Modeling the Cause of Fires

DATA 301

Fall 2020

Edward Du, Ethan Choi, Cal Schwefler

Introduction:

Each year, wildfires ravage the Pacific Coast of the continental United States. This is relevant to all members of our group, as we all live along the Pacific Coast and have seen the destruction that these fires cause, displacing many families and destroying homes and properties on an annual basis. We chose to attempt to model some aspect of these wildfires to aid in the understanding of, and ultimately reduction of, these hazardous forces of nature. A Kaggle dataset containing data from 1.9 million wildfires from 1992 to 2015 was chosen as our dataset [1]. We hoped that with this data, we could generate a model with meaningful insight into the relationships between various recorded attributes of the fires. The following is a brief description of the notable data included for each fire listed in the data.

Dataset Contents:

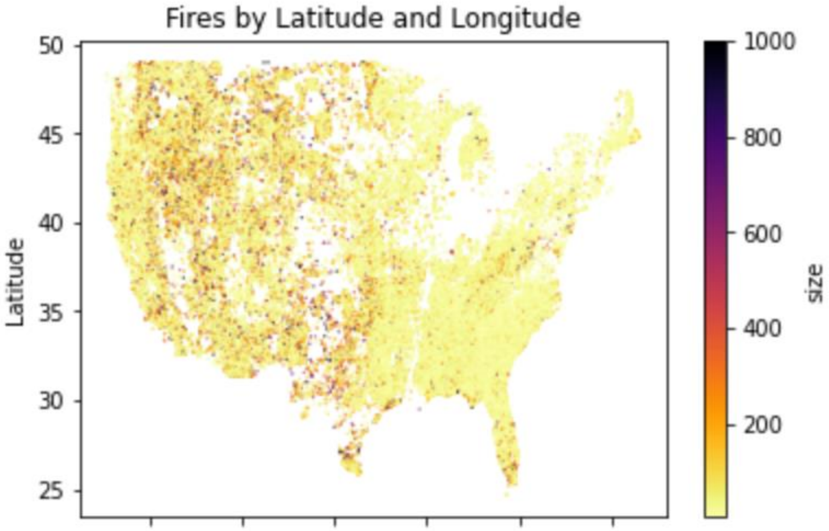


Figure 1. Fire Size across the United States.

The following is a list of descriptions for a number of categories in the dataset:

FIRE_YEAR: The calendar year the fire was discovered.

DISCOVERY_DOY: The day of the given year when the fire was discovered.

STAT_CAUSE_DESCR: The code for the cause of the fire.

FIRE_SIZE: The final size in acres of the given fire.

LATITUDE: The latitude of the given fire.

LONGITUDE: The longitude of the given fire.

STATE: The state where the given fire is located.

OWNER_CODE: Who is responsible for the land on which the fire occurs.

Given the contents of our data, we first attempted to model fire size based on the longitude and/or latitude of the given fire. After this we moved to predicting the cause of the fire given a multitude of fire attributes. The following sections will discuss the results from both efforts.

Data Understanding:

Before attempting to create a model for the fire data, we first visualized our data to note some simple trends, see Figures 2-4 for a sample of the visualization completed.

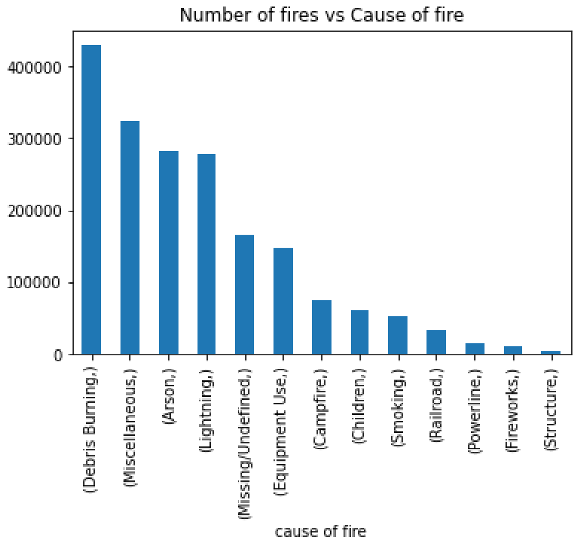


Figure 2. The frequency of fires for each fire cause.

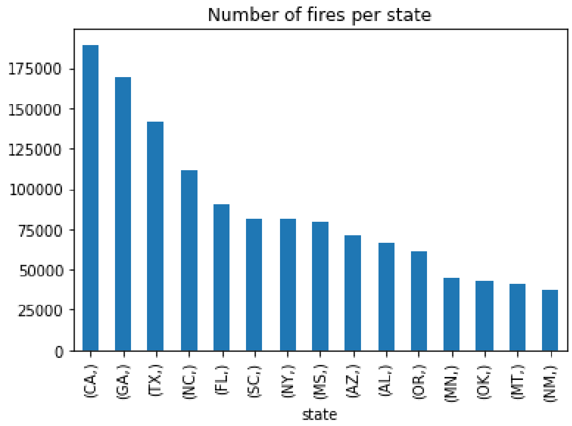


Figure 3. The frequency of fires for each state (most frequent 15 are displayed).

It can be noted that there are fires occurring in every state and there are several dominant causes for fires from Figures 2 and 3. From Figure 4, we can note that there are many more small fires than large fires that occur in the United States. This heavy concentration of smaller fires in the data presented challenges as we began training our models because these larger fires acted as outliers to the data (the median fire size is about 15 acres, however, the mean is about 60 acres).

First Attempts:

Our first trials included attempting to predict the size of fires from latitude and longitude using a knn (k-nearest neighbors) model. However, we were not able to move forward with these methods as they never produced meaningful results. Chiefly, the data in the set contained an irregularly large number of very small fires as mentioned earlier. Figure 4 demonstrates the clearly skewed distribution of fire sizes throughout the data set, as an incredibly large percentage of the fires appear to be between 0 and 100 acres in size.

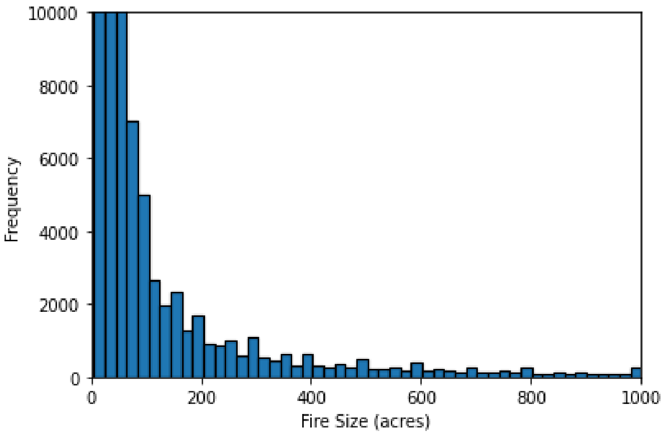


Figure 4. Frequency of fires by fire size, note the large number of fires below 100 acres.

The predicted fire size results from the knn model plotted against the actual fire size can be seen in Figure 5 (if the model was working correctly, we would expect a linear trend from bottom left to top right). We note that the larger fires are consistently predicted to be smaller fires. This is due to the heavy skew presented by the smaller fires (there are simply many more small fires than large fires).

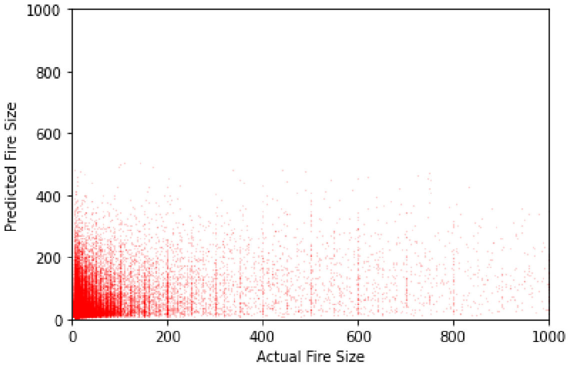


Figure 5. Predicted fire size plotted against actual fire size using a knn clustering model.

As a group, we determined to approach the modeling of fires using a different method than the simple knn method. We had determined that due to the skewed nature of our data, the typical clustering methods were not sufficient and instead chose to use the Random Forest algorithm.

Finalized Method:

The final method chosen for analysis was the Random Forest algorithm, which aided in the development of a classification model that attempts to predict the cause of a fire as a function of the fire's other features [2]. This algorithm utilizes decision trees, as illustrated in Figure 6. By using a large committee of decision trees, the model can make predictions with minimal errors from outliers in the data. This point is key to why we chose to use the Random Forest model. In our first attempt we struggled to effectively model the data because of the much less frequent large fires, which functioned as outliers to the data composed of predominantly small fires.

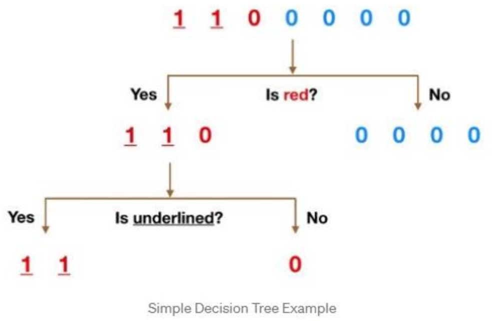


Figure 6. Depiction of decision tree used in Random Forest algorithm.

With our new model, we chose to focus on a different attribute of the fire data. We began modeling based on the fire cause instead of the fire size. The fire causes, charted in Figure 7, were combined into 5 general categories: unknown cause, unintentional cause, preventable cause, natural cause, and purposeful cause, see Figure 8. We tested using multiple variations of data and found the highest accuracy for our model when using input data that included fire year, fire discovery date and time, the containment date, fire size, latitude, longitude, and owner code. This yielded more promising results and was less prone to the error introduced by the larger fires.

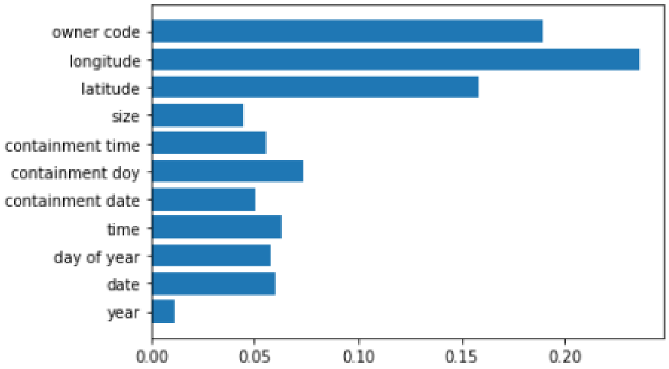
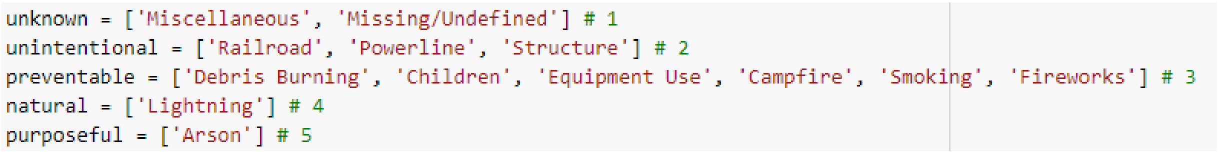


Figure 7. Each of the fire causes charted for their relative frequency.



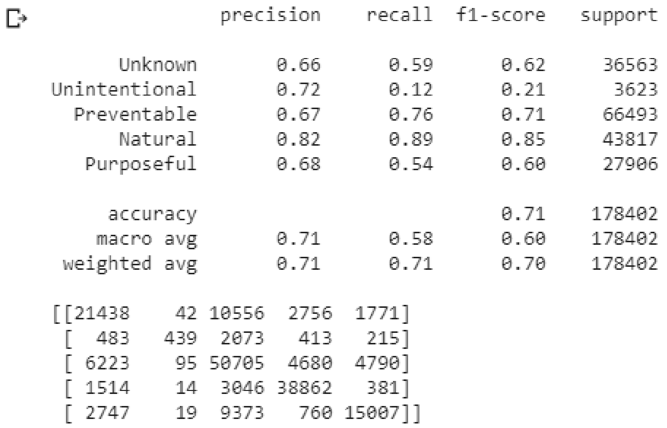
```
unknown = ['Miscellaneous', 'Missing/Undefined'] # 1
unintentional = ['Railroad', 'Powerline', 'Structure'] # 2
preventable = ['Debris Burning', 'Children', 'Equipment Use', 'Campfire', 'Smoking', 'Fireworks'] # 3
natural = ['Lightning'] # 4
purposeful = ['Arson'] # 5
```

Figure 8. The 5 categories used for our modeling and the fire causes that fit into them.

With our chosen model and categories of analysis, we trained our model and ran it against validation data to develop final conclusions about the data.

Results:

Our results from using the random forest classifier were gathered from using 100 estimators. It was found that doubling the estimators only yielded a 0.2% increase in mean accuracy. Using our model on the testing data gave an accuracy of 70.9% (higher than the accuracy for the training data). An accuracy output generated within the code is depicted in Figure 9.



```
                precision    recall  f1-score   support

       Unknown       0.66      0.59      0.62     36563
 Unintentional       0.72      0.12      0.21      3623
   Preventable       0.67      0.76      0.71     66493
       Natural       0.82      0.89      0.85     43817
    Purposeful       0.68      0.54      0.60     27906

      accuracy                           0.71    178402
     macro avg       0.71      0.58      0.60    178402
  weighted avg       0.71      0.71      0.70    178402

[[21438    42 10556  2756  1771]
 [  483   439  2073   413   215]
 [ 6223    95 50705  4680  4790]
 [ 1514    14  3046 38862   381]
 [ 2747    19  9373   760 15007]]
```

Figure 9. Accuracy output for our model predicting to which category a fire belongs.

The results from our validation data in Figure 9 are extremely similar to the testing results, as to be expected. We found that most of the labels that were incorrectly classified to the “Preventable” group we defined, as seen in Figure 10.

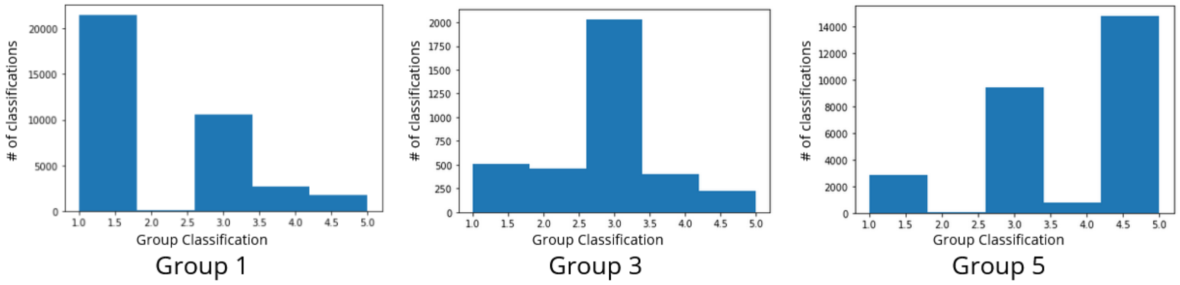


Figure 10. Group 1, 3, and 5 (unknown, preventable, and purposeful) with frequencies at which fires are classified to each group.

Conclusion:

Ultimately, we have demonstrated that our model can predict the cause of a fire with decent accuracy based on other fire attributes. As a group, we have learned much about data organization and the Machine Learning process throughout this project. A few lessons we have learned is that organizing data can be challenging (especially if the data is not naturally normal), that using different models within Machine Learning can greatly affect model results, and the importance of understanding data before trying to apply a model to it.

References:

[1] R. Tatman, “1.88 Million US Wildfires,” *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/rtatman/188-million-us-wildfires. [Accessed: 06-Oct-2020].

[2] “Random Forest - Overview, Modeling Predictions, Advantages,” *Corporate Finance Institute*, 28-Apr-2020. [Online]. Available: https://corporatefinanceinstitute.com/resources/knowledge/other/random-forest/. [Accessed: 15-Nov-2020].

[3] Yiu, Tony. “Understanding Random Forest.” *Medium*, Towards Data Science, 14 Aug. 2019. [Online]. Available: towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed: 15-Nov-2020]