Fire prediction with logistic regression on territory of Bosnia and Herzegovina

Mursel Musabašić, Denis Mušić and Elmir Babović
"Džemal Bijedić“ University of Mostar, Faculty of Information Technologies

{mursel.musabasic} {denis} {elmir.babovic}@edu.fit.ba

Abstract. The Canadian Fire Weather Index system [1] has been used worldwide by many countries as classic approach in fire prediction. It represents system that account for the effects of fuel moisture and weather conditions on fire behaviour. It numerical outputs are based on calculation of four meteorological elements: air temperature, relative humidity, wind speed and precipitation in last 24h. In this paper meteorological data in combination with Canadian Fire Weather Index system (CFWI) components is used as input to predict fire occurrence using logistic regression model. As logistic regression is a supervised machine learning method it’s based on user input in the form of dataset. Dataset is collected using NASA GES DISC Giovanni web-based application in the form of daily area-averaged time series in period of 31.7.2010 to 31.7.2020, it’s analysed and pre-processed before it is used as input for logit model. CFWI components values are not imported but calculated in run-time based on pre-processed meteorological data. As a result of this research windows application was developed to assist fire managers and all those involved in studying the fire behaviour.

1. Introduction

Fire can be a process caused by a natural sequence of events or it can occur as a human-caused activity. Fire also represents one of the key elements in forest renewal and succession. Beside it's natural behaviours, fire is global issue which affects, not only ecological and economical environments, but also affects all living things. How great the danger of fire is can be shown through recent events in the world. During 2019 – 2020 Australian bushfires burnt over 12 million hectares, destroying thousands of homes, and killing at least 30 people [2]. There were over hundreds of millions wild animals killed. In California 2020 fire season there were almost ten thousand fires detected resulting in at least 30 deaths and over four million hectares burnt [3]. Thousands of buildings were destroyed. Economists estimated that both countries suffered cost over hundreds of million US dollars.

Although smaller in territorial size, Bosnia and Herzegovina has problems with fires during it's fire seasons. In period of 2010 to 2020, European Forest Fire Information System (EFFIS) established by Joint Research Centre (JRC) and Directorate General for Environment of the European Commission EC) [4] reported at least 970 number of fires resulting in 380.000 hectares of burnt area (about 7.5% of country size).

Of the total burnt area, over 30 thousand hectares fall on agriculture areas, about 145 thousand hectares of forest areas, and over 100 thousand on other natural land and artificial surfaces. Worst fire seasons were recorded in 2012, 2017, 2019 and 2020 with 767 number of fires and about 305 thousand hectares of burnt area.
Table 1. Last decade fire seasons statistics for Bosnia & Herzegovina

| Year | Burnt areas (ha) | Number of fires |
|------|-----------------|-----------------|
| 2010 | 3692            | 10              |
| 2011 | 23676           | 41              |
| 2012 | 92809           | 147             |
| 2013 | 3616            | 10              |
| 2014 | 3203            | 5               |
| 2015 | 15318           | 26              |
| 2016 | 34538           | 89              |
| 2017 | 83449           | 146             |
| 2018 | 3155            | 12              |
| 2019 | 28839           | 134             |
| 2020 | 100107          | 340             |
| Total| 392402          | 960             |

Fire season of 2020 recorded the greatest number of fires and areas burnt. A reason for this number of fires can be found in the situation where the population left urban areas and went to rural areas thus escaping possible corona-virus infection. From an economic aspect, the fires that occurred in the last decade have cost Bosnia and Herzegovina almost 60 million US dollars.

Even 25 years after the war country is struggling with formal strategic plans in fire prevention. In addition to the fact that firefighters use old equipment, there are almost no indications that the early warning system will improve soon [5]. The aim of this study is to show that it is possible to provide at least a step towards better fire risk management with the aid of supervised machine learning methods.

![Fire trend shows faster grow during 2020.](image)

2. Data and methods

2.1. Data preparation

Dataset for logistic regression model consists of three datasets acquired through NASA GESDISC Giovanni web application. Air temperature (°C) and relative humidity (%) are collected from AIRS3STD dataset [6]. Precipitation (mm) is collected from GPM_3IMERGDL_06 dataset [7] and wind speed (m/s) from M2T1NXFLX_5_12_4 dataset [8]. Air temperature, relative humidity and precipitation raw values are in daily temporal resolution as time series, area averaged. Wind speed data is based on hourly temporal resolution, and it is subsequently converted to daily temporal resolution including unit of measure conversion from m/s to km/h. Time period of dataset ranges from 31.7.2010 to 31.7.2020 counting in 3654 rows where every row corresponds to single day. There are two phase of data preparation. The first step is done using Microsoft Excel, but any other spreadsheet application can...
be used, and second through application. During data loading Canadian Fire Weather Index components are calculated. As many other countries Federal Hydrometeorological Institute of B&H also use CFWI system [9] to calculate six components through WRF-NMM model created by NOAA's National Weather Service. First three components (FFMC, DMC and DC) are so-called „fuel moisture codes“ and other three components (ISI, BUI and FWI) are called „fire behaviour indices“ (Wagner 1987). Every code and index is represented by numerical value. All six components can contain own scale of relative values, but they are structured in way to indicate potential fire burning behaviour. With higher values of fuel moisture codes, the lower are the values of indices and vice versa.

Figure 2. CFWI structure [1]

FWI (Fire weather index) is the result based on previous components calculation outputs. As per EFFIS, FWI output value is divided in six fire danger classes as shown in Table 2.

Table 2. Fire danger classes per EFFIS

| Fire danger classes | FWI value ranges |
|--------------------|-----------------|
| Very low           | < 5.2           |
| Low                | 5.2 – 11.2      |
| Moderate           | 11.2 – 21.3     |
| High               | 21.3 - 38       |
| Very high          | 38 - 50         |
| Extreme            | > 50            |

In final step of data preparation output values for dependent variable „Fire“ are calculated based on following condition: if final value of FWI is over or equal 11.2 it will generate 1 as fire present indication else 0 as there is no fire present.

2.2. Model training

By default, application is designed to split prepared dataset in ratio 80:20 within which 80% of data is used as training set and 20% of data is used for model testing. There is option for users to input custom trainset size in range 0.0 to 1.0. Besides splitting data in train and test dataset's, there is an option for user to get insight of correlations between independent variables. Before training user needs to select appropriate columns as input for a logistic model.

2.3. Logistic regression

Logistic regression - is a method coming from classic statistics. It is used to model binary dependent variables. With logistic regression we can predict if something is true or false, right or wrong, is someone ill or not, 1 or 0, etc. Logistic regression assumes existence of linear relationship between dependent variable and independent variables [10]. This relationship is expressed by equation as shown below:
log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1x_1 + \cdots + \beta_mx_m \tag{1}

p - \text{probability of event being true}
\beta_0 - \text{bias}
\beta_1 \ldots \beta_m - \text{beta coefficients}
x_1 \ldots x_m - \text{independent variables, predictors}

Left of equation is called log of odds and its main step in linear transformation from non-linear form. On the right side we have beta coefficients which are also called weights and are in direct association with independent variables. Weights are real numbers and give importance to independent variables in classification decision and determine how predictors influence on classification decision if is something true or false. This linear form is written as:

\[ z = \left( \sum_{i=1}^{m} w_i x_i \right) + b \tag{2} \]

\[ w_i - \text{weights} \]
\[ x_i - \text{predictors} \]
\[ b - \text{bias} \]

To be able to solve linear combination of predictors and weights (2) we need a function which is called sigmoid or logistic function:

\[ y = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3} \]

Logistic function outputs values in range from 0 to 1. The sigmoid function results in effect that if one is divided by a big number, it will get closer to zero, and if it’s divided by number a little bit bigger than one it will get closer to one. Sigmoid function is used to map predictions values to probabilities values.

2.4. Calculating the parameters
To find best parameters for our model we use maximum likelihood estimation. First, we determine log-likelihood function then we find our weights. As sigmoid function (3) outputs our values as probabilities we can interpret them as Bernoullies variables, Y = Ber(p), where:

\[ p = \sigma(\omega^T x) \tag{4} \]

Given Bernoulli probability function for single point:

\[ P(Y = y | X = x) = p^y(1-p)^{(1-y)} \tag{5} \]

we can write likelihood function for complete data as:

\[ L(\omega) = \prod_{i=1}^{n} P(Y = y_i | X = x_i) = \prod_{i=1}^{n} p^{y_i}(1-p)^{(1-y_i)}, p = \sigma(\omega^T x_i) \tag{6} \]

The reason we are multiplying all the data points is that we are assuming they are independent and not related (0/1, fire/no fire, etc). So, we need to maximize our likelihood function and to avoid multiplication form we take log of likelihood function.
\[ LL(\omega) = \sum_{i=1}^{n} y_i \log(p) + (1 - y_i) \log(1 - p), p = \sigma(\omega^T x_i) \]

As we want to maximize our log-likelihood we take derivative of our function and move in direction of gradient using gradient ascent algorithm.

\[ \frac{\partial LL(\omega)}{\partial (\omega)} = (y - p)x, p = \sigma(\omega^T x) \]  \hfill (8)

Taking in account derivative of LL our update weight rule becomes like this:

\[ \omega_{i+1} = \omega_i + \frac{\partial LL(\omega)}{\partial (\omega)} \alpha \]

\[ \alpha = \text{learning rate, scaling size} \]  \hfill (9)

To practically solve our problem and as it's more convenient, update weight rule is calculated through our software application iteratively including the creation of complete logit model.

2.5. Fire prediction software application

Specifically, for this study we have created software application which can be used by people who are involved in wildfires research area. In this early version of software application, users can choose multiple options in fire analysis/prediction, such as correlation, maximum number of iterations, step size/learning rate, regularization penalty (only L2), data split, data shuffle (Fisher-Yates method) by iteration and data normalization using Min-Max scaler.

Besides training process setup application results contains values for: cost, McFadden pseudo-R2, accuracy on train/test data, sensitivity on train/test data, confusion matrix, AIC and BIC model selection methods and final regression equation.
Figure 4. Fire prediction input

Figure 5. Correlation matrix window

3. Results

Three models were compared during model training. In first model (model A) meteorological data is combined with fuel moisture code components (FFMC, DMC and DC), second model (model B) is combination of meteorological data and fire behaviour indices (BUI, ISI and FWI) and the third model (model C) includes only meteorological data (temp, RH, wind, and rainfall). As rainfall values deviate in higher percent than others variable values data needed to be normalized before training process.

Following table shows prediction power of tested models:

|                         | Model A | Model B | Model C |
|-------------------------|---------|---------|---------|
| Cost                    | 0,1567  | 0,0813  | 0,1969  |
| McFadden $R^2$          | 0,6691  | 0,8283  | 0,5841  |
| Accuracy on trainset    | 0,93    | 0,97    | 0,92    |
| Accuracy on testset     | 0,62    | 0,86    | 0,88    |
| Sensitivity / Specificity (test) | 0,63 / 0,94 | 0,70 / 0,95 | 0,58 / 0,94 |
| AIC                     | 929,78  | 489,17  | 1158,93 |
| BIC                     | 971,64  | 531,03  | 1182,85 |

Looking at the Table 3, we can conclude that model B have most desirable results. Compared to other models it’s loss function returned smallest error value. With largest pseudo-$R^2$ value and by sensitivity and specificity percentages which are derived from confusion matrix, shows more accurate
prediction compared to model A and C and therefore the best-chosen logistic regression model. AIC and BIC selection methods returned lowest values for model B and thus confirm the above statements. These results tell the fire manager that fire behaviour indices (including wind as weather element) have most significant impact on the occurrence of fire and its behaviour.

4. Conclusion

“All models are wrong, but some are useful”. This citation is attributed to the British statistician George Box. It means that models will always be wrong because of the complexity of the nature of the things being modelled, but they can be useful. Although machine learning methods can be very helpful, they still require a larger amount of data to create models as accurately as possible. This means that countries (especially Bosnia & Herzegovina) must in some way establish a stable and reliable network of many meteorological stations where measurements will be made constantly, and the stored data will be used for various types of model evolution. Despite the above, the model created as part of the research, which is, as stated in the research, based on average satellite data for 4 basic meteorological elements, the evaluation of the model still showed almost 87% accuracy. This percentage is not so small, if we consider missing variables which are almost always taken into account when evaluating fires, such as distances from populated areas, vegetation index, elevation, forest types, etc.

References

[1] Canada N R Canadian wildland fire information system: Canadian forest fire weather index (FWI) system [Internet]. Canadian Wildland Fire Information System | Canadian Forest Fire Weather Index (FWI) System [cited 2020Oct12] https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi
[2] Elisabeth P L Y 2020 The 2019-2020 Australian bushfires IDMC 2020 [cited 2020Oct15]. https://www.internal-displacement.org/publications/the-2019-2020-australian-bushfires-from-temporary-evacuation-to-longer-term
[3] California Department of Forestry and Fire Protection (CAL FIRE) 2020 incident archive Cal Fire Department of Forestry and Fire Protection [cited 2020Oct17] https://www.fire.ca.gov/incidents/2020/
[4] European Forest Fire Information System Gwis statistics [cited 2020Oct5] https://effis.jrc.ec.europa.eu/apps/gwis.statistics.portal
[5] Fire risk management: UNDP in Bosnia and Herzegovina [Internet] UNDP 2020 [cited 2020Oct10] https://www.ba.undp.org/content/bosnia_and_herzegovina/en/home/climate-and-disaster-resilience/FireRisk.html
[6] AIRS S T and Teixeira J 2013 AIRS/Aqua L3 Daily Standard Physical Retrieval (AIRS-only) 1 degree x 1 degree V006 (Greenbelt: Goddard Earth Sciences Data and Information Services Centre - GES DISC) [cited 2020Sep10] doi:10.5067/Aqua/AIRS/DATA303
[7] Huffman G J, Stocker E F, Bolvin D T, Nelkin E J and Jackson T 2019 GPM IMERG Late Precipitation L3 1 day 0.1 degree x 0.1 degree V06 (Greenbelt: Goddard Earth Sciences Data and Information Services Centre - GES DISC) [cited 2020Sep12] doi:10.5067/GPM/IMERGDL/DAY/06
[8] Global Modelling and Assimilation Office 2015 MERRA-2 tavg1_2d_flx_Nx: 2d, 1-Hourly, Time-Averaged, Single-Level, Assimilation, Surface Flux Diagnostics V5.12.4 (Greenbelt: Goddard Earth Sciences Data and Information Services Centre - GES DISC) [cited 2020Oct6] doi:10.5067/7MCPBJ41Y0K6
[9] Federal Hydrometeorological Institute of B&H Prognoze Index opasnosti od požara – FWI [cited 2020Oct20] http://www.fhmzbih.gov.ba/latinica/AGRO/pozar.php
[10] Sampson M G 2018 Statistical Analysis in JASP – A Students Guide Logistic regression 2nd Edition JASP v0.9.1 p 70