

A Study on forest fire risk

Using forest fires in Portugal to build a model that aids risk assessment and preparedness



December 11, 2018

Intro to machine learning

Michael Griffith

Wildfires are a dangerous, expensive, and unpredictable part of life for many areas, my goal is to try to use machine learning to help prevent and alleviate some of the damage caused by these natural disasters. Often public officials and emergency workers are caught by surprise by the intensity and area burned from a forest fire, especially out of season, or they over prepare for “fire season” and waste resources when the frequency and intensity of burns is abnormally low. Through machine learning we can attempt to minimize this waste without sacrificing readiness and shift those saved resources to be more prepared for irregular fires outside of the expected windows where these fires occur. To build a model for forest fire prediction I used a set of data gathered from over five hundred data points based on fires in Portugal.

At first, I had attempted to use stochastic regression to solve this problem and estimate the area affected on a week by week basis by considering the FWI (Forest Fire Weather Index) system variables. This was only somewhat successful. Although accurate at predicting the average size of fires over the course of a month or a given time period, it was not valuable in the practical use case I was trying to solve. I needed to be able to determine risk for that particular data point (that week of data) in the moment and send an alert. Regression models were simply not effective at changing quickly from expecting a low risk to expecting a high risk, it was only effective in tracking the expected average fire intensity over a time. Since I was interested in developing a more efficient and responsive warning system, this simply did not satisfy my goal.

To make a more practical system I turned to classifying each measurement of FWI variables into risky or not risky categories using K Nearest Neighbor algorithms. The first portion of my analysis will focus on what separates the times when fires are frequent and intense from the “slow seasons” where fires seem to be rare and/or tame.

Fig. 1 – What time of year do the biggest (more than one standard deviation above the mean) fires occur?

|  |  |
| --- | --- |
|  |  |

It seems that all of the large fires (those that have greater than one standard deviation above the mean for all the fires in this data set) happen in the range of July-September, with only two fires over fifty hectares (which is about six tenths of a standard deviation above the mean) happening outside of this range with one occurring in February, and one in June. However, as we know forest fires do not generally use calendars so there must be other variables contributing to why these months are much more common for large fires. Let’s look at the FFMC (Fine Fuel Moisture Code), DMC (Duff Moisture Code), DC (Drought Code), ISI (Initial Spread Index), and the temperature for these months compared to the mean across the data we have for all the fires that occurred throughout the year. For more information about these codes you can visit <http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi> to better understand the FWI system that these codes are used for.

Fig. 2 – FWI variable average vs. month as standard deviations above the mean

We can see in Figure 2 that the DC (Drought Code) rating is high in the months where fires are most common and most intense, while these months also have higher than average ratings in all other variables as well (apart from September having a slightly below average ISI). DC rating means that the deep organic layers (e.g. large logs/fallen trees) are low in moisture from a long period of lower than average rainfall which is expected near the end of the warmer summer months. This provides a huge source of fuel for the forest fires, but it is not the only variable that can affect the ability for a forest fire to rage out of control. We also see a high DC rating in October, however we show no large fires in that month, so we must also look at the combination of other variables that contribute to a frequent and devastating string of fires, and the variables that contribute to calm months so that we can isolate high-risk times in the slow months, and low-risk times in the months that are generally more dangerous so that we can be less surprised at abnormally timed blazes, and more aware of when fires will happen during the hotter Summer months.

In the months leading up to September we see a high rating across all the FWI variables, while in October we see a steep drop in DMC (Duff Moisture Code) which means that the loose and medium sized woody materials have much more moisture than in September, normally indicative of rainfall after a dry period. This prevents the forest fires from latching on and becoming uncontainable as we can see similar values for all other variables in October that we see in the previous three months that are our “hot spot” for large fires. Therefore, we can assume that large fires require the right combination of dryness throughout all layers of the fire fuel ecosystem as well as average to above average wind conditions, but with the lack of any of these features the fires are stunted. For example, we can see high wind and temperatures in June but a low DC rating, so we see fewer intense fires than the following months where the DC rating continually rises without dropping in other measurements on the FWI system. Our prediction model will be centered around the points we have gleaned from the initial data analysis by focusing on the FWI ratings to predict large forest fires. Also, since an exact measurement is not practically useful in the case of forest fires (response and resources available, luck of spotting a fire early, etc. all play into the area burned before containment that are outside of the measurement in this data set) , we will focus on classifying our prediction of the area affected into small (zero hectares burned), moderate-low (from above zero to ten hectares burned), moderate (from ten to fifty hectares burned), and severe (more than fifty hectares burned) fires.

Using K Nearest Neighbor to train the data, and a k-fold cross-validation scheme and a k-value of fifty we find that the prediction system is usefully accurate for our goal of warning when conditions could be dangerous, correctly classifying forty-eight out of the fifty data points used as the validation set and not under-estimating the risk of any time section. Each data point is classified into a set of four different classes based on the risk that the current condition of the forest poses for a fire. By comparing the standard deviations above or below the mean I was able to predict the nature of risk based only on quantitative data of forest conditions independent of temperature, season, or even sector the measurements are taken in. Of the two that were not classified correctly both are what we would consider “high-risk” for a larger fire, both data points were in September, the most dangerous month for forest fires, and both had very high numerical ratings across the FWI system. Those two data predictions, even though they did not materialize into dangerous forest fires, had all the markers we found in our earlier analysis to give the system the impression they could potentially be higher-risk times for an intense and widespread forest fire.

Here is the most accurate (called Sample 1 here) and the least accurate (called Sample 2 here) of all the ten validation tests:

Building a system that helps to utilize resources effectively still must err on the side of caution over efficiency to avoid unnecessary loss of life and property by skewing risk reports towards a higher risk rather than focus on the most accurate prediction of what will happen. This is the most prudent course of action since fires can be devastating and underestimating a risk is much worse than overestimating. I believe that the classification system I have in place is a good compromise of increasing efficiency and having a high rate of true positive as well as minimizing false positives with the highest priority on eliminating false negatives for severe events.