Learning the Cause of Wildfires

By Carter Portwood and Christine Miller

STA 208 Final Project, spring 2019

# Introduction

Wildfires are an increasingly common threat to human lives, infrastructure, and wildlife. While occasional wildfires are often a natural part of an ecosystem, the majority of wildfires in the US are human-caused. Determining how fires start is important in suggesting future preventive measures and in bringing those responsible to justice. Currently firefighters, law enforcement, and land managers use a variety of clues to ascertain the cause of a fire. These include burn patterns, location, and local weather.

Our question is whether machine learning can be used to predict the cause of a fire based on the characteristics of a fire. In particular, can we create a model that will accurately predict whether a fire was caused by arson or not? Characteristics available for prediction include the duration, location, size, land owner, and population density. This project aims to create an additional tool to help prevent wildfires.

# Methods

### Data

We will be using data on 1.88 million US wildfires from Kaggle: <https://www.kaggle.com/rtatman/188-million-us-wildfires>

This dataset contains information on the location, timing, duration and final size of the fire, along with identifying information about each fire and the source of the information. A full description of each of the variables included in the dataset can be found at the Kaggle link.

In addition to the wildfire dataset we also used information on locations of urban areas from the 2010 census from DATA.gov: <https://catalog.data.gov/dataset/tiger-line-shapefile-2017-2010-nation-u-s-2010-census-urban-area-national>

Fire departments often use information about human presence and influence on the fire's location to suggest the cause of a fire, and we noticed that this information was missing from the data set. The Urban Areas data contains geographic polygons that define urban areas with two categories of population density: urbanized areas (UAs) that contain 50,000 or more people and urban clusters (UCs) that contain at least 2,500 people, but fewer than 50,000 people. We used the latitude and longitude coordinates in the fire data set to extract the population density for each fire (Code: ExtractUrban.R).

Because we were limited by computing time we first randomly selected 10% of the entire dataset (over 171,000 observations) to perform all exploration and analysis. We have not submitted our data because it is well over 10MB.

### Exploratory Data Analysis

Using the subset of the data we explored patterns in when and where wildfires have occurred and how those variables relate to the cause of the fire.

### Multiclass Cause Classifier

To try to predict the cause of a fire based on the available characteristic we fit and tuned KNN, Random Forest, and Gradient Boosting classifiers.

The KNN model was tuned over distance metrics (Manhattan, Euclidean) and the number of neighbors, K (3, 5). The random forest was tuned over the number of max features (‘auto’, ‘sqrt’), max depth (3, 5, None), and the minimum samples needed to split a leaf (2, 10) all with 20 trees in the forest. The best tuning parameters were selected by three-fold cross validation on the training data.

Gradient Boosting was not tuned due to computation time, but fit using the default parameters.

These models were then compared based on overall accuracy on the test data.

### Arson Classifier

Fires caused intentionally through arson require additional investigation and legal action. Thus it is of special interest to predict which fires are caused by arson.

This classification will be evaluated based on a weighted f score with beta = 2. This metric weights the recall of the prediction higher than the precision. We chose this metric because it is preferable to investigate a few extra fires that do not end up being caused by arson than to miss fires that are due to arson.

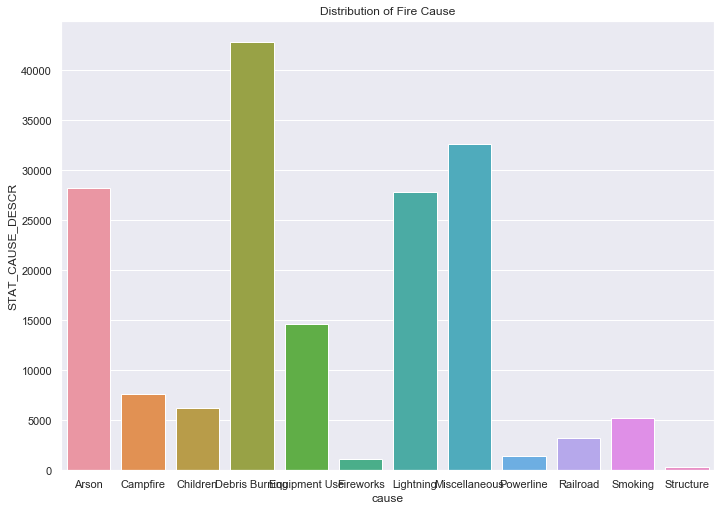
Because of the size of the data set and the time limitation of this project we chose to fit a KNN classifier and a random forest classifier to see if we could learn those fires that were caused by arson. To provide a point of reference for all of the other arson prediction models, we will set the benchmark with a model that predicts every case with the most common fire cause (not arson).

The KNN model was tuned over a range of K’s (1 to 10). The random forest model was tuned over the number of trees (ranging from 200 to 1000) and the minimum leaf samples size from (10 to 40). The upper limit was placed on the number of trees to limit computational time, and the leaf sample size was limited to 10 to avoid overfitting of the trees. Prediction accuracy of each fit was evaluated with f2 score and 5 fold cross validation on the training set.

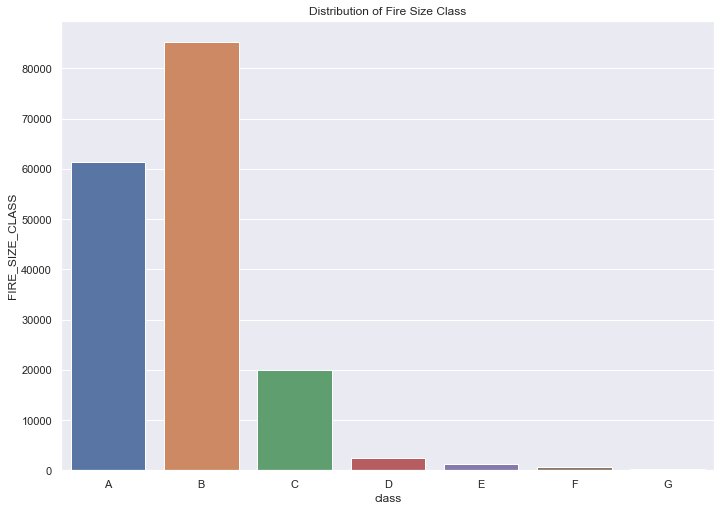
# Results

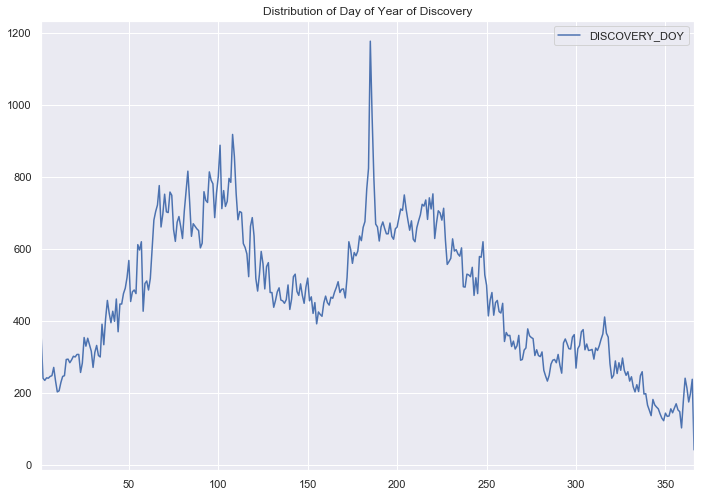
## Exploratory Data Analysis

We first investigated the distribution of the causes of wildfires. Debris Burning and Arson are the most common causes

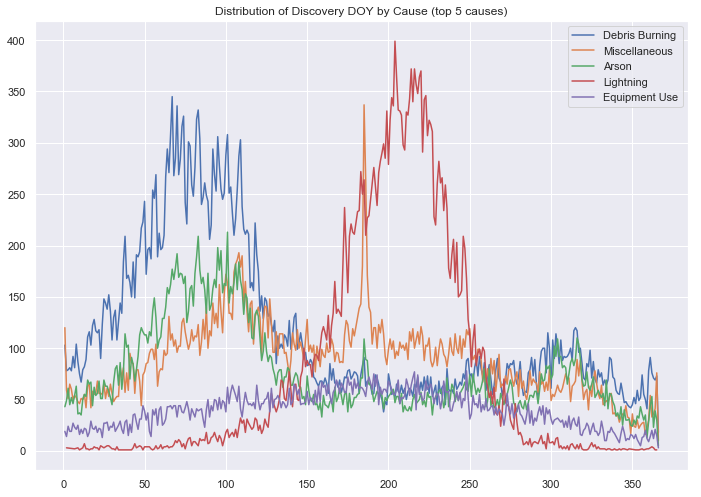


We looked at how the size of fires is distributed. As seen in the figure below most fires are quite small, but there are a few very large ones. Size goes A (small) to G (large). The size classifications are nonlinear (see [here](https://www.nwcg.gov/term/glossary/size-class-of-fire) a description of size class definitions).

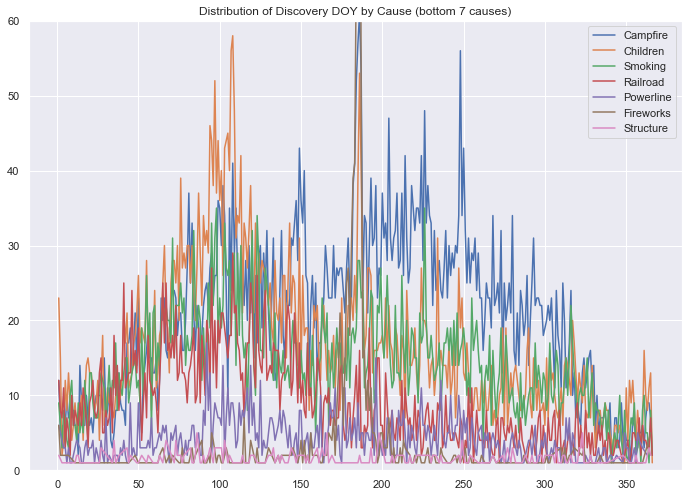


We then looked at what time of year fires happen, as seen in the graphs below. The spike between 150-200 is 4th of July. Interestingly, the data is bimodal, with a peak in the spring and a peak in the summer. 

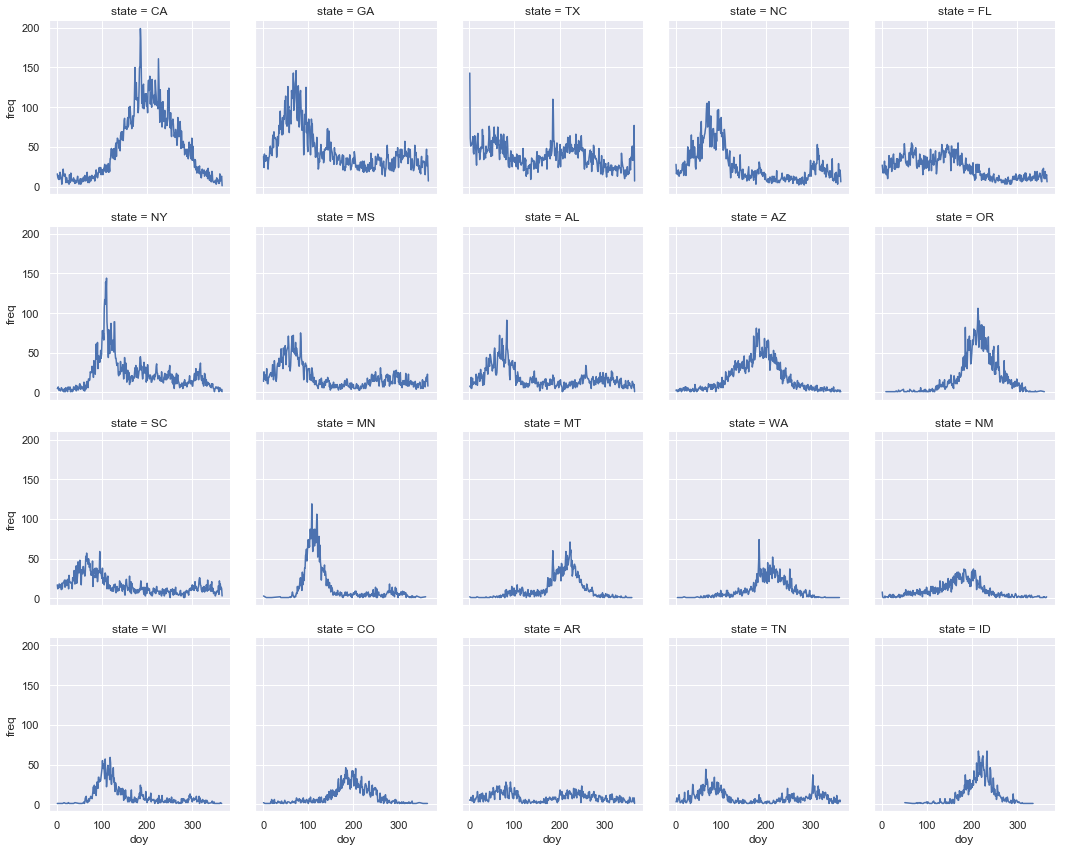
In the next graph of fire frequency by day of year and cause, it is clear that the spring-time peak corresponds to human-caused fires, while the summer peak is largely from lightning-caused fires.



The next graph shows the less common causes distributed by day of year.

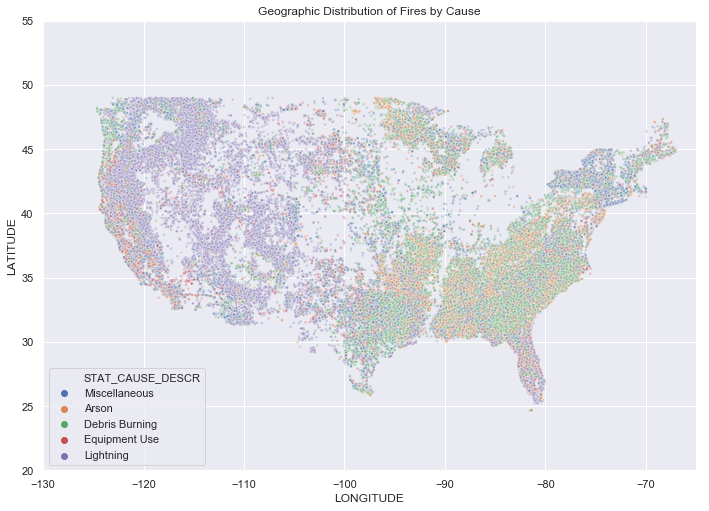


Now let's see what the distribution by day of year looks like for the top 20 states.



As we can see, the East-Coast states tend to have most frequent fires in the spring, while the western states tend to have summer fire seasons. Perhaps this is due to summers being drier than winter and spring in the West (while the East tends to have more consistent precipitation year-round), and a higher proportion of fires being lightning-caused in the more sparsely populated West.

Finally we investigated where fires occur. We can see that there are some distinctive East vs. West Coast trends in the cause of fires. Lighting seems to be a major cause in the West (excluding the coast) but not in the East. Arson and debris burning appear to standout in the East. Generally, state borders and other political boundaries are not visible, which is a good indicator that the data source is not biased by differences in reporting practices. The one exception is New York State, which stands out next to its neighbors Pennsylvania and Vermont and appears to have an inordinate proportion of miscellaneous fires. This indicates that the state may have a different fire reporting or classification system.

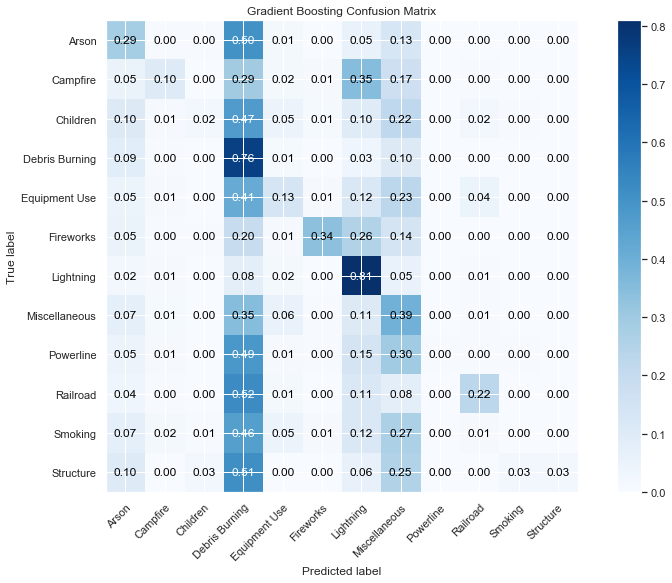
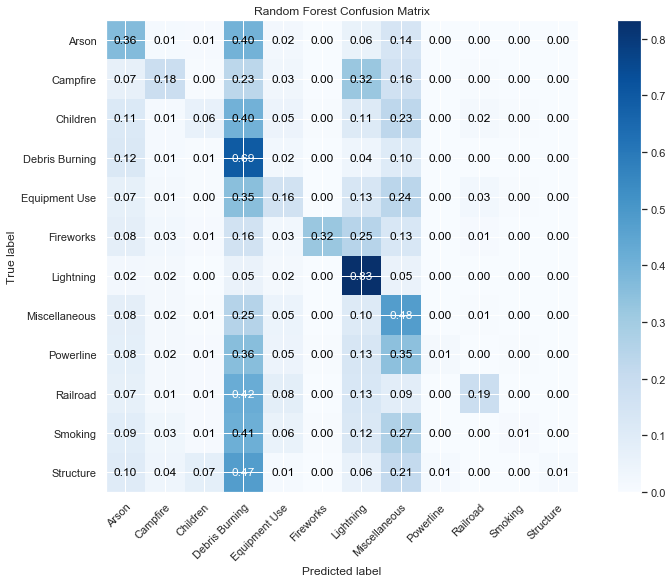


## Multiclass Classification

The random forest classifier performed best with an accuracy of 48.9%. The un-tuned gradient boosting model had an accuracy of 46.5%, and the KNN model had an accuracy of 45.4%.

Tuning and the amount of data used were limited by computation time and power (the above is only using 10% of the data). More exhaustive parameter cross-validation searches would've been possible with greater computation power.

The models did not fit the data very well, as we can see with the confusion matrices below. Classification with 12 classes is a difficult task, and the lower-frequency causes likely serve only to degrade the performance of the model. Below are the confusion matrices from the 3 models. One can see that the most common causes clearly dominate the predictions.

As we saw in the above exploratory data analysis, there are clear geographic and temporal differences in lightning-caused wildfires and human-caused wildfires. This is echoed in the confusion matrices of the models – lightning was not the predicted cause for many of the non-lightning causes despite its frequency. The exception to this is the campfire-caused wildfires, which each model mistook for lightning-caused fires in a plurality of cases. A possible explanation for this is that, as shown in a prior graph, campfire-caused wildfires tend to happen in the summer, when human-caused wildfires are far less frequent than lightning-caused wildfires. An interesting direction to expand on this analysis would be to analyze the relative importance of variables in these models. 

## 

## Arson Classifier

Fires caused intentionally through arson require additional investigation and legal action. Thus it is of special interest to predict which fires are caused by arson.

The F2 score of the benchmark model is zero because the model never predicts that the cause of the fire was arson. The benchmark prediction accuracy was .84 on the test set because the class imbalance in the data.

The tuning for KNN selected k = 1. The KNN prediction accuracy on the test set is 0.79 and the f2 score is 0.39. The accuracy of this fit is decreased compared to the bench mark set, but it has a much higher recall, indicating that it is misclassifying a fair number of fires as caused by arson that are not.

The random forest with 200 trees and a minimum leaf size of 10 was selected. The final random forest arson prediction accuracy is 0.85 and the f2 score is 0.18, which improved slightly on the benchmark model.

The KNN and random forests had different advantages when predicting wildfires caused by arson. The random forest improved the prediction accuracy compared to the benchmark overall, but had a lower f2 score. This indicates that it was missing many of the cases that were caused by arson. In comparison the KNN model, decreased the prediction accuracy compared to the baseline but had a higher f2 score, indicating that it was wrongly assigning many fires to arson that were not caused by arson. If the goal is to identify fires that warrant further investigation for arson the KNN model is preferable.

Overall, neither of the models was particularly successful in separating those fires caused by arson.

# Conclusions

Given all the interesting patterns the wildfire data set contained it was surprising that the tested classifiers where not able to learn the causes of the fires. The class imbalances are a major barrier to both the multiclass and the arson classification.

It is possible that the fits could improve with the addition of more of the data set, or with more computationally intensive methods such as SVMs or neural networks, that might be able to capture non-linear aspects of the classification more accurately.

Future directions could include fitting a human-caused vs. non-human-caused model, which may be more promising based on the given predictors. Additional weather information for all the fires could also be a valuable predictor.