Predicting Wildfires

A Machine Learning Project

By David Cavanaugh

1 Introduction

1.1 Background

Forests, shrub-land, and grassland cover more than half of the land in the United States, and while fires can be a vital part of maintaining a healthy ecosystem, when fires burn out of control or when fires are caused unnaturally - by human causes - the natural status quo can be disrupted leaving a mark on the ecosystem which can persist years after the wildfire. In addition, there is increasing amounts of research which suggest that climate change has exacerbated the wildfire problem, with evidence showing that wildfires are increasing in duration, size, and frequency - in part due to the effects of climate change [*citation*]. The National Inter-agency Fire Center compiles statistics on wildfires which occur within the United States, they combine reports from local, state and federal agencies that are involved in fighting wildfires. According to The National Wildfire Coordinating Group (NWCG), a wildfire is "a wild-land fire originating from an unplanned ignition, such as lightning, volcanoes, unauthorized and accidental human caused fires and prescribed fires that are declared wildfires [*citation*]".

1.2 Significance

From 1993 to 2018, the average number of wildfires per year was 80,250 with an average total area of 6.098 million acres burned, for comparison, that would be like burning the entire state of Vermont, every year, since 1993. These wildfires account for billions of dollars of damages and sometimes even lead to mass evacuations and loss of life. Wildfires also decrease air quality with smoke impacting not only local regions but large swaths of the country as it is swept across the U.S. by the west-east prevailing winds. With all of that in mind, the frightening trend is that wildfires appear to be increasing in size, but not decreasing in volume. This trend brings to the forefront of millions of Americans minds the risk posed to them and their livelihoods by wildfires. With the expectation that in the near future the risk posed by wildfires will increases there is an increased motivation to understand when and where these wildfires will be taking place to that proper preventative measures can be taken.

1.3 Research Question

In this project it is our goal to utilize over 25 years of wildfire data, with observations on over 1.5 million individual wildfires, to predict the probability of a wildfire occurring in a given county, during a given month, which will burn more than 5,000 acres (a fire classification code of G). The modeling will be a mix of time series and and categorization methods which will produce probabilities from 0 to 1. From those probabilities we can identify locations and months which are most likely, or most susceptible, to a large wildfire, and compare them to the data from 2019 and 2020 as a validation for our predictions.

2 Data

2.1 Source

The source of the data is the U.S. Department of Agriculture and is published in the Research Data Archive on their website under the title of *Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617] (5th Edition)*. The data was published in 2021 by Karen C. Short. A full citation can be found in the bibliography (*coming soon*). For additional details see the website `https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009.5`

2.2 Overview

For a view of the data see the table below which details a single record from the table (coming soon - see the Jupyter Notebook for now). The table itself contains 2,166,753 records, each of which corresponds to a unique wildfire - identified by the FOD_ID. Most of the columns in the table are id columns which help to identify the fire and categorize it. We will select only a few of these columns for our analysis.

2.3 Features

Here we will summarize each of the features that we will use, what the feature is, and how it will be used. Not every feature will make it to the final dataset, however, they are all features which may be used for analysis or possibly to further classify the data so that we can predict on like subsets of data.

1. **FOD_ID:** This is the fire id, it serves as a unique id which can identify each fire. It is not used in analysis, just as a primary key for reference.

2. **DISCOVERY_DATE** and **DISCOVERY_TIME:** This is the date and time that the fire was first discovered or confirmed to exist. It will be used to indetify what month the fire took place in, this will become the primary source for the time series aspect of this project. The time specifically could be used in conjunction with the contained date and time to calculate fire duration more precisely.

3. **NWCG_CAUSE_CLASSIFICATION:** A categorical variable (described more on Notebook) which indicates the cause of the fire. It will be used as an additional feature for prediction - it may be that natural vs. human caused fires are different to predict or are able to be predicted better (or worse).

4. **CONT_DATE** and **CONT_TIME:** The contained date and time. Could be used to calculate fire duration which could be interesting to analyze, but I am unsure if duration will actually be used as a feature since it will likely be very similar to FIRE_SIZE and overall not entirely useful since duration is only known *after* the fire happens.

5. **FIRE_SIZE** and **FIRE_SIZE_CLASS:** Fire size and fire class size are effectively the same information since the class is based on the acres burned, which is the fire size column. For our analysis we will focus primarily on the number of class G fires, which is fires that consume 5000+ acres.

6. **LATITUDE** and **LONGITUDE:** These are ancillary geographical identifiers. They can be used (with the govt. API) to get the FIPS_CODE when it is missing. This is especially useful for the G class fires which are missing only the FIPS_CODE since there are only a limited number of that size fires we want to retain as many of those records as possible.

7. **FIPS_CODE:** The FIPS_CODE identifies the state and county as a five digit code (first two for state, last three for county) where the fire originated. This will be used to identify the geographic regions (state & county) level for where the fire took place. Some fires span multiple counties so this identifies the origination county.

2.4 Missing Data

Any missing data will be dropped from out dataset. This is only a concern when we are dropping observations which are fires of size class G, since there are fewer than 5,000 total of that size class. We can use the FCC API to fill in gaps in the for the missing FIPS codes, however, it is likely unfeasible to fill in all 600,000 plus missing FIPS, so many of them will likely be dropped from the dataset. From preliminary analysis it doesn't appear that any particular type of fire or size of fire is more likely than another to be missing data so dropping data shouldn't skew our dataset in any particular direction. Beyond missing FIPS codes, the only other vital column is the discovery date which has no missing data.

2.5 Formatting

The final formatting will come by aggregating the data down to month and FIPS code, where for every month and every FIPS code present in the final data will have a value of zero or one, indicating whether or not there was a size class G fire in that FIPS during that month. For a more clear depiction of what that dataset will look like observe the dataset in the Jupyter notebook. One important observation in terms of future modelling concerns is that this data is particularly sparse. There are hundreds of FIPS and hundreds of dates, so, the matrix created will have hundreds of thousands of values, however, there are fewer than 5,000 fires which are classified in the G size. We could potentially loosen this description to include both F and G to reduce the sparsity of the matrix, but for now G is the only class we are considering.