FOREST FIRES IN PORTUGAL

Problem Definition

Binary Classification Problem

Original Attributes:
"id","region","district","municipality","parish","lat","lon","origin","alert_date","alert_hour","extinction_date","extinction_hour","firstInterv_date","firstInterv_hour","alert_source","village_area","vegetation_area","farming_area","village_veget_area","total_area".

Target Variable: "intentional_cause"

- 0 -> no
- 1 -> yes

Output of the Classification Model: probability of a fire being intentional

Data Understanding

Type of Data

- Tabular
- Nondependency-oriented data

Types and Scales of Attributes

id	region	district	municipality	parish	lat	lon	origin	alert_date	alert_hour	extinction_date
Numerical	Categorical	Categorical	Categorical	Categorical	Numerical	Numerical	Categorical	Numerical	Numerical	Numerical
Ratio	Nominal	Nominal	Nominal	Nominal	Ratio	Ratio	Nominal	Interval	Ratio	Interval

extintion_hour	fistInterv_date	firstInterv_hour	alert_source	village_area	vegetation_area	farming_area	village_veget_area	total_area	intentional_cause
Numerical	Numerical	Numerical	NA	Numerical	Numerical	Numerical	Numerical	Numerical	Categorical
Ratio	Interval	Ratio	NA	Ratio	Ratio	Ratio	Ratio	Ratio	Nominal

Data Preparation | Data Quality Issues

Missing values

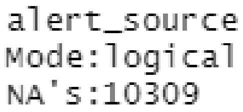
- Variables with some missing values (region, extinction_date, firstInterv_date).
- In the case of alert_source, all the values are missing values.

Inconsistent or incorrect values

- Values of region having "-".
- Diferent ways of naming the same district ("Viana do Castelo" and "Viana Do Castelo").
- The coordinates are not represented in the same way.
- Some of the coordinates values have "," instead of ".", which is not used in R.

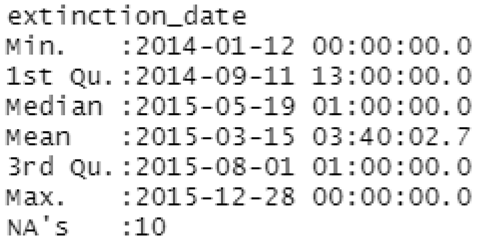
Data Preparation | Data Pre-processing

Data Cleaning – Handling Missing Values



```
alert_source
Mode:logical
NA's:10309
```

"Alert_source" has been withdrawn since all values are missing values.

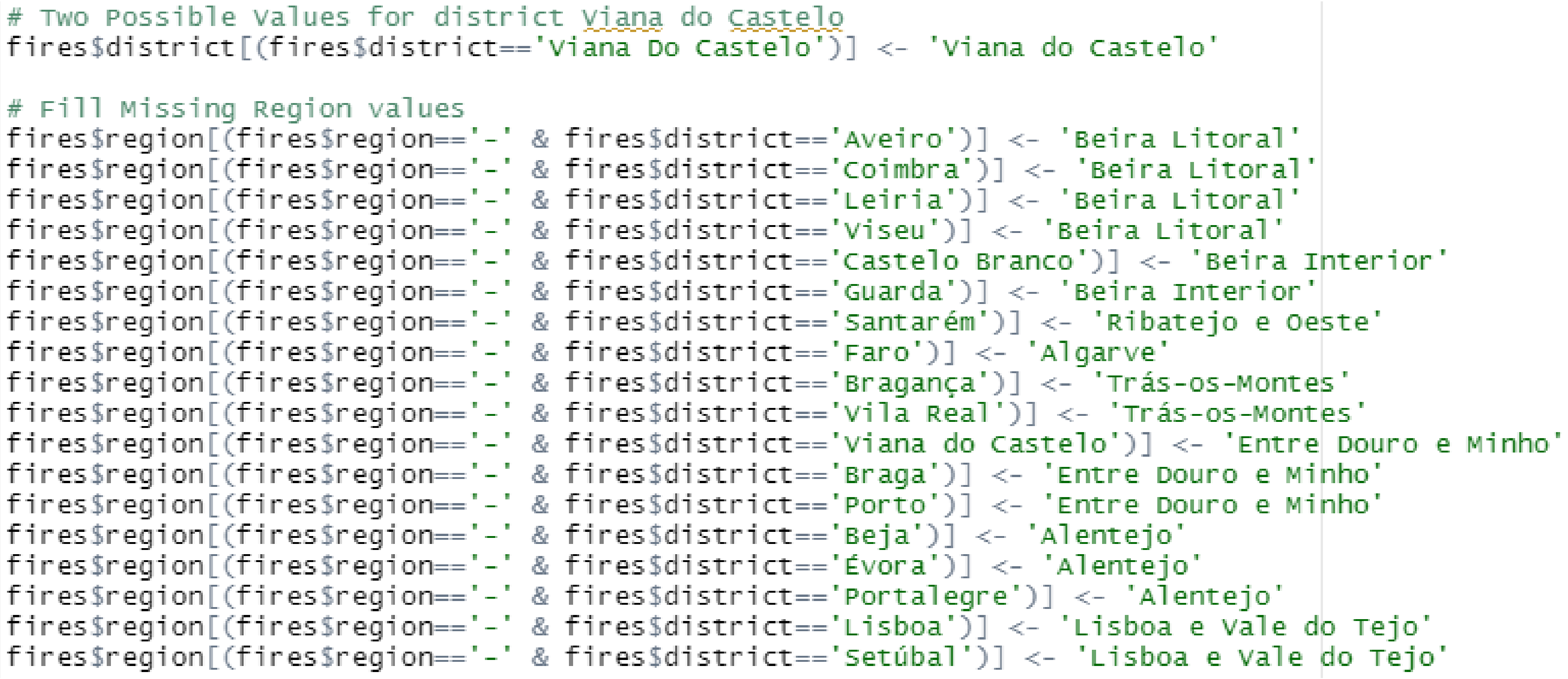


```
extinction_date
Min.   :2014-01-12 00:00:00.0
1st Qu.:2014-09-11 13:00:00.0
Median :2015-05-19 01:00:00.0
Mean   :2015-03-15 03:40:02.7
3rd Qu.:2015-08-01 01:00:00.0
Max.   :2015-12-28 00:00:00.0
NA's   :10
```

The average duration of a fire was calculated, the result of which was about 0.96 (one day). In this way, it was added one day to "alert_date", to fill the missing values of "extinction_date".

Data Preparation | Data Pre-processing

Data Cleaning – Handling Incorrect Values

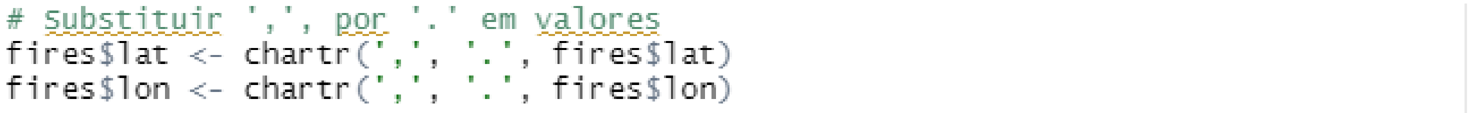


```
# Two Possible Values for district Viana do Castelo
fires$district[(fires$district=='Viana Do Castelo')] <- 'Viana do Castelo'

# Fill Missing Region values
fires$region[(fires$region=='-' & fires$district=='Aveiro')] <- 'Beira Litoral'
fires$region[(fires$region=='-' & fires$district=='Coimbra')] <- 'Beira Litoral'
fires$region[(fires$region=='-' & fires$district=='Leiria')] <- 'Beira Litoral'
fires$region[(fires$region=='-' & fires$district=='Viseu')] <- 'Beira Litoral'
fires$region[(fires$region=='-' & fires$district=='Castelo Branco')] <- 'Beira Interior'
fires$region[(fires$region=='-' & fires$district=='Guarda')] <- 'Beira Interior'
fires$region[(fires$region=='-' & fires$district=='Santarém')] <- 'Ribatejo e Oeste'
fires$region[(fires$region=='-' & fires$district=='Faro')] <- 'Algarve'
fires$region[(fires$region=='-' & fires$district=='Bragança')] <- 'Trás-os-Montes'
fires$region[(fires$region=='-' & fires$district=='Vila Real')] <- 'Trás-os-Montes'
fires$region[(fires$region=='-' & fires$district=='Viana do Castelo')] <- 'Entre Douro e Minho'
fires$region[(fires$region=='-' & fires$district=='Braga')] <- 'Entre Douro e Minho'
fires$region[(fires$region=='-' & fires$district=='Porto')] <- 'Entre Douro e Minho'
fires$region[(fires$region=='-' & fires$district=='Beja')] <- 'Alentejo'
fires$region[(fires$region=='-' & fires$district=='Évora')] <- 'Alentejo'
fires$region[(fires$region=='-' & fires$district=='Portalegre')] <- 'Alentejo'
fires$region[(fires$region=='-' & fires$district=='Lisboa')] <- 'Lisboa e Vale do Tejo'
fires$region[(fires$region=='-' & fires$district=='Setúbal')] <- 'Lisboa e Vale do Tejo'
```

Naming the same district in the same way.

Assigning respective regions to region values that have "-".



```
# Substituir ',', por '.' em valores
fires$lat <- chartr(',', '.', fires$lat)
fires$lon <- chartr(',', '.', fires$lon)
```

Substitution of "," by "." in the coordinate values.

Data Preparation | Data Pre-processing

Data transformation

- Conversion of all coordinates to the decimal form.

Dimensionality Reduction - Feature Selection

- id
- firstInterv_hour
- extinction_hour
- alert_source

Irrelevant variables were removed.

Feature Engineering

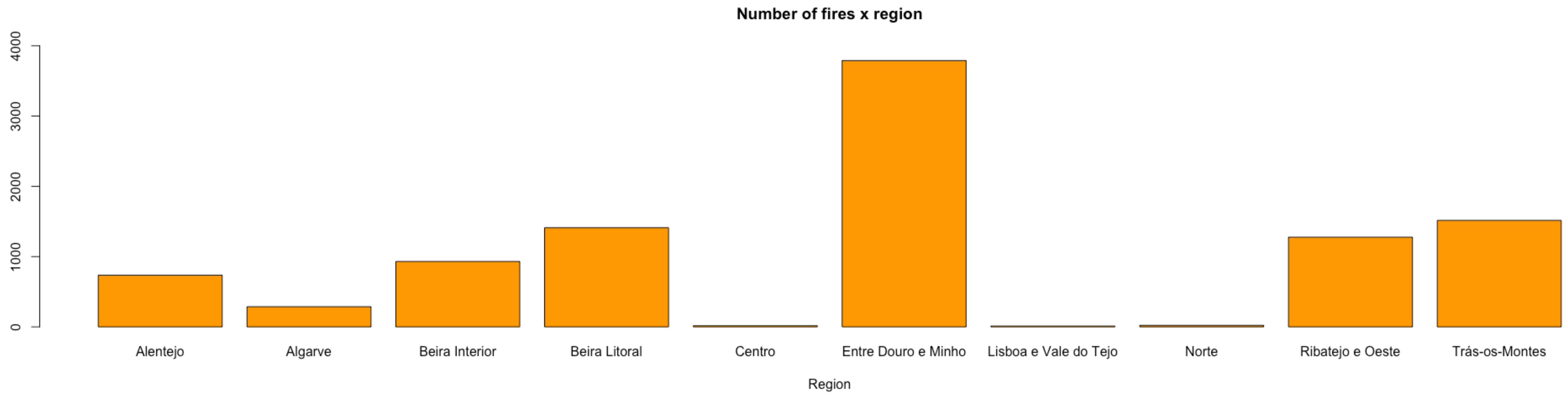
- temp (average temperature)
- tempMax (maximum temperature)
- windGust (wind gust)
- windVelocity (wind velocity)

Through the values of the coordinates ("lat" and "lon") and the alert date ("alert_date") a few more variables were added.

Data Visualization

Which is the region where most fires take place?

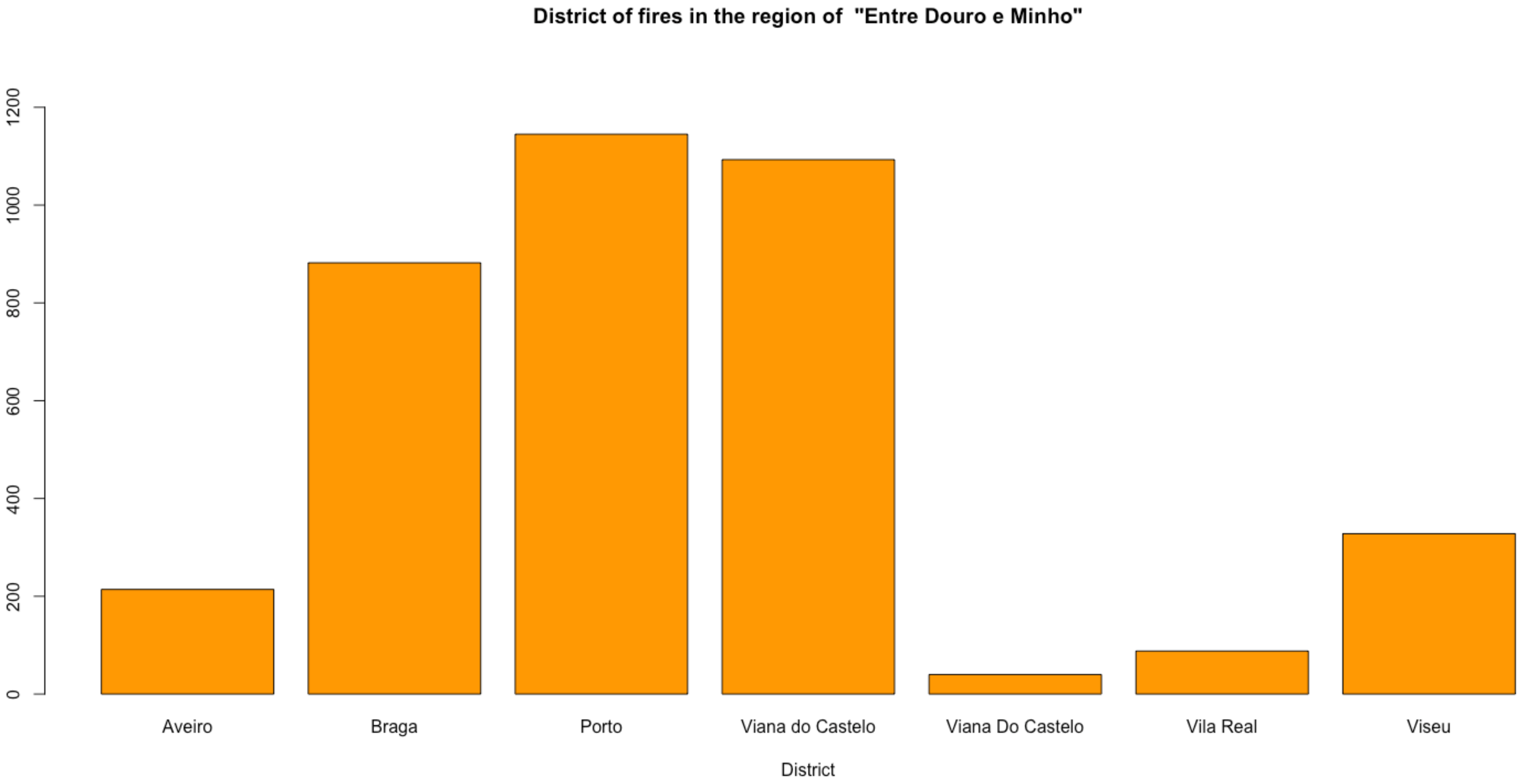
Number of fires that occur on each region



Data Visualization

What are the districts of the region "Entre Douro e Minho" where most fires take place?

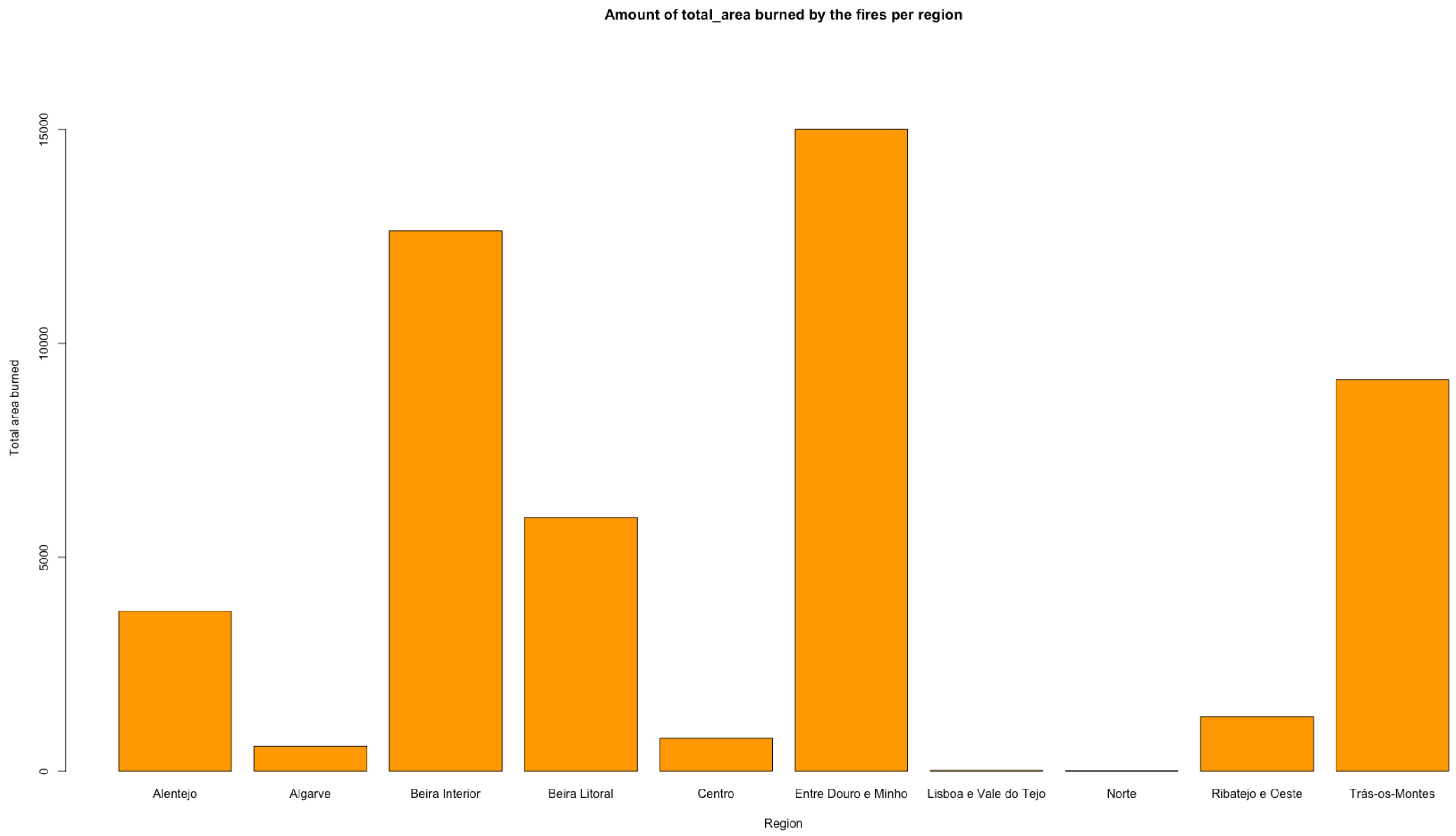
Number of fires that occur on region "Entre Douro e Minho"



Data Visualization

What are the regions where the fires have burned the most area?

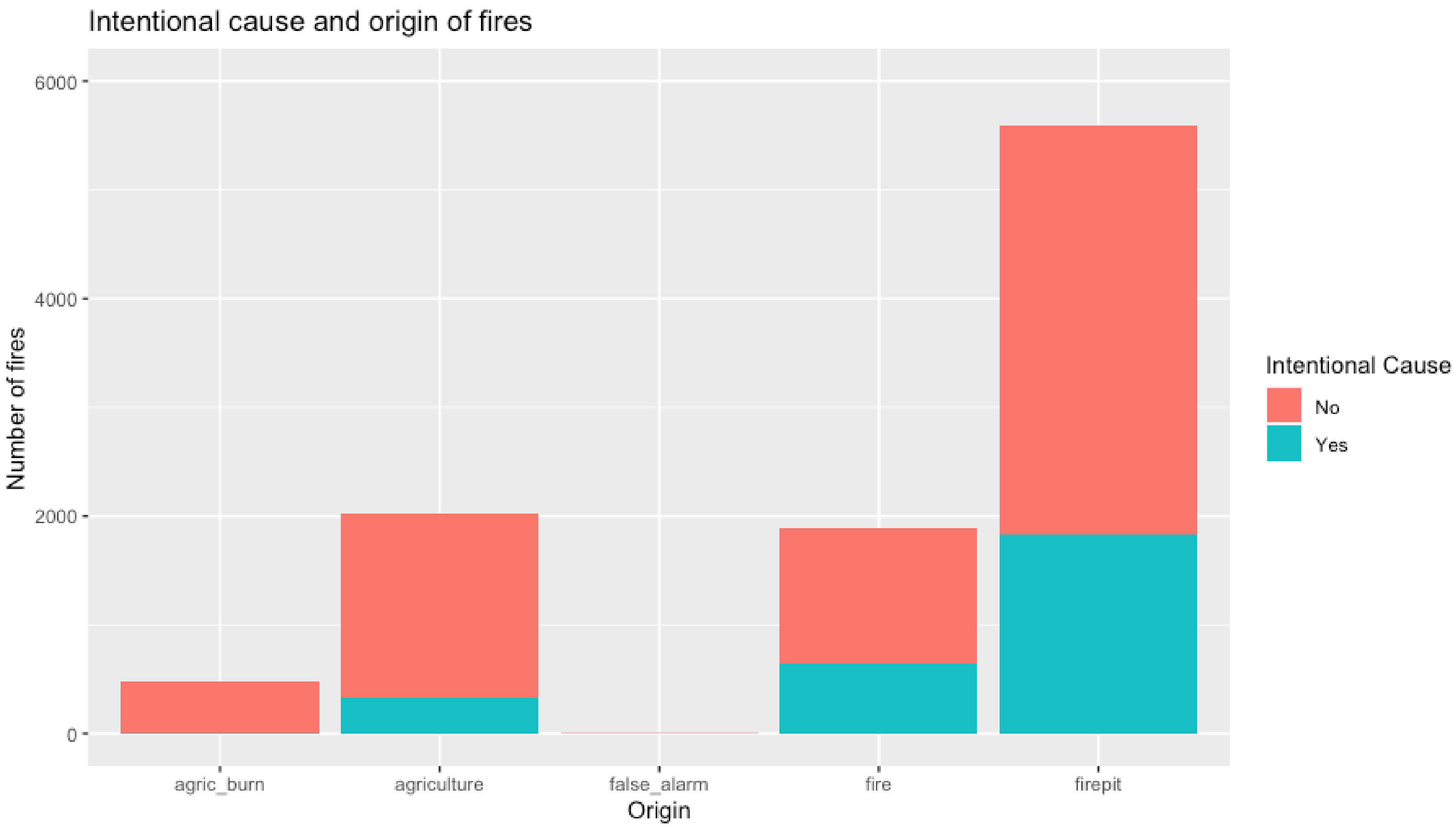
Amount of total area burned by the fires per region



Data Visualization

What are the origin of the fires that were intentional?

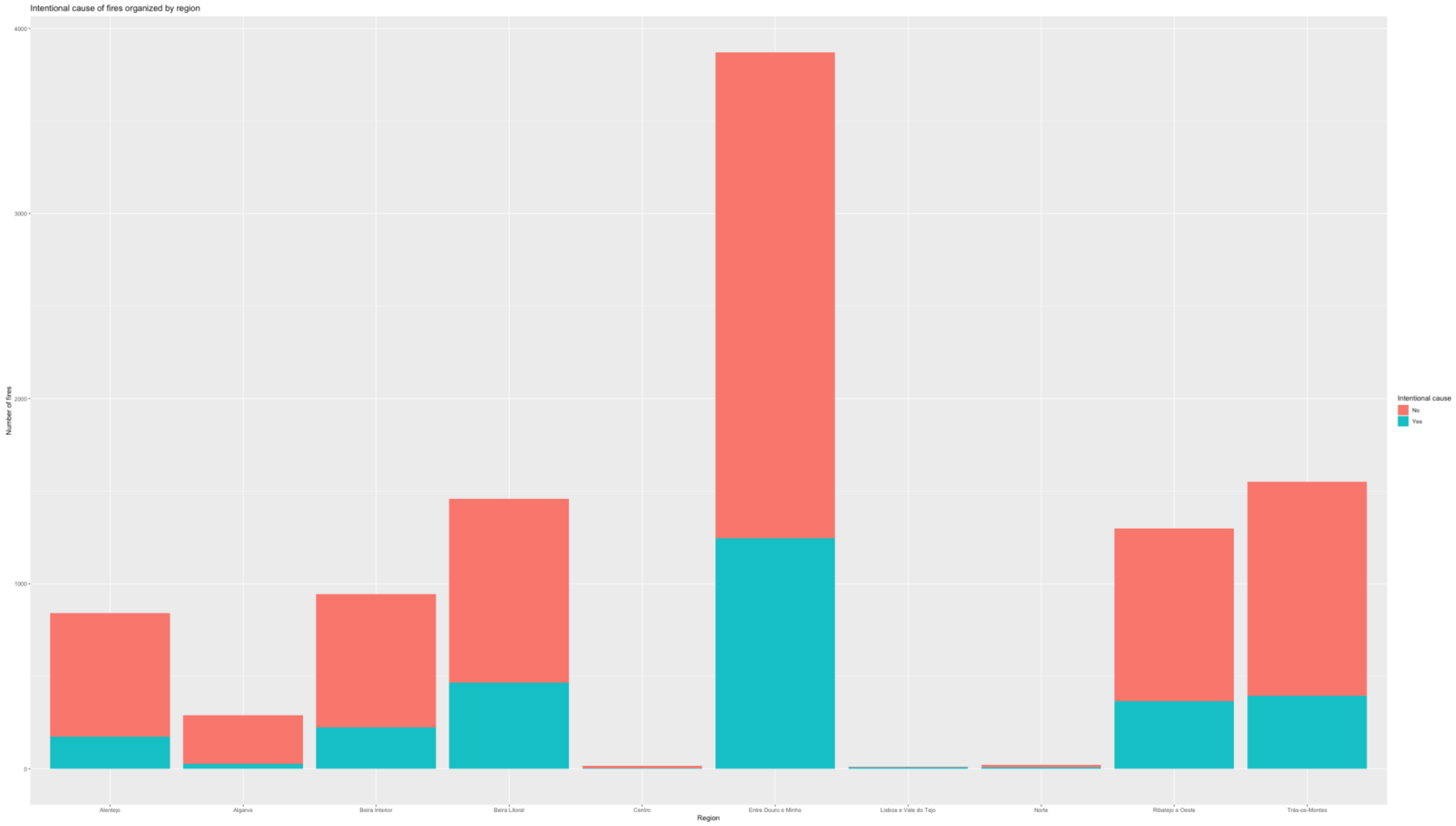
Intentional cause and origin of fires



Data Visualization

How many fires were intentional on each region?

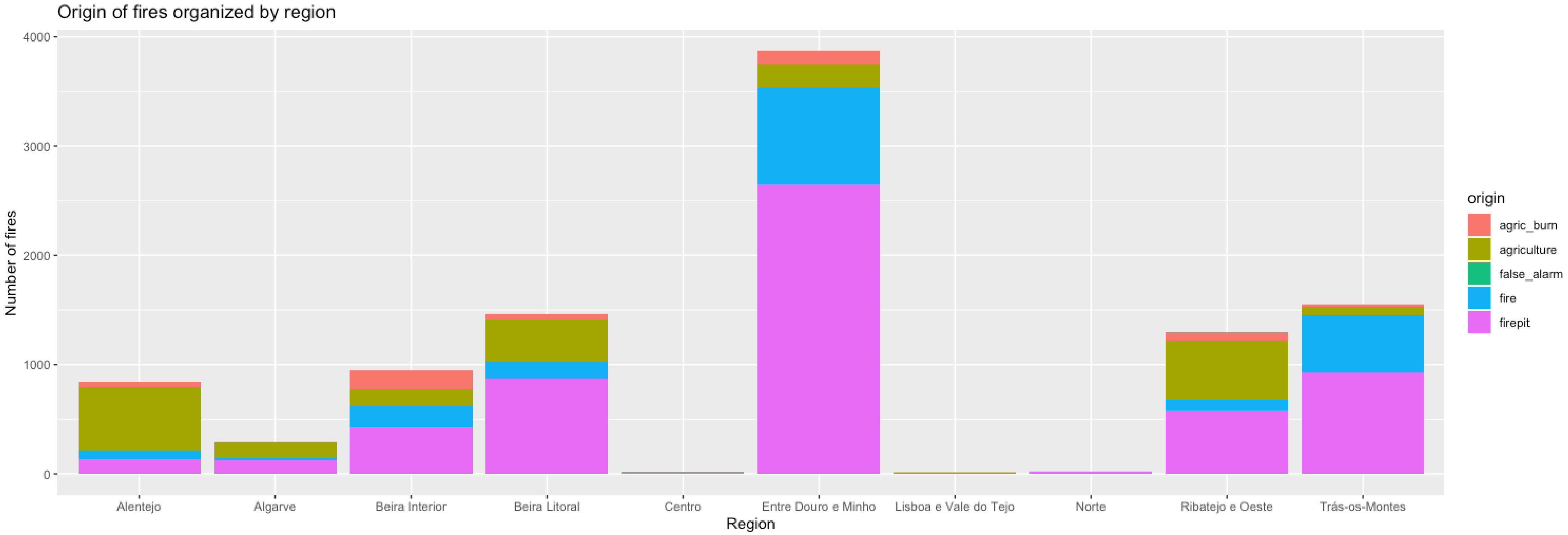
Intentional cause of fires per region



Data Visualization

What are the origin of the fires on each region?

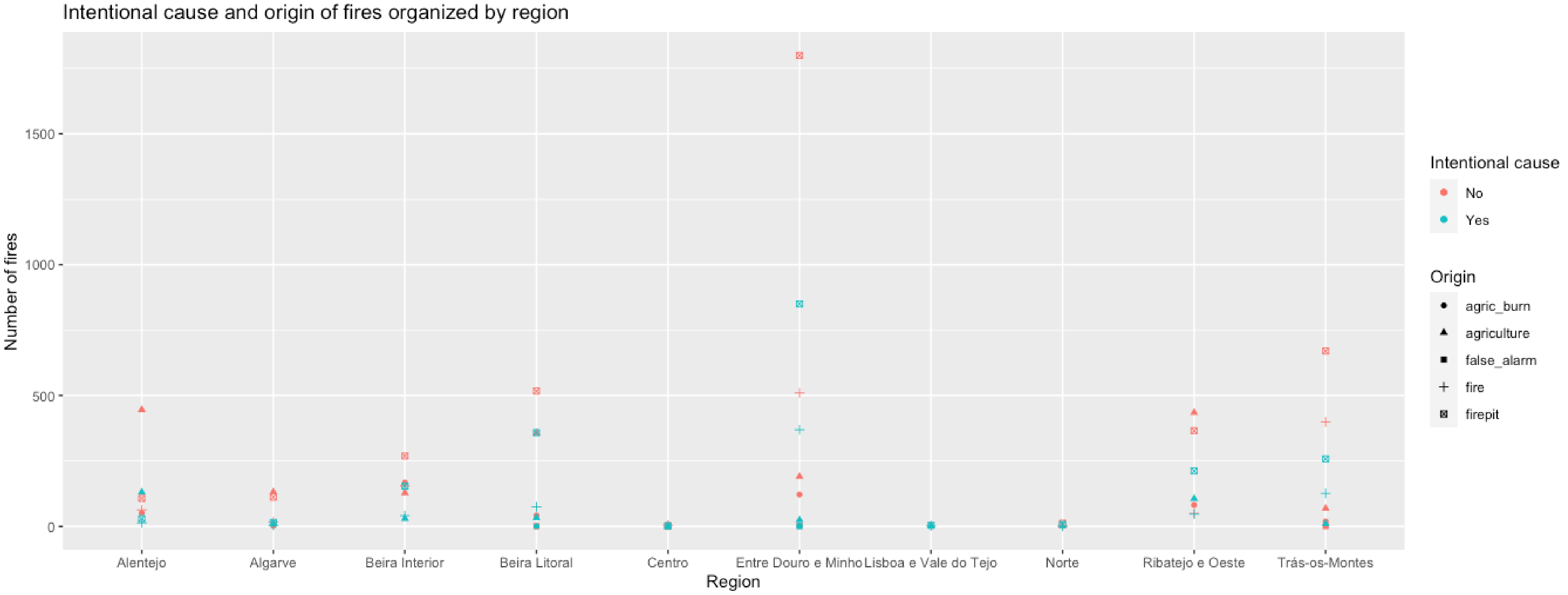
Origin of fires per region



Data Visualization

What are the origin and intentional cause of the fires on each region?

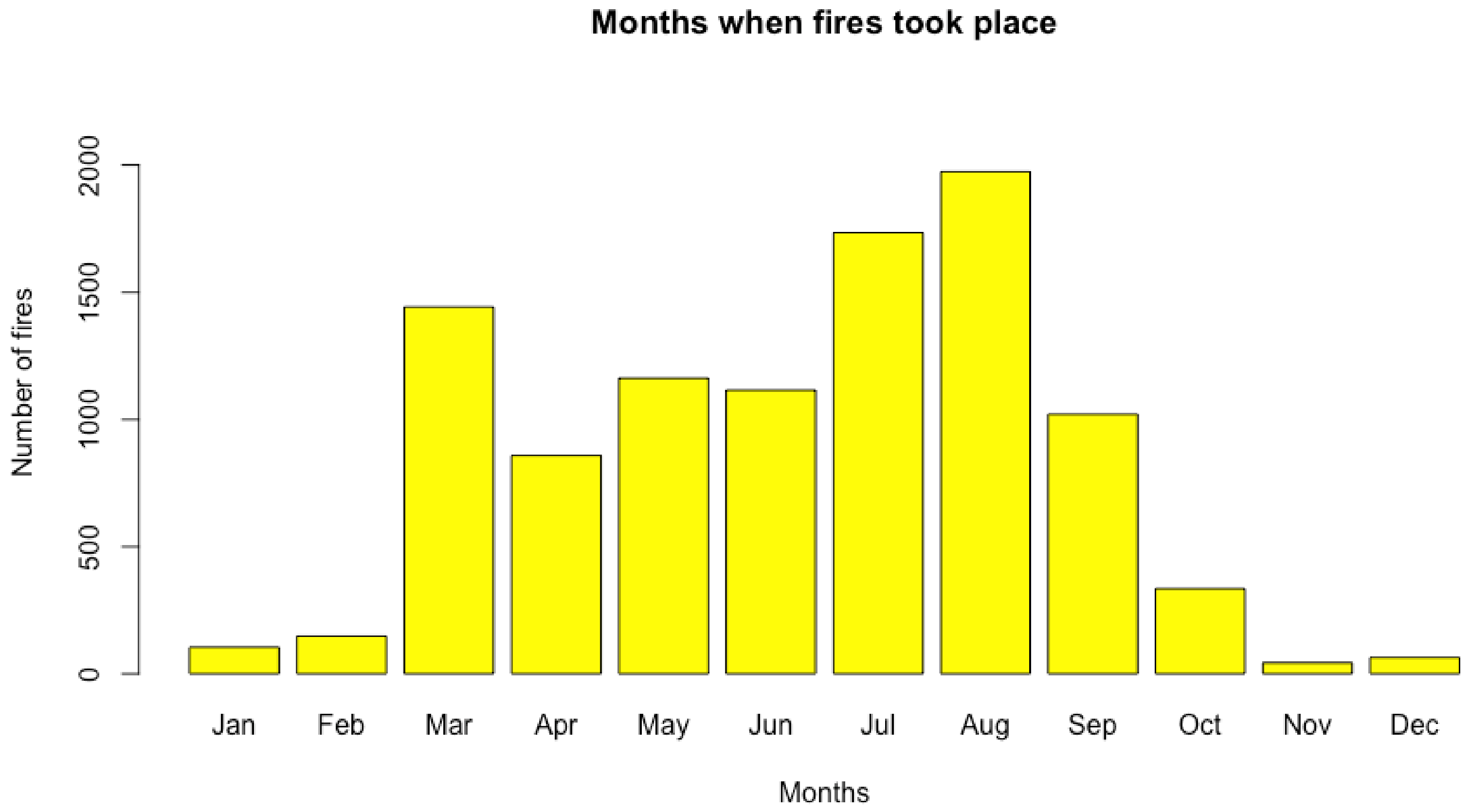
Intentional cause and origin of fires per region



Data Visualization

What are the times of the year where most fires take place?

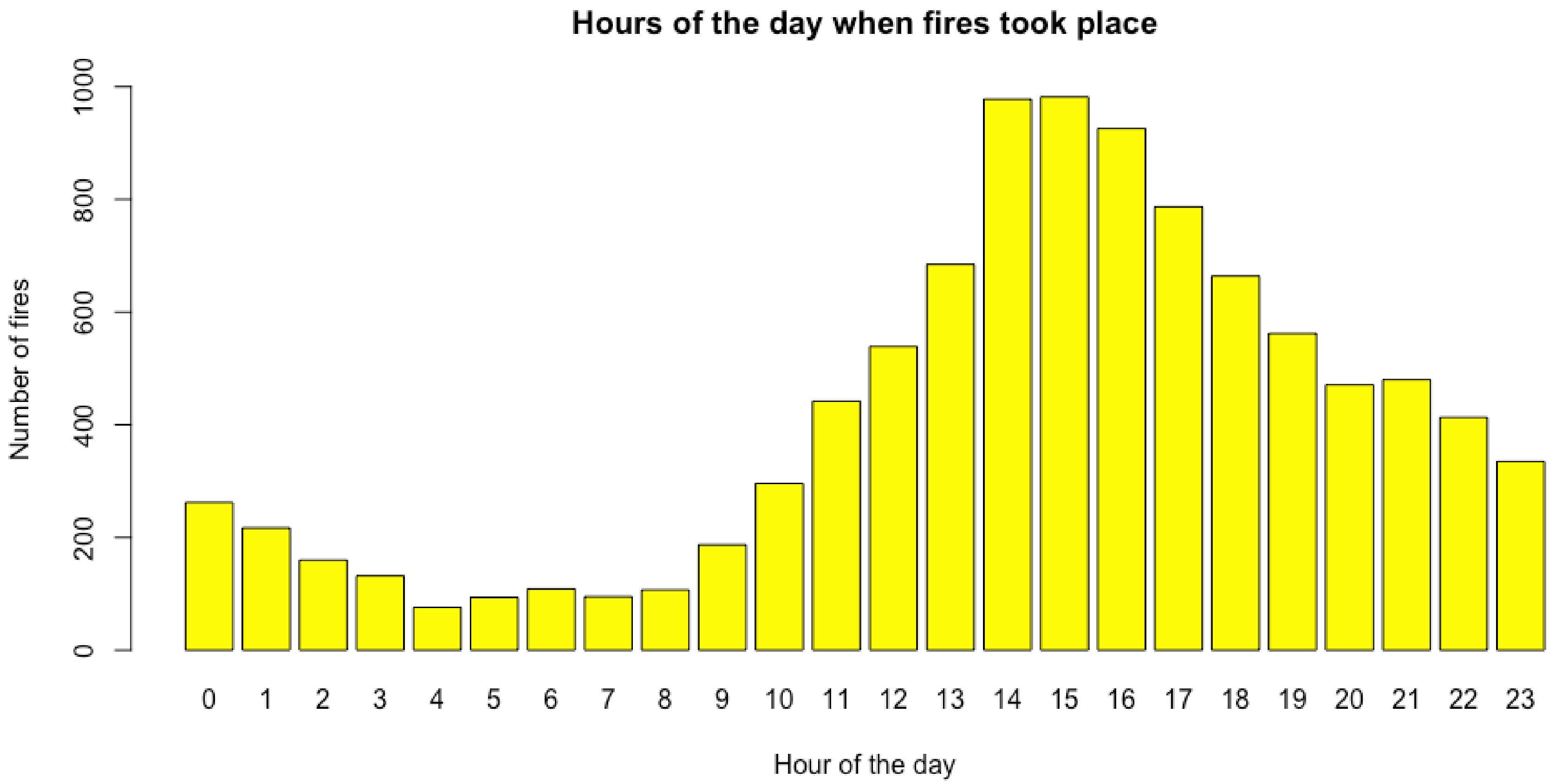
Months of the year when fires took place



Data Visualization

What are the hours of the day where most fires take place?

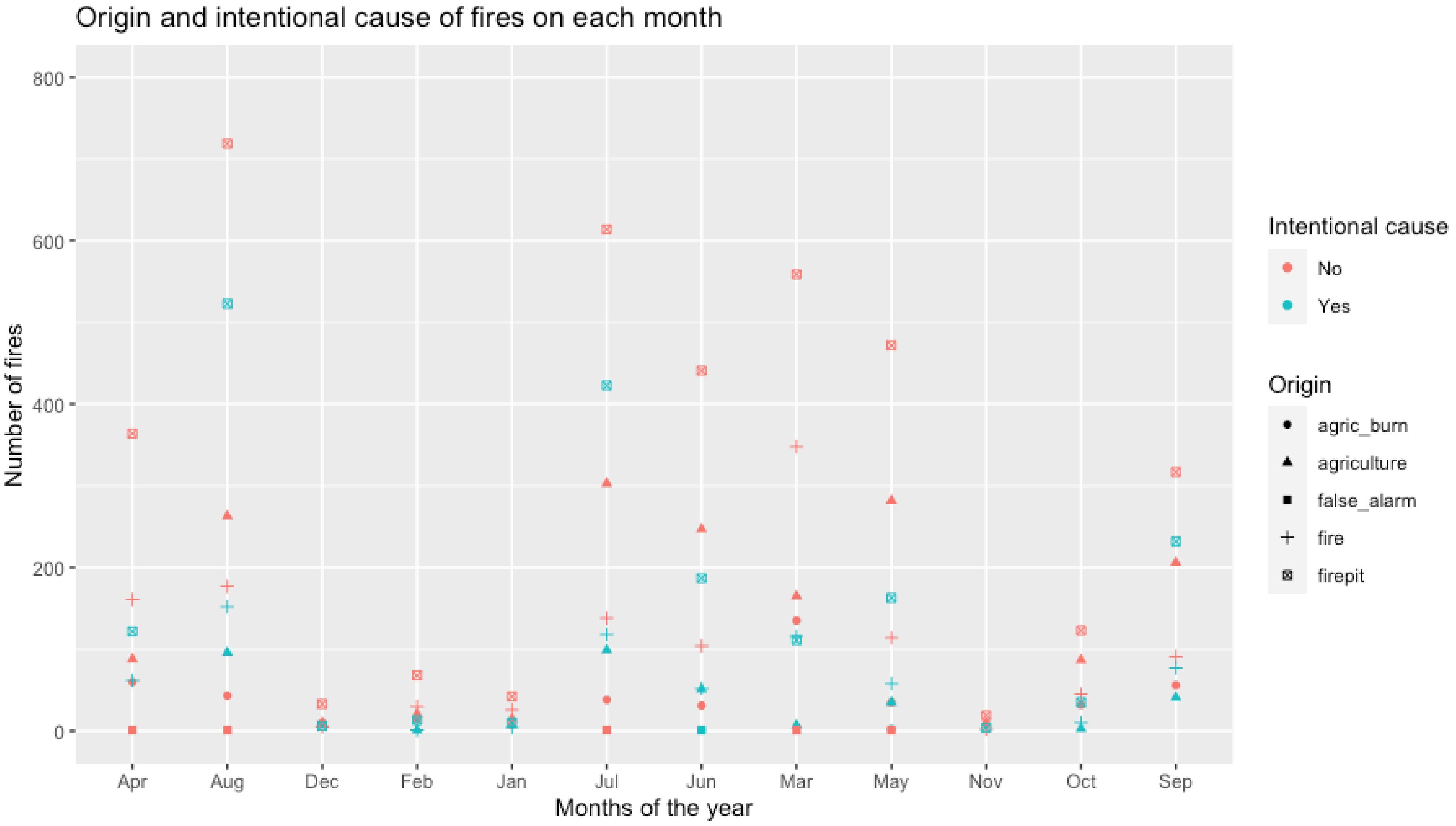
Hours of the day when fires took place



Data Visualization

What are the origin and intentional cause of fires during the year?

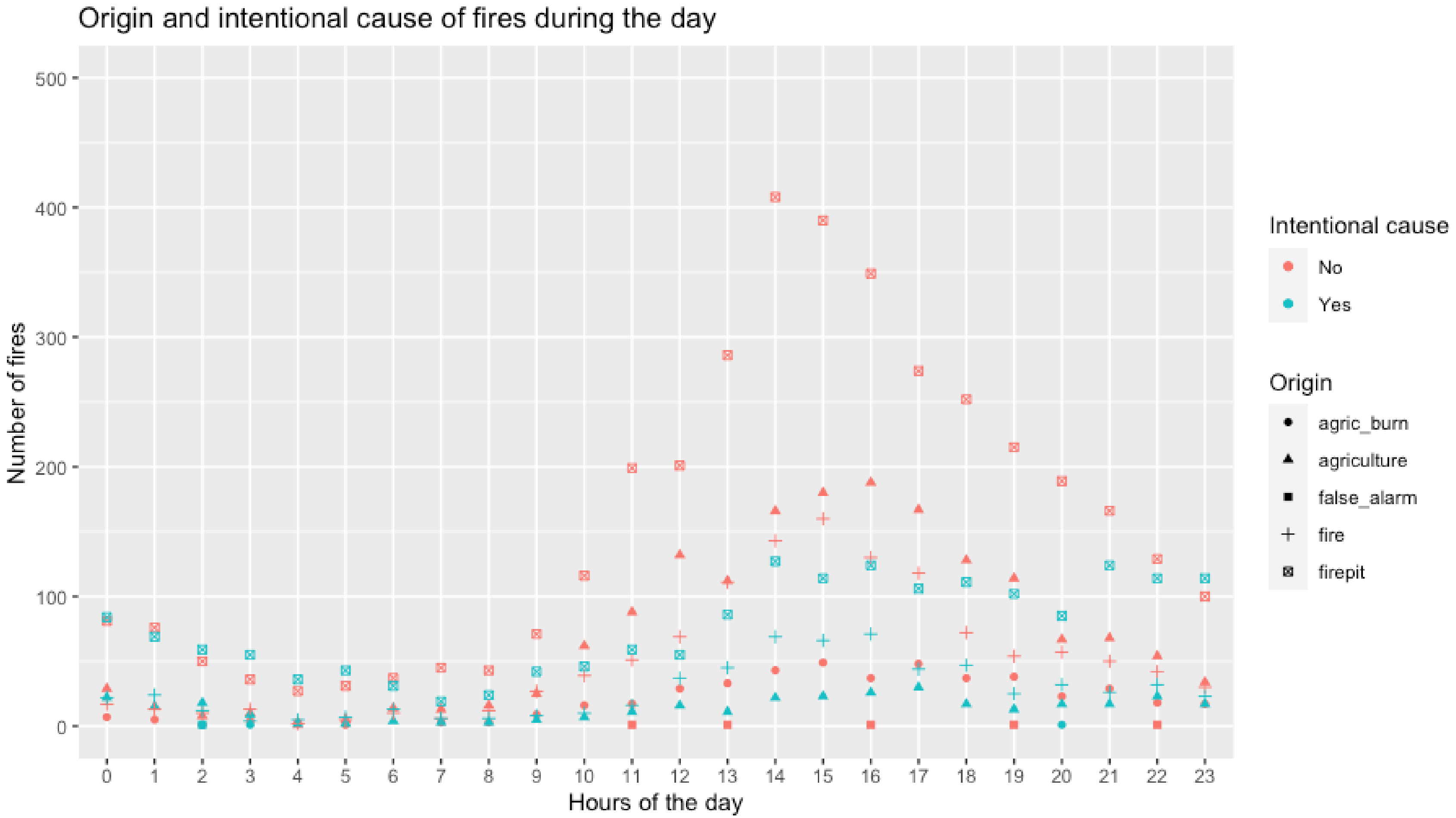
Origin and Intentional Cause of fires that occur during the month



Data Visualization

What are the origin and intentional cause of fires during the day?

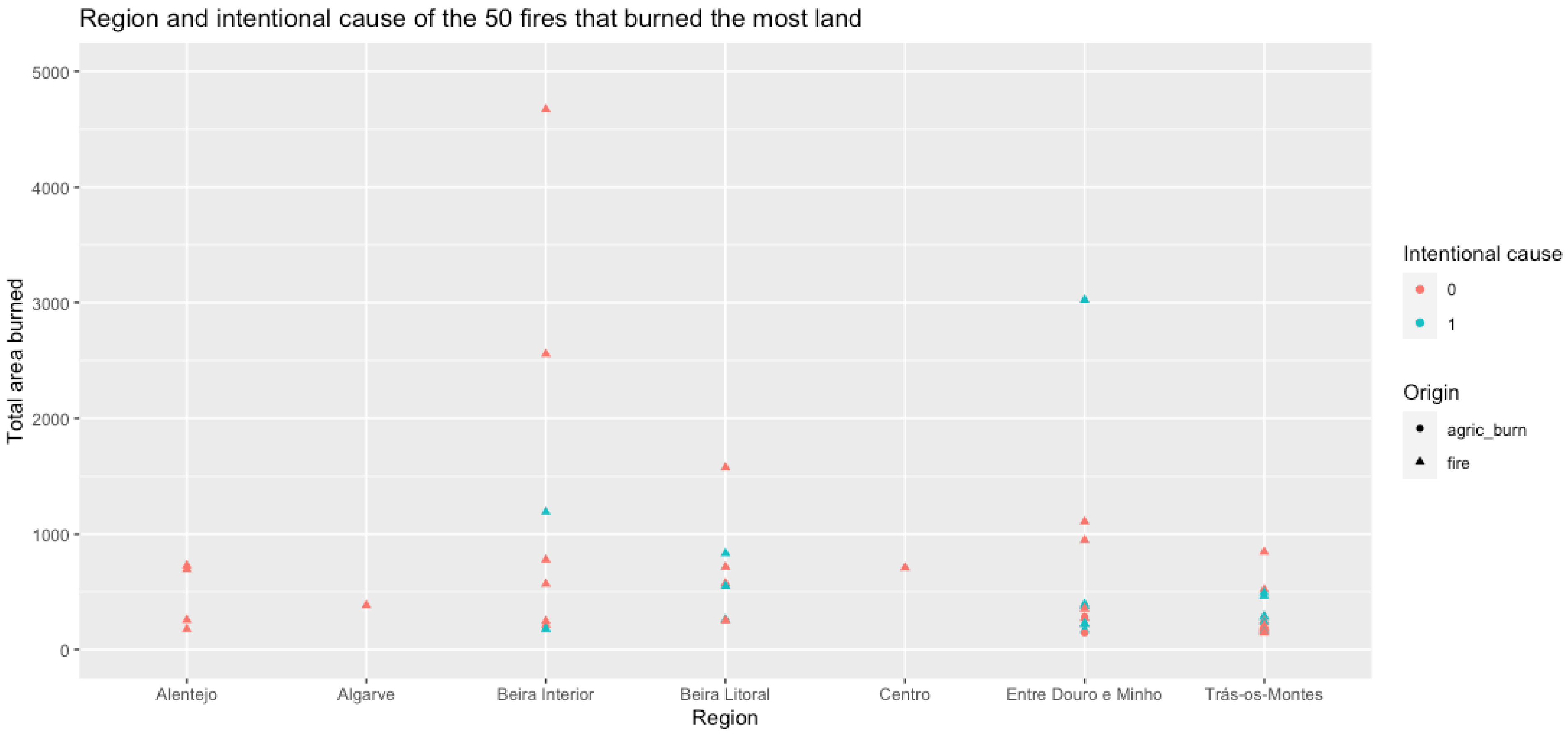
Origin and Intentional Cause of fires that occur during the day



Data Visualization

What are the origin and intentional cause of the fires that burned the most land?

Region and intentional cause of the 50 fires that burned the most area



Predictive Modelling

- Evaluation Metric: AUC (Area under the Curve)
- Train and test Split (70% - 30%)
- k-fold Cross Validation (10 folds)
- Applied recipes where:

  + Irrelevant predictors-> removed

  + Categorical predictors -> converted to numeric values

  + Numeric predictors -> centered and scaled

  + Date predictors -> sometimes included (depends on the model)

  + Variables with large correlations to others -> removed

Predictive Modelling | Best Results

Model	Engine	Parameters			Roc_Auc
		Tuned		Decided	
Logistic Regression	glmnet	Penalty: 0.00053		-	0.731835
Decision Trees CART	rpart	Tree_depth: 4	Min_n: 2	-	0.555673
K-Nearest Neighbors	kknn	Neighbors: 10	Dist_power: 1	-	0.720147
Neural Network	nnet	Hidden_units: 7	Penalty: 1	Epochs: 10	0.723100
Naive Bayes	klaR	Smoothness: 0.75	Laplace: 0	-	0.694749
Random Forest	ranger	Mtry: 4	Min_n: 20	Trees: 100	0.763972
Boosted Trees	xgboost	Mtry: 4	Min_n: 11	Trees: 100	0.744530

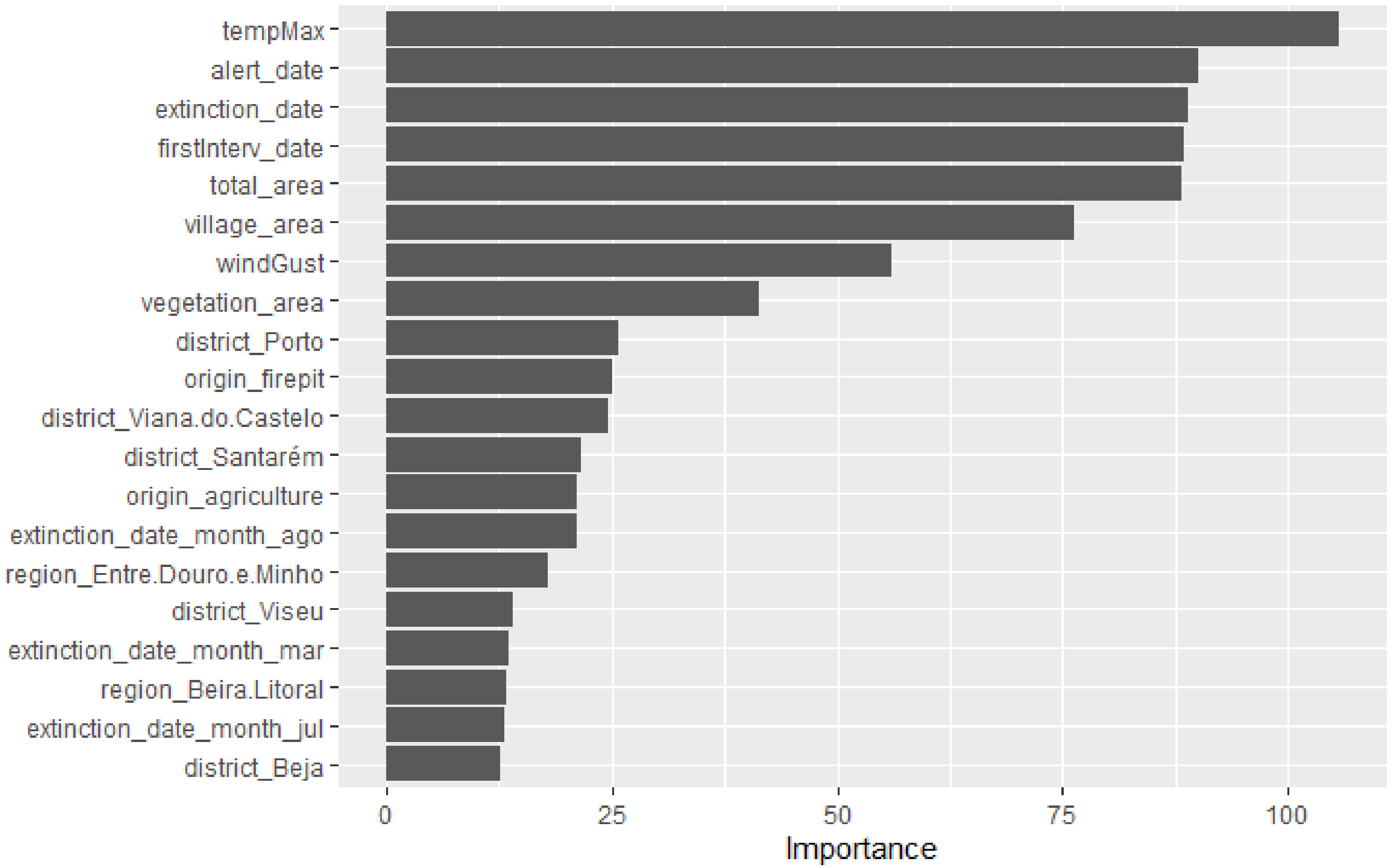
Predictive Modelling | Last Fit

Random Forest

- mtry = 4
- min_n = 20
- trees = 100
- Engine: ranger

-> Roc_auc: 0.7627880

20 Most Important Features

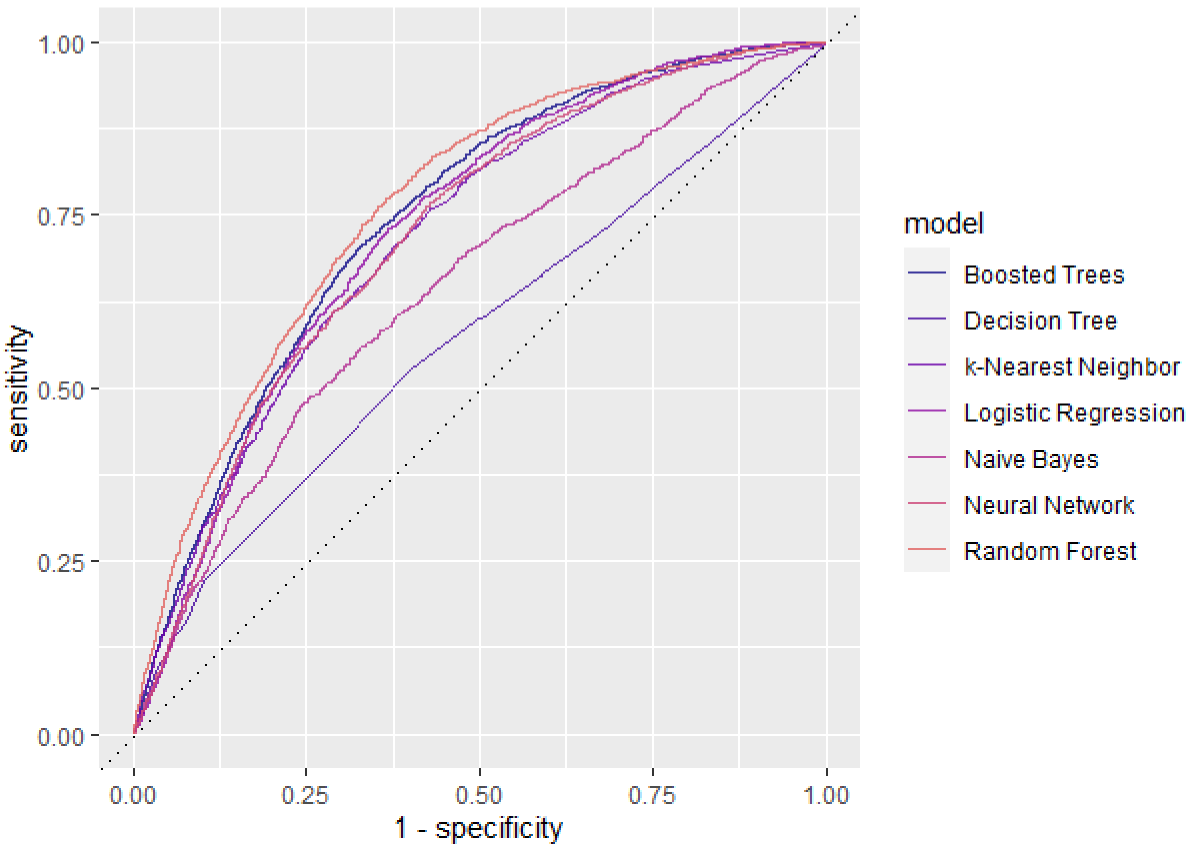


Conclusions, Limitations and Future Work

- The model that achieved the most AUC was Random Forest.
- According to the last fit model, the variables that matter the most (above 25% of importance) are: tempMax; total_area; village_area; windGust; vegetation_area.
- One of the limitations or difficulties was making the data tidy and working in the correct formats that the models needed.
- For future work, more variations of features could be selected for other models to produce better results. Also, different tuning of the parameters could be performed.

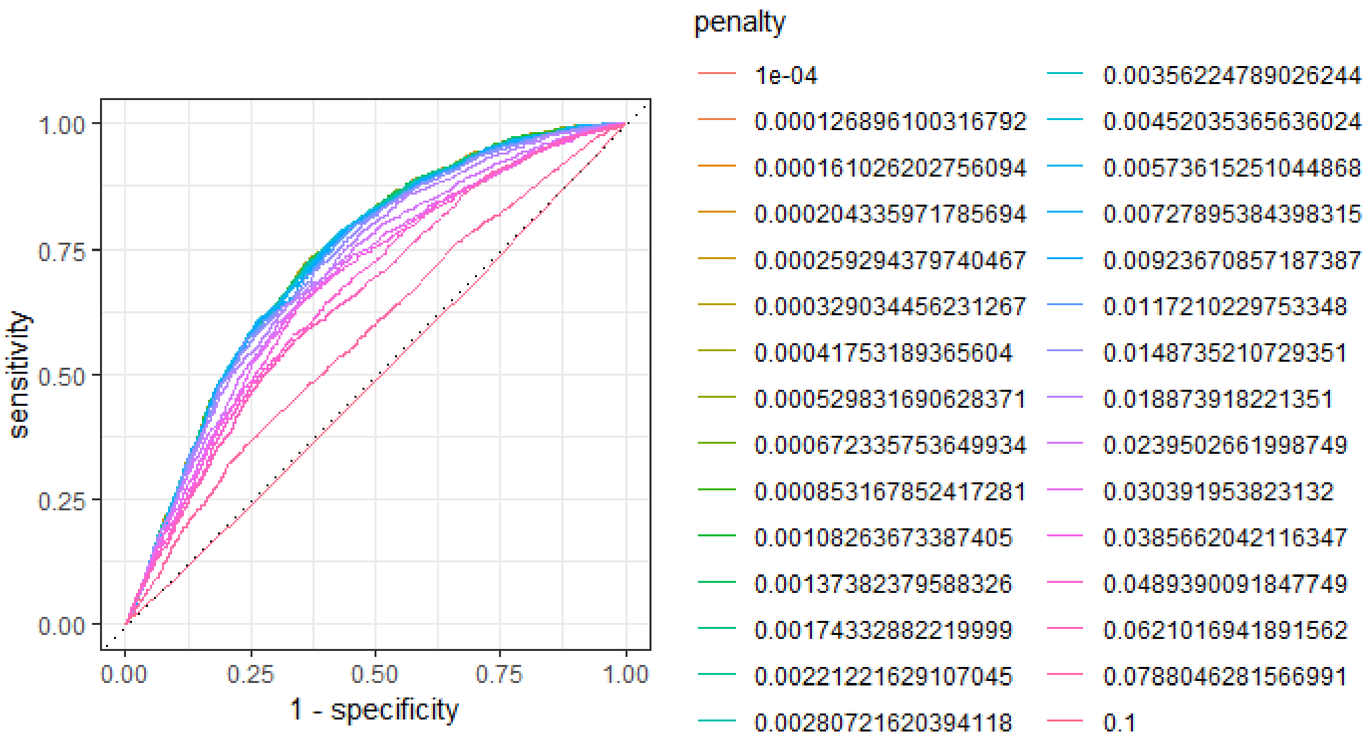
Annexes

AUC of the best model of each type



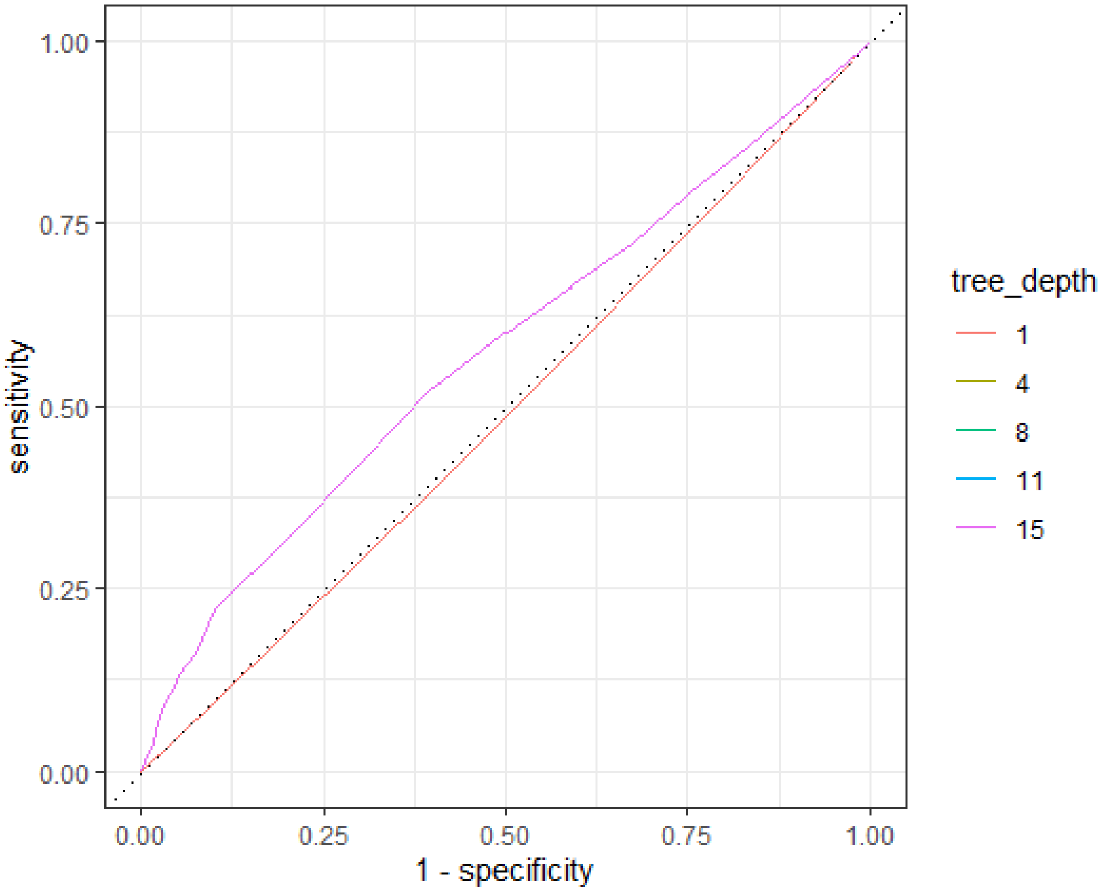
Annexes

Logistic Regression Results



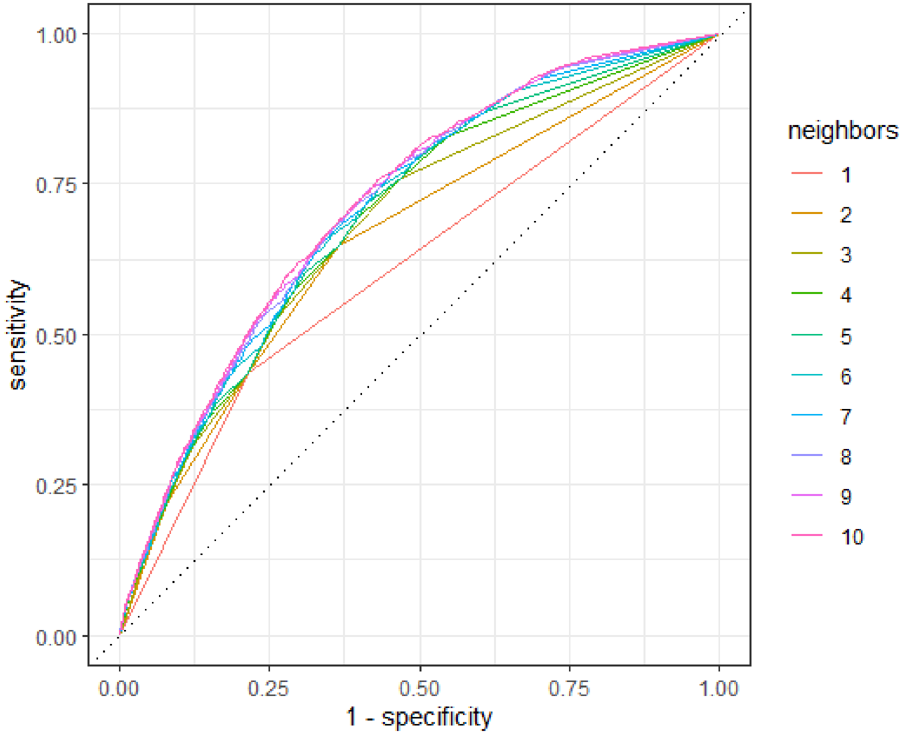
Annexes

Decision Trees CART Results



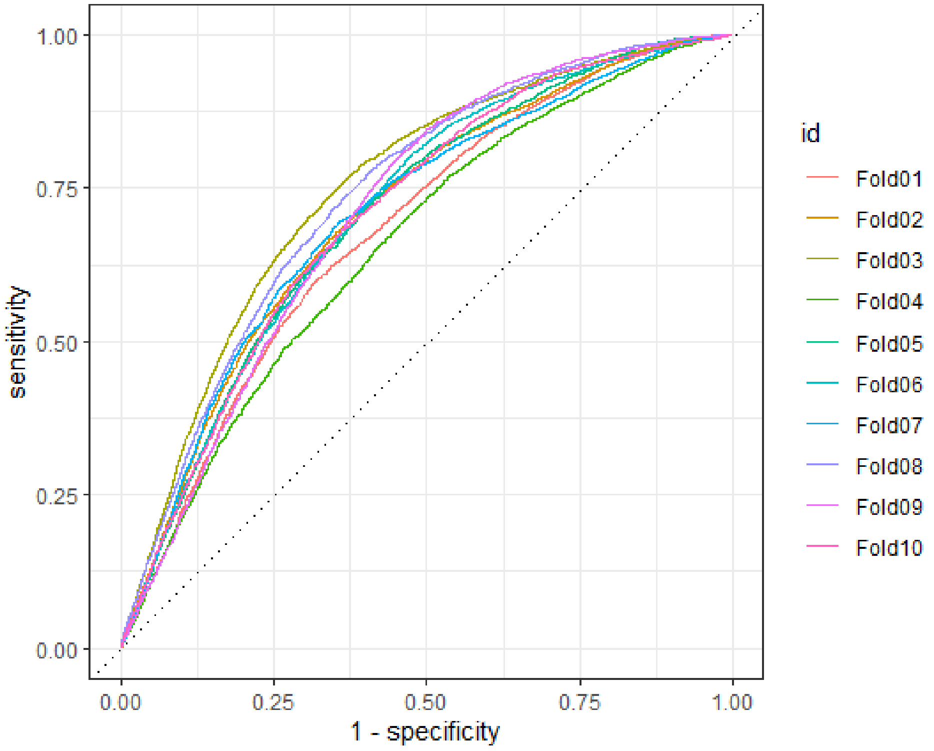
Annexes

K-Nearest Neighbors Results



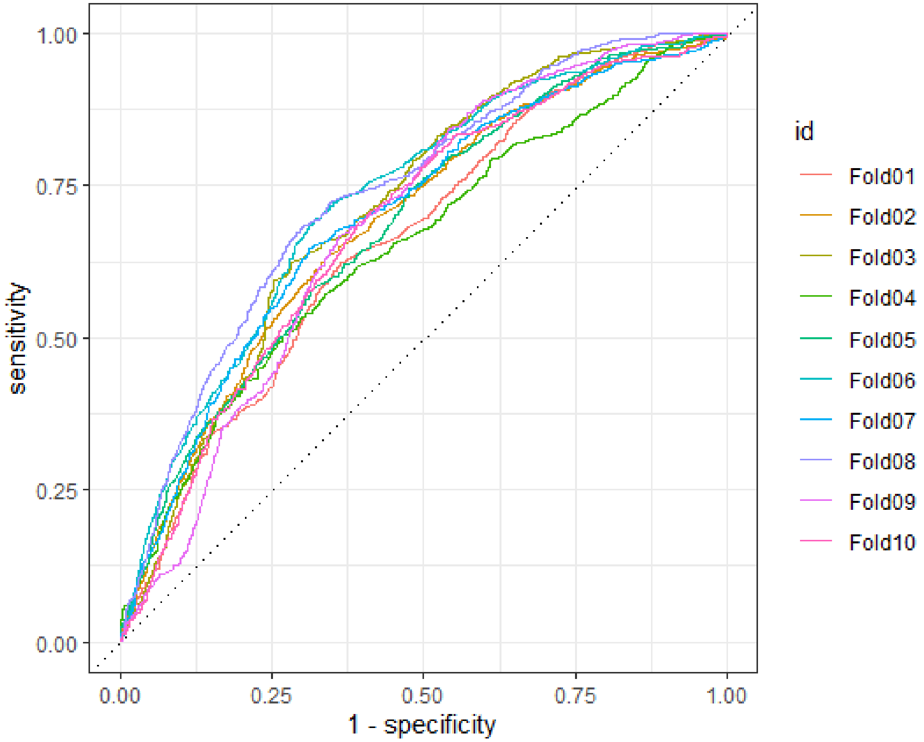
Annexes

Neural Network Results



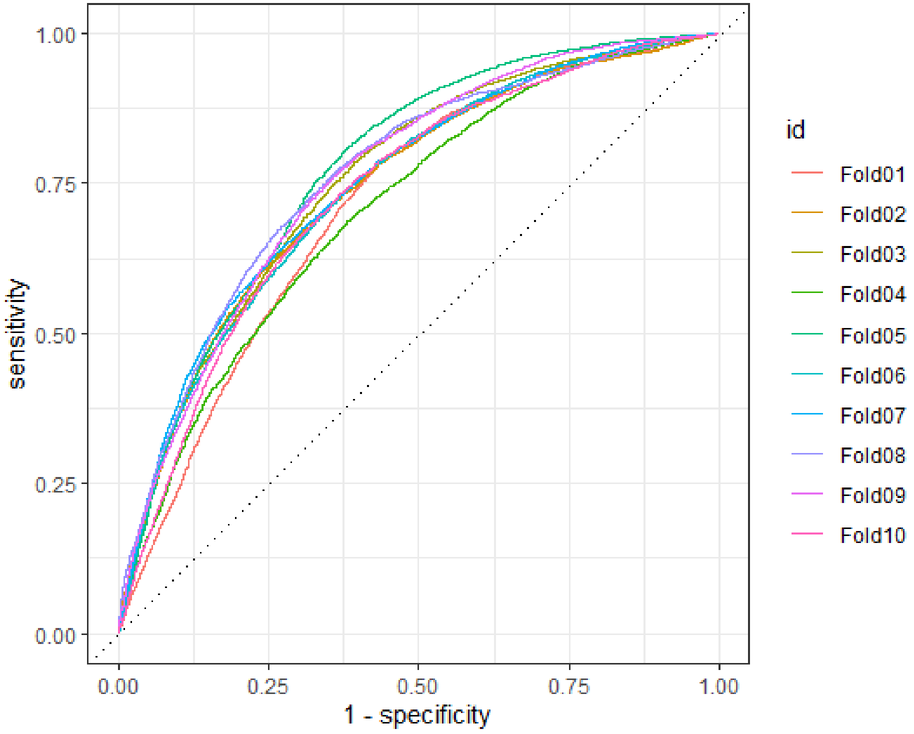
Annexes

Naive Bayes Results



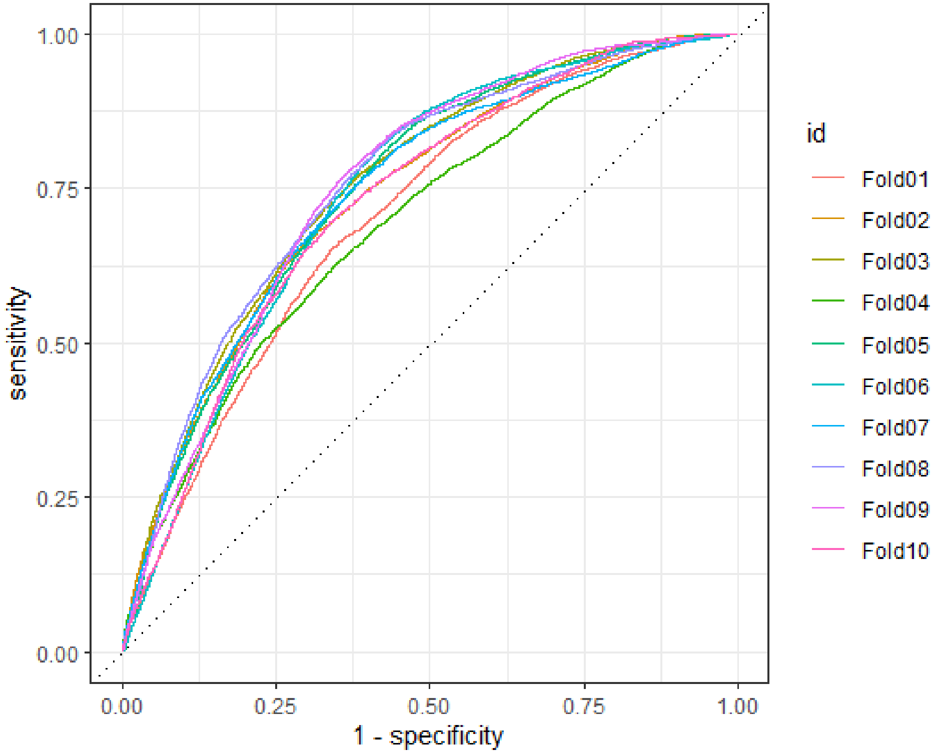
Annexes

Random Forest Results



Annexes

Boosted Trees Results


sensitivity
1 - specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10