Group 21: Predicting the causes of wildfires

COMS 4995: Applied Machine Learning

Xinyu He
Dieter Joubert
Ritvik Khandelwal
Ethan Tucker
Keli Wang

Intro

Wildfires can have severe economic and social consequences in the United States, and the threat of global warming could impact the severity of wildfires.

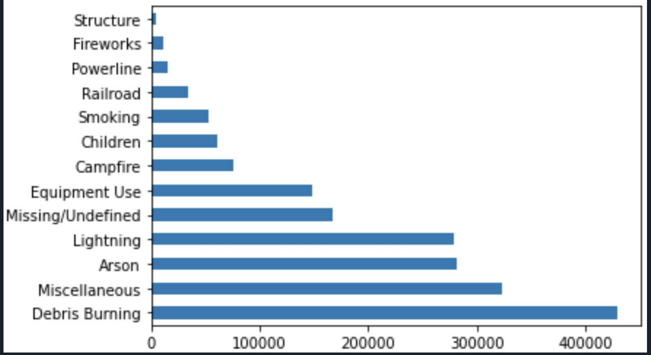
We will look at a spatial database (hosted on Kaggle) which consists of 1.88 million wildfires that occurred in the United States from 1992 to 2015.

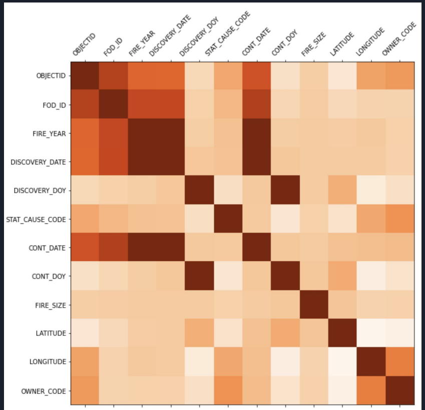
Questions to consider:

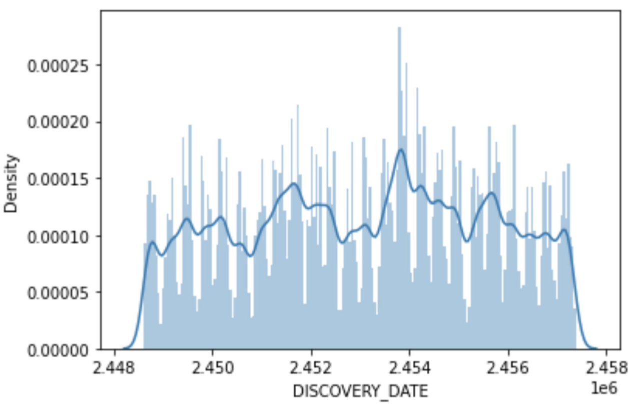
- Are the causes of wildfires predictable?
- What level of risk do various geographic areas have for wildfires?
- Has wildfires frequency increased over time?
- How quickly are different regions able to get wildfires under control?

Initial Exploration

- **(top right)** Histogram of wildfire causes in dataset
- **(bottom right)** Concept of a yearly fire season is clear from the 24 peaks in the 24-year spanning density plot of wildfire events.
- **(bottom)** Heatmap detailing the correlations amongst dataset features







Cleaning and Sampling

Feature Encodings:

- STAT_CAUSE_DESCR: Cause of the fires - Ordinal encoding
- STATE: State where fires occurred - Ordinal encoding

Repeated Data for Same Wildfire?

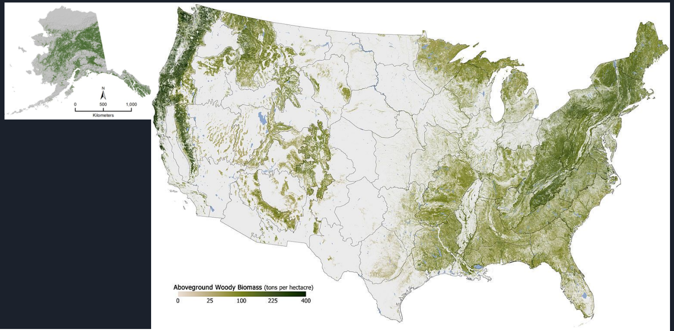
- Do we have multiple samples for the same wildfire reported at the same time? An insignificant amount of data, 8766 samples (0.4%)) were found to have non-unique DISCOVERY_DATE feature, but the repeats per non-unique sample was no greater than 1 for any case.

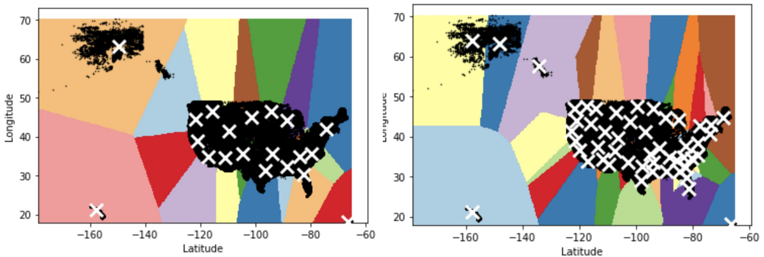
Cleaning and Sampling

Latitude and Longitude Data

To understand the effect of latitude and longitude without having issues of feature scale in the raw values we can use the standard scaler to set the values between -1 and 1. However, because the range of raw latitudinal and longitudinal distances considered in the dataset is different, after scaling we can weight each by their relative range. This will allow later comparison of these two spatial features, which may be important for example due to longitudinal temperature differences, and latitudinal differences in forest density.

Due to variations in forest coverage throughout the US, it will also be useful to use a clustering algorithm, such as k-means to create a new feature that can classify fire locations within a set of regions having contiguous forest coverage. In the top figure, we see the forest coverage for the contiguous US, and Alaska. The bottom figure to the right shows an implementation of K-means, which creates classifications for fires occurring in different regions. This is a hyperparameter to explore further in order to understand how it relates to the continuity of the forest coverage. If wildfires are often spreading between clusters, its likely they're connected, and should be classified as a single cluster.





K-means with 20 (left) and 50 (right) clusters

Insight from Data Exploration

Size of Wildfire

- Most of the fire size based on the number of acres within the final fire perimeter expenditures are between 0.26-9.9 acres.
- The estimate of acres within the final perimeter of the fire tends to be larger in spring (Feb. and Mar.) and winter (Nov., Dec. and Jan.), while it becomes smaller in summer.
- Arson is the cause of fire that results in the largest fire size on average, while campfire results in the smallest fire.

Frequency of Wildfire

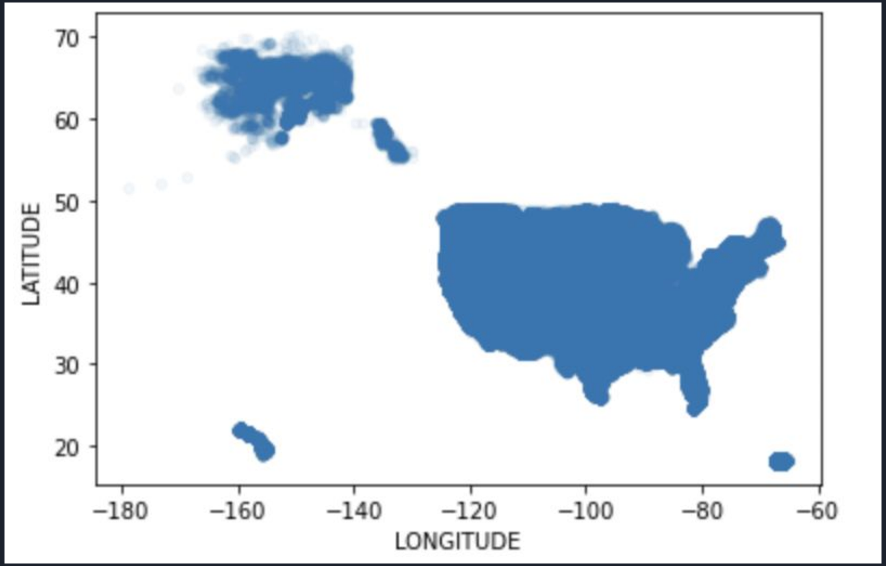
- More wildfires occurred in spring and summer.
- Slightly more wildfire occurred during weekends.

Wildfire Management

- Most of the wildfire were contained within 7 hours or 1.5 days after discovered.

Insight from Data Exploration

- Distribution of wildfires according to longitude and latitude



Machine Learning Techniques

To predict the causes of the wildfire, we plan on experimenting with the following machine learning techniques:

- ➔ [D] Multinomial Logistic Regression - predict categorical placement in or the probability of category membership on a dependant variable based on multiple independent variables.
- ➔ [D] Support Vector Machines - deterministic algorithm which maximize the margin the between the closest support vectors whereas logistic regression (probabilistic) maximizes the posterior class probabilities.
- ➔ [R] Decision Tree - a greedy algorithm following a rule-based approach to find the non-linear decision boundary with minimal preprocessing making it easier to interpret and deploy. It is also invariant to the scale of the data

- Ensembling - With the hope of decreasing the model variance we will combine several weak learners to make the final prediction. We can also check the correlation between individual learners based on the performance.
- ➔ [R] Random Forest - To generalise better on unseen data, we combine a whole bunch of simpler trees. The random forest performs ensembling while also randomly selecting a subset of the features on which to seek optimal splits at each node.

Machine Learning Techniques

- [K][X] Boosting - weak learners are learned sequentially with early learners fitting simple models to the data and then analyzing the data for errors. Fundamentally, output of the tree is adjusted based on the misclassifications.

➔ Gradient Boost Classifier - potentially not that helpful as our dataset is very large nevertheless it would be interesting to see how it performs on our data.

➔ Histogram Gradient Boosting Classifier - Histogram based split finding in tree learning. Could be very useful for us as it is exponentially fast on large datasets and it natively supports categorical features.

➔ XGBoost - If the need arises that we want to make sure certain features don’t interact while making a prediction, then we can use XGBoost as it enforces monotonicity constraints.

- [X][R][K] Artificial Neural Networks - It will be interesting to explore how neural networks handle tabular data. We can potentially experiment with different state-of-the-art neural network architectures and observe how they impact the model performance.

Machine Learning Techniques

To predict the likelihood of fires occurring over time, and to understand quantitatively the temporal-dependance of increasing frequency and scale, we intend on doing time series analysis wherein the following techniques can be used:

1) Autoregressive integrated moving average (ARIMA) [optional?]
2) Convolutional Neural Networks (CNN) [optional?]
3) [E] Vanilla Recurrent Neural Network (RNN)
4) [E] Long Short Term Memory (LSTM)
5) Bidirectional LSTM (BiLSTM) [optional?]

So the goal of using various different techniques mentioned above is to achieve the best possible model while inferring why a particular model performs better or worse than the other.