CS 6220: Data Mining Final Report

Timothy Chen, Li McCarthy, Kai Wu

I. Abstract

Machine learning methods, typically associated with computational intelligence (i.e. artificial intelligence), are now widely prevalent in the environmental sciences. Motivated by the 2020 California and Oregon wildfires which were two of the largest to date in their respective states (1,2), we created models to predict wildfire sizes. Using aggregated data about U.S wildfires from 1992 to 2015 (6), we selected the fire size class (A-G) as our target label. Within the dataset, we used the longitude, latitude, state, and the time delta between discovery and containment time to predict the fire size class. A decision tree was used to model the data, as the transparency of the reasoning for the predictions can be valuable in identifying potential reasons for why a wildfire size might be severe. When fitting the data to the decision tree, we found that our engineered feature, hours to containment, had a high feature importance using the Gini criterion. With the decision tree model as a basis, we further used a random forest classifier to finally achieve a testing accuracy of about 70%.

Introduction

Fire has been a natural occurrence since the dawn of time. In spite of this, it's still something that's far from being under our control. In recent years, communities in the Western United States have been struck by a series of devastating wildfires. The combination of thunderstorms, forceful winds, and drought-ridden dry terrain resulted in record-breaking megafires, burning more than 8.2 million acres of land (3), destroying more than 10 thousand buildings (3), and killing at least 35 people. (4,5) The unprecedented fires created smoke so thick that California governor Gavin Newsom compared breathing it to smoking 400 cigarettes a day (4).

The United States is not alone in the fight against this particular natural disaster - there are victims all across the world. In 2019, India was struck with massive forest fires that erupted across Karnataka's Bandipur National Park. Approximately 10,920 acres were burned, with additional damage spreading to a nearby forest range in Tamil Nadu. (12) India also experienced severe wildfires in Uttarakhand a few years earlier in 2016, resulting in massive damages - the average temperature of northern India was reported to have increased by 0.2 °C. (14) In 2009, Australia fell victim to a series of bushfires, widely referred to as Black Saturday, resulting in Australia's highest-ever loss of human life from a bushfire with 173 deaths, 414 injured, and thousands of wildlife killed. More than 1.1 million acres were burnt, and 3,500 structures destroyed across multiple towns (12).

It has become abundantly clear that wildfires are extremely dangerous. Experts are indicating that the increase of this natural disaster is due to a combination of poor forest management and climate change. Significant bodies of research relating to wildfire likelihood, risk, and reduction already exist. The National Fire Data Rating System (NFDRS) (10), which the well known "Smokey Bear Signs" (Fig.0) originate from, allows U.S. local residents, tourists, and fire management authorities to make their own judgments on how much they want to prepare for wildfires in a certain area. The NFDRS is based on the following environmental variables: fuels, which broadly encompass burnable components such as grass or shrubs present in the area,

weather, which includes precipitation and humidity, topography, which includes features such as the slope of the land in the area, and risk, which encompasses other factors including human involvement. These features are then fed into a software such as the Weather Information Management System (WIMS), or private vendor software. It is worth note that historically these models have drawn from slightly different features or combinations thereof, resulting in different results from the same inputs (11). From the inputs, software using the NFDRS predicts a Low, Moderate, High, Very High, or Extreme fire risk.

Fig. 0: Smokey Bear Signs across the USA

From left to right: Woodfords, CA, Keweenaw County, MI, Susquehannock State Forest, PA
http://smokeybearassociation.com/smokey-bear-signs-across-the-usa/

Given the devastating impact of wildfires, being able to continuously evolve and adapt our understanding of what causes a severe wildfire and predict this can be valuable both for wildfire prevention and response. The ability to predict the severity of a wildfire based on certain features can allow for more informed preventative measures, hopefully reducing human and material losses. We investigated a large wildfire dataset (6) with two primary goals: to gain a better understanding of the features correlated with wildfire severity, and to examine possible predictive models for severity.

II. Methodology

Our core dataset, "1.88 Million US Wildfires Dataset" (6), aggregates U.S. wildfire data from 1992 to 2015. The dataset consists of the following core data elements:

- Global unique identifiers.
- Unique identifiers which link back to the original datasets.
- Information about the group reporting the fire.
- The cause of the fire
- Dates and times of both the fire discovery and containment.
- Estimate of the acres within the final perimeter of the fire.
- Location information: latitude and longitude, state, county, and geographic area.

We selected fire size class as our target variable, and evaluated the remaining feature importances using Gini impurity to decide on what features made it to our final model. Our first step to addressing our topic was selecting only the subset of data that contained the desired targets. Given the immense size of the dataset, removing rows which had missing information still left us with a sizable amount of data. Then, we isolated any available features that had the potential to be useful for predicting fire severity. Our final set of columns comprised of the following:

- Fire Size Class: a standard fire size reporting metric, ordinally encoded. Alphabetical labels correspond to a scale where A is the smallest size and G is the largest. This was treated as one possible label. The fire size classes were mapped to ordinal integers (A=1, B=2, …, G=7) which represent the data numerically which is easier to train models on.
    - A: greater than 0 but less than or equal to 0.25 acres
    - B: 0.26 to 9.9 acres
    - C: 10.0 to 99.9 acres
    - D: 100 to 299 acres
    - E: 300 to 999 acres
    - F: 1000 to 4999 acres
    - G: 5000+ acres

- Hours to Containment: the total number of hours between fire discovery time and fire containment. This was a new feature calculated using Discovery Time and Containment Time.
- Discovery Month: month of the year, represented as integers 1-12 or one-hot encoded. The inclusion of month did not seem to have a high impact on accuracy, even when encoded.
- U.S. State: 52 US states (including DC and PR), ordinally encoded.
- Fire Stat Cause Code:
    - 1: Lightning
    - 2: Equipment Use
    - 3: Smoking
    - 4: Campfire
    - 5: Debris Burning
    - 6: Railroad
    - 7: Arson
    - 8: Children
    - 9: Miscellaneous
    - 10: Fireworks
    - 11: Powerline
    - 12: Structure
    - 13: Missing/unknown

Of the thirteen potential fire causes, we found that most did not contribute much to model accuracy after using one-hot encoding on the column. The causes that had the most feature importance were Lightning and Campfire. In Fig. 1B, it is likely that Campfire cause is associated with small fire size classes (lowest fire size class mean), while in Fig. 1A Lightning is the most numerous cause of wildfires while associated with a medium fire size class mean in Fig. 1B.

- Latitude and Longitude: continuous numerical values.

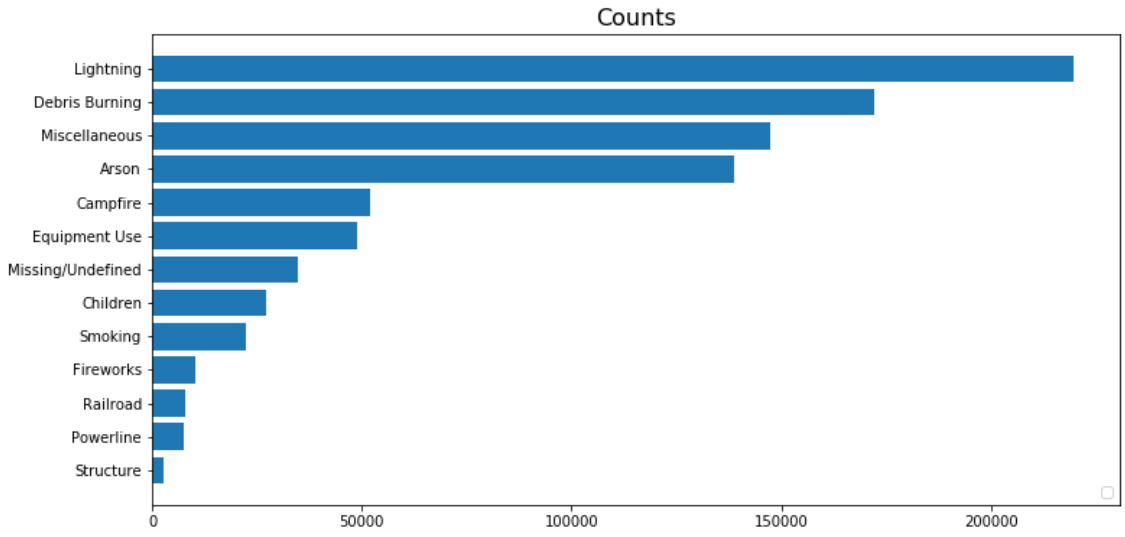
The dataset was also heavily skewed toward fire size classes A and B, with the total dataset size of almost 900,000 instances, and roughly 760,000 of them being a part of class A or B. Meanwhile, fire size class G only accounted for 3,151 of the instances. Given how skewed the

data was, we performed undersampling on classes A and B since we had so much data to start with. This was noted to have improved our testing accuracy for the decision tree by about 3-5%.

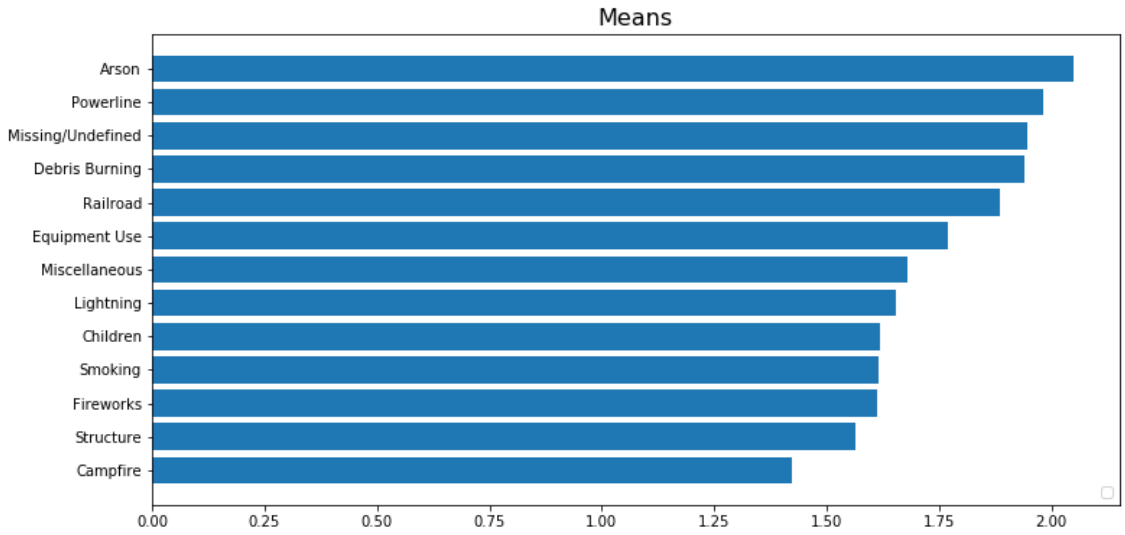
Exploratory figures for Fire Causes. Counts of each Cause, and Fire Size Class Mean by Cause

Fig.1: Exploratory figures for fire causes.

1A: Total count of fires aggregated by type of cause.



1B: Mean severity for each type of cause, with a higher value indicating greater severity.



III. Code

Our code was collaboratively written in Python 3.7+ and Jupyter Notebook, with primary usage of the pandas library for data processing, scikit-learn for modeling, git for version control. A partial code sample is included below with commentary.

To obtain our working dataset, we narrowed down our imports from the original data via a sqlite connection.

Load in data

```
# Select only pertinent features (this should make loading in data faster)
con = sqlite3.connect("wildfire_data.sqlite")
query="""
SELECT
FIRE_NAME,
FIRE_SIZE_CLASS,
STAT_CAUSE_DESCR, STAT_CAUSE_CODE,
STATE, COUNTY,
LONGITUDE, LATITUDE,
DISCOVERY_DATE,
DISCOVERY_TIME,
CONT_DATE,
CONT_TIME,
FIRE_YEAR
from Fires
"""
query=query.strip()
df = pd.read_sql_query(query, con)
con.close();
```

As we were working with two targets, one of which we derived from existing features, we dropped the rows for which we could not derive these targets.

Drop rows with missing data

Given our data is not sparse at all, we have the freedom to just drop all rows that are missing data we care about.

```
# drop the following rows if they have missing data for the following features
needed_cols = ['FIRE_SIZE_CLASS', 'DISCOVERY_DATE', 'DISCOVERY_TIME', 'CONT_DATE', 'CONT_TIME', 'STAT_CAUSE_CODE', '
df = df.dropna(subset=needed_cols) # remove rows where both of these are missing
df.shape
```

```
(892007, 13)
```

Categorical data was converted into ordinal labels.

Create new columns to work with

Map fire size class to integers so they can be ordered.

```
di = {"A": 1, "B": 2, "C": 3, "D": 4, "E": 5, "F": 6, "G": 7}
df['FIRE_SIZE_CLASS'] = df['FIRE_SIZE_CLASS'].map(di)
```

Map states to indices to help handle categorical

```
states = df.STATE.unique()
ind = states.argsort(axis=0)
state_di = {states[i]: i for i in ind}
df['STATE_CODE']=df['STATE'].map(state_di)
```

To calculate the time to containment, we re-interpreted julian dates into datetime and took the delta from time of discovery to time of containment.

Convert the date/time columns to datetime objects. Originally they are in julian time. Also calculate the time to containment (time delta of containment date - discovery date)

```
#To make these dates and times easier to manage, let's convert them to datetime. We can add new columns DISCOVERY_D
df['DISCOVERY_DATETIME'] = df['DISCOVERY_DATE'];
df['CONT_DATETIME'] = df['CONT_DATE'];
```

```
#To populate those two rows, let's convert them into datetime.
df['DISCOVERY_DATETIME'] = df['DISCOVERY_DATETIME'].apply(lambda x: julian.from_jd(x, fmt="jd"))
df['CONT_DATETIME'] = df['CONT_DATETIME'].apply(lambda x:julian.from_jd(x, fmt="jd"))

```

```
#Let's also add the time
temp_df = pd.DataFrame();
temp_df['dt'] = df['DISCOVERY_TIME'].apply(lambda x: dt.timedelta(hours=int(x[0:2]), minutes=int(x[2:5])))
df['DISCOVERY_DATETIME'] = df['DISCOVERY_DATETIME'] + temp_df['dt']
df['DISCOVERY_DATETIME'].head()
```

```
0   2005-02-02 13:00:00
1   2004-05-12 08:45:00
2   2004-05-31 19:21:00
3   2004-06-28 16:00:00
4   2004-06-28 16:00:00
Name: DISCOVERY_DATETIME, dtype: datetime64[ns]
```

```
#Do the same thing for CONT_DATETIME
temp_df = pd.DataFrame();
temp_df['dt'] = df['CONT_TIME'].apply(lambda x: dt.timedelta(hours=int(x[0:2]), minutes=int(x[2:5])))
df['CONT_DATETIME'] = df['CONT_DATETIME'] + temp_df['dt']
df['CONT_DATETIME'].head()
```

```
0   2005-02-02 17:30:00
1   2004-05-12 15:30:00
2   2004-05-31 20:24:00
3   2004-07-03 14:00:00
4   2004-07-03 12:00:00
Name: CONT_DATETIME, dtype: datetime64[ns]
```

```
df['CONT_DATETIME'] = df['CONT_DATETIME'].apply(lambda x: dt.datetime.strptime(x, '%Y-%m-%d %H:%M:%S'))
df['DISCOVERY_DATETIME'] = df['DISCOVERY_DATETIME'].apply(lambda x: dt.datetime.strptime(x, '%Y-%m-%d %H:%M:%S'))
df['TIME_TO_CONT'] = df['CONT_DATETIME'] - df['DISCOVERY_DATETIME']

df['TIME_TO_CONT'].describe()
```

```
count                    892007
mean      1 days 06:44:52.524004
std      13 days 19:33:02.720135
min             0 days 00:00:00
25%             0 days 00:30:00
50%             0 days 01:28:00
75%             0 days 04:45:00
max          3653 days 01:30:00
Name: TIME_TO_CONT, dtype: object
```

To ensure that the dataset used was consistent, our processed data was saved as a .pkl and reloaded on demand.

IV. Results

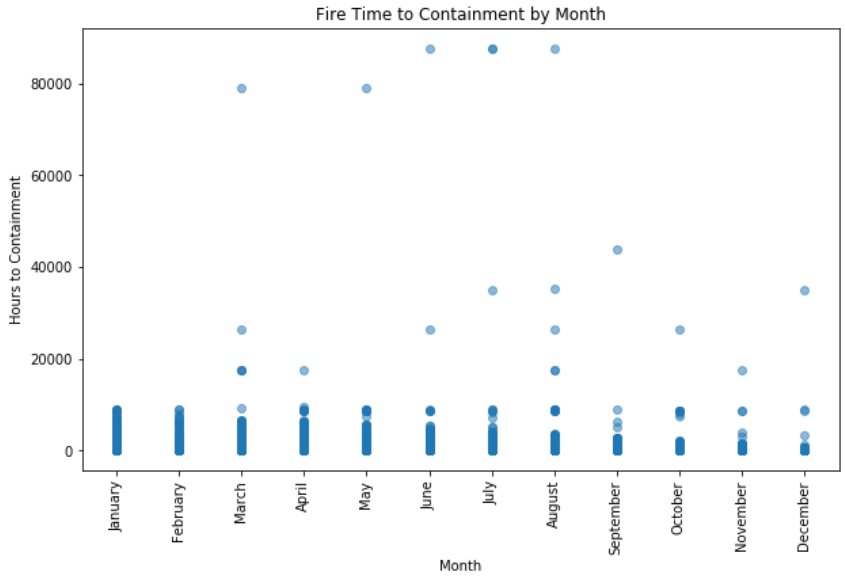
Feature Analysis

As described in the Methods section, we started exploratory work by further examining some of the individual features in the dataset. Out of the features available, three columns of categorical features intuitively stood out as possible fire severity indicators-- the U.S. state in which the fire originated, the month of the year that the fire was discovered, and the Fire Cause Code. Many of these results may seem self-explanatory: the longest fire time to containment, for example, was intuitively discovered in the months of June-August (Fig 2A). There are some elements which could be surprising, however. For example, it was surprising to find any fires that could take multiple years (over 8760 hours) to contain. Without more information, it is hard to determine whether this reporting is valid, as fire times this long are quite rare and far beyond the 3rd quartile of data (Fig. 3).

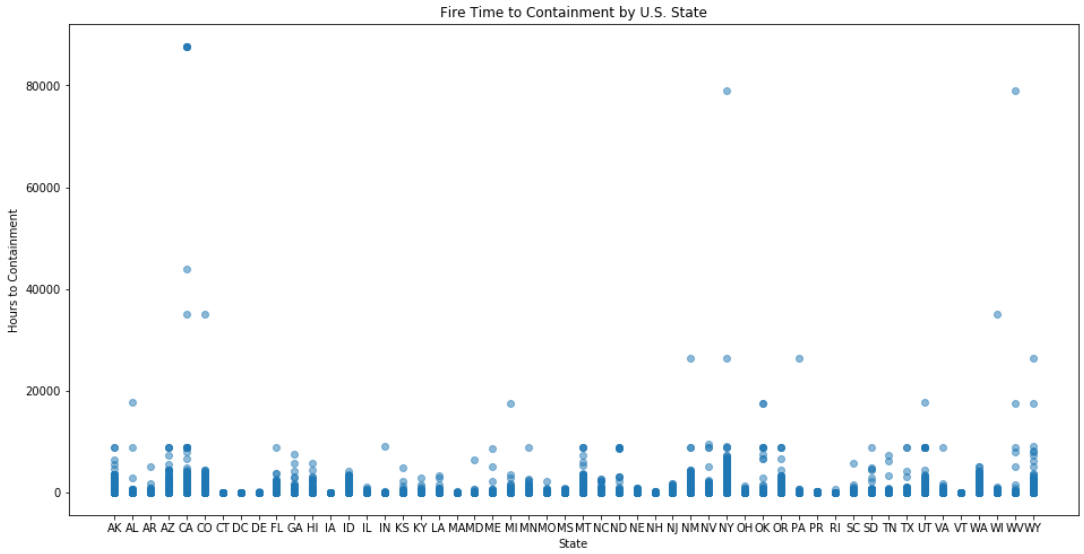
Given the relatively small number of instances with fire containment times over a year, and the lack of precedent for fires longer than 5 months in duration (9), we chose to treat those rows as outliers and exclude them from analysis, removing any rows with that had a discovery to containment delta of 3000 hours (125 days) or more. Doing so still left us with 891101 rows, which we used to recreate the scatter plots of hours to containment (Fig.4).

Fig. 2: Scatter plots of hours to containment

2A: Hours to containment organized by month of the year. Each point represents one fire.



2B: Hours to containment organized by U.S. State, including Puerto Rico and District of Columbia. Each point represents one fire.



2C: Hours to containment organized by fire cause code. Each point represents one fire.

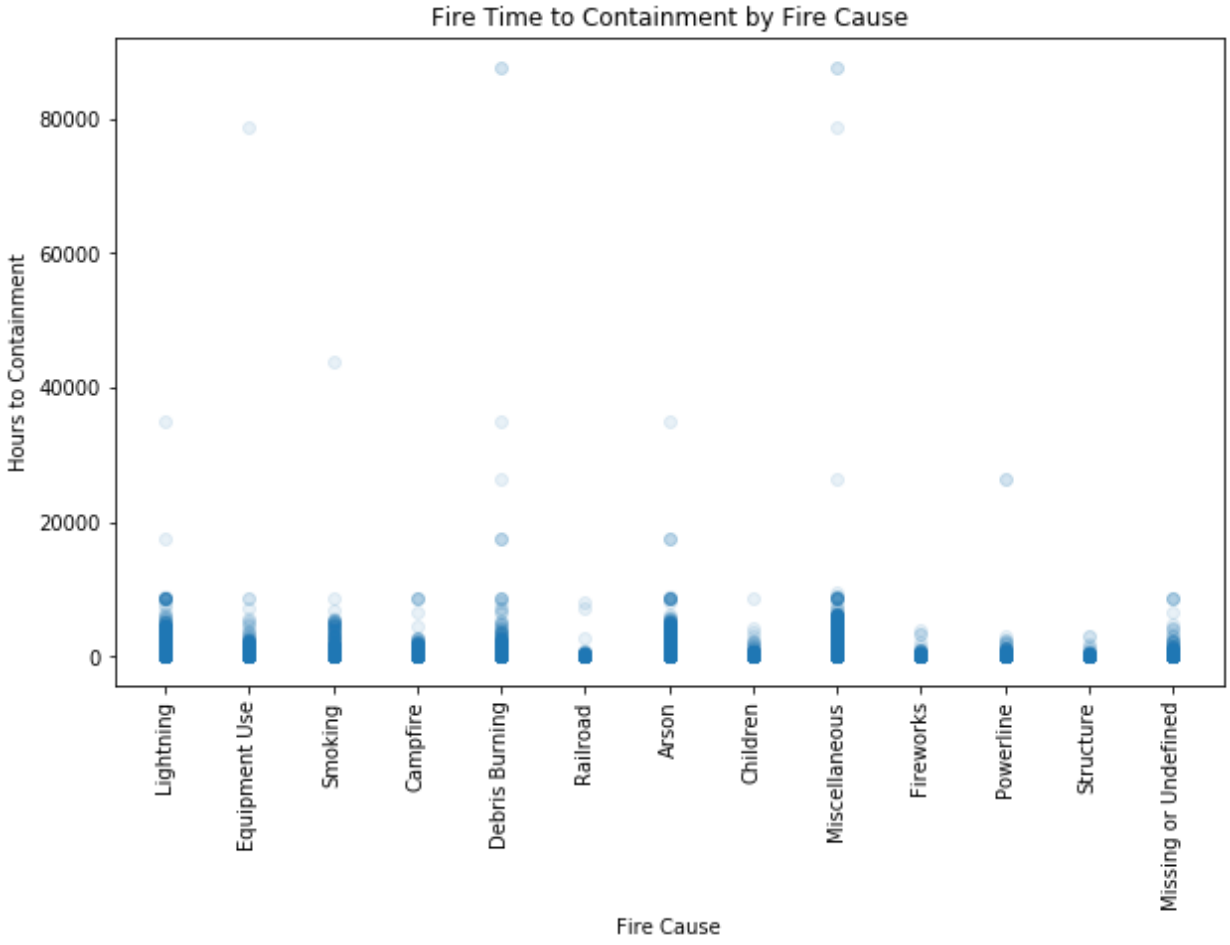
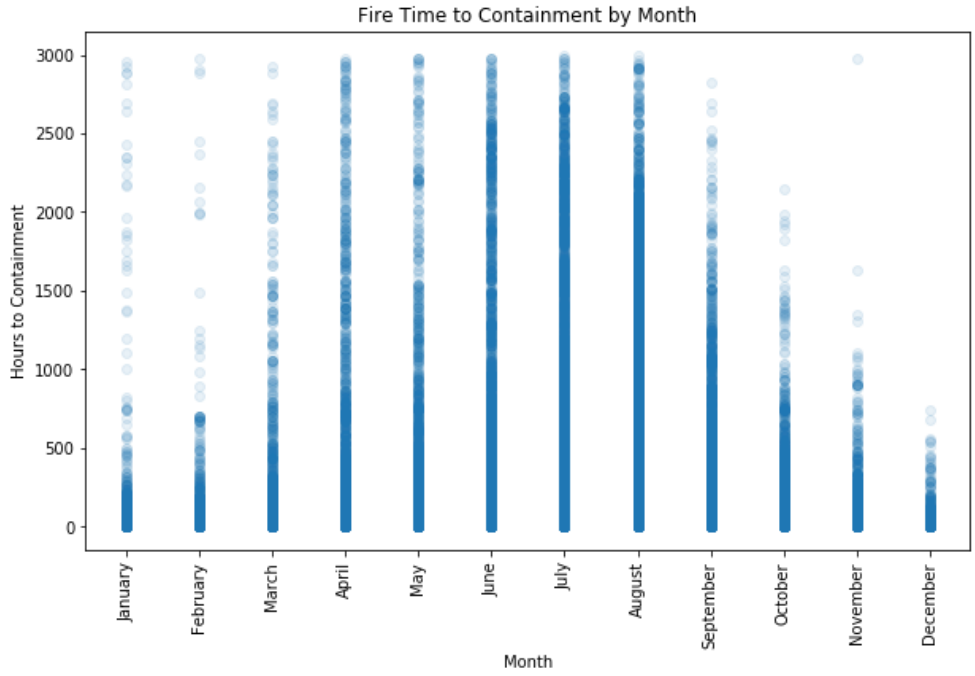


Fig. 3: Time to Containment metrics.

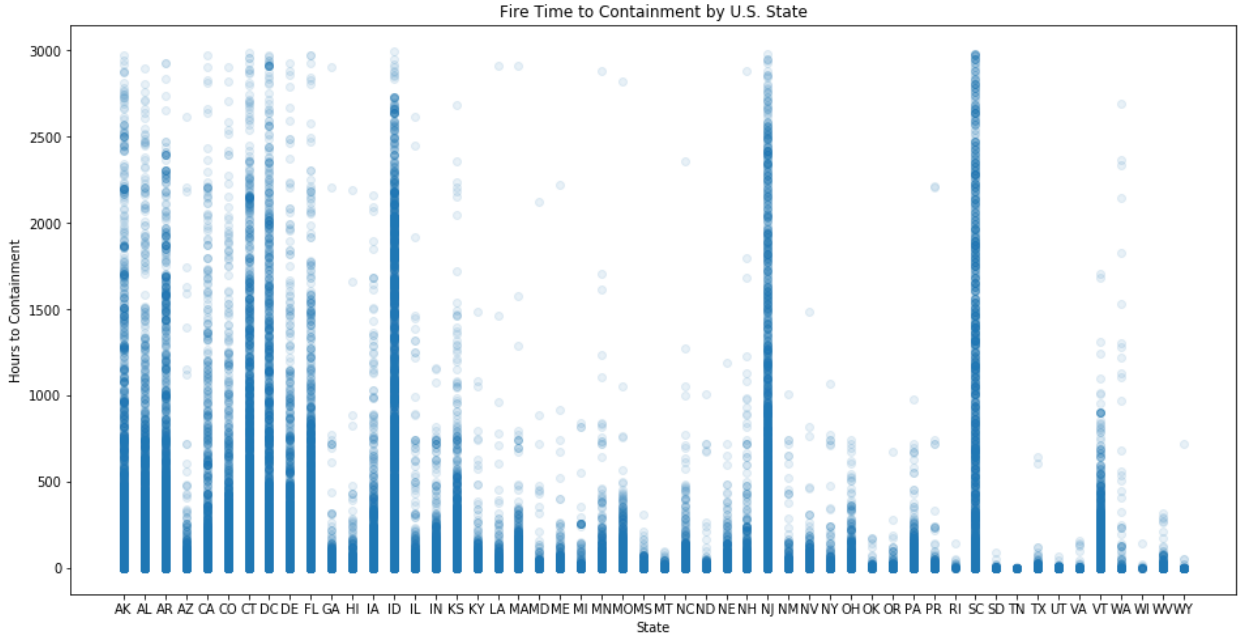
Count	892007
Mean	1 days 22:18:16.178348
Standard Deviation	13 days 19:29:31.562665
Min	0 days 00:01:00
25%	0 days 15:24:00
50%	0 days 18:40:00
75%	0 days 23:59:00
Max	3653 days 18:45:00

Fig. 4: Scatter plots of hours to containment with time to containment 125 days (3000 hours) and over removed.

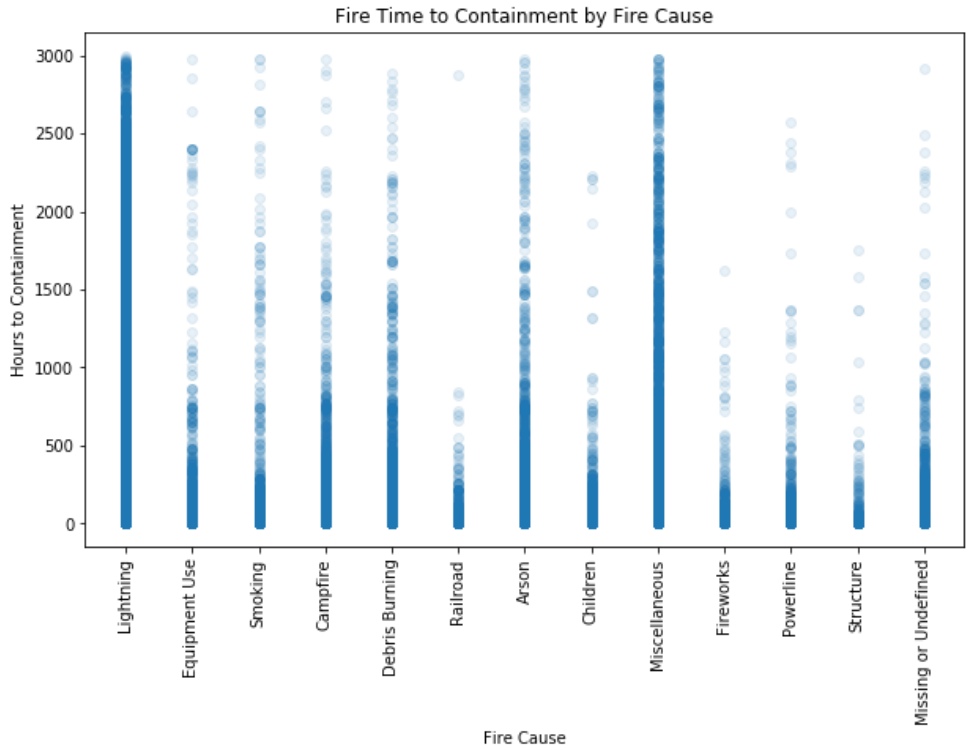
4A. : Hours to containment organized by month of the year. Each point represents one fire.



4B. Hours to containment organized by U.S. State, including Puerto Rico and District of Columbia. Each point represents one fire.



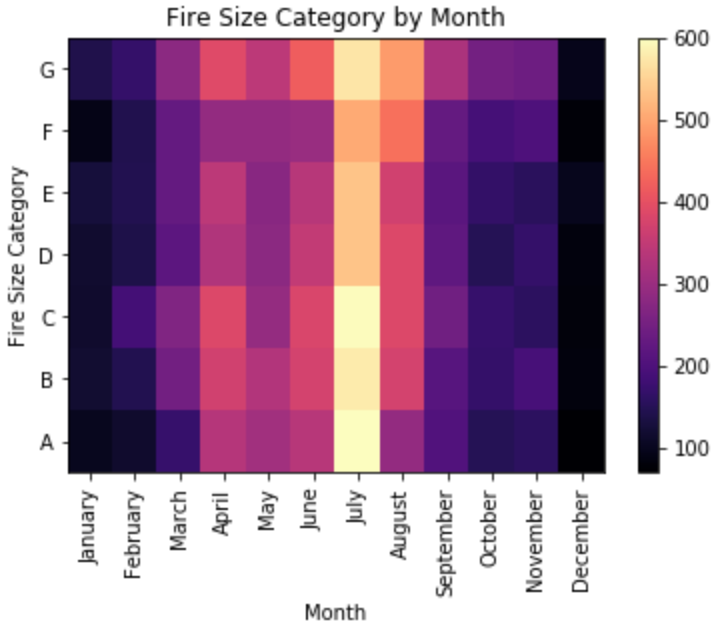
4C. Hours to containment organized by fire cause code. Each point represents one fire.



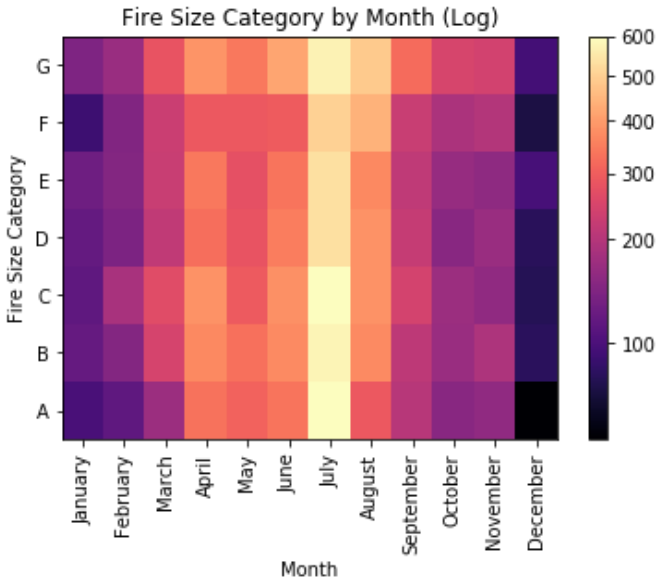
As another preliminary analysis, we repeated this process with our alternate measure of severity, Fire Size Class.

Fig. 5: Heatmap of Fire Size Categories

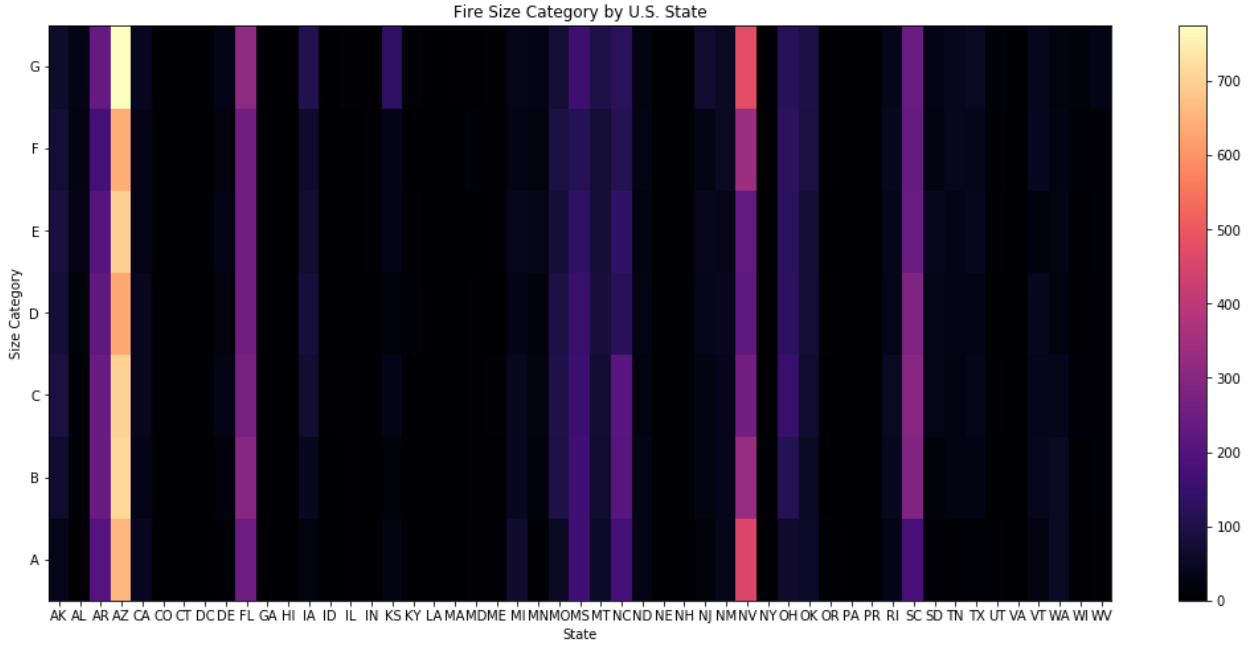
5A: Fire size category organized by month. Colors represent fire frequency and are interpolated on a linear scale.



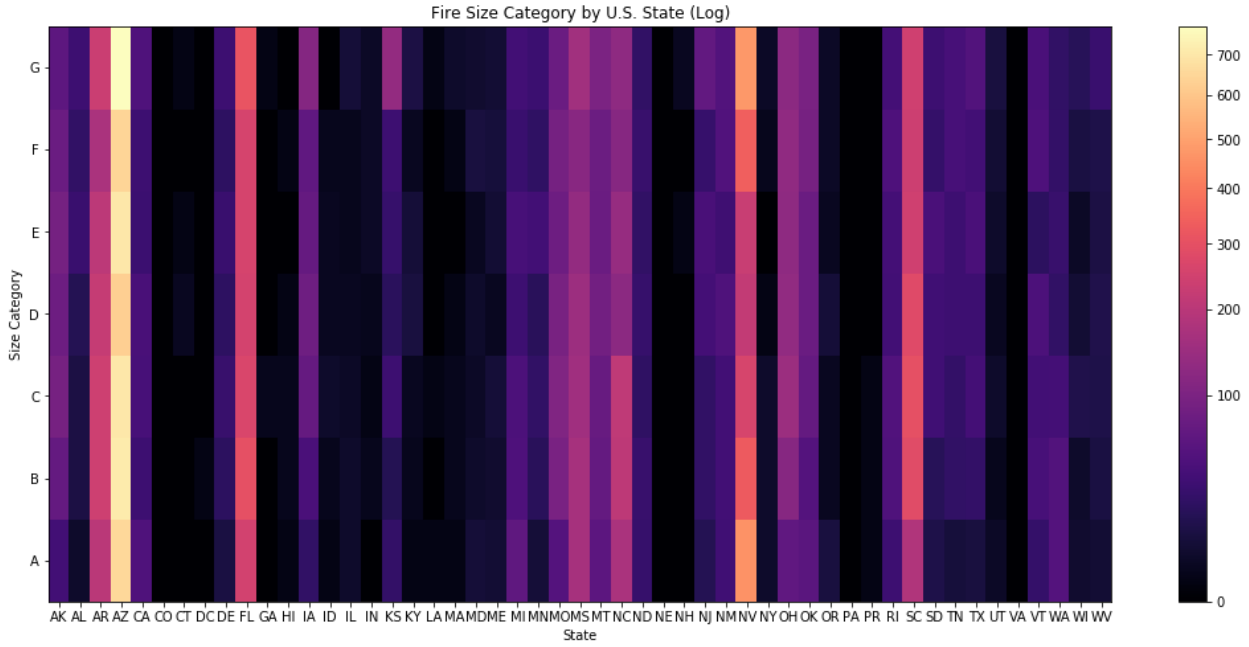
5B: Fire size category organized by month. Colors represent fire frequency and are interpolated on a logarithmic scale.



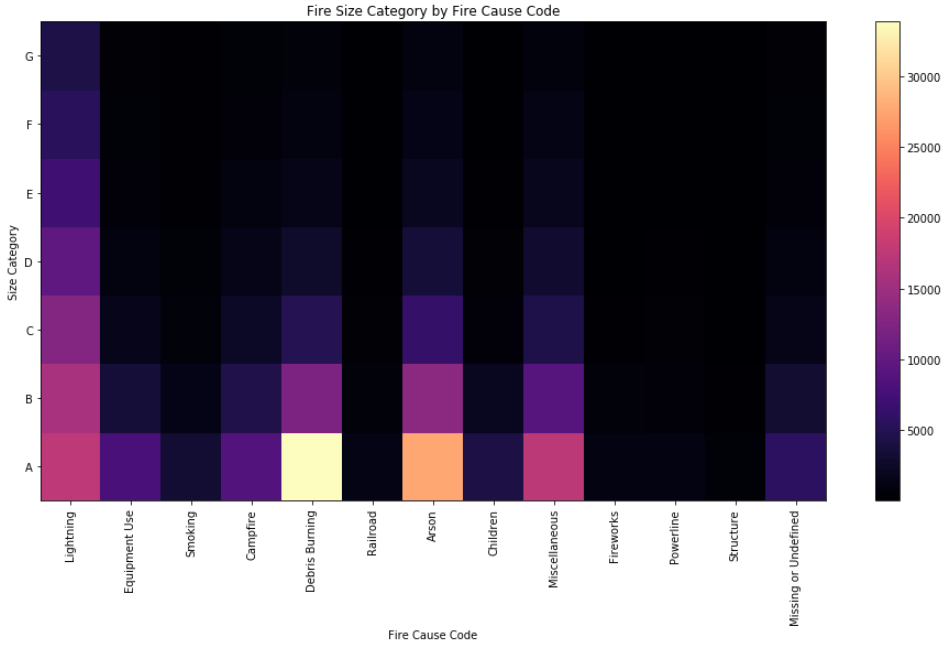
5C: Fire size category organized by U.S. State. Colors represent fire frequency and are interpolated on a linear scale.



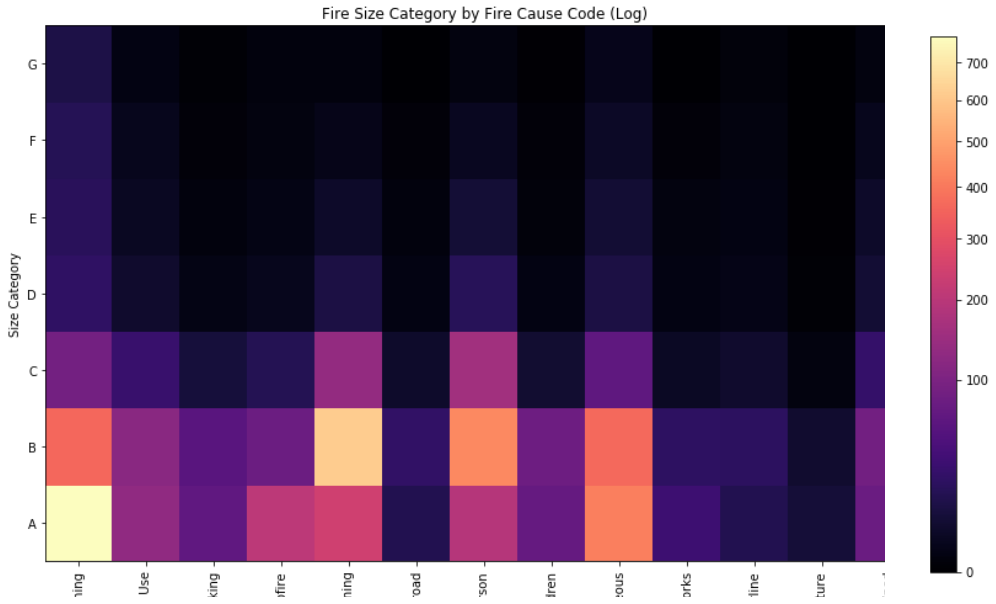
5D: Fire size category organized by U.S. State. Colors represent fire frequency and are interpolated on a log scale.



5E: Fire size category organized by Fire Cause Code. Colors represent fire frequency and are interpolated on a linear scale.



5F: Fire size category organized by Fire Cause Code. Colors represent fire frequency and are interpolated on a log scale.



Both targets appeared to have visible trends with the single input features of Fire Cause Code, Month, and State. The hours to containment correlated visually with the Fire Size Category for Fire Cause Code and Month, but did not with State.

Models

Linear Regression

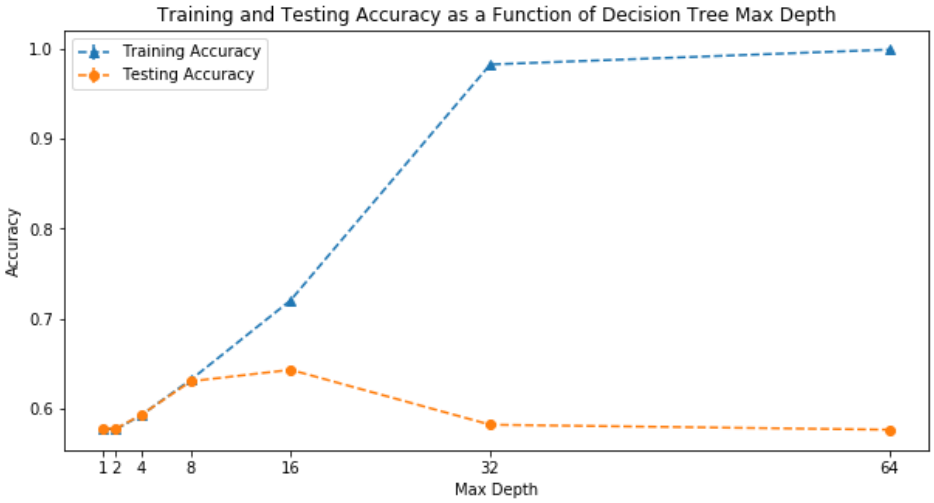
As a first pass at classification of fire severity based on the Hours to Containment target, a Linear Regression was modeled using all of the features. This predictor provided a solid baseline to indicate that we do not have sufficient feature data to predict on this target. Across 5 folds, with 33% of the data reserved for testing, training R^2 measured at a mean of 0.101 +/= .008 and testing R^2 at .102 +/= .001.

Decision Tree

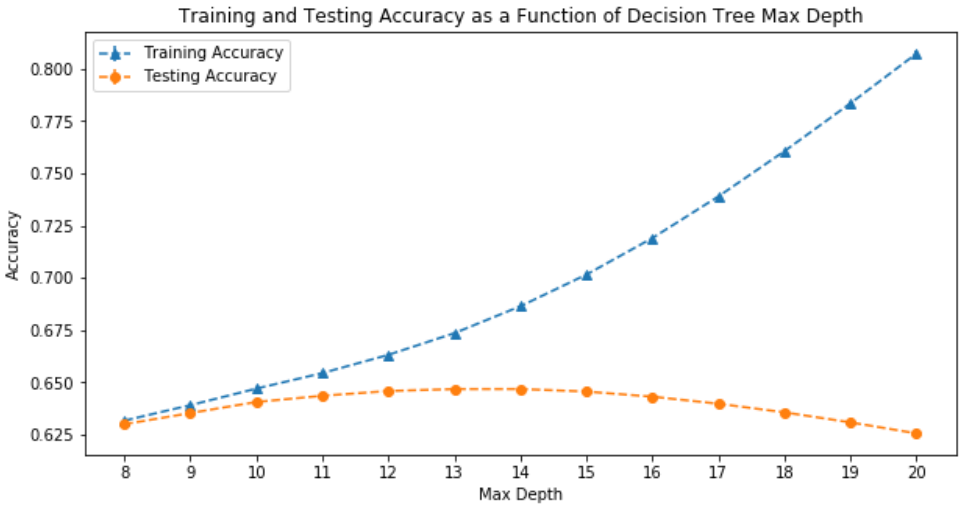
To begin classification of fires based on the Fire Size Category target, a Decision Tree was trained on the features and labels, using 5-fold cross-validation and a testing set of 33%. As the first tuning parameter, we varied the maximum depth for the tree amongst a set of seven values: [1, 2, 4, 8, 16, 32, 64]. We can see the training and testing accuracy increase fairly tightly in step from depths 1-8, at which point the model starts to overfit, leading to a clear diversion at depth 16.

Fig. 6: Comparison between Training and Testing accuracies as a function of max tree depth. Standard deviations between -fold validations are plotted, but so small as to be negligible.

6A: Decision tree classification accuracy created using max depths [1, 2, 4, 8, 16, 32, 64]



6B: Decision tree classification accuracy creating using max depths [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]

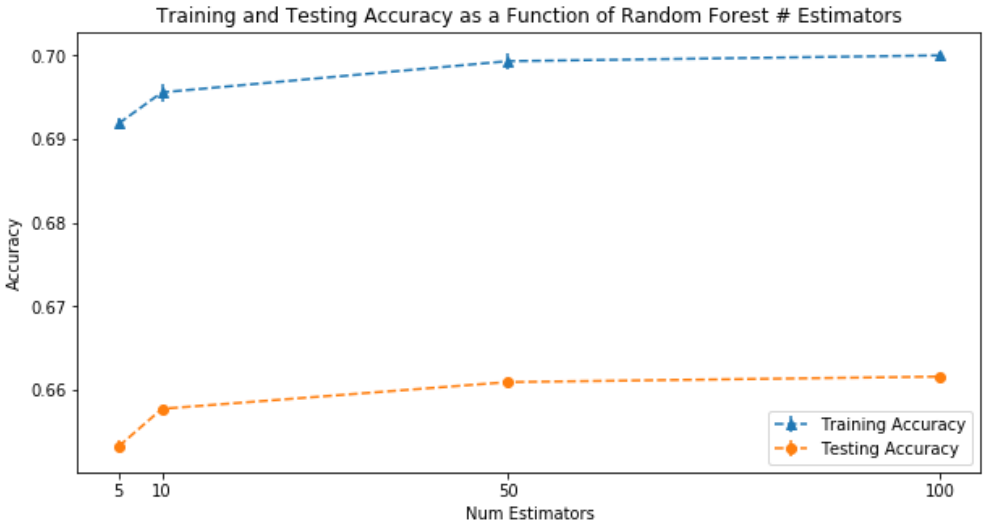


From the decision tree model, we were able to produce a maximum testing accuracy at depth 14 of around .646 +/= .0005.

Random Forest

We also trained a Random Forest on the Fire Size Category target, first doing a rough tuning by number of estimators (number of estimators in the forest). Again, each condition was repeated across 5 folds with a test proportion of .33.

Fig. 7: Random Forest with maximum depth 14 and variable numbers of estimators.



Our maximum testing accuracy for the Random Forest measured 0.67 +/- .0008, with a corresponding training accuracy of .70 +/- .0003, our highest accuracy thus far.

V. Discussion

Our Decision Tree and Random Forest models performed above chance when classifying each instance to a corresponding Fire Size Category. To investigate our overall results, we used an output for the Random Forest classifier with maximum depth 14 and 50 estimators. We can tell from examining both training and testing datasets that the model is overpredicting our most

common classes (fire size classes A and B, the less severe classes) and under-predicting the more severe classes. (Fig. 8). As one might expect, the accuracy of a true prediction for each of the size classes relative to the total true labels for each size class is higher for size classes A and B (Fig. 9), which have more instances to estimate with.

It is interesting to note in Fig. 8 that the more extreme fire size classes (A,B and G) are more accurately classified than the middle classes. It could be that the medium fire size classes have no defining attributes described in our dataset, and are therefore difficult to predict. These middle classes (D, E, F) are also much less numerous than the smaller classes (A, B, C), and are only slightly more numerous than G.

The features we found most important in our model were longitude, latitude, and hours to containment. Longitude and latitude are likely important because they relate to the climate, and certain longitude and latitude values can act as a proxy for describing the type of weather common to a location. Hours to containment is also a very reasonable predictor for fire size, and is positively correlated with fire size class as the longer it takes to contain a fire, the more time it has to spread and cover more area.

Fig. 8: Frequency of each class predicted by Random Forest classifier.

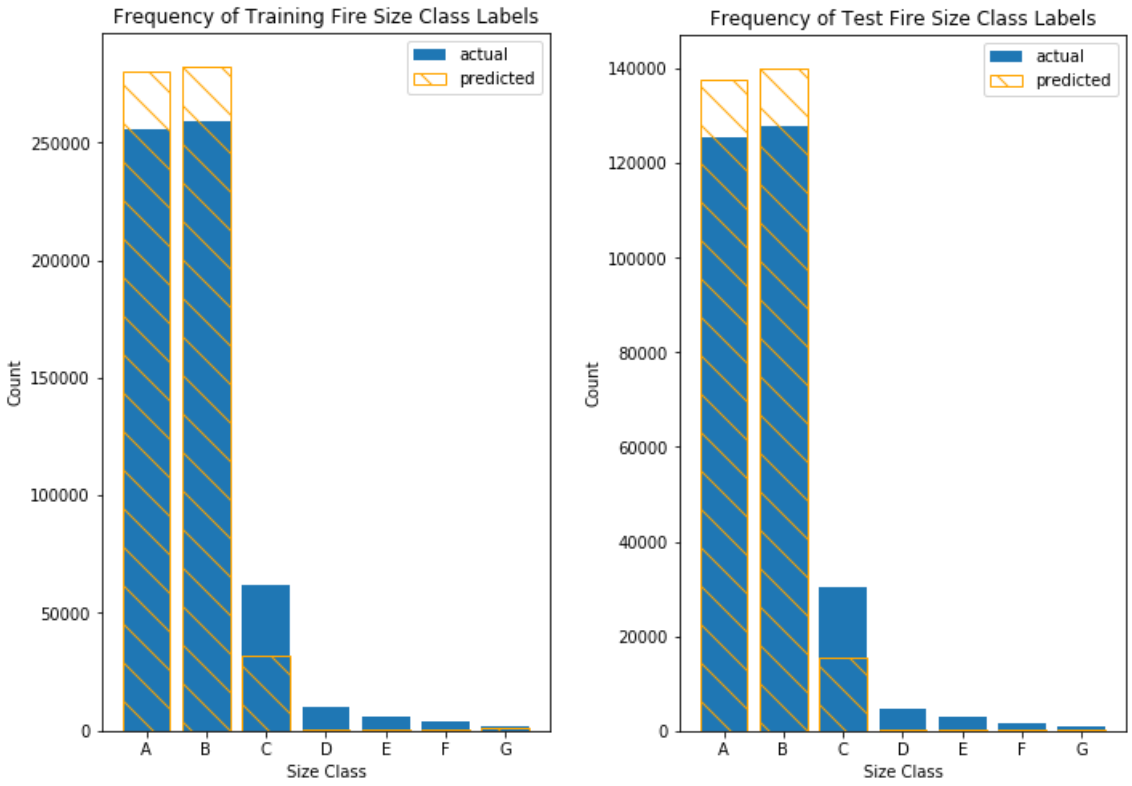
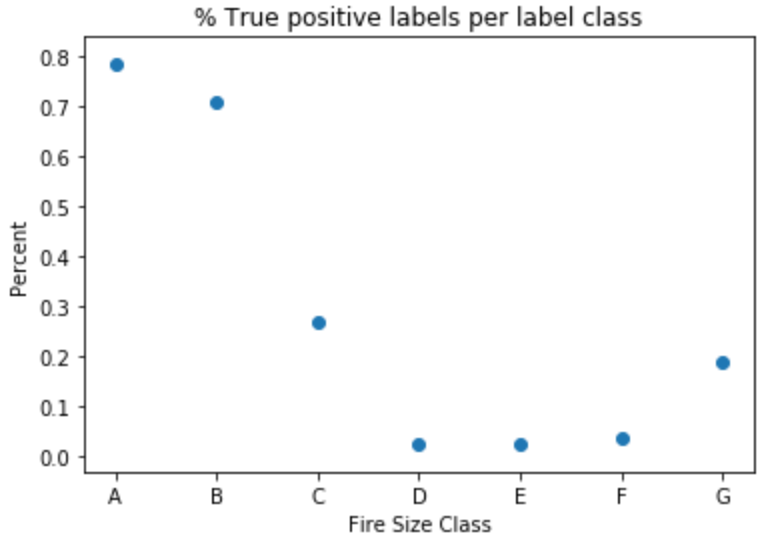


Fig. 8: Percentage of correctly classified size classes over total count of each size class



VI. Conclusions / Future Work

In conclusion, this project aimed to analyze the time delta between discovery and containment time to predict the fire class size. In initial attempts to classify fire severity based on our Hours to Containment target, a Linear Regression model indicated that we didn't have sufficient feature data to predict on said target. Across 5 folds with a testing size of 33%, the training R^2 measured at a mean of 0.101 +/= .008 and testing R^2 at .102 +/= .001.

We then used a Decision Tree to model the data as it'd allow for the reasoning for the predictions to be deduced, and could thus help in identifying why a potential wildfire might be severe. The Decision Tree was trained on the features and labels, using 5-fold cross-validation and a testing size of 33%. The training and testing accuracy increased fairly tightly in step from depths 1-8, at which point the model started to overfit, resulting in a clear diversion at depth 16. The Decision Tree model produced a maximum testing accuracy at depth 14 of around .646 +/= .0005. We also found out via Gini criterion that our engineered feature, hours to containment, had a high feature importance.

Next using the Decision Tree model as a basis, we trained a Random Forest on the Fire Size Category target. We produced a maximum testing accuracy for the Random Forest of 0.67 +/- .0008, with a corresponding training accuracy of .70 +/- .0003, our highest accuracy thus far.

The features we found most important in our model were longitude, latitude, and hours to containment. As mentioned earlier, longitude and latitude are likely significant as they relate to the climate and have the potential to serve as a representation of the type of weather common to a certain location. Hours to Containment is also a fairly robust predictor for fire size as it's positively correlated with Fire Size Class given that a larger Hours to Containment value typically means the fire has had time to cover more ground.

Multiple changes could be made in order to amplify the results we obtained. Model accuracy could be improved by finding and adding more significant features. Some additional features that

could be added include weather conditions, such as temperature, precipitation, humidity, etc. An obvious correlation we could draw would be that high temperatures and low humidity would result in a higher likelihood for the occurrence of both more numerous and more severe wildfires. We could also incorporate the already utilized longitude and latitude features to add the functionality of location to more accurately describe the climate in the area with the fire.

In addition, the switching of fire size class labels to "low-med-high" danger levels could potentially simplify multi-class predictions, as our model faced difficulty when predicting the middle fire size classes. The low danger level could include A, B, C - the med D, E, F - and the large G.

Further usage of more complex supervised learning techniques for example Artificial Neural Networks and Support Vector Machines could provide more advanced insight in predicting the occurrence of wildfires.

VII. References

1. CNBC: At least 33 dead as wildfires scorch millions of acres across Western U.S.-- 'It is apocalyptic' *https://www.cnbc.com/2020/09/12/fires-in-oregon-california-and-washington-spread-death-toll-rises.html*

2. CNN: California sets new record for land torched by wildfires as 224 people escape by air from a 'hellish' inferno *https://www.cnn.com/2020/09/05/us/california-mammoth-pool-reservoir-camp-fire/index.html*

3. National Large Incident Year-to-Date Report *https://gacc.nifc.gov/sacc/predictive/intelligence/NationalLargeIncidentYTDReport.pdf*

4. MSN: 'I never could have envisioned this': At least 35 dead as nearly 100 wildfires rage across 12 Western states *'https://www.msn.com/en-us/news/us/i-could-never-have-envisioned-this-at-least-35-dead-as-nearly-100-wildfires-continue-to-rage-across-12-western-states/ar-BB18ZJTe?__hstc=26100450.ed71421e05751a5b16ce163231ca6993.1474329600098.1474329600100.1474329600101.2&__hssc=26100450.1.1474329600101&__hsfp=2025384311*

5. NYT: Historic Wildfires Rage in Western States
https://www.nytimes.com/article/wildfires-photos-california-oregon-washington-state.html

6. Kaggle: 188 Million US Wildfires
https://www.kaggle.com/rtatman/188-million-us-wildfires

7. Westerling, Anthony L., and B. P. Bryant. "Climate change and wildfire in California." Climatic Change 87.1 (2008): 231-249.
https://www.researchgate.net/profile/A_Westerling/publication/225397044_Climate_Change_and_Wildfire_in_California/links/547fd09a0cf25b80dd7039f0/Climate-Change-and-Wildfire-in-California.pdf

8. National Wildlife Coordinating Group: Size Class of Fire
https://www.nwcg.gov/term/glossary/size-class-of-fire

9. Guinness World Records: Largest Forest Wildfire
https://www.guinnessworldrecords.com/world-records/601916-largest-forest-wildfire-single-fire

10. USDA: National Fire Danger Rating System
https://www.fs.usda.gov/detail/cibola/landmanagement/resourcemanagement/?cid=stelprdb5368839

11. National Wildfire Coordinating Group: Gaining an Understanding of the National Fire Danger Rating System
https://www.nwcg.gov/sites/default/files/products/pms932.pdf

12. Rediff: The devastating effects of the Uttarakhand fires
https://www.rediff.com/news/report/the-devastating-impact-on-glaciers-animals-due-to-uttarakhand-fires/20160503.htm