Forest Fires in Portugal

Pedro Mota, Tatiana Araújo

Department of Computer Science
University of Porto

January, 2022

Overview

The problem

```
## Rows: 10,309
## Columns: 21
## $ id                 <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
## $ region             <chr> "Trás-os-Montes", NA, "Entre Douro e Minho", "Trás-~
## $ district           <chr> "Bragança", "Viseu", "Aveiro", "Viseu", "Braga", "C~
## $ municipality       <chr> "Bragança", "Oliveira de Frades", "Vale de Cambra",~
## $ parish             <chr> "Zoio", "União das freguesias de Oliveira de Frades~
## $ lat                <chr> "41°44'17''", "40°46'55''", "40°51'11''", "41°4'16'~
## $ lon                <chr> "6°53'34''", "8°14'33''", "8°19'50''", "7°46'22''",~
## $ origin             <chr> "firepit", "firepit", "firepit", "firepit", "firepi~
## $ alert_date         <dttm> 2014-03-16 00:00:00, 2014-03-17 00:00:00, 2014-03-~
## $ alert_hour         <time> 16:15:00, 20:53:00, 12:55:00, 14:22:00, 12:07:00, ~
## $ extinction_date    <dttm> 2014-03-16 00:00:00, 2014-03-17 00:00:00, 2014-03-~
## $ extinction_hour    <time> 17:47:00, 22:46:00, 15:30:00, 15:25:00, 13:14:00, ~
## $ firstInterv_date   <dttm> 2014-03-16 00:00:00, 2014-03-17 00:00:00, 2014-03-~
## $ firstInterv_hour   <time> 16:35:00, 21:05:00, 13:10:00, 14:42:00, 12:07:00, ~
## $ alert_source       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ village_area       <dbl> 0.520, 0.000, 0.000, 0.000, 0.000, 0.073, 0.900, 0.~
## $ vegetation_area    <dbl> 0.0000, 0.0200, 0.0200, 0.0500, 0.2000, 0.0000, 0.0~
## $ farming_area       <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0~
## $ village_veget_area <dbl> 0.5200, 0.0200, 0.0200, 0.0500, 0.2000, 0.0730, 0.9~
## $ total_area         <dbl> 0.5200, 0.0200, 0.0200, 0.0500, 0.2000, 0.0730, 0.9~
## $ intentional_cause  <dbl> 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, ~
```

Data Pre-Processing

Missing values and duplicate data

```
##      division                metrics   value
## 1        size           observations   10309
## 2        size              variables      21
## 3        size                 values  216489
## 4        size            memory size 2998376
## 5  duplicated duplicate observation        0
## 6     missing  complete observation        0
## 7     missing   missing observation   10309
## 8     missing     missing variables        6
## 9     missing        missing values   12157
## 10  data type               numerics       5
## 11  data type               integers       1
## 12  data type        factors/ordered       2
## 13  data type             characters       6
## 14  data type                  Dates       0
## 15  data type               POSIXcts       3
## 16  data type                 others       4
```

Data Cleaning - Missing values

- As one can see, there's no duplicate data, but there are a lot of missing values.
- In the table below, we can find which columns have missing values and how many values are missing, per attribute:

```
## # A tibble: 6 x 6
##   variables         types  missing_count missing_percent unique_count unique_rate
##   <chr>             <chr>          <int>           <dbl>        <int>       <dbl>
## 1 region            chara~          1206          11.7             11   0.00107
## 2 extinction_date   POSIX~            11           0.107          549   0.0533
## 3 extinction_hour   hms               11           0.107         1258   0.122
## 4 firstInterv_date  POSIX~           309           3.00           549   0.0533
## 5 firstInterv_hour  hms              311           3.02          1256   0.122
## 6 alert_source      logic~         10309         100                1   0.0000970
```

Data Cleaning - Missing Values

- The column **alert_source** is all missing values, so we can immediately drop it.
- Regarding the **alert_data**, **extinction_date** and **firstInterv_date** datetime attributes, we assumed that the time field is wrong and we substituted them by the attributes **alert_hour**, **extinction_hour** and **firstInterv_hour** and we called these new attributes **alert_datetime**, **extinction_datetime** and **firstInterv_datetime**, respectively. After that, we imputed 3 missing values that appear after the tranformation, using k-nearest neighboors method.

Data Cleaning - Outliers

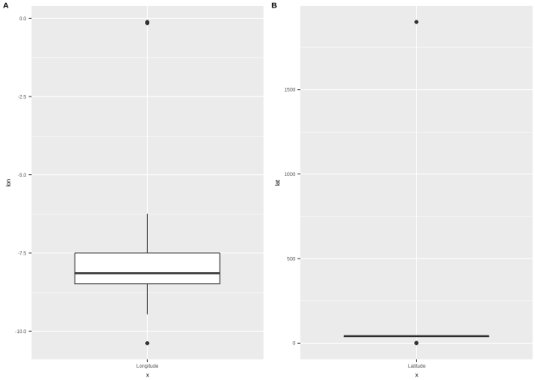
- Regarding the outliers, we found them in the 5 attributes listed below:

```
## # A tibble: 1 x 5
##     lat   lon village_area vegetation_area farming_area
##   <int> <int>        <int>           <int>        <int>
## 1   936   876         2378            1753         2395
```

- We just treated the outliers found before as outliers in the **lat** and **lon** attributes, since we think the rest aren't wrong values and may have important information.
- Also, we found two districts "Viana do Castelo" and "Viana Do Castelo" which are the same district, but were consider different, due to capitalization.

Data Cleaning - Outliers

- For the outliers regarding the attributes **lat** and **lon**, we imputated them using the attributes **region** and **parish**, using the k-nearest neighboors method.



Data Cleaning - Redundant Features

- We've found that the attribute **village_veget_area** and **total_area** are redundant, since they are just the sum of the feature **village_area** and **vegetation_area** and the the sum of the feature **village_area**, **vegetation_area** and **farming_area**, respectively, so we may drop them.

Data Transformation

- Regarding the attributes **lat** and **lon**, the latitude and longitude coordinates of the location of the fire, respectively, they are represented as characters, which isn't optimal for comparisons purposes.
- We thought of:
    - Converting the coordinates into a 3D coordinate space, where we would only have 3 features. Also, in the 3D coordinate space, close points are also close in reality, unlike in the coordinate system, where two extreme values can, actually, be very close together.
    - Converting the coordinates into a decimal representation. In this case, we have just 2 features to represent the coordinates and it's already in the form that will need later, in order to get the temperatures.

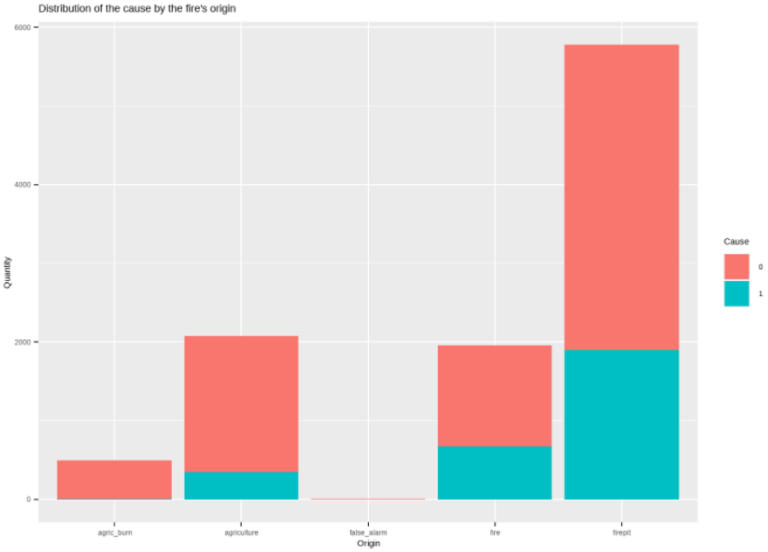
With this in mind and since, in our case, we are only working with latitudes and longitudes within Portugal, which means there no extreme coordinates that are very close in reality, we the chose the decimal representation. Also, we need the coordinates in these form for extracting the temperature and the precipitation.

Feature Engineering

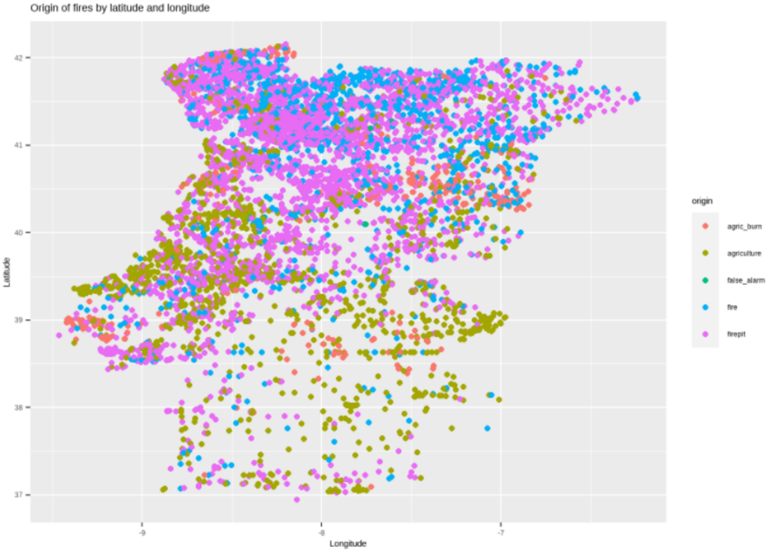
- As we said before, we created new attributes **alert_datetime**, **extinction_datetime** and **firstInterv_datetime** and we think that they can provide relevant information, but not as they are. That is, it's irrelevant the datetime by itself, but the difference, in minutes, within them, may be useful. So we created the **burning_time**, which is the duration, in minutes, of the fire and it's given by the difference between the attribute **extinction_datetime** and **alert_datetime**.
- We created also the **weekday**, **date**, **year**, **hour**, **burned_village_area**, **burned_green_area** attributes and also gather information relative to in the maximum temperature and the precipitation, in the attributes **max_temp** and **prcp**, respectively.

Data Exploration

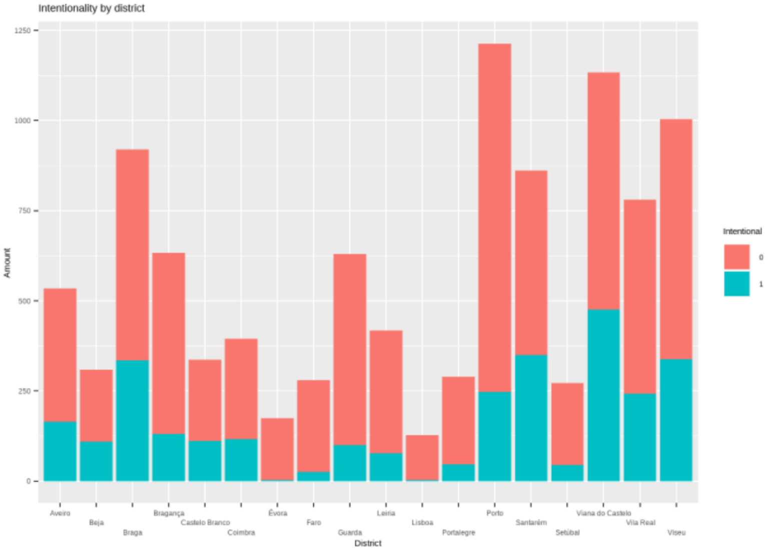
Distribution of the cause by the fire's origin



Origin of fires by latitude and longitude



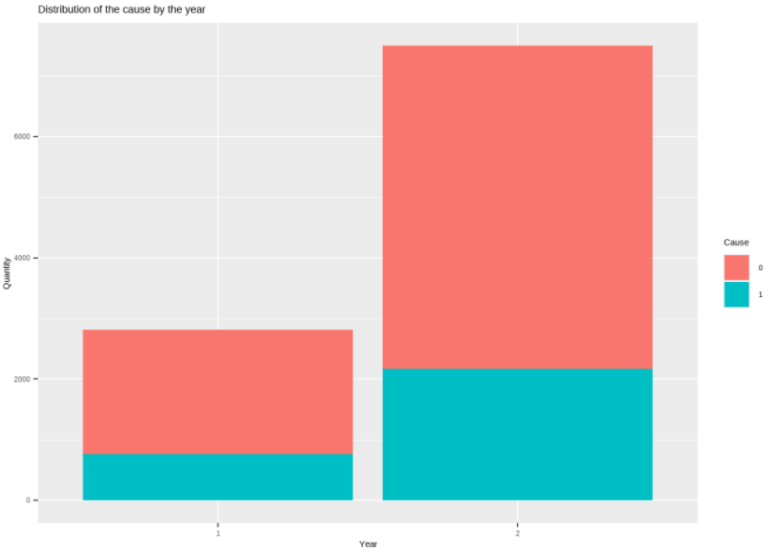
Cause by district



Other statistics

- The “preferred” hour for a fire to start is at 2PM.
- On average, fires burned about 2 hours and half.
- On average, fires burned about 1.85km of village area and 3.07km of green area.
- On average, the temperatures got higher in 2014 that in 2015.

Other statistics



Predictive Modelling - Random Forest

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1982  454
##          1  234  422
##
##                Accuracy : 0.7775
##                  95% CI : (0.7624, 0.792)
##     No Information Rate : 0.7167
##     P-Value [Acc > NIR] : 9.019e-15
##
##                   Kappa : 0.407
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8944
##             Specificity : 0.4817
##          Pos Pred Value : 0.8136
##          Neg Pred Value : 0.6433
##              Prevalence : 0.7167
##          Detection Rate : 0.6410
##    Detection Prevalence : 0.7878
##       Balanced Accuracy : 0.6881
##
##        'Positive' Class : 0
##
```

Conclusion

The biggest challenge was in the data pre-processing and feature engineering part. Especially, in the feature engineering part, were we tried to use some domain knowledge. Future work could pass from creating new features, re-check the discard predictors and gather more data relatively to the fires, in order to improve our classifications models. With a good model, the authorities could use this in their research, in order to combate the criminals and the deforestation due to fires.

Appendix

- In the appendix, we will show the results we obtained for the other models used, besides the Random Forest, which were the model with the highest accuracy.

k-Nearest Neighbors

```
## k-Nearest Neighbors
##
## 10309 samples
##    13 predictor
##     2 classes: '0', '1'
##
## Pre-processing: centered (37), scaled (37)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 9278, 9278, 9278, 9278, 9278, 9278, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    7  0.7190799  0.2240025
##    9  0.7197581  0.2168339
##   11  0.7225717  0.2150496
##   13  0.7192728  0.1961471
##   15  0.7221833  0.1922980
##   17  0.7220866  0.1866709
##   19  0.7202435  0.1735947
##   21  0.7206321  0.1707769
##   23  0.7198554  0.1635645
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 11.
```

As we can see, the highest accuracy we can get is of about 72%, with k = 17.

Naive Bayes

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1596  620
##          1  384  492
##
##                Accuracy : 0.6753
##                  95% CI : (0.6585, 0.6918)
##     No Information Rate : 0.6404
##     P-Value [Acc > NIR] : 2.481e-05
##
##                   Kappa : 0.2606
##
##  Mcnemar's Test P-Value : 1.202e-13
##
##             Sensitivity : 0.8061
##             Specificity : 0.4424
##          Pos Pred Value : 0.7202
##          Neg Pred Value : 0.5616
##              Prevalence : 0.6404
##          Detection Rate : 0.5162
##    Detection Prevalence : 0.7167
##       Balanced Accuracy : 0.6243
##
##        'Positive' Class : 0
##
```

Decision Tree

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2095  121
##          1  677  199
##
##                Accuracy : 0.7419
##                  95% CI : (0.7261, 0.7573)
##     No Information Rate : 0.8965
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2135
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7558
##             Specificity : 0.6219
##          Pos Pred Value : 0.9454
##          Neg Pred Value : 0.2272
##              Prevalence : 0.8965
##          Detection Rate : 0.6776
##    Detection Prevalence : 0.7167
##       Balanced Accuracy : 0.6888
##
##        'Positive' Class : 0
##
```

AdaBoost

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2002  497
##          1  214  379
##
##                Accuracy : 0.7701
##                  95% CI : (0.7548, 0.7848)
##     No Information Rate : 0.7167
##     P-Value [Acc > NIR] : 1.033e-11
##
##                   Kappa : 0.3725
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9034
##             Specificity : 0.4326
##          Pos Pred Value : 0.8011
##          Neg Pred Value : 0.6391
##              Prevalence : 0.7167
##          Detection Rate : 0.6475
##    Detection Prevalence : 0.8082
##       Balanced Accuracy : 0.6680
##
##        'Positive' Class : 0
##
```

XGBoost

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1925  422
##          1  291  454
##
##                Accuracy : 0.7694
##                  95% CI : (0.7541, 0.7842)
##     No Information Rate : 0.7167
##     P-Value [Acc > NIR] : 1.822e-11
##
##                   Kappa : 0.4053
##
##  Mcnemar's Test P-Value : 1.124e-06
##
##             Sensitivity : 0.8687
##             Specificity : 0.5183
##          Pos Pred Value : 0.8202
##          Neg Pred Value : 0.6094
##              Prevalence : 0.7167
##          Detection Rate : 0.6226
##    Detection Prevalence : 0.7591
##       Balanced Accuracy : 0.6935
##
##        'Positive' Class : 0
##
```

The End