Tackling Wildfires with Big Data

by Julie Bougard and Nicholas Parkes

Live site version: https://medium.com/@niparkes/tackling-wildfires-with-big-data-7a1fe42a1029

Background

For years, wildfires have destroyed millions of acres of land and homes in the United States, especially in California. While some areas are more prone to fire propagation (given land cover types, vegetation health, and land surface temperature and precipitation), 90% of fires are caused by humans, according to the U.S. Department of Interior.

This dataset contains data on wildfires in the United States ranging from 1992 to 2015 created to support the US Fire Program Analysis. This set contains almost 2 million total wildfires and corresponding metrics such as location, date, size, cause, time to extinguish.

We first used exploratory data analysis (EDA) to better understand the relationship between each metric, both across the country and only in California. We then developed models to answer the following two questions:

1. Can we predict the cause of a fire?
2. Can we predict the time it will take for a fire to be extinguished?

Exploratory Data Analysis

Wildfires over Time

The number of fires in the US has remained quite constant overall, with the trend slightly increasing (Figure 01). We can see peaks in 2000, 2006 and 2011.

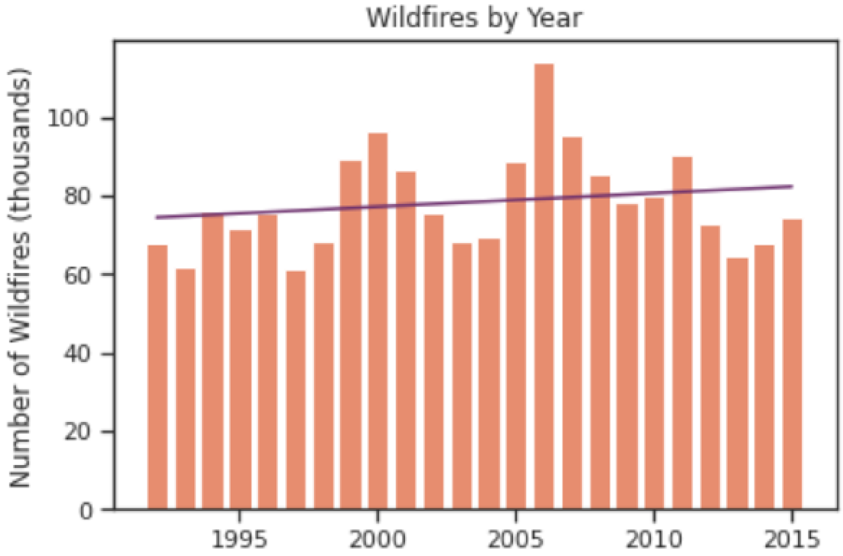

Figure 01: Number of Fires per Year 1992–2015

However, wildfires are a phenomenon affected by seasonality, as they occur most frequently in Spring and late Summer.

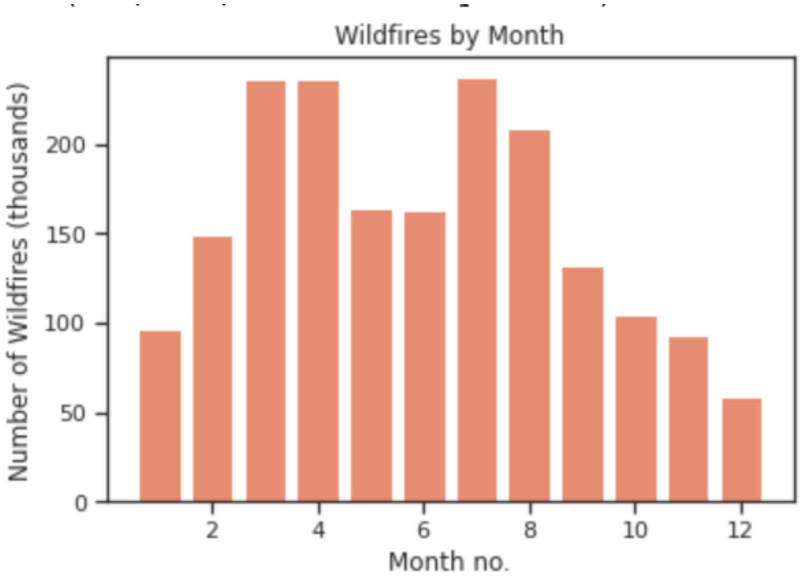

Figure 02: Total Number of Fires by Month

More interestingly, there is a massive peak in wildfire occurrence on the 4th of July, which may indicate that most of these fires are borne of holiday celebrations.

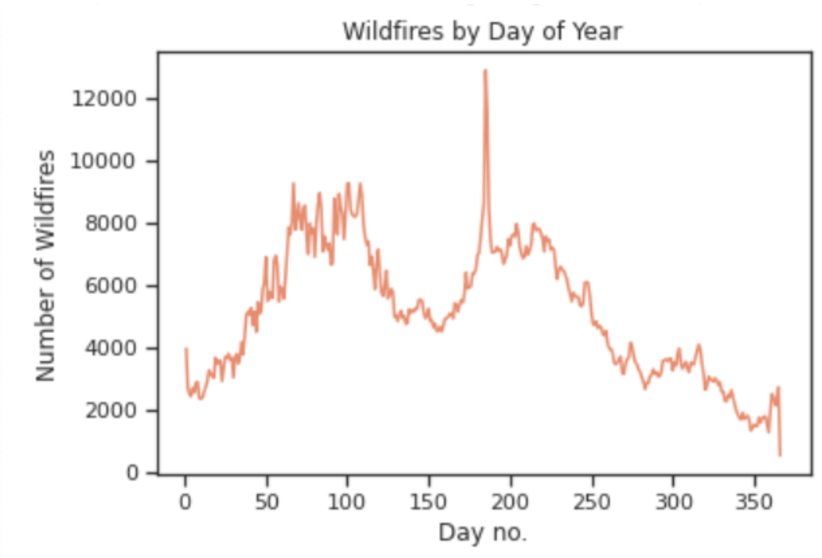

Figure 03: Total Number of Fires per Day of Year

Wildfires by State

Some states are more affected by wildfires than others, as shown in Figure 04 below. The top three states with the highest number of fires are California (189.55 thousands), Georgia (168.87 thousands) and Texas (142.02 thousands).

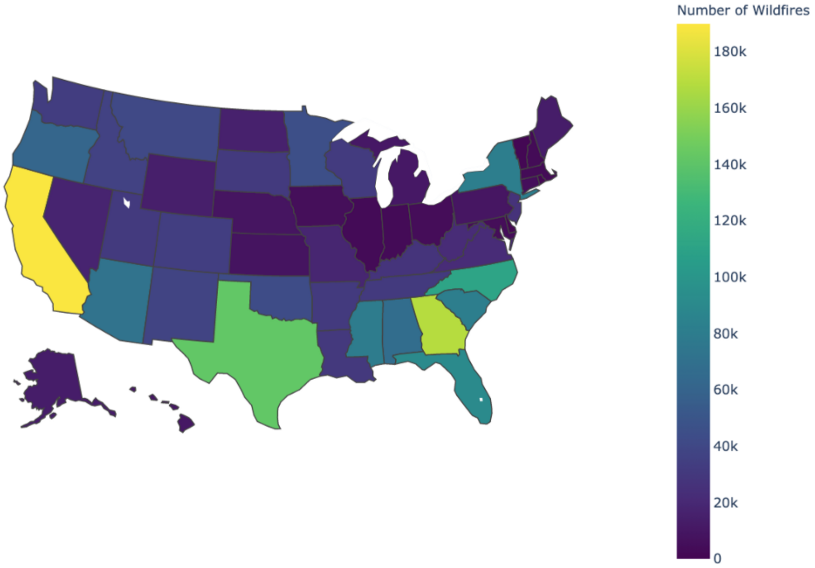

Figure 04: Total number of wildfires 1992–2015 per state

Wildfires by Cause

We can see that Debris Burning is the most common cause of wildfires countrywide.

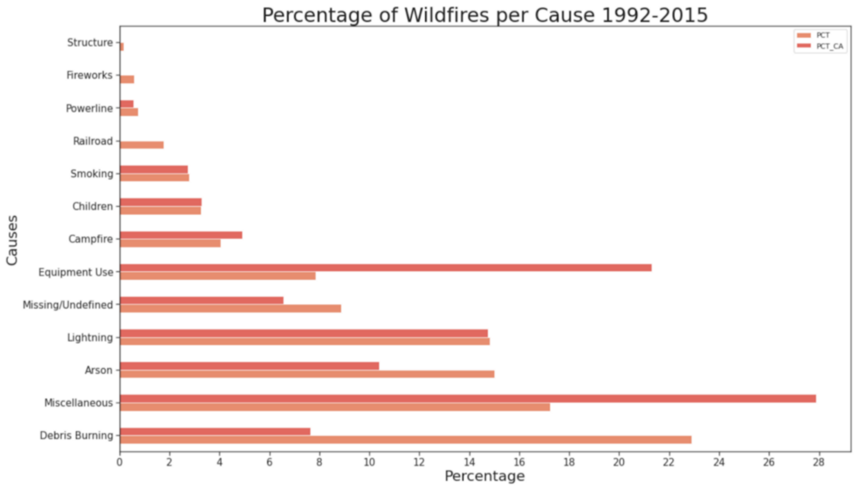

Figure 05: Percentage of Wildfires by Cause USA/California

“Lightning” is the only natural cause and only represents 15% of wildfires both across the country and in California, which means that 85% of all fires were caused by humans. This data validates the U.S. Department of Interior’s claim that 90% of fires are caused by humans. In addition, “Arson” is the third leading cause of wildfires at 15% country-wide and 10% in California.

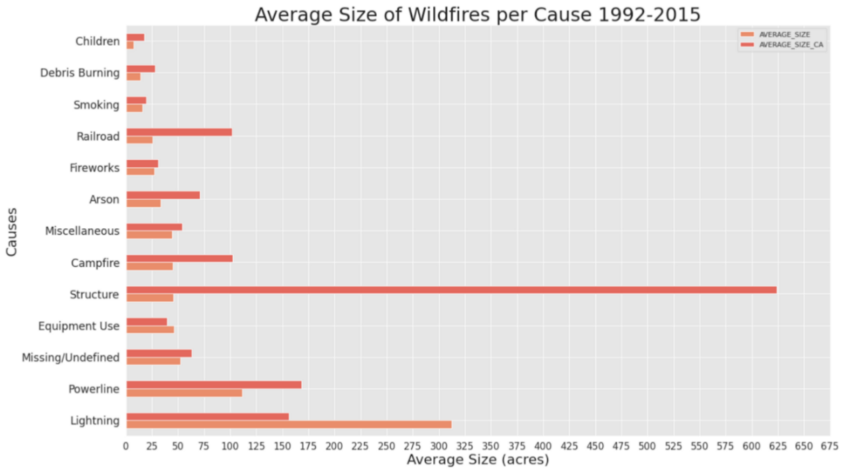

Figure 06: Average size of wildfires per cause countrywide and only in California

Around the US, fires caused by Lightning (natural cause) are the largest, with an average of 310 acres, followed by power lines at an average of 110 acres. In California however, Structure caused fires (accidents) burn 625 acres of land on average (this is most likely due to outlier data).

Length of Fires

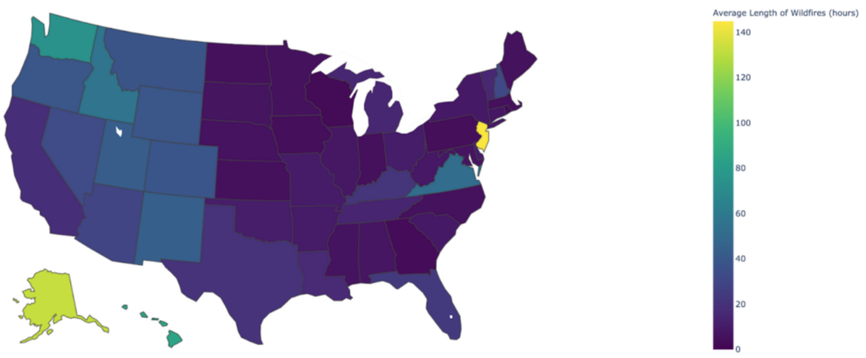

Figure 07: Average Length of Wildfires 1992–2015 per state

From this plot, we can see New Jersey wildfires last the longest at an average of 144 hours between discovery and containment time. The discrepancy between NJ and its neighboring states' burn time (PA has an average of 3.2 hours) may be due to outliers or poorly entered data.

Size of Fires

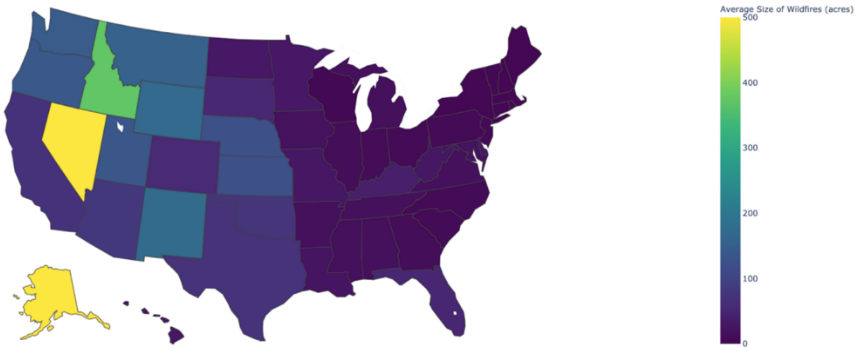

Figure 08: Average Size of Wildfires 1992–2015 per state

Alaska's wildfire size far surpasses any other states at an average of 2500 acres. That is not surprising given it is the US's largest state. However, California and Texas (also some of the largest states) only have average sizes of 68 acres (even though they have the 1st and 3rd highest counts of wildfires).

Predicting Wildfire Causes

Wildfires have thirteen causes as identified in our exploratory analysis. The largest portion of these are taken up by Debris Burning, as we can see in a quick summary of our chart from above.

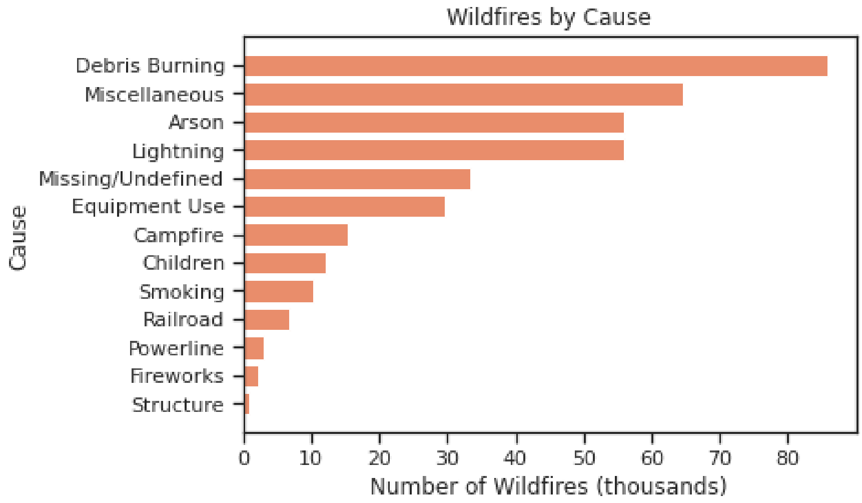

Figure 09: Wildfires by Cause (again)

For this classification task, we opted to use decision trees, since their explanatory nature lends itself well to our end goal.

Preprocessing

In order to preprocess the data, we first need to narrow our dataset down to usable columns. For this predictive task, we settled on the following list of features:

[Fire Year, Fire Month, Fire Day of Year, Fire Size, Latitude, Longitude]

We debated using state or county data, but since these would need to be one-hot-encoded and would greatly increase the size of our dataset, we opted instead to use coordinates. Their values as floats make them ideal to encode geographic information.

Model 1: Benchmarking

We can establish a baseline to make sure that our fancier models are improving our ability to make predictions.

The simplest benchmark to use will just be to guess the most common cause: Debris Burning.

Accuracy: 25.0%.

This is the accuracy our models are looking to beat.

Model 2: Simple Decision Tree

Let's use a single explanatory feature (Fire Size) to build out a very simple model.

Accuracy: 30.2%

It seems that by just using the size of a fire, we can improve our prediction by just a bit. However, our score still isn't very good.

We can look at a confusion matrix to see what predictions our model is making. A perfect model would have a rating of 1.0 on a downward-sloping diagonal line and 0.0 everywhere else.

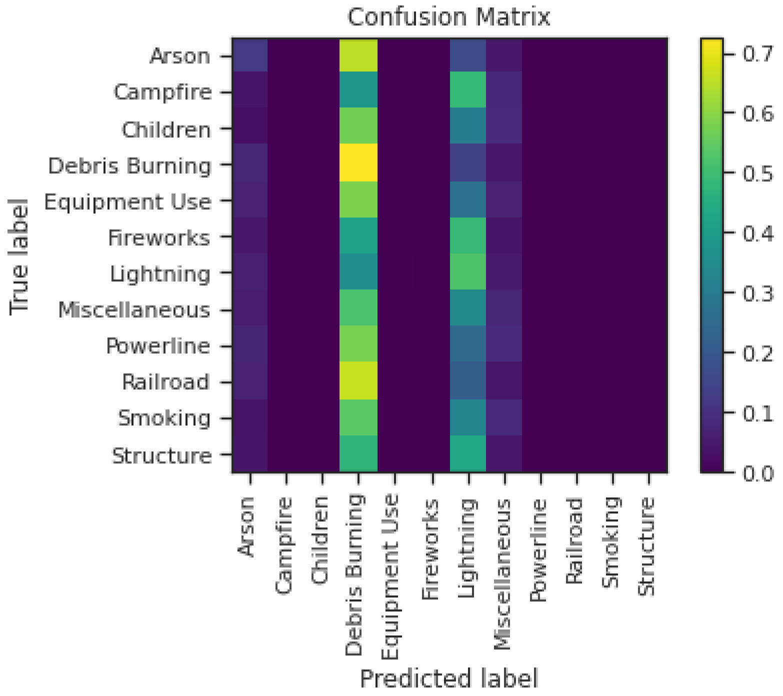

Figure 10: Simple Decision Tree Confusion Matrix

It looks like we are predicting Debris Burning and Lightning for nearly every case. Since we discovered during our EDA that these two are highly dependent on fire size, this makes sense.

Model 3: Adding More Features

Let's add in the rest of our seemingly relevant features into our tree model to see how it improves our prediction accuracy.

Accuracy: 43.6%

This is a great improvement on our previous models! Let’s take a look at our new confusion matrix.

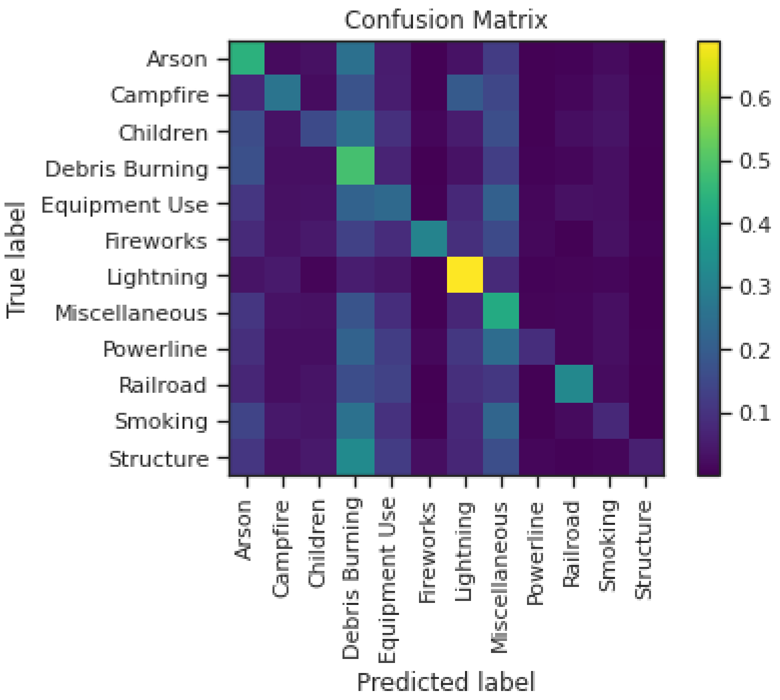

Figure 11: Advanced Decision Tree Confusion Matrix

We can see a diagonal line developing! And we are no longer simply predicting debris or lightning for every case. In particular, we are very accurately predicting lightning strikes at a rate of ~70%.

Method 5: Random Forest

Let’s now see if using a fancier method can yield a better accuracy.

Accuracy: 46.7%

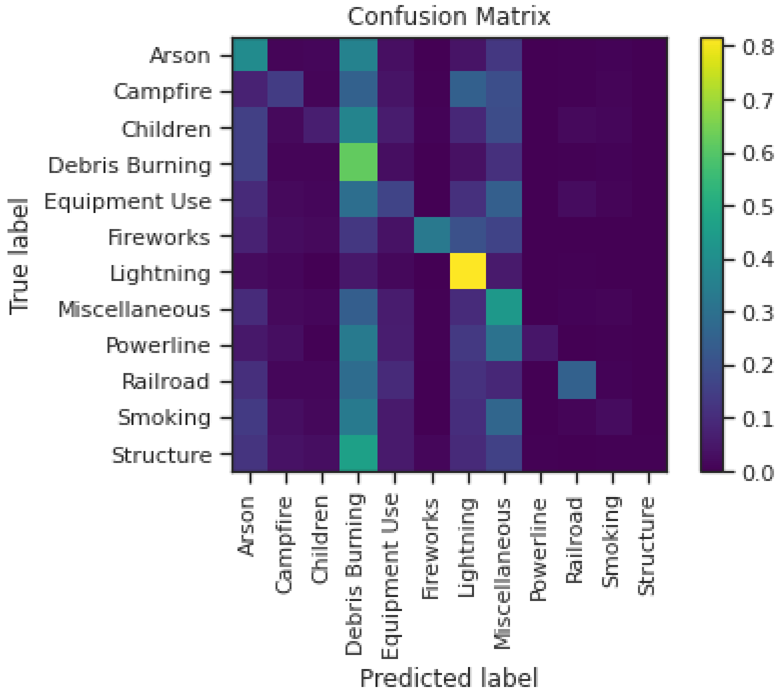

Figure 12: Random Forest Confusion Matrix

It does! Even if it is an ever so slight of an improvement from our regular decision tree model.

One great insight we can make given this model's accuracy at detecting lightning strikes is to know whether a given fire was manmade or natural. Since lightning strikes are the only prevalent natural cause of wildfires, running our model can give high-accuracy insight into whether a manmade cause needs to be investigated further.

Predicting Time to Extinguish After Detection

Let's tackle this regression problem with a linear model before moving on to some more advanced methods. But first:

Preprocessing

Given latitude, longitude, cause, type of property owner, month and day of discovery, can we predict how long it will take to extinguish (contain) the fire? We first checked for multicollinearity between our features, as shown in Figure 13.

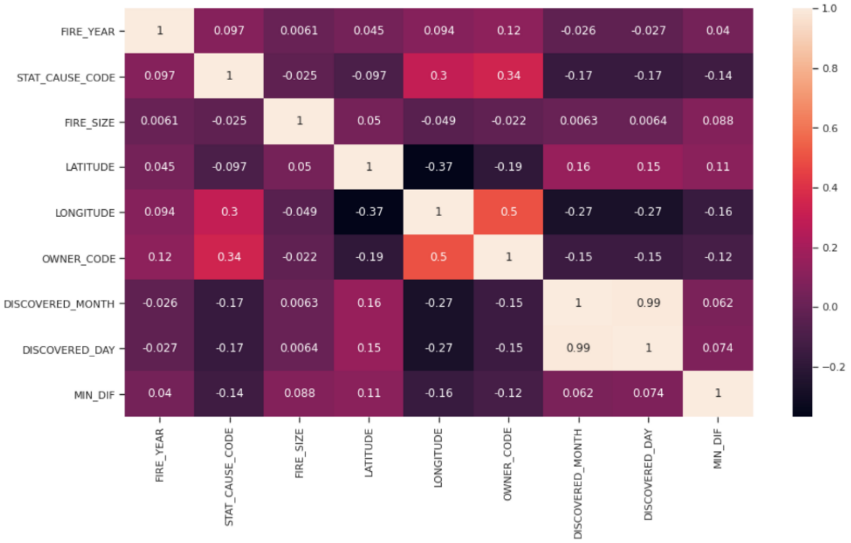

Figure 13: Correlation Matrix Heatmap of Dataset Features

We can see there is little correlation between our features (apart from discovered day and month which is expected). We therefore do not need to run a Principal Component Analysis (PCA) on our dataset to reduce the number of features.

Linear Regression

This model had an R^2 of 0.143. Pretty egregious but it is a simple model.

Elastic Net Regularization

This model actually performed more poorly than the unregularized model. A purely L1 regularization clocked out an R^2 of 0.126 and an equally weighted L1/L2 model yielded 0.055.

Random Forest Regressor

For the Random Forest Model, we ran a grid search to find the best hyperparameters to use.

This model improved performance somewhat with an R^2 of 0.214, but our regression models did not end up yielding any useful insights.

Obstacles

DateTime format

The original date format was in Julian format (For example, January 1, 2008 is represented as 2008001 and December 31, 2007 is represented as 2007365) and we had to convert it to Gregorian format.

Geographic Data

We really wanted to plot out our data on the county level, rather than the state level. And while we had running code to do so, it took an intractable amount of time to run. (Despite all that we learned in this class, our skills were not up to the task of speeding up a few million calls to the FCC FIPS Code API.) We also found great difficulties in using Geopandas or any other geography-based plotter. We finally settled on Plotly after a long haul.

Improving Accuracy

Our classification errors in answering the first question were not perfect, although they improved enough during our iterative process to yield some useful results.

Our regression models and time estimates, however, never reached a useful degree of accuracy. Despite many attempted ways of improving hyperparameters and kinds of regression models, this kept being a sticking point. It may simply be the case, as it often is in machine learning, that outcomes (in our case, fire burn times) are driven by exogenous factors that we do not have access to in our dataset.

Potential Next Steps

Network Models

We shied away from neural network models in this assignment to use tools with less of a “black box” paradigm, but we may have lost the malleability and accuracy that networks often provide.

Adapting our model to 2015–2021

Unfortunately, our dataset was rather out of date. If a new dataset is gathered, it would be interesting to expand our current models to see if they still maintain some of their explanatory power.

Real-Time Fire Tracking

One of our initial inspirations for this project was the real-time fire tracking system WIFIRE developed by the University of California San Diego.

> *"WIFIRE is an integrated system for fire analysis. Using computational techniques including signal processing, visualization, modeling and data assimilation, the web-based platform merges satellite imagery and real-time data from cameras and sensors to assemble a picture of the fire, the conditions around it, and its trajectory."—Marty Graham, Dell Technologies*

The scope of this project was very small compared to the possible space of wildfire-related data exploration there is to be done.