Your project is expected to have the following components at minimum:

1. Introduction/ Background
2. EDA (Exploratory Data Analysis)
3. Modeling
4. Description of Challenges/ Obstacles Faced
5. Potential Next Steps/ Future Direction

1. Introduction/Background

For years, wildfires have destroyed millions of acres of land and homes in the United States, especially in California. While some areas are more prone to fire propagation (given land cover types, vegetation health, and land surface temperature and precipitation), 90% of fires are caused by humans[1], according to the U.S. Department of Interior.

This dataset contains data on wildfires in the United States ranging from 1992 to 2015 created to support the US Fire Program Analysis. This set contains almost 2 million total wildfires and corresponding metrics such as location, date, size, cause, time to extinguish.

We first used exploratory data analysis (EDA) to better understand the relationship between each metric, both across the country and only in California. We then developed models to answer the following two questions:

1. Given available information, can we predict the cause of a fire?
2. Given available information, can we predict the time it will take for a fire to be extinguished?

2. EDA

a) Evolution of fires over time

The number of fires in the US has remained quite constant overall, with the trend slightly increasing (Figure XX). We can see peaks in 2000, 2006 and 2011. However, wildfires are a phenomenon affected by seasonality, as they occur more frequently in March, April, July and August. More interestingly, there is a massive peak in wildfire occurrence on the 4th of July, which may indicate that most of these fires are borne of holiday celebrations.

[1] https://www.iii.org/fact-statistic/facts-statistics-wildfires

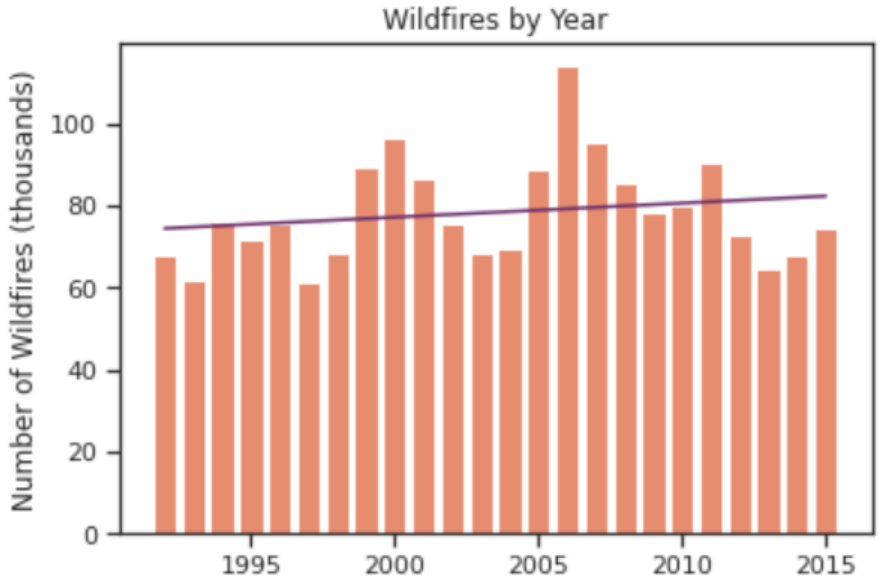


Figure XX: Number of fires per year

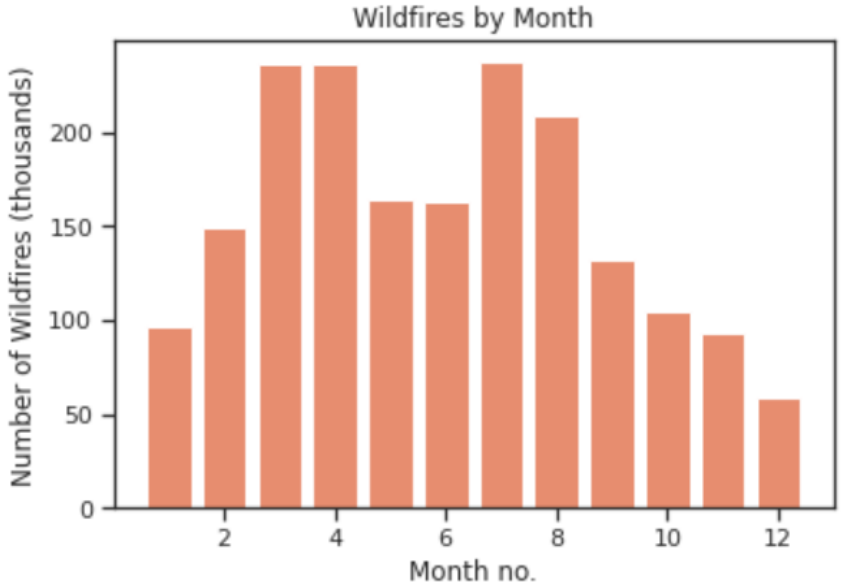


Figure XX: Total number of fires per month from 1992 - 2015

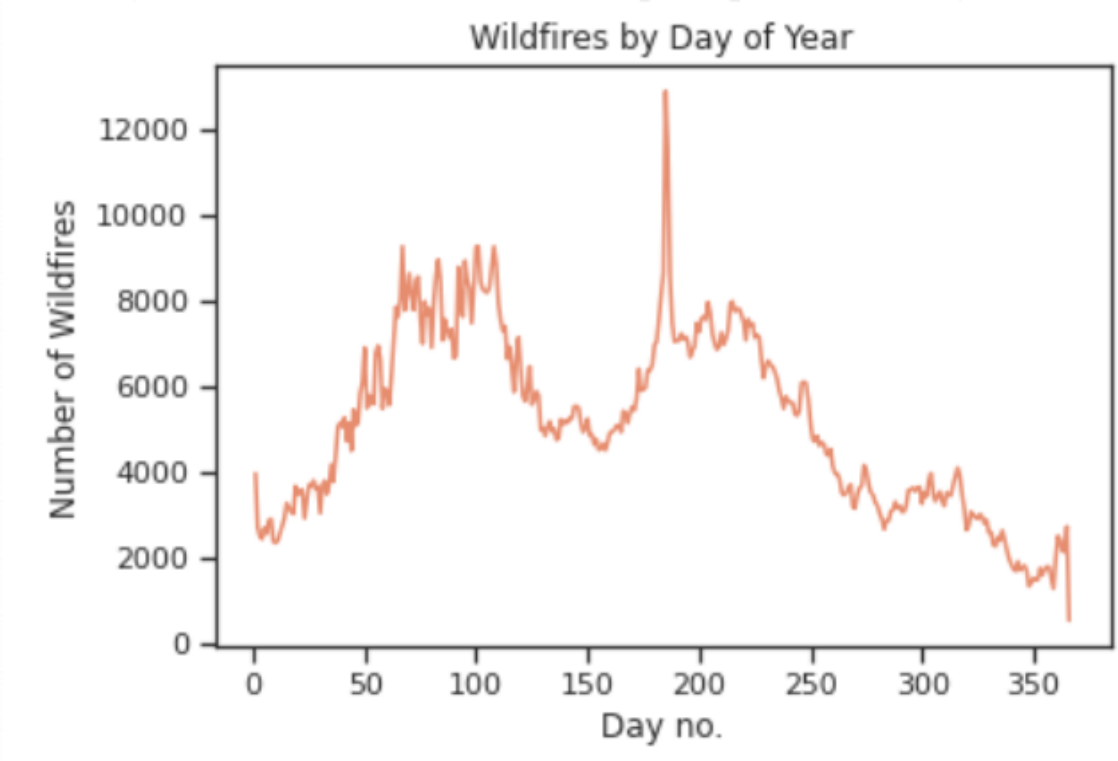


Figure XX: Total number of fires per day from 1992-2015

b) Which states have the most fires?

Some states are more affected by wildfires than others, as shown in figure XX. The top three states with the highest number of fires are California (189.55 thousands), Georgia (168.87 thousands) and Texas (142.02 thousands).

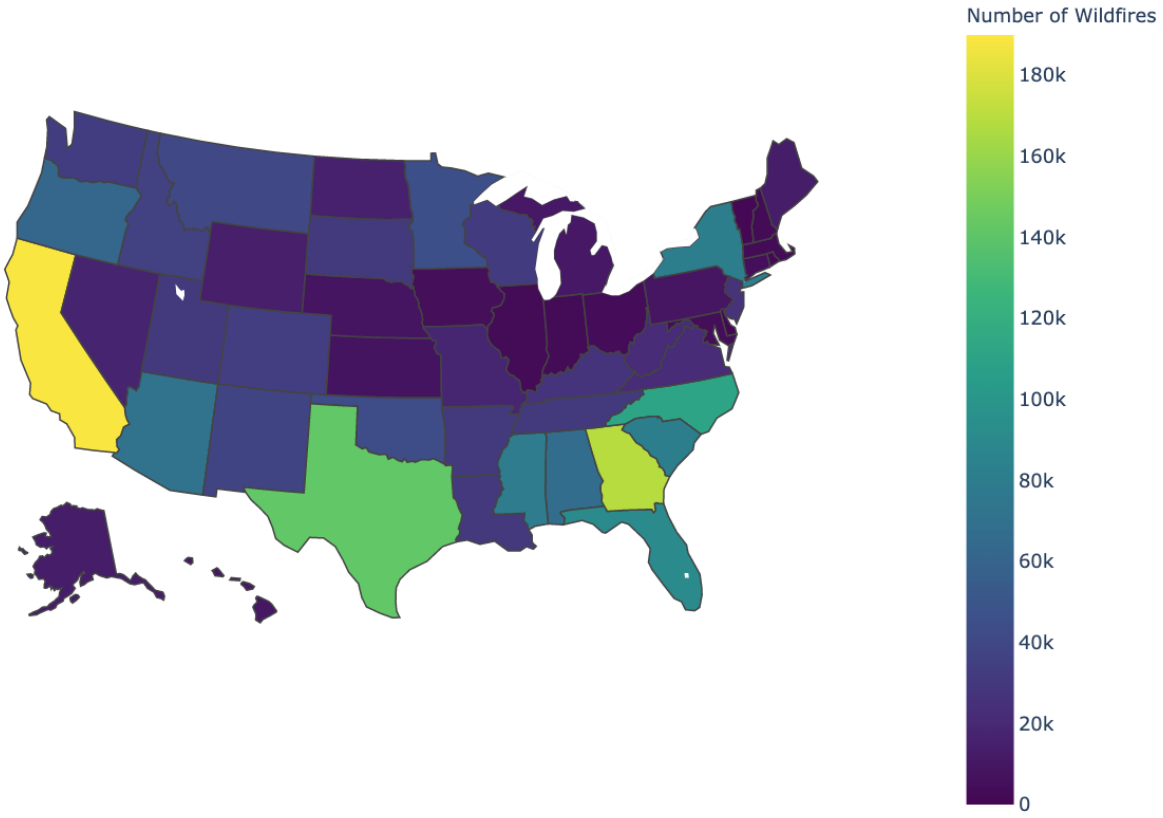


Figure XX: Total number of wildfires 1992-2015 per state

c) Main causes of wildfires

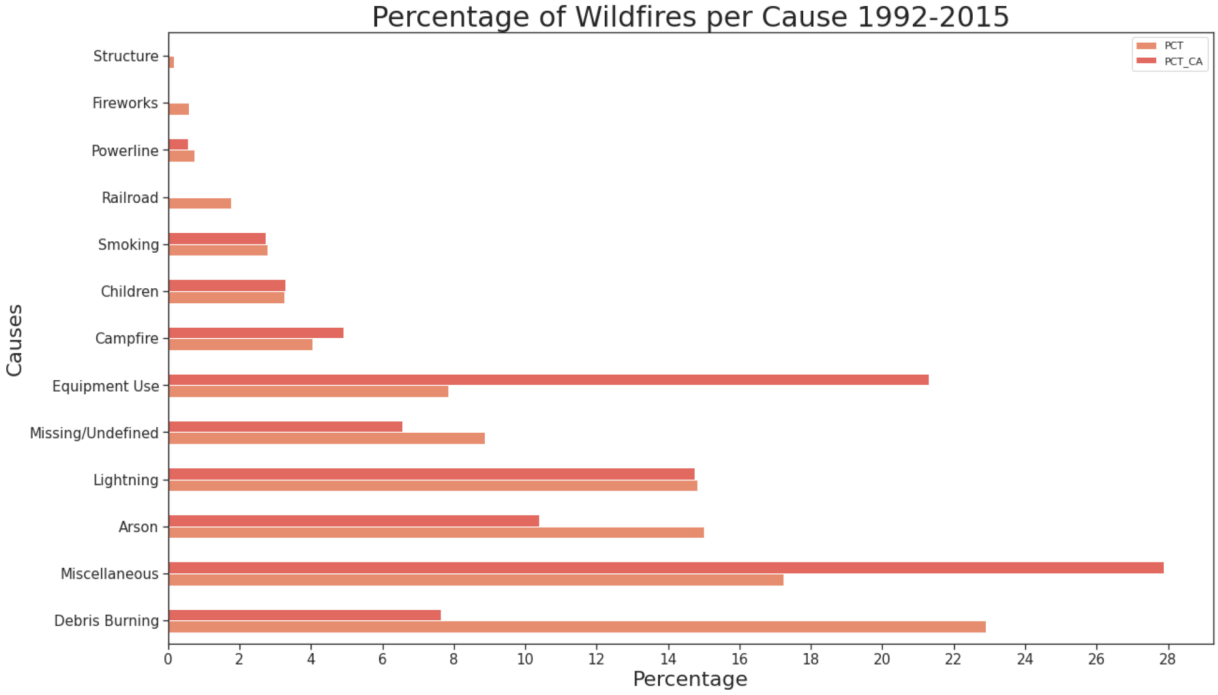


Figure XX:

We can see that Debris Burning is the most common cause of wildfires countrywide however "Miscellaneous" is the most common in California. "Lightning" is the only natural cause and only represents 15% of wildfires both across the country and in California, which means that 85% of all fires were caused by humans. This data validates the U.S. Department of Interior's claim that 90% of fires are caused by humans[2]. In addition, "Arson" is the third leading cause of wildfires at 15% country-wide and 10% in California.

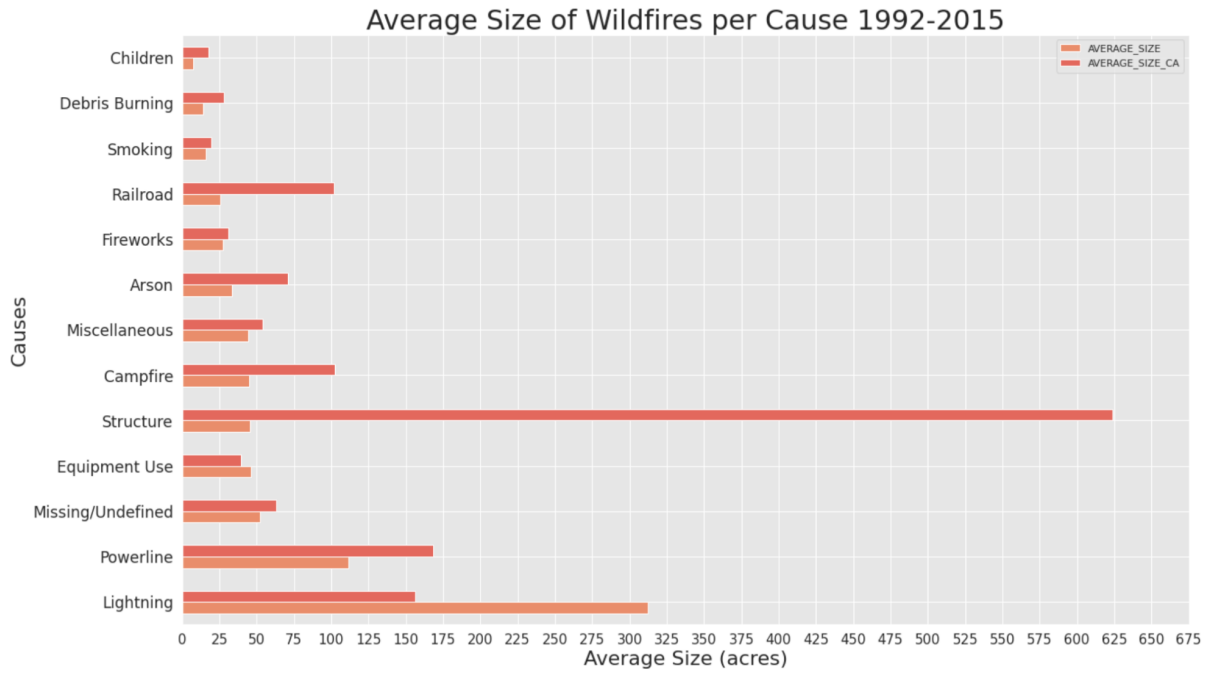


Figure XX: Average size of wildfires per cause countrywide and only in California

Around the US, fires caused by Lightning (natural cause) are the largest, with an average of 310 acres, followed by power lines at an average of 110 acres. In California however, Structure caused fires (accidents) burn 625 acres of land on average.

d) Length of fires

[2] https://www.iii.org/fact-statistic/facts-statistics-wildfires

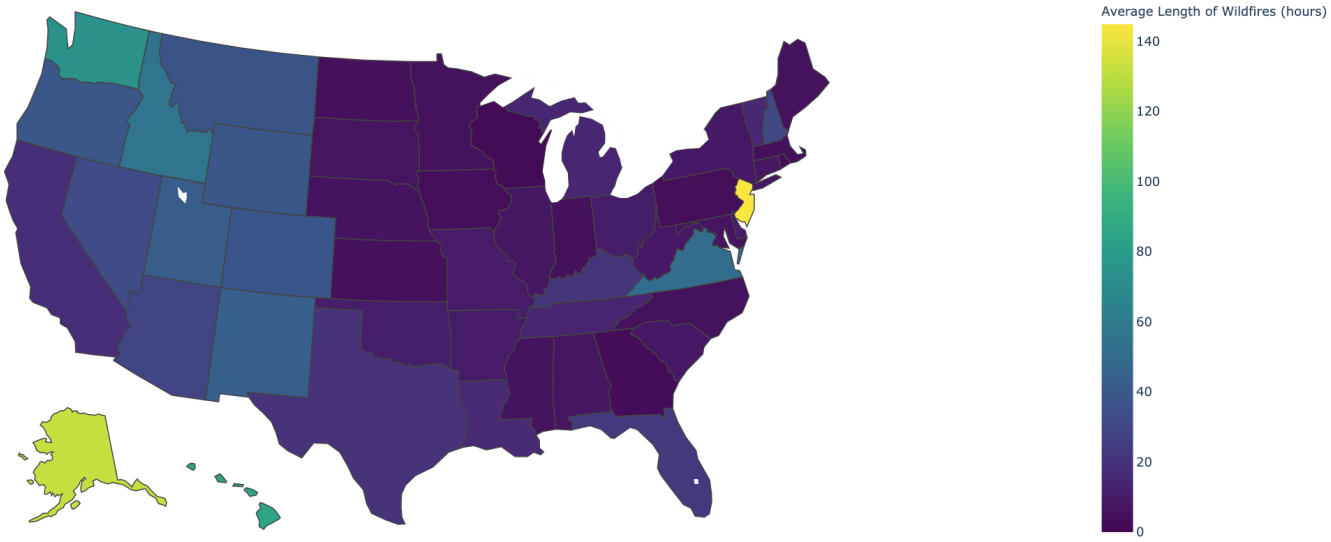


Figure XX: Average Length of Wildfires 1992-2015 per state

From this plot, we can see New Jersey wildfires last the longest at an average of 144 hours between discovery and containment time. The discrepancy between NJ and its neighboring states' burn time (PA has an average of 3.2 hours) may be due to outliers or poorly entered data.

e) Size of fires

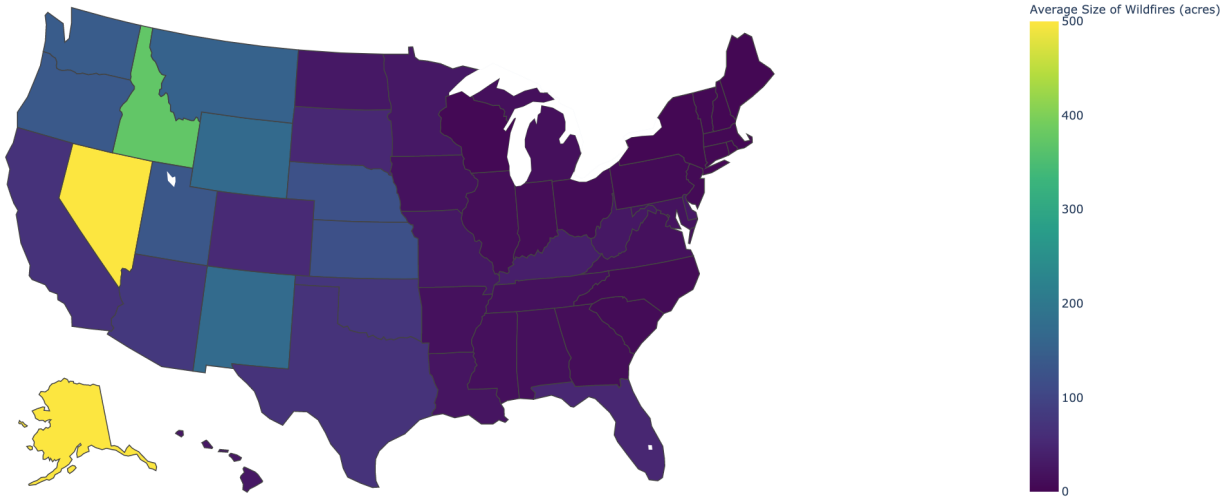
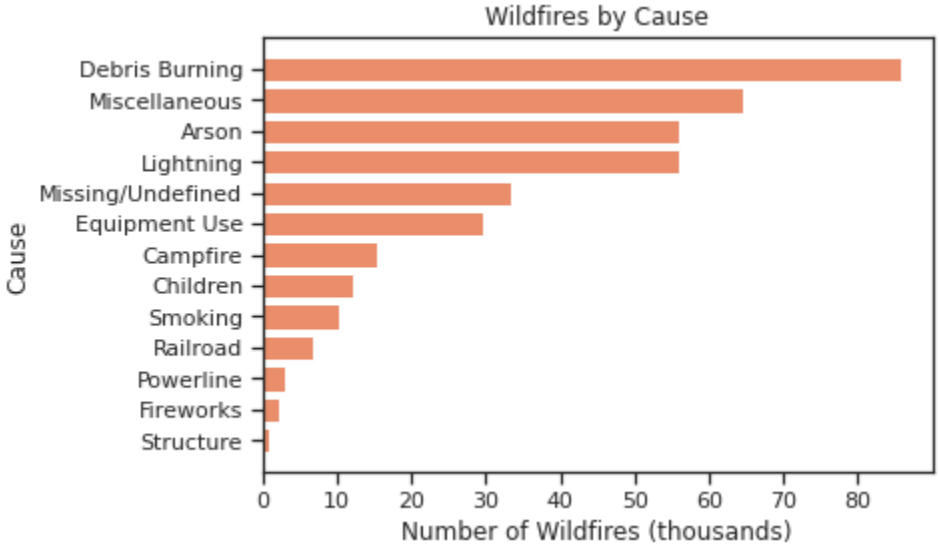


Figure XX: Average Size of Wildfires 1992-2015 per state

Alaska's wildfire size far surpasses any other states at an average of 2500 acres. That is not surprising given it is the US's largest state. However, California and Texas only have average sizes of 68 acres (even though they have the 1st and 3rd highest counts of wildfires).

3. Modeling
    a) Predicting Wildfire Causes

Wildfires have thirteen causes as identified in our exploratory analysis. The largest portion of these are taken up by Debris Burning, as we can see in a quick summary of our chart from above.



For this classification task, we opted to use decision trees, since their explanatory nature lends itself well to our end goal.

Preprocessing
In order to preprocess the data, we first need to narrow our dataset down to usable columns. For this predictive task, we settled on the following:
Fire Year, Fire Month, Fire Day of Year, Fire Size, Latitude, and Longitude

We debated using state or county data, but since these would need to be one-hot-encoded and would greatly increase the size of our dataset, we opted instead to use coordinates. Their values as floats make them ideal to encode geographic information.

Model 1: Benchmarking
We can establish a baseline to make sure that our fancy models are actually doing something useful

The simplest benchmark to use will just be to guess the most common cause: Debris Burning.

Accuracy: 25.0%.

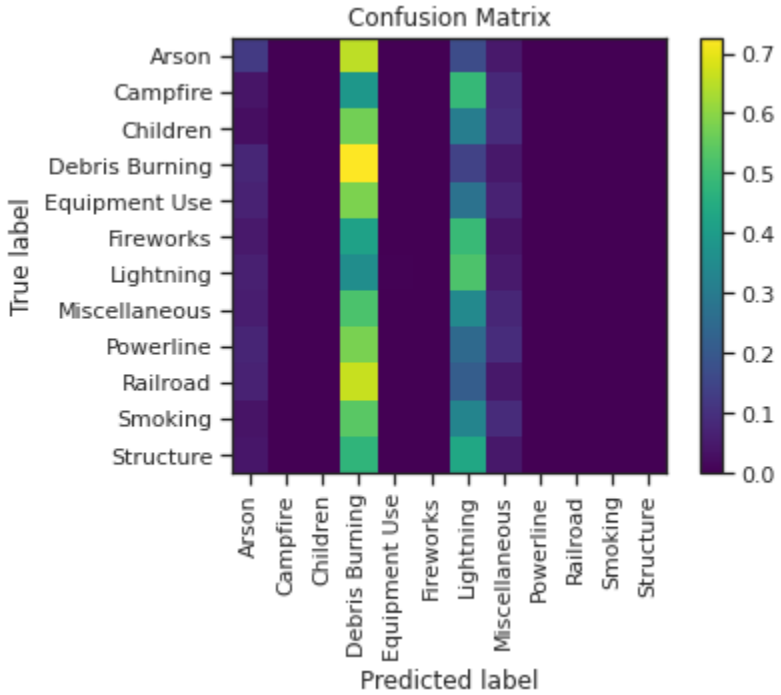
As long as we can beat this accuracy, we will have a useful model.

Model 2: Simple Decision Tree
Let's use a single explanatory feature (Fire Size) to build out a very simple model.

Accuracy: 30.2%

It seems that by just using the size of a fire, we can improve our prediction by just a bit. However, our score still isn't very good.

We can look at a confusion matrix to see what predictions our model is making. A perfect model would have a rating of 1.0 on a downward-sloping diagonal line and 0.0 everywhere else.

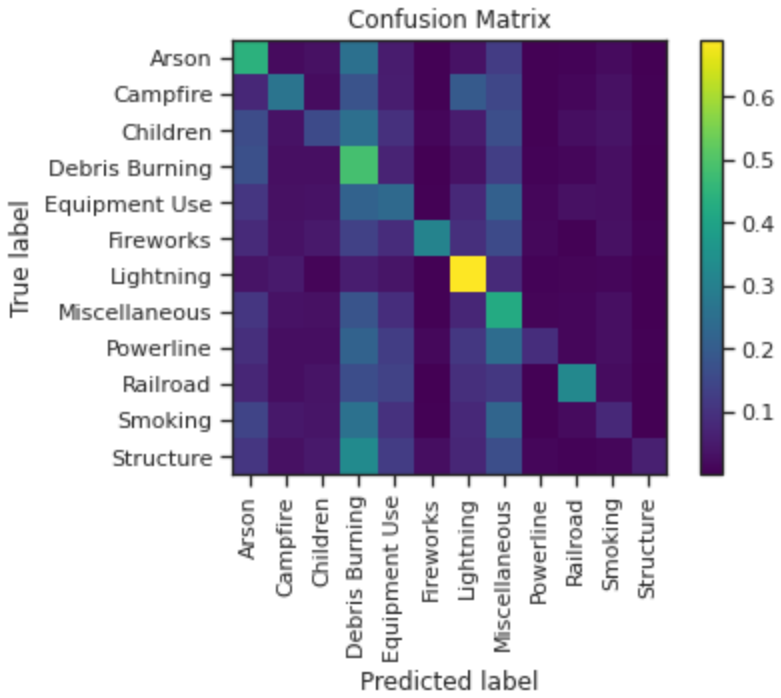


It looks like we are predicting Debris Burning and Lightning for nearly every case. Since we discovered during our EDA that these two are highly dependent on fire size, this makes sense.

Model 3: Add more features
Let's add in the rest of our seemingly relevant features into our tree model to see how it improves our prediction accuracy.

Accuracy: 43.6%

This is a great improvement on our previous models! Let's take a look at our new confusion matrix.

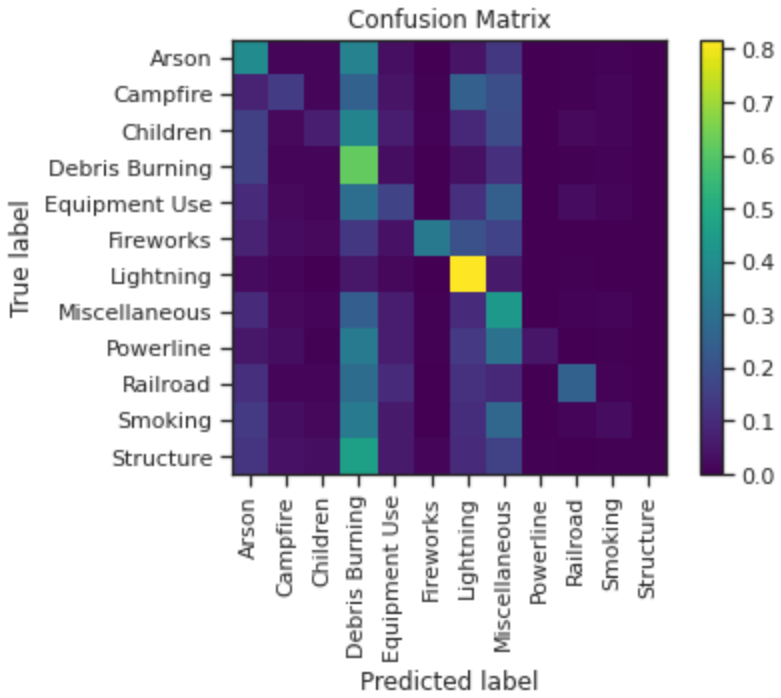


We can see a diagonal line developing! And we are no longer simply predicting debris or lightning for every case. In particular, we are very accurately predicting lightning strikes.

Method 5: Random Forest

Let's now see if using a fancier method can yield a better accuracy.

Accuracy: 46.7%



Ever so slight of an improvement from our regular decision tree model.

b) Predicting Time to Extinguish After Detection

This is a linear regression problem. Given latitude, longitude, cause, type of property owner, month and day of discovery, can we predict how long it will take to extinguish (contain)

the fire? We first checked for multicollinearity between our features, as shown in figure XX.

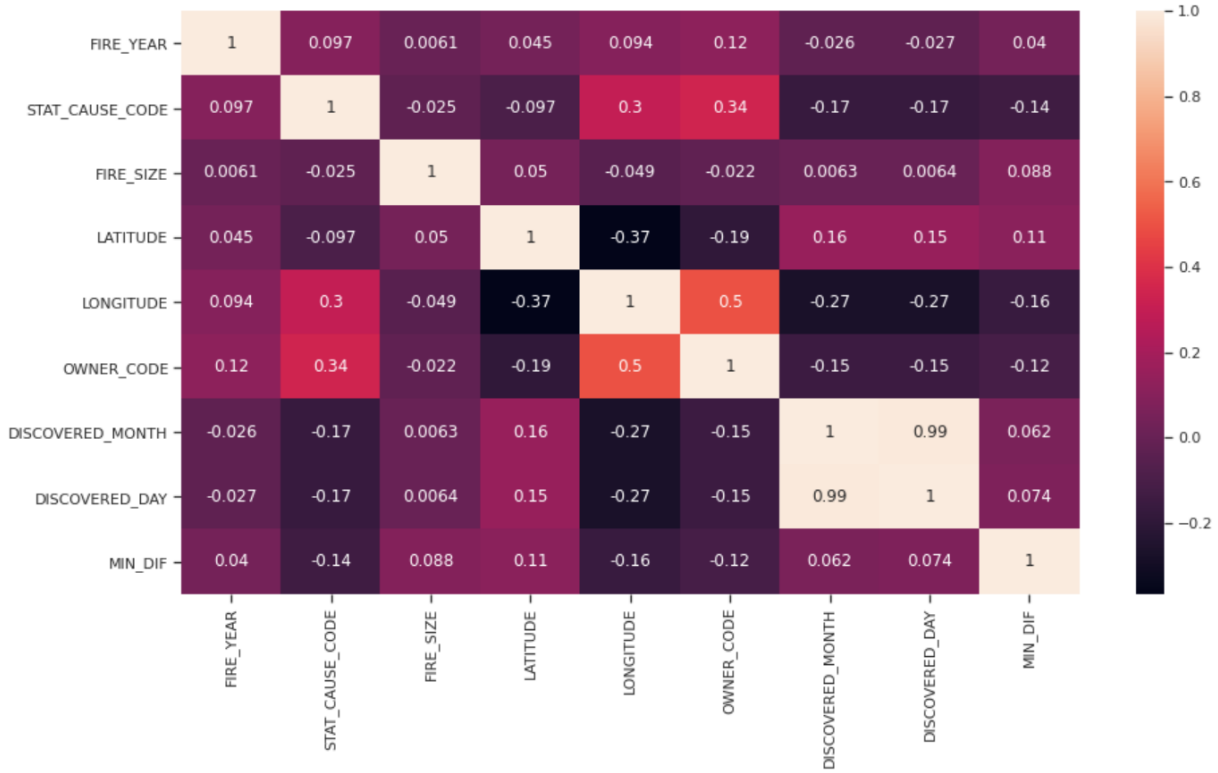


Figure XX: Heatmap of correlation matrix between features

We can see there is little correlation between our features (apart from discovered day and month which is expected). We therefore do not need to run a Principal Component Analysis (PCA) on our dataset to reduce the number of features.

Because our data does not suffer from multicollinearity, we ran four different normalization models: simple Logistic Regression, Lasso/L1, Elastic Net, and a Random Forest Regressor on sklearn.

The simple LR model had an R^2 of 0.143, the Lasso model had an accuracy of 0.126, the Elastic Net one of .055 and the Random Forest one of 21.4%.

4. Description of Challenges/ Obstacles Faced

a) Datetime format

The original date format was in Julian format (For example, January 1, 2008 is represented as 2008001 and December 31, 2007 is represented as 2007365) and we had to convert it to Gregorian format.

b) Accuracy of our models

Our classification and regression errors are very bad. The MSE for our Linear regression model is also very bad. Since the dataset is so huge, we had to sample data to make it go faster. We also tried running our models on more recent data (2010-2015 as opposed to 1992-2015). However, as we pointed out in our EDA (Figure XX), the trends have not changed that much (fire count per year, leading causes of fire countrywide and in California, etc.)

5. Potential Next Steps/ Future Direction

a) Finding better hyperparameters
b) Consulting wildfire specialists
c) Joining our data with meteorological data for same location
d) Adapting our model to 2015-2020

6. Appendix

```
LR Score: 0.14269633455188746
LR MAE: 1772.2029452905817
LR MSE: 20544255.19324879
LR RMSE: 4532.577102846546
```

L1

```
L1 Score: 0.12645137114802973
L1 MAE: 1754.982989319965
L1 MSE: 20933546.277872104
L1 RMSE: 4575.319254202061
```

EN:

```
Score: 0.05486016006981276
MAE: 1735.8403807674572
MSE: 22758949.925213367
RMSE: 4770.634121918528
```

Random Forest

```
RF Score: 0.21439236947580778
RF MAE: 1435.7792082483597
RF MSE: 19087621.951310482
RF RMSE: 4368.938309396286
```