CS 760: Machine Learning Fall 2020

Wildfires in the United States

AUTHORS: Joseph Mohr

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED OR MULTIPAGE.

Wildfires can devastate communities and landscapes. As temperatures rise due to climate change, it makes sense to think that wildfires will grow in frequency and intensity. Is this supported by data up to this point? This project will explore this proposition. It will then try to determine with what accuracy we can predict the general acreage a wildfire will burn in its lifespan. This type of information could hopefully help tell important factors of larger-burning wildfires that could help direct focus on certain fires or be a good point towards additional funding for fighting wildfires.

1 Introduction

Machine learning has been applied to nearly everything imaginable, from object recognition to predicting the outcomes of sporting events. It has also been applied to more natural events, such as earthquakes. Can it also then be applied to determining the severity of wildfires based on information such as cause, temperature, and location? This project will work towards exploring that as well as analyzing a dataset of over 1.8 million wildfires occurring in the United States from 1992 through 2015.

This project aims to explore if it is possible to determine important aspects of a wildfire that play a role in determining its severity to better focus resources on certain new fires. This is done through data analysis and using several machine learning models to predict the severity of a wildfire giving data that would generally be available at the start of a wildfire. The impact that differences causes can have will be explored in the data analysis section, along with other explorations. Then the results of the machine learning experimentation will be presented to see how the models performed at predicting the fire size class of wildfires.

The code is located at https://github.com/jvmohr/cs760_final/tree/master.

2 Related Work

While I wasn't able to find similar work trying to predict wildfire size, other machine learning tasks were performed in various notebooks at https://www.kaggle.com/rtatman/188-million-us-wildfires/notebooks. These usually tried to predict the cause of the fire as opposed to the severity given data that would be available at the start of a wildfire.

3 Datasets

Two datasets were used for this project, one relating to wildfires in the United States and one relating to Earth surface temperature.

The dataset regarding US wildfires, "1.88 Million US Wildfires", can be found at https://www.kaggle.com/rtatman/188-million-us-wildfires. It contains various information on over 1.8 million wildfires across the United States from 1992 to 2015.

The dataset regiarding Earth surface temperature, Climate Change: Earth Surface Temperature Data, can be found at https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data. It contains average temperature per month and some other data for various cities and each state for a wide range of years, up through 2013.

The temperature dataset was joined to the wildfires dataset based on the year, the month, and the state. So the average temperature for the month gets matched to the month in which a given wildfire was first discovered in. At this point, the dataset has feature columns for fire year, latitude, longitude, month, average temperature, average temperature uncertainty, the cause of the fire, and state in which the fire occurred. The dataset also contains a continuous and a discrete column for fire size.

The discrete column for fire size is categorized as follows:

- A = 0.0 - 0.25 acres
- B = 0.26 - 9.9 acres
- C = 10.0 - 99.9 acres
- D = 100 - 299 acres
- E = 300 - 999 acres
- F = 1000 - 4999 acres
- G = 5000 + acres

4 Approach

After the two datasets are joined as described in the previous section, the data is almost ready to be fed to machine learning algorithms, but there are several more things to mention first.

4.1 Preprocessing

First, at this point, all the wildfires that occurred after the end of the temperature dataset are dropped as they don't have a temperature value associated with themselves. The state, average temperature uncertainty, and year columns have also been dropped.

One hot encoding is then done over the cause of the fire column to create thirteen binary columns. The cause column is then dropped.

The dataset is then split into training and testing datasets with seventy-five percent of the data going towards the training set. During this step the fire size class is designated at the y feature.

Standard scaling is then performed on the data. It is fit using the training set and then applied to both the training and testing sets. Standard scaling is subtracting the mean of a column and then dividing by the standard deviation for each data point.

4.2 Machine Learning

I have used two classification machine learning algorithms, k-nearest neighbors (KNN) and random forest. The goal of both of these algorithms is to accurately classify new wildfires as having one of seven fire size classes. KNN is used with k=5, neighbor weight being the inverse of its distance, and Euclidean distance being used to compute distance. Random forest is used with 100 decision trees and balanced class weights due to the uneven distribution of fire size classes.

For these two models, two metrics have been used. Obviously, accuracy is the first. Then, also due to the uneven distribution of fire size classes, balanced accuracy is also computed. This is computed by taking the average of the recall obtained for each class.

This whole process was repeated with removing those samples that have a fire size class of A. A is a very big class, so this speeds up the process while also helping the accuracy. I believe this can safely be done as this class A corresponds to a fire that burned less than a quarter of a acre of land. These fires shouldn't last long enough for there to be any need to predict its severity.

For the implementation, I used the Python library sci-kit learn in large part due to the large size of the dataset (over one million rows). In a normal semester I likely happily would've spent a lot of time optimizing my own algorithms so they would run in a reasonable amount of time. My own general implementations are included in my repo as well. Anyways, I will describe the machine learning algorithms used now.

4.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a supervised machine learning. As the name implies, the general idea is that any new sample is matched with its k "nearest neighbors." Its predicted y-value, whether for classification or regression, comes from those neighbors then.

From this description, there's two big questions that arise. First off is "What value does k take?" and second is "How are these nearest neighbors decided?" For this problem, I've set k equal to 5 for the final results. My general intuition is that k can't be too large for this problem due to the massive number of samples with class A or B. If k is too small, though, it's very likely the new sample will be near a few of those class A or B points, no matter what the new point's class is.

As for the distance, standard Euclidean distance between two points is used. In addition, the inverse of these distances among the k neighbors is used so that closer neighbors are weighed more heavily than the further neighbors.

KNN, as implemented in scikit-learn, also has a speedup (potentially) in how it computes the neighbors. It has three possible modes for how to compute them: a K-D Tree, a Ball Tree, or brute force. Brute force will search through all the samples to find the neighbors (as my implementation of KNN that isn't used here does), but that is quite slow when dealing with a dataset of this size. The other two both used a tree to speed up this process. When creating these models, I didn't set this parameter so that it would chose the best on its own. It chose a K-D Tree, so I will quickly describe that as well.

In short, a K-D Tree is a binary tree that separates points that it knows are far apart. For example, if points A and B are close and points A and C are far apart. Then we can safely assume that B and C are far apart as well without needing to calculate them. A K-D Tree takes advantage of information like this to reduce

the total number of computations necessary to compute the nearest neighbors.

4.4 Random Forest

Random forests are an ensemble supervised machine learning model that is essentially a forest of decision trees working together to provide a prediction. As KNN, it can be used for both classification and regression. The power in random forests is that it isn't reliant on just one prediction; it is an aggregate of many predictions. The only things to talk about in regards to random forests and not decision trees is how many decision trees to make and how to make them such that they're all unique.

In this case, I've set the number of decision trees to be constructed to 100. This seems to be a good number of trees without destroying run-time. Different decision trees are then each built with a random subset of the training data.

Also of note is that class weights have been set to balanced to provide some adjustment for the imbalanced classes. This adjusts the weight of a class to be inversely proportional to its frequency in the training data.

4.4.1 Decision Trees

As mentioned, the foundation of a random forest is its constituent decision trees. Decision trees are based on strategic splitting which are guided by information gain. A sample implementation can be found in the lone python file in the repo.

5 Results

First, results will be presented from data analysis done on these datasets. Then results from the machine learning models will be presented.

5.1 Data Analysis

Figure 1 plots the number of fires per year as well as the average fire size per year. There does seem to be some correlation between the number of fires and their average size. One possible explanation is that while many firefighters are fighting the largest fires, more smaller fires are able to occur and grow in size.

Figure 2 shows how the percent of fire size classes change as the temperature increases. Of note, at the higher temperatures there aren't very many data points, so that section may not be representative of an actual trend. Overall, it does seem as though fires go grow a bit more dangerous at temperature increases. The four largest groups (100+ acres burned) do stay fairly low throughout, with a little bit of an upwards slope at the end, although that can possibly be explained by the lack of fires at those temperatures, as noted.

Figure 3 shows the average fire size by cause of fire compared against the overall average. For both of these, the largest class G has been excluded as a few fires would otherwise dominate the averages. Lightning and powerlines are the two causes furthest above average. One possible explanation for this is that wildfires caused by either of these things can happen in more remote areas, while a fire caused by a campfire or a child is happening near civilization and under the watchful eye of at least some people.

Figure 4 plots the same thing as Figure 3 but without excluding any data. The overall average jumps from around 20 to around 75. Lightning and powerline both remain above average causes.

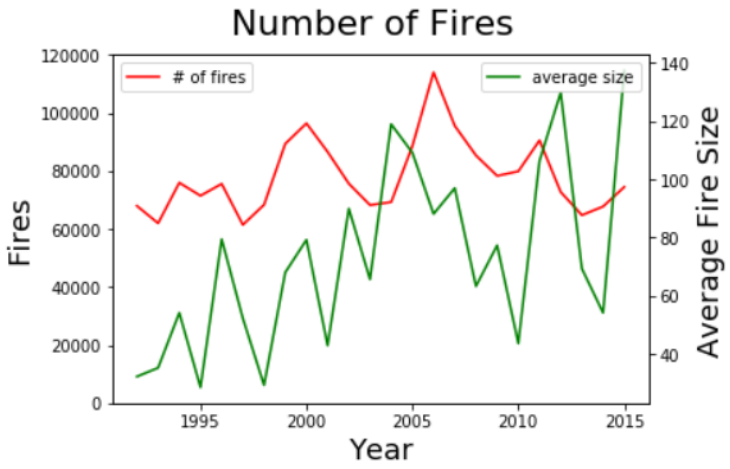

Figure 1: Number of Fires vs Average Size per Year

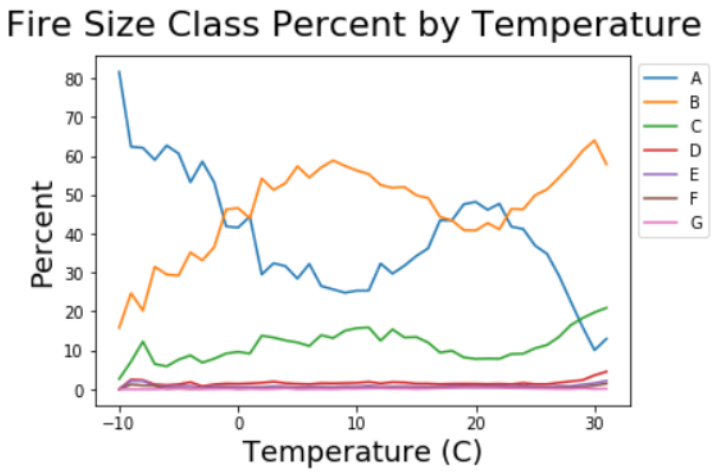

Figure 2: Fire Size Class Percent by Temperature

Cause	# of Fires
Lightning	149
Miscellaneous	11
Equipment Use	8
Arson	6
Missing/Undefined	4
Campfire	3
Debris Burning	2
Powerline	1

Table 1: Number of Fires With Over 100,000 Acres Burned

Lastly for this section, Table 1 shows the number of fires with over 100,000 acres burned by cause. Clearly

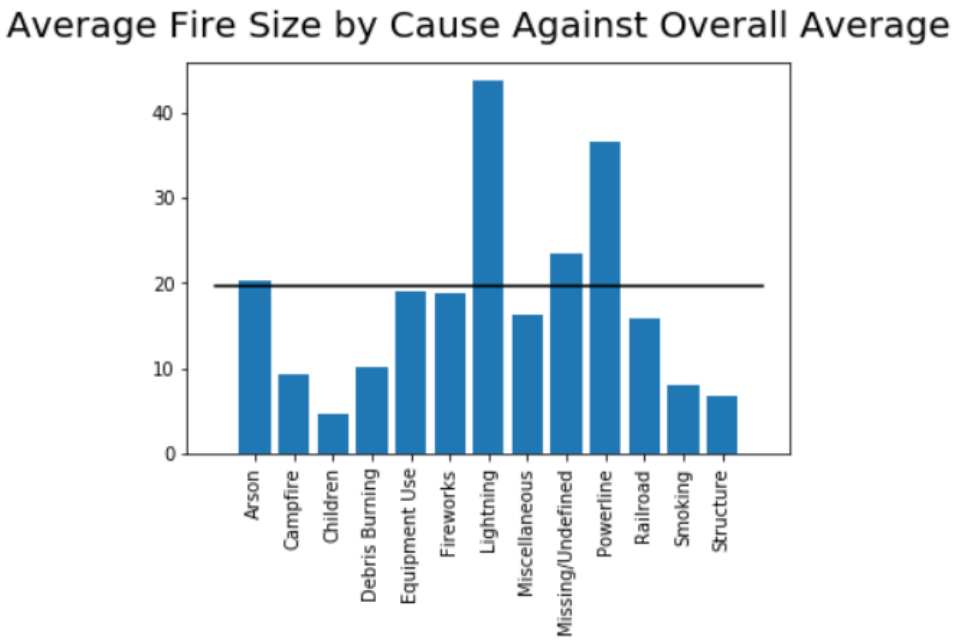

Figure 3: Average Fire Size by Cause Against Overall Average: Excluding class G

lightning is the main cause of these large fires as it is the cause of 149 of 184 such fires. That helps explain its incredible average fire size. Surprisingly powerlines is very low on this list with only one, while it was the cause with the second highest average fire size.

More plots and information can be found in the 'Fires Visualization and ML Exploration' notebook in the repo.

5.2 Machine Learning

As previously mentioned, k-nearest neighbors and random forest were ran on the data as described. The results of this can be seen in Table 2.

	All Classes		w/o Class A	
Model	Acc	BalAcc	Acc	BalAcc
KNN	58.749%	25.712%	73.732%	23.439%
RForest	61.096%	25.229%	75.726%	22.391%

Table 2: Machine Learning Results

In terms of normal accuracy, the random forest model performs slightly better, while the KNN model performs slightly better in terms of balanced accuracy. While both achieve a respectable accuracy without class A, their balanced accuracies are still fairly low.

Table 3 shows the counts of each class for the dataset before the train/test split. The models do predict a similar distribution, as can be seen in Figure 5, although it is still a little heavy on predicting B. Random forest goes a bit further in the "predicting B often" direction.

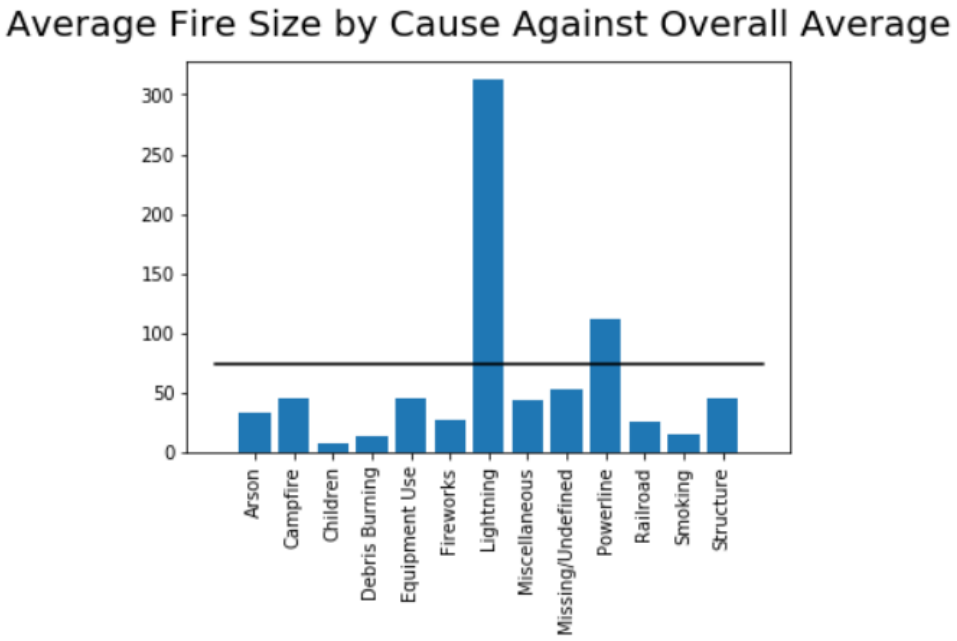

Figure 4: Average Fire Size by Cause Against Overall Average: All Data

Class	A	B	C	D	E	F	G
Count	602,234	853,958	201,911	26,268	13,102	7,220	3,377

Table 3: Fire Size Class Counts

Overall, these models show that there is some promise in attempting to predict the severity of a wildfire given some initial information. Adding in other information would likely help push these models to higher accuracies.

6 Future Work

One important area to refine what was done for this project would be to use a more exact temperature dataset and to better join it. Right now, the temperatures dealt with are average temperatures for the whole state over a month. Then that is joined to the wildfires based on the discovery month of the wildfire. Perhaps joining a temperature for the discovery date or month and then some midpoint would help for the joining problem.

Another interesting area to explore would be to find a dataset for a different part of the world, such as Brazil. It would be interesting to see if the trends there are similar or not. Similarly, adding a column that describes the terrain where the fire originated or where a majority of it took place would be another interesting avenue as a terrain like rainforests may be quite different than a normal forest.

Exploring what factors cause a fire in class E to move on to class F (or some similar progression for large fires) would be interesting as well.

Finally, expanding this dataset to include things like landscape, surrounding landscape, and weather would

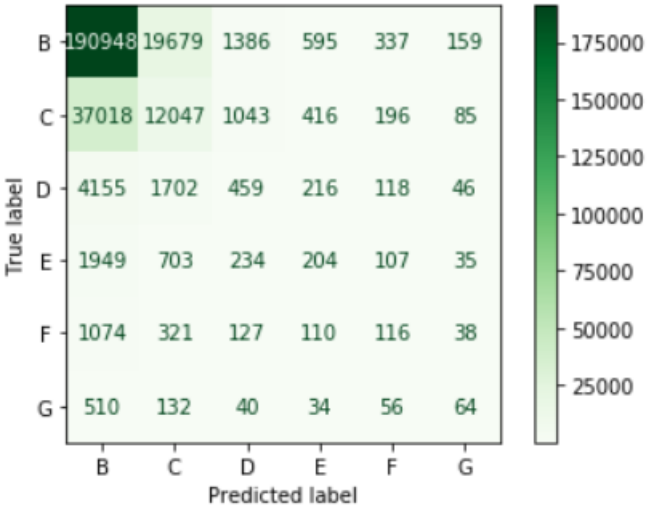

Figure 5: Confusion Matrix for KNN w/o A

be very important for a more full and accurate analysis of wildfires and their spread.

7 Conclusion

The two models presented here are a good start towards predicting the severity of wildfires. Given more information about a fire, such as the landscape and weather conditions upon its start, I believe the models will be able to achieve a higher accuracy.

The data analysis yielded some interesting insights into wildfires, such as the prevalence of lightning-caused wildfires among the largest wildfires. Other interesting correlations among the data were explored as well, but that's only the surface. There's certainly much more information to be gleaned from this data. Further digging into this dataset and finding information about wildfires worldwide would yield even more informative graphs.

8 References

Harris et al. Array programming with NumPy, Nature, 585, 357–362 (2020), DOI:10.1038/s41586-020-2649-2

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010)