Applied Competitive Lab in Data Science - Final Report

Group: Rotem Agmon, Kim Yekutiel, Gal Getz, Moral Bootbool

Introduction:

The dataset contains records of wildfires in the United States from 1992 to 2015, totaling 570K entries and covering an area of approximately 140 million acres burned.

Our task is to **model and predict the cause of the fire** based on key attributes such as discovery date, final fire size, and precise location.

Our Process:

🔥Data Preprocessing:

Features analysis

First, we started to overview every single feature, and decide which feature passes to the next stage.

The Reason of dropping features	List of features we dropped	Remarks
Leakage	SOURCE_SYSTEM_TYPE SOURCE_SYSTEM NWCG_REPORTING_AGENCY NWCG REPORTING UNIT ID NWCG REPORTING UNIT NAME CONT DATE CONT DOY CONT TIME	Also, we thought that the DURATION of the fire (feature we wanted to create), and FIRE SIZE are leakage too. But our mighty lecturer approved us to use those features :) **Below is an explanation of the features we identified as potential sources of leakage.**

Large amount of missing data	ICS209INCIDENT NUMBER ICS 209 NAME MTBS ID MTB FORENAME COMPLEX NAME	Due to the large amount of missing data, we couldn't use those features, even if we wanted to. If we filled in the data, we would cause a bias.
Categorical features that have too many values	FIPS_CODE COUNTY FIRE CODE LOCAL INCIDENT ID LOCAL FIRE REPORT ID FPA ID FOD_ID	Those features are categorical, and we won't make them dummies in order to avoid too many columns, which could overfit our model.

*** Why are the features above considered as potential sources of leakage?**

Because they provide information that would not be available at the time of prediction in a real-world scenario. For example, NWCG_REPORTING_AGENCY provides information about the agency or unit reporting the wildfire, so if the Department of Defense made the report, it would imply that the fire is probably arson.

Another example is the features: CONT_DATE, CONT_DOY, and CONT_TIME. They represent the containment date, day of year, and time of containment of the wildfire, respectively. Including these features in the model would essentially use future information to predict past events, leading to data leakage.

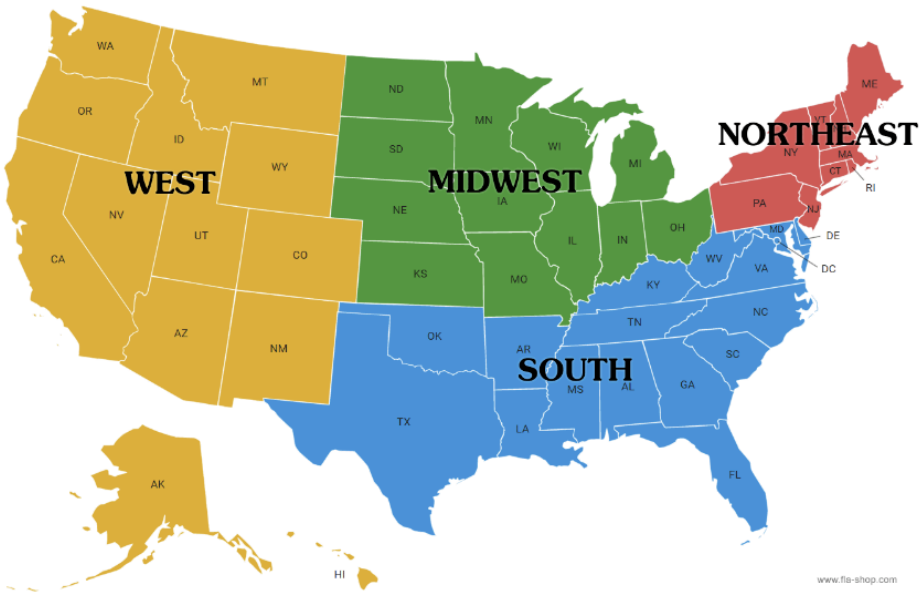
Feature engineering

Here we will explain about the features we engineered:

- FIRE NAME - this feature could perform leakage in the name itself, so we chose to make this into a binary feature - does the fire have a name or not.
- YEAR - the year the fire began, taken from DISCOVERY_DATE.
- DAY_OF_WEEK - We engineered a new feature that symbolizes the day in the week (Sunday, Monday, etc.), and we did it as a Cyclic feature with Sin and Cos,

as we learned in the lecture. Eventually, we didn't use this feature as we will explain later.

- MONTH - just as DAY_OF_WEEK, this too is a Cyclic feature.
- DURATION - new feature we designed for the duration of the fire. When we tried to make this feature, we discovered that there were a lot of missing values, and because of the leakage, we decided to drop this feature.
- REGION - a more general geographical feature instead of the STATE feature.



Map of the U.S. with 4 regions

- LOG_FIRE_SIZE - log of FIRE_SIZE, explanation will appear in Transformations and Outliers sections.
- FOREST_UNIT - a binary feature that indicates whether “forest” is in the SOURCE_REPORTING_UNIT_NAME feature.

One Hot Encoding:

We convert the categorical features that remain in our work to one hot encoding. The features are: REGION, OWNER_DESCR.

Missing Values and Outliers

- The features that remain have no missing values, so we didn't need to handle it.
- To identify outliers in the numeric features, we examined the histograms and other EDA graphs. However, we did not observe any outliers except for FIRE_SIZE.

- FIRE_SIZE exhibited outliers in the form of very large fires, many of which were Lightning fires. Given that it is a large part of the Lightning records, we want to keep them, so we apply log transformation to this feature.
- Regarding LATITUDE and LONGITUDE, no outliers were detected since all records are within the geographical bounds of the USA.

External data

Weather Data: We believe that integrating weather data can enhance our ability to predict fires, especially considering indications from our EDA (as can be seen in the EDA section, suggesting a relationship between the labels and warmer months like July and August, in some states).

Due to the difficulty of manually extracting data from all years, we chose to source the weather dataset from 1992, the earliest year in our data, to avoid potential leakage.

We aggregated the weather data for each state and month, resulting in the creation of **two additional features:**

1. MAX_TEMP: This feature indicates the average maximum temperature observed in each state and month.
2. PRECIPITATION: This feature represents the average precipitation observed in each state and month.

Special Dates: We decided to incorporate special dates into our analysis based on their potential impact on fire occurrences. These dates were manually gathered from various online sources.

We initially added the IS_SPECIAL_DATE ,a binary feature that indicates whether a given date is considered special based on our collected dates (such as Christmas, Halloween and Memorial Day). However, upon conducting our exploratory data analysis (EDA), we decided to replace it with **IS_FOURTH_JULY_RANGE**. This feature identifies whether a date falls within the range of the Fourth of July holiday. We will provide further explanation on why we selected this date in the EDA section.

Aggregations

We conducted several aggregations, one of which involved the weather data discussed above in the External Data section. This aggregation was performed at the state and month levels. Additionally, the REGION feature is sort of an aggregation based on mapping the state to regions in the USA.

Transformations

We performed several transformations on the data, including the following:

1. LOG_FIRE_SIZE: As discussed earlier, we applied a logarithmic transformation to the FIRE_SIZE feature to handle outliers and improve its distribution (graph in EDA section).
2. Cos and Sin Transformations: We applied cyclic transformations to the Day of Week and Month features using cosine and sine functions. This transformation helps capture the cyclical nature of time-related features, such as days of the week, by representing them in a continuous and periodic manner.

Interactions

For each of the regions, we created an interaction feature with longitude and latitude. We did this in order to explore the influence of longitude and latitude in each region. For example, we saw that there were more fires caused by lightning on the shores - both regions and longitude-latitude won't be able to catch them, but in this approach, the model could infer this.

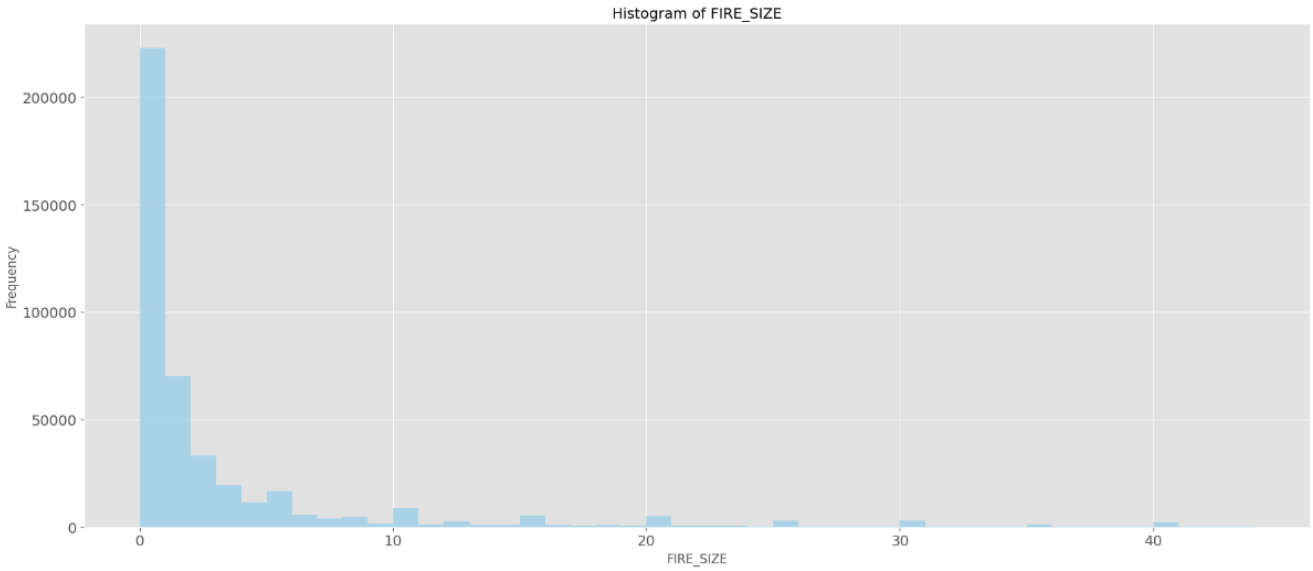
🔥Exploratory Data Analysis (EDA):

Our Exploratory Data Analysis (EDA) involved several key steps:

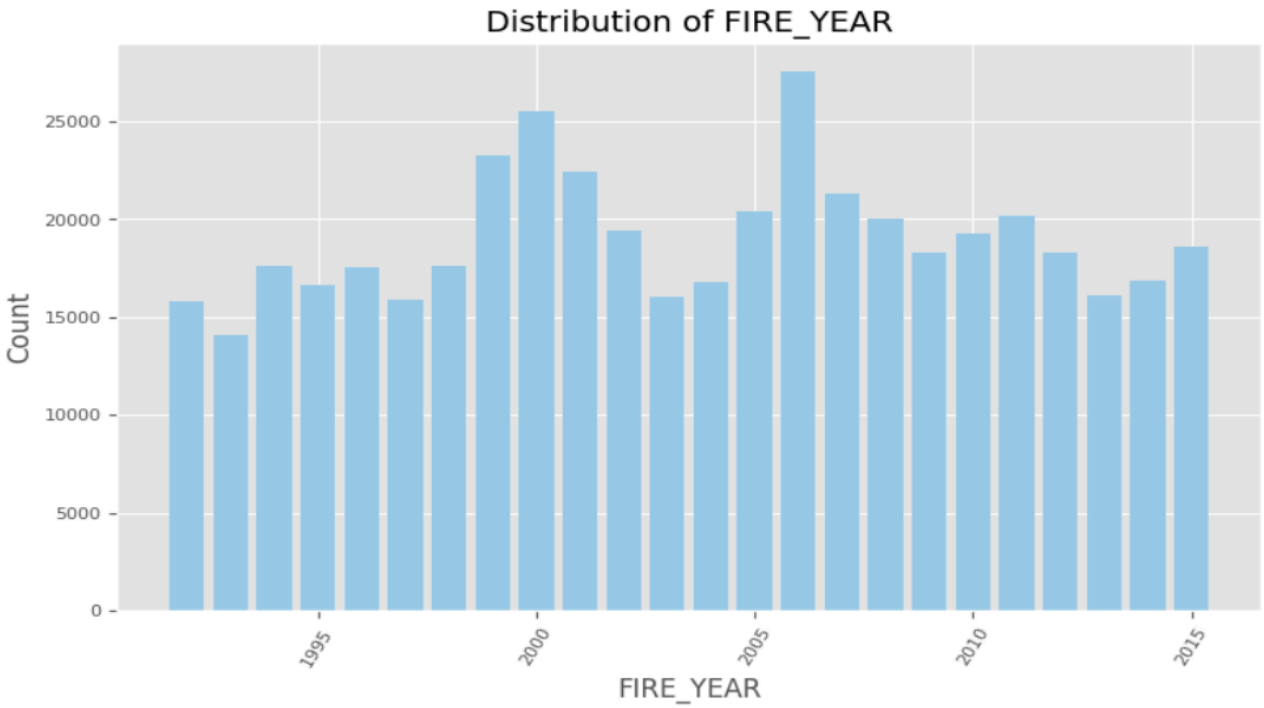
- **The EDA was performed on the train data only**.
- We began by categorizing features as either categorical or numerical for analytical purposes.
- Numerical features were represented visually using histograms to depict their distribution, while categorical features were examined using bar plots to understand their variability.
- We did comparisons of the features with the target variable, to examine whether the feature behaves differently for each value of the target (using different plots).
- Deep dived with analysis for features that required further investigation, like interactions between features.
- Correlations between different features were explored.

During our analysis of the features, we observed:

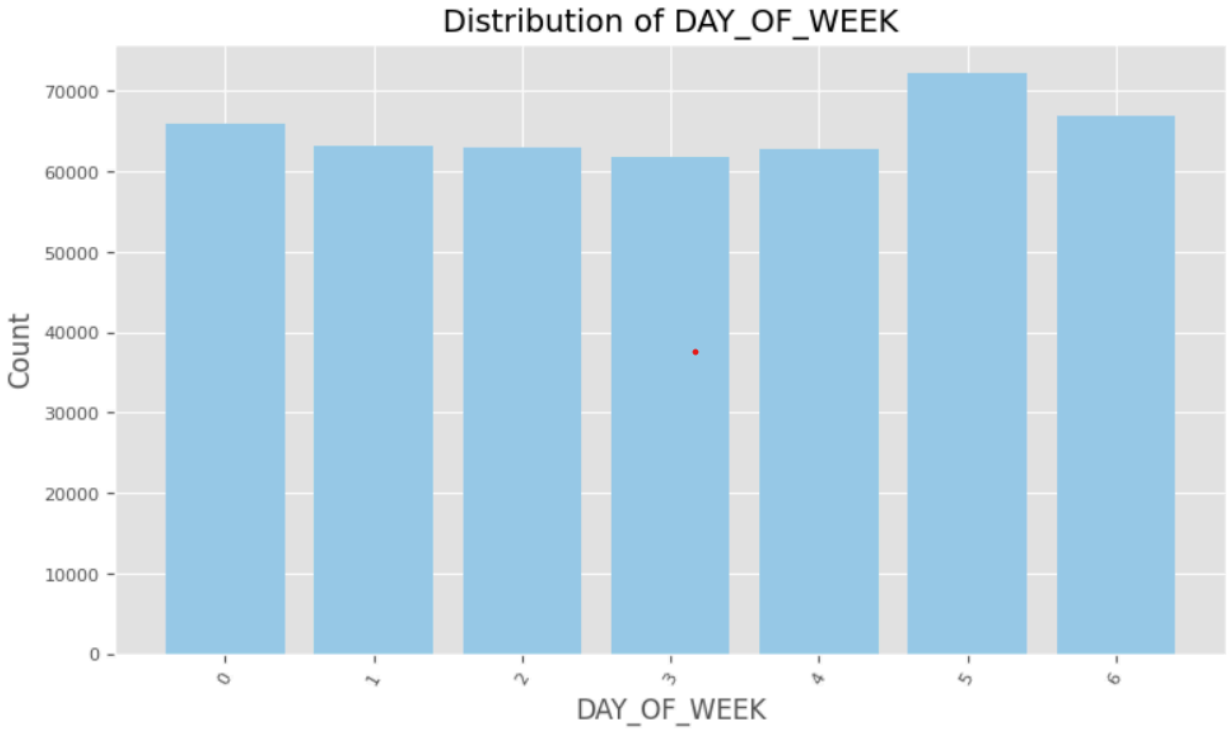
- There were mostly small fires in the data, only a few larger ones.



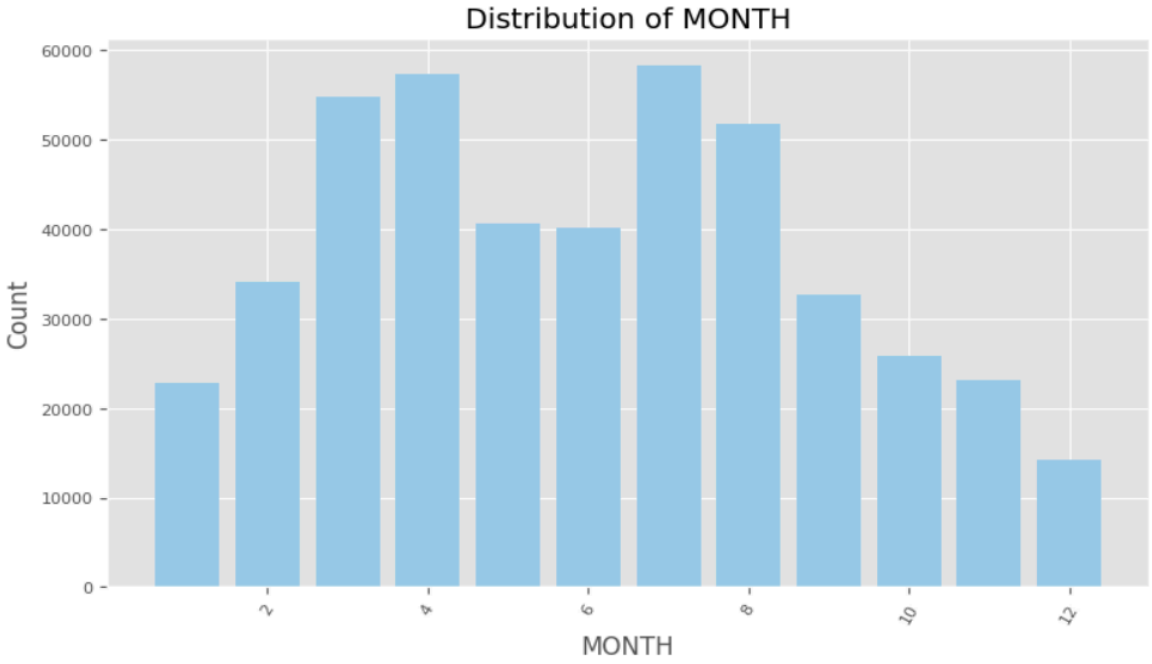
- The year 2006 was particularly noteworthy due to a significant increase in fire occurrences, which may have been linked to a heatwave in California during that period.



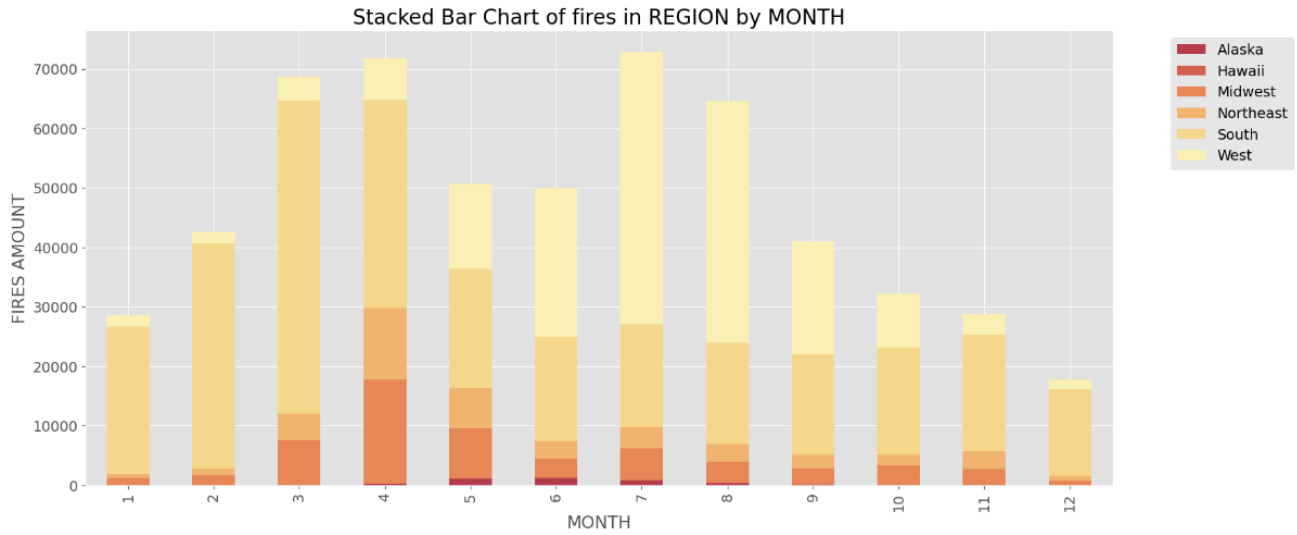
- We found that weekend days had a slightly higher frequency of fires.



- Looking at the data month by month, we noticed two main peaks happening in March/April and July/August.

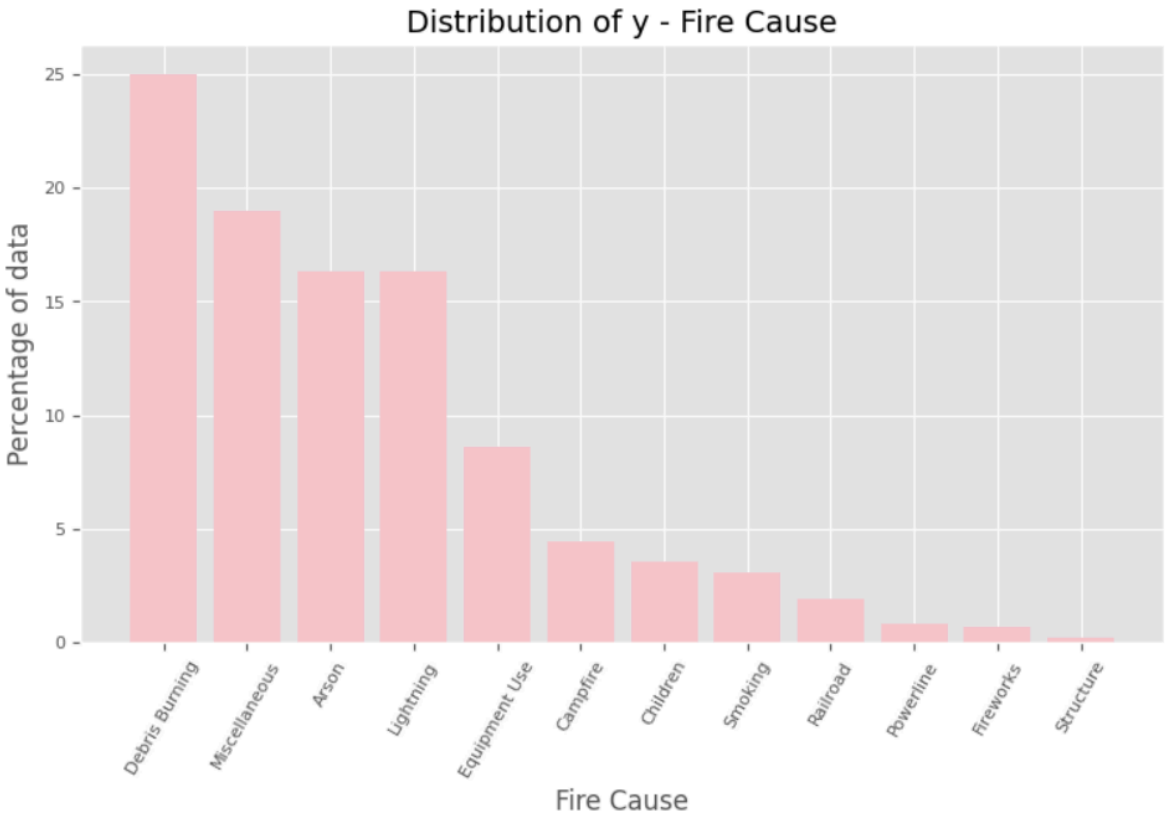


This made us curious to dig deeper and find out if these months were particularly related to specific locations (in terms of climate). Breaking the data into regions, we can observe that the March/April peak is related to the South of the US, whereas the July/August one is related to the West of the US. At this point we took under consideration that the Region feature might be useful.



The target variable:

- The Fire Cause variable has 12 different labels, 25% of train data is Debris Burning, ~20% is caused by Miscellaneous, and the rest can be observed in the plot below.

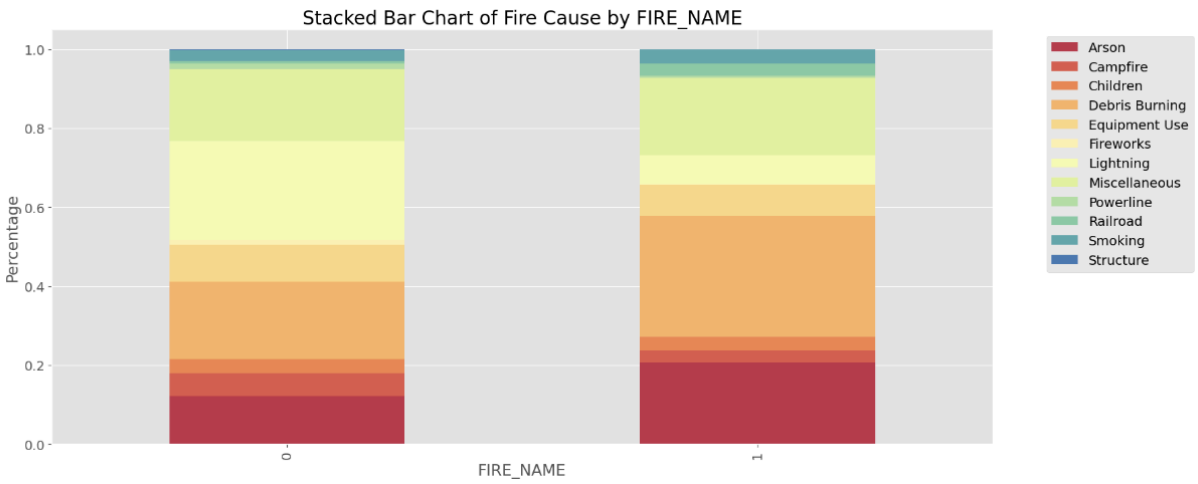


- We saw that the distribution of this label wasn't uniform in the data (as in identically distributed), so we decided to keep the train-test partition according to this distribution in further steps.

Comparisons with the target variable:

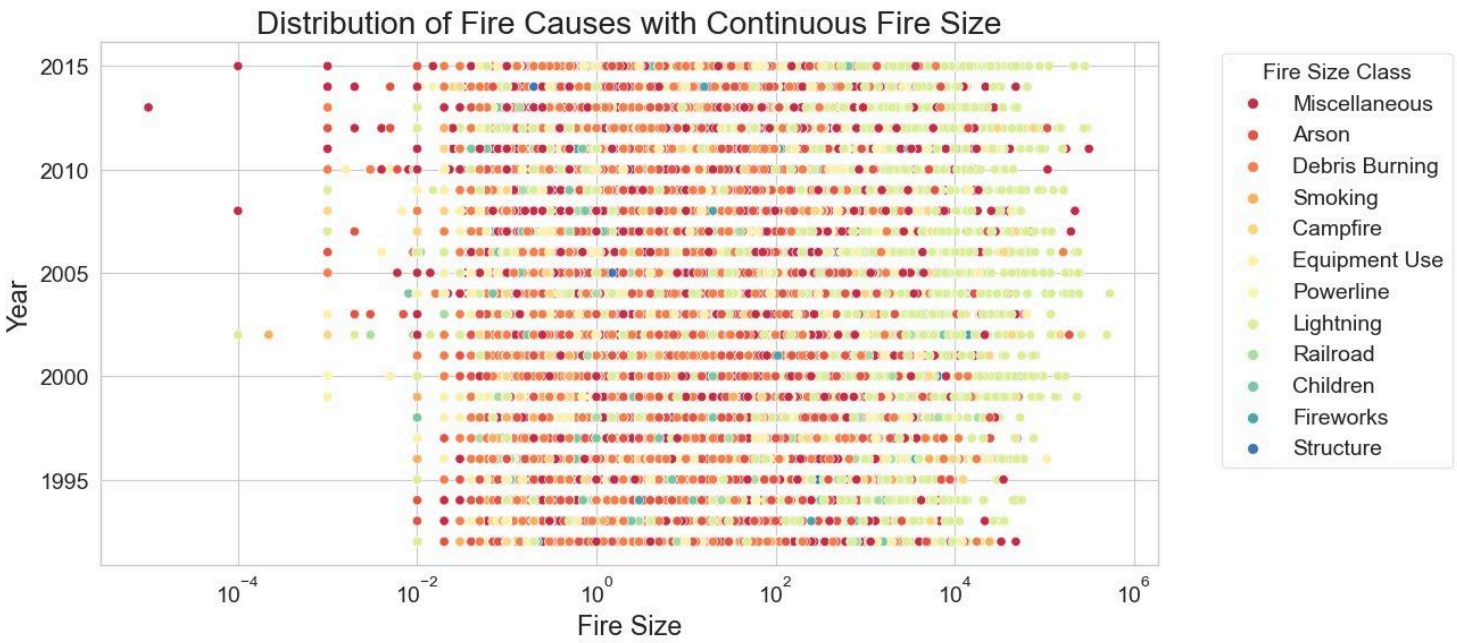
In general, we wanted to seek for features that behave differently between the target variable labels, and to examine their relationship with the target. We'll show here some of the interesting plots.

- When we looked whether fires had **names** (1) or not (0) as a binary feature, we found an equal amount of fires for both cases (~50%). We found that having a fire name has a connection to the Lightning cause.



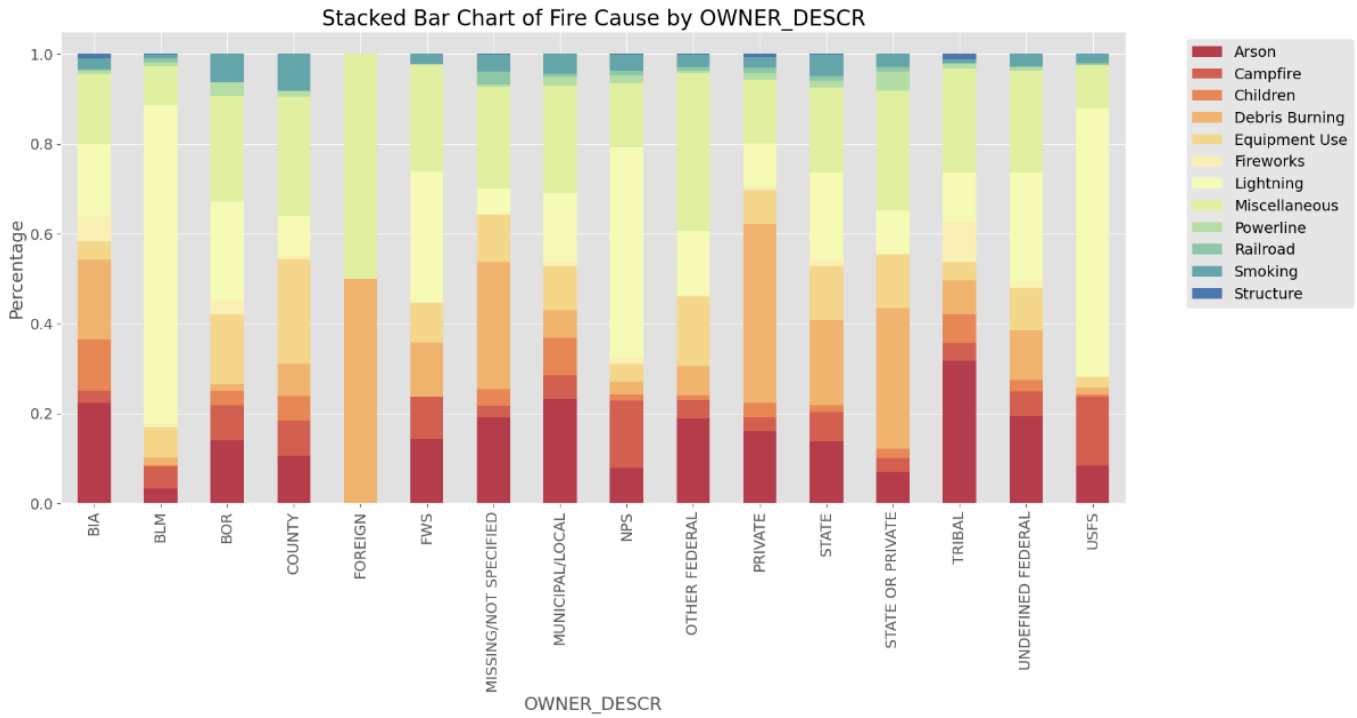
- Upon examination of fire types over the years to check if there are trends in new fire types, we didn't find a specific trend.

- Large fires were predominantly caused by lightning strikes. This observation was only accrued after performing log transformation over the Fire Size, as we can see in the plot below.



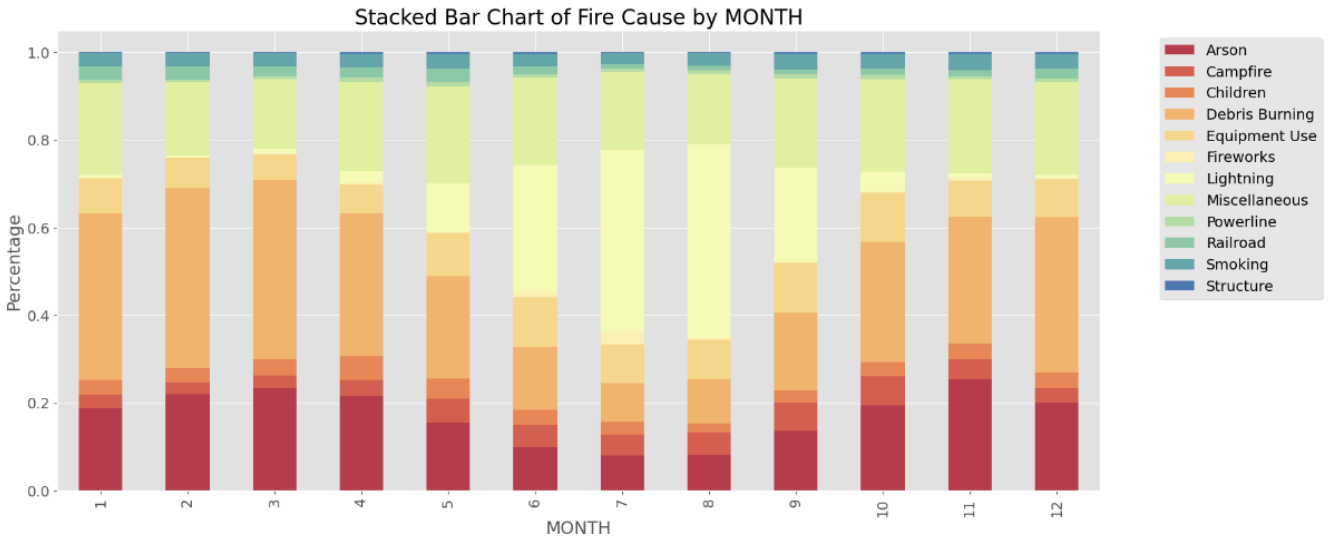
- The distribution of fire types varied significantly across different **states**, but we didn't find a specific trend. The number of states was too large, so we decided not to use it in the models and to find a different way to indicate location.

- **The landowners description** was connected to the type of fire, so we decided to keep this feature (we saw the feature importance was relatively high, further on).

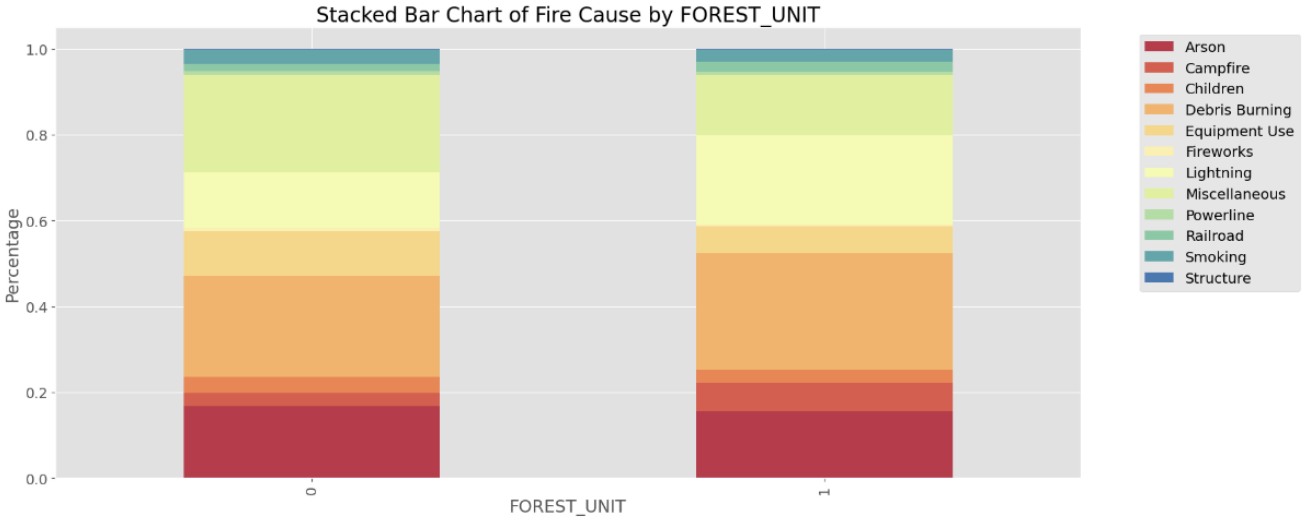


- **Day of the week** did not significantly influence the type of fire, suggesting it could potentially be excluded from further analysis.

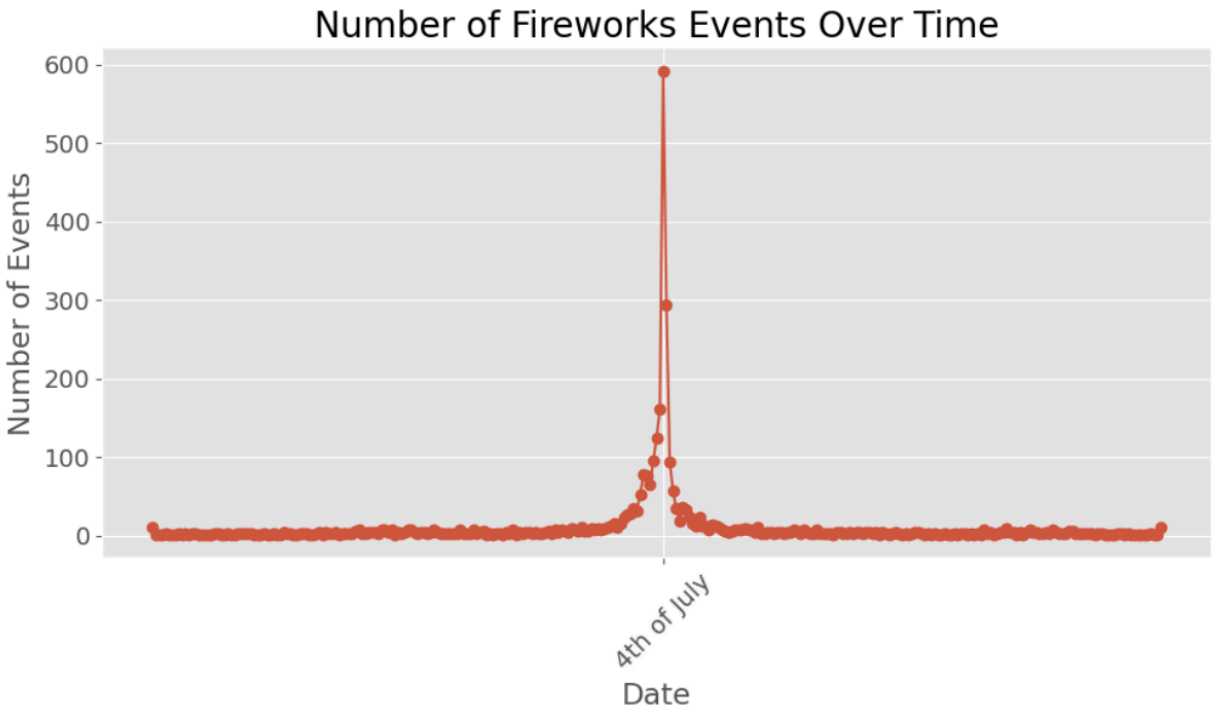
- Comparison of **Month** by the target revealed a higher incidence of lightning fires in July and August, for example, suggesting that the month of the year may have a significant impact for the prediction.



- Plotting the engineered **Forest Unit** feature revealed connection to Lightning, Miscellaneous and slightly with Equipment Use and Debris Burning. The connection of forests to all of those causes made sense to us as well.



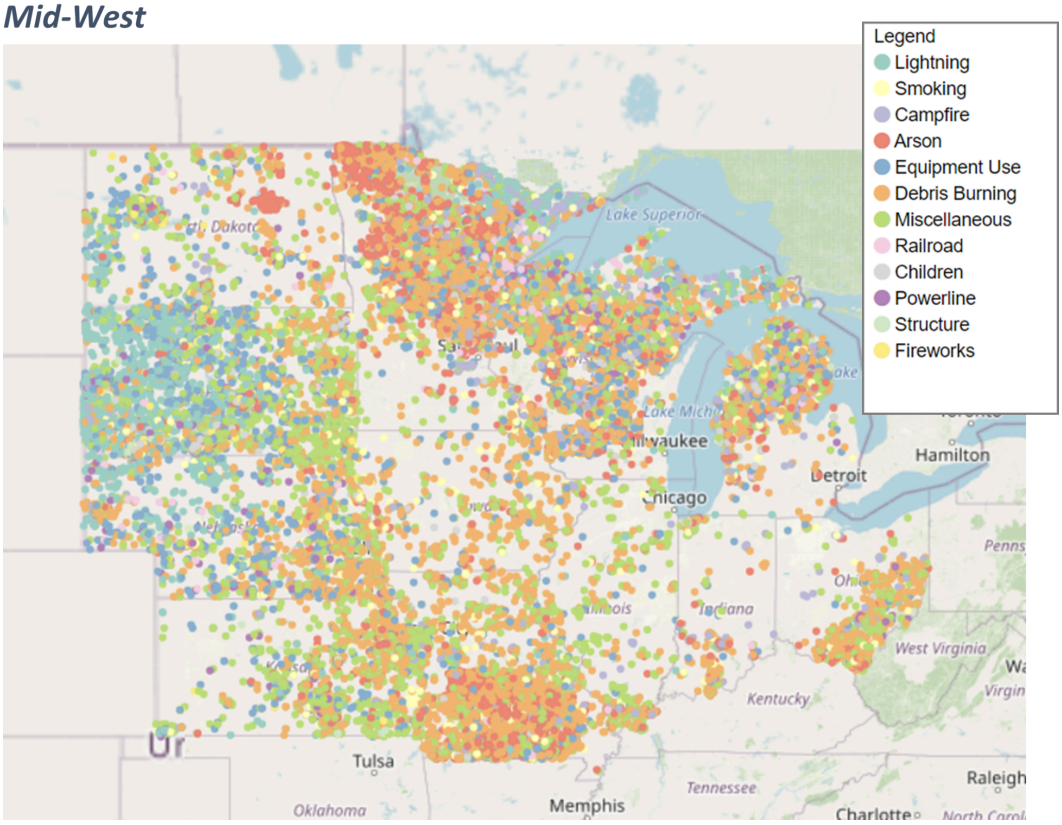
- The **Is Special Date** feature was found not interesting enough, so we decided to remove it. It was slightly interesting for Fireworks, so we decided to explore Fireworks fires over time, and found the following results:



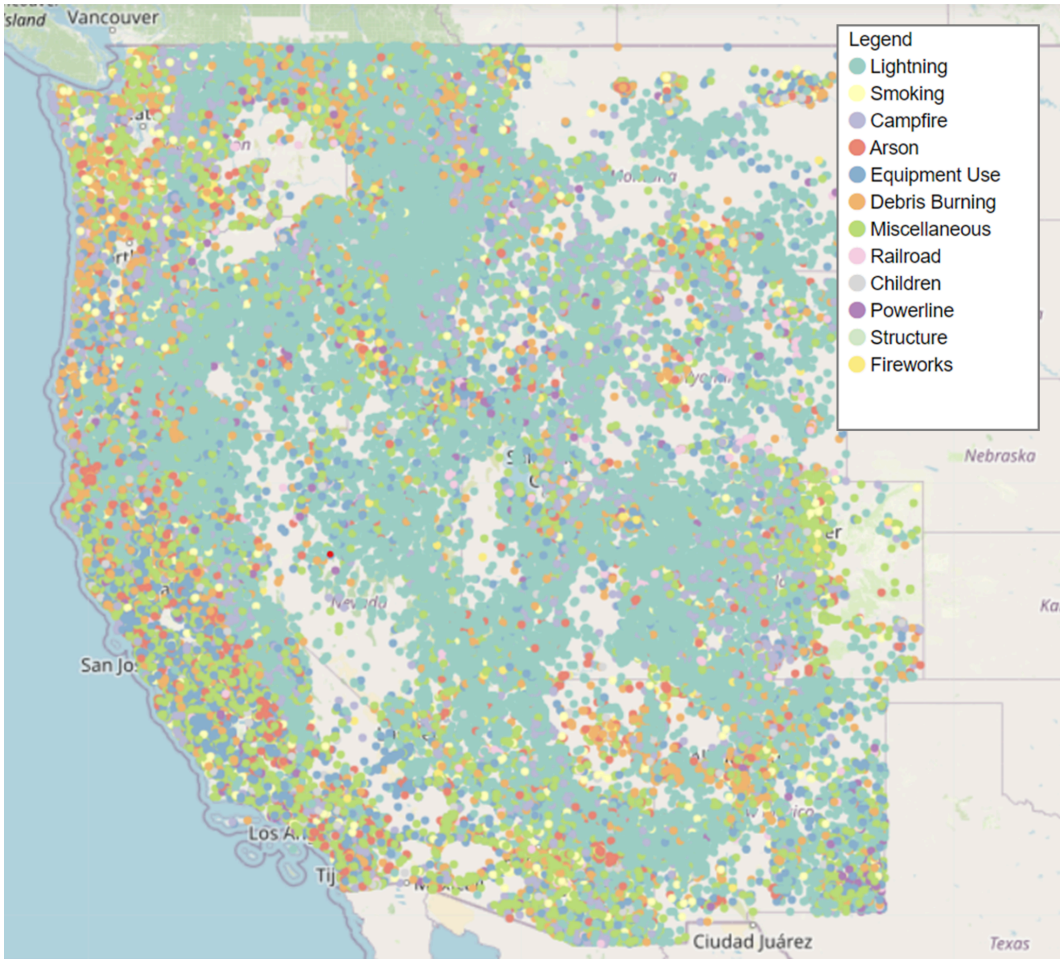
We found that most of the Fireworks fires occurred in the days close to 4th of July, American Independence Day. So we decided to add the feature '**Is Fourth Of July Range**' to the data. By performing the same analysis to different fire causes, we found that also Campfire, Arson, Miscellaneous and Children peaked in the number of fires on the 4th of July. Moreover, while exploring each label over time, we found that each label had a different trend, so we decided to add the **Day Of Year** feature.

- By making the same stacked bar plot from before to **Regions - West, South, Midwest and Northeast**, we found that the percentage of each fire cause varied significantly between each region. We also decided to split **Alaska** and **Hawaii** from the West region, because they had a unique behavior.
  In the following plots we can observe the differences geographically (we only show some of the plots):
  - Midwest - The West area of the Midwest is more prone to Lightning fires, whereas the middle part is more prone to Miscellaneous. The North and South areas have a high concentration of Arson and Debris Burning.
  - West - We can see that in total there are mostly Lightning fires in the West of the US.

Mid-West



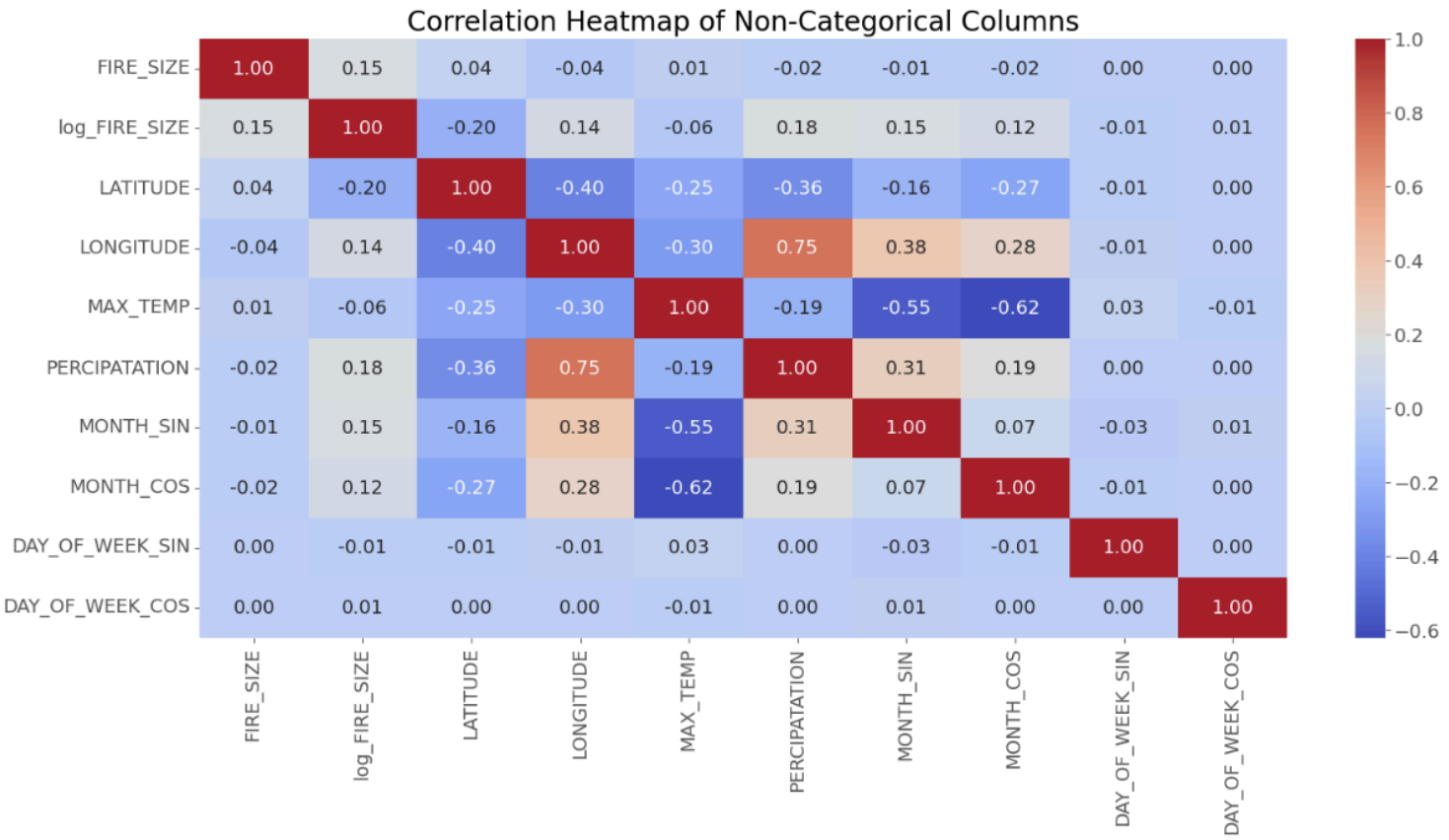
West



- For **Max Temp** and **Precipitation** we found that the distribution of the features looks different for every Fire Cause.

Correlations:

The Correlation Heatmap below (of only the non-categorical features) shows that there is a high and positive correlation of **0.75** between **Longitude and Precipitation** (which make sense because it depends on the location), and a high and negative correlation of **-0.62** between **Month and Max Temperature** (which also reasonable, of course).



🔥Model Training and Evaluation:

Splitting the data to train and test

After dropping unnecessary features and before the EDA we split the data to train and test, we decided to keep the distribution of the Y feature because as we explained earlier we saw that the distribution is not uniform, and some features have significantly less events than others.

Evaluation Metric:

The metric we used to evaluate the models is **ROC-AUC.** The Area Under the ROC Curve is the scalar ROC-AUC.

Baseline Model:

We applied two baseline models: one computes the frequency of fire classes in the training set and gets a random label based on this distribution, while the second is a simple decision tree. The resulting ROC-AUC scores were 0.5 and 0.57, respectively.

Insights from the training stage:

- After assessing the feature importance scores and observing that 'DAY_SIN', 'DAY_COS', 'STATE', and 'IS_SPECIAL_DATE' had low importance, we decided to remove these features from our models. Additionally, we examined the SHAP values for these features, which were, for the most part, centered around zero. We then reran our models and evaluated their performance.

- Running time vs model performance - we decided to drop the ‘STATE’ dummies, because the model is very large, it took us a lot of time and resources to fit and predict, moreover, the delta in performance was very small and even worse. So in this situation the running time won the tradeoff anyway.

🔥Model Selection:

Models:

We implemented several models:

- Logistic Regression: A linear model commonly used for binary classification tasks. We tried it because of its interpretability.
- XGBoost**:** An ensemble learning method known for its scalability and efficiency in handling large datasets. We tried it because we wanted to handle the high dimensionality of the data.
- Random Forest: Another ensemble learning technique that constructs a multitude of decision trees during training and outputs the mode of the classes of individual trees as the final prediction. Random forests are robust against overfitting and handle high-dimensional data well.

Cross Validation

We used the 'Optuna' library for hyperparameter tuning through cross-validation, aiming to find the optimal combination of parameters for each model. Following optimization, we compared the performance of these models against the baseline on the test set.

Comparison of model performances

We tested the models performances on the **test set** and found that the **XGB model** performed the **best** on both ROC-AUC score and Accuracy score.

For the following results, the key-value mapping between number and fire type is:

key	0	1	2	3	4	5	6	7
value	Arson	Campfire	Children	Debris Burning	Equipment use	Fireworks	Lightning	Miscellaneous

key	8	9	10	11
value	Powerline	Railroad	Smoking	Structure

- **Overall Performance**

Model	ROC-AUC	Accuracy
Logistic Regression	0.609	0.357
Random Forest	0.827	0.492
XGBoost	0.865	0.541

- **Logistic regression**

```
ROC AUC Score: 0.608681171802928
Accuracy: 0.35725945167971357

Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00     18597
           1       0.00      0.00      0.00      5075
           2       0.00      0.00      0.00      4040
           3       0.32      0.94      0.48     28521
           4       0.00      0.00      0.00      9840
           5       0.00      0.00      0.00       773
           6       0.45      0.72      0.56     18594
           7       0.44      0.03      0.05     21637
           8       0.00      0.00      0.00       946
           9       0.00      0.00      0.00      2184
          10       0.00      0.00      0.00      3492
          11       0.00      0.00      0.00       249
```

- **Random Forest**

```
ROC AUC Score: 0.8266181239787639
Accuracy: 0.4922508512654895

Classification Report:
              precision    recall  f1-score   support

           0       0.52      0.34      0.41     18597
           1       0.55      0.08      0.14      5075
           2       0.29      0.01      0.01      4040
           3       0.44      0.79      0.56     28521
           4       0.44      0.07      0.12      9840
           5       0.55      0.20      0.29       773
           6       0.62      0.83      0.71     18594
           7       0.47      0.46      0.46     21637
           8       0.00      0.00      0.00       946
           9       0.40      0.25      0.31      2184
          10       0.00      0.00      0.00      3492
          11       0.00      0.00      0.00       249
```

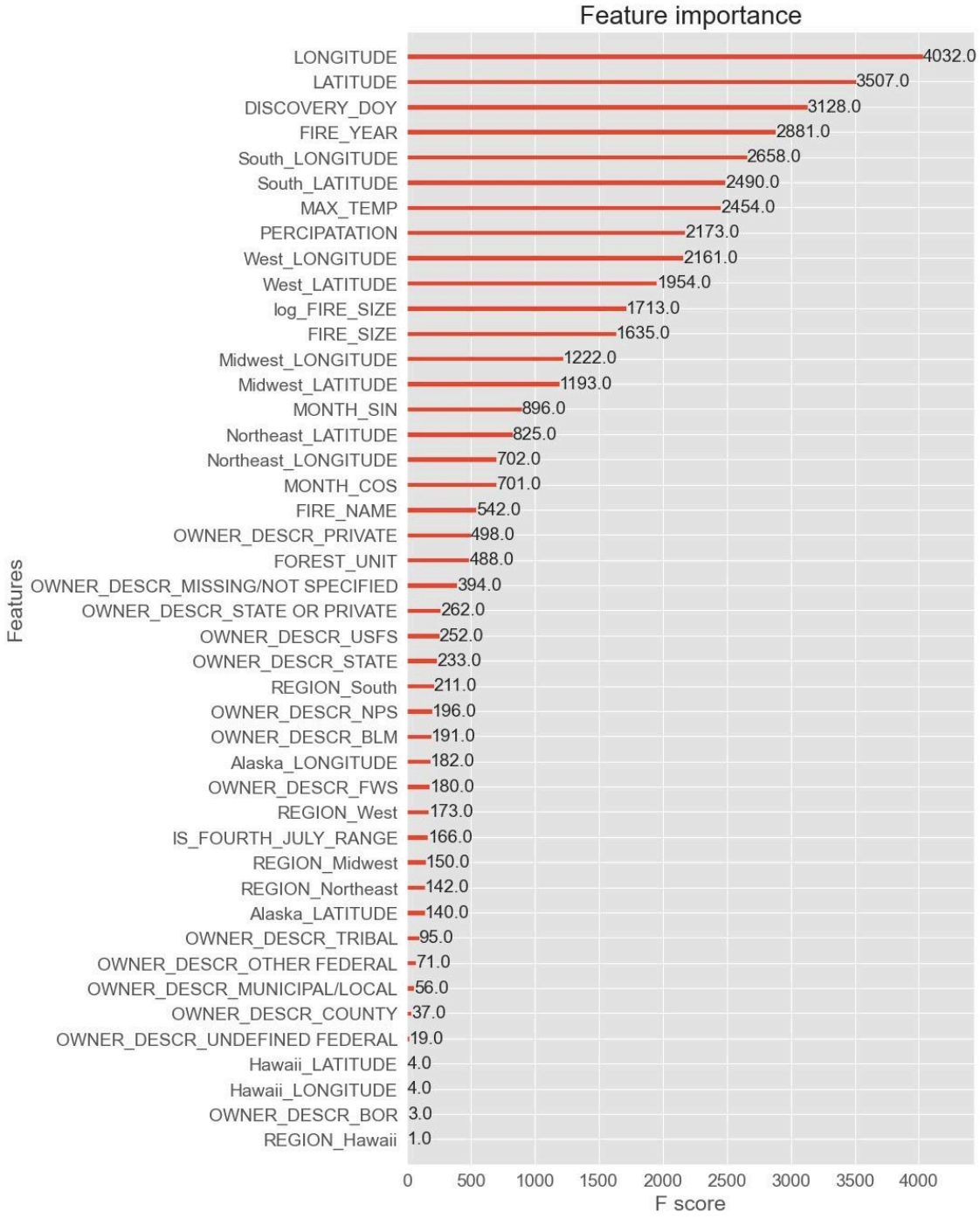
- **XGB model**

```
ROC AUC Score:
0.8648947263158694
Accuracy: 0.5405184821146488

Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.48      0.51     18597
           1       0.50      0.23      0.32      5075
           2       0.37      0.08      0.13      4040
           3       0.49      0.74      0.59     28521
           4       0.41      0.19      0.26      9840
           5       0.54      0.47      0.51       773
           6       0.70      0.84      0.76     18594
           7       0.51      0.52      0.52     21637
           8       0.38      0.02      0.04       946
           9       0.43      0.39      0.41      2184
          10       0.41      0.01      0.01      3492
          11       0.38      0.01      0.02       249
```

Model Interpretation:

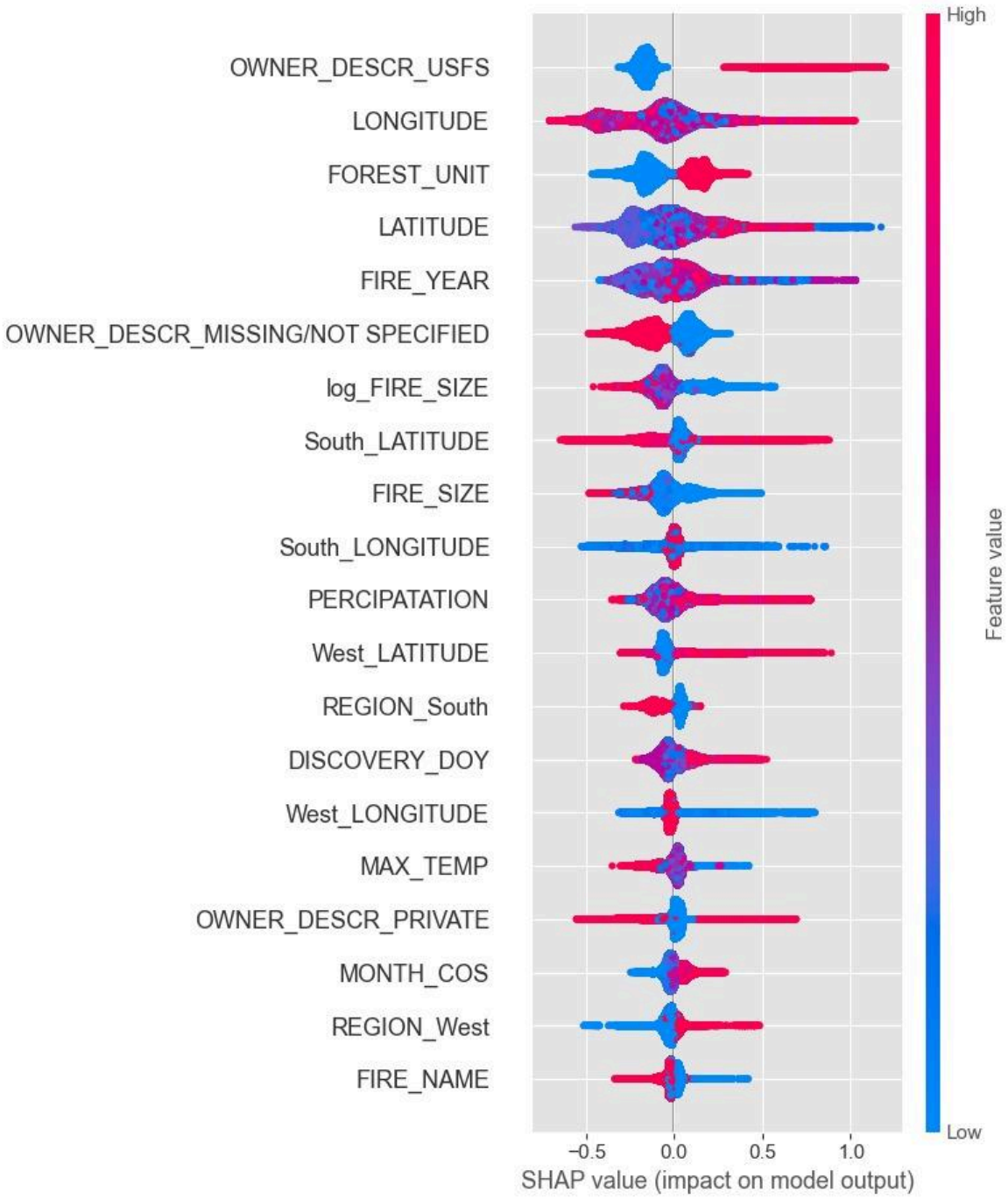
- **Feature importance analysis (over XGBoost)**

  We did a feature importance analysis several times during our model assessments, as explained above. As can be observed from the final feature importance, LONGITUDE and LATITUDE are the most influential features. Another influential feature is DISCOVERY_DOY, which is the day of year feature. This shows that as we suspected, the day in the year cycle has an influence over the result of the model.



- **SHAP analysis**

  This plot combines feature importance with feature effects. We can observe that some features with low feature importance values also have a low SHAP value - mostly centered around zero. For example, the REGION_south feature.
  We saw in the feature importance plot that LONGITUDE and LATITUDE are both with high scores. We can see here that indeed those features have a good impact in the graph below.



References:

- External dataset of weather taken from: https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/