Forest Fires in Portugal - What Are The Causes?

Practical Assignment of Data Mining I

By Robson Teixeira, Eduardo Rodrigues and Claudio Rocha

M:CC – FCUP, 10/01/2021

Contents

Chapter 1

Introduction

In this project, we try to find the best machine learning model that more accurately predicts whether a forest fire occurs negligently, intentionally, naturally or recurrently. From a database that was given to us, we divided the work into several parts.

The remainder of this report is organized as follows: in chapter 2, we describe the importance of predicting forest fires that are a big problem actually; in chapter 3 is described the causes ofnthe ocurrences that is the variable that the model will be predict and a table with all the variables of the original dataset; chapter 4 is dedicated to the exploration, cleaning and engineering of the data. Some graphics are plotted in this chapter and they help to visualize the origins and locations of the forest fires; the models used to test the dataset are described and compared in chapter 5; in chapter 6, We finalize with the main conclusions and the last chapter includes references.

Chapter 2

Problem Definition

Forest fires are a very important issue that negatively affects climate change. Typically, the causes of forest fires are those oversights, accidents and negligence committed by individuals, intentional acts and natural causes. The latter is the root cause for only a minority of the fires.

Their harmful impacts and effects on ecosystems can be major ones. Among them, we can mention the disappearance of native species, the increase in levels of carbon dioxide in the atmosphere, earth's nutrients destroyed by the ashes, and the massive loss of wildlife.

Data mining techniques can help in the prediction of the cause of the fire and, thus, better support the decision of taking preventive measures in order to avoid tragedy. In effect, this can play a major role in resource allocation, mitigation and recovery efforts.

Chapter 3

Forest Fire Dataset

The Institute for Nature Conservation and Forests (ICNF) is the governmental body responsible for the nature and forest policies, including the management of protected areas and state managed national, municipal, and communal forests of mainland Portugal. The ICNF has been maintained a database with data of all forest fires that occurred in Portugal over several years. The data set used in this study is a subset extracted from this database regarding the fires that occurred over 2015. It consist of **7511** records of fires and for each one, there is relevant information such as the GPS coordinates (latitude and longitude) where occur the fire, the date and time of fire alert, the date and time of the first intervention, and the date and time of fire extinction, besides the origin of the ignition, the affected area, and the cause type. The table 3 describes all variables contained in `Forest Fires` data set:

Table List of variables in `Forest Fires` data set.

Variable	Type	Description
id	integer	id number
region	character	region name
district	character	district name
municipality	character	municipality name
parish	character	parish name
lat	character	latitude value
lon	character	longitude value
origin	character	how the fire started
alert_date	character	date when fire started
alert_hour	character	alert hour
extinction_date	character	date of the end of fire
extinction hour	character	hour of the end of fire
firstInterv_date	character	date of intervention
firstInterv_hour	character	hour of intervention
alert_source	logical	alert source
village_area	numeric	village area affected
vegetation_area	numeric	vegetation area affected

Variable	Type	Description
farming_area	numeric	farming area affected
village_veget_area	numeric	total village+veget affected
total_area	numeric	total area affected
cause_type	character	cause of the fire

A classification for causes types are presented in table 3.2.

Table 3.2: Classifications of causes of forest fires.

Cause	Description
Unknown	absence of suficient objective evidence to determine the cause of the ignition of fire
Natural	lightning generated in thunderstorms
Negligence	the misguided use of fire in activities such as burning trash, mass burning of agricultural and forest fuels, fun and leisure activities; failure to properly extinguish cigarettes by smokers; the dispersal and transport of incandescent particles from chimneys; etc.
Intentional	incendiarism and arson, mostly resulting from behaviors and attitudes reacting to theconstraints of agroforestry management systems and to conflicts related to land use
Rekindling	reburning of an area over which a fire has previously passed, but where fuel has been left that is later ignited by latent heat, sparks, or embers

A glimpse of the structure of the `Forest Fires` data set is provided below:

Table: A glimpse of the structure of the data set.

```
## Rows: 7,511
## Columns: 21
## $ id                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...
## $ region            <chr> "Entre Douro e Minho", "Entre Douro e Minho", "T...
## $ district          <chr> "Viana do Castelo", "Porto", "Vila Real", "Vila ...
## $ municipality      <chr> "Ponte de Lima", "Marco de Canaveses", "Boticas"...
## $ parish            <chr> "Serdedelo", "Vila Boa de Quires", "Cerdedo", "G...
## $ lat               <chr> "41:44:48.5663999999878''", "41:12:58.4280000000...
## $ lon               <chr> "8:31:12.3276000000027''", "8:12:28.378800000002...
## $ origin            <chr> "fire", "fire", "fire", "firepit", "firepit", "f...
## $ alert_date        <chr> "2015-03-24", "2015-03-24", "2015-03-24", "2015-...
## $ alert_hour        <chr> "17:01:00", "17:10:00", "21:40:00", "16:00:00", ...
## $ extinction_date   <chr> "2015-03-24", "2015-03-24", "2015-03-25", "2015-...
## $ extinction_hour   <chr> "18:09:00", "18:47:00", "05:45:00", "17:00:00", ...
## $ firstInterv_date  <chr> "2015-03-24", "2015-03-24", "2015-03-24", "2015-...
```

```
## $ firstInterv_hour   <chr> "17:10:00", "17:16:00", "22:00:00", "16:14:00", ...
## $ alert_source       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ village_area       <dbl> 2.50, 0.00, 0.50, 0.00, 0.10, 0.00, 0.35, 0.50, ...
## $ vegetation_area    <dbl> 0.000, 1.350, 38.000, 0.010, 0.000, 0.100, 14.82...
## $ farming_area       <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, ...
## $ village_veget_area <dbl> 2.500, 1.350, 38.500, 0.010, 0.100, 0.100, 15.17...
## $ total_area         <dbl> 2.5000, 1.3500, 38.5000, 0.0100, 0.1000, 0.1000,...
## $ cause_type         <chr> "negligent", "negligent", "negligent", "negligen...
```

A summary for each variable present in dataset is provided below. The metrics displayed are: quantity and percentage of zeros, quantity and quantity and percentage of NA's, data type and quantity of unique values.

Table 4: A summary of variables of the dataset.

```
##              variable q_zeros p_zeros q_na   p_na q_inf p_inf      type unique
## 1                  id       0    0.00    0   0.00     0     0   integer   7511
## 2              region       0    0.00  501   6.67     0     0 character     10
## 3            district       0    0.00    0   0.00     0     0 character     19
## 4        municipality       0    0.00    0   0.00     0     0 character    297
## 5               parish       0    0.00    0   0.00     0     0 character   2270
## 6                  lat       0    0.00    0   0.00     0     0 character   5858
## 7                  lon       0    0.00    0   0.00     0     0 character   5867
## 8               origin       0    0.00    0   0.00     0     0 character      5
## 9           alert_date       0    0.00    0   0.00     0     0 character    317
## 10          alert_hour       0    0.00    0   0.00     0     0 character   1312
## 11    extinction_date       0    0.00    9   0.12     0     0 character    319
## 12    extinction_hour       0    0.00    9   0.12     0     0 character   1201
## 13   firstInterv_date       0    0.00  214   2.85     0     0 character    318
## 14   firstInterv_hour       0    0.00  215   2.86     0     0 character   1202
## 15       alert_source       0    0.00 7511 100.00     0     0   logical      0
## 16       village_area    5349   71.22    0   0.00     0     0   numeric    591
## 17    vegetation_area    2648   35.25    0   0.00     0     0   numeric   1052
## 18       farming_area    5976   79.56    0   0.00     0     0   numeric    650
## 19 village_veget_area    1413   18.81    0   0.00     0     0   numeric   1377
## 20         total_area       8    0.11    0   0.00     0     0   numeric   1781
## 21         cause_type       0    0.00    0   0.00     0     0 character      4
```

A sample the first observations is provided below:

```
## # A tibble: 6 x 21
##      id region district municipality parish lat   lon   origin alert_date
##   <int> <chr>  <chr>    <chr>        <chr>  <chr> <chr> <chr>  <chr>
## 1     1 Entre~ Viana d~ Ponte de Li~ Serde~ 41:4~ 8:31~ fire   2015-03-24
```

```
## 2     2 Entre~ Porto    Marco de Ca~ Vila ~ 41:1~ 8:12~ fire   2015-03-24
## 3     3 Trás-~ Vila Re~ Boticas      Cerde~ 41:3~ 07:5~ fire   2015-03-24
## 4     4 Trás-~ Vila Re~ Montalegre   Gralh~ 41:5~ 7:42~ firep~ 2015-03-25
## 5     5 Trás-~ Vila Re~ Valpaços     Alger~ 41:3~ 07:2~ firep~ 2015-03-12
## 6     6 Entre~ Vila Re~ Mondim de B~ Ermelo 41:2~ 07:5~ firep~ 2015-03-13
## # ... with 12 more variables: alert_hour <chr>, extinction_date <chr>,
## #   extinction_hour <chr>, firstInterv_date <chr>, firstInterv_hour <chr>,
## #   alert_source <lgl>, village_area <dbl>, vegetation_area <dbl>,
## #   farming_area <dbl>, village_veget_area <dbl>, total_area <dbl>,
## #   cause_type <chr>
```

Chapter 4

Data Preparation

Data preparation consists of the process of cleaning and transforming raw data in a form that can be used by machine learning algorithms. Next sections, we exploit the `Forest Fires` dataset in order to perform the steps of cleaning a transforming, when need.

4.1 Data Cleaning

4.1.1 Latitude and Longitude

The `Forest Fires` dataset store the latitude and longitude of the place where occurred the fire into variables `lat` e `lon` respectively. These values are in format of *Degrees^{o}Minutes'Seconds"* and for the reason contain special characters `º`, `'`, `:` and `"`. Besides, there are wrong values into variables as dates between the coordinates and values with scientific notation `E-12`, `E-11` and `E-02`. A sample of these inconsistencies is provided in the tables below:

```
## # A tibble: 1 x 2
##   lat                     lon
##   <chr>                   <chr>
## 1 41º41'25.821599999997'' 8º20'37.446000000002''
```

```
## # A tibble: 1 x 2
##   lat                 lon
##   <chr>               <chr>
## 1 1900-01-01 14:19:38 07:30:27
```

```
## # A tibble: 1 x 2
##   lat                         lon
##   <chr>                       <chr>
## 1 38:36:5.11590769747272E-12 8:35:49.9999999999972
```

A cleaning and transformation steps were perfomed on `lat` and `lon` variables to remove the special caracters and scientific notation. For the values wrongs where there is a date among the coordinates, it was performed an data imputation based on another observations that has the same `region`, `district`, `municipality` and `parish`. After the cleaning steps, the values were transformated from GPS coordinates to decimals coordinates in order to be able retrieve historical data from nearest weather stations using the RNOAA package

The data imputation and transformation generated 8 NA's in `lat` and `lon` variables for parishes listed below:

```
## # A tibble: 8 x 6
##   region   district municipality    parish                    lat   lon
##   <chr>    <chr>    <chr>           <chr>                     <chr> <chr>
## 1 Alentejo Évora    Mora            Cabeção                   <NA>  <NA>
## 2 Alentejo Évora    Montemor-o-Novo Cortiçadas de Lavre       <NA>  <NA>
## 3 Alentejo Évora    Montemor-o-Novo Ciborro                   <NA>  <NA>
## 4 Alentejo Évora    Mourão          Granja                    <NA>  <NA>
## 5 Alentejo Évora    Évora           Horta das Figueiras       <NA>  <NA>
## 6 Alentejo Évora    Montemor-o-Novo Cortiçadas de Lavre       <NA>  <NA>
## 7 Alentejo Évora    Estremoz        São Lourenço de Mamporcão <NA>  <NA>
## 8 Alentejo Évora    Mora            Brotas                    <NA>  <NA>
```

In order to fixing this, the latitude and longitude values for these parishes were imputed directly from the localization data retrieved from the internet.

4.1.2 District

Mainland Portugalis is divided into 18 districts and the variable `district` from `Forest Fires` dataset refer the place where occurred the fires. As seen in table @ref(tab:summary_data), this variable has 19 unique values, so there are some inconsistent data. The table below display the unique values for this variable:

```
##  [1] "Viana do Castelo" "Porto"           "Vila Real"       "Bragança"
##  [5] "Braga"            "Portalegre"      "Santarém"        "Viseu"
##  [9] "Guarda"           "Leiria"          "Castelo Branco"  "Aveiro"
## [13] "Évora"            "Faro"            "Coimbra"         "Viana Do Castelo"
## [17] "Lisboa"           "Beja"            "Setúbal"
```

As seen in table above, there are two references for the same district: *Viana do Castelo* and *Viana Do Castelo*. So a step of cleaning was performed into this variable values.

4.1.3 First Intervention and Extinction

The variables `firstInterv_date` and `firstInterv_hour` store the date and time that occured the the first intervention by autorities after the fire alert. As seem in table 4, these variables have a total of NA's values equals 214 and 215, respectively. In order to reduce these quantity, a data imputation were performed based on values of `extinction_date` and `extinction_hour` assumption that if there are values for extinction date and time it because some intervention was realized. After data imputatio the quantity of NA's was reduced to 7 in both variables as can be seen below:

```
##           variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
## 1 firstInterv_date       0       0    7 0.09     0     0 character    318
## 2 firstInterv_hour       0       0    7 0.09     0     0 character   1209
## 3  extinction_date       0       0    9 0.12     0     0 character    319
## 4  extinction_hour       0       0    9 0.12     0     0 character   1201
```

The remaining quantity of NA's values in `firstInterv_date`, `firstInterv_hour`, `extinction_date`, and `extinction_hour` represent 0.9% and 1.2% respectively of the total of observations. As these values are relatively low, these observations were removed from dataset.

Alert Source

As can be seen in table below, the variable `alert_source` has 100% of values with NA's, so this variable were removed from dataset.

```
##   variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1      var       0       0 7502  100     0     0 logical      0
```

4.2 Data Transformation

In the `Forest Fires` dataset, the variables `region`, `district`, `municipality`, `parish`, `origin`, `cause_type` are with the data type as `<character>` when the appropriataded should be a `<factor>`. Thus, a step to transform these variables into the most appropriated data type was performed.

Besides the variables aforementioned, others also were transformed. The variables `alert_date`, `alert_hour`, `firstInterv_date`, `firstInterv_hour`, `extinction_date`, and `extinction_hour` contains date and time data, and thus, these variables were appropriated merged and transformed in POSIXct object in order to able to handle it as a date. The variables `alert`, `firtsInterv` and `extinction` were created based on corresponding variables, and the ones were removed from the dataset.

4.3 Feature Engineering

Weather Data

As additional information, let's use weather data in order to build features that can help the models obtain better accuracy. The weather data can be obtained from the National Oceanic and Atmospheric Administration (NOAA) Climate Data Sources using RNOAA package whicu provides free access to National Climatic Data Center's (NCDC) archive of global historical weather and climate data in addition to station history information. These data include quality controlled daily, monthly, seasonal, and yearly measurements of temperature, precipitation, wind, and degree days as well as radar data and 30-year Climate Normals.

For the propose this analysis, it was created 5 new features: `tavg`, `tavg15d`, `tmax`, `tmin`, and `prcp` based on the following weather measurements:

Measurement	Description
tavg	the average temperature of day (degrees C)
tavg15d	the average temperature of the last 15 days (degrees C)
tmax	the maximum temperature of day (degrees C)
tmin	the minimum temperature od day (degrees C)
prcp	the precipitation (mm)

In order to retrieve the weather data it was considered the alert date of fire. The sample of the first rows of created features it provided in table below:

```
## # A tibble: 6 x 6
##   alert                tavg tavg15d  tmax  tmin  prcp
##   <dttm>              <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1 2015-03-24 17:01:00  12.3   11.1   19.1   5.7  11.1
## 2 2015-03-24 17:10:00  15.1   13.3   21.6   5.7   3
## 3 2015-03-24 21:40:00  12.3   11.1   19.1   5.7   3
## 4 2015-03-25 16:00:00  13.7   11.5   16.2  11.5   0.4
## 5 2015-03-12 17:20:00  11.1    9.85  16.7   4.9   0
## 6 2015-03-13 14:48:00  12.7   11.4   16.2   9.6   0
```

After load weather data into `Forest Fire` dataset, some steps of data imputation were performed to reduce the quantity od NA's and zero values. The reasoning for imputation was calculate the value based on other variables. For instance, `tavg` variable with NA or zero value, check if there is values for `tmax` and `tmin` and calculate the average temperature. For `tmax` variable with NA or zero value, check if there is values for `tavg` and `tmin` and calculate the maximum temperature, and so forth. Below it provided the summary of the new features after the data imputation:

```
##   variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
```

```
## 1     tavg       8    0.11    0    0     0     0 numeric    335
## 2  tavg15d      56    0.75    0    0     0     0 numeric   1549
## 3     tmax     133    1.77    0    0     0     0 numeric    325
## 4     tmin     133    1.77    0    0     0     0 numeric    323
## 5     prcp    4694   62.57    0    0     0     0 numeric    138
```

Based on Features

The variable `alert` provides relevant information for the goal of this analysis. The quantity of unique values for this variable is 7313. This quantity can represent noise for the predictive model. For this reason, it was created two new features based on it: `alert_month` and `alert_period`. The first one is a numeric variable with the month of a fire alert. The latter one is an ordinal feature with 4 values *Morning*, *Afternoon*, *Night*, *Dawn* with a slot of 6 hours each one. After creation of these new features, the `alert` variable was removed from dataset.

Another new feature called `duration` was created based on `alert` and `extinction` variables. Its value is the elapsed time between fire alert and its extinction. We consider that information can be useful help to predict the fire cause type as rekindling.

After this process of feature engineering the `Forest Fires` dataset now he has 24 variables as against 21 from the original dataset. A summary for the new dataset is shown the table below:

```
##              variable q_zeros p_zeros q_na p_na q_inf p_inf           type
## 1                  id       0    0.00    0 0.00     0     0        integer
## 2              region       0    0.00  501 6.68     0     0         factor
## 3            district       0    0.00    0 0.00     0     0         factor
## 4        municipality       0    0.00    0 0.00     0     0         factor
## 5               parish       0    0.00    0 0.00     0     0         factor
## 6                  lat       0    0.00    0 0.00     0     0      character
## 7                  lon       0    0.00    0 0.00     0     0      character
## 8               origin       0    0.00    0 0.00     0     0         factor
## 9         firstInterv       0    0.00    0 0.00     0     0 POSIXct/POSIXt
## 10          extinction       0    0.00    0 0.00     0     0 POSIXct/POSIXt
## 11       village_area    5342   71.21    0 0.00     0     0        numeric
## 12    vegetation_area    2646   35.27    0 0.00     0     0        numeric
## 13       farming_area    5968   79.55    0 0.00     0     0        numeric
## 14 village_veget_area    1412   18.82    0 0.00     0     0        numeric
## 15         total_area       8    0.11    0 0.00     0     0        numeric
## 16               tavg       8    0.11    0 0.00     0     0        numeric
## 17            tavg15d      56    0.75    0 0.00     0     0        numeric
## 18               tmax     133    1.77    0 0.00     0     0        numeric
## 19               tmin     133    1.77    0 0.00     0     0        numeric
## 20               prcp    4694   62.57    0 0.00     0     0        numeric
## 21         cause_type       0    0.00    0 0.00     0     0         factor
## 22        alert_month       0    0.00    0 0.00     0     0         factor
```

```
## 23       alert_period       0    0.00     0 0.00      0      0         factor
## 24           duration       4    0.05     0 0.00      0      0        numeric
##    unique
## 1    7502
## 2      10
## 3      19
## 4     297
## 5    2270
## 6    5805
## 7    6213
## 8       5
## 9    7185
## 10   7120
## 11    590
## 12   1051
## 13    650
## 14   1375
## 15   1779
## 16    335
## 17   1549
## 18    325
## 19    323
## 20    138
## 21      4
## 22     12
## 23      4
## 24    598
```

4.4 Data Analysis

Understanding the structure of the data, the distribution of the variables, and the relationships between them is fundamental to build a solid model.

Based on the dataset we ploted some graphics that helped us to get some conclusions and showed a general notion about the problem of the forests fires.

Figure 4.1 depicts the bar graphic of the distribution of forests fires during 2015. The x-axis represents the months along the year 0f 2015 and the y-axis represents the total of fires that occurred by month.

This graphic showed us that july and august are the months with the largest ocurrences and the period between march and september needs more atention. Probably we will consider the variable month as important to the analisys.

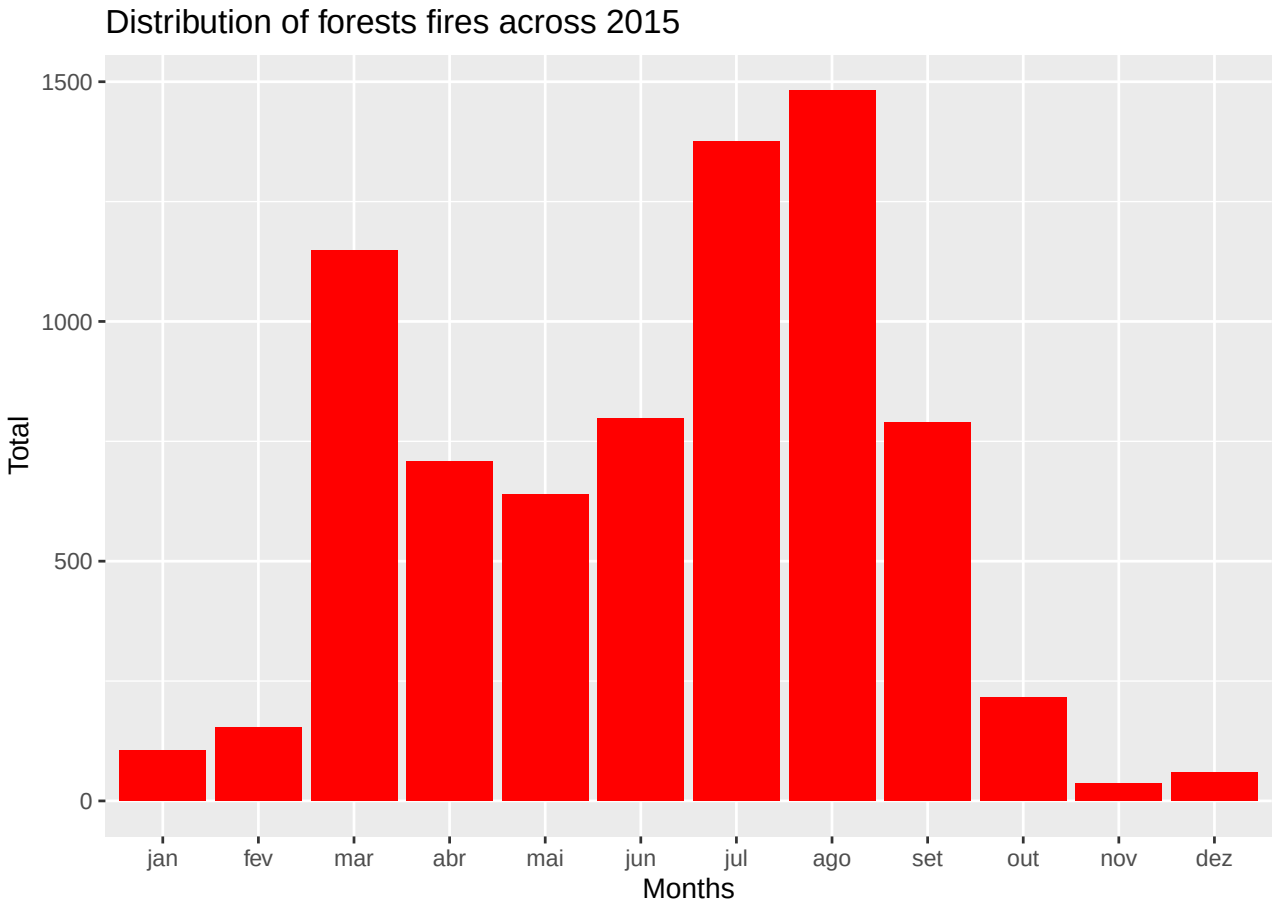


Figure 4.1: Barplot of the distribution of forests fires during 2015.

4.4.1 Region and District

These two variables represents the areas of the occurrences and we can observe by summary that the variable region has a lot of NA´s (501) that corresponds to 6,67% of the total of lines. The distribution of the occurrences by region can be observed in the grafic 4.2. The y-axis represents the regions and the x-axis represents the total of fires that occurred by region.

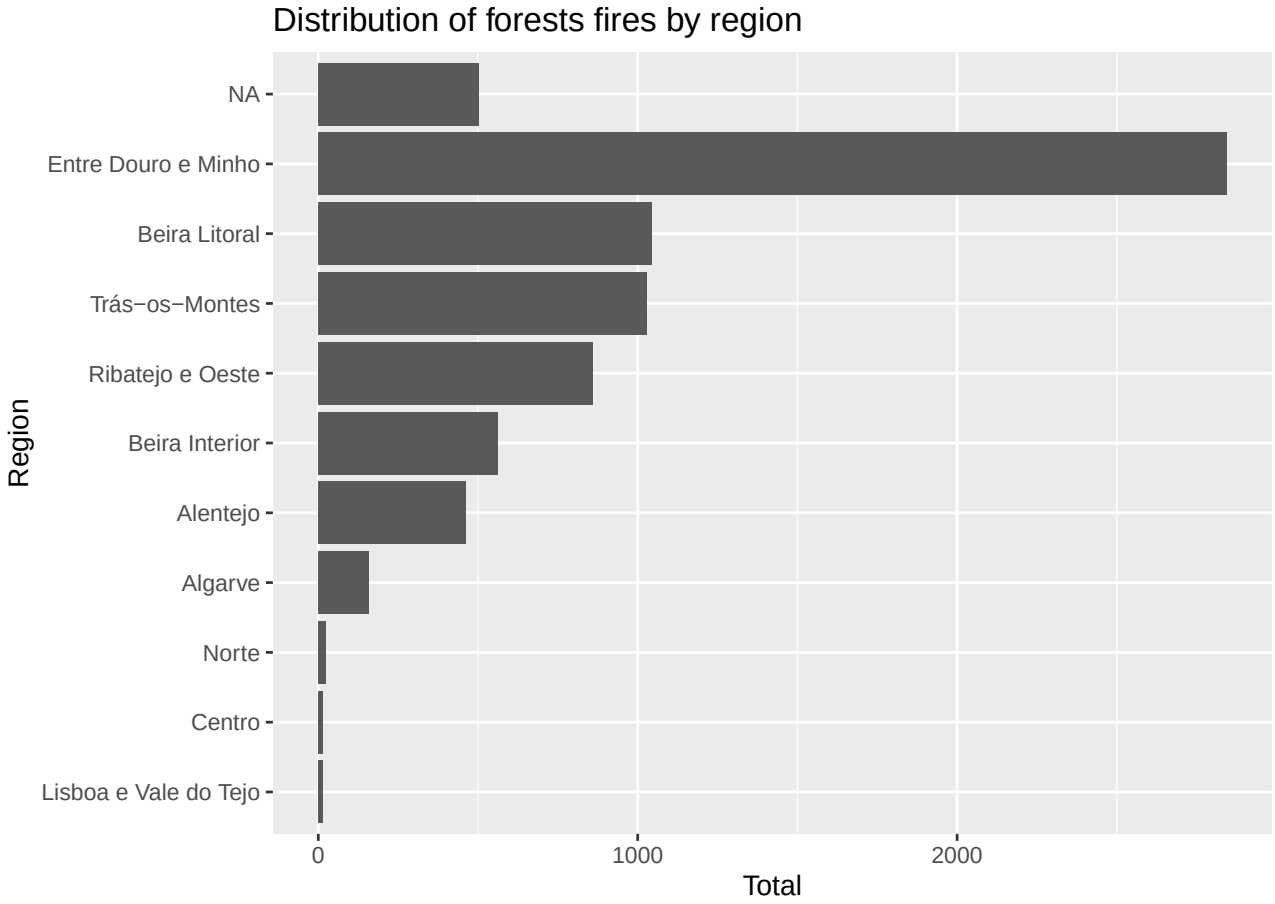


Figure 4.2: Barplot of the distribution of forests fires by regions.

Observing this grafic we saw that Entre Douro e Minho was the region with more forests fires and other regions like Centro, Lisboa and Norte were with minimum occurrences.

The relationship between region, month and causes is represented on figure 4.3. The x-axis includes the diferent regions, y-axis represents the months of ocurrences and the variable cause is showed by colours listed on the labels.

Another important graphic is the figure 4.4 that corresponds to the distribution of forests fires by district. The y-axis represents the districts and the x-axis represents the total of fires that occurred by region.

This graphic indicates that Porto and Viana do Castelo were the districts with more forests fires.

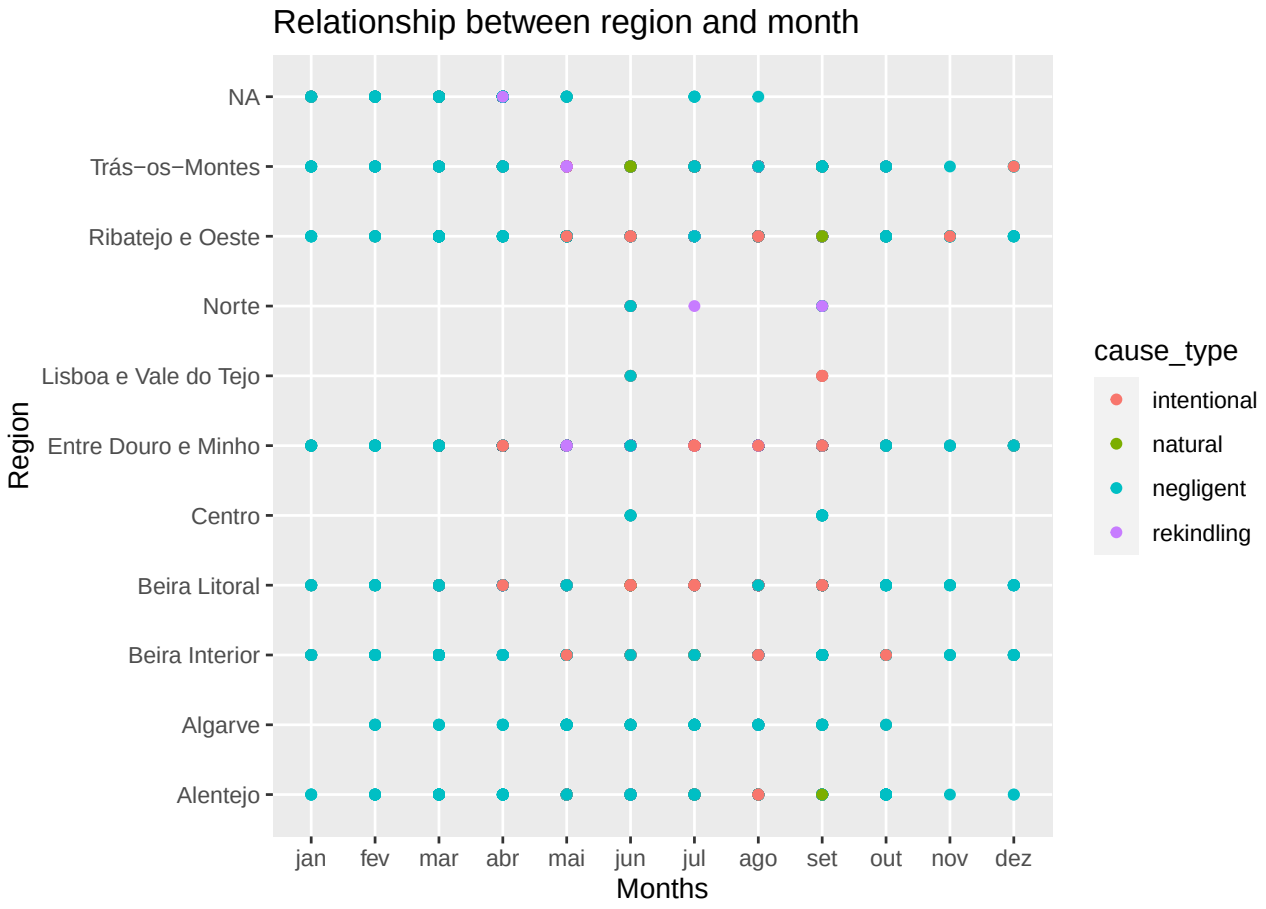


Figure 4.3: Distribution of forests fires relating region, month and causes.

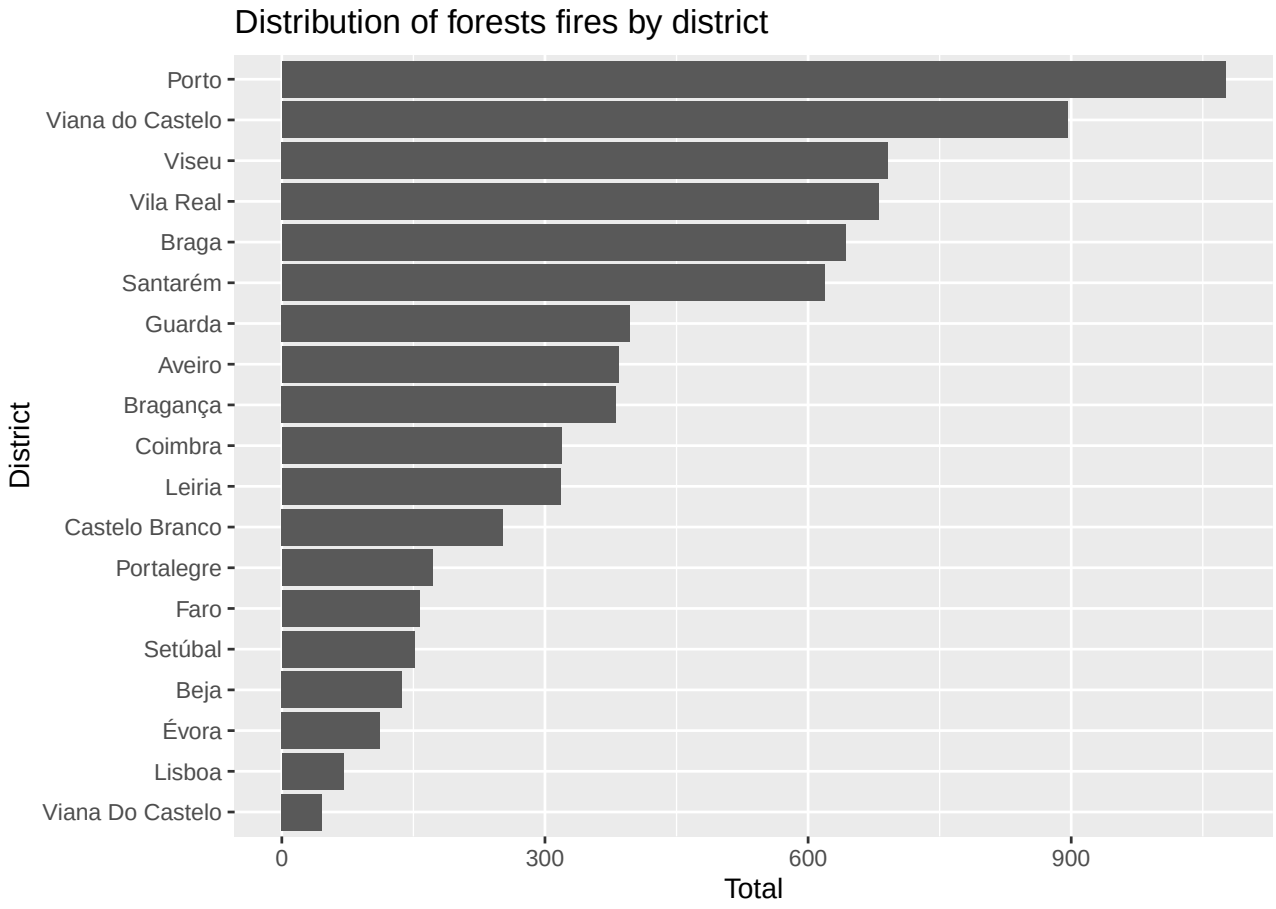


Figure 4.4: Barplot of the distribution of forests fires by districts.

The relationship between district, month and causes is represented on figure 4.5. The x-axis includes the diferent districts, y-axis represents the months of ocurrences and the variable cause is showed by colours listed on the labels.

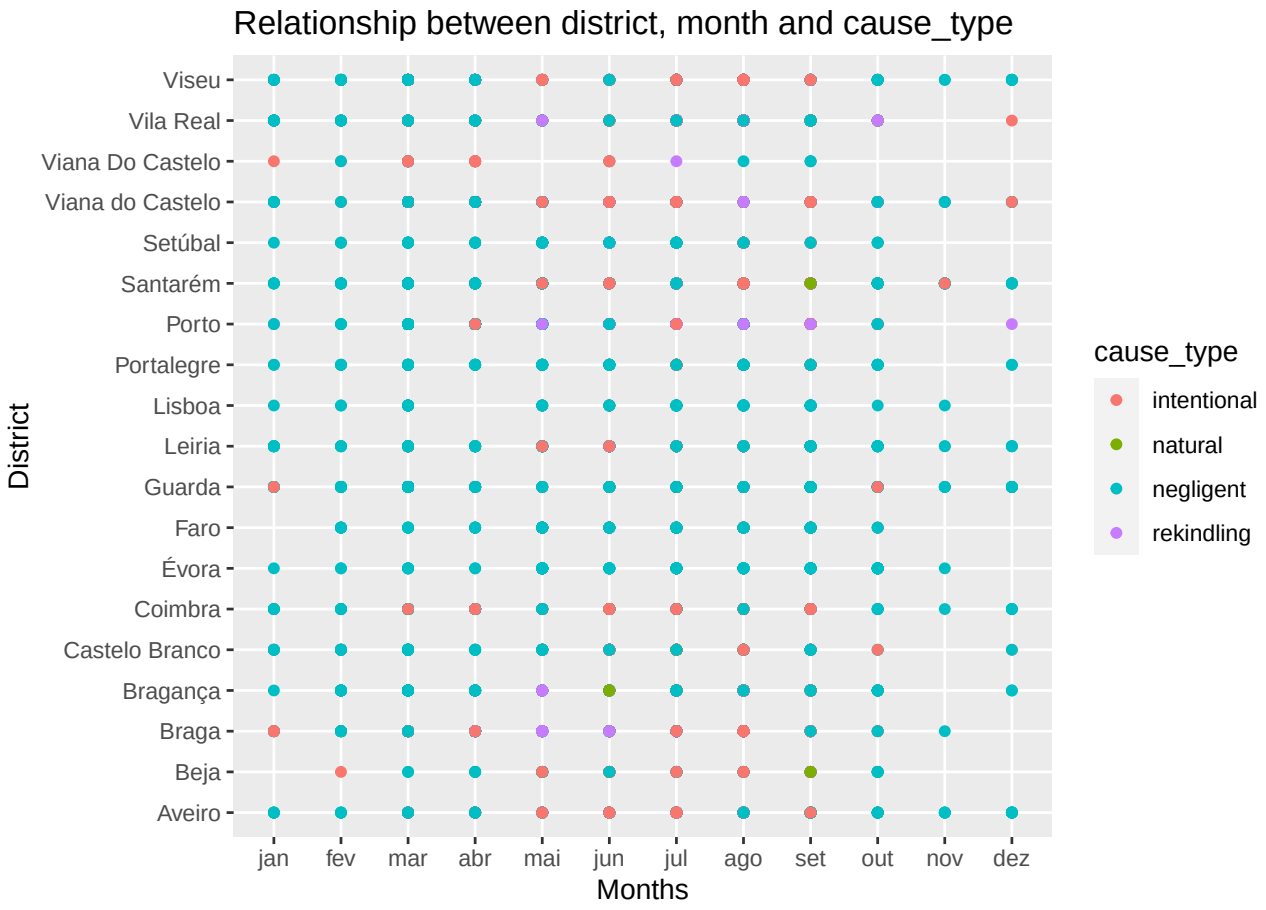


Figure 4.5: Distribution of forests fires relating district, month and causes.

4.4.2 Origin

Origin informs the reason that the fire started and apparently appears to be an important observation for evaluation.It can be observed in figure 4.6.

On the x-axis are listed the diferent origins and on y-axis represents the total of fires that occurred.

The firepit was the origin of the most forests fires comparing it with the other origins.

4.4.3 Cause type

This is the variable to be predicted by the model that will be chosen.It shows the four causes of the occurrences: intentional, natural, negligent and rekindling. They can be observed in the

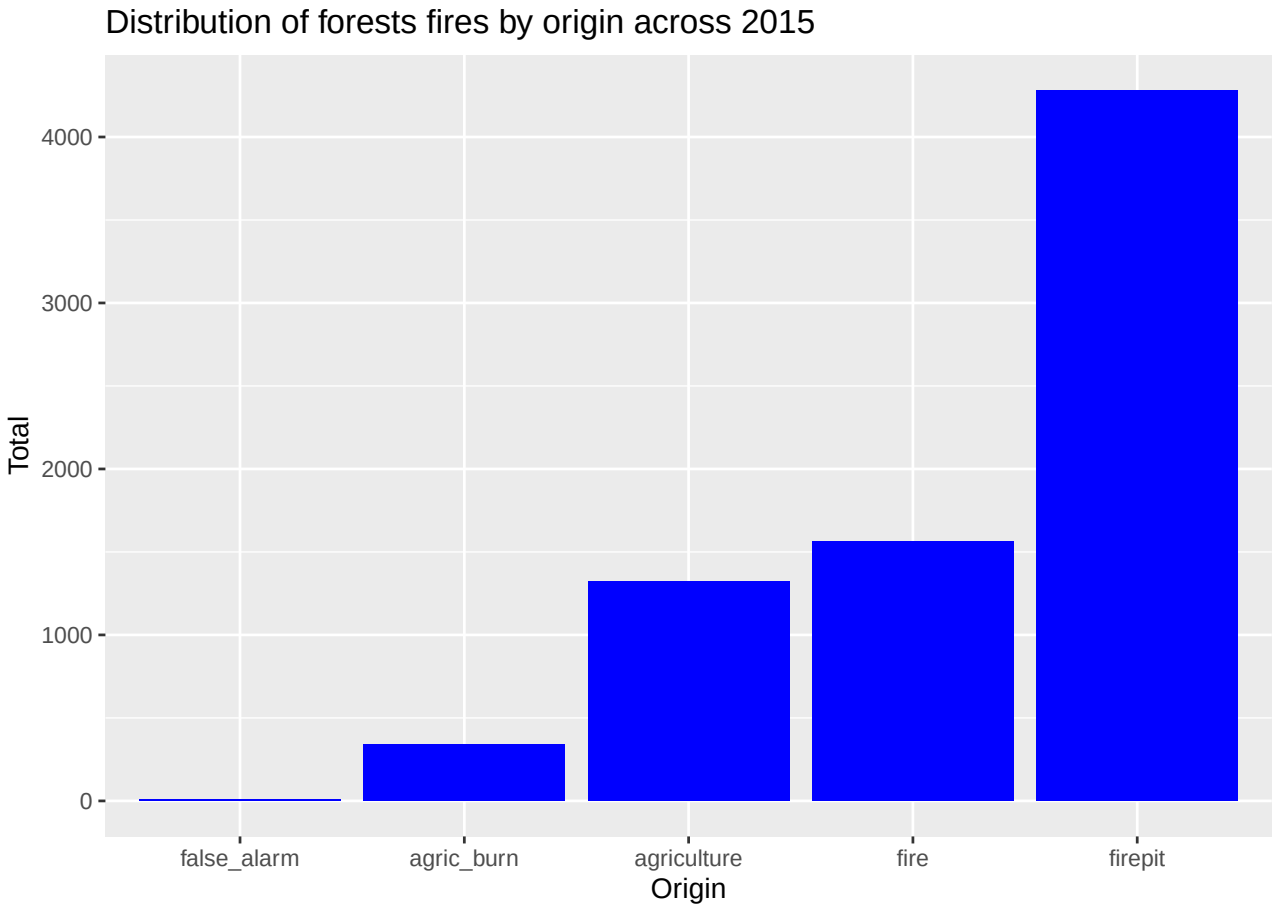


Figure 4.6: Barplot of the distribution of forests fires by origins.

graphic 4.7 below. On the x-axis are listed the diferent causes and on y-axis represents the total of fires that occurred.

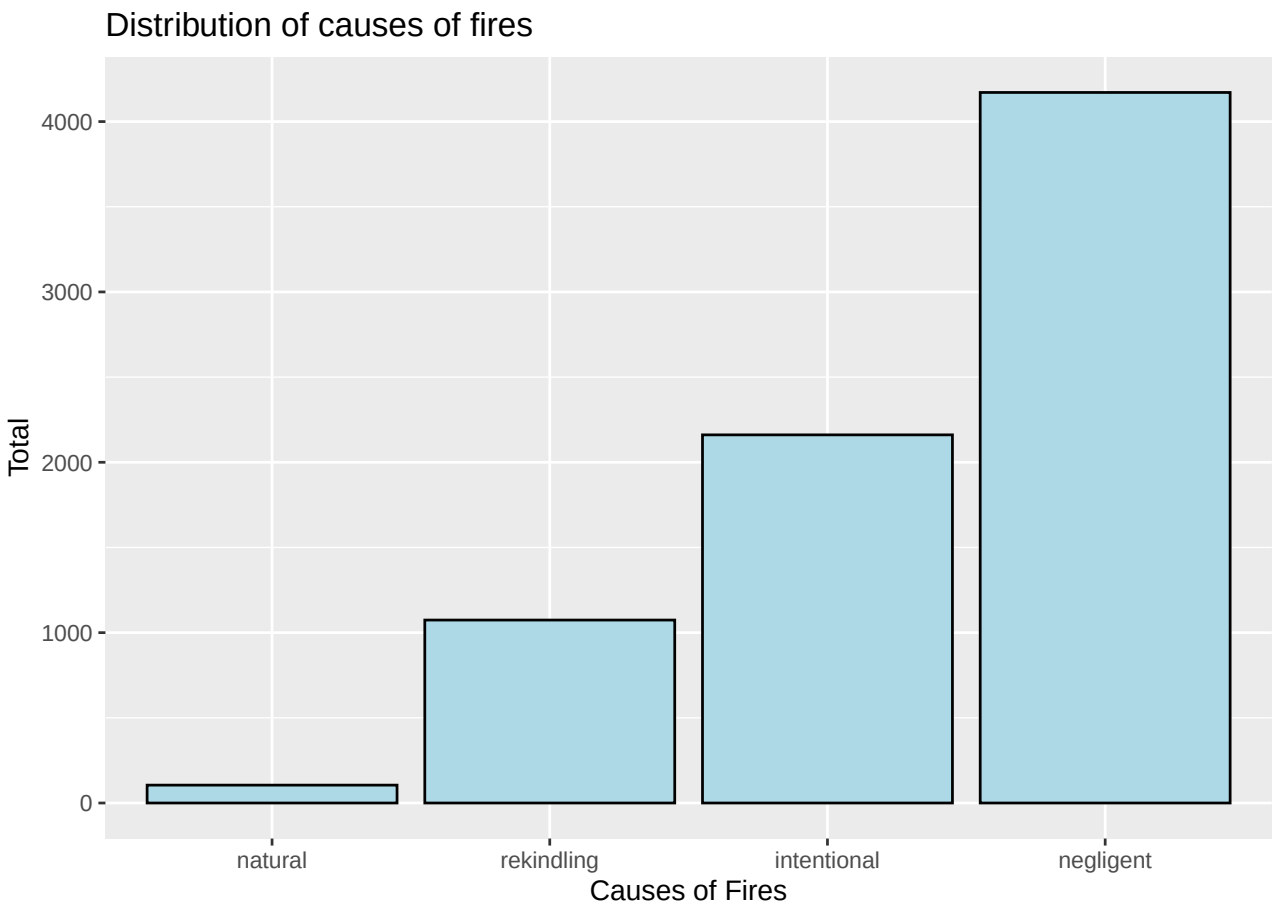


Figure 4.7: Barplot of the distribution of forests fires by causes.

A thing that calls our attention is the difference between the number of fires that were caused by negligence and by natural causes. The number of forest fires caused intentionally were almost the half of the negligent causes what is an alarmant number.

4.4.4 TAVG

This variable is the average temperature of each day of the occurrences. It was included from the station datas and your variation is described in the graphic 4.8.

The graphic 4.8 shows us that the large number of occurrences happened when the average temperature was between 24º e 26º.

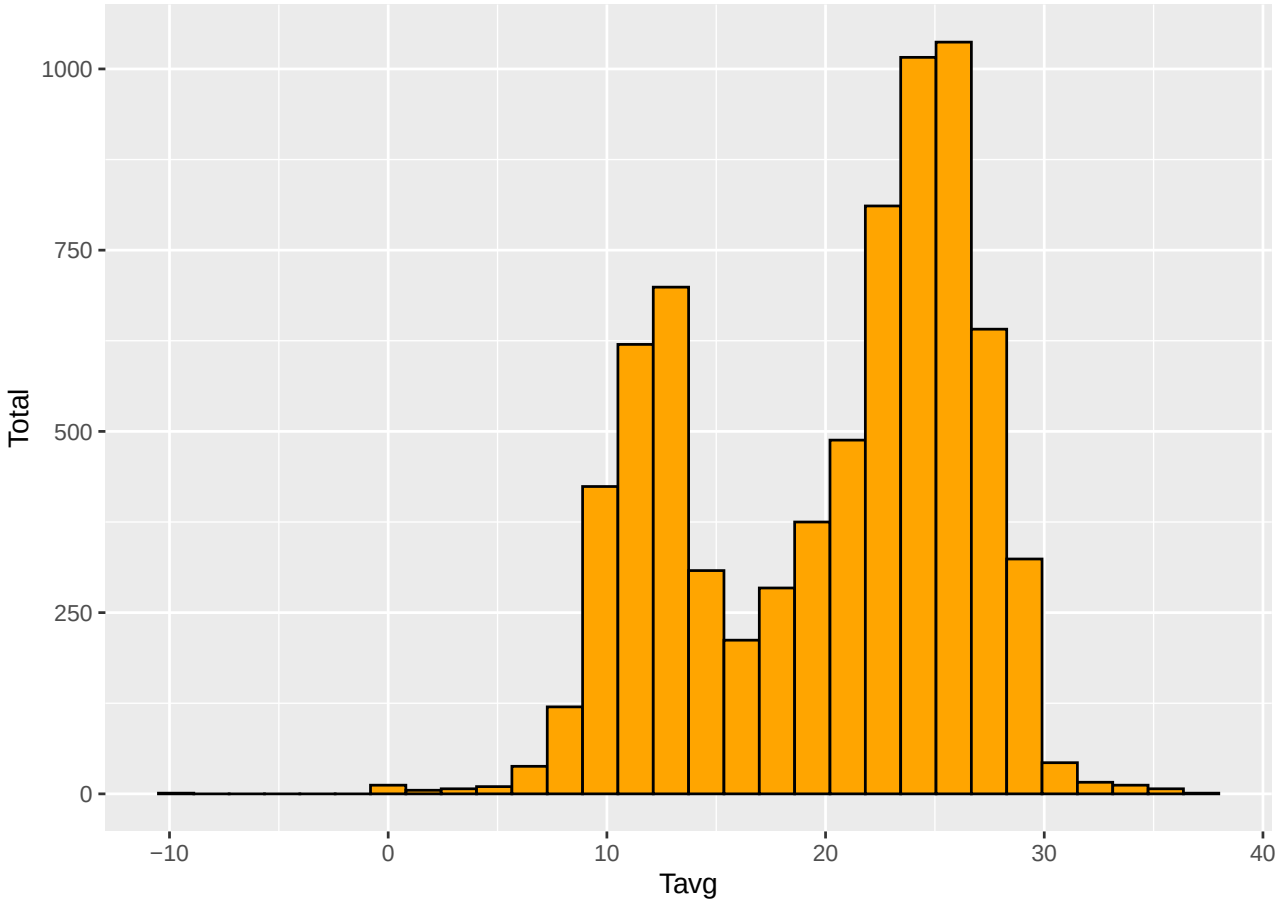


Figure 4.8: Histogram of Average Temperature by occurrences

4.4.5 PRCP

This variable is the maximum precipitation volume of rain of each day of the occurrences. It was included from the station datas and your variation is described in the graphic 4.9.

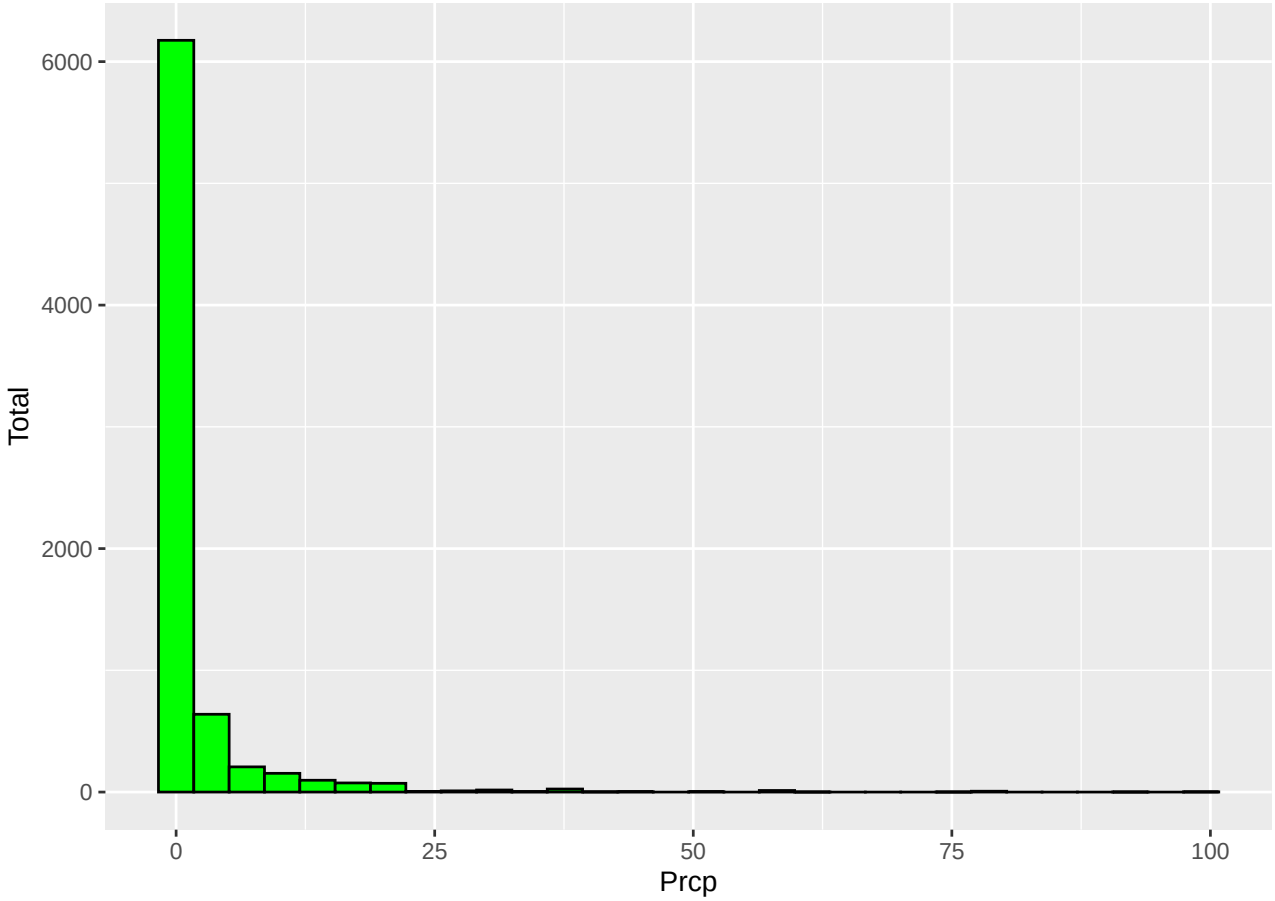


Figure 4.9: Histogram of PRCP by occurrences

The graphic 4.9 shows us that the large number of occurrences happened when the prcp was near to zero.

4.5 Dimensionality Reduction

Dimensionality reduction consists of the process of reducing the number of features in a dataset in order to eliminate the noise and features redundancy and create a more compact projection of the data. After previous preprocessing steps performed, the `Forest Fires` dataset has now 24 variables and some of them are redundant or irrelevant to the context of this problem.

The variables `region`, `district`, `municipality` and `parish` has a high correlation once that one is a subdivision of another. A set of parishes is part of a municipality. In turn, a set of municipalities is part of a district. Thus, as the variable `region` has a lot of number of NA, and the

variables `municipality` and `parish` are subdivisions of districts, they are irrelevant/redundant for the context and they were removed from the dataset.

The variables `village_veget_area` and `total_area` are redundants because contains a somatory of values of the variables `village_area`, `vegetation_area` and `farming_area`. Thus, they were removed from the dataset.

The variables `lat` and `lon` were important to retrieve the weather data from the nearest station. Once these data were inserted into the dataset, these variables became irrelevant to the context and they were removed from the dataset.

The variables `tmax` and `tmin` were important in order to performe data imputation into the `tavg` variable. After that they became irrelevant to the context and they were removed from the dataset.

The variables `id`, `firstInterv`, and `extinction` were considered irrelevants to the context once their values dont't provide useful information that can help to predict cause of forest fire, thus, these variables were removed from the dataset.

Thus, the dimensionality of the `Forest Fires` was reduced and has now 12 features considered relevants to the context. These variables are displyed below:

```
##  [1] "district"       "origin"          "alert_month"     "alert_period"
##  [5] "duration"       "village_area"    "vegetation_area" "farming_area"
##  [9] "tavg"           "tavg15d"         "prcp"            "cause_type"
```

4.6 Feature Selection

Feature selection refers to techniques that select a subset of the most relevant features for a dataset.

Once made the dimensionality reduction, it needs to check which the previously selected variables are really relevant. For this, it was used the Recursive Feature Elimination (RFE), the most widely used wrapper-type feature selection algorithm.

RFE is popular because it is easy to configure and use and because it is effective at selecting those features in a training set that are more or most relevant in predicting the target variable.

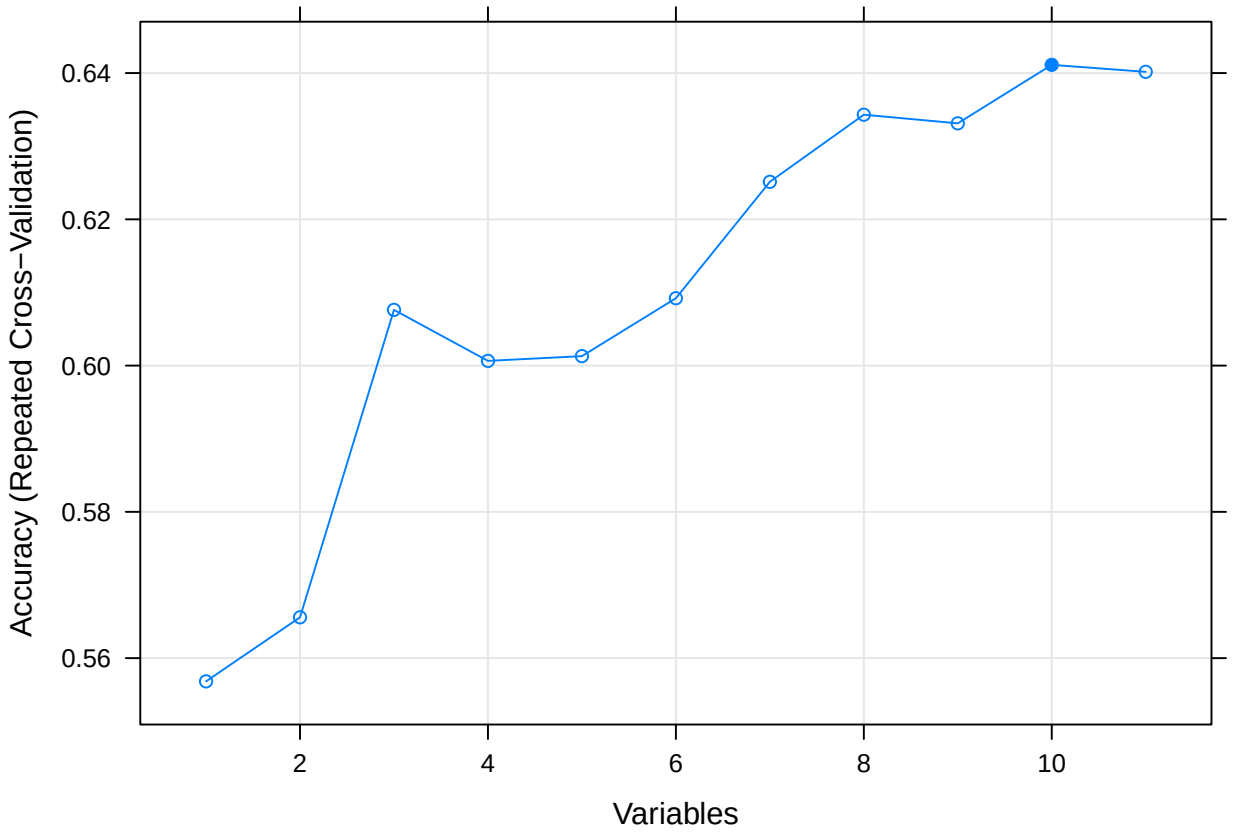
There are two important configuration options when using RFE: the choice in the number of features to select and the choice of the algorithm used to help choose features.

- `size`: a integer vector for the specific subset sizes that should be tested (which need not to include ncol(x))
- `rfeControl`: a list of options that can be used to specify the model and the methods for prediction, ranking etc.

In order to performe the feature selection, it was used the RFE algorithm implemented by caret package. The hyperparameters values used was`size=c(1:11)` that correspond all the predictors features and `rfeControl(functions=rfFuncs, method="repeatedcv", repeats=5)` that mean the Random Forest method was use with 5 rounds of 10-Fold Cross-Validation.

The result of execution of RFE algorithm it provided below:

```
##
## Recursive feature selection
##
## Outer resampling method: Cross-Validated (10 fold, repeated 5 times)
##
## Resampling performance over subset size:
##
##  Variables Accuracy   Kappa AccuracySD KappaSD Selected
##          1   0.5568 0.05332   0.008738 0.02409
##          2   0.5656 0.12550   0.015684 0.03324
##          3   0.6076 0.29187   0.017560 0.03127
##          4   0.6006 0.31349   0.019507 0.03202
##          5   0.6013 0.30798   0.025636 0.04427
##          6   0.6092 0.31493   0.018765 0.03178
##          7   0.6251 0.33683   0.016502 0.02639
##          8   0.6343 0.34991   0.016802 0.03035
##          9   0.6331 0.35030   0.017825 0.03092
##         10   0.6411 0.35921   0.016906 0.03076        *
##         11   0.6402 0.35563   0.015462 0.02840
##
## The top 5 variables (out of 10):
##    district, origin, alert_month, tavg15d, alert_period
```



The variables selected were:

```
##  [1] "district"        "origin"          "alert_month"     "tavg15d"
##  [5] "alert_period"    "tavg"            "village_area"    "farming_area"
##  [9] "vegetation_area" "duration"
```

Chapter 5

Prediction Models

Predictive modeling is a technique that uses mathematical and computational methods to predict an event or result. The model is used to predict a result in some future state or moment based on changes in model inputs. The model parameters help to explain how the model inputs influence the result.

5.1 `Caret` Package

The R `caret` allows us to create several powerful predictive models using a simplified and consistent modeling syntax. There are more than 200 different models available in `caret` which allows us with very little change to set up the resampling approach and the parameter tuning. Behind the scenes, `caret` automatically resamples the models and conducts parameter tuning. This way it is possible to build and compare models with very little overhead.

The `train()` function from `caret` was used to modeling all the predictive models needs for this studied. Besides one, the function `trainControl()` which allows to set up several aspects of the model, like the resampling method, it was defined withe `method=cv` and `number=10` what means that was used the 10-fold cross-validation for all the predictive models created.

```
ctrl <- trainControl(method = "cv", number = 10, savePredictions = "final", classProbs =
```

5.1.1 Data Splitting

As a result of the data preparation and preprocessing, the forest fires dataset final has now 7502 observations. In order to build the predictive models, it was split into training and test sets where the first one containing 70% of the data, or precisely 5253 observations, and the second one containing 30% of the data corresponding to 2249 observations. The training of models was performed using the training set and the test set was used to assess the performance of them.

5.2 Distance-based Approach

Distance-based algorithms are machine learning algorithms that classify queries by computing distances between these queries and a number of internally stored exemplars. Exemplars that are closest to the query have the largest in uence on the classification assigned to the query. Two speci c distance-based algorithms, the nearest neighbor algorithm and the nearest-hyperrectangle algorithm, are studied in detail. It is shown that the k-nearest neighbor algorithm (kNN)

(Citeable URL: https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/zw12z7835)

5.2.1 K-Nearest Neighbor

Distance-based algorithms are machine learning algorithms that classify queries by computing distances between these queries and a number of internally stored exemplars. Exemplars that are closest to the query have the largest influence on the classification assigned to the query. The distance-based algorithm used in this study was the k-nearest neighbor algorithm (kNN).

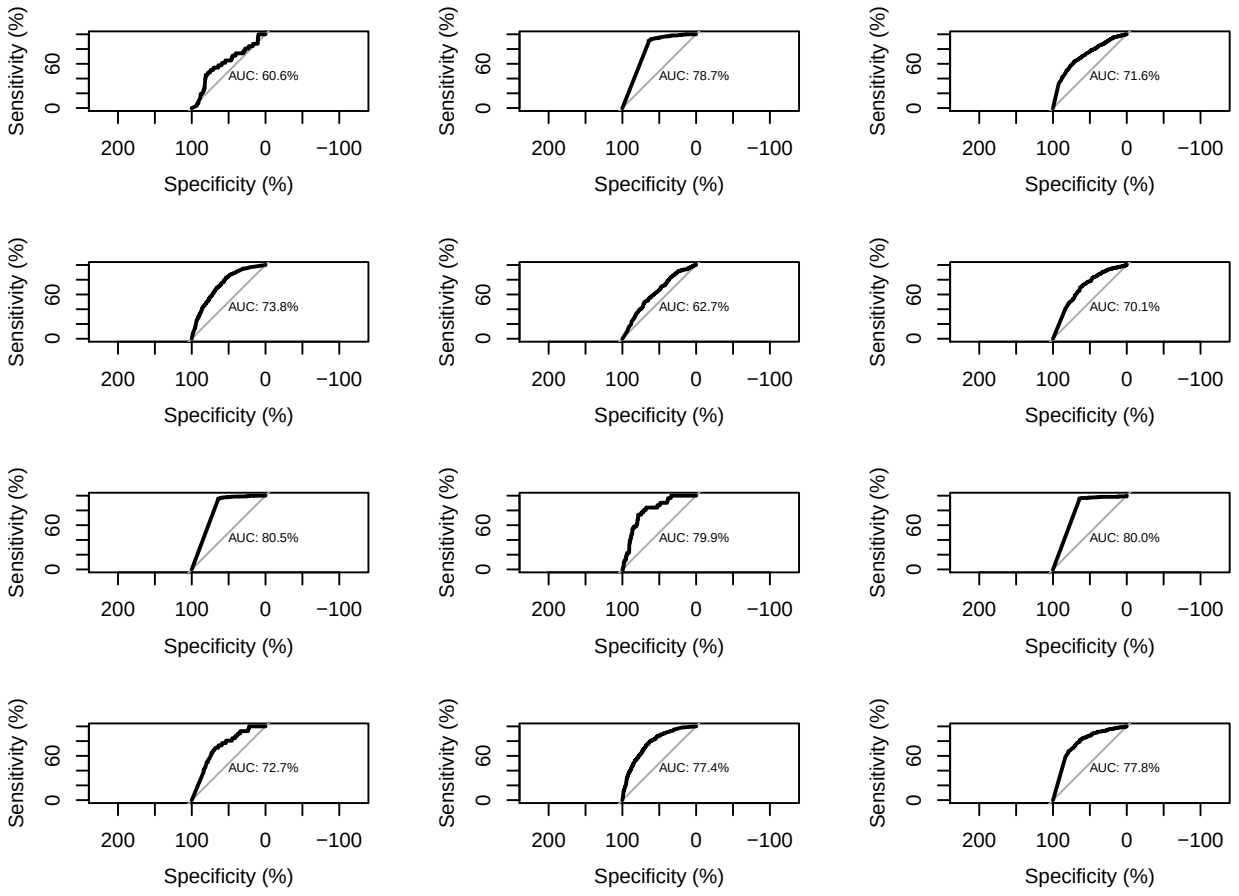
The results for KNN were:

```
## k-Nearest Neighbors
##
## 5253 samples
##   10 predictor
##    4 classes: 'intentional', 'natural', 'negligent', 'rekindling'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4727, 4729, 4727, 4728, 4727, 4727, ...
## Resampling results across tuning parameters:
##
##   kmax  Accuracy   Kappa
##   5     0.5530169  0.2418421
##   7     0.5808171  0.2771270
##   9     0.5891959  0.2863009
##
## Tuning parameter 'distance' was held constant at a value of 2
## Tuning
##  parameter 'kernel' was held constant at a value of optimal
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were kmax = 9, distance = 2 and kernel
##  = optimal.
```

Confusion matrix:

```
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##              Reference
## Prediction    intentional natural negligent rekindling
##   intentional        12.8     0.5       9.7        4.2
##   natural             0.3     0.2       0.3        0.2
##   negligent          11.7     0.6      41.3        5.4
##   rekindling          3.9     0.1       4.2        4.5
##
##  Accuracy (average) : 0.5892
```

Performance:




```
## Multi-class area under the curve: 73.82%
```

5.3 Probabilistic Approach

Discriminant analysis is used to predict the probability of belonging to a given class (or category) based on one or multiple predictor variables. It works with continuous and/or categorical predictor variables.

Compared to logistic regression, the discriminant analysis is more suitable for predicting the category of an observation in the situation where the outcome variable contains more than two classes. Additionally, it's more stable than the logistic regression for multi-class classification problems.

5.3.1 Naive Bayes

According to Bayes' theorem, it is possible to find the probability that a certain event will occur, given the probability of another event that has already occurred. Naive Bayes is a particular class of Bayesian classifiers that predicts the probability that a case belongs to a certain class. Due to its simplicity and high predictive power, it is one of the most used algorithms. This algorithm assumes that there is no dependency relationship between the attributes. However, this is not always possible. The algorithm reads the database and builds a probability table.In Bayesian classification, the main interest is to find the posterior probabilities, the probability of a label given some observed features.

5.3.2 Logistic Regression

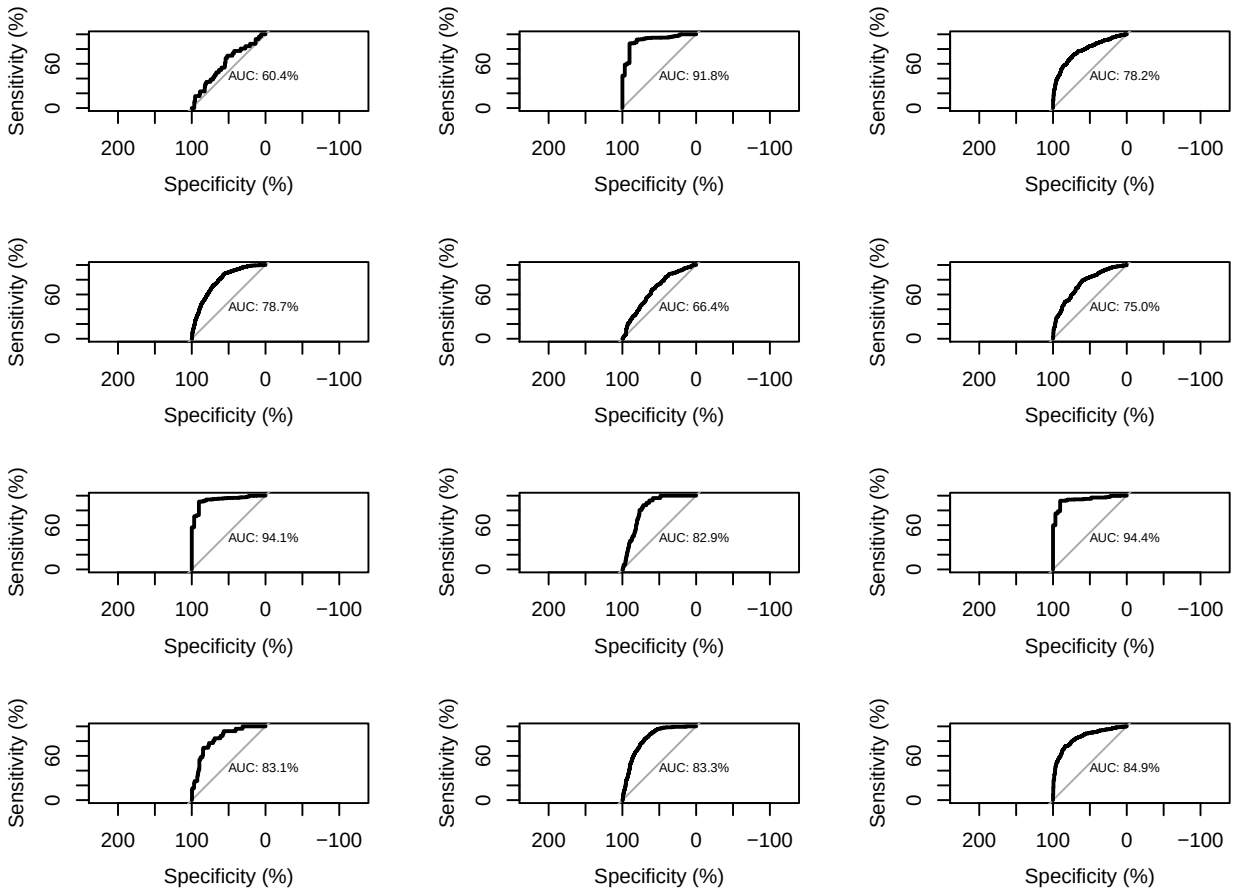
The results for Logistic Regression were:

```
## Penalized Multinomial Regression
##
## 5253 samples
##   10 predictor
##    4 classes: 'intentional', 'natural', 'negligent', 'rekindling'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 4728, 4726, 4725, 4728, 4729, 4727, ...
## Resampling results across tuning parameters:
##
##   decay   Accuracy   Kappa
##   0.0000  0.6264913  0.3265991
##   0.0001  0.6259203  0.3257233
##   0.1000  0.6274405  0.3262704
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was decay = 0.1.
```

Confusion matrix:

```
## Cross-Validated (10 fold) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##              Reference
## Prediction    intentional natural negligent rekindling
##   intentional        13.5     0.4       7.8        3.7
##   natural             0.1     0.1       0.1        0.0
##   negligent          12.8     0.9      44.9        6.4
##   rekindling          2.5     0.0       2.7        4.2
##
##  Accuracy (average) : 0.6275
```

Performance:




```
## Multi-class area under the curve: 81.1%
```

5.4 Mathematical Formulas

5.4.1 Linear Discriminants Analisys (LDA)

Linear discriminant analysis (LDA): Uses linear combinations of predictors to predict the class of a given observation. Assumes that the predictor variables (p) are normally distributed and the classes have identical variances (for univariate analysis, p = 1) or identical covariance matrices (for multivariate analysis, p > 1).

The LDA algorithm starts by finding directions that maximize the separation between classes, then use these directions to predict the class of individuals. These directions, called linear discriminants, are a linear combinations of predictor variables.

LDA assumes that predictors are normally distributed (Gaussian distribution) and that the different classes have class-specific means and equal variance/covariance.

Before performing LDA, consider:

Inspecting the univariate distributions of each variable and make sure that they are normally distribute. If not, you can transform them using log and root for exponential distributions and Box-Cox for skewed distributions. removing outliers from your data and standardize the variables to make their scale comparable. The linear discriminant analysis can be easily computed using the function lda() [MASS package].

5.4.2 Penalized Discriminants Analisys

Penalized logistic regression imposes a penalty to the logistic model for having too many variables. This results in shrinking the coefficients of the less contributive variables toward zero. This is also known as regularization.

The most commonly used penalized regression include:

ridge regression: variables with minor contribution have their coefficients close to zero. However, all the variables are incorporated in the model. This is useful when all variables need to be incorporated in the model according to domain knowledge. lasso regression: the coefficients of some less contributive variables are forced to be exactly zero. Only the most significant variables are kept in the final model. elastic net regression: the combination of ridge and lasso regression. It shrinks some coefficients toward zero (like ridge regression) and set some coefficients to exactly zero (like lasso regression).

5.5 Logical Approaches

5.5.1 Decision Trees

The Decision Tree classification method works as a tree-shaped flowchart, where each node indicates a test done on a value. The connections between the nodes represent the possible

values of the upper node test, and the leaves indicate the class to which the record belongs. After the decision tree is assembled, to classify a new record, just follow the flow in the tree starting at the root node until reaching a leaf. Due to the structure they form, decision trees can be converted into Classification Rules.

5.5.2 Tree Bag

Bagging (Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees.

5.6 Optimization Approaches

5.6.1 Neural Networks

A neural network is an adaptive system that learns using interconnected nodes or neurons in a layered structure that resembles the human brain. A neural network can learn from the data - so it can be trained to recognize patterns, classify data and predict future events.

A neural network divides the input into layers of abstraction. He can be trained using many examples to recognize patterns in speech or images, for example, just like the human brain does. Their behavior is defined by the way their individual elements are connected and the strength, or weights, of those connections. These weights are automatically adjusted during training according to a specified learning rule until the artificial neural network performs the desired task correctly.

5.6.2 SVM

It is used for both classification and prediction tasks. It consists of separating classes that can be separated by a straight line, called linearly separated classes. The model tries to trace the separation based on the best distance between the closest points. There are variations of SVM, such as the Kernell trick, which allows applying SVM to a set of nonlinearly separable data. Support vector machines for binary or multiclass classification.

Linear

It is an extremely fast machine learning algorithm for solving multiclass classification problems from ultra large data sets that implements a cutting plane algorithm for designing a linear support vector machine.

Radial

Radial kernel support vector machine is a good approch when the data is not linearly separable.The idea behind generating non linear decision boundaries is that we need to do some non linear transformations on the features which transforms them to a higher dimention space.We do this non linear transformation using the Kernel trick.

Polinomial

The polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of training samples in a feature space over polynomials of the original variables, allowing learning of non-linear models. Intuitively, the polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. In the context of regression analysis, such combinations are known as interaction features.

5.7 Ensemble Approaches

The ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

5.7.1 Random Forests

Random forest is a supervised learning algorithm which is used mainly used for classification problems. This algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It gets a random sample with replacement from the training set, select some features at random and reduce dimensionality of the set, accordingly and train a tree model without pruning. It predicts the class obtained by majority vote averaging the output of each tree.

5.7.2 XGBoost

Chapter 6

Conclusions, Shortcomings and Future Work

Chapter 7

References

https://www.mathworks.com/?s_tid=gn_logo

(Citeable URL: https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/zw12z7835)

PDA Reference http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/