Predicting Wildfire Severity in California Leveraging Google Earth Engine and Machine Learning

Allan Kapoor
Springboard School of Data
October 2021

The Need for Wildfire Prediction

Wildfires increasing in severity

Factors: Drought, climate change, development, fire suppression, fuel buildup

Limited firefighting resources

Many fires at once, difficult to prioritize

Beneficiaries: CALFIRE, National Forest Service, National Parks Service, Bureau of Land Management, county fire departments, communities

Problem Statement

In order to maximize the efficiency of California's limited wildland firefighting resources, how can historical wildfire records, along with spatial environment and weather data, be leveraged to predict the severity of wildfires when they occur?

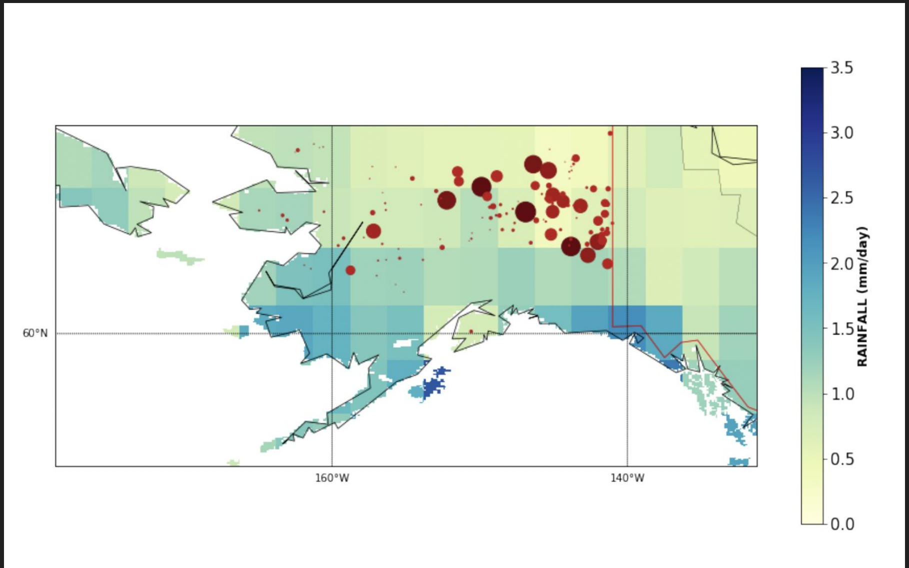
Challenges

Many complex factors, very localized

Influence of weather conditions after fire has started

No explanatory features in wildfire records - need granular spatiotemporal data

Class imbalance



Low spatial/temporal granularity leads to poor predictive power
Credit:
http://noiselab.ucsd.edu/ECE228/projects/Report/75Report.pdf

Data Sources: Wildfire Records

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

- Discovery date, location (latitude and longitude), total burned area
- 1.88 million records total
- Using California only, 2005-2015: 80,000 wildfires

Data Sources: Explanatory Features

Variable	Source	Format
Elevation	USGS National Elevation Dataset (via Google Earth Engine)	2D grid, 10.2 m resolution
Slope	Calculated from elevation dataset	2D grid, 10.2 m resolution
Aspect	Calculated from elevation dataset	2D grid, 10.2 m resolution
Temperature	PRISM Daily Spatial Climate Dataset ("tmax" band)	Gridded time series, 4 km resolution
Dew point	PRISM Daily Spatial Climate Dataset ("dtmean" band)	Gridded time series, 4 km resolution
Precipitation	PRISM Monthly Spatial Climate Dataset ("ppt" band)	Gridded time series, 4 km resolution
Wind speed	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset ("vs" band)	Gridded time series, 4 km resolution
Energy Release Component (ERC)	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset ("erc" band)	Gridded time series, 4 km resolution
Burning index (BI)	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset ("bi" band) v	Gridded time series, 4 km resolution

Variable	Source	Format
100-hour dead fuel moisture	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset ("fm100" band)	Gridded time series, 4 km resolution
1000-hour dead fuel moisture	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset ("fm1000" band)	Gridded time series, 4 km resolution
Vegetation Type	CALFIRE Forest and Rangeland Assessment	2D grid, 30 m resolution
Level III Ecoregions	Ecoregions of the Continental United States, US EPA	Shapefile (polygon vector)
Burn probability	Wildfire Hazard Potential for the United States, US Forest Service	2D grid, 270 m resolution
Fire intensity level (1-6)	Wildfire Hazard Potential for the United States, US Forest Service	2D grid, 270 m resolution

Data Wrangling, Feature Generation

Extract records from SQLite database

Processing Julian dates

Spatiotemporal data: Google Earth Engine

Other spatial data: GeoPandas, Rasterio

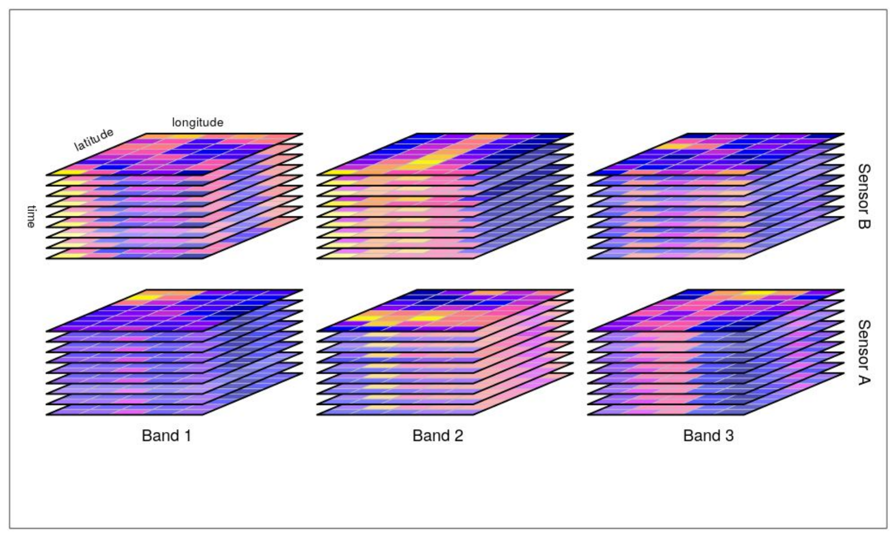
Leveraging Google Earth Engine

Cloud hosting, cloud processing of spatial data

4D multiband gridded spatiotemporal weather data - temporal averaging

Topographic data - aspect and slope calculations

earthengine-api

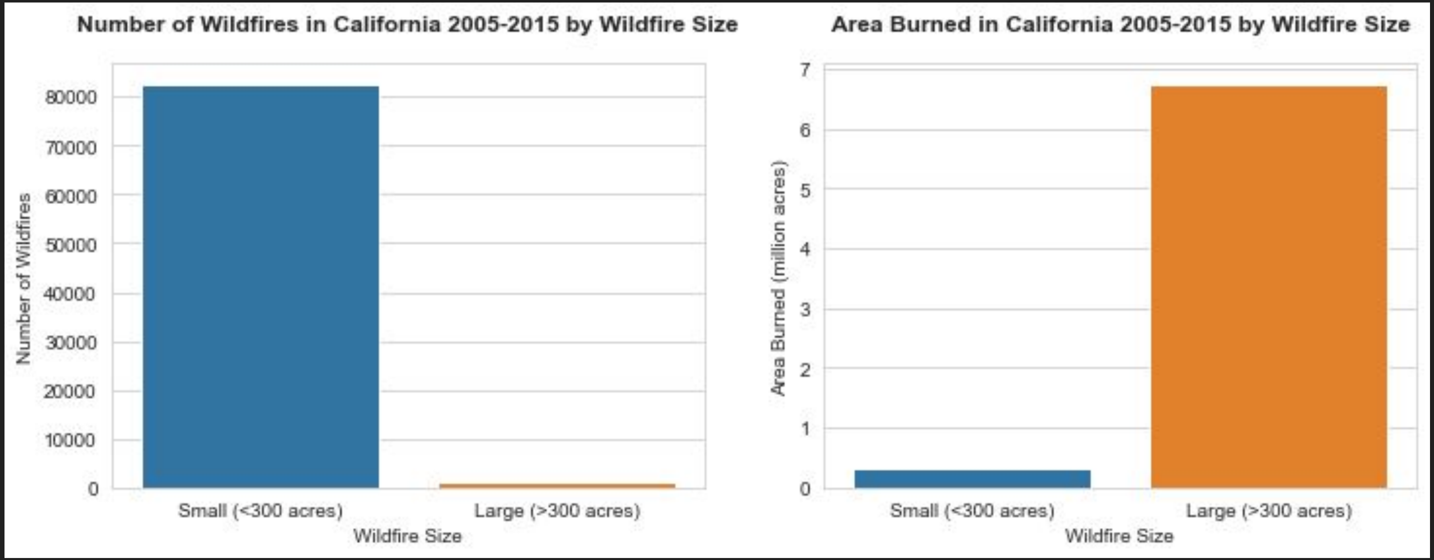


Credit: https://r-spatial.github.io/stars/

EDA: Wildfire size

Numerous small fires: 51% 0.25 acres or less

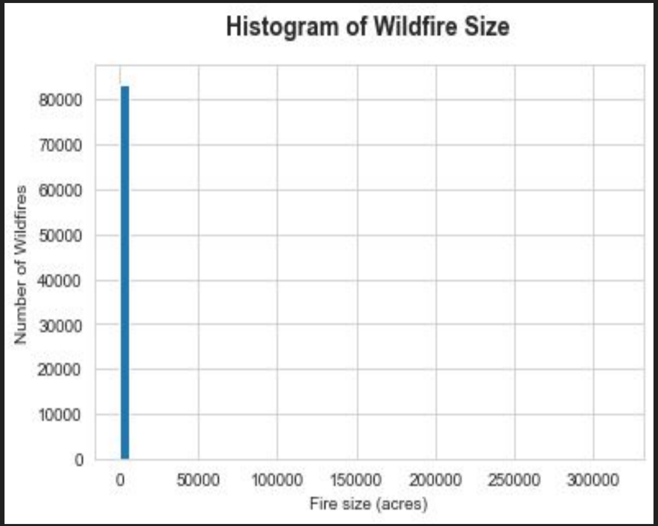
Large fires do majority of damage: 1.2% of wildfires make up 96% of area burned

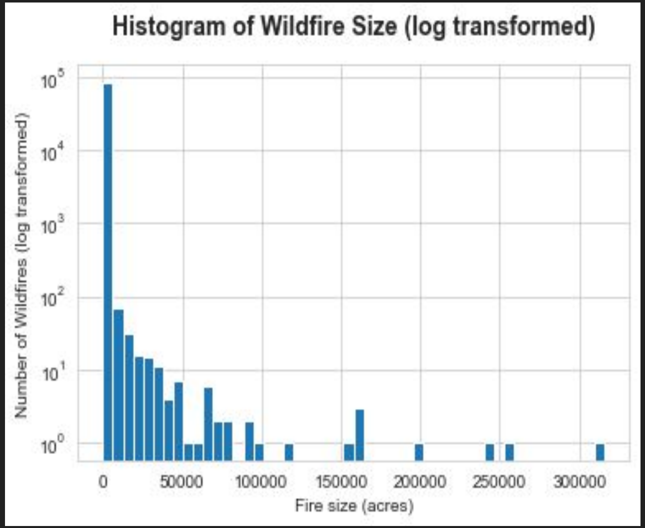


EDA: Wildfire size

Distribution of wildfire size very skewed

Many small wildfires, a few massive outliers





EDA: Wildfire size

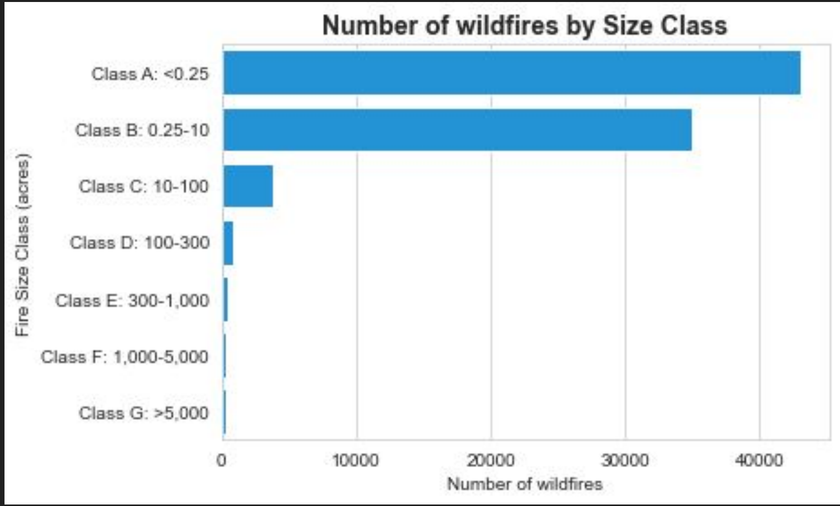
National Wildfire Coordinating Group: Fire Size Classes

Dropping Class A (<0.25 acres)

Small = Class B, C, D (0.25 - 300 acres)

Large = Class E, F, G (>300 acres)

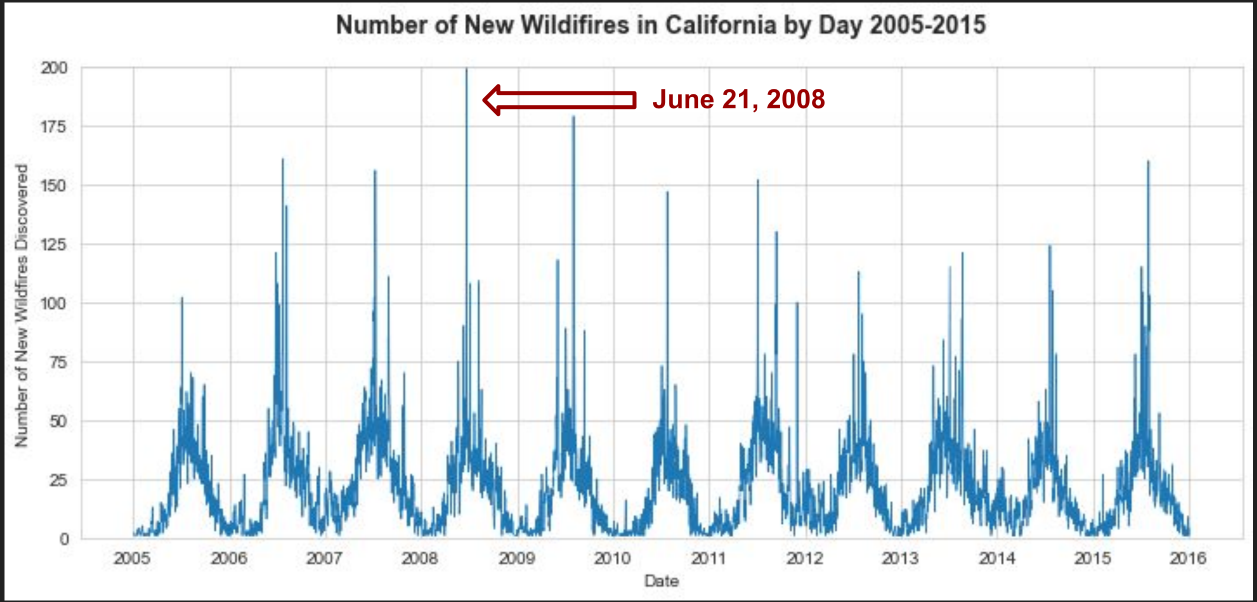
Binary classification: will fire burn >300 acres?

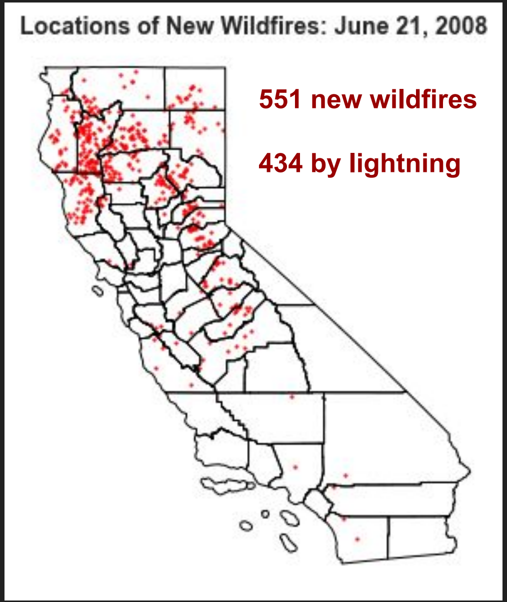


EDA: Wildfire Frequency

Seasonal oscillations, summer peaks

33 days with 100+ new wildfires. 38 new wildfires per day in August.





EDA: Explanatory Features vs Fire Size Class

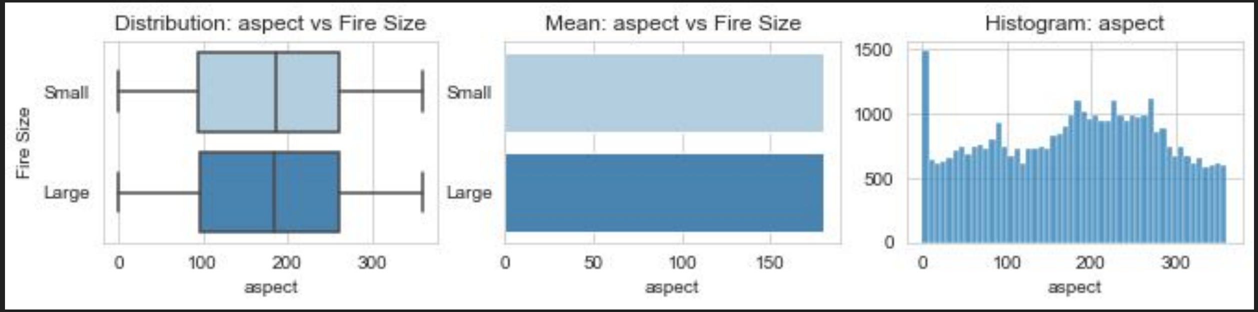
Elevation

p<0.0001

Aspect

p=0.8398

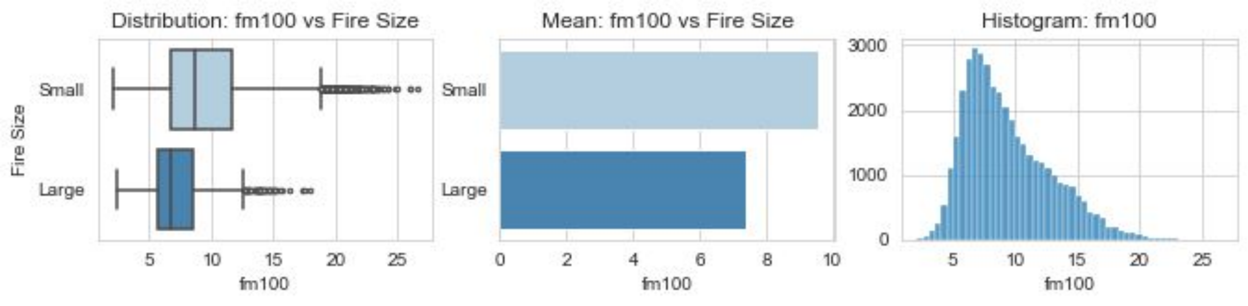
(later processed as a cyclical variable)



EDA: Explanatory Features vs Fire Size Class

100-hour fuel moisture

p<0.0001



Wind speed

p=0.0093

EDA: Explanatory Features vs Fire Size Class

Two-tailed T test

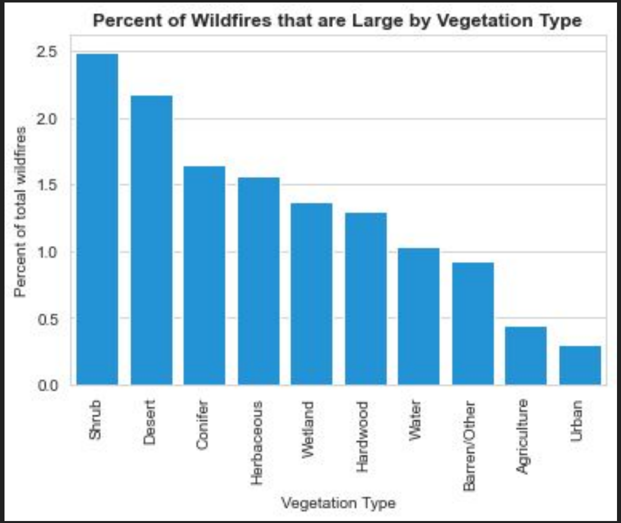
Small vs large wildfires

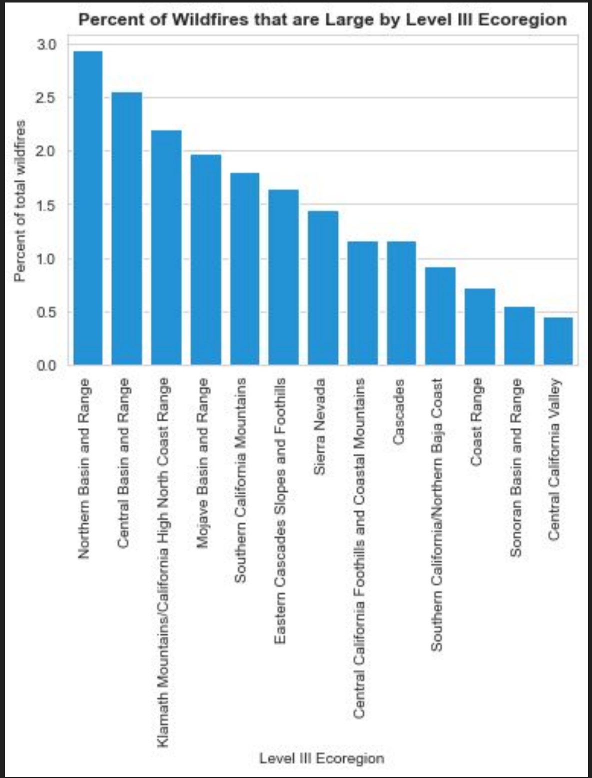
Most variables have statistically significant difference

Feature	T-Score	P Value		Feature	T-Score	P Value
elevation	-13.2816	<0.0001		**fm100**	27.1172	<0.0001
slope	-17.3167	<0.0001		**fm1000**	23.7555	<0.0001
aspect	-0.2021	0.8398		**burn probability**	-8.2959	<0.0001
day of the year	-2.1209	0.0342		**fire intensity level 1**	1.9139	0.0559
precipitation	-2.9386	0.0034		**fire intensity level 2**	-9.2578	<0.0001
dew point	6.8842	<0.0001		**fire intensity level 3**	-10.3161	<0.0001
max temp	-13.2021	<0.0001		**fire intensity level 4**	-4.4639	<0.0001
wind speed	-2.6042	0.0093		**fire intensity level 5**	-1.2976	0.1947
energy release component	-20.6889	<0.0001		**fire intensity level 6**	-1.0156	0.31
burn index	-7.5722	<0.0001				

EDA: Explanatory Features vs Fire Size Class

Categorical features





Modelling Approach

Preprocessing

- Transformations by feature
- For each feature, transformation that reduces skew the most
- Cyclical variables > dual harmonic features (aspect and discovery day of the year

Evaluation Metric

- F2 Score (most important to correctly classify positive samples)
- Confusion matrices (on validation set)

Evaluation Method

- 10-fold cross validation

Modelling Approach

Hyperparameter Tuning

- RandomizedSearchCV
- Optuna

Addressing Class Imbalance

- Class weighting
- SMOTE Oversampling/Random Undersampling - Imbalanced Learn

Algorithms tested

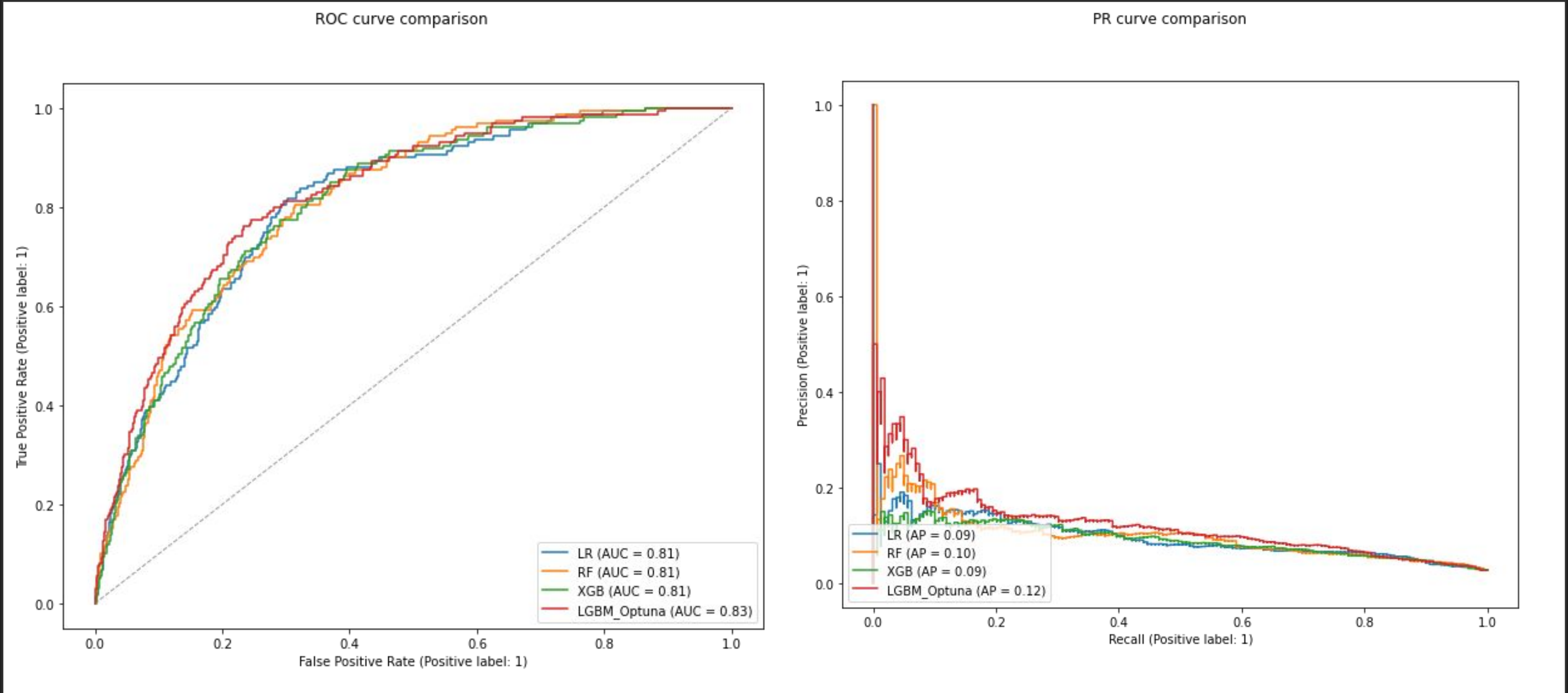
- Logistic regression, random forest, XGBoost, LightGBM
- With/without SMOTE

Model Evaluation & Selection

Ordered by CV Mean F2 score:

	Cross Validation Mean			Validation Set		
	f2	**recall**	**roc auc**	**f2**	**recall**	**roc auc**
LGBM (Optuna)	0.296	0.457	0.688	0.287	0.434	0.678
LGBM	0.280	0.447	0.68	0.259	0.415	0.662
XGB	0.260	0.55	0.699	0.258	0.535	0.694
RF	0.255	0.413	0.661	0.268	0.447	0.674
LGBM (SMOTE)	0.248	0.476	0.674	0.243	0.440	0.662
RF (SMOTE)	0.240	0.507	0.678	0.234	0.522	0.678
XGB (SMOTE)	0.239	0.493	0.674	0.249	0.509	0.683
LR	0.232	0.748	0.731	0.243	0.780	0.749
Dummy	0.025	0.025	0.500	0.019	0.019	0.496

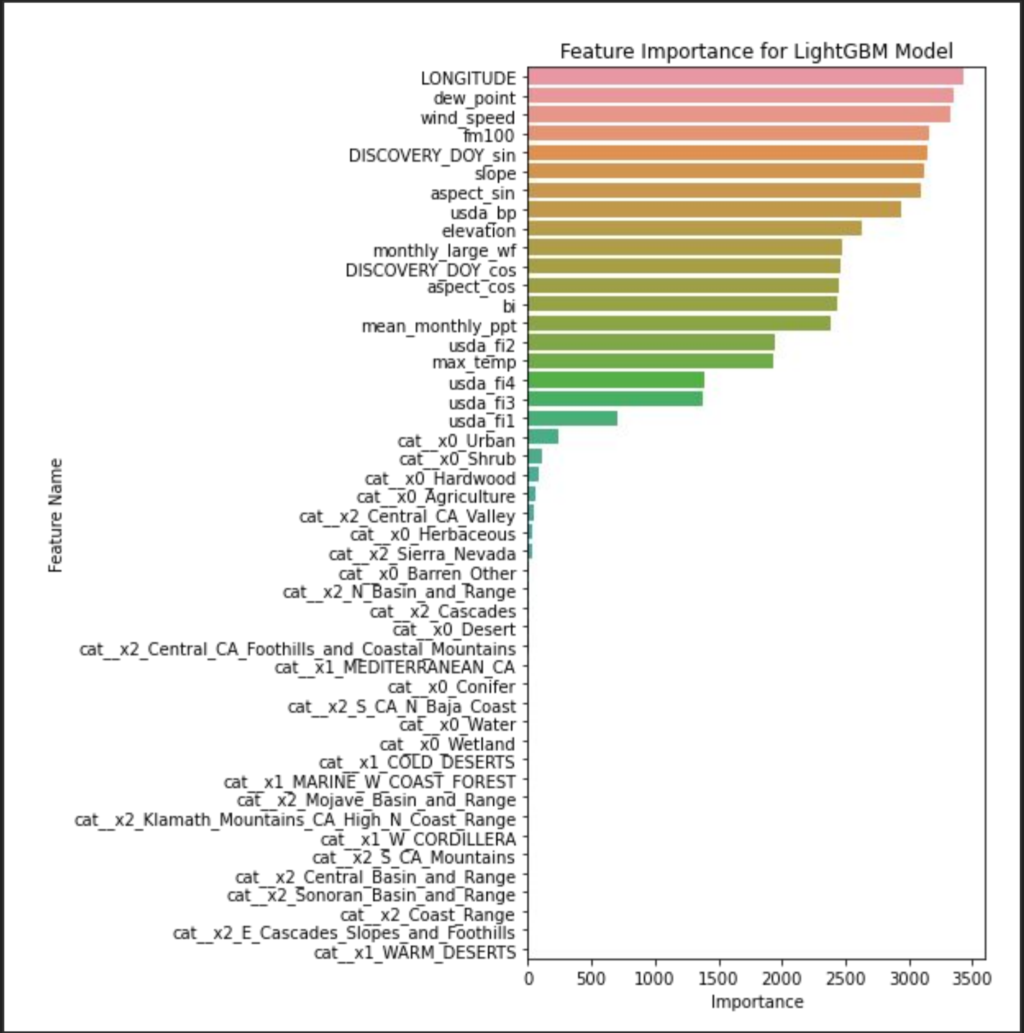
Model Evaluation & Selection



Selected Model

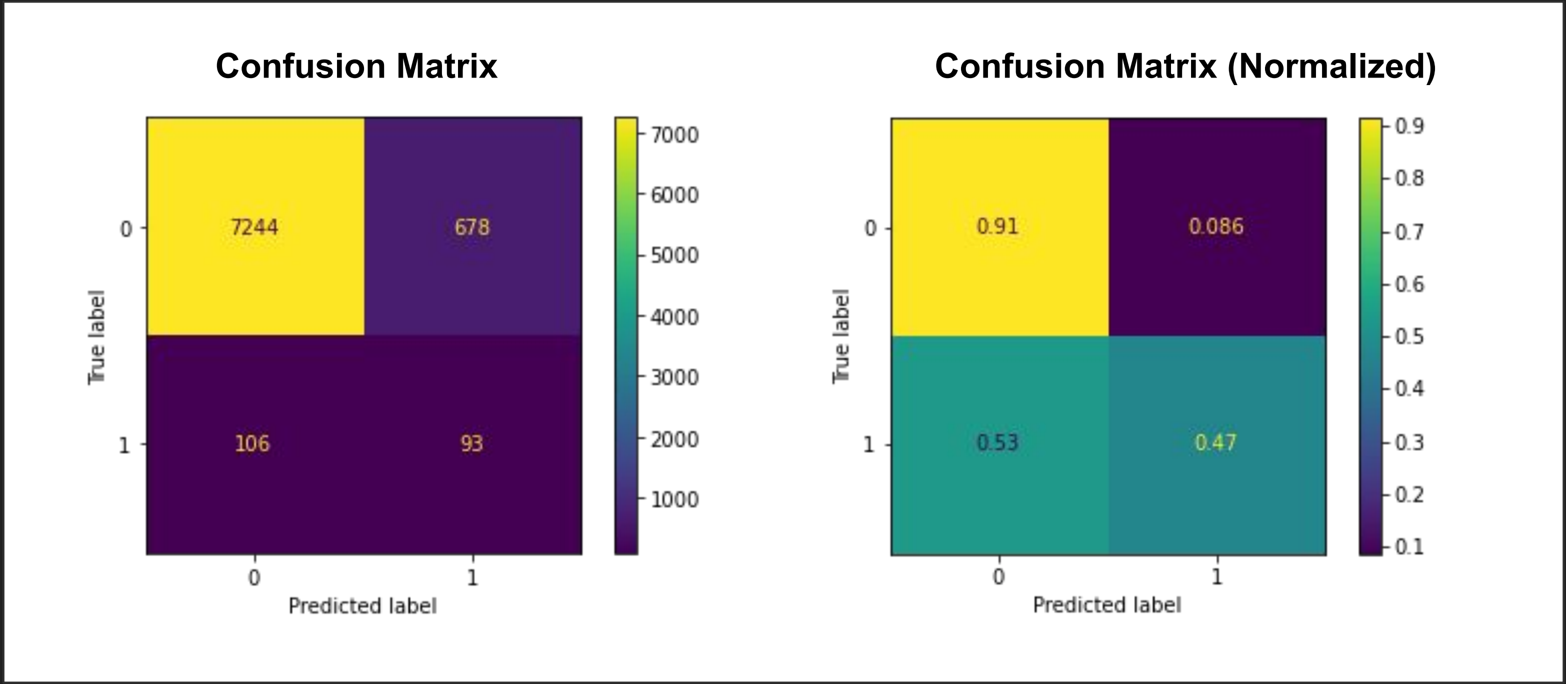
LightGBM model turned with optuna

Hyperparameter	Value
n_estimators	850
learning_rate	0.037
num_leaves	200
max_depth	12
min_data_in_leaf	200
lambda_l1	50
lambda_l2	100
min_gain_to_split	0.322
bagging_fraction	0.800
bagging_freq	1.000
feature_fraction	0.500



Selected Model: Performance on Test Set

f2	recall	roc auc
0.297	0.467	0.691



Conclusion: Opportunities

Refine/add explanatory features:

- Revisit time frames for weather features
- Proximity to roads, CALFIRE airports
- Quantitative environmental variables: Normalized Difference Vegetation Index (NDVI), canopy density, fuel load, etc.

Scope/training:

- Expand timeframe from 2005-2015 to 1992-2018 (new data available)
- Limit model to peak fire wildfire season months and/or high hazard areas

Conclusion: Using the Model

Make predictions for future wildfires as they occur.

Front-end interface:

- Wildfire location via map
- Date
- Automatic generation of explanatory features

Batch processing: upload shapefile, geojson, or csv of new wildfire locations