Sagnik Rana & Daniel Lesser

2/16/19

Data Science and Big Data 95-885

Project 1: Project Proposal

Forest fires are increasingly becoming a prevalent topic within the national dialogue. Every year, there appears to be a greater number of fires at an ever-increasing level of severity. Forest fires can have many negative impacts, including the loss of natural habitats, the endangerment of native species, the destruction of homes as well as immense economic costs.

The primary purpose of this project is to help curb forest fires through the analysis of historical trends. Recent media reports are quick to highlight the damage and the impact of forest fires, though less focus is typically given to the causes of these fires and how to reduce the risk going forward. With our research, we aim to identify any underlying phenomena regarding forest fire frequency, location, time of year, and cause. With this information, we hope to better arm policy makers so that they can employ targeted solutions in the United States.

This topic is relevant today due to the increased political discourse on wildfires and the immense costs recent events have inflicted on communities and businesses. In California, the recent Camp and Woolsey wildfires led to the largest power company in the state, Pacific Gas and Electric (PG&E), to declare bankruptcy. Meanwhile, some political leaders have enflamed tensions by blaming the U.S. Forest Service for the recent disasters. As such, it is imperative for us to better understand the increase in wildfires so that we can attack this problem with a concerted effort.

Taking a step back, one might ask how the team became aware of this issue. While researching for a project topic, it was hard to overlook the frequent news stories on forest fires. For instance, a 2018 article in the New York Times detailed how 88 people perished and 200 people went missing in the California Camp fires. Along with the recent Bankruptcy of PG&E, this presented a compelling topic for us to delve further into.

There is a plethora of prior research on this topic, including several publications from the University of California on how home landscaping and building materials impact wildfires. In addition, the U.S. Forest Service has produced research papers on the impact of Climate Change on Wildfires. Finally, there are several kernels on Kaggle where previous data enthusiasts have performed exploratory data analysis.

The core dataset chosen to analyze wildfires is stored in an SQLite database. The database consists of two tables. The first table includes 1.9 million records, with each record referring to an individual forest fire that took place between 1992 and 2015 in the United States. Of the 39 columns in the table, many are bookkeeping in nature and will not be used in the analysis. The remaining features cover several important topics, including the date, time and duration of the fire, the state, county and GPS coordinates of the fire, the cause of the fire and who the owner of the land was at the time of the fire. Overall, the data in this table is clean and will not require significant transformation. The dataset comes from Kaggle.

In addition to the data on fires occurrences, there is another table on the agencies that comprise the National Wildfire Coordinating Group. This data includes descriptive labels about each agency's location, geographic area covered, and role in fighting wildfires. Overall, there are 5,876 records and 13 columns in the dataset. However, at this time we do not feel this additional information will be essential to answering our questions enumerated later in the proposal.

Time permitting, there are three additional sources of information we intend to pursue in this project. The first comes from the National Oceanic and Atmosphere Administration (NOAA) and is a record of hourly rainfall in the United States. Climate change has been cited in other research projects, such as one presented in 2016 in the Proceedings of the National Academy of Sciences, as a potential factor in the rise of wildfires. The dataset from the NOAA is dispersed across hundreds of text files and will require meaningful effort to parse, clean and merge with the wildfire data.

The second additional piece of information of interest is the annual budget of the U.S. Forest Service. This budget can be narrowed to Wildland Fire Activities, as outlined in the annual U.S. Department of Agriculture budget. The goal of looking at this data would be to determine any relationship between the available budget and wildfire rates. Finally, we intend to use the land area of each state to normalize the number of wildfires when comparing across states.

As outlined at the outset of this proposal, the motivation for looking at this dataset is to better understand the phenomenon of wildfires in the U.S. and how we might adjust policy to prevent them. To that end, we propose the following research questions:

1. How has the frequency of wildfires trended over time? Assuming the frequency has increased, is the rate of increase accelerating? Similarly, are fires becoming more severe, as measured by size and duration?
2. What states have seen the greatest increase in wildfires? Once the size of the state is considered, are there certain hot spots that have been overlooked?
3. If there are certain states that can be designated as hot spots, can we identify a common cause of the fires? If there are significant differences in the cause by state, there may be an opportunity to employ targeted policy by location.

4. What times of year are most prone to wildfires? Has the overall increase in wildfires been due to more frequent fires during the riskiest part of the year or an expansion of the window in which forest fires take place?

Given the breadth of this topic, there is a significant opportunity for further research. Time permitting, we will consider the following additional questions:

5. Is there a strong relationship between rainfall and forest fire frequency? Has rainfall changed over time in the areas hit hardest by wildfires?
6. Is there a relationship between the U.S. Forest Service budget and forest fire frequency? Would additional funding help reduce forest fires, or is there a confounding factor?

We plan to address these questions through analysis of the core wildfire dataset in a Jupyter Notebook as well as in Tableau. Tableau will be particularly useful when charting out wildfire frequencies based on latitude and longitude coordinates. Our initial tasks will be to summarize the data and identify any underlying trends. We will then pursue our 3-4 core questions and prepare our presentation around those. Time permitting, we will seek to answer questions 5 and 6 with the additional datasets.

Coming into this project, we have several preconceived theories and assumptions about wildfires. We believe that wildfire frequency has been rising over time and at an accelerating rate. In addition, recent media attention has led us to the hypothesis that the size and duration of fires have increased. The national focus has been on California, but even a preliminary review of the data indicates there may be other hot spots in the country, such as in Georgia, Texas and North Carolina. We hypothesize that the cause of fires may differ by region, offering the opportunity to have targeted policies at the state level. Finally, we expect that the window for wildfires has expanded beyond the summer months into the spring and fall, thus increasing the risk of wildfire during more times of the year.

In summary, there is a national interest in wildfires due to the significant impact it has on human life, wildlife, the environment, property, and the economy. With the increased focus on the issue, there have been several recent research projects dedicated to the topic. We intend to use the individual records of wildfires to identify whether wildfires have been increasing, what the cause may be and if there are opportunities for targeted policy adjustments at the state and national level.