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
Data Scientist at Samsung Electronics America

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.

The Problem

- Since 2000, every year “an average of 72,400 wildfires burned an average of 7.0 million acres” in the US (Hoover & Hanson, 2019, p.1)
- The number of fires is decreasing but the area burnt and hence fire size is increasing
- California is one of the hardest hit states in the US: 2.821 fires burning about 70,719 acres per year (5-year average; Cal Fire, 2020)
- Fire activity is related to weather patterns and climatic changes (Reinhardt, 2015; Merzdorf, 2019)

→ **Can we use machine learning and weather data to predict the likelihood of a fire to grow “big”?**

→ **Which machine learning algorithm performs best?**

Blue Cut Fire in California (2016)

For source of picture see last slide

Who might care?

Firefighter in the Santa Ynez Mountains, CA

Person watching a fire

Destroyed neighborhood in Paradise, CA

Fire fighters

- California Department of Forestry and Fire Protection (Cal Fire) is responsible for fire management in California
- Fire fighters working at the front line

People living in California

Limit loss of life and property
Timely evacuation

Insurers & NGOs

Loss or damage of property
Impact on people
Impact on wildlife

For sources of pictures see last slide

Data

Data source
1.88 Million US Wildfires database (SQL) obtained from Kaggle
https://www.kaggle.com/rtatman/188-million-us-wildfires

NOAA GSOD weather database (queried via Google's BigQuery API)
https://www.kaggle.com/noaa/gsod

Variables
Fire size (in acres and as classes A to G)
Information of fire (e.g., location, cause, discovery date)
Weather data (e.g., temperature, precipitation, wind)

Fire

Location

Date

Size

Cause

Weather station

Location

Temperature

Rain

Wind

For sources of pictures see last slide

Data preparation

- Obtain data from different sources
  → Extract relevant fire data from SQL database
  → Query NOAA GSOD database to obtain weather data (Google's BigQuery API)

- Clean weather data
  → Recode missing (e.g., 999.9 to NaN)

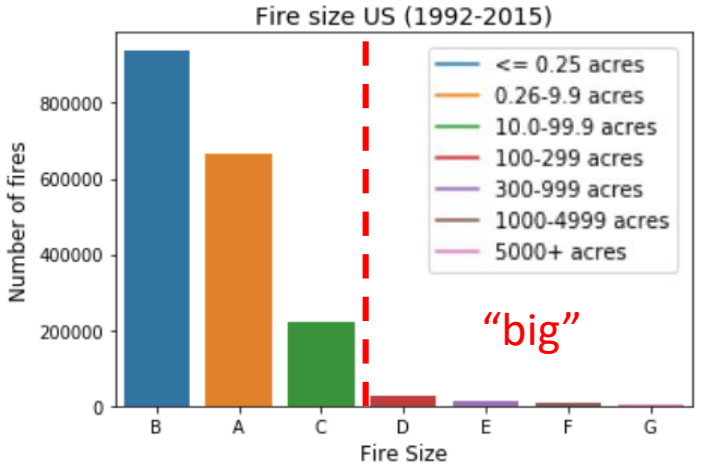
- Extract California data
  → Extract California weather data
  → Extract California fire data

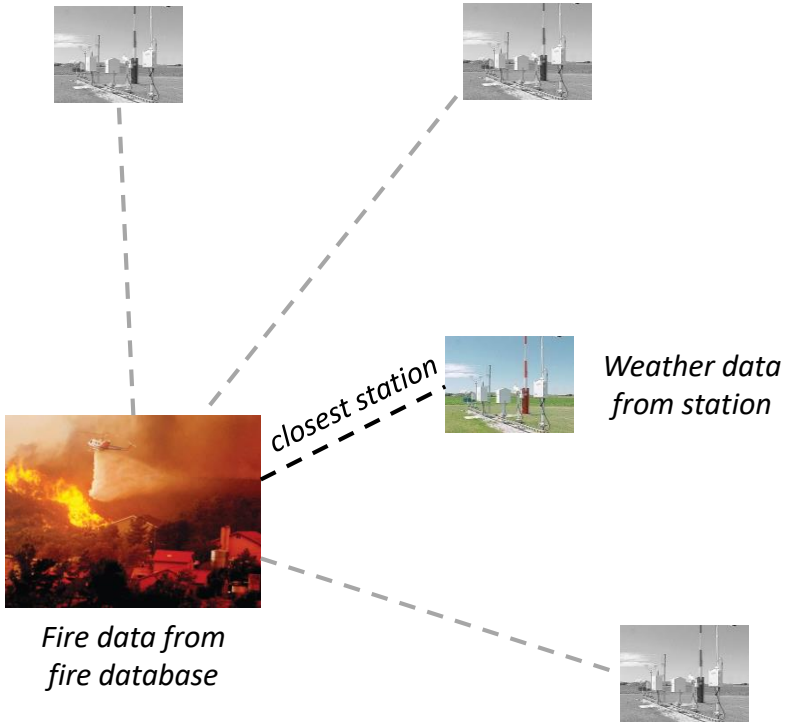
- Select sample of California fires
  → Include all big fires (classes D to G)
  → Include an equal amount of non-big fires at random (classes A to C)

- Find relevant weather data for each fire in sample
  → Identify weather station closest to fire based on longitude and latitude
  → Calculate and extract relevant weather variables

- Calculate additional fire variables for sample
  → How many fires burn simultaneously?
  → How many fires were in the same area in the previous years? How many acres did these burn?

- Clean data & exclusion of missing for sample
  → Some variables needed to be cleaned again
  → For the machine learning part, cases with missing values were excluded (e.g., fires in the first years that did not have data on fires in previous years)





Data overview

Datasets relevant for analysis

US fire data
Num. fires: 1,880,465
Num. "big" fires (classes D to G): 54,093
Num. variables: 25

- *Overview of all reported fires in the US*
- *Used for comparison*

California fire data
Num. fires: 189,550
Num. "big" fires (classes D to G): 4474
Num. variables: 25

- *All reported fires in California*
- *Subset of the US fire data*
- *Used for exploration*

California weather data
Num. observations: 1,644,717
Num. variables: 19

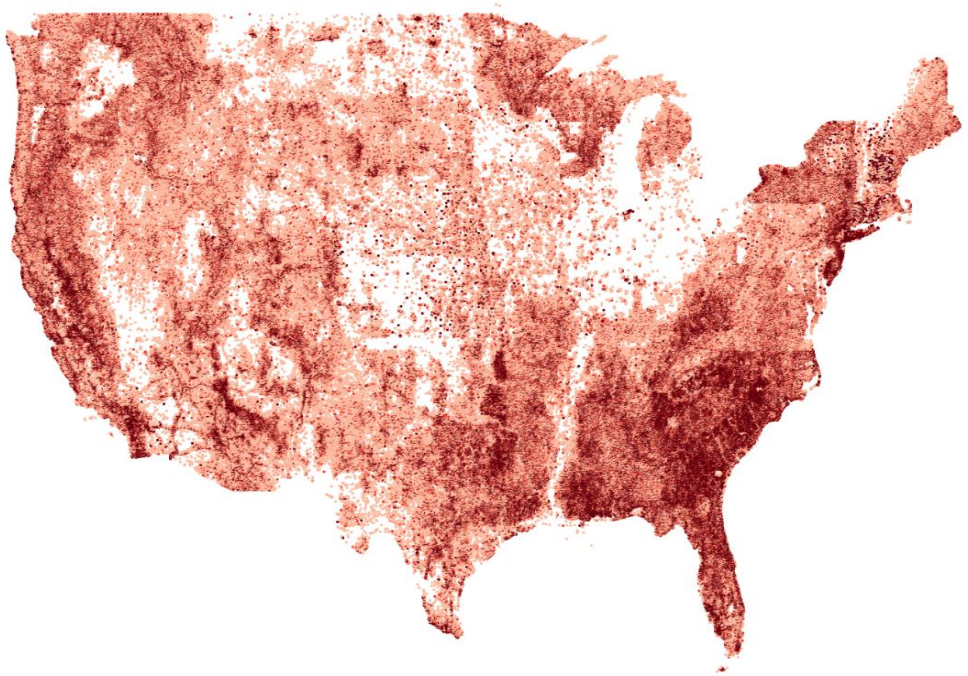
- *All weather reports for California (1992-2015)*
- *Used for exploration*

California fire & weather data sample
Num. fires: 8948
Num. "big" fires (classes D to G): 4,474 (50%)
Num. variables: 51

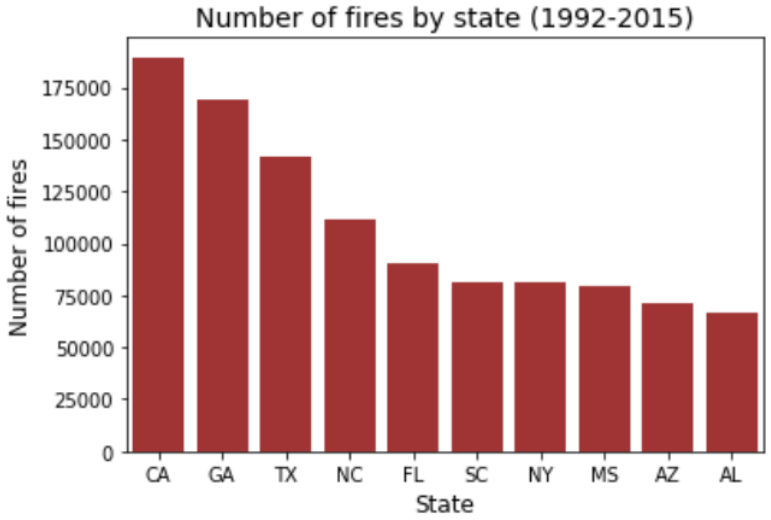
- *Fires in California and corresponding weather data*
- *Sample includes all "big" fires and an equal amount of non-big fires*
- *Used for exploration, inferential statistics, and machine learning*

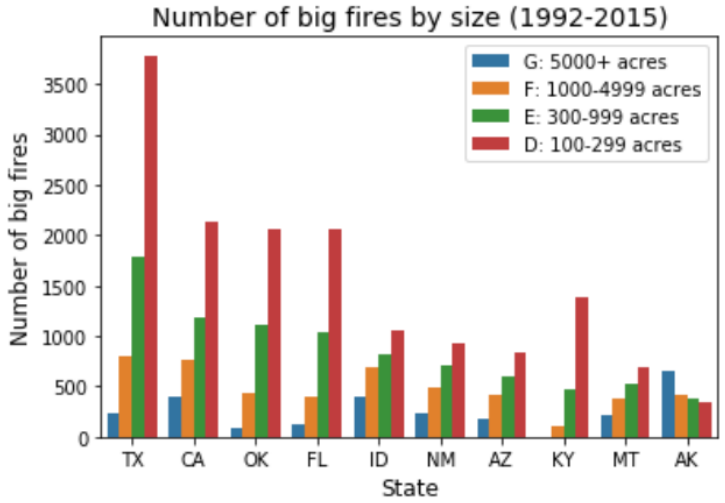
Data Exploration: California in the national comparison (1992-2015)

- California has 189,550 recorded fires. This is the highest number nationally.
- California has 4,474 “big” fires (classes D to G) this is the second highest number nationally (after Texas)



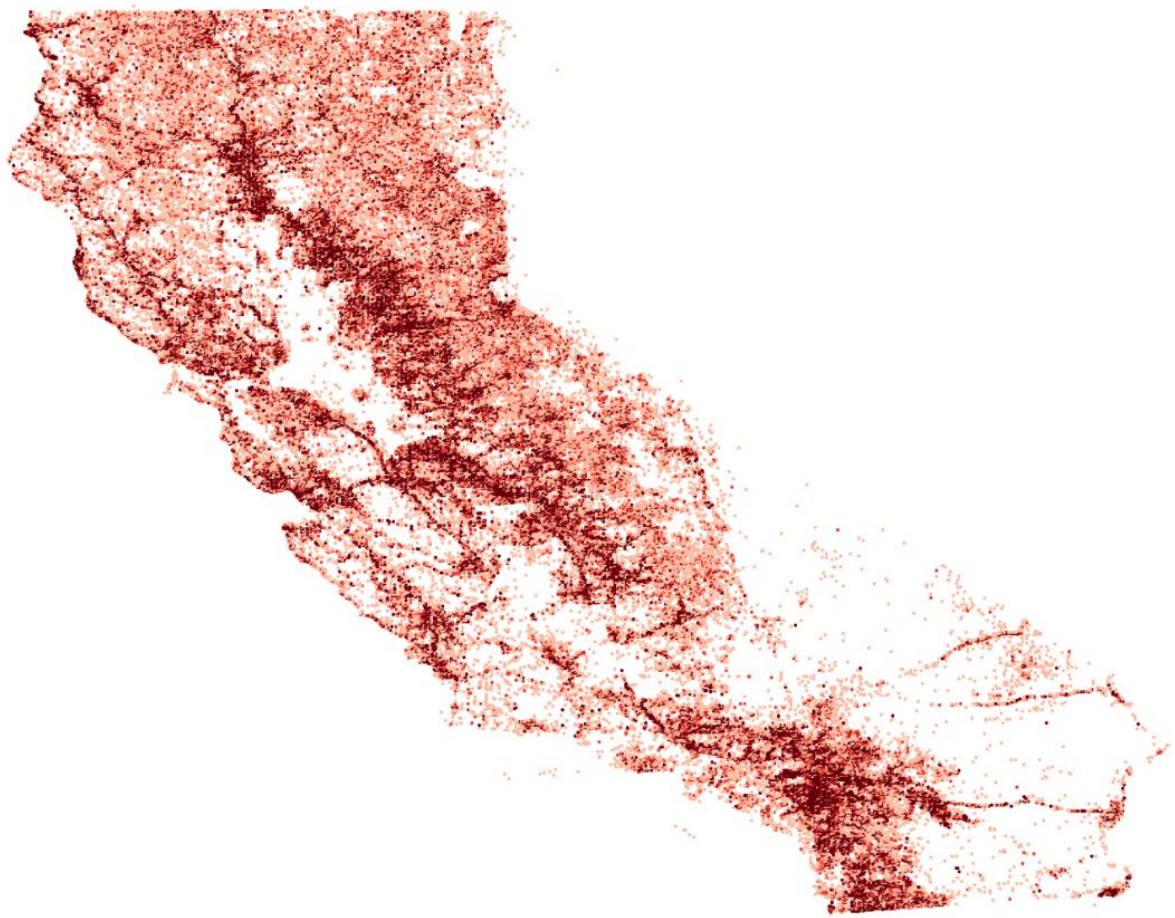
Prevalence of wildfires in the US (1992-2015)
Darker color indicate more fires





Data Exploration: Fires in California

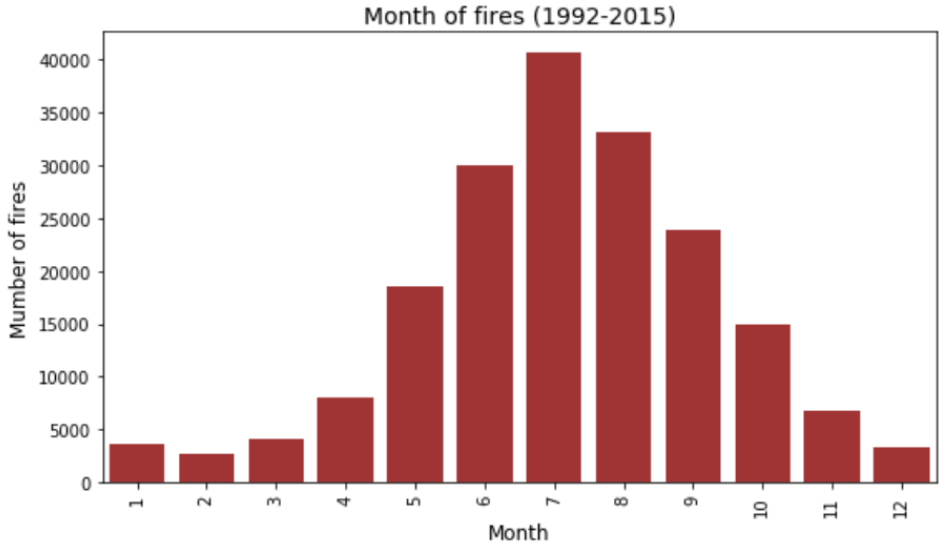
- Regional disparities: more fires in mountainous areas of the Central Valley and the South Coast

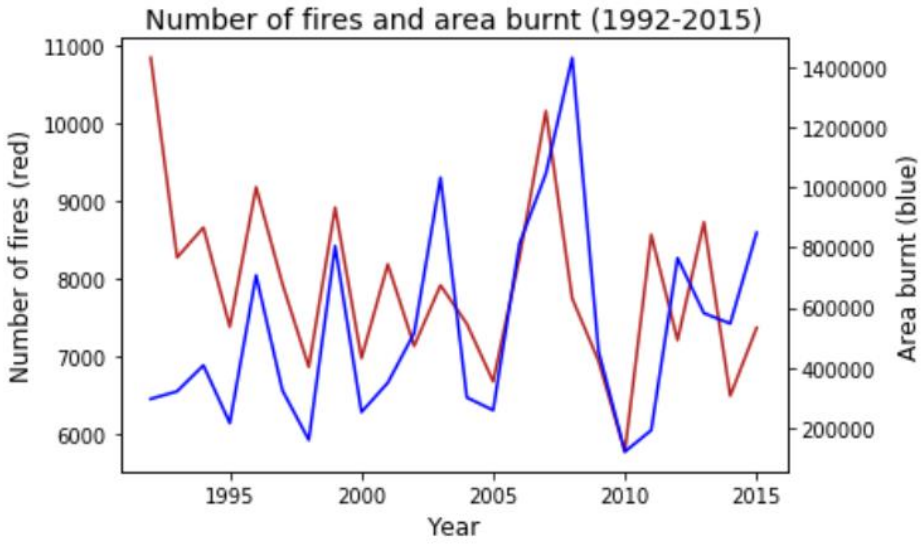


Prevalence of wildfires in California (1992-2015)
Darker color indicate more fires

Data Exploration: Fires in California

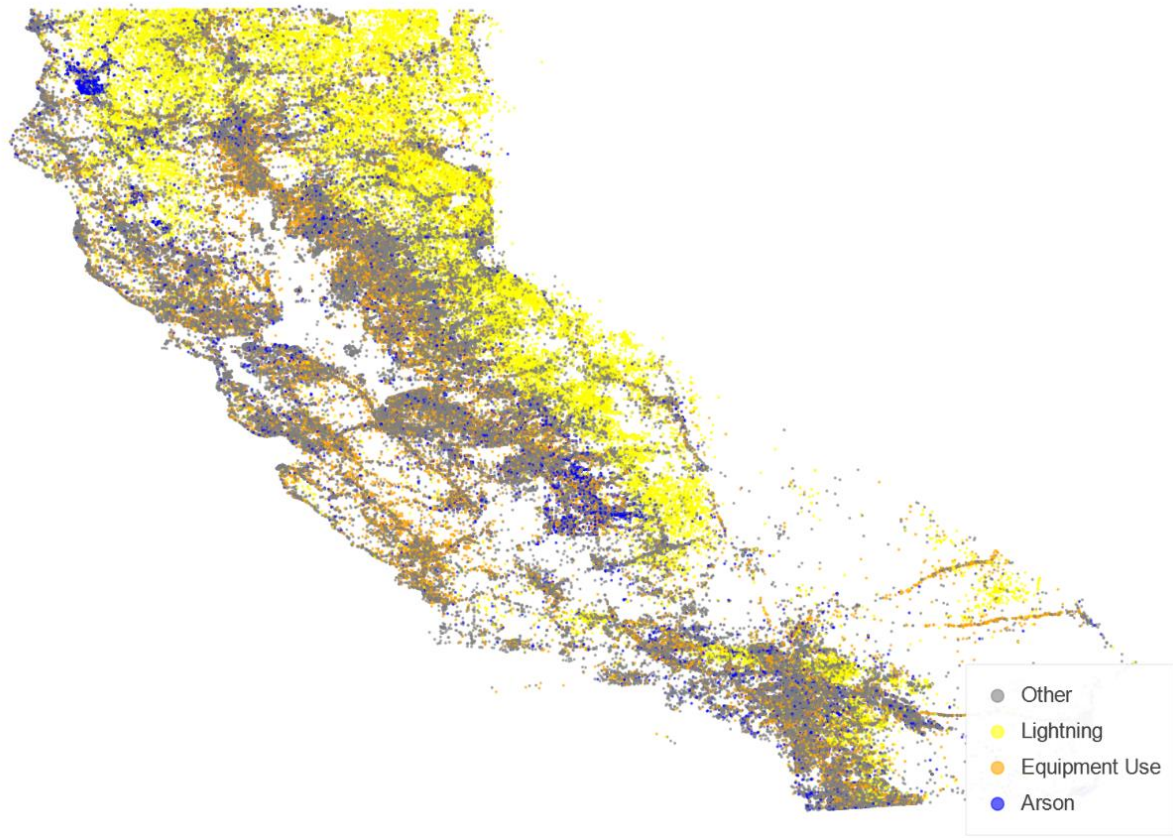
- Regional disparities: more fires in mountainous areas of the Central Valley and the South Coast
- Most fires occur in the summer months (May to September)
- Trend towards less fire but more acres burnt → consistent with national trends
- Considerable variation in number of fires and acres burnt between years. Potential reasons: (1) weather effects (e.g., dry year) and (2) vegetation (e.g., time to regrow)





Data Exploration: Fires in California

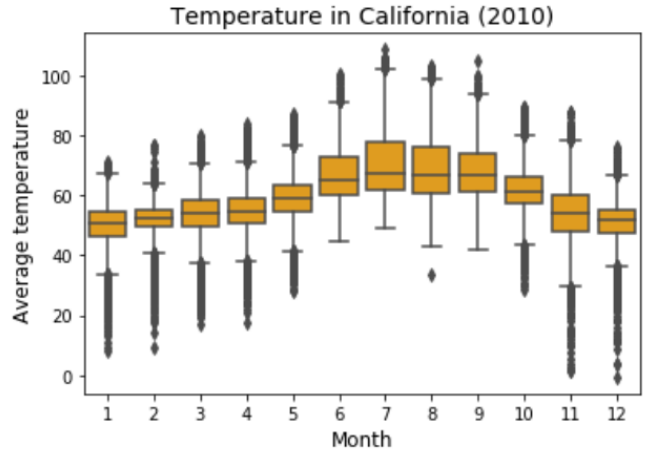
- Regional disparities: more fires in mountainous areas of the Central Valley and the South Coast
- Most fires occur in the summer months (May to September)
- Trend towards less fire but more acres burnt → consistent with national trends
- Considerable variation in number of fires and acres burnt between years. Potential reasons: (1) weather effects (e.g., dry year) and (2) vegetation (e.g., time to regrow)
- Most common known fire causes include lightning, equipment use, and arson. The causes of many fires are not specified (e.g., miscellaneous or undefined).

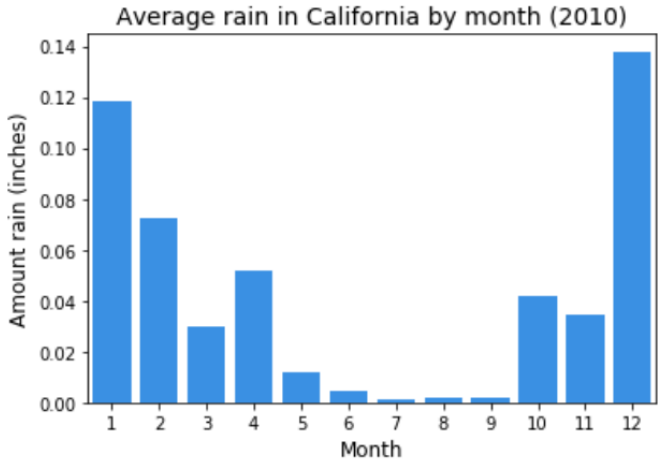


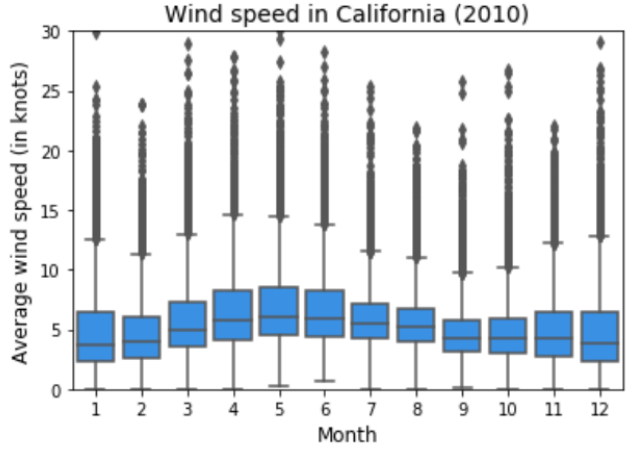
Most common specified fire causes in California (1992 -2015)

Data Exploration: Weather California

- Example year: 2010
- Temperature: Hot summer months and cooler winter months
  → Most “big” fire take place during the summer
- Rain: Most rain during the winter months, relatively little rain during the summer months
  → Few “big” fires during the winter
- Wind: No distinct wind pattern







Data Exploration: Fires in California and weather

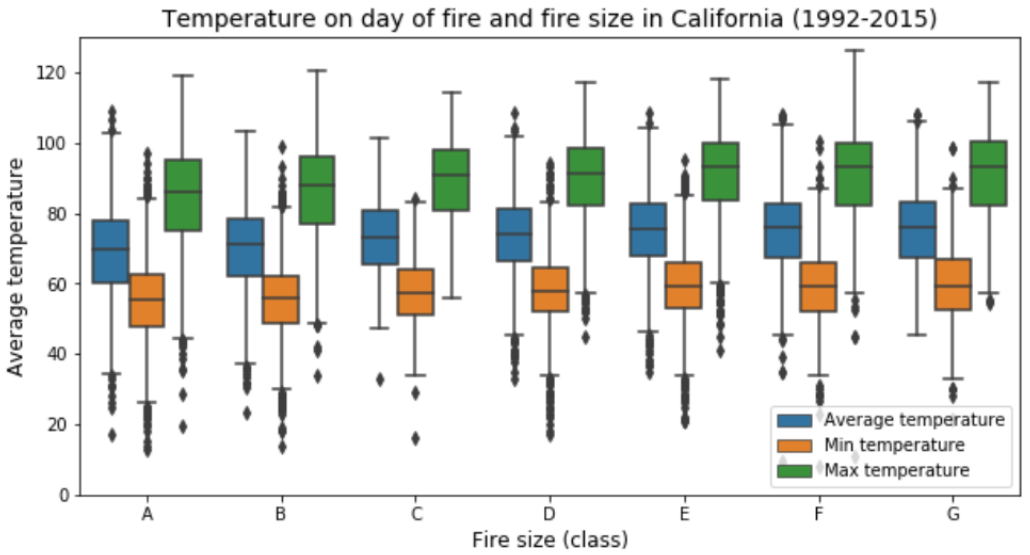
- Larger fires are associated with higher temperatures (average, min, max) on both the day of fire and the 30 days prior to the fire

Average temperature on day of fire
Big fire (classes D to G): 74° (sd = 11.1)
Non-big fire (classes A to C): 69.7° (sd = 12)

Max temperature on day of fire
Big fire (classes D to G): 90° (sd = 12.4)
Non-big fire (classes A to C): 85.3° (sd = 13.8)

Average temperature 30 days prior to fire
Big fire (classes D to G): 71.1° (sd = 9.6)
Non-big fire (classes A to C): 68.6° (sd = 10.5)

→ 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 average temperature (Spearman's rho = .20). Generally, effects are stronger for temperature on day of fire compared to 30-days prior.



Data Exploration: Fires in California and weather

- Larger fires are associated with higher temperatures (average, min, max) on both the day of fire and the 30 days prior to the fire
- Larger fires are associated with less rain in the 30 days prior to the fire

Rain 30-days prior to fire (average, in inches)
Big fire (classes D to G): 0.21 (sd = 0.6)
Non-big fire (classes A to C): 0.40 (sd = 1.1)

→ The relations between amount of rain (30-days) and fires size is negative and statistically significant ($p < .01$). The strength of the relationship is rather weak (Spearman's rho = .-.08).

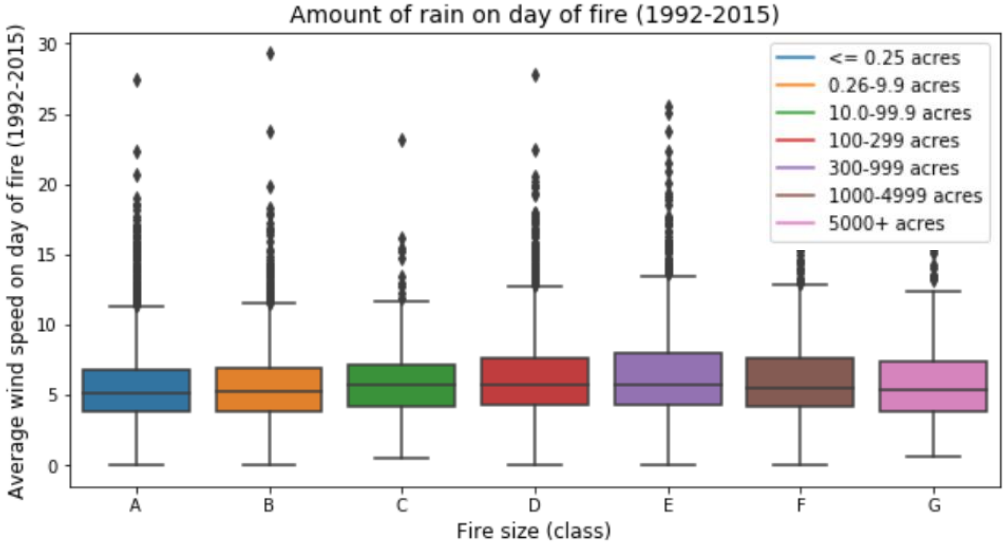
Data Exploration: Fires in California and weather

- Larger fires are associated with higher temperatures (average, min, max) on both the day of fire and the 30 days prior to the fire.
- Larger fires are associated with less rain in the 30 days prior to the fire.
- Larger fires are associated with more wind in terms of both average wind speed and max wind speed on the day of fire

Average wind speed on day of fire (in knots)
Big fire (classes D to G): 6.3 (sd = 3.2)
Non-big fire (classes A to C): 5.6 (sd = 2.8)

Max wind speed on day of fire (in knots)
Big fire (classes D to G): 13.3 (sd = 5.2)
Non-big fire (classes A to C): 12.3 (sd = 4.5)

→ The relations between wind speed (average, max) and fires size are positive and statistically significant (p < .01). The strength of the relationship is rather weak (Spearman's rho = . 09 for both).



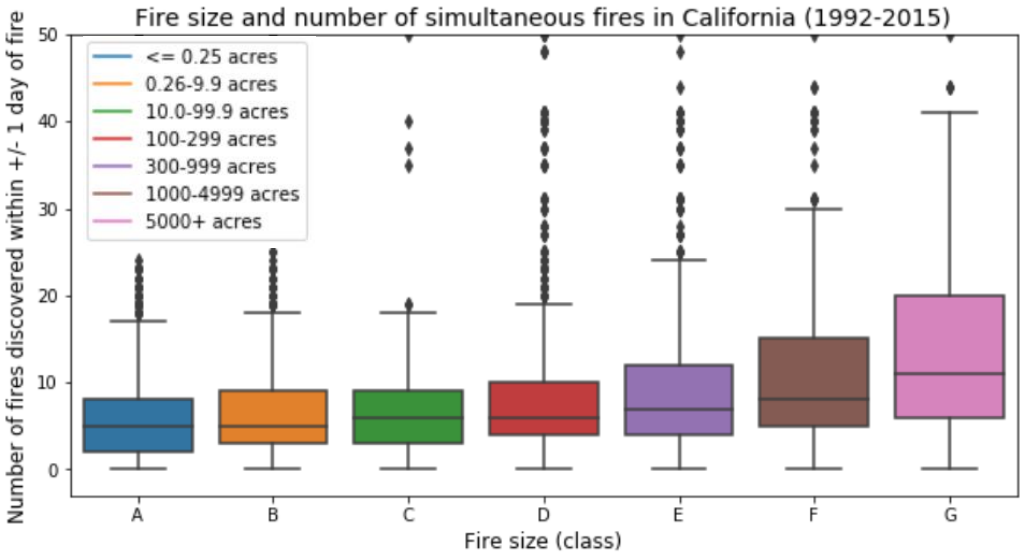
Data Exploration: Fires in California and other fire characteristics

- There is a strong relationship between the size of a fire and the number of simultaneously burning fires in the state. This hints at fire fighting capacity as an important determinant of fire size.

Number of simultaneous fires (discovery +/- 1 day of fire)
Big fire (classes D to G): 12.9 (sd = 22.6)
Non-big fire (classes A to C): 6.7 (sd = 8.5)

Number of simultaneous fires (discovery +/- 3 days of fire)
Big fire (classes D to G): 22.7 (sd = 26.3)
Non-big fire (classes A to C): 14.4(sd = 13.2)

→ 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.



Data Exploration: Fires in California and other fire characteristics

- There is a strong relationship between the size of a fire and the number of simultaneously burning fires in the state. This hints at fire fighting capacity as an important determinant for fire size.

- The number of previous fires (past 3 years) in the same area (5km^2) is slightly lower for big fires. This relation is not significant.

- There is a relation between the area burnt by previous fires (past 3 years) that took place in the same area (5km^2) and fire size. If many acres were burnt previously, a fire is more likely to grow big. This may be explained by remoteness of fires.

Number of fires within 5 km² (past 3 years)
Big fire (classes D to G): 0.85 (sd = 1.3)
Non-big fire (classes A to C): 1 (sd = 1.8)

Area burnt by fires within 5 km² (past 3 years, in acres)
Big fire (classes D to G): 954 (sd = 5489)
Non-big fire (classes A to C): 455 (sd = 4806)

→ The relations between number of close fires in the past 3 years and fire size is not significant. However, the relation between area burnt by these fires and fire size is significant ($p < .01$). The relationships is rather weak (Spearman's rho = .7).

Modeling Overview

Goal
Predict fire size (big vs. non-big) based on weather data and fire characteristics

Models
- Type: Supervised learning
- Binary classification: 1 for “big” fire (classes D to G), 0 for “non-big” fire (classes A to C)
- Models:
  - Random Forest (GradientBoost, XGBoost)
  - Logistic Regression
  - (K Nearest Neighbors)

Target and feature variables
- Target: Fire size (binary)
- Features
  - Weather
  - Fire characteristics
  - Location

Target variable
Fire size (binary, “big” vs. “non-big”)

Feature variables

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.)

Fire 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.)

Location
Latitude
Longitude

Feature selection

Target variable

Fire size (binary, big vs. non-big)

Feature variables

Weather

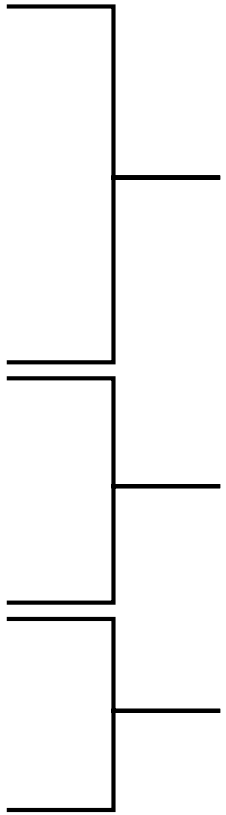
Temperature on day of fire (min, max, average) (3 var.)
Temperature 30 days prior to fire (min, max, average) (3 var.)
Rain on day of fire
Rain 30 days previous to fire
Wind speed on day of fire (average, max) (2 var.)

Fire 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.)

Location

Latitude
Longitude

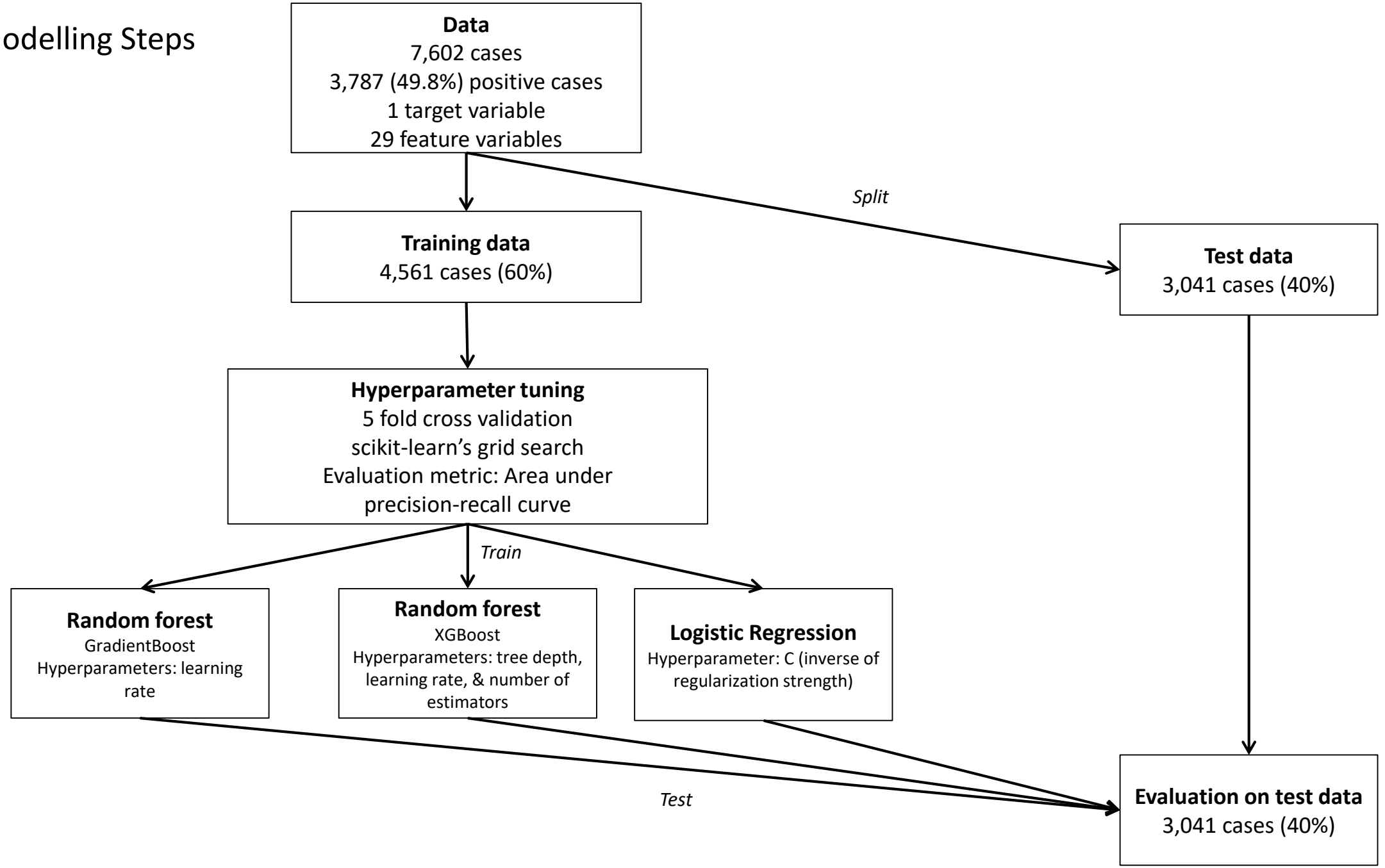


- *Weather characteristics found to be relevant during exploratory analysis.*
- *Weather characteristics reasonably related to fire*

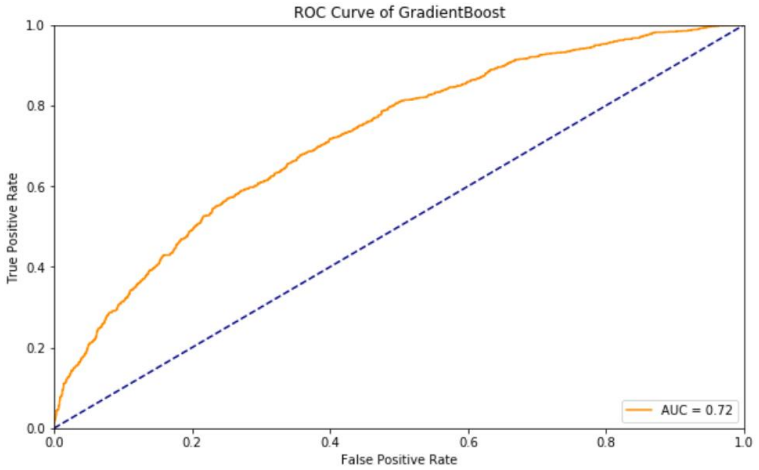
- *Fire characteristics that can reasonably be known or guessed when a fire is discovered*
- *Fire characteristics found to be associated with fire size during exploratory analysis*
- *Characteristics found to have considerable variation and that may reasonably affect fire size (e.g., month)*

- *Spatial disparities (exploratory analysis)*
- *Account for general geographic, terrain, and resource characteristics not included*

Modelling Steps

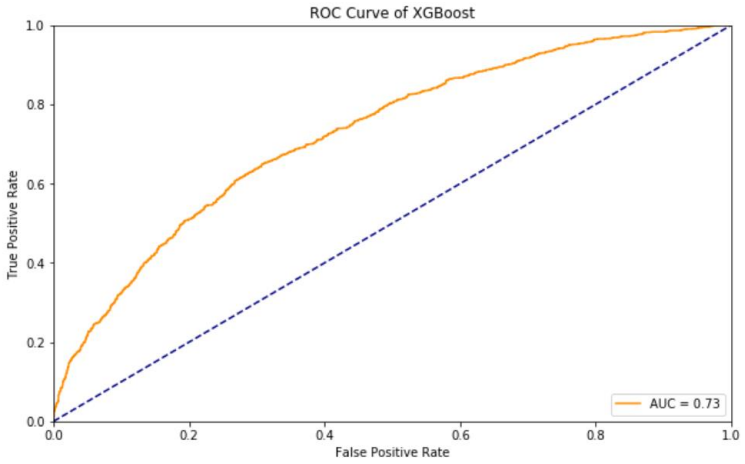


Model Comparison: Performance



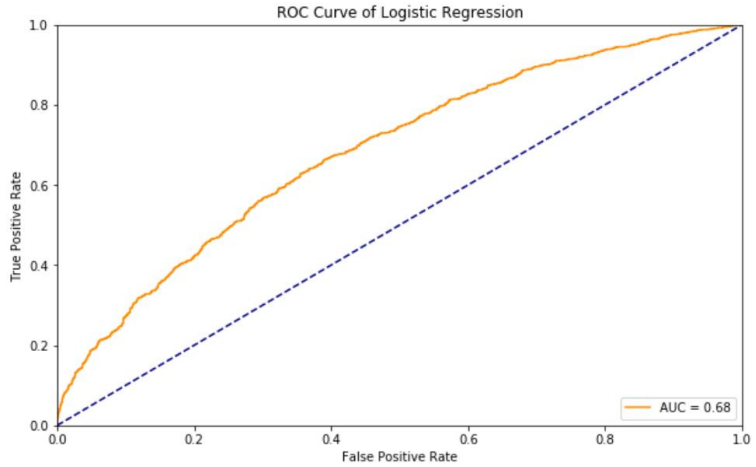
Random Forest (GradientBoost)

- Performance
  - Accuracy 0.65
  - Recall (positive cases): 0.64
  - Precision (positive cases): 0.66
  - f1-score (positive cases): 0.65
- Most important features
  - Simultaneous fires (+/- 1 day)
  - Area burnt by prior fires (3 years)
  - Latitude
  - Longitude
  - Average temperature on day of fire



Random Forest (XGBoost)

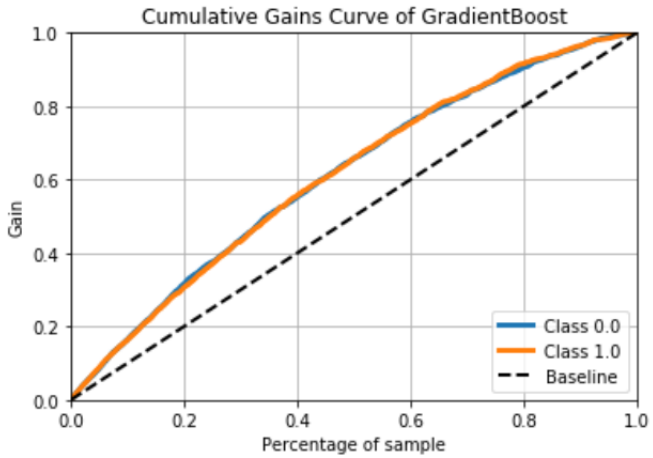
- Performance
  - Accuracy 0.67
  - Recall (positive cases): 0.66
  - Precision (positive cases): 0.67
  - f1-score (positive cases): 0.66
- Most important features
  - Simultaneous fires (+/- 1 day)
  - Area burnt by prior fires (3 years)
  - Fire caused by lightning
  - Month of June
  - Month of July

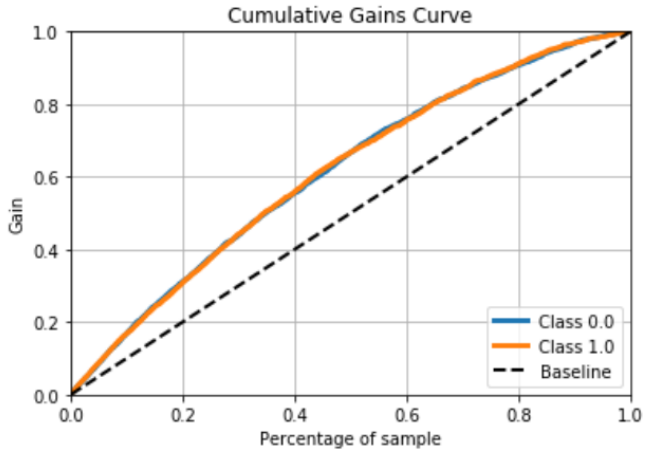


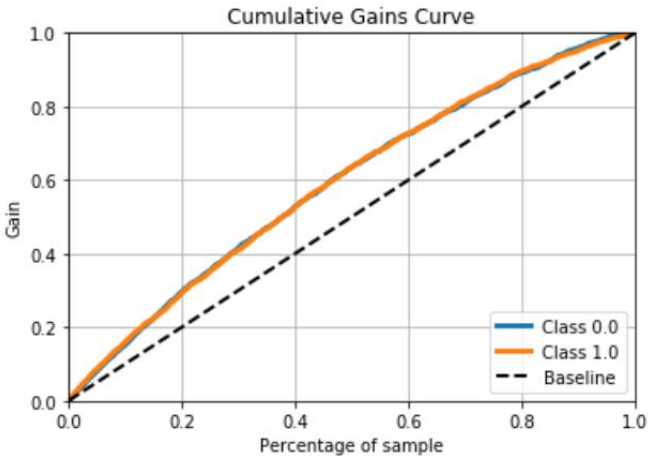
Logistic regression

- Performance
  - Accuracy 0.63
  - Recall (positive cases): 0.61
  - Precision (positive cases): 0.64
  - f1-score (positive cases): 0.63
- Most important features
  - Month of January
  - Month of August
  - Month of December
  - Month of September
  - Fire cause lightning

Model Comparison: Cumulative Gains Curve







- Some gain but overall mediocre performance
- Random forest performs better than logistic regression
- Limited opportunity to improve performance by adjusting the cut-off value

Model generalizability: Predicting “big” fires in Arizona

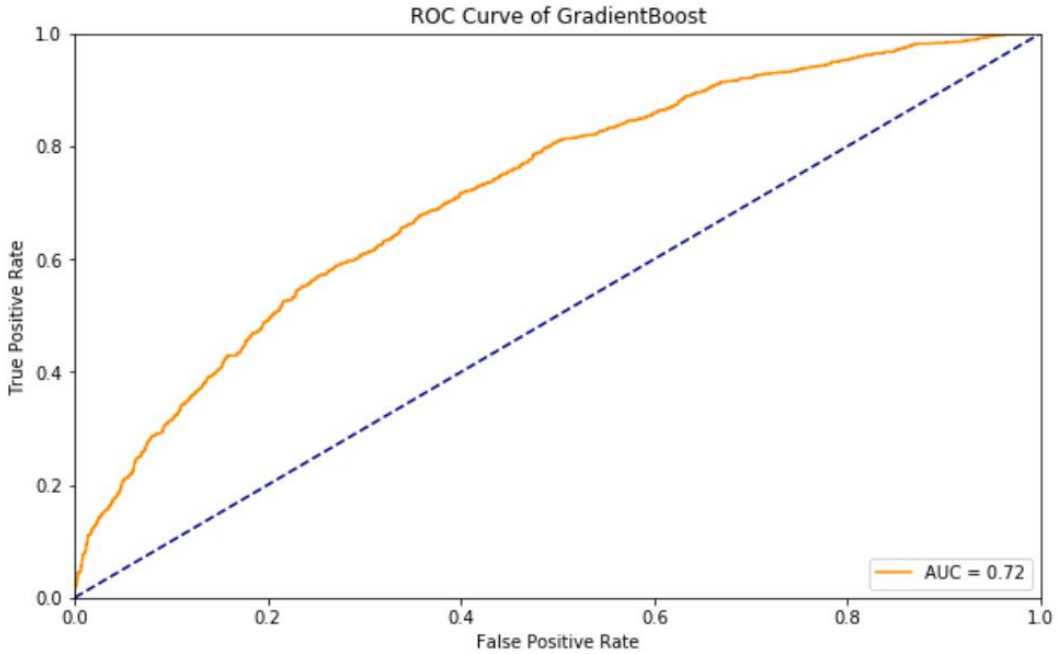
- Sample of 3,222 wild fires in Arizona 1,611 (50%) positive cases (“big” fires)
- Overall performance rather comparable to California sample
- GradientBoost performs best

→ Indicates potential for generalizability

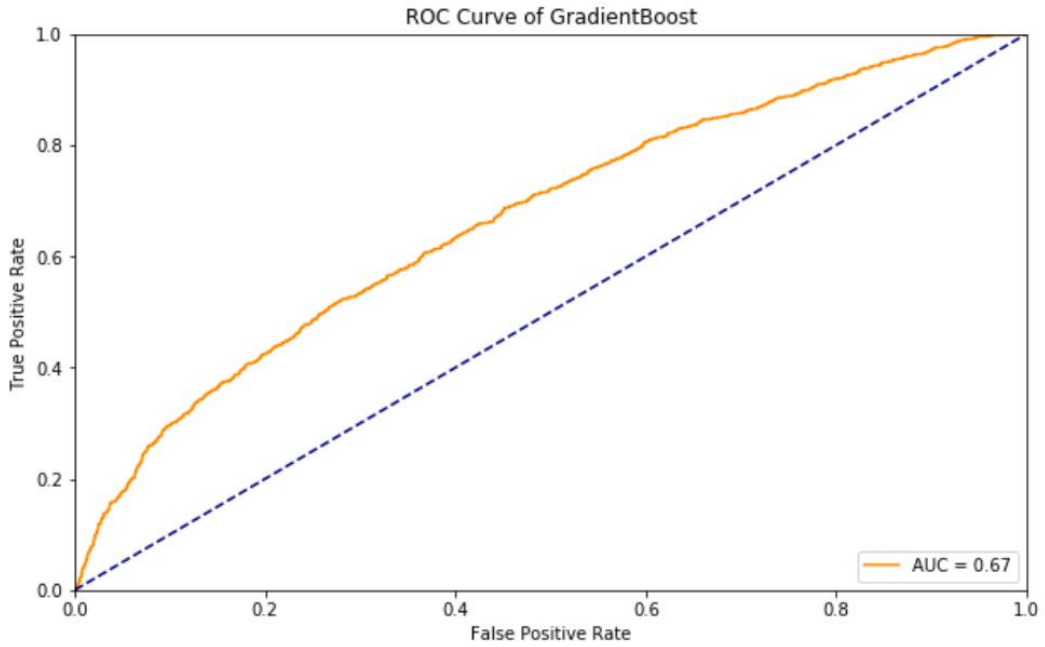
Model	Class	Values			Accuracy	ROC AUC
		Precision	***Recall***	***F1-score***		
Random Forest (GradientBoost)	0	0.70	0.34	0.46	0.58	0.67
	1	0.55	0.85	0.66		
Random Forest (XGBoost)	0	0.66	0.28	0.40	0.55	0.63
	1	0.53	0.84	0.65		
Logistic regression	0	0.63	0.49	0.55	0.59	0.64
	1	0.56	0.69	0.62		

Performance measures for predicting “big” fires in Arizona using the California model

Model generalizability: Predicting “big” fires in Arizona



ROC curve GradientBoost California data



ROC curve GradientBoost Arizona data

Assumptions, Limitations, & Potential Usage

Assumptions & Limitations

- ***Model simplicity***: Using weather and fire data from the day of fire discovery only
- ***Data accuracy***: Weather data from the nearest station may not account for local phenomena (e.g., localized rain); some errors in the data (e.g., wrong dates)
- ***Binary target variable***: Variation in the categories "big" and "non-big" fires not accounted for
- ***Narrow feature variables***: Model mainly focuses on weather variables and some fire characteristics, other variables such as vegetation or preventive measures are not included
- ***Exploratory analyses:*** Were conducted on the whole dataset instead of on the training data only

Potential use

- Limited use due to mediocre model performance
- Use depends on purpose, i.e., might be useful as a decision supporting tool for employment of limited resources or for informing citizens to facilitate evacuation preparation
- Not suited as sole or major decision making tool; ideally combined with other techniques
- Simplistic model with large opportunities for further elaboration → e.g., management variables, resource variables, and vegetation variables

Conclusions

- Despite their relative simplicity the models provide some gain in predicting fire size; BUT overall performance is mediocre
- Random forest (boosted) good model choices
- Importance of weather variables lower than expected (exploratory analyses & feature importance)
- Analyses hint at importance of fire management aspects (e.g., number of simultaneous fires)
- Models not suited as sole or major decision making tool; need to be combined with other techniques

Contained Bayliss Fire, California (2019)

For source of picture see last slide

References

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

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

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/.

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

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

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

Picture sources

Firefighter in the Santa Ynez Mountains, CA: http://archive.boston.com/bigpicture/2008/07/californias_continuing_fires.html

Person watching a fire: https://www.inc.com/business-insider/california-wildfires-homeowners-fire-insurance.html

Destroyed neighborhood in Paradise, CA: https://www.axios.com/california-camp-wildfire-death-toll-paradise-04db9de9-8354-439f-b7d9-65734b01fa7a.html

Contained Bayliss Fire, California: https://patch.com/california/campbell/sj-firefighters-battling-blaze-south-san-jose

Weather station: https://www.weather.gov/rah/virtualtourasos