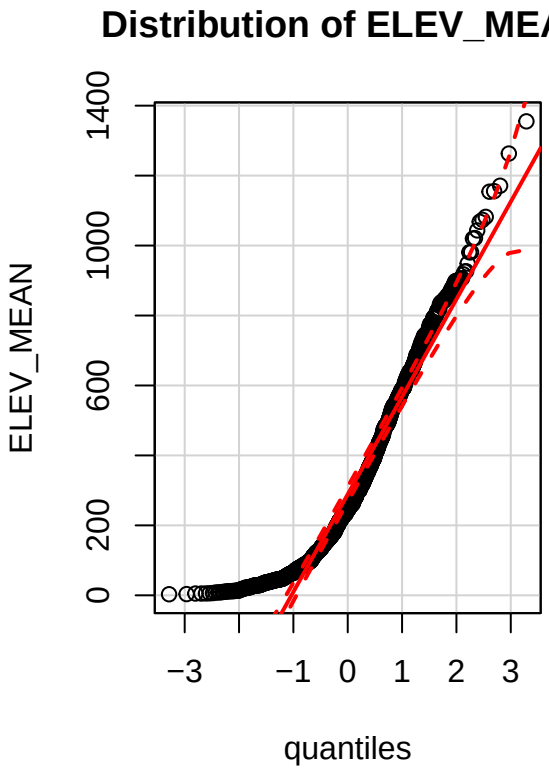
3. Distribution of the values

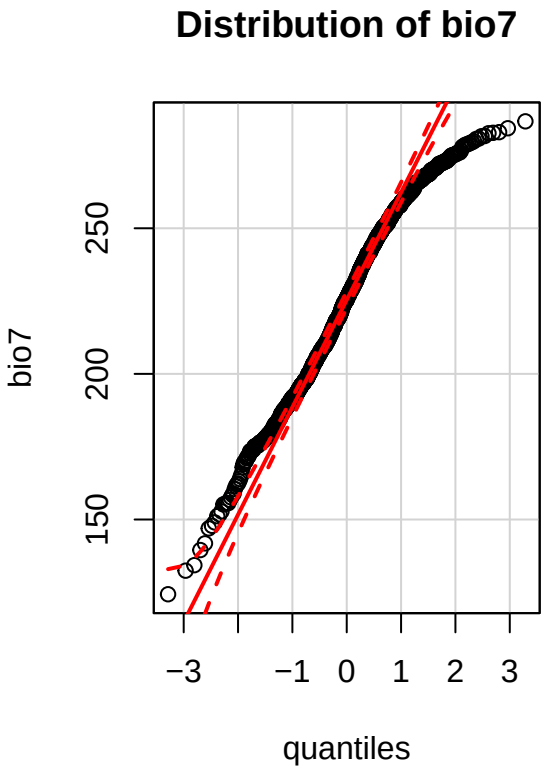
The most common graphical tool to assess the normality of the data is a Quantile-Quantile (Q-Q) plot.

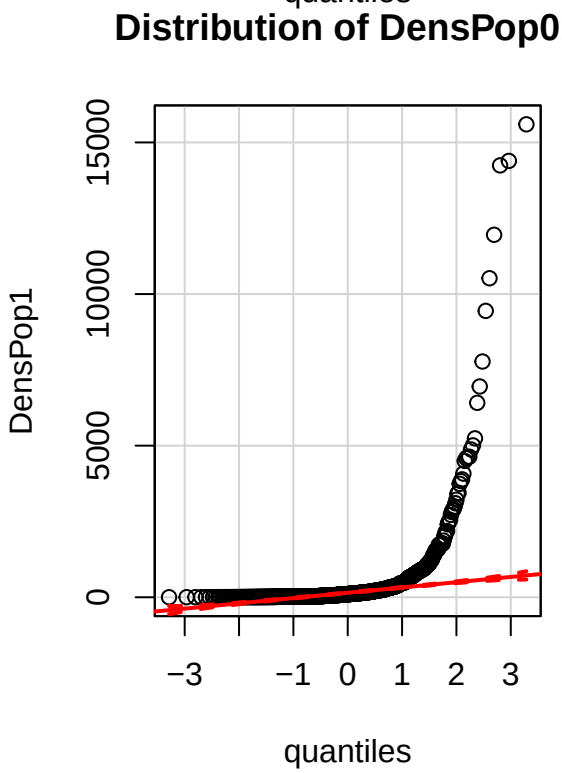
In a Q-Q plot quantile the values of a theoretical distribution are plotted against quantile values of the observed sample distribution. Normaly the normal distribution is used to make a judgment if the 2 quantiles are in the same distibution(Razali and Wah 2011).

In the following plots we analyse the variables with the e qqplot to check if the values follow the normal distribution.









We can see that most of the variables follow in most of the values of a normal distribution. However, and due to the high presence of outlier the **DensPop01** those not follow this distribution.

4. Standard Deviation

The Standard Deviation is a measure of how spread out numbers are. If the standand deviation is low than the observation values are close to the median, otherwise the data are spread out across a wide range of values.

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

In 3 of 5 cases show previously showed the standard deviation is to higher. This values (ELEV_MAX, ELEV_MEAN and DensPop01) must influence the results of the target variable.

Target Variable Summary

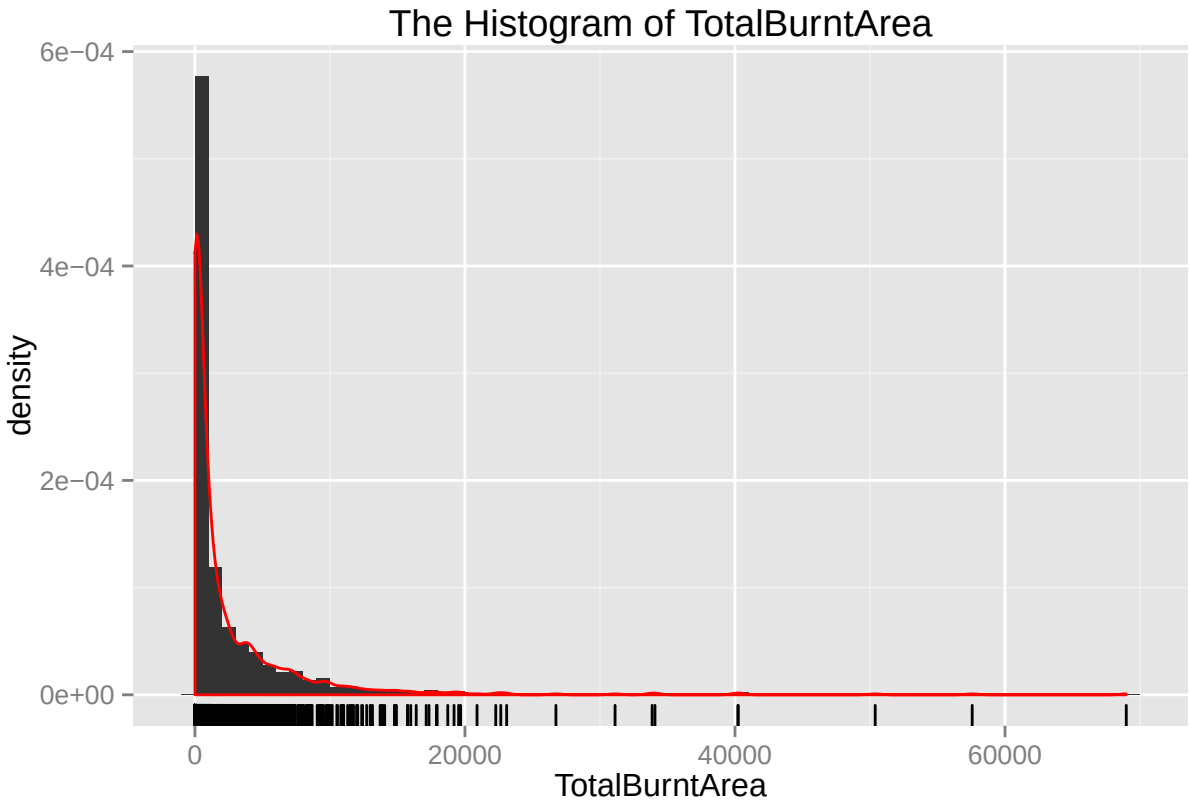
In this chapter we will make an exploratory analisys of the target variable, **TotalBurntArea**.

1. Summary

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

Table 3: Summarization of target variable

2. Histogram



Through the histogram above we can see that there is a large concentration of values below 20,000 square meters burned.

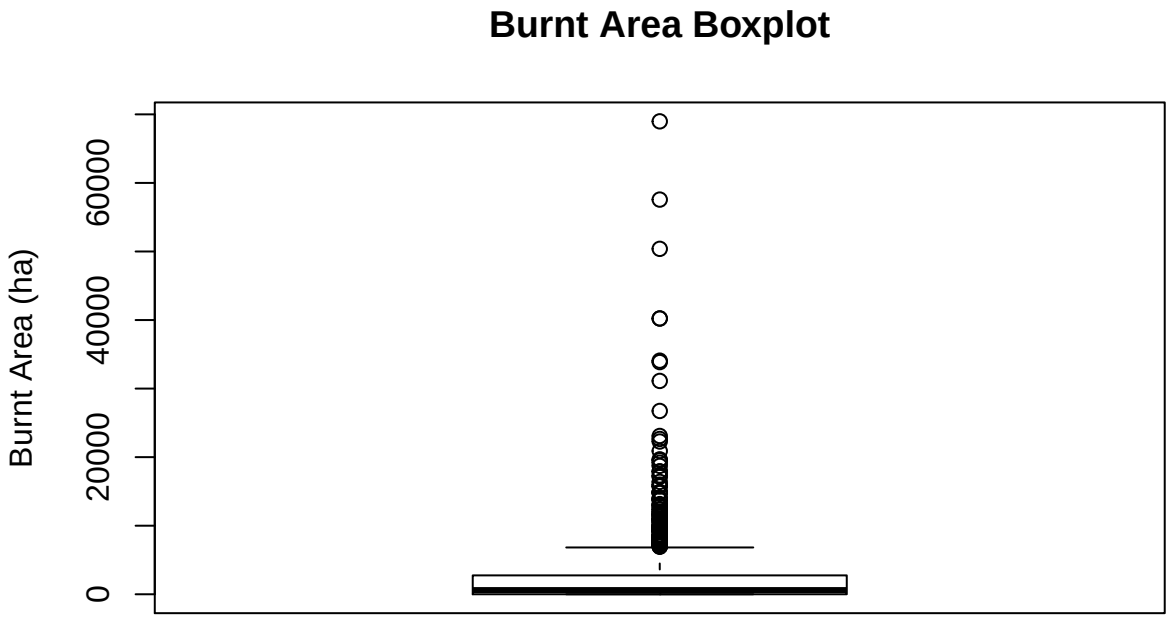
3. Outliers

- TotalBurntArea: 106 (10.71%)

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

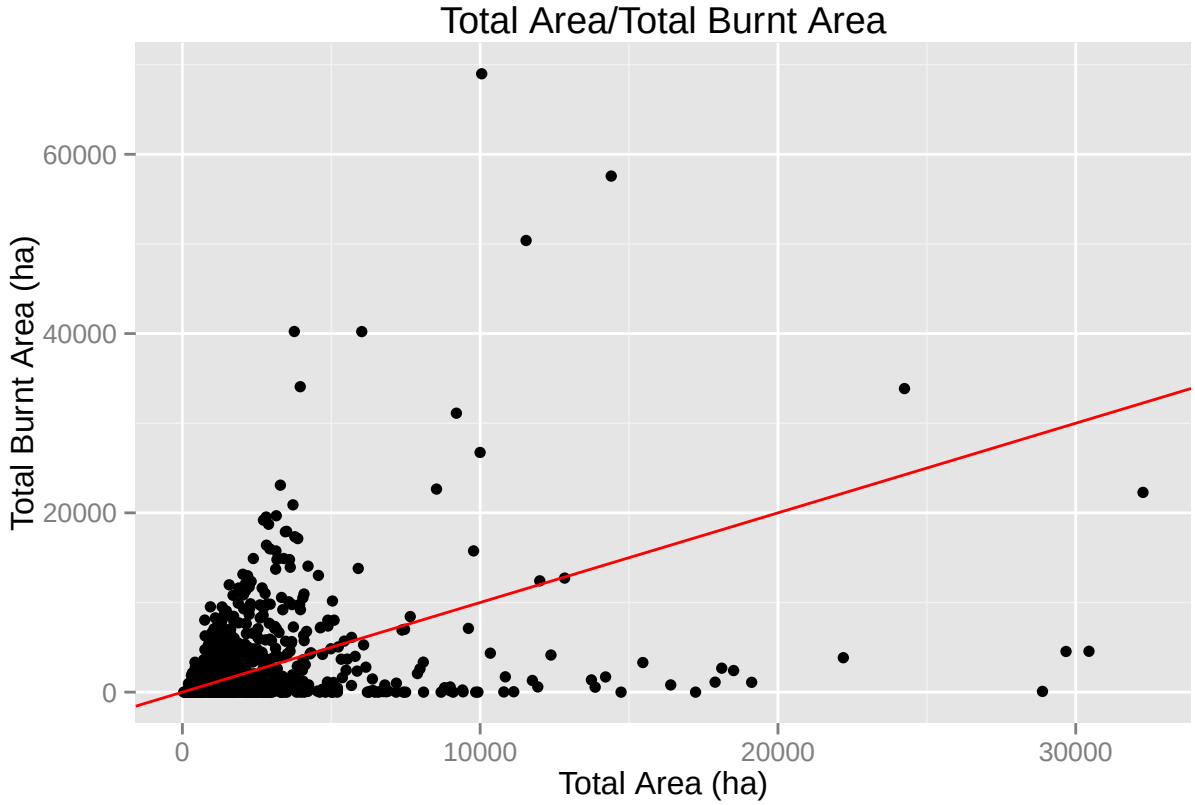
On the one hand despite containing a large number it means that the overall burned area values are low.

On the other hand some outlier values are pretty high which means that there are situations where the area of burned forest is very high.

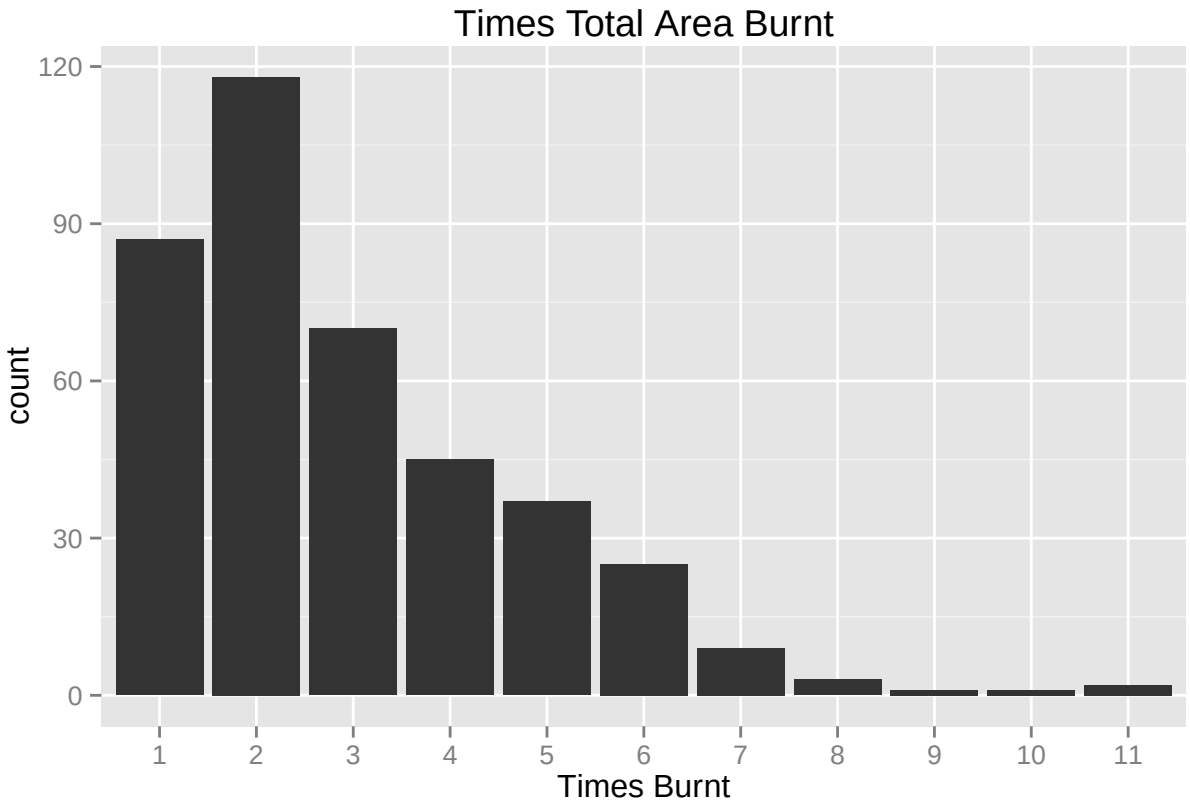


4. Total Area vs. Total Burnt Area

Compare the total burnt area for each of the locations of the data set with the total area may be able to draw some conclusions.



We checked out from the graph that there are some cases (40.2 %) where the burnt area is greater than the total area. This means that in these cases the land area was completely burned repeatedly over time.



By the graphic the majority of areas analyzed burns between 1-3 times its total area. However, there are regions where their area has been completely decimated by fire repeatedly over time which allows us to conclude that these regions are more conducive to fires.

Conclusions

From the exploratory analisys made to the data we can take several conclusions.

- The fact that there are many variables for analysis can somehow undermine the process of data analysis. The analysis of only the variables with the greatest impact came speed up this process;
- The number of outliers is significantly low;
- Most of the analyzed variables follow, or is very close to the normal distribution;
- Although in most presented cases the total burnt area is less than 20,000 square meters there's a significant number of places the that the fire consumed a considerable area of its total area or has been completely burned several times;

Data Pre-Processing

Remove None importance Variables

In order to reduce the size of data to analyze/evaluate, we can determine as was done in chapter 2.2, the importance of each variable in the training data in the target variable.

In the following table we have the variables with no relevance to the calculation of the target variable:

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

Table 4: Variables with no importance

This way we will remove these 8 variables that will help to improve the processing efficiency in building models since we eliminate 8 columns of 990 entries.

With this pre-processing that does not prejudice the information analysis could be reduced 9.88% of the total data.

From now on we will no longer use the *train_data* provided. We will use the same data frame but without the data previously removed.

Normalizing Value

Data normalization is one of the pre-processing techniques that we will use for the analisys of the data.

This technique allows the values and variables of the dataframe are all on the same scale (typically with mean 0 and standard deviation 1).

This way, the prediction of some regression models can be more efficient since the range of values can influence the performance of the data is provided.

Not all models that we will present ahead perform this procedure internally, so we will normalize the training dataframe.

```
norm.new_train_data <- scale(new_train_data[,])
colnames(norm.new_train_data) <- colnames(new_train_data)
norm.new_train_data <- data.frame(norm.new_train_data)
```

Comparing the values before and after the normalized data we can see that before the pre-processing the scale of values was very dispersed (e.g. comparing a 88 with a 706), while after normalization the values are in a much more “harmonic” scale.

```r
head(new_train_data[,c(1:6)], 3)
```

```
##   ELEV_MAX ELEV_MEAN ELEV_STD SLOPE_MAX SLOPE_MEAN SLOPE_STD
## 1      396  168.5080  76.7385   44.9590   18.87750  11.90550
## 2      706  604.4890  42.7725   39.1152    8.99396   6.03132
## 3       88   34.2032  23.7021   14.3287    2.32026   1.76678
```

```r
head(norm.new_train_data[,c(1:6)], 3)
```

```
##     ELEV_MAX ELEV_MEAN   ELEV_STD  SLOPE_MAX SLOPE_MEAN   SLOPE_STD
## 1 -0.2493703 -0.569479  0.2047242  0.5829093  1.0775585  1.73288356
## 2  0.6648126  1.160624 -0.4090371  0.2295771 -0.3401967 -0.09093957
## 3 -1.1576551 -1.102441 -0.7536369 -1.2690826 -1.2975129 -1.41499956
```

Predictive Models

In this chapter we will look several forecasting models in order to figure out which the best regression model that can predict the target variable of the test data set with less error.

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.

The Mean Absolute Error (MAE) measures the average absolute deviation between the predictions and the true values. The MAE is measured in the same unit as the original variable scale. That means if the values are scaled the mae is also scaled.

For evaluate the modles 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.

Multiple Linear Regression

Description

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.

Evaluation

For evaluate this method we use a central imputation for a initial pre-processing and only have in count the positive values generated with this module.

We obtain the following error:

MAE: 2310.785.

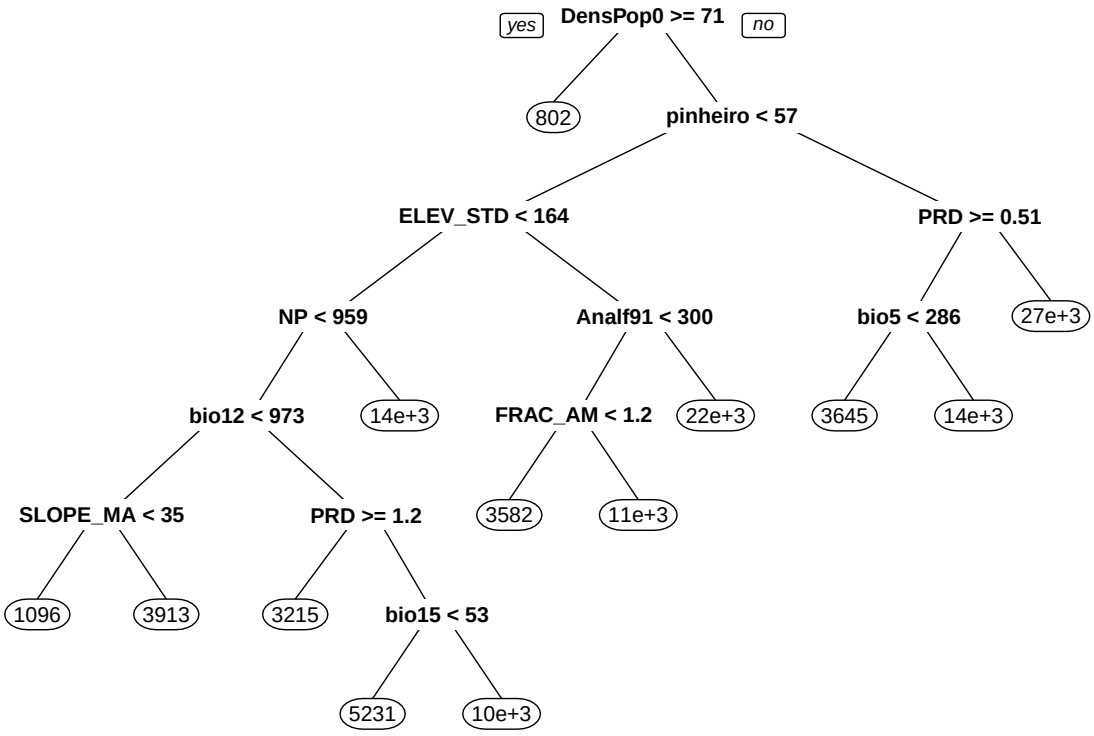
Regression Trees

Description

Tree-based models are models that provide as result a model based on logical tests on the input variables.

This partitioning is defined based on carefully chosen logical tests on these variables. Within each partition all cases are assigned the same prediction (either a class label or a numeric value).

Here is the tree regression generated for this case study:



Evaluation

For evaluate this method we use 2 diferent parameters, the **se** for the number of standard errors to use in the post-pruning and the **minsplit** to control the stopping criteria used to stop the initial tree growth.

Parameter	Values
se	1,2,3
minsplit	1,15,30

With this module the error **2802.628** was obtained with the parameters:

se	**minsplit**
1	15

Table 6: Regression Tree Parameters

k-Nearest Neighbors(KNNs)

Description

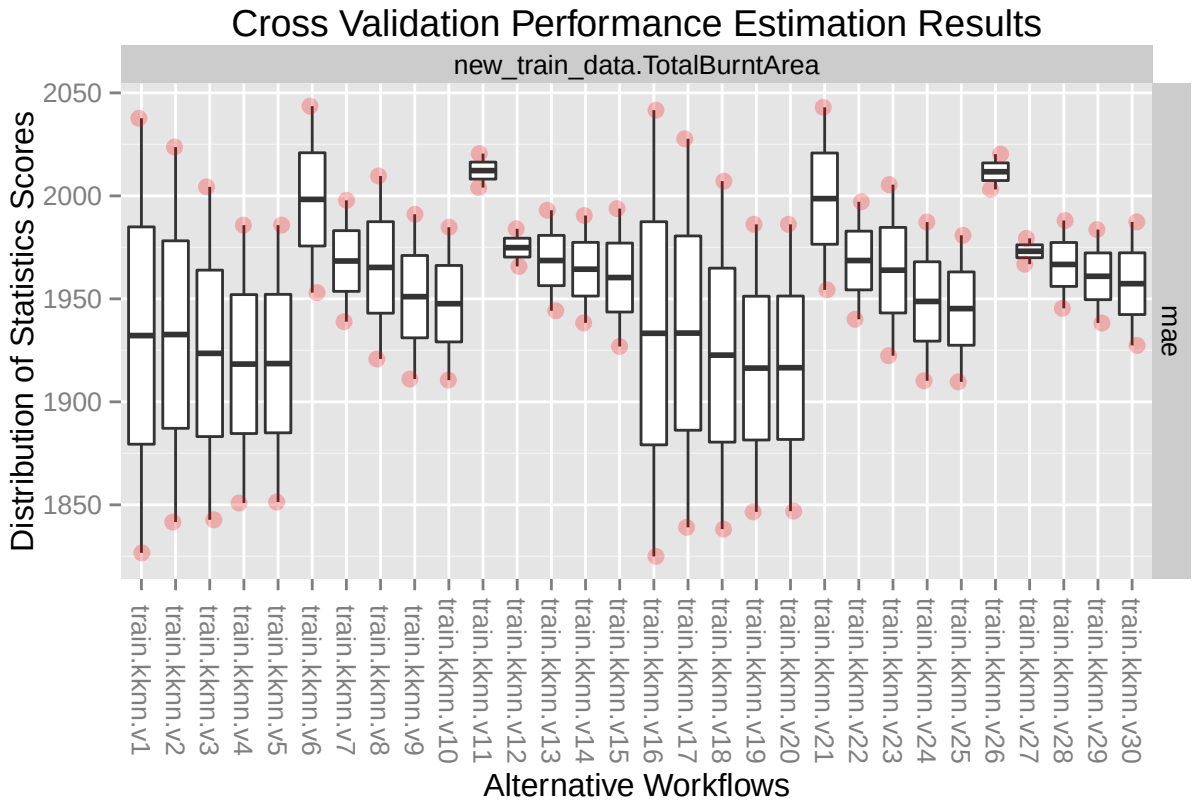
The purpose of the k Nearest Neighbours (kNN) algorithm is to use a data set where the data points are separated into several separate classes to predict the classification or regression of a new sample point. This sort of situation is best motivated through examples.

Evaluation

We made several tests to the KNN model, changing the value of the number of the neighbours considered (**k** parameter) the distance between neighbors (**distance** parameter) and the kernel to use (**kernel** parametes).

Parameter	Values
k	5,7,9,11,13
distance	1,2,3
kernel	epanechnikov, triangular

It is possible to verify that the result of estimation with the different values is very similar for the values of the diferent tested parameters.



With this module the error **1916.375** was obtained with the parameters:

scale	**k**	**distance**	**kernel**
TRUE	11	1	triangular

Table 8: Knn module Parameters

Support Vector Machines(SVMs)

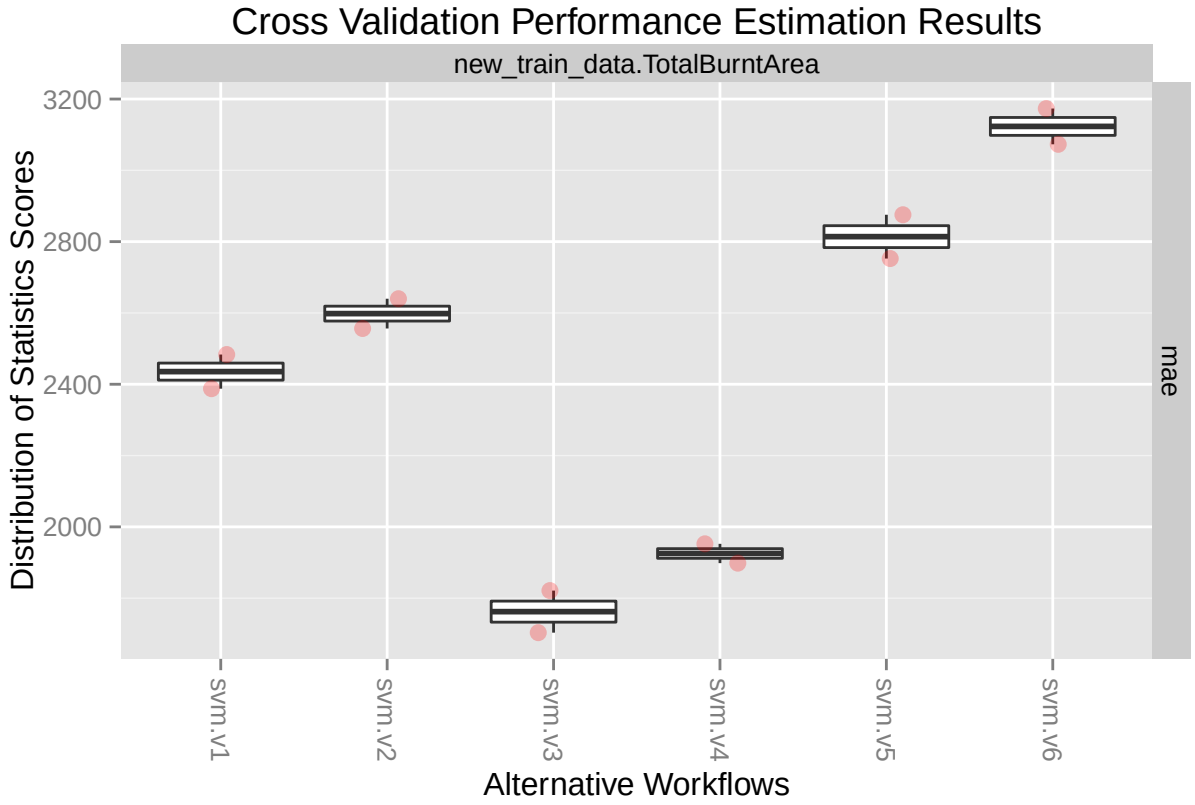
Description

The main goal of SVMs model is mapping the original data into a new, high-dimensional space, where it is possible to apply linear models to obtain a separating hyper plane. The mapping of the original data into this new space is carried out with the help of kernel functions. (Torgo 2010)

Evaluation

We made several tests to the SVM model, changing the value of cost of constraints violation (**cost** parameter) and the value needed for kernel's formula (**gamma**).

Parameter	Values
cost	1,10
gamma	0.1, 0.01, 1



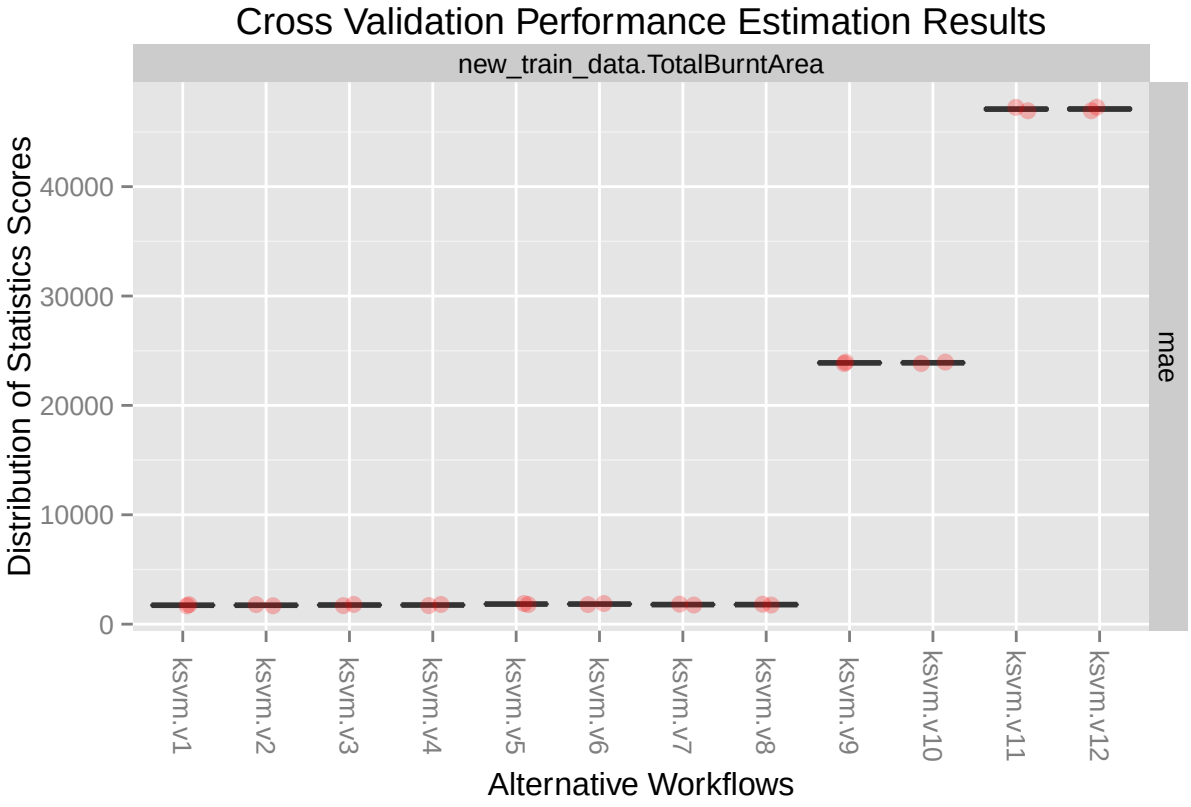
The MAE error **1762.223** was obtained with the parameters:

cost	**gamma**
1	0.01

Table 10: SVM module Parameters

It was made a different approach using SVMs package **ksvm** (Karatzoglou et al. 2004) which allows us to take advantage of more parameters and a greater variety of kernels.

Parameter	Values
epsilon	0.01, 10^-9
C	1, 2
kernel	rbfdot, laplacedot, besseldot



With this module the error **1732.284** was obtained with the parameters:

epsilon	**C**	**kernel**
1e-09	1	rbfdot

Table 12: KSVM module Parameters

Artificial neural networks(ANNs)

Description

Artificial neural networks are models with a strong biological inspiration composed by a set of units (neurons) that are connected between them by associated weights.

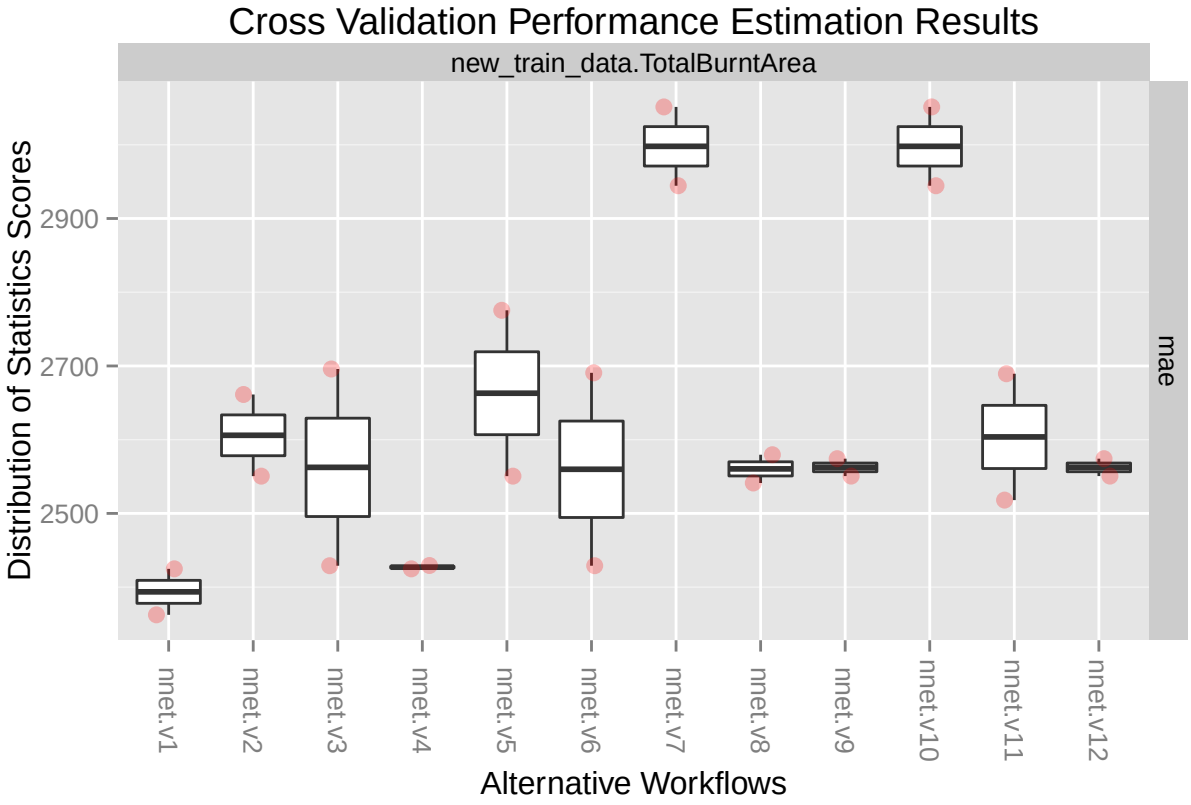
These basically consist of inputs which are multiplied by weights and then computed by a mathematical

function which determines the activation of the neuron ANNs combine artificial neurons in order to process information(Rojas 1996).

Evaluation

To evaluate the performance of the Ann's we made several tests with diferent parameters. We varied the the number of the hidden layers (**size** parameter), the maximum number of iterations (**maxit** parameter) and the weight decay (**decay** parameter).

Parameter	Values
size	2,4,6
maxit	200,300
decay	0.1, 0.4



With this module the error **2393.626** was obtained with the parameters:

size	**maxit**	**decay**	**scale**	**trace**	**linout**
2	200	0.1	TRUE	FALSE	1

Table 14: Artificial Neural Networks module Parameters

Random Forest (Ensembles)

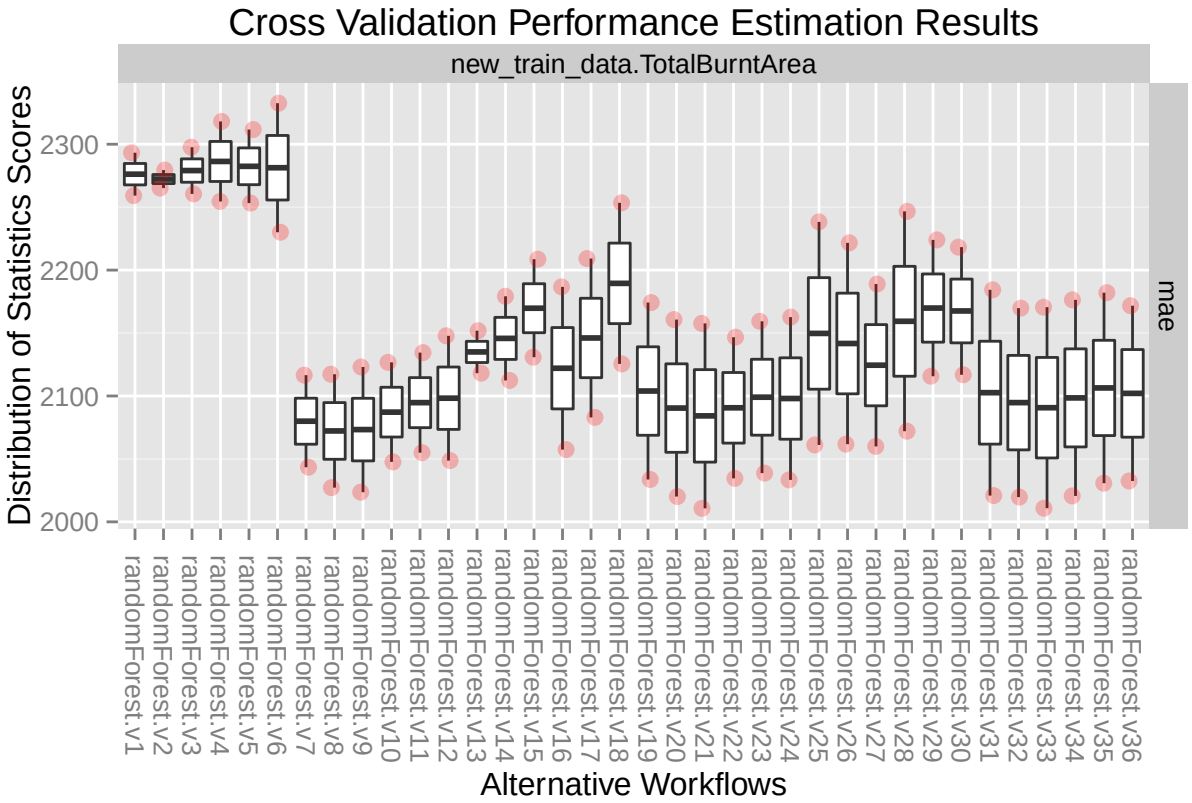
Description

The random forest is a ensemble model The individual decision trees are generated using a random selection of attributes at each node to determine the split. More formally, each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest(Han, Kamber, and Pei 2011).

Evaluation

To evaluate this module we 4 diferent parameters, the **ntree** to define the umber of trees to grow, the **nodesize** that sets the the minimum size of terminal node, the **corr.bias** to perform (or not) bias correction for regression and **mtry** who define the number of variables randomly sampled as candidates at each split.

Parameter	Values
ntree	250, 500, 1000
nodesize	5, 10
corr.bias	T (True), F (False)
mtry	3,6, 9



With this module the error **2393.626** was obtained with the parameters:

ntree	nodesize	corr.bias	mtry
500	5	FALSE	3

Table 16: Random Forest module Parameters

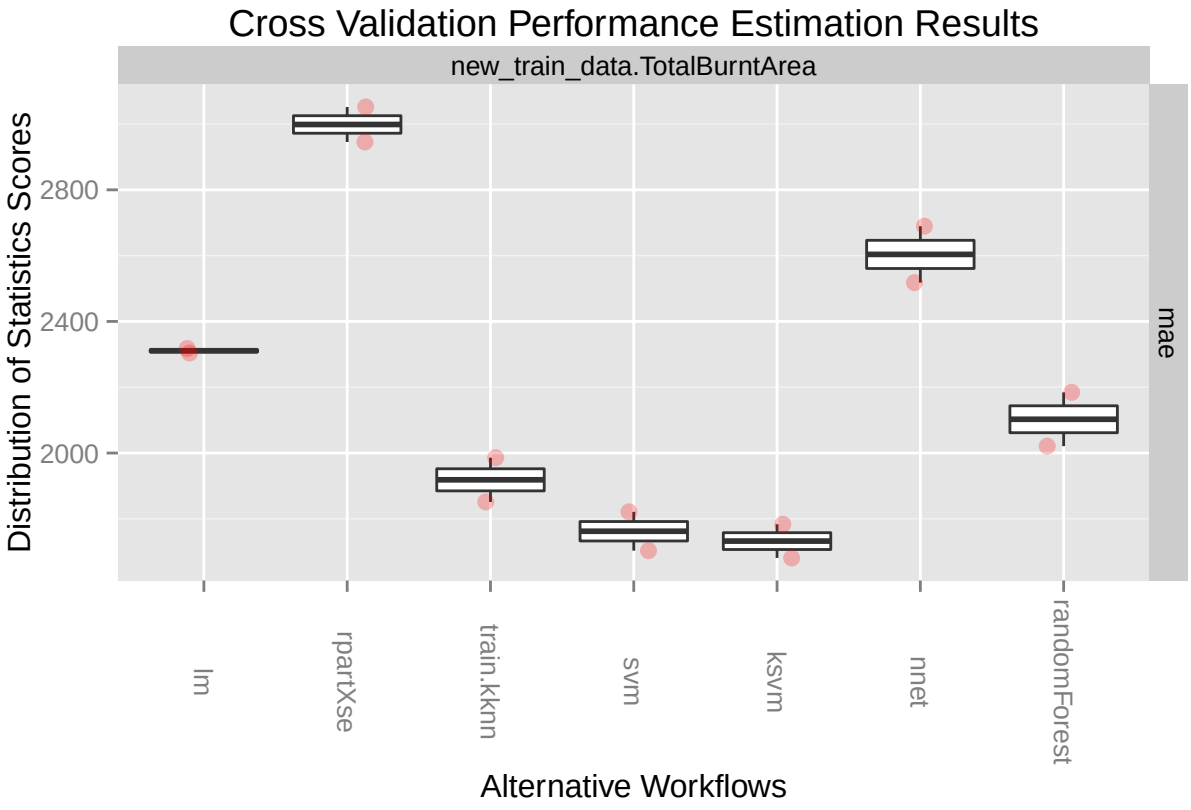
Conclusions

After evaluating different models with different parameters it is possible to make a comparison between the results obtained by different predictive models. The following table shows for each model/forecast algorithm tested the **MAE** taken and the associated parameters.

Model	MAE ERROR	Parameters
Multiple Linear Regression	2310.785	Default Parameters
Regression Trees	2802.628	**se**:1, **minsplit**:15
k-Nearest Neighbors	1916.375	**scale**:TRUE, **k**:11, **distance**:1, **kernel**:triangular
SVM	1762.223	**cost**:1, **gamma**:0.01
SVM (ksvm)	1732.284	**epsilon**:1e-09, **C**:1, **kernel**:rbfdot
ANN	2393.626	**size**:2, **maxit**:200, **decay**:0.1, **scale**:TRUE, **trace**:FALSE, **linout**:1
Random Forest	2393.626	**ntree**:500, **nodesize**:5, **corr.bias**:FALSE, **mtry**:3

We can see throw the table that Suport Vector Machine model with ksvm package produces a better MAE value which allows us to ensure that this model is the best to predict the total burnt area for a given set data.

In the plot we have the cross validation results for all the models that we have tested and the can conclude that the top3 model are the **knn**, **svm** and the **ksvm** being the ksvm the best.


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

Clustering

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.

K-Means Method

The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration(Han, Kamber, and Pei 2011).

We used the following script to find the best k (from 2 to 10) value for clustering the data using the kmeans method:

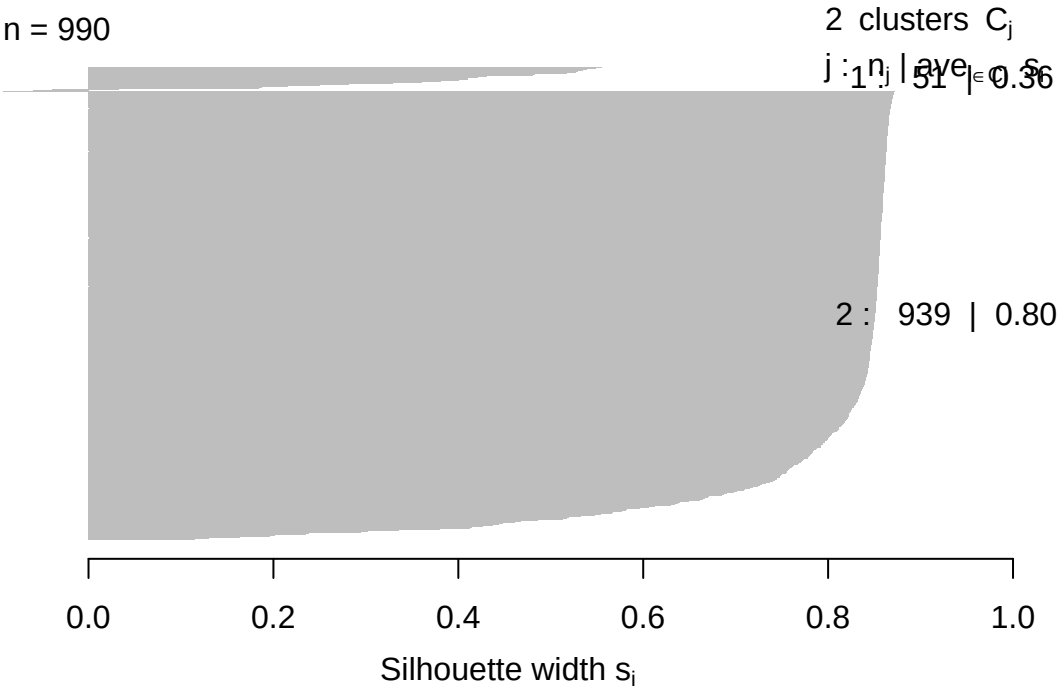
```
set.seed(1234)
d = dist(new_train_data[,-73])
avgS = c()
for(k in 2:10){
  cl = kmeans(new_train_data[,-73],centers=k,iter.max=300)
  s = silhouette(cl$cluster,d)
  avgS = c(avgS,mean(s[,3]))
}
```

The best k is given by the silhouette coefficient closer to 1:

```
  nClusters    SilhCo
1         2 0.7792156
2         3 0.6994996
3         4 0.6889785
4         5 0.5841365
5         6 0.5495380
6         7 0.3890778
7         8 0.3873062
8         9 0.3810128
9        10 0.3327353
```

If we plot the silhouette of the clustering using kmeans with the k=2 centers, we can see that we only get a very small percentage of cases in one of the clusters, and all the remaining cases in the other, this means that the distance of observations in the data set is very small.

Silhouette plot of (x = k2$cluster, dist = dist(new_train_data[, ·



Average silhouette width : 0.78

References

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Karatzoglou, Alexandros, Alexandros Smola, Kurt Hornik, and Achim Zeileis. 2004. "kernlab - an S4 Package for Kernel Methods in R." *Journal of Statistical Software* 11 (9): 1–20. http://www.jstatsoft.org/v11/i09.

Leite, Flora Ferreira, António Bento Gonçalves, Luciano Lourenço, and Xavier Úbeda. 2013. "Grandes Incêndios Florestais Em Portugal Continental Como Resultado Das Perturbações Nos Regimes de Fogo No Mundo Mediterrâneo." http://hdl.handle.net/1822/25046.

Razali, Nornadiah, and Yap B. Wah. 2011. "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests." *Journal of Statistical Modeling and Analytics* 2 (1).

Rojas, R. 1996. *Neural Networks: A Systematic Introduction*. Springer Berlin Heidelberg. http://books.google.pt/books?id=txsjjYzFJS4C.

Torgo, Luis. 2010. *Data Mining with R: Learning with Case Studies*. 1st ed. Chapman & Hall/CRC.