Forest Fires in Portugal - What are the causes?

Data Mining I (CC4018) - 2020/2021: Practical Assignment

Rita P. Ribeiro

November, 2020

Description

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.

The ICFN - Nature and Forest Conservation Institute has the record of the list of forest fires occurred in Portugal for several years. For each fire, there is information such as the site, the alert date/hour, the extinction date/hour, the affected area and the cause type (intentional, natural, negligent, rekindling or unknown). In the file fires2015_train.csv, you have data on reported forest fires during 2015, for which the cause is known.

The attributes have information regarding the forest fire's alarm point and the total affected area:

- 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
- cause_type

The goal of this practical assignment is to build a machine learning model to predict the cause type of a forest fire.

As additional information, you can choose to use weather data. The script getTemperatureNOAA.R helps you with that. It allows you to get data from the National Oceanic and Atmospheric Administration (NOAA) Climate Data Sources. Namely, you can have access to the 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. In the station_data.RData file you have the data collected from all the stations and from which you can fetch weather information.

Tasks

Using the above data set, you have a set of main tasks to accomplish as described next. Still, you are free to include other tasks to increase the value of your assignment.

Task 1: Data importation, clean-up and pre-processing

In this part of your work you should focus on importing the provided data into an appropriate R format so that your posterior analysis is made simpler. You should also check if it is necessary to carry out any data clean-up and/or pre-processing steps.

Task 2: Data exploratory analysis

This part involves summarising and visualising the data to provide useful insights. Think about questions that could be interesting to check with the available data, and provide answers either using textual summaries or data visualisation.

Task 3: Predictive modelling

From the available data, you should define the data set that will be used for the classification task at hand. Different models should be considered and the choice of the final model should be justified.

Task 4: Kaggle Competition

Additionally, you should submit your solution for the fires2015_test.csv data set into the Kaggle Competition. Your accuracy rank will be accounted for the final grade.

Tools

In your work, you can use R or Python. In case you choose to use R, you can find material for dynamic reporting in R with markdown here. If you choose to use Python you can use the colab research.

Deliverables

The practical assignment is **mandatory** and should be performed by groups of, **preferably, three students**.

Until the next **December, 9th, 2020** you should inform me of the group constitution: full names and student numbers. After this date, no more groups are accepted.

Your assignment should be sent to me by email with the subject "[DMI] Group X", where X is the number that I will assign to your group, and including the following items in a compressed file:

- a final report [1] in PDF or HTML generated dynamically with the identification of group members, and with a structure similar to the following:
  - introduction;
  - problem definition;
  - data pre-processing;
  - exploratory data analysis;
  - predictive modelling: experimental setup and obtained results;
  - conclusions, shortcomings and future work;
  - appendix (optional).
- the source of a ready to execute dynamic report that produces your final report with all the code that is necessary to run to obtain the results you present;
- any complementary files needed to execute your report (e.g. data files, data objects).

Grades

The grading of the practical assignment is distributed as follows: Task 1 (20%), Task 2 (20%), Task 3 (20%), Task 4 (20%), quality of the report (10%) and presentation (10%). Presentation is mandatory.

Important Notes

- Any data pre-processing steps must be presented and justified.
- All the algorithms and parameters used should be indicated.
- Organization of discourse and presentation, clarity of language and ideas are rewarded.
- Long sequences of poorly formatted output dumps are penalized.
- The report should also refer to any source you used and make explicit which part of the work it influenced.
- It is important that your code does not rely on an absolute path so that it can be run on any computer.
- The R/Python code should not appear on the report, just the output of it.
- If your report contains code that takes too much time to run, you should include the code but do not execute it in the report generation. You can load its output as a binary file.

Deadline

The deadline for submitting the practical assignment is $\boxed{\textbf{January, 10th, 2021}}$.

[1] Please keep your report under 4-5 pages. You can freely attach appendices with additional information you consider relevant, but your work should be perfectly understood by the report alone.