Springboard Capstone Project

Wildfires in California:

Using machine learning to predict the likelihood of a fire to grow 'big'

Lena Berger

Mentored by Nik Skhirtladze

Disclaimer

This project was done as part of a data science boot camp. All stakeholders mentioned are fictional. This study was not commissioned by any agency or stakeholder. The project makes use of fire data collected by Short (2017) and data provided by NOAA. Some of the findings presented in this document are based on a subsamples of these datasets and are not representative of and should not be generalized to the entire database without further due. Please contact the author previous to use or further distribution of the findings or analyses. The author does not take any responsibility or liability for consequences resulting from use of these analyses.

Contents

"Wildfire will increase throughout the United States, causing at least a doubling of area burned by the mid-21st century"
(Vose, Peterson, & Weynand, 2012, p.V)

1. Introduction

Even though the number of wildfires in the United States has slightly decreased over the last 30 years, the number of acres burnt has increased (Hoover & Hanson, 2019). Since 2000, every year "an average of 72,400 wildfires burned an average of 7.0 million acres" (Hoover & Hanson, 2019, p.1). As recent examples of fires in Australia, New Zealand, and the United States illustrate, wildfires can have devastating consequences for both human and nature including loss of life, property, and natural habitat. While the size of a fire is not necessarily an indicator for its impact on human settlements or eco-systems (Hoover & Hanson, 2019), the increase in fire activity and fire size is a matter of concern.

One of the hardest hit states in terms of wild fires is California. On average California experiences 2.821 fires burning about 70,719 acres per year (5-year average; Cal Fire, 2020). Fire activity is related to weather patterns and changing climates (Reinhardt, 2015). The relationships are complex as climate—e.g., dry climate—enhance fire risks while fires also have climatic implications (Merzdorf, 2019). The goal of this project is to predict the likelihood of a fire to grow big in California. More specifically, we aim to develop machine learning models to predict fire size using weather data and data on fire characteristics. Such models could be used by fire managers to support decision making on use of limited resources and to facilitate timely evacuation. They may also be relevant to stakeholders impacted by fires such as residents, insurances, and non-government organizations focused on protection of ecosystems.

2. Data Acquisition and Cleaning

We used two data sources for this project. The *1.88 Million US Wildfires database* is an SQL database containing information on wild fires that occurred in the United States from 1992 and 2015. The wildfire information was initially derived from reporting systems of federal, state, and local fire organizations (Short, 2017). The dataset was compiled by Short (2017) and is publically available on Kaggle (https://www.kaggle.com/rtatman/188-million-us-wildfires).

The *NOAA GSOD database* provides global surface summaries from over 9000 weather stations. This database is provided by NOAA and is publically available via Google's BigQuery API (see https://www.kaggle.com/noaa/gsod?select=gsod2019).

Data cleaning and preparation involved the following steps:

1) ***Data cleaning***: Some variables needed cleaning, particularly recoding missing in the weather data. Cleaning was done initially before the weather data was used. An additional cleaning step was taken later after data was compiled and new variables created. This is mainly because creation of new variables revealed miscoding (e.g., wrong dates) that were not previously apparent.
2) ***Merging***: The biggest step consisted in merging the fire and the weather data. Due to computational limitations, weather data was extracted and additional weather variables

computed for a sample of all California fires only. The sample was constructed by including all 'big' fires (fire classes D to G) and an equal amount of non-big fires (fire classes A to C). The definition of "big" fire was established empirically and is based on fire frequency. Fires of classes A and B are most common. Fires of classes D to G are rather rare (< 3% of all fires). For every fire in the sample, we identified the closest weather station and extracted weather data for the day of fire and computed additional variables for the time prior to the fire based on data from the respective station.

3) ***Derivative dataset***: We created several derivative dataset from the initial fire and weather data that are used for different purposes during exploration and machine learning. The final datasets include the US fire data, California fire data, California weather data, and California fire and weather data sample. The datasets are described in detail below.
4) ***Creating new variables***: We created new fire and weather related variables. Weather related variables include, for example, precipitation and average temperature in the 30 day prior to the fire. Weather variables were created during data merging. Fire related variables include, for example, a binary variable indicating whether a fire is "big" (classes D to G) or not "big" (classes A to C) and a variable indicating number of days from discovery to containment. Fire related variables were created before or after data merging, depending on need. For example, the "big" fire variable was used for sample selection and thus needed to be created prior to sample selection. Other fire variables were only compiled for the fires in the sample and hence were created post sampling.

It is important to note that the process of data preparation and exploration is not strictly linear (i.e., first preparation and then exploration). For example, interesting insights during an initial exploratory analysis (e.g., discrepancies in number of fires across years) motivated the creation of new variables (e.g., number of simultaneous fires). Similarly, the differentiation of "big" vs. "non-big" fires, which was essential for sampling, was based on an initial exploration of fire frequency.

3. Data Exploration

3.1 Introduction to the cleaned data

The processes above yielded four derivative datasets. The *US fire data* contains information on all fires reported for the US. This dataset is used for comparison and to situate the case of California in the broader national context. This dataset contains 1,880,465 fires of which 54,093 are classified as "big" and 25 variables.

The *California fire data* is a subset of the *US fire data* and includes fires reported for California only. This is the main dataset used for exploratory analysis. This dataset allows us to gain an overview of major characteristics of fires in California. This dataset contains 189,550 fires of which 4,474 are classified as "big" and 25 variables.

The *California weather data* contains weather information from all stations located in California for the years 1992 to 2015. This dataset is used for an exploratory analysis of weather patterns. The dataset contains 1,644,717 observations and 19 variables.

Finally, the *California fire and weather data sample* includes fire and weather data for all "big" fires reported in California and an equal amount of "non-big" fires. This dataset is used for exploratory and inferential analysis on the relation between fire size and weather patterns. This is also the main input dataset for machine learning. This dataset contains 8948 fires of which 4,474 (50%) are classified as "big" and 51 variables. Note that this dataset contains some missing values due to characteristics of variables (e.g., number of fires in 3 years prior to a fire is not available for fires in the first three years). Cases with missing values were removed prior to using the data for machine learning. Machine learning draws from this data but does not include all variables.

In the following sections, we will first situate the case of California in the broader national context. Next, we provide an overview of key characteristics of fires and weather variables for California. Finally, we examine the relation between fire size and weather as well as fire size and other fire characteristics. For fire size, we will use the categorical fire size class variable (classes A to G) as well as the binary "big" fire variable with "big" referring to classes D to G. The analyses reported here are a selection of the exploratory analyses done and focus on the key findings.

3.2 California in the national context

Figure 1 provides an overview of fire frequency across the United States. California is one of several regions with relatively high numbers of fires. For the years 1992 to 2015, California has 189,550 recorded fires. This is the highest number nationally (see Figure 2). Other states with high fire numbers include Georgia, Texas, North Carolina, and Florida. Figure 3 shows the number of fires by fire class. Over 85% of the fires are less than 10 acres (i.e., classes A and B). Fires larger than 100 acres (i.e., classes D to G) are rather rare with less than 3% of all fires. In the following, will refer to fires over 100 acres as "big" fires. Figure 3 shows the 10 states with the highest number of "big" fires. California has 4,474 fires classified as "big". This is the second highest number nationally (after Texas; see Figure 4 for more details).

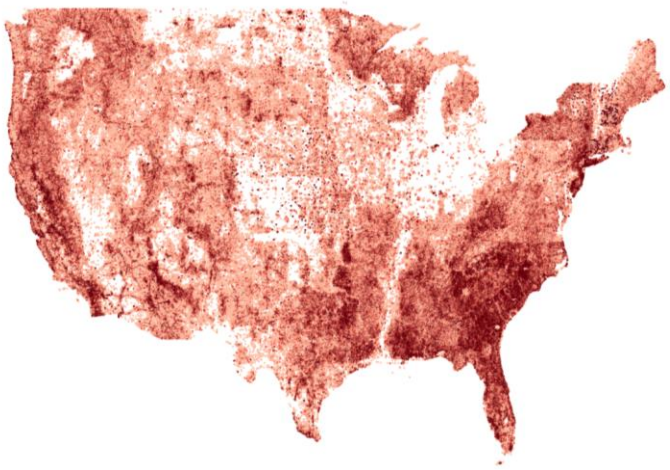


Figure 1. *Prevalence of wildfires in the US (1992-2015)*
Darker color indicate more fires

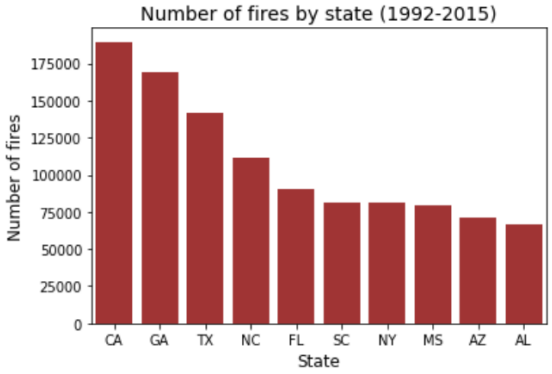


Figure 2. *Number of fires by state for the 10 states with the highest number of fires (1992-2015)*

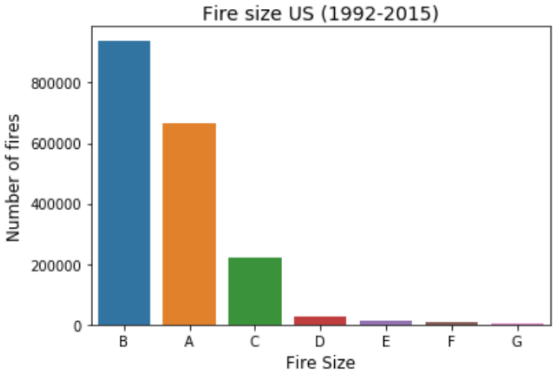


Figure 3. *Number of fires by fire class.*

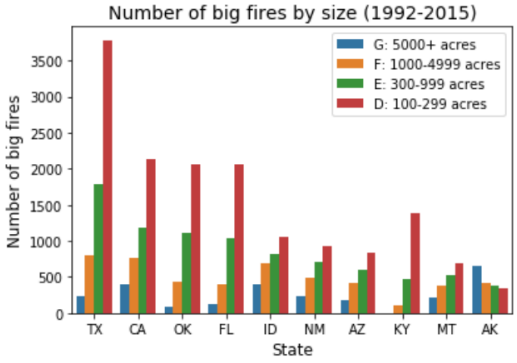


Figure 4. Number of fires by size for the 10 states with the highest number of "big" fires (1992-2015)

3.3 Fires in California

Figure 5 shows the spatial distribution of fires in California. Wild fires are most common in the mountainous areas of the Central Valley and the South Coast. Fires are least common in the deserts in the East. Most fires occur in the summer months (May to September; see Figure 6). Figure 7 shows number of fires and amount of acres burnt by year. The trend mimics the overall national trend. The number of fires decreases over time but the number of acres burnt increases. Interesting is the considerable variation in number of fires and acres burnt by year. Some years observe more fires and higher numbers of acres burnt compared to other years. Potential reasons for this lie in weather patterns (e.g., dry vs. wet year) as well as in time needed for vegetation to regrow after a fire.

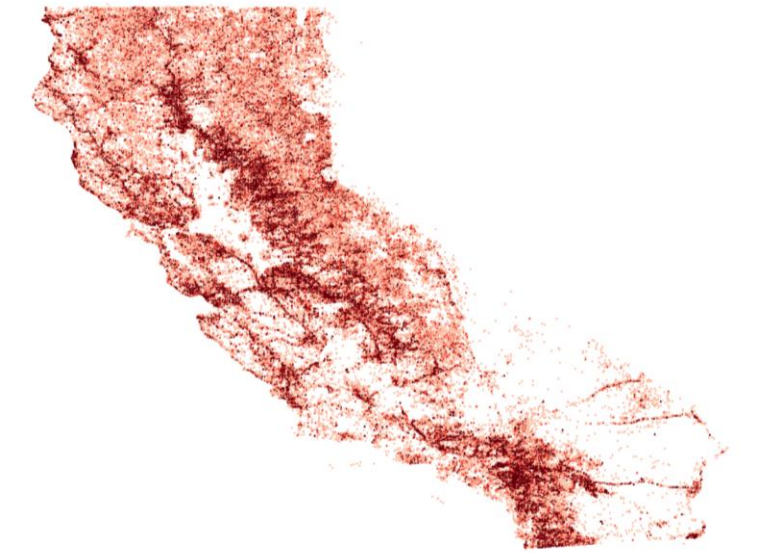


Figure 5. *Prevalence of wildfires in the US (1992-2015)*
Darker color indicate more fires

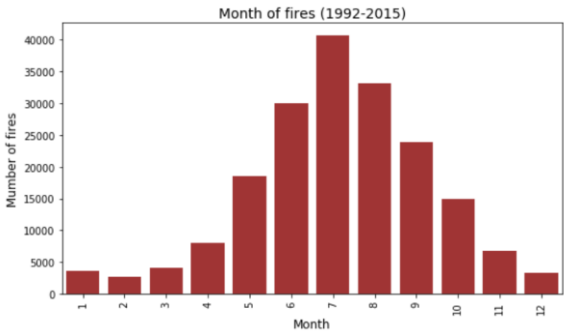


Figure 6. *Fire frequency by month (1992-2015)*

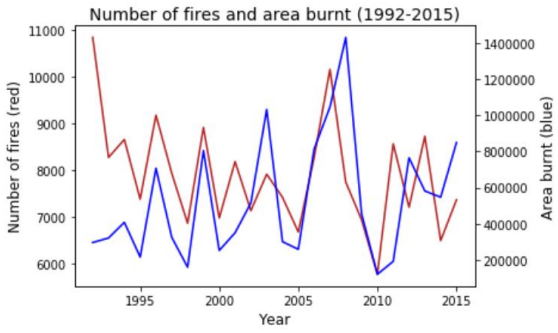


Figure 7. *Number of fires and acres burnt by year (1992-2015)*

Information on fire cause is incomplete with many fires coded as miscellaneous or undefined. The available data indicates that most common fire causes include lightning, equipment use, and arson. Figure 8 shows the spatial disparity of these fire causes. Fires occurring in the eastern and northern part of the country tend to be caused by lightning. Fires occurring in the western part are more likely caused by equipment use.

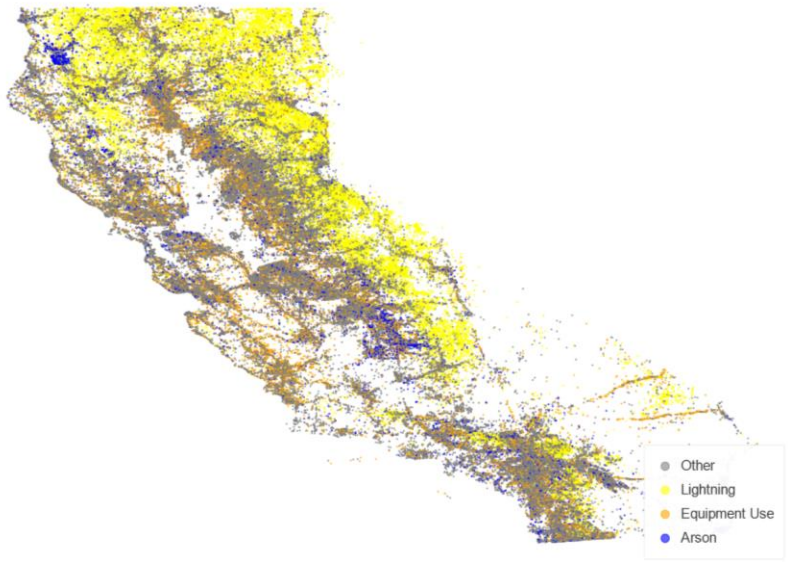


Figure 8. *Most common specified fire causes in California (1992 -2015)*

3.4 Weather in California

To gain a better impression of weather patterns in California, we discuss average temperature, rain, and wind speed for the example year of 2010.[1] Figures 9 to 11 show the average temperature, amount of rain, and wind speed for the year 2010 by month. The summer months (May to September) tend to observe the highest temperatures. Most rain falls in the winter months. This corresponds with the observation of fires with most fires occurring in the summer months. There is no distinct yearly pattern in terms of wind speed.

[1] We chose an example year rather than data aggregation across all years because the number of observations and stations varies across the years. For example, we have 52,336 observations for the year 1997 and 90,553 observations for the year 2012. This may yield misrepresentation.

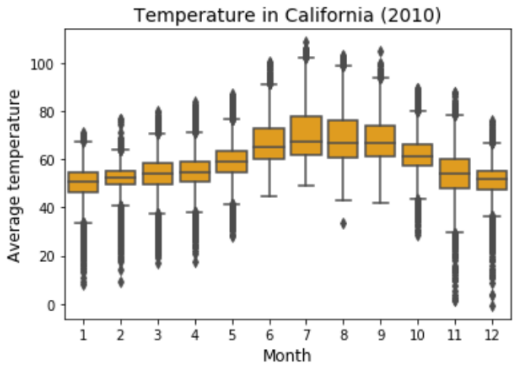


Figure 9. *Average temperature in California by month for the year 2010*

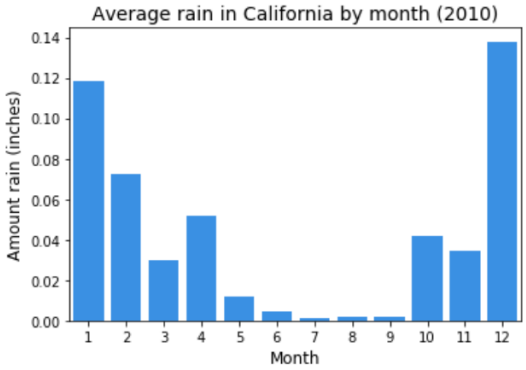


Figure 10. *Amount of rain in California by month for the year 2010*

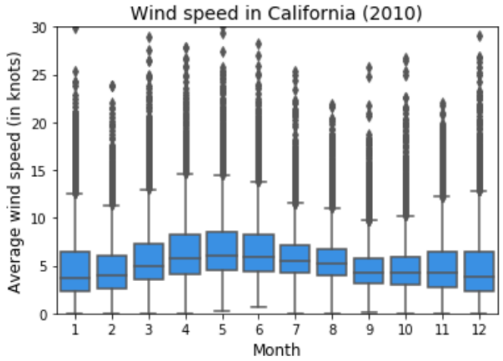


Figure 11. *Average wind speed in California by month for the year 2010.*

3.5 Relation between fire size and weather in California

To examine the relation between fire size and weather variables, we will use two fire size variables. The first variable is binary and distinguishes between “big” and “non-big” fires. The second variable is ordinal and consists in the fire size classes A to G. Inferential statistics are based on the ordinal variable and Spearman’s rank correlation.

We would expect fire size to be positively associated with temperature because higher temperatures lead to a hotter and dryer environment. Figure 12 shows the relationship between fire size class (A to G) and temperature on the day of fire. The average, max, and min temperature tend to be slightly higher on days when large fires are discovered. The relations between all temperature measures and fires size are positive and statistically significant ($p < .01$). The strength of the relationship varies. It is strongest for the average temperature (Spearman’s rho = .20).

We also tested the relationship between fire size and temperature 30 days prior to the fire. The relationship is positive and significant ($p < .01$) but tends to be weaker compared to the effect of temperature on the day of fire (Spearman’s rho ≤ .12).

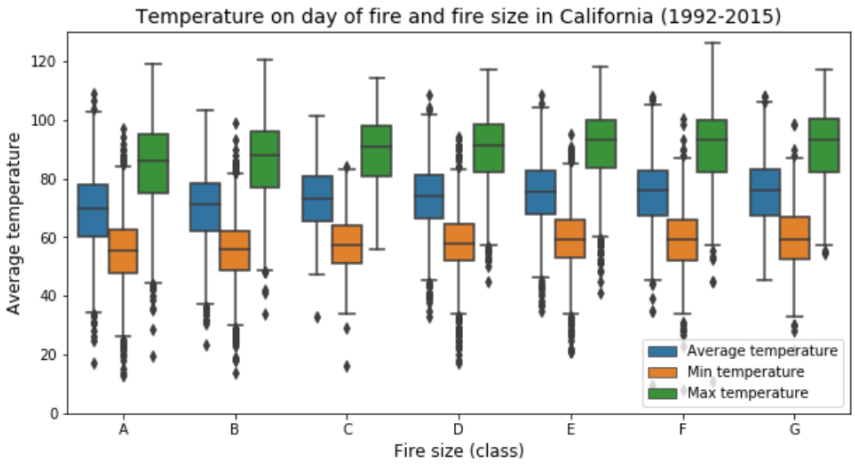


Figure 12. *Relation between temperature on day of fire and fire size.*

In contrast to temperature, we expect precipitation to be negatively associated with fire as it creates a wet environment. We find that the amount of precipitation in the 30 days prior to the fire is lower for "big" fires (mean = 0.21 inches, sd = 0.6 inches) compared to "non-big" fires (mean = 0.40 inches, sd = 1.1 inches). This relationship between amount of rain (30-days) and fires size is negative and statistically significant ($p < .01$). However, the strength of the relationship is rather weak (Spearman's rho = .-.08).

Finally, we examined the relationship between fire size and wind. Higher wind speed can be conducive to fire spread and hence we expect a positive relationship between wind and fire size. Figure 13 shows the relationship between fire size class and average wind speed. Average wind speed on the day of fire is higher for "big" fires (mean = 6.3 knots, sd = 3.2 knots) compared to "non-big" fires (mean = 5.6 knots, sd = 2.8 knots). Similarly, the maximum wind speed on the day of fire is higher for "big" fires (mean = 13.3 knots, sd = 5.2 knots) compared to "non-big" fires (mean = 12.3 knots, sd = 4.5 knots). The relationships are positive and statistically significant ($p < .01$). The strength of the relationship is again rather weak (Spearman's rho = . 09 for both).

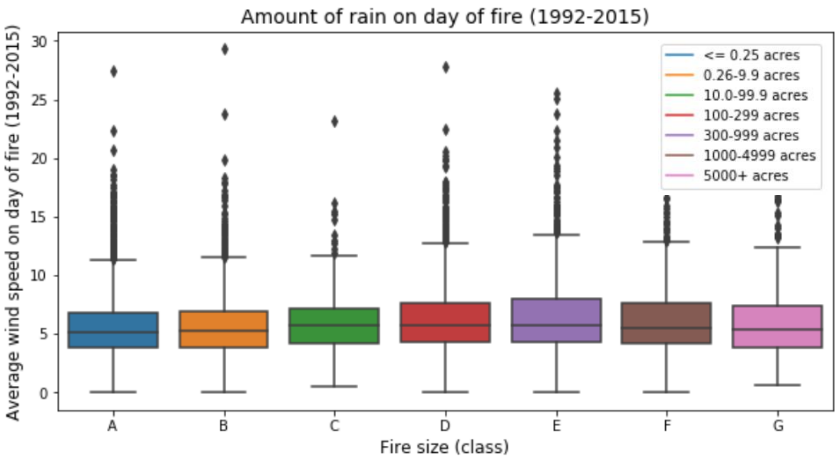


Figure 13. *Fire size and average wind speed on day of fire.*

3.6 Relation between fire size and other fire characteristics

In addition to weather events, we explore the impact of two fire related characteristics on fire size, namely simultaneousness of fires and previous fires. Given the number of fires taking place in California in a year (i.e., ca. 2821; Cal Fire, 2020), there are several days where we observe two or more fires taking place simultaneously. This can influence availability of firefighting resources for any given fire and thus fire size. To capture this influence, we use two variables namely the number of fires that were discovered +/- 1 day after a fire and the number of fires discovered +/- 3 days after a fire. Figure 11 shows the relation between fire size and number of fires discovered within the +/- 1 day range. The graph indicates a positive relation between fire size and number of simultaneously burning fires. “Big” fires are more likely to occur during times when many fires are burning simultaneously. The relations between number of simultaneously burning fires (+/- 1 day, +/- 3 days) and fires size are positive and statistically significant ($p < .01$). The relationships are strong with Spearman’s rho = .24 for +/- 1 day and .23 for +/- 3 days.

The size of a fire may also be influenced by the number of fires that previously took place in the area and the area burnt. As fire feed on vegetation, previous fires may limit the amount of material that a fire can feed on. To measure this effect, we counted the number of fires that took place within $5km^2$ of a fire in the previous 3 years and the amount of acres burnt by these fires. We find that the number of fires is slightly higher for “non-big” fires (mean = 1, sd = 1.8) compared to “big” fires (mean = 0.85, sd = 1.3). However, the area burnt is larger for “big” fires (mean = 954 acres, sd = 5489 acres) compared to “non-big” fires (mean = 455 acres, sd = 4806 acres). The relations between number of close fires in the past 3 years and fire size is not significant. The relation between area burnt by these fires and fire size is significant ($p < .01$) and positive (Spearman’s rho = .7). If many acres were burnt previously, a fire is more likely to grow big. Accordingly, the direction of the relation is counter to what we expected. This may be a result of remoteness of fires (e.g., fires in mountainous areas are more likely to grow big).

3.7 Summary

Our exploratory and inferential analyses indicate association between fire size and weather variables as well as between fire size and other fire characteristics. Fire size is positively associated with temperature and wind and negatively associated with amount of rain. The relations are statistically significant but the effect sizes tend to be rather small. The strongest relationship identified relates to amount of simultaneous fires. Large fires are more likely to occur when there are multiple fires discovered around the same time. This hints at the importance of firefighting capacity and management for fire size.

4. Modeling

The goal of our modelling is to use machine learning algorithms to predict the size of a fire. We use a binary fire size variable (i.e., "big" vs. "non-big") as our target variable. Given this setup, we use supervised, binary classification algorithms. The model is trained on 60% (3,787 cases) of the data and tested on the remaining 40% (3,041 cases) of the data. When generating the California fire and weather sample data we constructed a balanced sample, i.e., an equal amount of "big" and "non-big" fires. While some cases were excluded due to missing values at a later stage, our data remains well-balanced.

4.1 Data Pre-processing

Before applying machine learning techniques to the dataset, we needed to pre-process the data. The following pre-processing steps were taken:

1. **Target variable**: The target variable for our classification consists in a binary variable indicating whether a fire is "big" (1) or not (0). This variable was constructed earlier.
2. **Feature variables/dummy variables**: Feature variables consist in weather variables, fire characteristics, and location. Most of these variables were created in earlier steps. The exception to this are the variables on fire cause and fire month. The fire cause variables are three dummy-variables indicating if a fire is associated with one of the three most common causes (i.e., lightning, equipment use, and arson). The month variables consist of 12 dummy variables that capture in which month the fire was discovered.
3. **Missing**: Some cases contained missing values and were removed. This mainly applies to cases of the first three years 1992 to 1994. Variables on fire activity in the 3 years prior to the fire are missing for these cases because the database starts in 1992.
4. **Train-test split**: The initial dataset was split in a training and test dataset. We reserved 60% of the data for training and 40% for testing.
5. **Normalization**: Some of the variables were measured on ratio scale. Since one of the machine learning algorithms used here, namely K Nearest Neighbors (KNN), is sensitive to scaling, the ratio scaled variables were rescaled before used with KNN. We rescaled the ratio variables to have a mean of 0 and a standard deviation of 1. The respective variables were also rescaled in the test data set. For this, the mean and standard deviation of the training dataset were used. This is to account for the fact that the model aims to predict new sets of cases whose mean and standard deviation are not necessarily known or, depending on the set, might be distorted.

4.2 Feature selection

An initial set of features was selected based on the exploratory analyses and background knowledge on fire. This selection was refined during the modeling process by evaluating model performance relative to excluding and including variables. Table 1 provides an overview of the features included in the final models and the justification for their inclusion.

Table 1. Feature overview.

Feature (29 variables total)	Motivation for inclusion
Weather • Temperature on day of fire (min, max, average) (3 var.) • Temperature 30 days prior to fire (min, max, average) (3 var.) • Rain on day of fire • Rain 30 days previous to fire • Wind speed on day of fire (average, max) (2 var.)	• *Weather characteristics found to be relevant during exploratory analysis* • *Weather characteristics reasonably related to fire*
System characteristics • Simultaneous fires (+/- 1 day) • Area burnt by close fires in 3 years previous to fire • Month in which fire happens (12 var.) • Fire cause (lightning, arson, equipment use) (3 var.)	• *Fire characteristics that can reasonably be known or guessed when a fire is discovered* • *Fire characteristics found to be associated with fire size during exploratory analysis* • *Characteristics found to have considerable variation and that may reasonably affect fire size (e.g., month)*
Location • Latitude • Longitude	• *Spatial disparities (exploratory analysis)* • *Account for general geographic, terrain, and resource characteristics not included*

4.3 Modeling and Evaluation Metric

Four machine learning models were trained on the data, namely two random forest models (GradientBoost & XGBoost), a logistic regression, and a K nearest neighbors (KNN) model. We will discuss the random forest and logistic regression models here. KNN was excluded due to poorest performance. The model is discussed in the appendix. The models were hyper-parametrized using 5-fold cross validation and scikit learn grid search. The area under the ROC curve was used as evaluation criterion for hyper-parametrization. The final models are evaluated on the test set using common performance measurers (i.e., accuracy, recall, precision, and f-1 score). We also compute ROC and Cumulative Gains Curves.

4.4 Random Forest

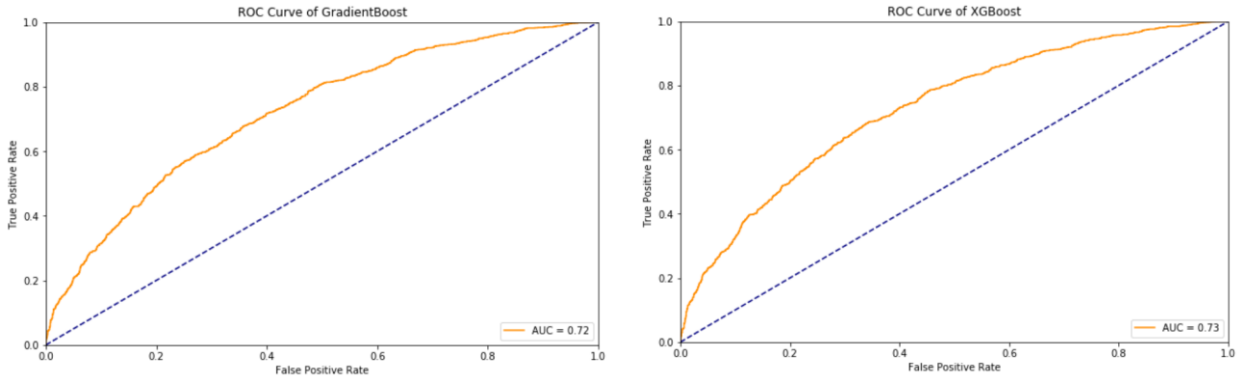
We trained two random forest models employing two different boosting algorithms, namely GradientBoost and XGBoost. For the GradientBoost the learning rate was determined using 5-fold cross validation and grid search. For the XGBoost, the learning rate, maximal tree depth, and number of

estimators were determined using 5-fold cross validation and grid search. The difference in hyper-parameterization is due to considerations of overfitting. We found the GradientBoost model to over-fit considerably when all hyper-parameters are determined using cross validation.

Table 2 shows the various metrics for the two models. The performance of the two models is comparable and overall mediocre. The GradientBoost model has an accuracy of 0.66, a precision for positive cases of 0.66, and a recall of positive cases of 0.64. The XGBoost model has an accuracy of 0.67, a precision for positive cases of 0.67, and a recall of positive cases of 0.66. Figure 14 shows the ROC curves of the two models.

Table 2. Random Forest model performance

Model	***Class***	***Values***			***Accuracy***	***ROC AUC***
		Precision	*Recall*	*F1-score*		
Random Forest (GradientBoost)	0	0.65	0.67	0.66	0.65	0.72
	1	0.66	0.64	0.65		
Random Forest (XGBoost)	0	0.67	0.68	0.67	0.67	0.73
	1	0.67	0.66	0.66		



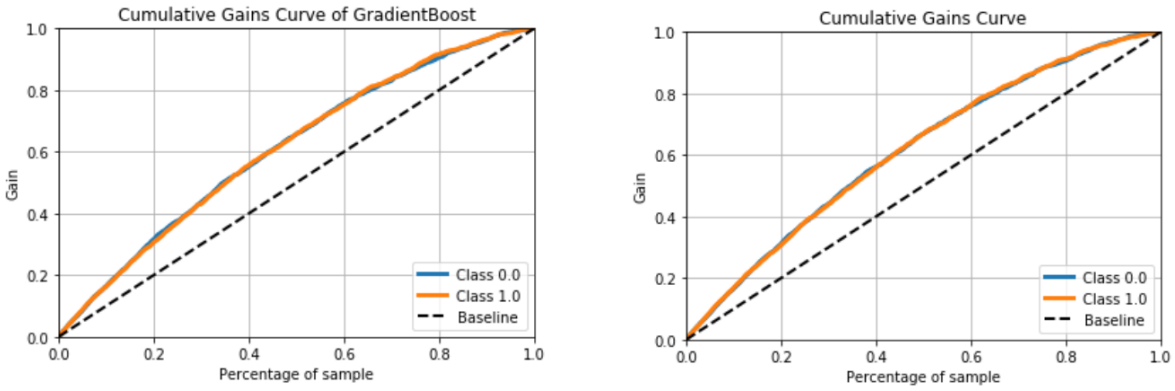
GradientBoost *XGBoost*

Figure 14. Random Forest ROC curve

The two models vary in terms of the most important features. The most important features in the GradientBoost model are number of simultaneous fires (+/- 1 day), latitude, area burnt by close fires in the 3 years prior to the fire, longitude, and average temperature on day of fire. The most important features in the XGBoost model are simultaneous fires (+/- 1 day), area burnt by close fires in the 3 years prior to the fire, fire caused by lightning, and the months of June and July. Both models share simultaneous fires (+/- 1 day) and area burnt by close fires in the 3 years prior to the fire. The former

was found to be strongly related to fire size in our exploratory and inferential analyses thus this is in line with our previous findings.

Machine learning algorithms commonly apply a cut-off value of 0.5 percent to make predictions. Fires with a likelihood of 0.5 or higher to grow "big" are classified as positive cases and systems with a likelihood of less than 0.5 are coded as negative cases. Performance statistics based on these cut-off predictions may obscure nuances in model performance. For example, predictions may be highly accurate for fires with a probability above 0.9 but poor for fires with a probability between 0.5 and 0.6. To gain a better grasp of nuances in model performance, we can look at the Cumulative Gains Curve. Cumulative Gains Curves are especially popular in marketing where they are used to evaluate model-based customer targeting relative to random customer selection. The application here is adopted. Figure 15 shows the curves for the random forest models. The Cumulative Gains Curve is based on the predicted probability and shows the percentage of true positive cases relative to the percentage of sample explored. Put differently, if we were to look at, for example, the 20% of fires with the highest predicted probability to grow "big" according to our model, how many would actually grow "big"? We can see that for 20% it is roughly 30% for both the GradientBoost and the XGBoost models. This is better than random where we would expect to capture 20% of *all big fires* when looking at 20% of the sample. This indicates that there is some benefit to our model in identifying big fires. The benefit is, however, rather small.



GradientBoost *XGBoost*

Figure 15. *Random forest Cumulative Gains Curve*

4.5 Logistic regression

Our logistic regression model was hyper-parametrized (parameter C) using 5-fold cross validation and grid search. Table 3 shows various metrics for the logistic regression model. The performance is slightly worse than the performance of the random forests. Model accuracy is 0.63. The recall and precision for positive cases are 0.64 and 0.61 respectively. Figure 16 shows the ROC curves of the model.

Table 3. Logistic regression model performance

Model	***Class***	***Values***			***Accuracy***	***ROC AUC***
		Precision	*Recall*	*F1-score*		
Logistic Regression (up-sampled training data)	0	0.63	0.65	0.64	0.63	0.68
	1	0.64	0.61	0.63		

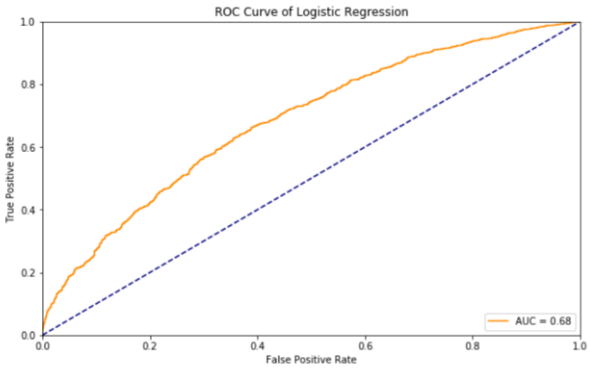


Figure 16. Logistic regression ROC curve

The most important features are the months of January, August, December, and September, and the fire cause lightning. These important features are considerably different from those observed in the random forest models. In the random forest models, fire characteristics and tend to be more prominent. Interestingly, weather characteristics tend to be underrepresented across all models.

Figure 17 shows the Cumulative Gains Curve for the logistic regression model. The curve is slightly flatter compared to the curves of the random forests. This indicates lower model performance.

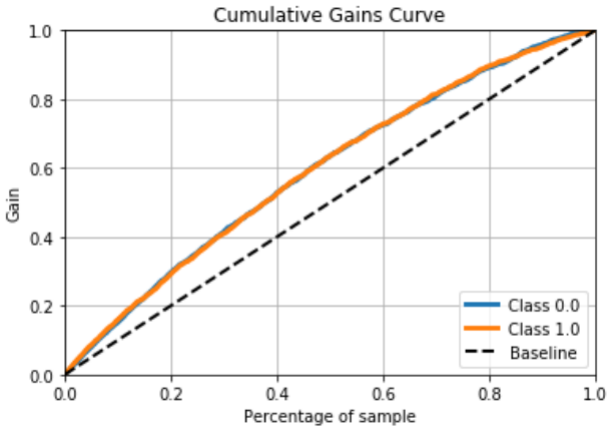


Figure 17. Cumulative Gains Curve for logistic regression model

5. Model generalizability: Predicting "big" fires in Arizona

To further evaluate the California model and explore its generalizability, we employed the different models to predict wild fires in the neighboring state Arizona. The Arizona wild fires represent a sample of 3,222 fires of which 1,611 (50%) are categorized as "big". Table 4 shows performance metrics for the three models. Overall performance of the models is comparable with their performance on the California data. The GradientBoost model performs best. Interesting is the relatively lower recall of negative cases of the random forest models (i.e., 0.34 for GradientBoost and 0.28 for XGBoost). These measures generally indicate potential to apply the model trained on the California data to other contexts. However, Arizona is a neighboring state with comparable climate. The models might not perform well on fires from states that are located further apart or different in their climate.

Table 4. *Model performance to predict "big" fires in Arizona.*

Model	***Class***	***Values***			***Accuracy***	***ROC AUC***
		Precision	*Recall*	*F1-score*		
Random Forest (GradientBoost)	0	0.70	0.34	0.46	0.58	0.67
	1	0.55	0.85	0.66		
Random Forest (XGBoost)	0	0.66	0.28	0.40	0.55	0.63
	1	0.53	0.84	0.65		
Logistic regression	0	0.63	0.49	0.55	0.59	0.64
	1	0.56	0.69	0.62		

6. Using model and recommendation

Before we discuss model use and recommendation, it is important to remind us of two characteristics. First, the simplicity of the model. The model predicts whether or not a fire grows "big" based on

weather data from the data of fire and simple fire characteristics. It includes 29 feature variables which is a relatively low number. Second, we use a 50% cut-off value meaning that fires with a likelihood of 50% or higher to grow "big" are classified as big and fires with a likelihood lower than 50% are classified as non-big. There is some room for adjusting this cut-off value to further tailor the model to more specific needs. For example, using a lower cut-off value would allow us to get a better recall for "big" fires but our predictions would be less precise. This scenario would be helpful when using the model for asking citizens to prepare or evacuation—i.e., the citizens would at times prepare for fires that do not grow big but would be ready for most fires that do grow big. We may also use a higher cut-off value. This would allow us to get a better precision but lower recall for "big" fires. Such a scenario is desirable when deciding about the use of limited resources—i.e., fire fighters would spend their resources on fires that are very likely to grow big. This being said, there are limitations to the optimization of precision and recall of big fires as indicated by the Cumulative Gains Curve indicate.

Given the simplicity and limitation of the model (see also below) it does a fair job at predicting whether or not a fire is likely to grow "big". The precision and recall are similar for positive ("big" fire) and negative ("non-big" fires) cases. The models might be helpful as support tool in making decisions on the use of scarce resources and information of citizens. As discussed, the models can support decisions on the use of limited resources or to enhance evacuation preparation. Further tailoring of the cut-off value to serve these needs is advisable.

However, given the mediocre performance, the current models should not be used as sole decision making tools. The models may be combined with other techniques. For example, they may provide a short list of fires that demand attention that is then adjusted based on experience and other fire management criteria. Furthermore, these relatively simple models may be used as a starting point to elaborate more complex prediction models such as models that predict direction of spread of the fire or models that update their prediction based on weather characteristics throughout the fire management process.

7. Limitations and assumptions

The following are key limitations and assumptions of our approach:

- ***Model simplicity***: As mentioned above, our models are relatively simple in that they predict fire size based on fire characteristics and weather events on the day of fire and 30 days prior to the fire. Given the relatively few feature variables (i.e., 29) the model does a fair job. However, performance could be improved by further elaborating the model. Most notably, we could include more nuanced weather variables (e.g., hourly weather data) as well as elaborate the model to update prediction based on weather events taking place during fire management.
- ***Data accuracy***: The weather data is retrieved by identifying the nearest weather station and then extracting relevant weather data from that station. This approach is limited in two main ways. First, it assumes that the coordinates of fires are accurate. Exploration of some coordinates show that they tend to be close to roads. While there are reasons for fires to occur close to roads, this may also indicate some inaccuracy in the coordinates. Second and more importantly, it assumes that the weather data of the closest station applies to the location of

fire. Some weather events (e.g., rain) can be fairly local and we can reasonably assume that there is some error in the weather data. Apart from the weather data, we found several incidents of wrong information in the data (e.g., containment date prior to discovery date). These discoveries urge to be mindful about data accuracy.

- ***Binary target variable***: We used a binary target variable (i.e., "big" vs. "non-big" fire). There is, however, considerable variation in fire size within these categories and particularly in the "big" category. Our current models to not account for this. We decided for a binary approach because of constraints on the sample size. Due to computational limitations, we were only able to obtain weather data for a limited amount of fires. Using a finer categorization may cause problems during training as the samples for each category are rather small.
- ***Big vs. problem***: Fire size may not directly correspond to threat to human life, settlements, or natural habitat. Put differently, some "big" fires may pose more of a problem than others. For example, firefighter may be more interested in managing a fire with the potential to growing "big" when the fire is close to a city vs. when the fire occurs in a relatively remote area. Our current models do not account for these differences.
- ***Narrow feature variables:*** The model mainly focuses on weather variables and some additional fire characteristics. Many other variables that are likely to affect fire size such as preventive measures or vegetation are not included in the model.
- ***Exploratory analyses***: We conducted the exploratory analyses on the whole dataset and only split the data into training and test data for the modeling. This is strictly speaking not correct because the patterns in the test data are leaking through in the exploratory analyses. Ideally, we should have split the data at the very beginning and conduct the exploratory analysis on the training data only.

8. Future work

The following is a selection of steps that could be taken to improve the models:

- ***Additional data***: As stated above, there are a number of factors that are relevant for fires that we did not include in our models. Most notably, data on vegetation could enhance model performance as vegetation differs in their potential to "fuel" fires. This is well-known among firefighters who at times try to systematically move a fire away from certain vegetation. Another important set of variables relates to fire management strategies. There are different strategies that fire managers can pursue to prevent fires in the first place such as forest thinning or removal of certain species. Data on employment of such strategies could further enhance model performance and also provide insights into the effectiveness of these strategies.
- ***More complex model***: The current models are relatively simple in that they use a limited number of feature variables and do not consider events during fire management. Weather events during management process tend to be a challenge that make management difficult and influence containment success. A more elaborate model could include such considerations and provide updated fire size predictions during the course of a fire.

- ***Interaction effects***: The current models contain variables capturing separate weather effects. It is possible that the effects of weather on fire size are driven by interacting events (e.g., dry wind). Including interaction effects of weather variable is one avenue to further refine the existing model.
- ***Combining different models***: Rather than relying on just one machine learning algorithm, we could combine the results of different algorithms, weighted by their relative performance. This approach is often found to enhance prediction and frequently used in competitions and practice.
- ***More nuanced fire size variable***: The target variable in the current model is binary (i.e., "big" vs. "non-big"). This could be further extended such that a model predicts the fire size class (i.e., A to G) or even the absolute fire size.

9. Conclusions

This project employed machine learning algorithms to predict fire size (binary variable) in California based on weather data and fire characteristics. We compared the performance of three different machine learning models, namely two random forest models (GradientBoost & XGBOost) and one logistic regression model. The models are based on 29 feature variables and were hyper-parametrized using 5-fold cross-validation and grid search. The two random forest models are comparable in terms of performance. The random forest models outperform logistic regression when tested on the California test data. Employing the models to a sample of fires from the neighboring state Arizona indicates potential for application beyond California. Similar to the California test data, random forest models perform best on the Arizona data.

The applicability of the models is limited due to mediocre performance. They may be used to support decision making related to use of limited resources or to inform citizens prior to evacuation. The models are not suited as sole or major decision making tool or as evacuation recommendation tool.

Apart from the actual models, our analysis yielded important insights into the factors affecting fire size. The effect of weather events is less pronounced than expected. We find that simultaneousness of fires tends to be a more important predictor. This hints at the key role of fire management capacity and efficient use of limited resources. Further work on these models may include refinement of variables to account for management as well as vegetation effects. The models may also be adopted to account for management and weather events taking place over the course of a fire. These steps would likely enhance model performance and use.

References

Cal Fire. (2020). *Current Year Statistics*. Available at: https://www.fire.ca.gov/stats-events/

Hoover, K., & Hanson, L.A. (2019). *Wildfire Statistics. Congressional Research Service.* Available at https://crsreports.congress.gov/product/pdf/IF/IF10244.

Merzdorf, J. (2019). A drier Future Sets the Stage for More Wildfires. *NASA Global Climate Change.* Available at https://climate.nasa.gov/news/2891/a-drier-future-sets-the-stage-for-more-wildfires/.

Reinhardt, E. (2015). Fire Management in a Changing Climate. *Fire Management Today, 74*(3), 15-18.

Short, K. C. (2017). *Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508].* 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

Vose, J.M., Peterson, D.L., & Patel-Waynand, T. (2012). Effect of Climatic Variability and Change on Forest Ecosystems: A Comprehensive Science Synthesis for the U.S. Forest Sector. *United States Department of Agriculture, Forest Service, Pacific Northwest Research Station.* Available at https://www.fs.fed.us/pnw/pubs/pnw_gtr870/pnw_gtr870.pdf

Appendix

KNN

Table 5 shows the model performance of the KNN model. Figures 18 and 19 show the ROC curve and the Cumulative Gains Curve. The KNN model performed slightly worse than the logistic regression model.

Table 5. KNN model performance

Model	Class	Values			Accuracy	ROC AUC
		Precision	Recall	F1-score		
Random Forest (upsampled training data)	0	0.61	0.68	0.64	0.62	0.67
	1	0.64	0.56	0.60		

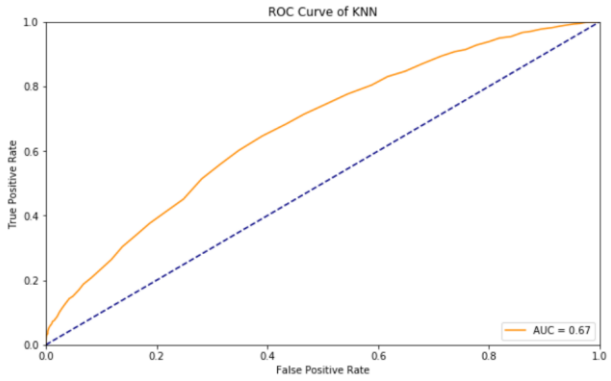


Figure 18. *ROC curve for KNN.*

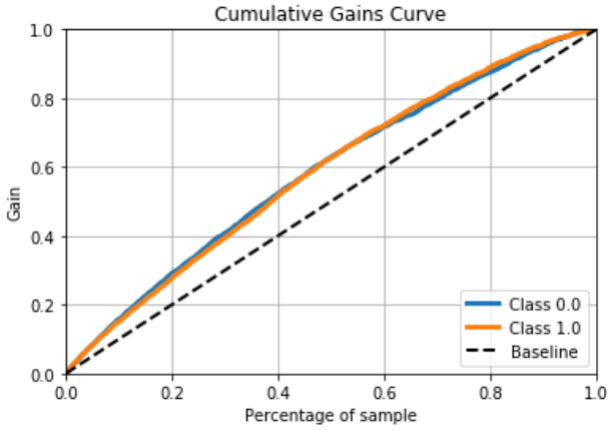


Figure 19. *Cumulative Gains Curve for KNN.*