Predicting the Cause of Wildfires Using Machine Learning

Brett Cotler

DATA 1030 - Final Report - December 3, 2020

1 Introduction

This machine learning experiment considers a data set titled ”1.88 Million US Wildfires” [1]. The data set is comprised of information about wildfires in the United States from 1992 to 2015. The data was collected for the national Fire Program Analysis system from state, local, and federal firefighting organizations. Using this data, I am hoping to determine whether the cause of a wildfire can be predicted by its date, location, and size. This is a classification problem, as the target variable, the cause of the fire, is categorical.

The importance of an experiment like this cannot be understated. As climate change fundamentally alters weather patterns around the world, the west coast of the United States continues to become hotter and drier. Over the past decade, the prolonged and intensifying fire season has devastated massive regions of California, Oregon, and Washington. Climate change is the underlying cause of these fires, and mitigating climate change will be necessary to save these forests. But climate change is already here, and it is therefore also necessary to incorporate forest management into climate adaptation practices. Predicting the individual actions which cause forest fires to start is a critical step in improving forest management.

As described in the title, this data set has 1.88 million data points. It includes a litany of features, but for the purpose of answering this question, I selected only the features which will provide potentially predictive information about the target variable. After this selection, the data set has 8 variables. The data set does have extensive documentation and the selected features are straightforward.

1. Fire Year: The year the fire occurred

2. Stat Cause Description: This is a categorical variable which describes the cause of the fire. Some examples of causes include ”Lightning” and ”Debris Burning.” This is the target variable for the experiment.

3. Latitude: The latitudinal location of the fire.

4. Longitude: The longitudinal location of the fire.
5. State: The state where the fire occurred.
6. Fire Size: A measure of the number of acres in the final perimeter of the fire.
7. Month: The month when the fire was discovered represented as the number of the month.
8. Day: The date when the fire was discovered.

This data set was posted to Kaggle in 2017 and there have been many notebooks published on its page. One interesting notebook was published by Troy Walters on September 17, 2017. This notebook conducts basic exploratory analysis on the wildfires data set using R. It specifically focuses on the question of fire duration, identifying areas which have poor fire response and management. A couple of states identified that experience long fire durations were Idaho and New Jersey. This analysis also includes a chart of the most common causes of fires and therefore could be useful in framing my analysis of the cause of fires [4].

This data set has also been cited in academic research. A 2018 article titled ”A weekly, continually updated data set of the probability of large wildfires across western US forests and woodlands” cites this data set in its efforts to make predictions of where and when large wildfires will occur. They used a random forest machine learning model to predict the probability that an individual pixel on a Google Earth image will be a part of a forest fire in a given week [3].

2 Exploratory Data Analysis

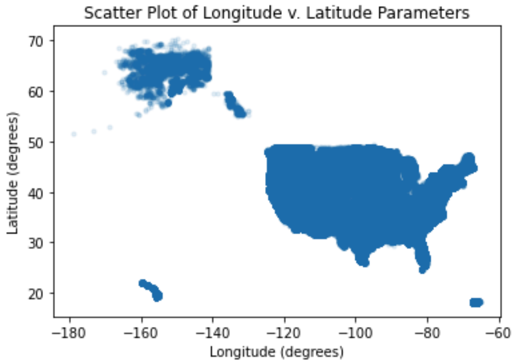

Figure 1: Scatter plot of longitude v. latitude of fire locations. This figure illustrates the precision of the data as the scatter clearly outlines the shape of the United States. Furthermore, this graphic is of importance in EDA because it illustrates that the forest fire data covers the entirety of the US; it is not limited to the regions which are traditionally considered to be fire prone.

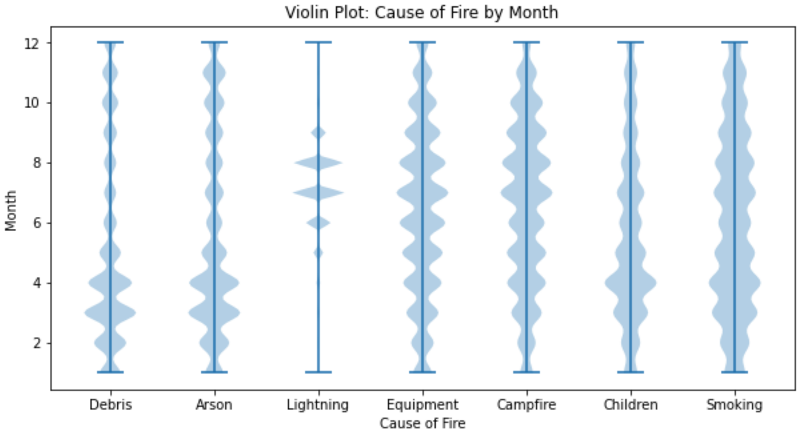

Figure 2: In this graphic, I selected 7 of the 9 most common fire causes to make a more readable violin plot. I compared the fire cause with the month of the fire. This figure is an interesting way to consider the distribution of each category of the target variable over time. It indicates that there is a temporal dependence of the cause of a fire for lightning and arson, but there is not a temporal dependence for smoking.

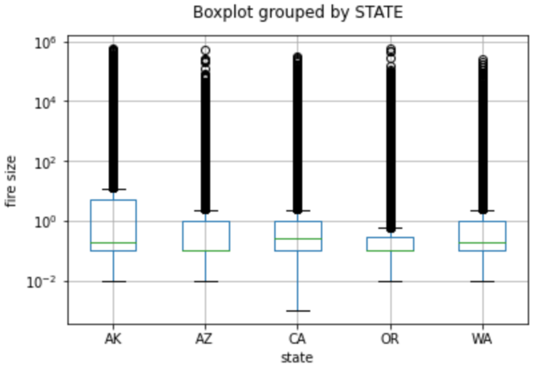

Figure 3: In this figure, I selected 5 fire prone states: California, Oregon, Arizona, Alaska, and Washington. I then created a boxplot distribution of fire size in these five states. This figure illustrates the difficulty in creating useful graphics over such large data sets, as the traditional median and inner quartile box is not visible on this plot. But I do think this figure is still valuable, as it emphasizes that a majority of the fires in the data set are small.

3 Methods

3.1 Splitting

Before I completed splitting to apply the machine learning algorithms, I decided to take a small random subset of the data to complete my experiment. I used a Stratified Shuffle Split to select one one-hundredth of my data. This reduced the number of data points (rows) from 1.88 million to 18,000. I decided this was necessary because it significantly reduced computation times.

After splitting out this subset, I chose to use a basic split for my data. Now that the dataset has only 18,000 data points, I found it most appropriate to use a 60-20-20 train-validation-test split.

When implementing the models, I first applied a 80-20 train test split to the data to create ”other” and ”test” sets. Then, as part of my GridSearchCV, I used a kfold split with 4 splits to split out the last 20 from the remaining 80 in the ”other” set.

Although I chose a basic split, it is possible that the data points are not IID, as a fire in a specific area at a specific time may increase the propensity of other fires in the area. But for this exercise it is necessary to treat the data as if it's IID because there isn't a straightforward way to measure that dependence. Furthermore, the data set could be grouped by the cause of the fire but there isn't a clear group structure.

3.2 Preprocessing

My preprocessed data has seven features. I used all four primary encoders across the seven features. I chose to preprocess the months using the ordinal encoder because it is a well-ordered feature with a finite number of categories. I chose to preprocess the state feature with the onehot encoder because it is a categorical feature with no order. The remaining features are continuous. The Fire Year and Day features were scaled using the MinMax scaler because both features have well-defined start and end values. Longitude, Latitude, and Fire Size were scaled using the standard scaler as these continuous features do not have well-defined start and end points. And finally, I used the label encoder to encode the different fire cause target values.

3.3 Machine Learning Pipeline

My machine learning pipeline utilized GridSearchCV to efficiently tune hyperparameters to determine the best parameters for a given input model.

I wrote a function called ”apply algo” which takes in a preprocessor, a machine learning algorithm and a random state. The function first splits the data into ”other” and ”test” using train test split. It then sets up the kfold and the preprocessor/machine learning algorithm pipeline. And finally the function defines the grid search and fits the algorithm to the input data. The model and the test data are the function outputs.

I decided to compare 5 different algorithms to find the best model for the fires data set. The five classification models I compared were Naive Bayes, Logistic Regression, Decision Tree, Random Forest, and K Nearest Neighbors. I also considered using a support vector machine, but the higher training time for a support vector machine would likely have been prohibitive.

The Naive Bayes and Logistic Regression models do not require hyperparameters. Logistic Regression used a default l2 regularization with C = 1 and tolerance = 0.0001. For both the Decision Tree and Random Forest models, I used a variety of different tree depths. I used the standard 1, 3, 10, 30 as the tree depths for the Decision Tree. I wanted to cast a wide net to ensure that the optimal model was discovered. For the Random Forest, I used depths of 1, 3, and 7 (the number of features). Because of the computational intensity of the Random Forest model, I found it necessary to reduce the size and quantity of the max depth parameters. The Random Forest model also requires a max features parameter and a random state parameter. I tried max features of 0.5, 0.75, and 1 to tune the number of features considered at each decision point. At least half of the features should be considered at a decision point to ensure that the splits are representative of several features. I set the random state to be 12 for reproducibility. And finally for K nearest neighbors, I tuned the number of neighbors parameter with the values 1, 3, 10, and 30 to again cast a wide net.

To evaluate model performance, I decided to simply use the accuracy score because this is a classification problem with fairly balanced data. GridSearchCV used the accuracy score to select the optimal parameters for each model I provided.

4 Results

4.1 Accuracy Scores and Uncertainty Estimates

The table below provides the baseline and model accuracy scores, model uncertainty, and the number of standard deviations the model is from the baseline.

Model	Baseline	Mean Acc.	Std. Acc.	Std. from Baseline
Naive Bayes	0.17	0.05	0.0030	70.40
Logistic Regression	0.17	0.39	0.006	131.37
Decision Tree	0.17	0.431	0.010	153.38
Random Forest	0.17	0.42	0.0	146.67
K Nearest Neighbors	0.17	0.427	0.0	150.27

One positive outcome of this experiment was the relative certainty of these models. The highest model standard deviation was 0.01 for the Decision Tree. Low standard deviation values suggest that the model performance is not dependent on the random split of the data.

Unfortunately, the accuracy scores are extremely low. The highest accuracy score comes from the Decision Tree model, a surprising outcome because a

Random Forest is an ensemble Decision Tree, and should therefore perform better. The decision tree accuracy was 0.431, which means that this model predicted the correct class of the test data less than half of the time. The accuracy levels of the models were a significant improvement upon the baseline accuracy of 0.17: a minimum of 70 standard deviations above the baseline. But a maximum accuracy of 0.431 is not high enough to suggest that the features of the data set predict the cause of a forest fire.

4.2 Feature Importance

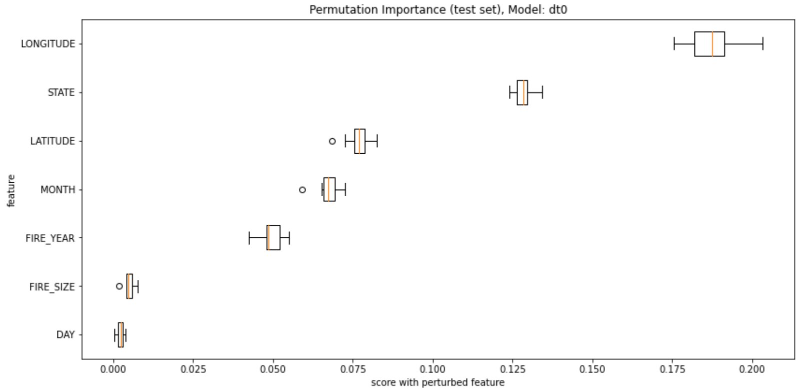

Figure 4

Figure 4 shows the global permutation importance of each feature for one iteration of the Decision Tree model. This figure illustrates that longitude is the most important predictor of fire size. Climatic and forest conditions vary widely longitudinally across the United States and therefore might account for the variation. The state and the latitude are the next most important predictors, further emphasizing the importance of location. The permutation importance also indicates that the Month when the fire occurred contributes to the class prediction. This is likely a consequence of seasonal variation in climate, which increases the propensity of certain causes of fires. Fire size and day were not important features because the fire size is not determined by fire cause started, and the day of a month does not capture seasonal variation.

5 Outlook

There is a lot of room for improvement in terms of the outcomes of this machine learning exercise. The most accurate model after cross validation and hyperparameter tuning didn't predict the classes correctly half of the time.

The first possible improvement should be including all of the data. Taking a small subset of the data limited the ability of the models to produce accurate fits. The second opportunity for improvement requires consideration of the dependence of the location features. Although there is some climatic and forest continuity across borders, the state where the fire occurs and the longitudinal and latitudinal location of the fire are highly dependent variables. Therefore, the model effectively has fewer features to distinguish fire cause. Another way to increase the precision of the model would be to include more hyperparameter values in the parameter grids that were inputs to GridSearchCV. Adding hyperparameter values would likely increase the resolution of hyperparameter tuning in the cross validation stage.

6 Github Repository

https://github.com/b-cotler/data1030-wildfire-project

References

[1] Tatman, Rachel. 2017. 1.88 Million US Wildfires: 24 years of geo-referenced wildfire records. Kaggle https://www.kaggle.com/rtatman/188-million-us-wildfires?

[2] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA FOD 20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

[3] Gray, Miranda E., Luke J. Zachmann, Brett G. Dickson. 2018. A weekly, continually updated dataset of the probability of large wildfires across western US forests and woodlands. Earth Syst. Sci. Data, 10, 1715–1727, 2018. https://doi.org/10.5194/essd-10-1715-2018

[4] Walters, Troy. Wildfire Exploratory Analysis. Kaggle. 2017 https://www.kaggle.com/captcalculator/wildfire-exploratory-analysis