# Applied Competitive Lab in Data Science (67818) Project Report

Nadav Alon, Shir Uziel, Noy Porat, Gilad Ticher and Noa Shavit

April 2024


"how not to save a catboost"

# Contents

# Introduction

Wildfires pose a significant threat to ecosystems, human life, and property worldwide. Understanding the factors that contribute to wildfire occurrences is crucial for effective mitigation and prevention strategies. In this project, we aim to analyze a comprehensive database of wildfires that have occurred in the United States between 1992 and 2015. The primary objective is to develop a predictive model that is capable of determining the cause of wildfires based on numerous factors such as time, environment, and geographic factors.

## Problem Definition

In this project, our main objective is to predict the cause of wildfires. The prediction has implications for resource allocation, policy formulation and emergency response planning. By accurately identifying the predominant causes of wildfires, stakeholders can take targeted actions to mitigate the risk of ignition and minimize the impact on affected communities. This report describes the methodology employed in the analysis process of the data, including data preparation, exploratory data analysis, feature engineering, model selection, and evaluation. Through the use of machine learning techniques and statistical analysis, we seek to gain insight into the underlying patterns and drivers of wildfires in the United States.

# Exploratory Data Analysis

## Data Description

The dataset used in this project includes records of wildfires reported in the United States between 1992 and 2015. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations. The dataset encompasses various attributes that describe each wildfire event, including geographic coordinates, date of occurrence, cause of ignition, and reporting systems of federal, state, and local fire organizations.

## Dataset Overview

The dataset consists of 571425 records. The columns include the following attributes:

- OBJECTID FOD_ID = Global unique identifier.
- FPA_ID = Unique identifier to track back to the original record in the source dataset.
- SOURCE_SYSTEM_TYPE = Type of source database or system that the record was drawn from
- SOURCE_SYSTEM = Name of or other identifier for source database or system.

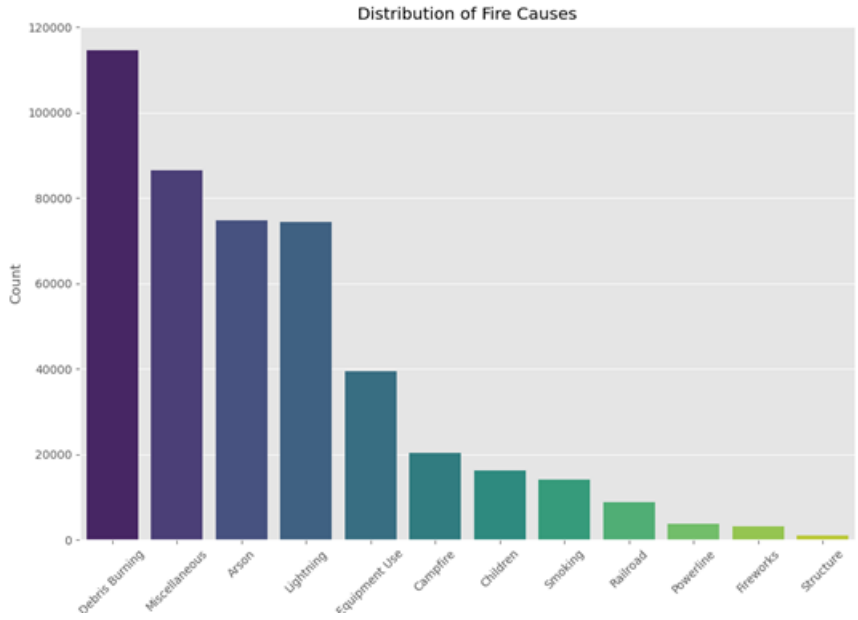
- NWCG_REPORTING_AGENCY = Active National Wildlife Coordinating Group (NWCG) Unit Identifier for the agency preparing the fire report.
- NWCG_REPORTING_UNIT_ID = NWCG Unit Identifier for the unit preparing the fire report. NWCG_REPORTING_UNIT_NAME = NWCG Unit Name for the unit preparing the fire report. SOURCE_REPORTING_UNIT = Code for the agency unit preparing the fire report, based on code/name in the source dataset.
- SOURCE_REPORTING_UNIT_NAME = Name of reporting agency unit preparing the fire report, based on code/name in the source dataset.
- LOCAL_FIRE_REPORT_ID = Number or code that uniquely identifies an incident report for a particular reporting unit and a particular calendar year.
- LOCAL_INCIDENT_ID = Number or code that uniquely identifies an incident for a particular local fire management organization within a particular calendar year.
- FIRE_CODE = Code used within the interagency wildland fire community to track and compile cost information for emergency fire suppression.
- FIRE_NAME = Name of the incident, from the fire report or ICS-209 report.
- ICS_209_INCIDENT_NUMBER = Incident (event) identifier, from the ICS-209 report.
- ICS_209_NAME = Name of the incident, from the ICS-209 report. MTBS_ID = Incident identifier, from the MTBS perimeter dataset.
- MTBS_FIRE_NAME = Name of the incident, from the MTBS perimeter dataset.
- COMPLEX_NAME = Name of the complex under which the fire was ultimately managed.
- FIRE_YEAR = Calendar year in which the fire was discovered or confirmed to exist.
- DISCOVERY_DATE = Date on which the fire was discovered or confirmed to exist.
- DISCOVERY_DOY = Day of the year on which the fire was discovered or confirmed to exist.
- DISCOVERY_TIME = Time of day that the fire was discovered or confirmed to exist.
- STAT_CAUSE_CODE = Code for the (statistical) cause of the fire.
- STAT_CAUSE_DESCR = Description of the (statistical) cause of the fire.
- CONT_DATE = Date on which the fire was declared contained or otherwise controlled.
- CONT_DOY = Day of the year on which the fire was declared contained or otherwise controlled.

- CONT_TIME = Time of day that the fire was declared contained or otherwise controlled.
- FIRE_SIZE = Estimate of acres within the final perimeter of the fire.
- FIRE_SIZE_CLASS = Code for fire size based on the number of acres within the final fire.
- LATITUDE = Latitude (NAD83) for the point location of the fire (decimal degrees).
- LONGITUDE = Longitude (NAD83) for the point location of the fire (decimal degrees).
- OWNER_CODE = Code for primary owner or entity responsible for managing the land.
- OWNER_DESCR = Name of primary owner or entity responsible for managing the land.
- STATE = Two-letter alphabetic code for the state in which the fire burned (or originated).
- COUNTY = County, or equivalent, in which the fire burned (or originated).
- FIPS_CODE = Three-digit code for representation of counties and equivalent entities.
- FIPS_NAME = County name from the FIPS publication. Shape = Geometry field for geospatial applications.

## Fire Causes and Patterns Recognition

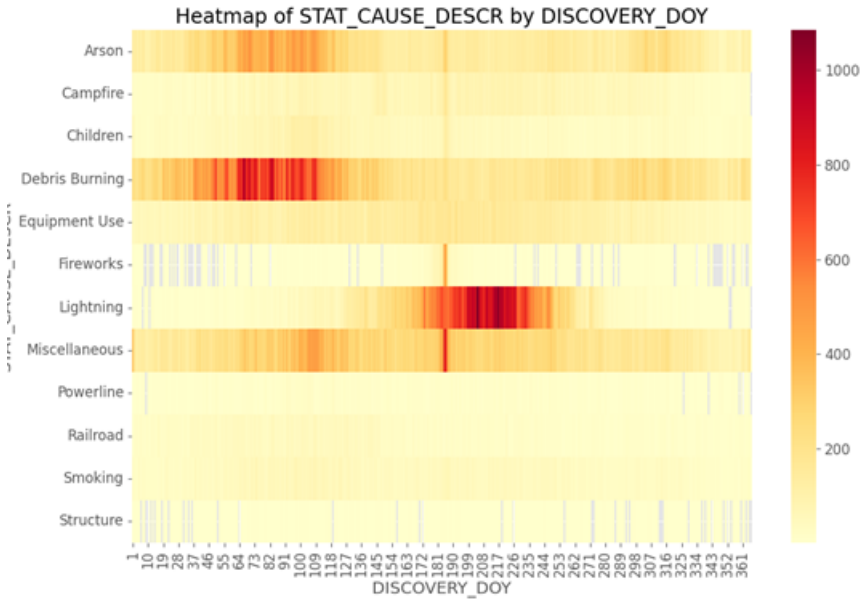
Cause of the fire, under the feature "STAT_CAUSE_DESCR": The cause of the wildfire serves as the target variable for the prediction. It includes several categories representing different ignition sources, including natural phenomena and human activities. The categories are:

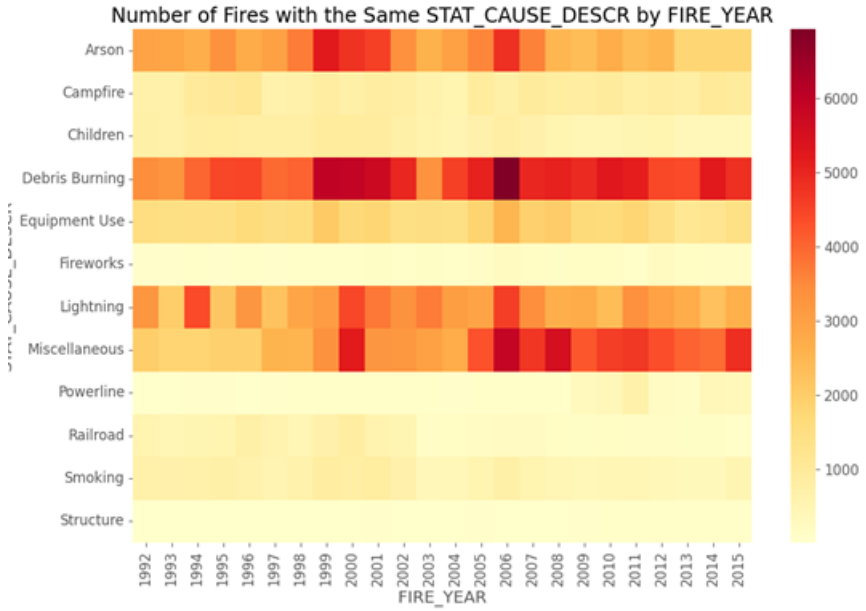
```
['Lightning', 'Miscellaneous', 'Arson', 'Debris Burning', 'Equipment Use',
'Campfire', 'Powerline', 'Children', 'Fireworks', 'Railroad', 'Smoking', 'Structure']
```




The data set exhibits class imbalance, with certain causes of fire occurring more frequently than others. The plot above highlights the imbalance, by showing the distribution of cases between different cause categories, and emphasizes the importance of considering class imbalance during model training and evaluation, which will be done later in the project. For campfires, there appears to be a relatively clear distribution of fire sizes, ranging from A to E and in some cases from C to B. For all designations, sizes D-F appear to be relatively widely scattered across the United States.

## Fire Causes by Date and Time



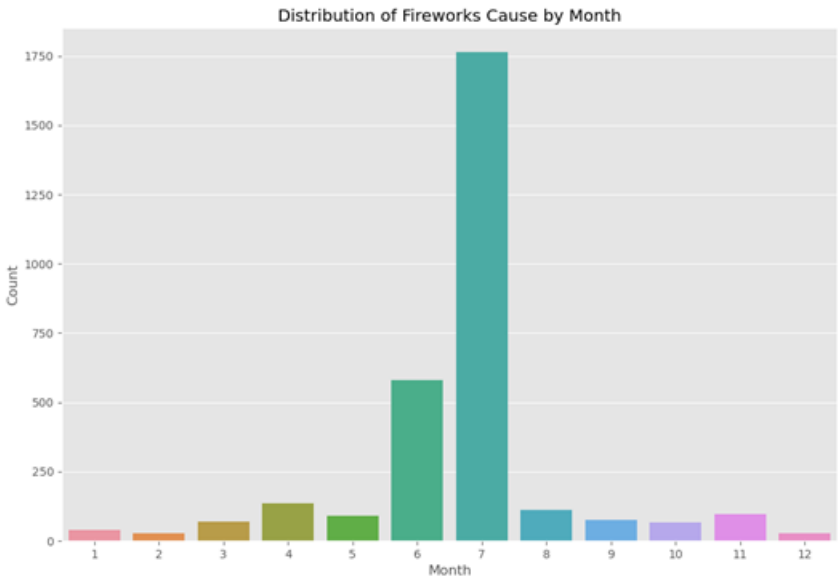




The first figure displays a heatmap of the causes of fires categorized by the discovery day. The heatmap shows the frequency or intensity of different fire causes across different discovery DOY. Darker shades represent higher frequency or intensity, while lighter shades indicate lower val-

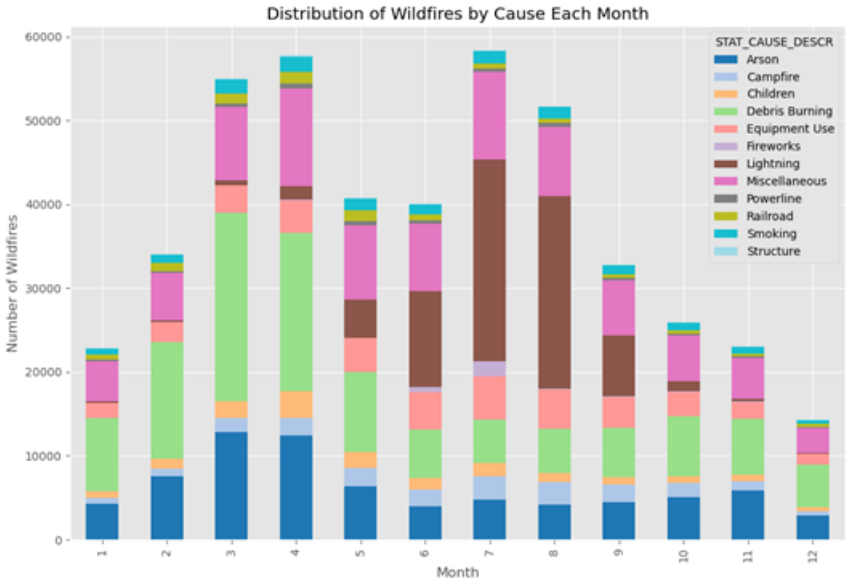
ues. The second figure presents a bar chart showing the number of fires with the same cause description, by the fire year.

- Debris Burning, Lightning, and Arson get quite high value in the first third of the year with a weak peak of about a month
- When looking by years, it gets higher to kind of a peak and gets a bit lower stable value.
- Lightning gets high value in the summer months, and is quite consistent over the years, as it depends on nature only.
- The Arson gets lower by the year.
- Miscellaneous gets a strong peak around the 4th of July and gets higher as the years pass.

We believe that the changes in state laws and public awareness could have an impact on human-caused wildfires and especially on causes of regulated fires such as 'Debris Burning'.

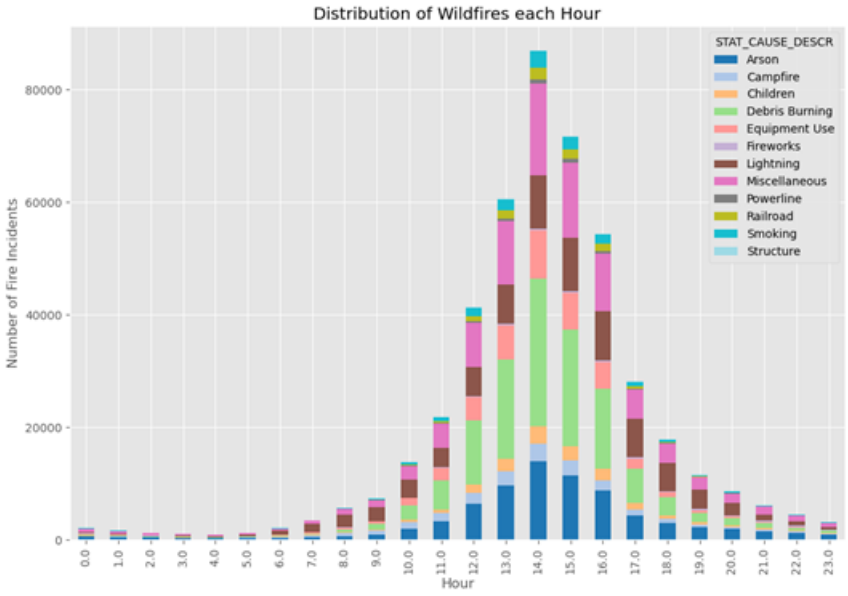



We find out that the number of wildfires caused by fireworks is higher in the summer months, particularly in July. This finding suggests that fireworks-related wildfires are more prevalent during the summer season when fireworks displays are more common.




When analyzing the number of wildfires caused each day in June and July, we found that the total number of wildfires caused is higher around the 4th and 5th of July, which is when Independence Day celebrations take place in the US. This finding indicates a correlation between wildfires caused by humans and Independence Day celebrations.
The plot shows the distribution of wildfires by cause each month. Some key observations from the plot:

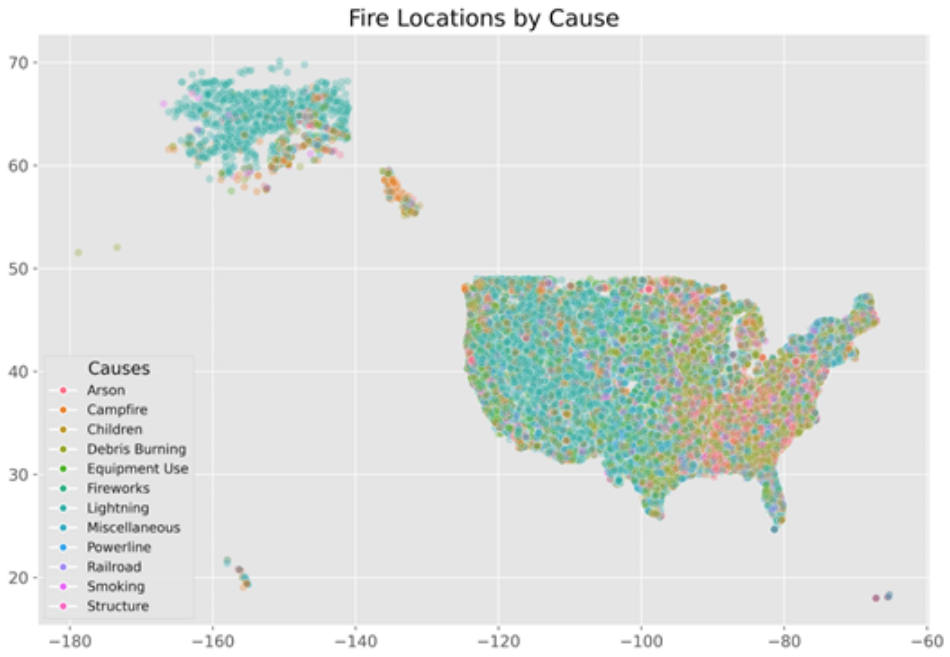
- Debris burning appears to be the leading cause of fire incidents across most months, with a notable peak in the months of March and April.
- Arson, Debris Burning, and Miscellaneous contribute significantly to fire incidents throughout the year, with relatively consistent levels across months.
- Causes like lightning, campfires, equipment use, structure, and fireworks show higher levels during the summer months, likely due to increased outdoor activities and celebrations.
- The overall number of fire incidents tends to be higher during the summer months, potentially due to dry conditions and increased outdoor activities.



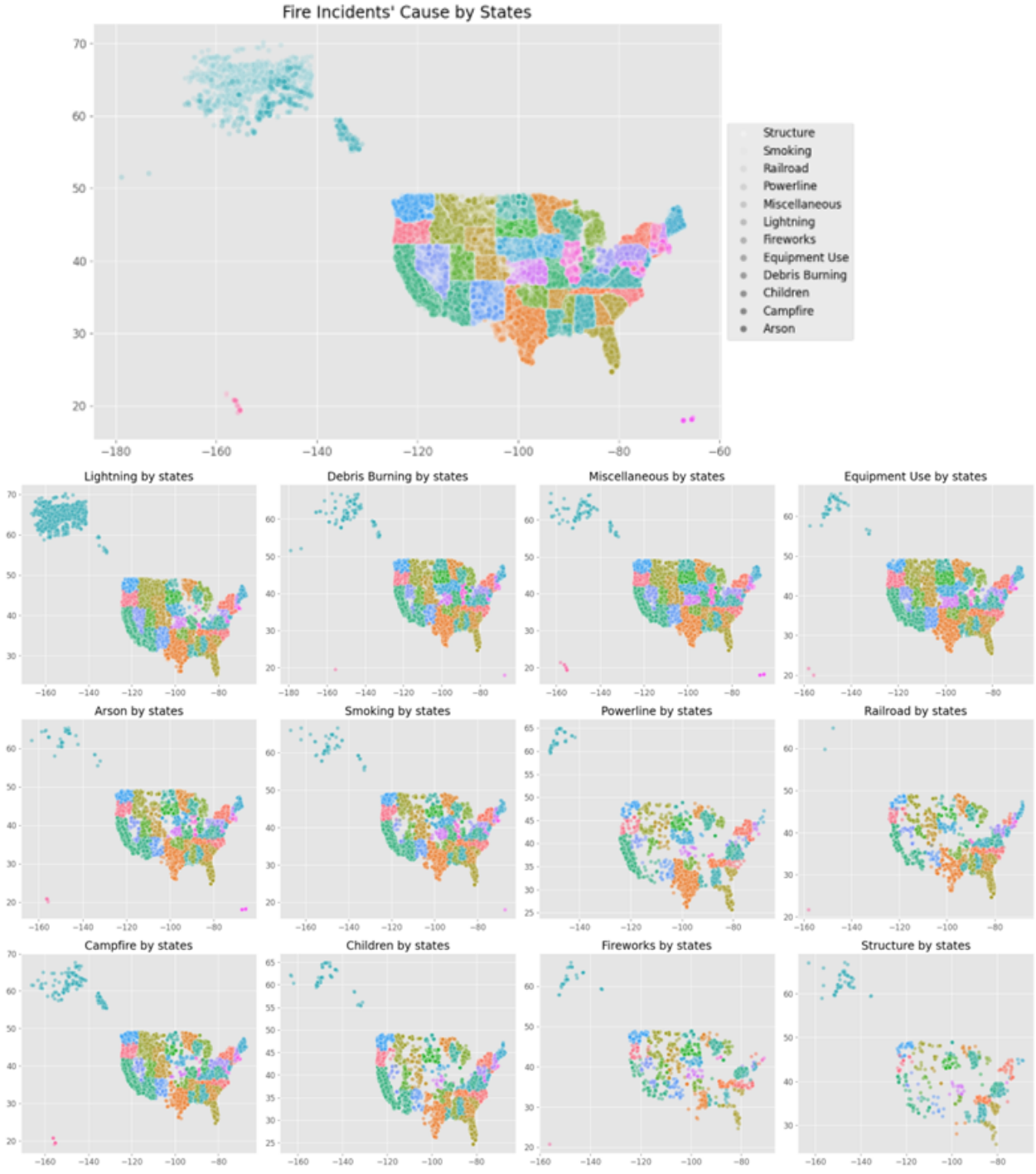

By analyzing the wildfire by the "hour in the day" it discovered, we can see most of the wildfires occurred around noon, and there was a peak at around 14. The ratio of the increase and decrease of wildfire is different between the causes, those caused by human activities or time specific have a higher ratio.

## Fire Causes by Location

We now turn to explore the distribution of fire caused by geographical location. There are 2 naive ways of looking at the geographical data: By pure coordinates ( Longitude and Latitude) and by state. First, we will examine the data by the coordinates of the incidents:




As can be seen in this graph, most cases in the West are caused by the "blue causes", i.e. Powerlines, Miscellaneous, but mostly Lightning. Most cases in the southeast are caused by the green and pink causes, i.e. Equipment Use, Fireworks, and Arson.
Let us now delve into the second naïve way and explore the various causes from the point of view of states. We have come to an understanding that encoding all 52 states separately will probably make it more difficult for our model to predict correctly, and it might be better to look for ways to group the states according to similar characteristics (such as geographical proximity).

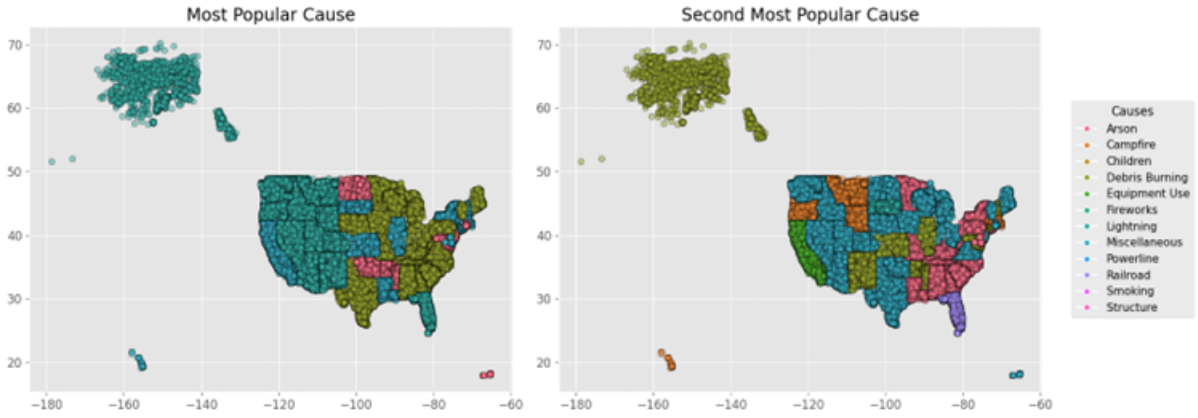



These graphs demonstrate the distribution of causes in each state. It shows the different states, each with its base color, and every dot (fire incident) on the map has a different intensity based on the cause of that fire. We can learn that in key areas in the middle of the US, there are more

incidents of arson and other "dark" causes, while the West Coast and Alaska are more likely to have fires caused by a more "mid-light" cause, like Lightning - which emphasizes the conclusion from the previous graph.
While exploring the geographic characteristics of the data, we decided to also examine how the size differs according to geographic data.


We made a diagram for each label according to its geographical location to see if there are areas with a particular distribution and allocation of labels according to the class size of the incident. It turns out that, on average, most fire incidents in Alaska and the Southeast are of size classes B and C, and only a few incidents are F and D sizes.

In Debris Burning, Other, Appliance Use, Arson, and Railroad, the fire size classes appear to be more mixed, with sizes from class A through class D predominating. For smokers, powerlines, children, fireworks and buildings, the predominant fire size class appears to be A - B (which are smaller fires), with few cases of other sizes.
Lastly, to conclude our exploration of fire incidents by geo-location attributes, we made these graphs below. On the left we can see the most common cause in each state, and on the right the second most common cause in each state.




These two graphs seal the deal with our observations from the previous graphs, and show how some labels tend to cluster together by geographic proximities (which goes hand by hand with climatic and sub-cultural similarities in those areas).

# Pre-Processing

## Irrelevant Features

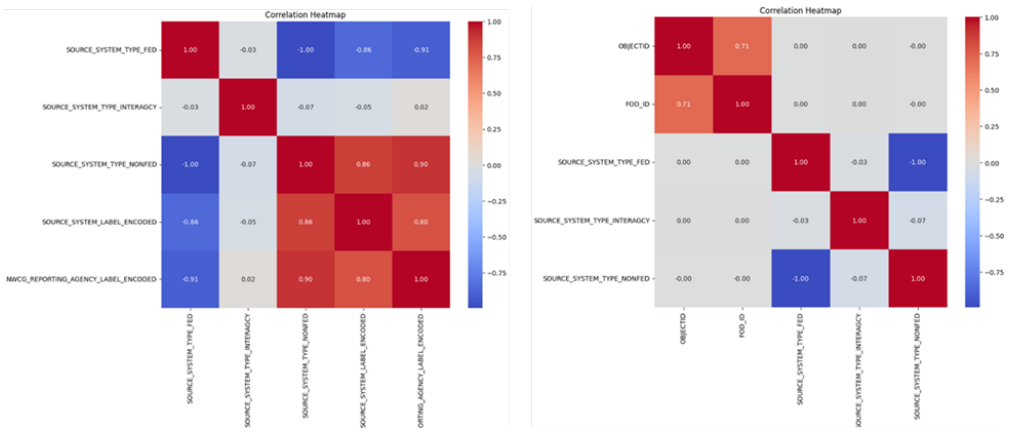
As searching for irrelevant features in the wildfire dataset, we considered the need of features with those properties:

- High missing value percentage
- Fire specific identification
- Text description of identifiers
- Part of features with high correlation

For the following features:

```
['OBJECTID', 'FOD_ID', 'FPA_ID', 'SOURCE_SYSTEM_TYPE', 'SOURCE_SYSTEM',
'NWCG_REPORTING_AGENCY', 'NWCG_REPORTING_UNIT_ID','NWCG_REPORTING_UNIT_NAME',
'SOURCE_REPORTING_UNIT', 'SOURCE_REPORTING_UNIT_NAME']
```

We saw a high correlation between 'Source_System_Type' , 'Source_System' and 'NWCG_Reporting_Agency'. As well as between 'Object_ID' and 'FOD_ID'. On the other hand, 'Source_System_Type' does not correlate with 'Object_ID' and 'FOD_ID'. From these connections, we can assume that the features: 'Source_System', 'Object_ID', and 'NWCG_Reporting_Agency' are irrelevant (we performed encoding to check the correlation):

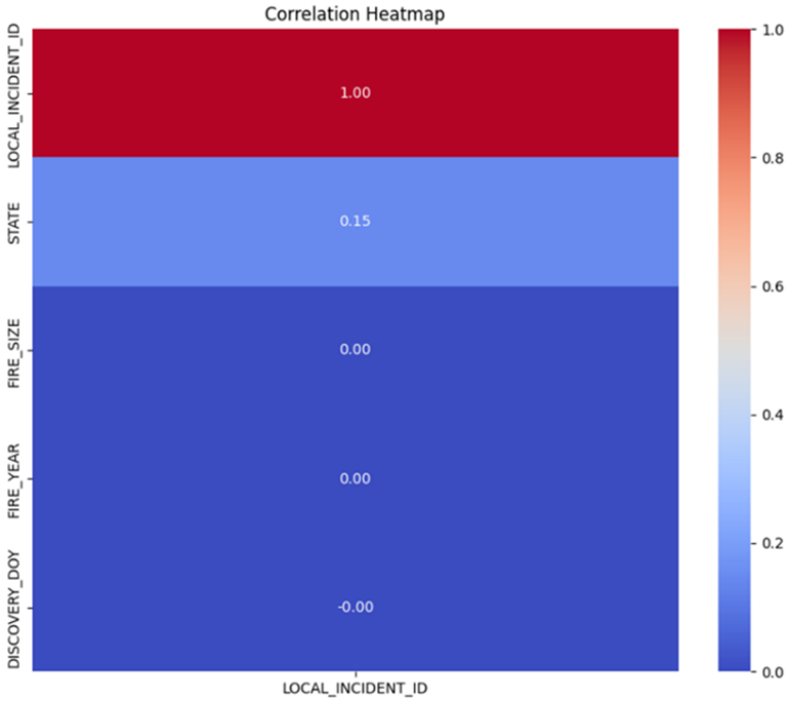



Moreover, since 'FOD_ID' and 'FPA_ID' have unique values, they won't contribute to feature engineering and are also irrelevant. We then examined the connection between 'NWCG_Reporting_Unit_ID' and 'NWCG_Reporting_Unit_Name'. We first assumed that each ID has only one name that doesn't contain any spelling mistakes. We found out that this assumption was correct. Lastly, we noticed that the features 'NWCG_Reporting_Unit_ID' and 'Source_Reporting_Unit' are similar in their values, and so do 'NWCG_Reporting_Unit_Name' and 'Source_Reporting_Name'. We used the library fuzzywuzzy to check how similar they are, based on the best-matching substring between the two strings (partial ratio). We decided that if the average ratio is over 70% then one feature in each pair is irrelevant for training. The first pair got a 73.67 % match and the second got a 79.17 % match. Since NWCG has a more organized format for its values- we prefer them over the other two.

For the following features: 'LOCAL_FIRE_REPORT_ID','LOCAL_INCIDENT_ID', 'FIRE_CODE', 'FIRE_NAME','ICS_209_INCIDENT_NUMBER', 'ICS_209_NAME', 'MTBS_ID', 'MTBS_FIRE_NAME','COMPLEX_NAME' These are the means of Nan values in each of those features:

| | |
|---|---|
| LOCAL_FIRE_REPORT_ID | 75.663067 |
| LOCAL_INCIDENT_ID | 41.201309 |
| FIRE_CODE | 81.369610 |
| FIRE_NAME | 49.170869 |
| ICS_209_INCIDENT_NUMBER | 98.653990 |
| ICS_209_NAME | 98.653990 |
| MTBS_ID | 99.403496 |
| MTBS_FIRE_NAME | 99.403496 |
| COMPLEX_NAME | 99.696498 |

We decided to remove: ['LOCAL_FIRE_REPORT_ID', 'FIRE_CODE', 'ICS_209_INCIDENT_NUMBER', 'ICS_209_NAME', 'MTBS_ID', 'MTBS_FIRE_NAME', 'COMPLEX_NAME'] features from our dataset, due to their high proportion of missing values, exceeding 75%. We also removed 'FIRE_NAME' due to almost 50% missing values and the risk of leakage.

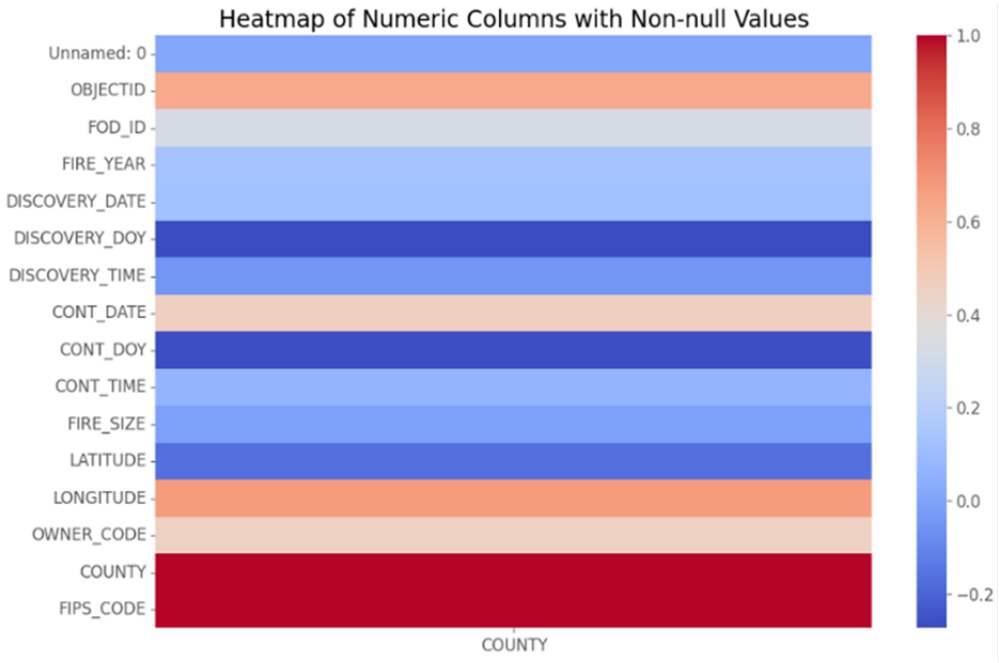



Additionally, we performed a correlation analysis including the feature 'LOCAL_INCIDENT_ID' and other variables of interest. The heatmap visualization of these correlations revealed that 'LOCAL_INCIDENT_ID' does not exhibit strong correlations with the other features, indicating that its inclusion may not significantly affect our analyses.

## Handle Missing Values

### FIPS Code and County

We noticed that FIPS_CODE has missing values and that COUNTY contains numeric values in some rows. We tried filtering the rows with numeric values to see if they could be used to fill in the missing values in FIPS_CODE, but we found that these values by themselves provided little information about the FIPS codes. We tried to find correlations between these numeric values in the COUNTY feature to other numeric features:




As can be seen in the heatmap diagram, there is a high correlation between these numeric COUNTY values and the FIPS_CODE. From this we can conclude that perhaps the COUNTY column can be omitted and we will use the FIPS_CODE instead. We grouped the columns: STATE, COUNTY and FIPS_CODE to check if each combination of state and county has a unique FIPS code. We created a dictionary where the keys are pairs of STATE and COUNTY, and the values are lists of all the different FIPS codes for each (state, county) pair. When we printed all the keys for which there is more than one FIPS code, we got an empty list. This means that there is a unique FIPS code for each combination (STATE, COUNTY). So we decided that we can get rid of the "COUNTY" column if there is a unique county code\name for each combination (STATE, FIPS_CODE). When we printed the county code, we did not get unique values, but variations of the same name.

That was enough for us to decide that the COUNTY feature can be dropped before training - STATE and FIPS CODE are informative enough.

Lastly, In our data set, the FIPS_CODE feature only represents the county part (as described in the dataset overview above). We want each FIPS code to be unique for each pair of (state, county) so we decided to use a 5-digit FIPS code, which is a concatenation of the 2-digit state FIPS code and the 3-digit county FIPS code. Using the US Counties data set, we filled in the FIPS code of records which had the 3 digits of the county FIPS code, and completed 2 digits of the state FIPS code. Then we filled the records with missing FIPS codes by the closest record by Euclidean distance, in the same state.

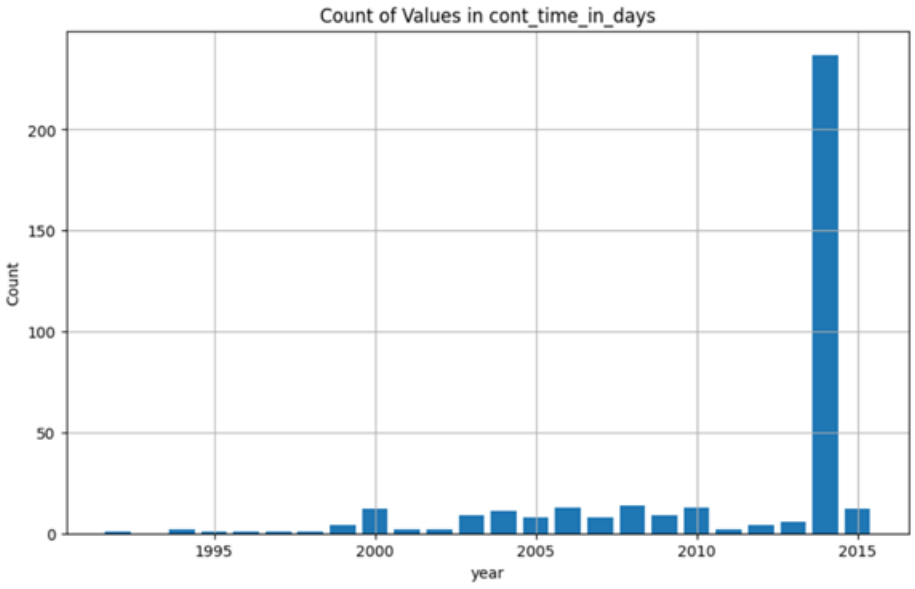
### Time

To make the predictions more accurate, we have split the time features into "hour" and "minute" features and ensured that each feature is represented by only two digits. Then we tried to fill in the missing values in the features: discovery time, contained date, contained day of week and contained time. It's done using a Random Forest Regressor model trained on the features: 'DISCOVERY_DATE', 'FIRE_SIZE', 'LATITUDE', and 'LONGITUDE'. The model was trained on the records with non-missing values in the target feature and used to predict the missing values. The model achieved an average accuracy score of 0.97 on the training set, but as some of the predictions were not valid, we got after corrections an accuracy score of about 0.9. The filling of time missing values caused some invalid values in the date and time columns, so we had to fix them by those rules:

- CONT_DATE should be after DISCOVERY_DATE, and with valid value
- CONT_DOY should be in correlation with CONT_DATE.
- All the time should be an integer in the valid range.

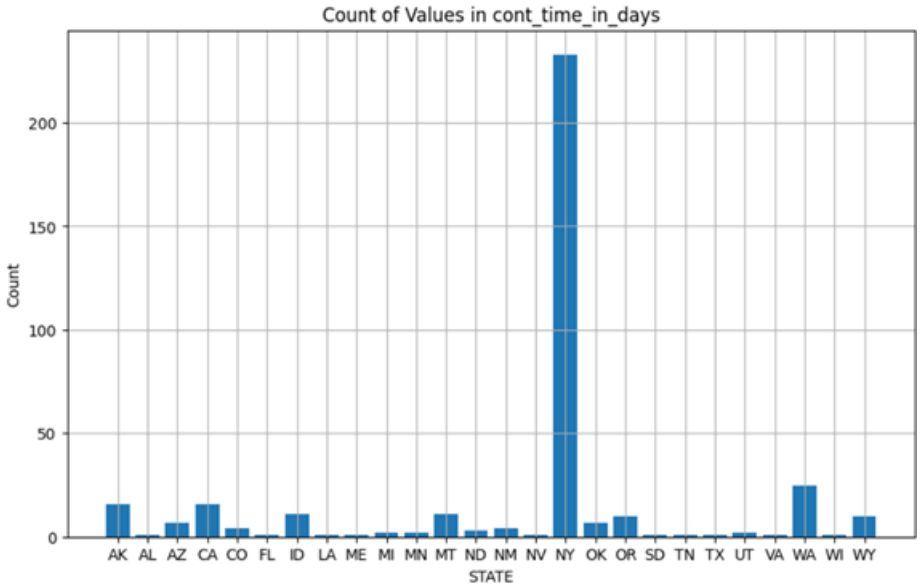
## Outliers Exploring

After plotting the distribution of the duration of fires, we encountered some unusual records of very long-lasting fires in the data. After checking the internet for the duration of wildfires in the US, we found that the average fire duration in 2013 was 37 days[1].
We wanted to check the veracity of those fires in our data before deleting them. This is a plot of the distribution of 4 months and above fires during the year:

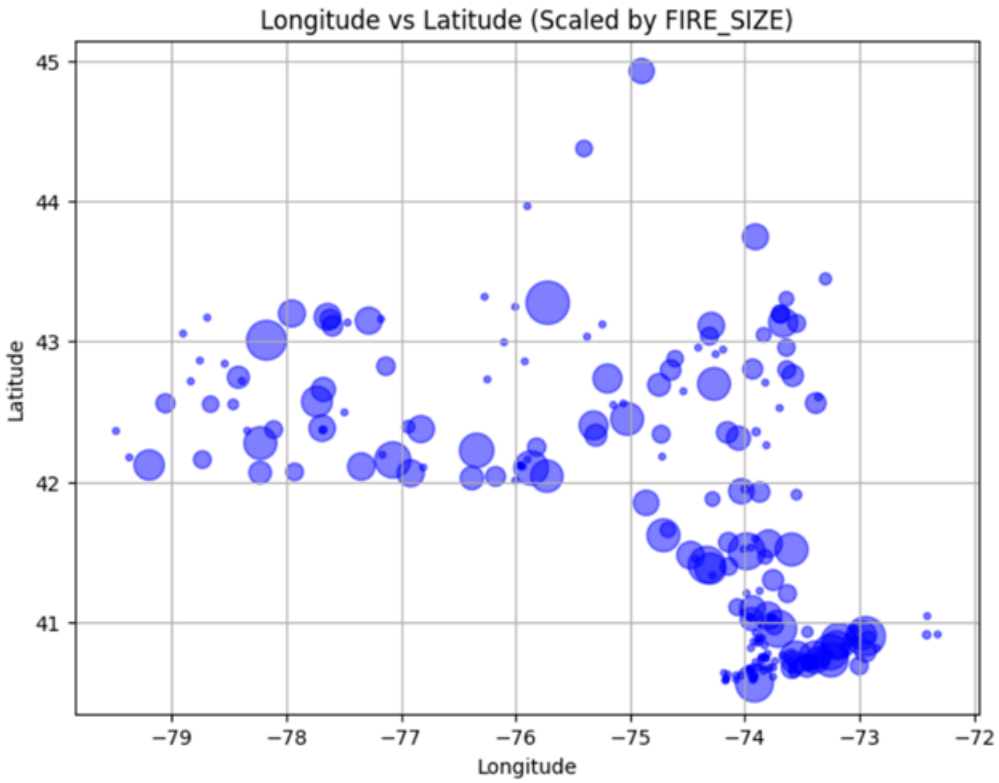



[1]https://wfca.com/wildfire-articles/how-long-do-wildfires-last/

Since most of them "occurred" in 2014, it became pretty obvious that these are outliers. But since it became interesting we continued plotting, now the distribution of those long fires over states:




looks like most of the longest fires are in New York State, which is unrealistic. And just to really make sure we are not deleting valid rows, or that these aren't rows representing the same fire, here's a plot of the longest fires (duration of 4 months or more) in MY by longitude and latitude (size changes with duration).



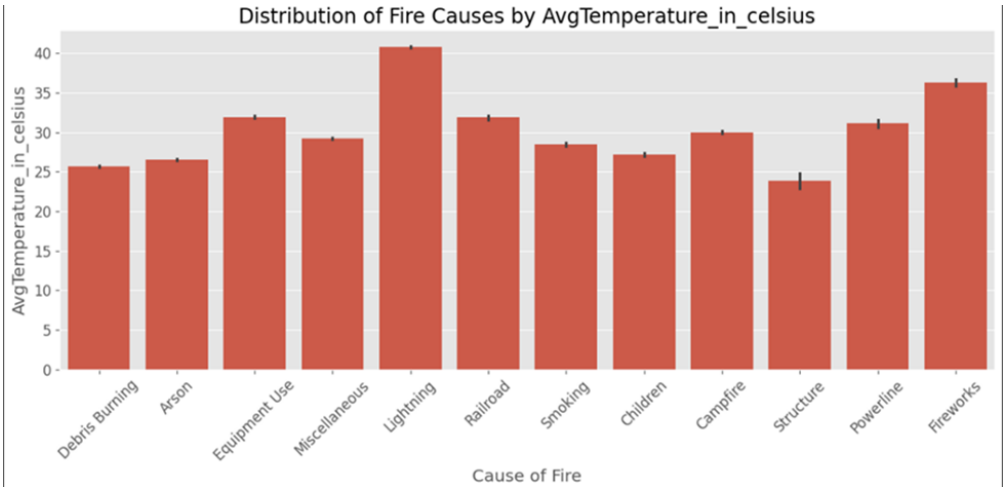

They are spreaded across the whole state, which is definitely unrealistic- we removed those rows.

# Feature Engineering

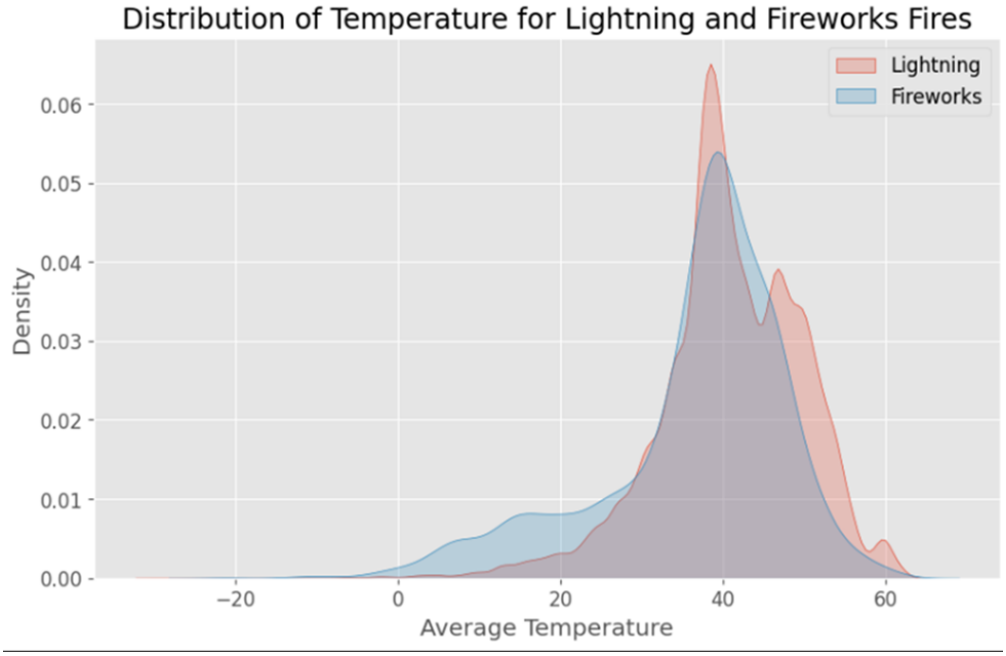
## Weather

When exploring the distribution of fire incidents by cause and month (plot is presented in the Data Analysis section) we found that Lightning, Fireworks and Debris Burning fires are high numbered in specific months. This fact led us to think that this might have to do with certain weather conditions in these months, so we tried to explore weather patterns and add new features related to this subject. We used an external dataset called "city_temperatures.csv" containing the following features: Region, Country, State, City, Month, Day, Year and AvgTemperature. This dataset contains the average temperatures of many states and regions around the world, in the range of years: 2015-2020. We decided to merge this dataset based on the State, Month, and Year features, so we can use the average temperatures in each state during the year for new features. We explored for missing values that might occur during the merge and we found that there are some missing years and states, so we completed them as explained below
First, we found that the years 1992-1994 weren't documented in the data, so we filled in their missing values based on the mean temperatures of each state and month of the year 1995,

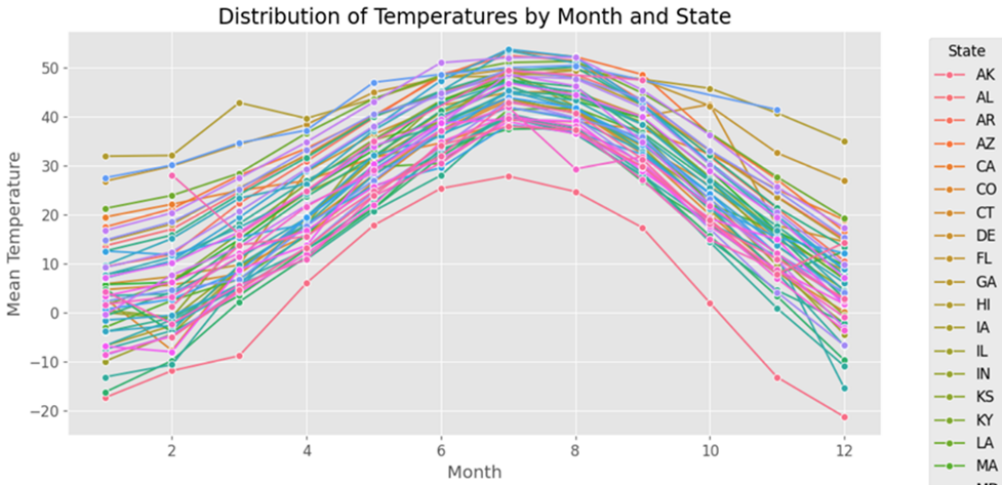
assuming that the mean temperatures of 1995 is the closest evaluation of the temperatures in the two years preceding.
Then, we found that the state "Puerto Rico" isn't documented as well. We calculated the mean longitude and latitude of Puerto Rico and any other US state from our data. We found that Florida has the closest Euclidean distance to Puerto Rico, and based on the assumption that nearby states have similar temperatures, we completed Puerto Rico's missing data based on Florida's data.
We added the average temperature feature as raw data to the train and test sets of our data. Since it was documented in Fahrenheit, we also added a feature of average temperatures in Celsius, for more comfortable exploring of patterns.




Since both lightning and fireworks fires are more common in higher temperatures, we checked if they are also affected by change of temperatures, but found that they are behaving almost in the same way:




Lastly, we added a few more features based on the new raw data we got from this dataset. Temperature Deviation and Is Summer, Is Winter which are binary features. We had to check first in which months the temperatures are considered high and low in the states:



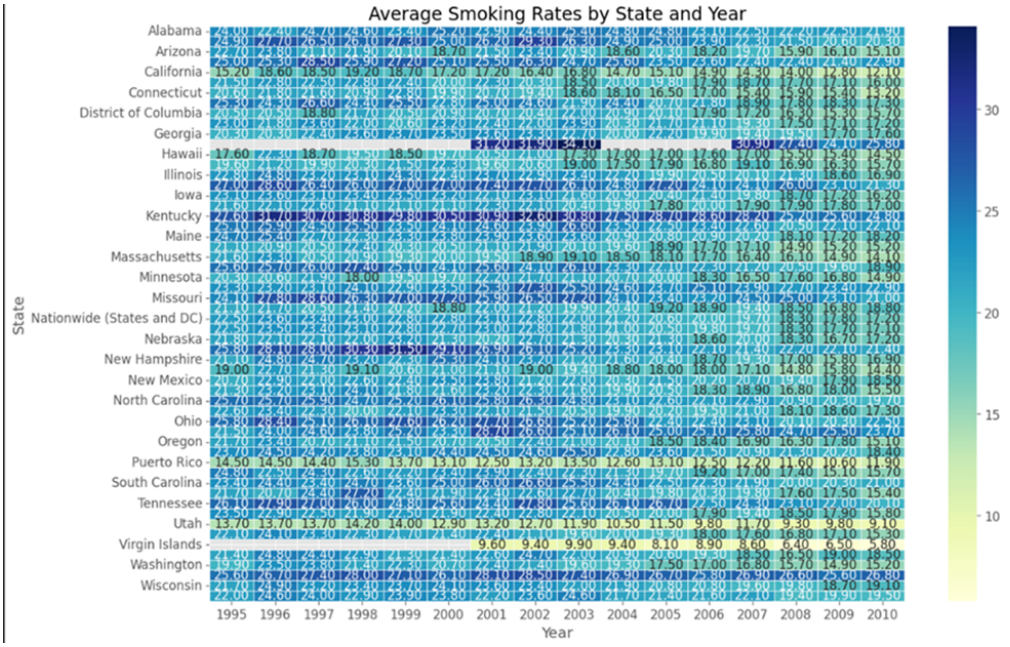

We can see that in all states- summertime is between June to August, and the average temperature is above 25 degrees, and winter is between December to February, and the average temperature is below 10 degrees. We created the features according to these results.

## Smoking

To explore smoking patterns that might help our model better understand the Smoking cause of fire, we used an external dataset called "tobacco.csv", which contains the prevalence and trends of tobacco use for 1995-2010. Percentages are weighted to population characteristics. This dataset contains the following features: Year, State, Smoke everyday, Smoke some days, Former smoker, Never smoked. We merged new features we extracted from this dataset based on the year and state features.

First, we found that there were some specific missing values in the following states and years: DC in the year 1995, Hawaii in 2004, Puerto Rico in 1995 and Utah in the years 1995-1996. We filled those missing values by the data in those states from the following year. For example, we filled Hawaii in 2004 based on the data from Hawaii in 2005.

Then, we decided to create a more comfortable features that we wanted to merge into our data instead of the 4 original features describing the prevalence of tobacco use: We added the "avg_smoking_rate" feature which is simply the mean by state and month of the sum of "Smoke every day" and "Smoke some days" percentages. It summarizes the average smoking rates in the states, without the need for data of the non-smokers (which is complementary to 1, hence there is no use for features about them as well).




Lastly, since the years 1992-1994, 2011-2015 are fully missing from this dataset, we continued to fill in the new features' missing values in those years. We decided to take the mean values of the three adjacent years of each of the missing years. It means that we had to do the filling process in a serial process, because for example the data of the year 2012 is based on the mean values

of 2009-2011, and 2011 is missing as well. Therefore, we divided it into two processes: forward process for the years 2011-2013. We started by filling each state in 2011 by the mean smoking rates in this state from the years 2008-2010, then continued in the same method for the years 2012 and 2013. And a backward process- filling 1994 from the years 1995-1997, then continuing to 1993 and lastly 1992. Then we could complete the "smoking category" feature based on the newly created data.

## Date and Time Features

The following features were extracted to capture datetime context of the wildfires:

- FIRE_MONTH - Calendar month in which the fire was discovered or confirmed to exist. Implemented using the DISCOVERY_DATE. From this feature we extracted two cyclic feature:
  - MONTH_SIN
  - MONTH_COS
- DISCOVERY_IS_WKND - True if the fire was discovered or confirmed to exist on weekend, otherwise false. Implemented using the DISCOVERY_DATE.
- DISCOVERY_DATETIME (to extract duration) - Date and time on which the fire was discovered or confirmed to exist. Implemented using the DISCOVERY_DATE, DISCOVERY_HOUR and DISCOVERY_MINUTE.
- CONT_DATETIME (to extract duration) - Date and time on which the fire was declared contained or otherwise controlled. Implemented using the CONT_DATE, CONT_HOUR and CONT_MINUTE.
- FIRE_DUR_DAYS - Time in days difference between the day fire was discovered and the day the fire was declared controlled. Implemented using the DISCOVERY_DATETIME and CONT_DATETIME.
- DISCOVERY_DOW - the discovery day of week. Splitted into two cyclic features, by using sine and cosine transformations on the relative part of week of the given day:
  - DAY_OF_WEEK_SIN
  - DAY_OF_WEEK_MONTH
- DISCOVERY_DOY - the discovery day of the year. Splitted into two cyclic features, by using sine and cosine transformations on the relative part of the year of the given day:
  - DAY_OF_YEAR_SIN
  - DAY_OF_YEAR_MONTH

- FOURTH_FIFTH_OF_JULY - true if the discovery day of the fire is the 4th or 5th of July, false otherwise.

## Geospatial Features

The following features were extracted to capture the geographical context of the wildfires:

- Distance_from_state_centroid - a feature that holds the distance of each fire from its county's centroid. The centroid is calculated by Longitude and Latitude, county is defined by the feature FIPS_CODE.
- Easting and Northing UTM grid coordinates represent positions on the Earth's surface using distances in meters to the east (easting) and north (northing) from a reference point. While latitude and longitude may not be critical, they are certainly important in the baseline model. This indicates that geographical location will likely have a significant impact on the target variable. We believe that by converting latitude and longitude coordinates to northing and easting coordinates, the model may be better equipped to identify spatial dependencies or patterns within the data.
- NEAR_FOREST - True if the reporting unit name contains the word "forest", otherwise false. Implemented using the NWCG_REPORTING_UNIT_NAME.

## Encoding and Clustering Categorical Features

- NWCG_REPROTING_UNIT_ID - this was one of the most important features to Catboost so we didn't want to give it up, but we had to decrease its dimension from 1286 categories. We decided to use the method we saw in class- take the most common categories, and categorize the other categories as "other". There is a drop in occurrences of IDs after the eighth most common ID, so we took the first eight most common IDs.

| NWCG_REPROTING_UNIT_ID | COUNT |
|---|---|
| USGAGAS | 44679 |
| USTXTXS | 29665 |
| USNCNCS | 24295 |
| USFLFLS | 21393 |
| USNYNYX | 20069 |
| USMSMSS | 17795 |
| USALALS | 17345 |
| USSCSCS | 15015 |
| USMNMNS | 7903 |

- FIRE_SIZE_CLASS - although this feature has only 7 categories, after printing the value count of it, we found that it could be reduced as well. The first 3 classes are very common

where the others are not. We decided to combine all D-G classes into D which will be a "more than 100 acres fires" category.

| FIRE_SIZE_CLASS | Count |
|---|---|
| A | 227406 |
| B | 163792 |
| C | 52875 |
| D | 6797 |
| E | 3417 |
| F | 1912 |
| G | 941 |

- STATE - We believe this is an important feature that could be reduced by geographic relations. With the information we gathered from the graphs in section 1, we decided to split the states into 6 groups, based on proximity and the probability of fire causes.

| Group Name | States |
|---|---|
| West Coast | CA, OR, WA, AK, HI |
| Western States | UT, ID, NV, CO, NM, MT, WY, AZ |
| Midwest | ND, SD, NE, KS, MN, IA, MO, WI, IL, MI, IN, OH |
| South | TX, LA, MS, AL, FL, GA, SC, NC, TN, KY, AR, VA, WV, OK |
| Northeast | ME, NH, VT, MA, RI, CT, NY, NJ, PA, DE, MD, DC |
| Territories | PR, GU, VI, AS, MP |

Lastly, we performed Label Encoding using LabelEncoder on all of these features.

# Model

After exploring the dataset and engineering relevant features, we proceed to build the model. We aim to develop a predictive model capable of accurately classifying the cause of wildfires based on various attributes. The model selection process involves training and evaluating multiple machine learning algorithms to identify the most suitable model for the task.

## Evaluation Metrics

We evaluate the model using the weighted ROC-AUC OneVsRest score, as required in the instructions of this project. We used various types of plots to visualize the results of our models, like ROC Curve and Confusion Matrices. These plots helped us to understand the models' performance in each class, and to identify where they're shinning and where their weak spots are. Model Selection Along our work process at this project, we implemented 2 types of models: CatBoost and XGBoost.

### Catboost

As we explained in the Baseline Model Section, Catboost was our first choice because of its ability to handle categorical features. Unfortunately, after adding the features we described above (or any combination of them), the AUC ROC score of the model decreased by 1-4% (depending on the combination of features we entered). Fortunately, we couldn't accept these results so at a very risky time before submission we have decided to move on to another model.

### XGBoost

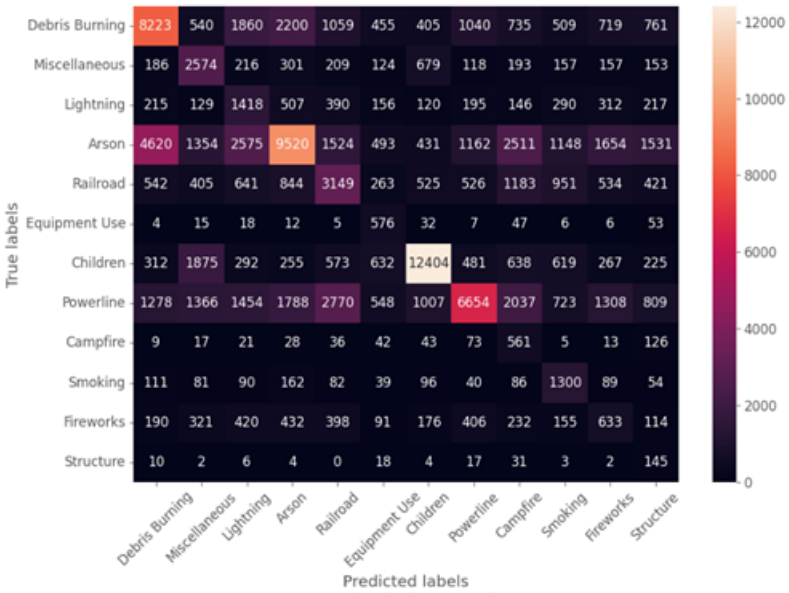
XGBoost is a popular gradient boosting framework, known for its performance relative to similar decision tree algorithms. Since it doesn't handle categorial features like Catboost we had to go back to some pre-processing and make some adjustments.

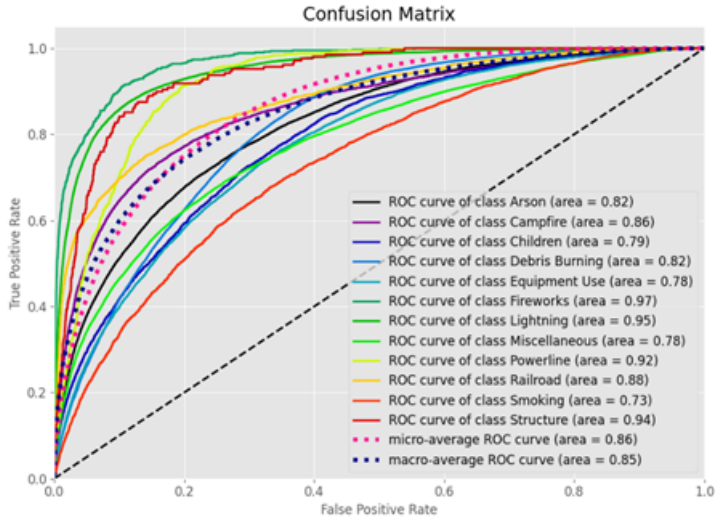
## Baseline Models

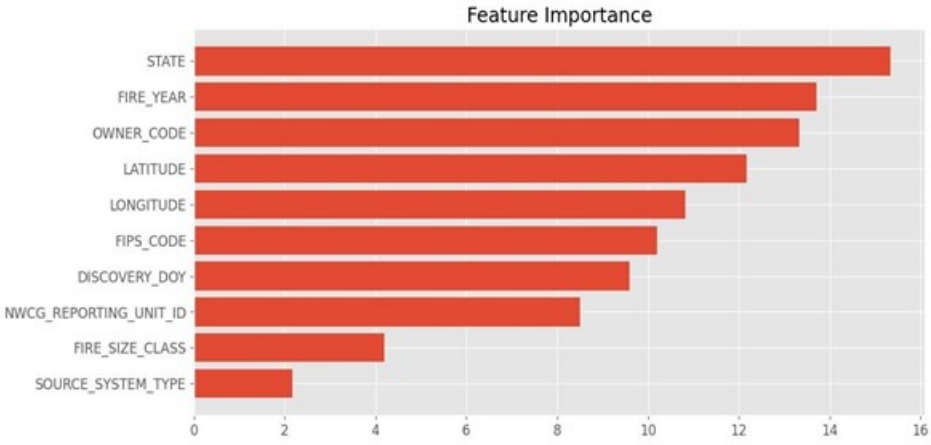
After investigating different learning methods and machine learning models, we have concluded that the boosting methods are currently the strongest learning methods in the field. At first, we thought that the CatBoost model would be a good fit for our dataset, which contains a mixture of numerical and categorical attributes, including the cause of wildfires as the target variable.
We decided to use in our baseline all the features that seemed "important enough" after preprocessing, but we left out all the datetime features that required further processing and feature engineering.
These are the features that we used:

```
['LATITUDE', 'LONGITUDE, 'STATE', 'FIPS_CODE', 'NWCG_REPORTING_UNIT_ID',
'OWNER_CODE', 'SOURCE_SYSTEM_TYPE', 'FIRE_SIZE_CLASS', 'DISCOVERY_DOY', 'FIRE_YEAR']
```

We used the train data and evaluation data and ran the CatBoost algorithm with 1000 iterations.


Debris Burning 8223 540 1860 2200 1059 455 405 1040 735 509 719 761
Miscellaneous 186 2574 216 301 209 124 679 118 193 157 157 153
Lightning 215 129 1418 507 390 156 120 195 146 290 312 217
Arson 4620 1354 2575 9520 1524 493 431 1162 2511 1148 1654 1531
Railroad 542 405 641 844 3149 263 525 526 1183 951 534 421
Equipment Use 4 15 18 12 5 576 32 7 47 6 6 53
Children 312 1875 292 255 573 632 12404 481 638 619 267 225
Powerline 1278 1366 1454 1788 2770 548 1007 6654 2037 723 1308 809
Campfire 9 17 21 28 36 42 43 73 561 5 13 126
Smoking 111 81 90 162 82 39 96 40 86 1300 89 54
Fireworks 190 321 420 432 398 91 176 406 232 155 633 114
Structure 10 2 6 4 0 18 4 17 31 3 2 145
True labels
Debris Burning
Miscellaneous
Lightning
Arson
Railroad
Equipment Use
Children
Powerline
Campfire
Smoking
Fireworks
Structure
Predicted labels
12000
10000
8000
6000
4000
2000
0

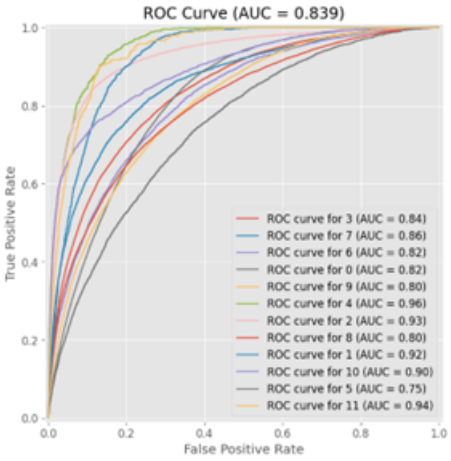

Confusion Matrix
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class Arson (area = 0.82)
ROC curve of class Campfire (area = 0.86)
ROC curve of class Children (area = 0.79)
ROC curve of class Debris Burning (area = 0.82)
ROC curve of class Equipment Use (area = 0.78)
ROC curve of class Fireworks (area = 0.97)
ROC curve of class Lightning (area = 0.95)
ROC curve of class Miscellaneous (area = 0.78)
ROC curve of class Powerline (area = 0.92)
ROC curve of class Railroad (area = 0.88)
ROC curve of class Smoking (area = 0.73)
ROC curve of class Structure (area = 0.94)
micro-average ROC curve (area = 0.86)
macro-average ROC curve (area = 0.85)

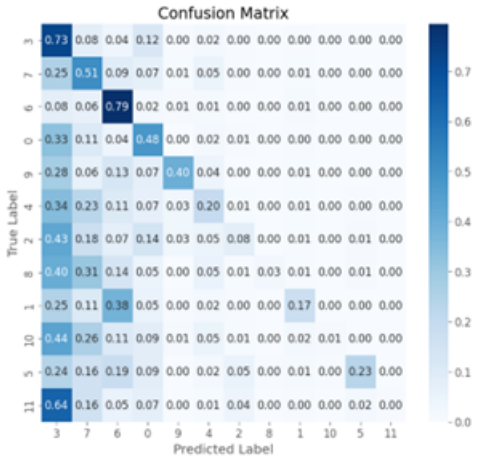

Feature Importance
STATE
FIRE_YEAR
OWNER_CODE
LATITUDE
LONGITUDE
FIPS_CODE
DISCOVERY_DOY
NWCG_REPORTING_UNIT_ID
FIRE_SIZE_CLASS
SOURCE_SYSTEM_TYPE
0
2
4
6
8
10
12
14
16

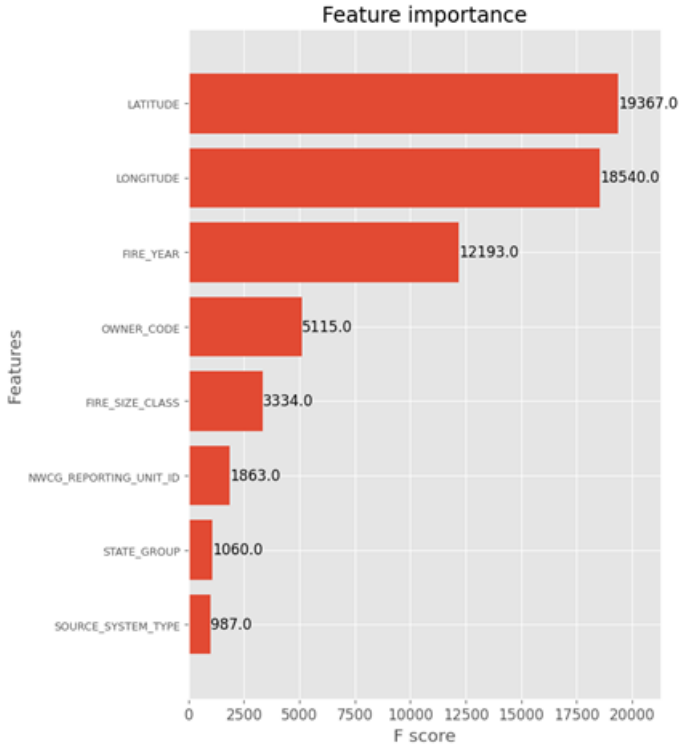
The model achieved a weighted ROC-AUC score of 0.83 on the test set, indicating a relatively high level of predictive accuracy. The confusion matrix reveals that the model performs well in some classes, with some variability in precision and recall scores across different causes of wildfires. There is high confusion between the pairs: Powerline-Railroad, and Arson and Campfire. The ROC plot illustrates the model's performance in distinguishing between different classes, where the highest predictive labels are Fireworks and Lightning, and the lowest predictive models are Smoking, Equipment Use, and Miscellaneous. In the experimentation, we found that any optimization to the baseline model only made it worse and hurt the predictions and the score. Therefore, we decided to give up and try another one.

## Final Model

We weren't satisfied with the CatBoost's results and decided to move on to a different model. We have settled down on XGBoost. For the comparative baseline, we opted to include all the features used in the later CatBoost model.




After incorporating features from the feature engineering process, we observed an improvement in the model's performance. Specifically, we achieved a validation score of 0.8525 and a test score of 0.8497.

## Hyperparameter Tuning

To optimize the model's performance, we conduct hyperparameter tuning using the "Optuna" library which is a hyperparameter optimization framework. The hyperparameter tuning process involves splitting the dataset into training and validation sets, and searching for the optimal combination of hyperparameters that minimize a multi-class log loss function using gradient boosting techniques. Those hyperparameters include:

`['n_estimators', 'learning_rate', 'max_depth', 'colsample_bytree', 'subsample', 'reg_lambda', 'reg_alpha']`

Before performing hyperparameter tuning the results were as following:

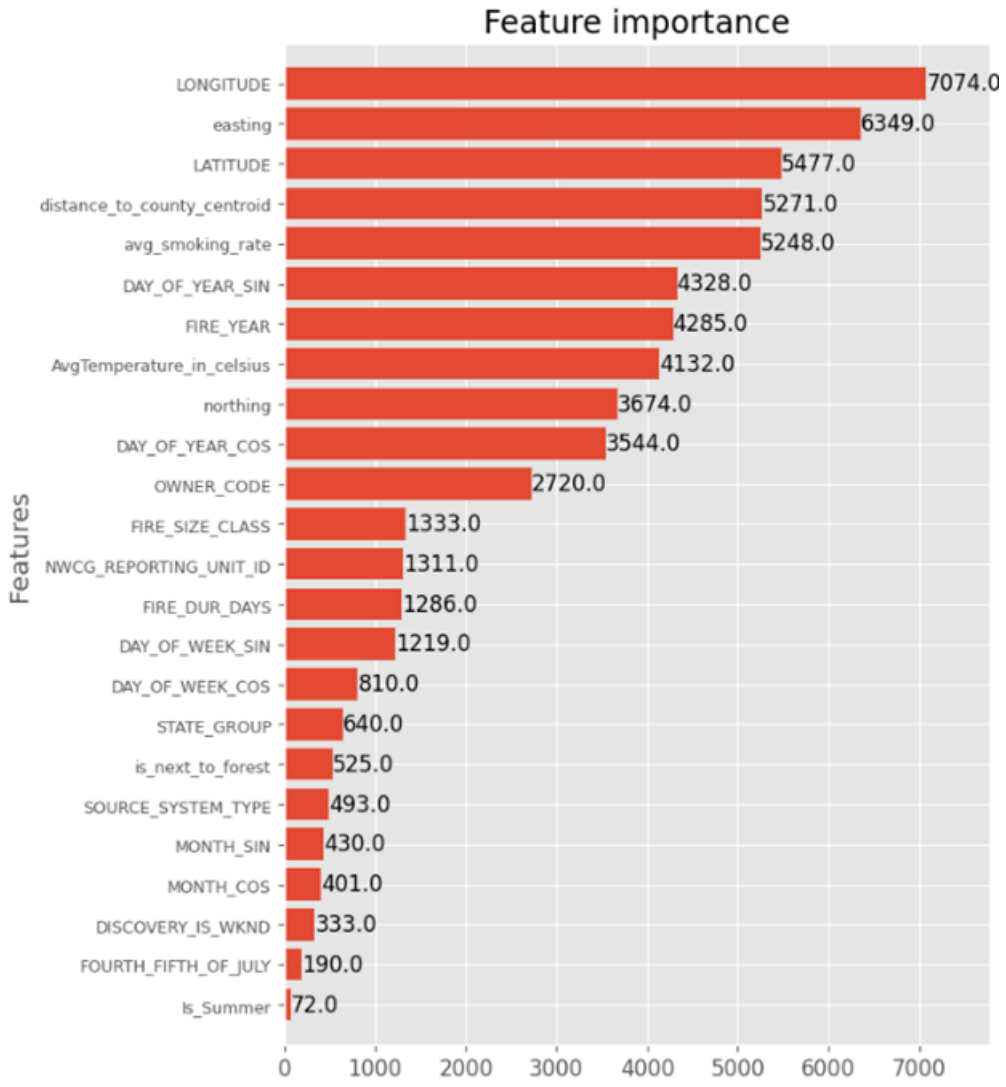
- On the validation set: 0.8525
- On the test set: 0.8498

Those are the hyperparameters that were chosen as best for out model:

| Parameter | Value |
|---|---|
| n_estimators | 790 |
| learning_rate | 0.0757 |
| max_depth | 7 |
| colsample_bytree | 0.8260 |
| subsample | 0.6990 |
| min_child_weight | 0.2426 |
| lambda | 0.9823 |
| alpha | 0.6229 |

## Feature Importance

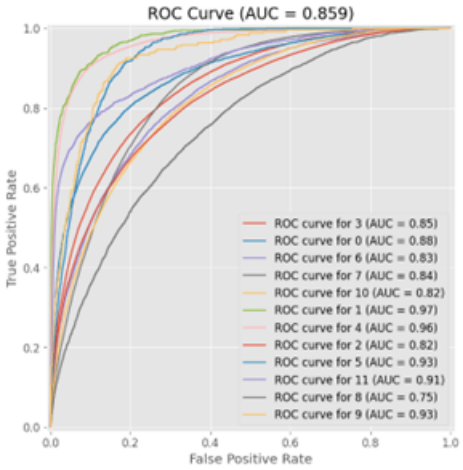
To identify the most relevant features for the predictive model we printed the feature importance:

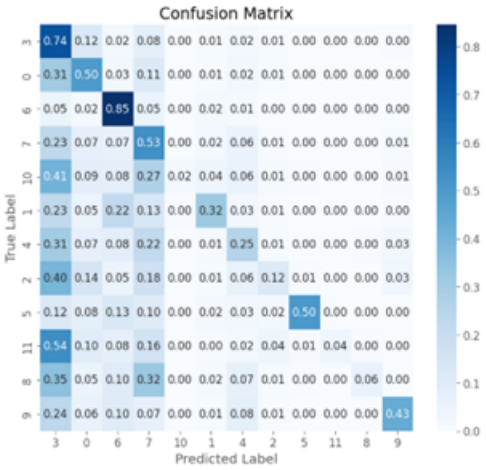



It seems that the geospatial features are the most important features of the model, where datetime features are the least important. We noticed already in the baseline model that fireworks for example are predicted well, it can explain why the fourth of july feature isn't significant.
Final Performance Results: Results after hyperparameters tuning:

- Score on the validation set: 0.8617
- Score on the test set: 0.8592

Final graphs on our test set:


ROC Curve (AUC = 0.859)
True Positive Rate
False Positive Rate
ROC curve for 3 (AUC = 0.85)
ROC curve for 0 (AUC = 0.88)
ROC curve for 6 (AUC = 0.83)
ROC curve for 7 (AUC = 0.84)
ROC curve for 10 (AUC = 0.82)
ROC curve for 1 (AUC = 0.97)
ROC curve for 4 (AUC = 0.96)
ROC curve for 2 (AUC = 0.82)
ROC curve for 5 (AUC = 0.93)
ROC curve for 11 (AUC = 0.91)
ROC curve for 8 (AUC = 0.75)
ROC curve for 9 (AUC = 0.93)


Confusion Matrix
True Label
Predicted Label

## Further Steps If We Had More Time

- Sub models on labels that the model couldn't divide - unfortunately, this is a method we started to implement but couldn't finish due to time constraints. In general the whole concept of Ensemble Learning methods is very strong in the Machine Learning field, and using the sub-model probabilities of prediction on smaller groups of labels would have improved our results almost definitely.

- A more delicate clustering - the initial separation of states into groups we made, which is the one we eventually ended up with, is quite coarse. This separation doesn't take into consideration the economical aspects mentioned in the geographic exploration part in this report. If we had the time we could probably make a more delicate and accurate clustering which would hopefully help us to classify the fires better.

- Better Geospatial features - one of the subjects we really wanted to explore is the distribution of wildfire causes as a function of location. We (desperately) searched for data about forests or national parks in the US, because we believed it could strengthen the results of Campfire predictions, and lead to other patterns we could explore and use. Another thing we could have done is geospatial clustering for states or counties since these are large categorical features.

- More is Less (side note) - we began this journey thinking "what more information we can add to our current data", not knowing that the baseline model will have pretty high results as it is. During the project we added a lot more features that we believed could be helpful, but we came to the conclusion that more features actually lowered the scoring of prediction so we had to give them up. If we had more time we would focus more on feature selection and on improving directly the results of the baseline model. Although we hoped for better scoring, we are proud of our learning process (see what we did there?) and of our improvement during this process.