Predicting California's Wildfires

By: Janzen Hui
Date: April 4, 2022

Background:

Forest fires seem to be getting increasingly prevalent and problematic. The United Nation's Environment Programme (UNEP), which sets the United Nation's environmental agenda, has predicted that "the likelihood of catastrophic wildfires events will increase by a factor of up to 1.57" by the end of the century [1]. As wildfires worsen, so do their economic, environmental, and societal impacts. Understanding the changing characteristics of these fires will be important to mitigate and adapt to these increased risks.

Currently, universities and governments use state-of-the-art physical simulators to predict forest fires. Unfortunately, these simulators are limited when there is a large number of parameters. They are also often biased toward the regions it was designed for. Data science and machine learning techniques are cheaper and offer increased flexibility when it comes to adding or removing features. Due to the plethora of weather, geographical, and satellite data, machine learning can consolidate different data types to determine the key driving forces of these fires. Previous machine learning studies have used these data types to develop linear and regression tree models [2].

Problem Area:

In California, the Department of Forestry and Fire Protection (CAL FIRE) expects that they will "experience longer wildfire seasons as a result of climate change" [3]. This study attempts to use machine learning models to predict the intensity of California's wildfires and to classify them as either large or small. Analyses will mainly utilize weather data to investigate its determining role in large forest fires. This predictive model will allow governments and first responders to effectively allocate resources to areas of need or for administrative planning such as budgeting and hiring.

Dataset:

California's Department of Water Resources (DWR) hosts a network of weather stations around the state. The weather data was downloaded from the DWR as CSV files. The weather dataset contained over 400 thousand rows and 44 columns of data. Details from 264 weather stations were collected as a JSON file using an API. An account with the DWR was also required. The weather and station data consisted of coordinates, elevation, and miscellaneous weather information such as precipitation and temperature. The forest fire data was downloaded as a CSV from the National Interagency Fire Center. It contained information on more than 2.1 million reported wildfires in the United States. The data tracked California's wildfires from 1992 to 2018. This dataset contained administrative information, fire start and end dates, location, and the fire size.

Data Processing:

The three datasets were individually processed, filtered, and irrelevant columns were removed. They were then combined based on two conditions. The fire start date was matched with the weather collection date and the fire's location was matched according to its nearest weather station. The final

cleaned dataset contained 234 thousand rows and 21 columns that could be used to classify the wildfire's significance. Exploratory data analysis (EDA) included examining the feature distributions independently to look for erroneous values and to check for normality. Variables with limited variance may be poor predictors within the model and were removed. Using the fire size, a new target binary variable was feature engineered to classify fires as large or not. Univariate analysis was performed on each independent feature with the dependent variable fire size and the binarized large fire. A time-series EDA was also performed to determine how the burn area and the number of large fires change over time.

Modelling Summary:

Various models were used to predict the fire size. A baseline time series model was used to determine how the fire size changes over time. Due to the fire size ranging from 0.1 acres to more than 410 acres, it was difficult to forecast the fire sizes with precision. Additionally, the majority of fires were less than 10 acres in size, causing a model imbalance. Using SMOTE-NC, the data imbalance was resolved by up-sampling the target minority class.

Five different machine learning models were built to categorize the fires. All models used grid search and cross-validation to optimize the hyperparameters. Randomized searches were also used for more computationally expensive models. Models were optimized and scored by the F1 scores as both precision and recall are important metrics that would ideally be maximized. False negatives would mean a fire is classified as small when in reality they are significant. False positives would result in fires being classified as large when in reality they are small. High precision and recall would mean low false positives and negatives.

First, logistic regression models were used as a baseline linear classifier. PCA was also combined with logistic regression to analyze features along a different dimension. Decision tree, random forest, and XGBoost models were also used, which utilize non-linear relationships. Models were assessed and compared using the F1, recall, precision and AUC scores according to Table 1. The random forest model has the highest AUC and F1 scores and was deemed the optimal model. The AUC score is representative of a model's ability to differentiate between the positive and negative classes.

Classification Model	F1 Score	Recall	Precision	AUC Score
Logistic Regression	0.15	0.55	0.09	0.57
Logistic Regression & PCA	0.14	0.66	0.08	0.55
Decision Trees	0.16	0.60	0.09	0.59
Random Forest	0.18	0.56	0.11	0.61
XGBoost	0.14	0.11	0.20	0.54

Table 1: Classification model's assessment metrics

The important features within the decision tree, random forest, and XGBoost models were assessed using the permutation feature importance calculations. The normalized importance percentage represents the predictive power of each feature within the model. The findings form the random forest model are described here and in Figure 1:

- **Cause of fire:** The cause of fire is the greatest predictor for large fires. 80% of large fires are started by humans.
- **Date:** The fires start date is a strong predictor for large fires. Interestingly, the month is a weak predictor. In contrast, the time series analysis shows that large fires are more prevalent during summer.
- **Location:** The regional features, such as longitude, latitude, and elevation, are moderate predictor for large fires.
- **Weather:** The average air temperature is a strong predictor, but many of the other weather features are fairly weak. Interestingly, precipitation does not have any predictive power in this model. As seen in its feature distribution, the precipitation in California is limited.

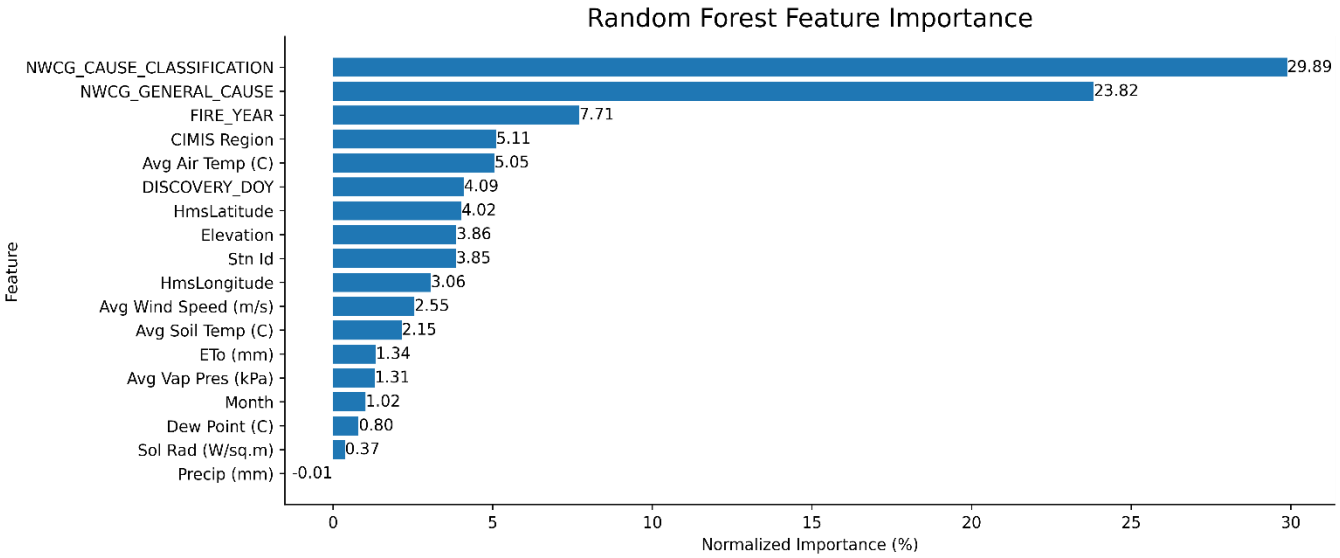

Figure 1: Optimized Random Forest Model Feature Importance Visualized

Conclusion and Next Steps:

A model has been generated that can be used to predict large fires. A 0.61 AUC score indicates that the model accurately differentiates between the large fires and small fires 61% of the time. There is room for improvement depending on the needs. Optimizing for precision may be useful for government bodies, as the majority of fires are small and having lots of false positives may lead to budget overinflation and gathering many unnecessary resources. Optimizing for recall may be more beneficial for first responders who want to limit the false negatives. Fires classified as false negative would mean large fires are classified as small. First responders may not have sufficient resources when responding to false negatives. Both precision and recall are important, but optimizing both may not be necessary depending on the target audience.

As machine learning improves and additional data is collected, the models will continue to improve. Many factors influence the intensity of the wildfires. Many of which are nuanced and are highly dependent on the specific microclimate. Dialing into a wide range of weather features, as was done here, may not be the best approach to generate a highly reliable model that fits all climate types. Models can be generated for specific climate regions around California and then applied to similar

environments around the world. Utilizing different data types in combination will also improve these models. Governments and communities will be able to use these models to plan for the future, by creating budgets, and gathering resources early. Locals can use these models to avoid areas with high wildfire risk. Firefighters can implement fire mitigation strategies in high-risk areas. As the climate changes, understanding the impact of wildfires is becoming more and more valuable to our society.

References:

1. Environment, U. N. (2022, February 23). *Spreading like wildfire: The rising threat of extraordinary landscape fires*. UNEP. Retrieved April 3, 2022, from https://www.unep.org/resources/report/spreading-wildfire-rising-threat-extraordinary-landscape-fires
2. Pourghasemi, H. R., Gayen, A., Lasaponara, R., & Tiefenbacher, J. P. (2020). Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling. *Environmental Research*, *184*, 109321. https://doi.org/10.1016/j.envres.2020.109321
3. California Department of Forestry and Fire Protection (CAL FIRE). (n.d.). *Incidents overview*. Cal Fire Department of Forestry and Fire Protection. Retrieved April 3, 2022, from https://www.fire.ca.gov/incidents/