CAUSE OF FIRE PREDICTION

A Data Science Project

Daniel Busbib, Nathanael Bohbot, Eli Chouatt, Aviv Lahat

The Hebrew University of Jerusalem
Applied Data Science Lab

1. Data Exploration and Feature Selection

The data set received had 39 columns and about 620,000 rows. The feature we needed to predict is the cause of the fire. In this data set, two features define the cause – the cause code, and the cause description.

The first action we took was to check that the objective feature columns are injective, meaning each code corresponds to one description, and vice versa. We were able to conclude that the cause code and description are injective and therefor do not require any data wrangling.

STAT_CAUSE_CODE	STAT_CAUSE_DESCR
1.0	Lightning
2.0	Equipment Use
3.0	Smoking
4.0	Campfire
5.0	Debris Burning
6.0	Railroad
7.0	Arson
8.0	Children
9.0	Miscellaneous
10.0	Fireworks
11.0	Powerline
12.0	Structure
13.0	Missing/Undefined

To start understanding the importance of each feature in the data set, we divided them equally between the group members. Each member explored the columns assigned to them by finding trends, unique values, creating visualizations and more, in order to determine if this feature is useful for further analysis.

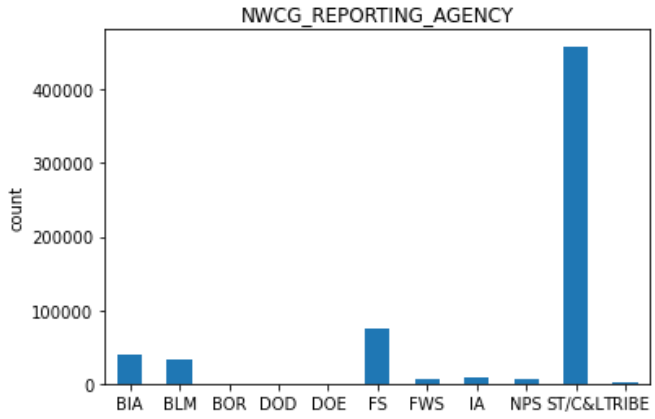
One observation was that in each fire incident reported, the features County, FIPS Name and FIPS Code were either all the same in that specific incident, or all None.

In the case where the County, FIPS Name and FIPS Code were not the same, we noted that 2 out of 3 were relatively similar.

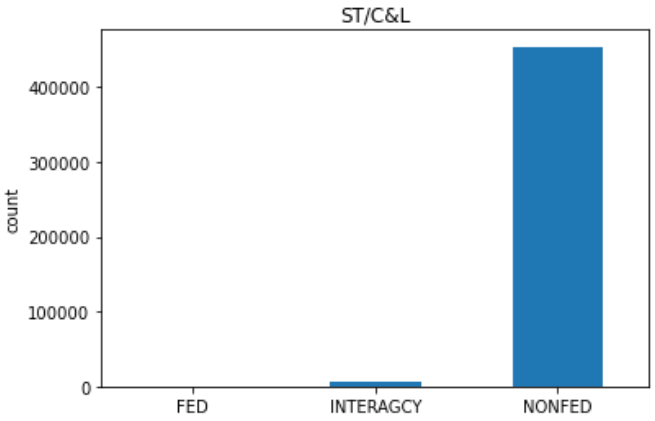
COUNTY	FIPS_NAME	FIPS_CODE
31	cook	031
17	navajo	017
orangeburg county	orangeburg	075
19	douglas	019
st lawrence	st. lawrence	089
jf	jefferson	073
5	coconino	005
47	okanogan	047
69	waupaca	135
stevens county	stevens	065

As a result of this observation, we decided to keep only one of these features.

A second observation rose from the exploration of the features grouped under the Overseeing Staff category. The fire records were drawn from either federal, interagency, or non-federal sources. Most of the fires were recorded by a state, country, or local organization, rather than by a national park service of bureau.



On top of that, the fires recorded by the states were mostly done by non-federal sources.



We wanted to look further into a possible correlation between the cause of the fire when reported by non-federal sources.

At this point, we decided to remove the features for which almost each entry had a unique value, and are not easily interpretable, such as Fire Name, Source Reporting Unit, Owner Description and more. After this initial feature selection, we kept 23 features out of 39.

2. Data Cleaning

Prior to the feature engineering process, we needed to decide how to deal with missing values in the features we want to insert to the model.

The features with missing values that most affected our progress were the missing time features.

About 50% of the reported fire incidents had missing values in the time features, so we chose to fill in the missing values instead of dropping them.

Our first approach was to fill in the missing values after the feature engineering. We used a linear regressor to predict the missing values based on all the fire incidents that had real values in our problematic features.

This approach was not percise due to two reasons:

a. Adding another model to the complete data analysis is heavy.
b. A few features we engineered are, for example, frequencies of occurrences. Filling in missing values with frequencies calculated before filling missing values results in both a bias and inaccuracies.

Therefore, we understood this process needed revising.

Our first idea was to use the linear regressor to fill the missing values before the feature engineering. This did not work because the raw data has non-numeric values that cannot be inserted into a model. Obviously, the model only worked after the feature engineering because all these values were encoded to numeric values.

The second idea also overcomes the inefficiency of a second model, which is to fill the missing values using estimation from the existing data.

To fill the missing values, we needed to make one assumption that would allow reverse engineering of all the other columns that have missing values.

The assumption we decided to make was to declare the discovery time of the fire as the average discovery time of fires in a certain state at a certain year in a certain fire class (and so on). From here, we engineered the other time features in order to fill in the missing values.

This method overcomes the disadvantages of our previous system to handle missing values.

We encountered an issue here that NaT values are not replaced in the pandas fillna method[1], these were later dropped.

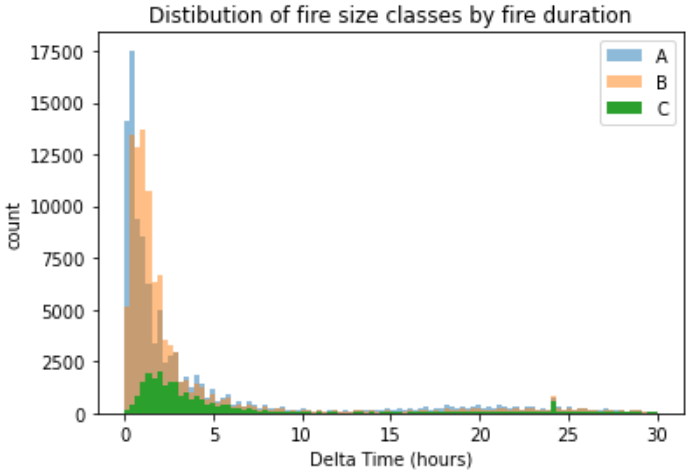
[1] https://github.com/pandas-dev/pandas/issues/11953

3. Feature Engineering

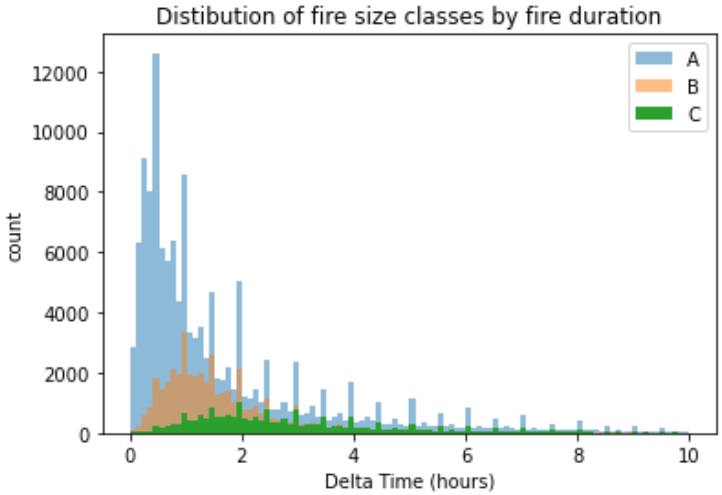
The feature engineering divided to three categories – time features and geographical features.

The first time feature we created is the duration of the fire. This feature logically seems useful to predict the cause of the fire because different causes result in different sized fires, leading to shorter or longer durations till the fire is contained.

This new feature led us to investigate the correlation between the duration of a fire and the fire size class that it is classified to. We found that the distribution of fires classified as A or B fires is too similar.



We tried different methods to make the distributions more different so that the classifying to fire classes will be significant. At first, we combined the fires classified A and B to one class, then we tried redefining the limit that defines which size fire is classified to each group A or B. We checked a couple of different boundaries between classes and chose to continue with a new upper bound for fire class A.

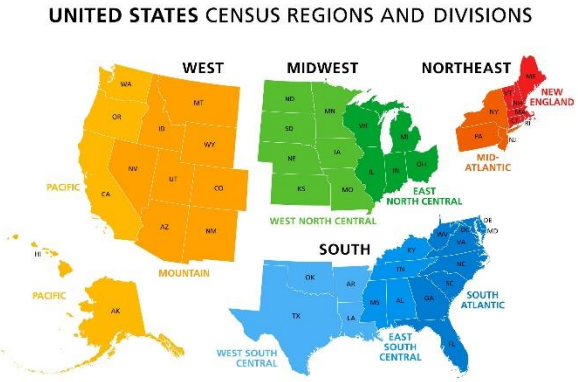


Now that we have a distinct distribution for each fire size class, we defined the median fire duration time of each class and used this value to fill in the missing delta times.

The encoding of the fire size classes is from 0-to-6, 6 being the class grouping the largest fires.

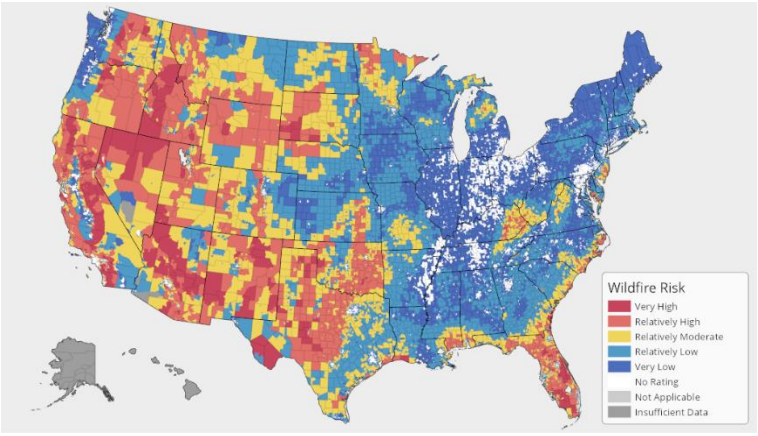
Next, we thought of creative features to describe the states, while preserving their properties, in order to have less dummies.

The first feature replaced the state with the region that state belongs to – North East, West, Mid-West, South, or Other. This feature preserves the states climate property. The encoding of this feature is the number of fires per region.



The second feature describe the state as the number of reported fire incidents in that state. This feature preserves the states terrain property.

The third feature describes the state by how fire prone it is. This data was found in the US Department of Homeland Security[2], and is in TIF files. We tried to work with the image data, but after a couple of hours requested permission to use the idea of the fire prone levels and define the states ourselves. The encoding of this feature is from 1-to-5, 5 being very high fire risk.



[2] https://hazards.fema.gov/nri/wildfire

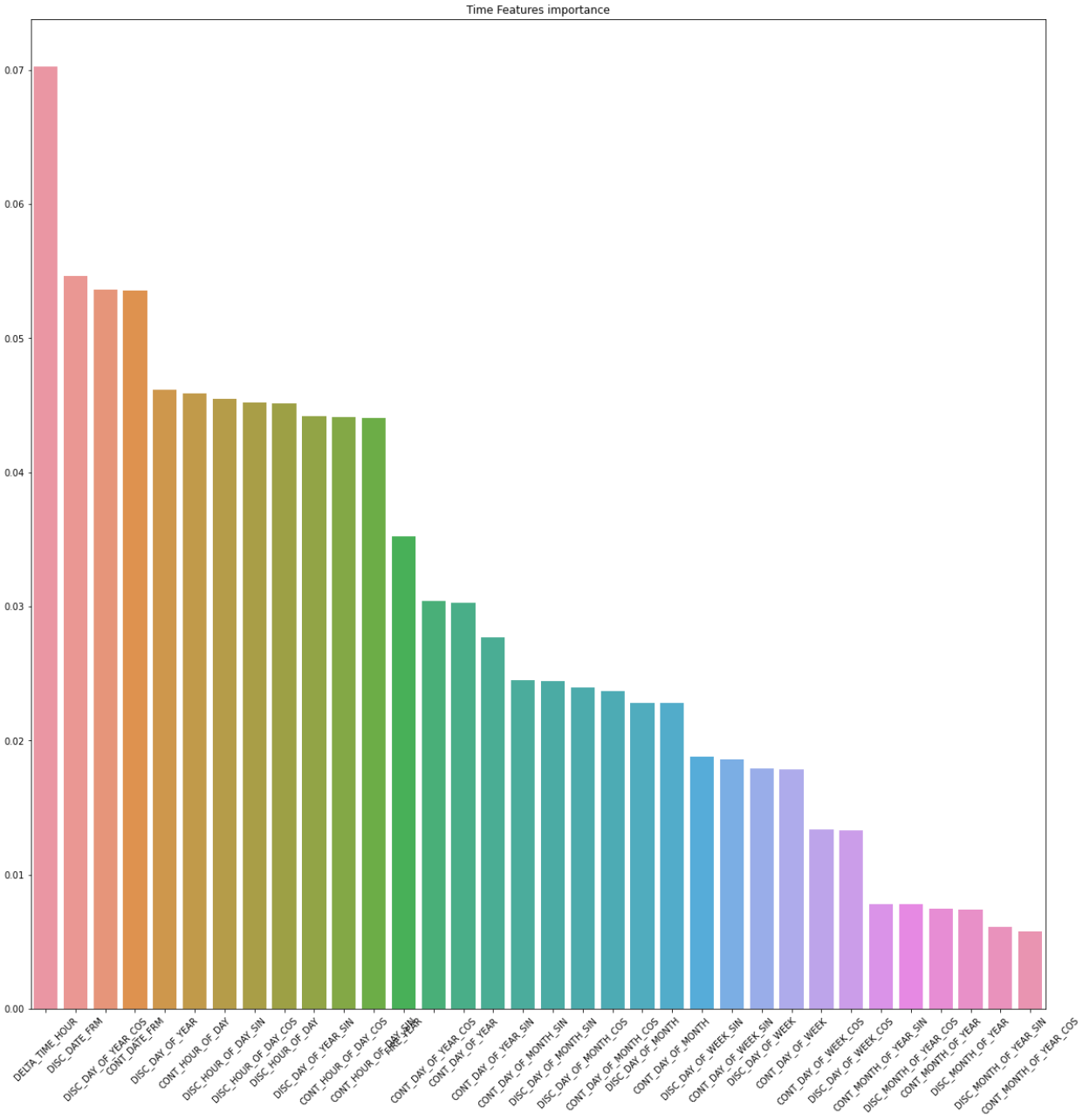
After this, we performed cyclic transformations to the time features to create more new time features such as the cosine and sine transformation of the year, month, week, and hour.

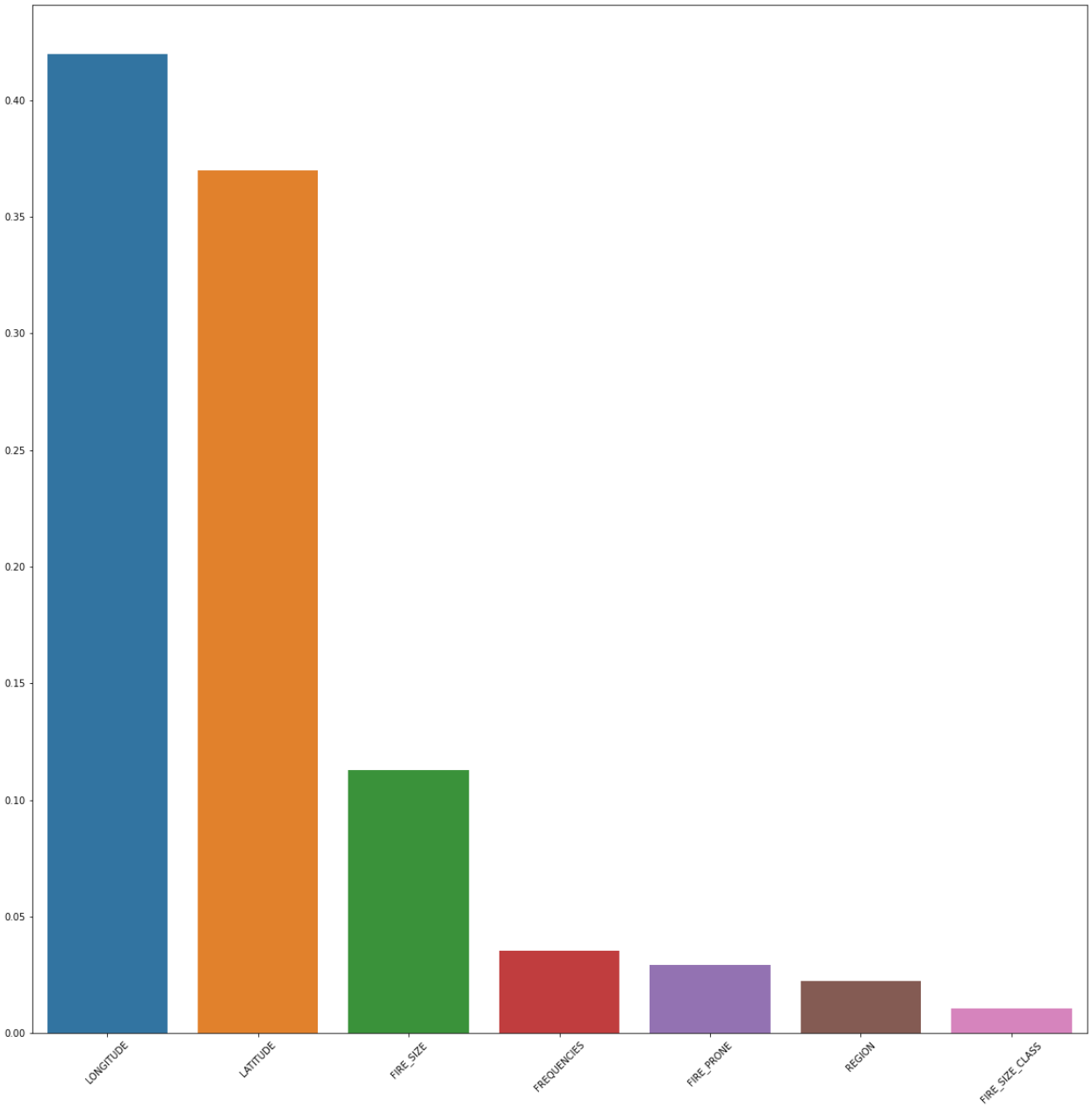
Finally, the third type of feature were weather condition features. To get this information, we find a dataset in Kaggle that report the storm events that happened in USA. And we added to our dataset a column that indicates if there were a storm event at the same time than the wildfire event in the state.

4. Feature Selection (again)

Our next step was to decide which of the features and cyclic transformations of features we want to use in our model.

We used a Random Forest Classifier seperatly on the time features and the geography features in order to declare which are most effective in each category.





At this turning point we made the decision to drop all the features in the Overseeing Staff category. We want the features in our model to be easy interpreted by being associated to the time and place of the fire.

The selected 14 final features for our model are – the latitude, the longitude, the fire size, the frequency of fires where the fire occurred, the fire prone level of where the fire occurred, the duration of the fire, the date when the fire was discovered, and the cosine and sine transformations of the hour of day when the fire was discovered and contained, and if there were a storm.

5. Model

In this project, we want to optimize the F1 score of our predictions.

After researching, we decided to use a Random Forest Classifier since it optimizes this metric.

To optimize our chosen model, we tried multiple models with a different number of estimators and accepted the model with 60 estimators.

The F1 score of our model averages at 0.7.

6. Model Evaluation

To evaluate our model, we used a confusion matrix in order to see which labels our model was able to predict correctly and on which labels it mixed up.

We can see that our model predicts the extreme edge labels correctly and that the center diagonal does have darker colors meaning more correct predictions, but in the middle areas the predictions are spread apart.