WildFires in the United States

CS - 396: Introduction to the Data Science Pipeline
Final Project Report

Team # G: Manikanta Mandlem, Nihaal Subhash, Devashri Naik, Rishita Jain

Github repo: https://github.com/ManikantaMandlem/Data-Science-Pipeline-Project

Datasets: https://drive.google.com/drive/folders/1Ol4ZhhuL9duYufnfnGSah9ahq0LxMmKw

Introduction:

Although wildfires are a natural occurrence within some forest ecosystems, climate change is causing fire seasons to become intense and widespread. Wildfires can increase air pollution, and carbon emission, cause a reduction in open forest spaces, and severe health conditions for people living nearby. Changes in these fire patterns can be observed in data that is reserved by the authorities. Fire can be caused by multiple factors like climate changes, human intervention, seasonal trends, etc. Analyzing data based on previous forest fires that occurred over the past years, can provide insights into some patterns and predict potential causes which can be helpful to prevent or identify the causes of wildfire. The project focuses on the analysis of forest fire data from 1992 - 2015 (24 yrs). It has been observed that since 1983, the National Interagency Fire Center believes that there were approximately 70,000 wildfires per year in America. Prediction and evaluation of the wildfires can also be extended to wildfire research that studies the effects of smoke from wildfires and determines the effects ranging from eye and respiratory tract irritation to more serious disorders, including reduced lung function, bronchitis, exacerbation of asthma, and heart failure, and premature death.

The analysis of this data opens new avenues for resource distribution, cause determination, and forest fire reduction. The objectives of this experiment are to determine the cause of the fire, the scale of the fire, and the class of fire based on the characteristics. We also intend to explore how different physical and chemical changes in the environment within an area will affect the spread of a particular wildfire. As a part of extended data analysis, we have built prediction models for fire-class, arson, and fire cause predictors. We believe that this can be useful to the authorities in the categorization and determination of the cause.

Dataset:

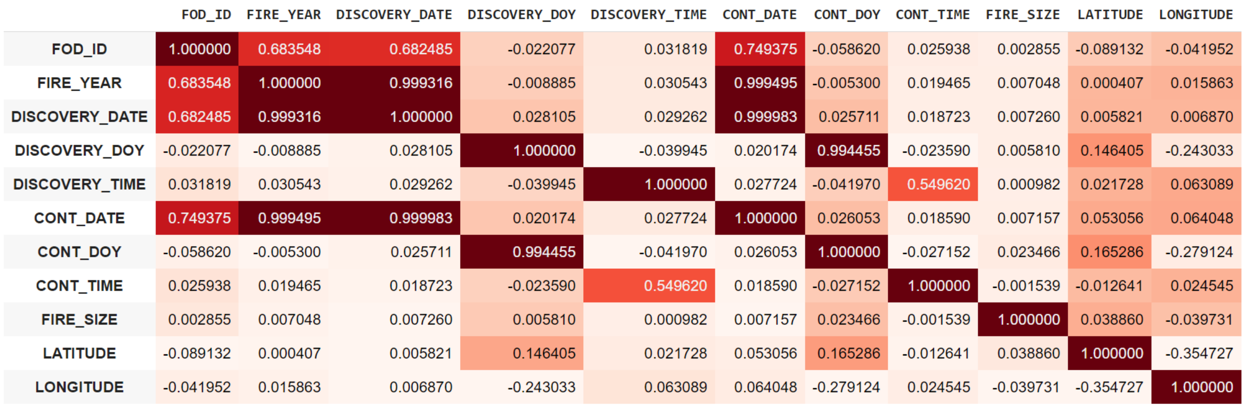
For the analysis of the data in this project, we used the open-source dataset from Kaggle https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires. The dataset consists of 1.88 million forest fires that occurred between 1992 to 2015 and varied across 50 different features. These records were acquired from reporting systems of federal, state, and local organizations. For our analysis, we considered 15 features that were relevant to our analysis including some categorical features (6) and numerical features (9).

Data Cleaning and Data Imputation

In order to process data for further analysis we had to clean the dataset. The dataset had multiple columns which had missing values. Ignorance of this could have been detrimental to the analysis. The columns 'CONT_DATE', 'CONT_TIME', 'CONT_YEAR', and 'DISCOVERY_TIME' had over 40% of missing values. We believed that 'Discovery_Time' could be an important feature for the prediction. Hence, we tried imputing the missing values for this feature. We used the knn-based semi-supervised imputation technique for data imputation. We found that the method explained in this article could be an effective way for data imputation in our case. For better clustering results, we divided the data into 50 clusters and assigned missing labels iteratively to maximize the expectation value. We had to be careful in using the imputed DISCOVERY_TIME feature as 40% of the values were missing. To make sure that imputation did not add additional bias to the predictive models, we trained and tested the model performance with and without discovery time and found out that the DISCOVERY_TIME feature is either improving the performance or is not affecting in any way. So, we decided to include the imputed DISCOVERY_TIME feature in our entire analysis. Furthermore, we dropped other columns with missing data and also dropped some other columns we believed were irrelevant to our analysis.

Exploratory Data Analysis

To understand and analyze the data better, we performed some EDA. Different features were observed to identify the connections within the data. Below are some plots and statistics that can help understand the nitty-gritty details.



	FOD_ID	FIRE_YEAR	DISCOVERY_DATE	DISCOVERY_DOY	DISCOVERY_TIME	CONT_DATE	CONT_DOY	CONT_TIME	FIRE_SIZE	LATITUDE	LONGITUDE
FOD_ID	1.000000	0.683548	0.682485	-0.022077	0.031819	0.749375	-0.058620	0.025938	0.002855	-0.089132	-0.041952
FIRE_YEAR	0.683548	1.000000	0.999316	-0.008885	0.030543	0.999495	-0.005300	0.019465	0.007048	0.000407	0.015863
DISCOVERY_DATE	0.682485	0.999316	1.000000	0.028105	0.029262	0.999983	0.025711	0.018723	0.007260	0.005821	0.006870
DISCOVERY_DOY	-0.022077	-0.008885	0.028105	1.000000	-0.039945	0.020174	0.994455	-0.023590	0.005810	0.146405	-0.243033
DISCOVERY_TIME	0.031819	0.030543	0.029262	-0.039945	1.000000	0.027724	-0.041970	0.549620	0.000982	0.021728	0.063089
CONT_DATE	0.749375	0.999495	0.999983	0.020174	0.027724	1.000000	0.026053	0.018590	0.007157	0.053056	0.064048
CONT_DOY	-0.058620	-0.005300	0.025711	0.994455	-0.041970	0.026053	1.000000	-0.027152	0.023466	0.165286	-0.279124
CONT_TIME	0.025938	0.019465	0.018723	-0.023590	0.549620	0.018590	-0.027152	1.000000	-0.001539	-0.012641	0.024545
FIRE_SIZE	0.002855	0.007048	0.007260	0.005810	0.000982	0.007157	0.023466	-0.001539	1.000000	0.038860	-0.039731
LATITUDE	-0.089132	0.000407	0.005821	0.146405	0.021728	0.053056	0.165286	-0.012641	0.038860	1.000000	-0.354727
LONGITUDE	-0.041952	0.015863	0.006870	-0.243033	0.063089	0.064048	-0.279124	0.024545	-0.039731	-0.354727	1.000000

Fig 1: Heatmap of Pearson correlation among the feature combinations with red gradient portraying the correlation.

We decided to start our EDA process with a heat map to get a general idea of how the attributes within our data are correlated. The heatmap above clearly demonstrates that though there are no meaningful correlations among features, there is a moderate correlation between fire contained time and fire discovery time. The possible reason for this could be because the fires that are discovered quickly are contained easily as the spread could be stopped from spreading to greater areas.

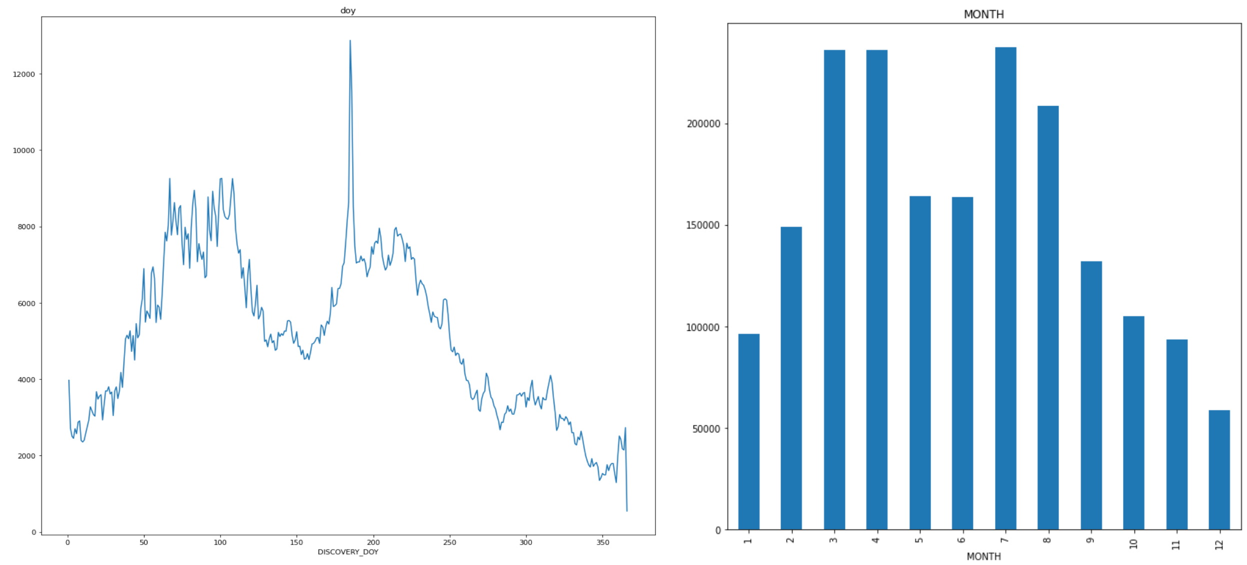


Fig 2: Number of wildfires corresponding to (a) each day in a year, and (b) each month of a year.

We can observe from plot (a) that most wildfires per day occurred in the month of July. This can be expected because of the heat and dryness in summers in California areas. In fig (b) we can see that the Spring months have the number of fires almost equal to the summer months. Upon some research, we found that wildfire occurrence is highest in spring because of the presence of driest leaves, warm temperatures, dry weather, and gusty winds. All these factors contribute to the rise in wildfires in the spring months. The increase in the summer months is due to the heat and dryness in the atmosphere as stated earlier.o8

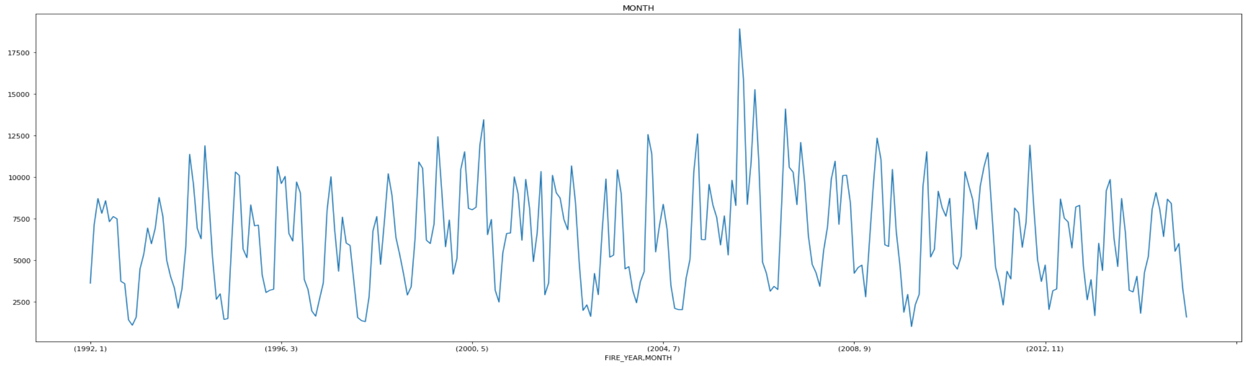


Fig 3: Seasonality of the wildfires throughout the years of collected data.

The plot in Fig 3 gives more insights into how the graphs in Fig 2 can be extrapolated for further prediction of wildfires. As expected, there is a clear cyclic pattern in the emergence of wildfires where the frequency of wildfires increases in the early spring and late-summer months of each year. Fig 4 shows that Debris burning is the first major cause of the wildfire.

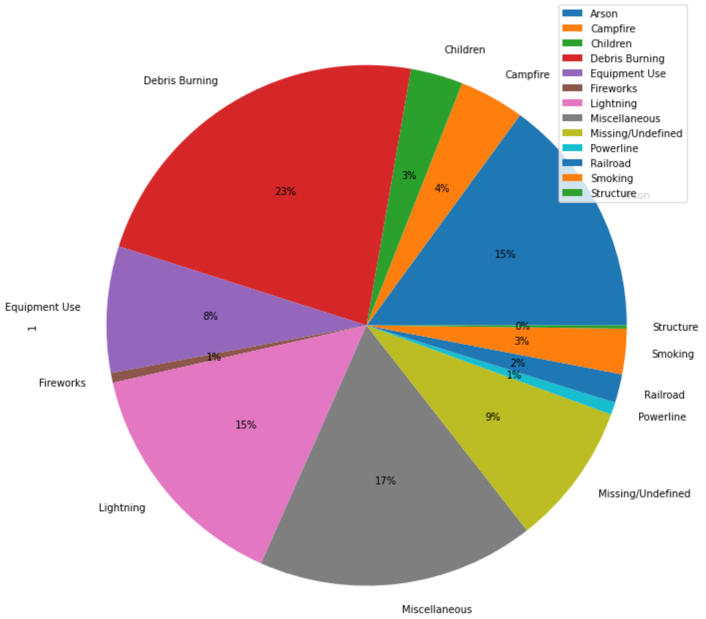


Fig 4: Causes of wildfires

Also, from a bit of in-depth analysis, we found that more than 80% of wildfires are caused by human activities. We can clearly see that humans are the major cause of wildfires. From a preliminary understanding of the data, we planned to formulate a problem statement that when solved, can help firefighting authorities in their operations.

Problem Statement

For this project we have devised three problem statements to address with the data at hand. The same is explained in detail in the following section.

1. Firstly, to predict the cause of the fire given the characteristics of the fire. There are a total of 12 different causes of the fire and so this is a 1/12 classification. A mere random guess would have an accuracy of around ~8%. Keeping this in mind, we believe that this is the hardest problem to be solved among all others.

Fire Cause	Count of Wildfires
stat_cause_descr_arson	281K
stat_cause_descr_campfire	76K
stat_cause_descr_children	61K
stat_cause_descr_debris burning	429K
stat_cause_descr_equipment use	148K
stat_cause_descr_fireworks	11K
stat_cause_descr_lightning	278K
stat_cause_descr_miscellaneous	324K
stat_cause_descr_missing/undefined	167K
stat_cause_descr_powerline	14K
stat_cause_descr_railroad	33K
stat_cause_descr_smoking	53K
stat_cause_descr_structure	4K

2. Secondly, to predict the cause of the fire if it is arson or non-arson. Wildfires set ablaze with malicious intent is a pressing issue that needs to be countered and we believe that a solution to this problem statement can help fire fighting authorities in identifying Arsons. Below is a table depicting the counts in each class.

Fire Cause	Count of Wildfires
Arson	281K
Non-Arson	1.6M

3. Finally, to predict the fire size class of the wildfire. In this dataset, the size of the fire is classified into 7 fire_size_classes, definitions of which are based on the acres of land affected by the wildfire as below:

Fire Size Class	Definition (Acres of Land)	Count of Wildfires
A	0 - 0.25	667K
B	0.26 - 9.9	939K
C	10 - 99.9	220K
D	100 - 299	28K
E	300 - 999	14K
F	1K - 5K	8K
G	5K+	4K

A model to predict the size of the fire can potentially help firefighting authorities to mobilize the available resources effectively in the event of a wildfire.

Feature Engineering

In order to simplify the data and get a set of meaningful features to train the prediction models, we performed feature engineering on certain features to create new features. The dataset consisted of information about the center of the wildfire-affected area as a topological feature in terms of latitude and longitude. While trying to use this information for our analysis, we understood that this method did not seem to be helpful. This is why we tried using these features by converting them into angle and distance. This distance is the distance between the (latitude, longitude) of the desired point and (mean_latitude, mean_longitude) point as considered in the US geography. Angle is calculated by joining the (latitude, longitude) point with the (mean_latitude, mean_longitude) point and measuring the angle made by this line with an imaginary horizontal line drawn on the US map.

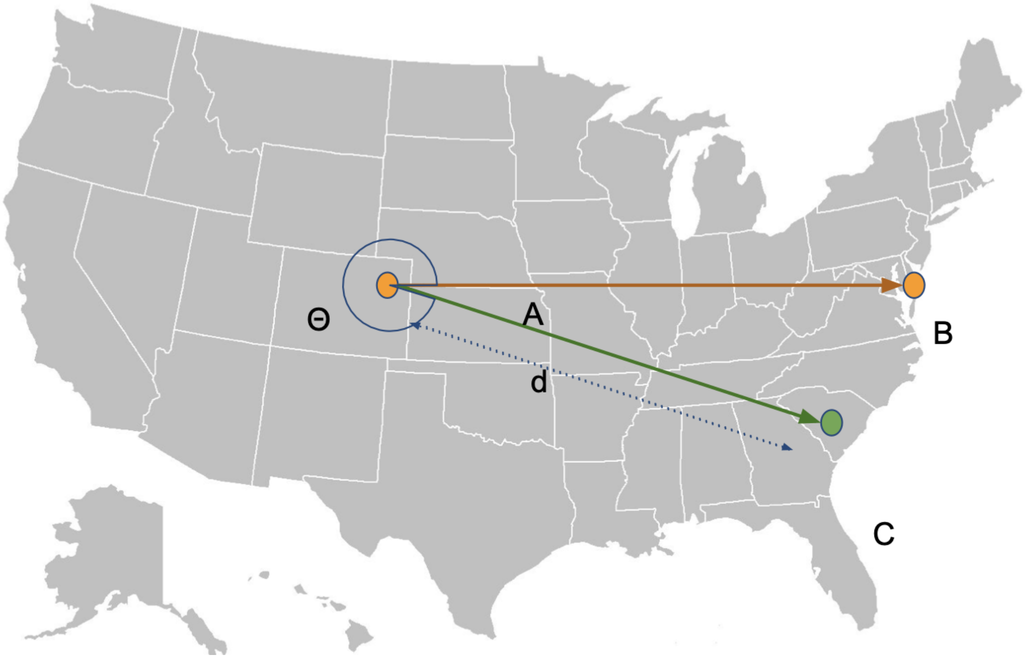


Fig 5: Calculation of distance and angle with respect to mean_latitude and mean_longitude point.

To make the data more manageable, we also mapped the 'STATE' column to 9 regions. The state-to-region information is gathered from the US State Forest Department [1]. The below map shows the division of states into regions.

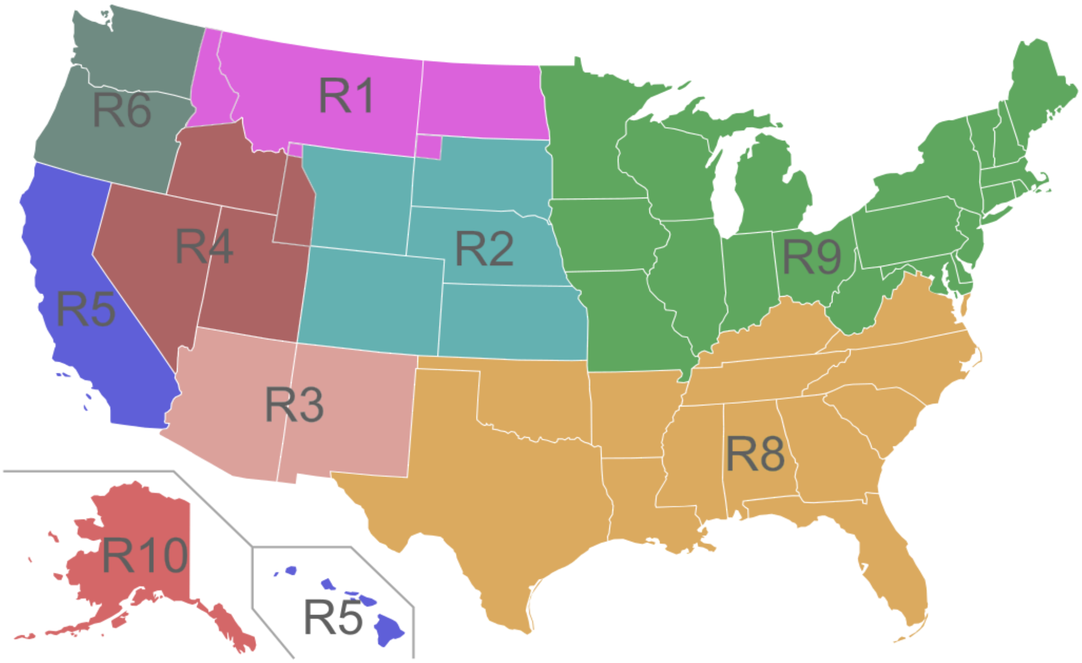


Fig 6: The division of states into regions based on the US Forest Department.

Finally, we clubbed Discovery Time and Discovery DOY into appropriate bins as shown in the below tables.

Time (24 Hr)	Bin
0 -4	Late Night
4 -8	Early Morning
8 - 12	Morning
12 - 16	Afternoon
16 - 20	Evening
20 - 24	Night

Months	Bin
Dec - Feb	Winter
Mar - May	Spring
June - Aug	Summer
Sep - Nov	Fall

Data Pre-processing

Before proceeding with any prediction tasks we converted categorical features into one-hot encoded features and scaled numerical features using MinMaxScaler so that they lie between 0 and 1. This helped in stabilizing the training process and for a better model fitting. We wanted to see if feature engineering is helping the models in any way and so we decided to create two datasets for model building, one is with raw data and the other is with feature engineered data. The below tables depict the outline of these two datasets.

Dataset type	Feature count
Raw Data	95
Feature Engineered	59

We split this data in a 90:10 ratio into Train and Test data. Given the huge dataset we had, we believed that 10% of the data is enough to evaluate the already cross-validated model performance. This is how the counts of train and test data points look like

Train Test Split	Percentage	# Data Points
Train	90%	1.7M
Test	10%	188K

We can see that even with 10% of data we have 188K data points for model validation

Data Modeling

We attempted to build different models for each of our different problem statements.
For the first model, we attempted to build a classifier to predict the cause of the fire given its characteristics. The output features are a 12-class classification problem. The input features to the model were Source System Type, Fire Year, Contained Day of Year, NWCG Reporting Agency, Discovery Date, Contained Date, Discovery Day of Year, Fire Size, Fire Size Class, Discovery Time, Latitude, Owner Description, State and Longitude

For the second model we attempted to build a classifier to predict whether or not the fire was caused by someone with malicious intent. The output feature is a binary value of 1 denoting arson and 0 denoting non-arson. The input features to the model were Source System Type, Fire Year, Contained Day of Year, NWCG Reporting Agency, Discovery Date, Contained Date, Statistical Cause Description, Discovery Day of Year, Fire Size, Fire Size Class, Discovery Time, Latitude, Owner Description, State and Longitude

For the third model we attempted to build a classifier to predict the size of the fire. The output feature is a classification with 7 different classes denoting the size of the fire.. The input features to the model were Source System Type, Fire Year, Contained Day of Year, NWCG Reporting Agency, Discovery Date, Contained Date, Statistical Cause Description, Discovery Day of Year, Discovery Time, Latitude, Owner Description, State and Longitude

We experimented with neural networks, random forest classifiers and gradient boosting classifiers. The idea of using neural networks was that they usually work really well when we have a large amount of data - and we had 1.88 million examples. The idea of using random forest classifiers was that decision tree-based methods are explainable(which would be useful to researchers working in the same area), and random forest models work really well with high dimensional data - since only a subset of features are given to each decision tree. The idea behind using Hist Gradient Boosting Classifier was that the estimator is much faster than the conventional Gradient Boosting Classifier for large datasets, where the number of examples is greater than 10,000.

The baseline values we got for the different models are:

Model	Fire Size Class Predictor	Fire Arson Predictor	Fire Cause Predictor
Hist Gradient Boosting Classifier	73.6%	87.2%	54.1%
Random Forest Classifier	**74.1%**	**88.7%**	**59.7%**

MLP	62.8%	86.8%	53.1%

We observed that the Random Forest Classifier attained the highest accuracies in each case. We believe that this is due to the fact that our data is high dimensional (95 features after one-hot encoding), and thus Random Forest Classifiers, which give only a subset of features to each Decision Tree, perform the best on our dataset.

We also attempted to model these classifiers both with and without the imputed discovery time values. There were doubts about the effectiveness of imputing the values as more than 40% of the values were missing. However, we observed that the final accuracy (after hyperparameter tuning) either improved or was unaffected by the imputation so we concluded that it was safe to do so.

After feature engineering we got the following accuracies:

Model	Fire Class Predictor	Fire Arson Predictor	Fire Cause Predictor
Hist Gradient Boosting Classifier	61.7%	86.9%	56.1%
Random Forest Classifier	62.4%	88.2%	58.0%

Some of the accuracies were lower, but with some fine-tuning and grid search we were able to get the same accuracies again.

We selected Random Forest as the model on which we would perform hyperparameter tuning because it had the highest accuracy in every problem. After tuning the best model performances for each task were as follows:

Problem	Fire Class Predictor		Fire Arson Predictor		Fire Cause Predictor	
	Raw Data	F.E. Data	Raw Data	F.E. Data	Raw Data	F.E. Data
Accuracy	75.1%	75%	89.2%	89.2%	61.8%	61.5%
F1 Score	71.3%	71.5%	88.3%	88.2%	60%	60%

-

Problem	Fire Class Predictor	Fire Arson Predictor	Fire Cause Predictor
Max Features	sqrt	0.7	0.7
Max Depth	50	30	30
Min Sample Split	30	10	10
Min Samples Leaf	1	1	1

Conclusion and Inferences

Our hypothesis that the Random Forest Algorithm would work well because the models have several input features held true. Random Forest did, in fact, have the highest accuracy, for all our different problem statements.

In most machine learning applications, geolocation-specific data such as latitude and longitude provide an extra context to the data. In our analysis, the location gives us the necessary information on where the fire took place in the North American Plain. Since these attributes are of importance in our data, we decide to perform feature engineering specific to the longitude and latitude values. In recent years, there have been many analytics tools developed for geo-location feature engineering. Geopandas has good built-in plotting functionality. There are a number of excellent Python libraries to visualize geodata such as Folium and Plotly. Libraries like GeoJson and Geopy are capable of generating new features with the provided latitude and longitude. They are also capable of converting the geodata into a physical address on a map. Geo-location data and its related features can be structured as regular tabular data with numerical or categorical variables which makes them easier to manipulate and analyze. With models like Random Forest and Gradient Boosting, the geo-location data doesn't need to be normalized.

Feature engineering did not have a significant difference in some cases. We believe this is because some of the engineered features do not give the model any new information. For example, the latitude and longitude features already give us good information about the location of the fire. The new angle and distance from center features do not add any new information to the model and thus do not improve accuracy. Dividing the states into 9 regions rather than 50 states should help reduce the number of features and make the model less likely to overfit, but perhaps with 1.88 million examples, overfitting did not turn out to be the concern we thought it would be. However, we stand by our notion that feature engineering was useful as it reduced the size of the dataset by 40% while maintaining similar performance.

Applying grid search CV allowed us to tune our parameters to produce the best possible models. The highest accuracies we have obtained are 75.1% at fire size prediction, 89.2% at flagging potential arsons, and 61.8% at fire cause prediction.

We believe these models could be potentially very useful. Fire cause predictions could help the fire fighting authorities in determining the cause of fires in various locations and prevent them from happening in the future based on the causes. Flagging potential arson cases could be useful for crime-fighting authorities. If a fire has been flagged as having the potential of being caused with malicious intent, the authorities could follow up and look into it, rather than scan every case individually. The fire size prediction model could be used to determine potentially how large fires could get. This could then be used to divert and distribute resources according to the potential dangers of each fire. For example, if a fire has the potential to become dangerously large, it could be prioritized over other fires in the area.

Challenges

One of the main challenges was the sheer size of the dataset. With 1.88 million rows and 95 columns (after one-hot encoding), we had various issues like RAM constraints and extremely slow training. Some of the methods like feature engineering regions and dropping the state columns helped with these issues.

Multiple columns had missing values. CONT_DATE, CONT_TIME, CONT_YEAR, and DISCOVERY TIME had over 40% missing values. We imputed the values for DISCOVERY_TIME and dropped the remaining columns. Even with DISCOVERY_TIME, we had to be careful to test with and without the imputed column since imputing 40% of values can be risky.

Some of the problem statements like fire cause prediction are inherently very difficult problems even for humans who are experts in the field. Therefore, it was difficult to get very high accuracy numbers.