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

Allan Kapoor[1]

Introduction

This report summarizes an effort to predict wildfire severity based on the location, time of the year, and environmental features of historic wildfires leveraging machine learning algorithms. Wildfire records (location, date, and burned area) were sourced from the United States Forest Service (USFS)[2]. These records were enriched with additional features from several spatiotemporal datasets via Google Earth Engine and then used to train a series of machine learning algorithms to predict, if given a location and date of discovery, a wildfire will burn greater than 300 acres[3].

The best performing model leverages LightGBM, a gradient boosting framework that uses tree based learning algorithms. This model achieves an F2 score of 0.297 and an ROC AUC of 0.691 on test data. Predictive features are limited to those that would be accessible for future wildfires shortly after they are discovered (therefore weather patterns following date of discovery and the cause of fire are not used). This model could be used by wildland firefighters in California (CALFIRE, county fire departments, US Forest Service, etc.) to prioritize active fires based on threat level in order to more efficiently allocate limited firefighting resources. The methodology could also be applied to other states across the US.

Problem Statement

Wildfires are a major natural hazard in California. Each year, wildfires cause widespread destruction of property and environmental resources, disrupt lives, contribute to poor air quality, and sometimes even deaths. While the landscape of California has always been fireprone (in fact, many California ecosystems are adapted to wildfire), the severity of wildfires has increased

[1] Thanks to Shmuel Naaman for mentorship and advice on feature engineering/algorithms, and thanks to Diana Edwards for help thinking through relevant explanatory features, appropriate time frames for weather variables, and sources for vegetation data.

[2] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

[3] For the purposes of this project, “large” wildfires are defined as those that burn 300 or more acres (National Wildfire Coordinating Group Class E or higher https://www.nwcg.gov/term/glossary/size-class-of-fire)

substantially in recent years. Six of the the top ten largest wildfires on record in California were in the last two years and 18 out of the top 20 where in the last 10 years[4].

This alarming trend is due to several interrelated factors. The State's history of suppressing wildfires coupled with multi-year drought conditions have led to a massive build-up of dry fuel and hazardous weather conditions. Historically, wildfires were a part of the natural cycle of forest ecosystems - clearing underbrush, leaving large trees intact, returning nutrients to the soil, and making way for new life. Now, fires are more likely to rage out of control, moving into the crowns of trees and decimating entire landscapes. CALFIRE's stated goal is to keep 95% of fires at 10 acres or less - but it is the other 5% that threaten to burn large swaths of forest and even entire towns.

A model that predicts which fires have the potential to become most severe would enable responders to make more informed decisions about how to allocate limited resources, protecting lives, property, and the environment.[5] Beneficiaries would include CALFIRE, responsible for fires within the State Responsibility Areas, federal agencies such as the National Forest Service, National Park Service, and the Bureau of Land Management, as well as individual county fire departments and public officials in communities with high fire risk.

> **Problem Statement:** In order to maximize the efficiency of California's limited wildland firefighting resources, how can we leverage historical wildfire records along with spatial environment and weather data to predict the potential severity of wildfires when they occur?

Challenges

Wildfire behavior is very complex and notoriously difficult to model. Initial spread of a fire is influenced by many factors, such as available fuels, humidity, wind speed and direction, to name a few. Many of these factors are very localized (for example, a fire may start in bushes near a tree with low-hanging branches, forming a "ladder" for the fire to travel up). Additionally, while weather conditions and human activities in the days following a wildfire's discovery have a strong influence on fire behavior, a useful predictive model should not require inputs that could only be known after the time of prediction.

A major challenge associated with this project's primary data source is that it does not include any of the explanatory weather or environmental features that are identified in the literature as being strong predictors of wildfire behavior. A previous attempt to predict wildfire size using the same dataset by graduate students in the Department of Electrical and Computer Engineering

[4] CALFIRE. 2021. Top 20 Largest California Wildfires. Accessed: https://www.fire.ca.gov/media/4jandlhh/top20_acres.pdf

[5] Noting that the prediction of wildfire behavior is an entire field of study and that robust wildfire models have been developed and are currently used by major firefighting agencies. However, these are generally model complex physical processes, rather than machine learning approaches. This project leverages outputs of one of these models as explanatory features. Recently, machine learning has been increasingly applied to wildfire science. See: https://cdnsciencepub.com/doi/full/10.1139/er-2020-0019

at the University of California, San Diego met limited success[6]. That study joined weather variables based on date/location, but the chosen weather data source consisted of monthly averages with very low spatial resolution. According to the authors, potential explanations for low model performance include "unpredictable human activities, low correlation from monthly-average climate data, and lack of geographical features including elevation, slope, and soil type." For this project, the latter two issues will be addressed by a) leveraging daily weather data with higher spatial resolution and b) joining spatial data on topography and vegetation.

Another anticipated challenge is that while the primary data source includes over 80,000 wildfires within California from 2005 to 2015, less than 1,000 of these are large wildfires. This severe class imbalance, with the positive class as the minority, is likely to negatively impact prediction. For this project, I attempt to address this through class weighting and over/ undersampling (see Modelling Approach).

Data Sources[7]

Wildfire Records

The primary data source used for this project is the official wildfire database maintained by the US Forest Service[8]. The wildfire records were acquired from the reporting systems of federal, state, and local fire organizations and have undergone a rigorous quality control process intended to eliminate duplicate records. At a minimum, each wildfire record includes discovery date, location (latitude and longitude), and final fire size. The full dataset spans the years 1992-2015 and includes 1.88 million records.

This project leverages data on wildfires in California only for the period 2005-2015. This limited the dataset to a size that could be processed on a personal computer, but it also has conceptual justifications as well. Wildfire behavior varies across different states, so a model that tries to make predictions for the entire US might be generally accurate but less accurate for a given state. Wildfire severity in California has become noticeably worse over the last 30 years, so a model trained on events from the 1990s may not predict future wildfire severity accurately. Filtering the wildfire dataset (see Data Wrangling) to California wildfires 2005-2015 still returns over 80,000 events.

[6] Xiong, Wu, and Chen. 2018. Machine Learning Wildfire Prediction based on Climate Data. University of San Diego, Department of Electrical and Computer Engineering. Accessed: http://noiselab.ucsd.edu/ECE228/projects/Report/75Report.pdf

[7] Data wrangling and feature generation notebook: https://github.com/allankapoor/wildfire_prediction/blob/master/Step1_DataWrangling-FeatureGeneration.ipynb

[8] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

Additional Explanatory Features

As mentioned previously, the wildfire dataset does not include explanatory features that could be used to predict wildfire behavior, so these features needed to be joined from other sources based on time and location. Table 1 below summarizes each of the explanatory features that were joined to the wildfire records. Links in the source column also include additional context information about each dataset.

Table 1: Sources of Explanatory Features

Variable	**Time frame***	**Description**	**Source**	**Format**
Elevation	n/a	Vertical elevation above sea level (NAVD 88), meters	USGS National Elevation Dataset (via Google Earth Engine)	2D grid, 10.2 m resolution
Slope	n/a	Vertical slope, degrees	Calculated from elevation dataset	2D grid, 10.2 m resolution
Aspect	n/a	Direction of slope face, degrees from North (clockwise)	Calculated from elevation dataset	2D grid, 10.2 m resolution
Temperature	Preceding 7 days	Maximum daily temperature, °C	PRISM Daily Spatial Climate Dataset (“tmax” band)	Gridded time series, 4 km resolution
Dew point	Preceding 7 days	Daily mean dew point temperature - a measure of air moisture, °C	PRISM Daily Spatial Climate Dataset (“dtmean” band)	Gridded time series, 4 km resolution
Precipitation	Preceding year	Monthly precipitation, millimeters	PRISM Monthly Spatial Climate Dataset (”ppt” band)	Gridded time series, 4 km resolution
Wind speed	Day of discovery	Wind speed, meters per second	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset (“vs” band) via Google Earth Engine	Gridded time series, 4 km resolution
Energy Release Component (ERC)	Day of discovery	The ERC is an index related to the available energy (BTU) per unit area (square foot) within the flaming front at the head of a fire. Each daily calculation considers the past 7 days.	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset (“erc” band) via Google Earth Engine	Gridded time series, 4 km resolution
Burning index (BI)	Day of discovery	A measure of fire intensity. BI has no units, but in general it is 10 times the flame length of a fire.	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset (“bi” band) via Google Earth Engine	Gridded time series, 4 km resolution
100-hour dead fuel moisture	Day of discovery	Represents the modeled moisture content of dead fuels in the 1 to 3 inch diameter class. Values can range from 1 to 50 percent.	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset (“fm100” band) via Google Earth Engine	Gridded time series, 4 km resolution
1000-hour dead fuel moisture	Day of discovery	Represents the modeled moisture content in dead fuels in the 3 to 8 inch diameter	GRIDMET: University of Idaho Gridded Surface Meteorological Dataset	Gridded time series, 4 km resolution

		class. Values can range from 1 to 40 percent.	(“fm1000” band) via Google Earth Engine	
Vegetation Type	n/a	Vegetation types, compiled from a variety of state/federal sources into a single comprehensive data set	CALFIRE Forest and Rangeland Assessment	2D grid, 30 m resolution
Level III Ecoregions	n/a	Ecoregions are areas where ecosystems (and the type, quality, and quantity of environmental resources) are generally similar.	Ecoregions of the Continental United States, US EPA	Shapefile (polygon vector)
Burn probability	n/a	Output from the FSim probabilistic wildfire model. The burn probability dataset is the simulated mean annual burn probability.	Wildfire Hazard Potential for the United States, US Forest Service	2D grid, 270 m resolution
Fire intensity level (1-6)	n/a	Output from the FSim probabilistic wildfire model. The fire intensity level dataset consists of six raster files, each representing the portion of all simulated fires that burned in the cell area at the specified flame length.	Wildfire Hazard Potential for the United States, US Forest Service	2D grid, 270 m resolution

Initial Dataset

Data wrangling, and all subsequent steps, was carried out using Python 3 in a Jupyter Notebook. The USFS wildfire dataset is available online as an SQLite database. The dataset has already gone through a rigorous vetting process to identify duplicates. Records with STATE = ‘CA’, FIRE_YEAR ≥ 2005 were queried and saved to a Pandas DataFrame. Discovery date was converted from Julian date format to YYYY-MM-DD. The DataFrame was then converted to a GeoPandas GeoDataFrame to enable spatial processing. Missing values for the COUNTY column were filled via a spatial join to county boundary polygons[9]. A few wildfire records with STATE= ‘CA’ actually had coordinates outside of California - these were dropped from the dataset.

Google Earth Engine

Topography variables (elevation, slope, aspect) as well as weather and environmental variables from the PRISM and GRIDMET datasets were accessed via Google Earth Engine (GEE). GEE is a spatial cloud computing platform that hosts a wide variety of geospatial datasets (with a focus on remote sensing/satellite imagery and weather data) and enables users to perform computationally intensive analyses on Google’s cloud via JavaScript and Python APIs.

Leveraging the earthengine-api Python package, a series of custom functions were written to extract relevant data for the date and location of each wildfire. Slope and aspect were calculated

[9] US Census. 2016. TIGER/Line Shapefile, 2016, state, California. Accessed: https://catalog.data.gov/dataset/tiger-line-shapefile-2016-state-california-current-county-subdivision-state-based

based on elevation and then extracted. Some weather variables were extracted for a single day (e.g., wind speed). For others (e.g., temperature, dew point), a series of data points from a range of days (or months) preceding the wildfire date were extracted and then averaged. Once processed in GEE, these variables were joined to the main wildfires table.

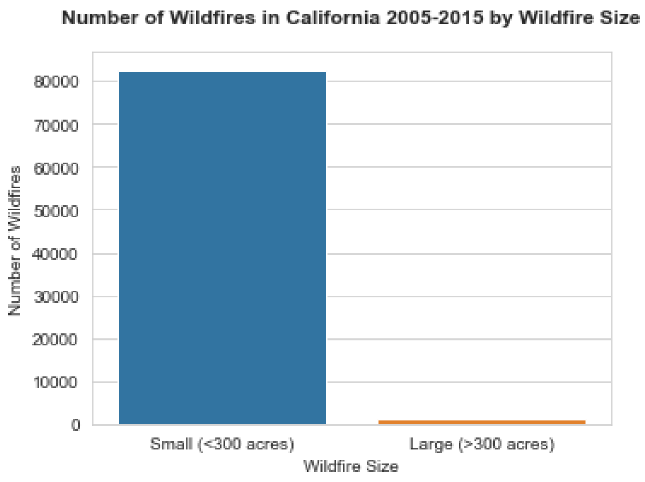
Other Spatial Datasets

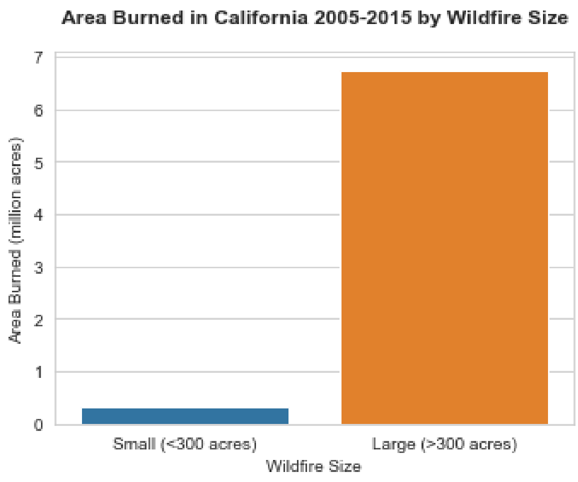
Additional spatial datasets such as Vegetation Type and Ecoregions were downloaded and added to the main table via spatial joins (leveraging GeoPandas for polygon vectors and Raster.io for gridded data). All necessary geographic coordinate system/projection conversions were performed before joining to ensure spatial accuracy.

Exploratory Data Analysis[10]

Before attempting to train a machine learning algorithm, it was important to validate some initial claims and to explore relationships between the target and explanatory features, as well as general trends in the data.

First, is it true that the majority of wildfires are small, but large wildfires cause most of the damage? Out of the 83,606 wildfires recorded in California 2005-2015, 51% each burned 0.25 acres or less (about a third the size of an American football field). Only 1.2% of the wildfires burned 300 acres or more. However, that 1.2% of wildfires contributed to 96% of the total area burned.

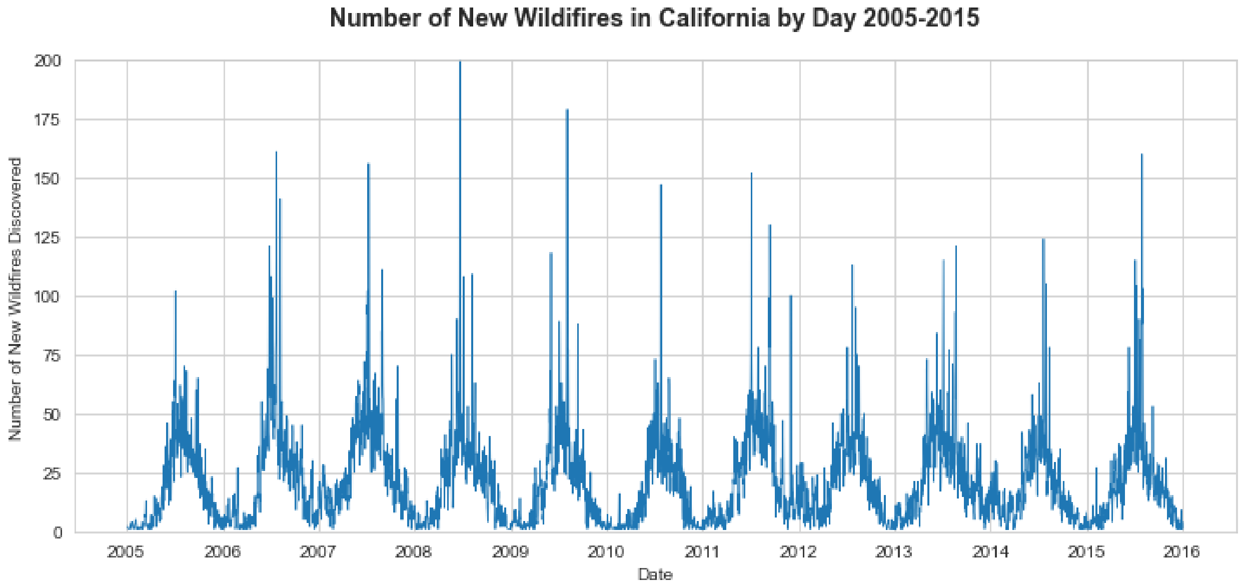

Number of Wildfires in California 2005-2015 by Wildfire Size


Area Burned in California 2005-2015 by Wildfire Size

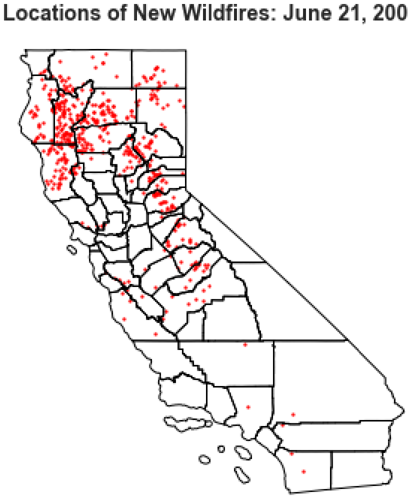
During peak wildfire season, there may be hundreds of small wildfires active at the same time throughout the state. How frequently do these fires start? The figure below displays the number of wildfires discovered in California each day from 2005 to 2015. Seasonal oscillations are apparent, with new wildfires per day peaking in the summer. There are 33 different days when

[10] EDA notebook: https://github.com/allankapoor/wildfire_prediction/blob/master/Step2_ExploratoryDataAnalysis.ipynb

100 or more new fires were discovered in a single day. Across all August months (peak wildfire season) throughout the timeframe, an average of 38 new wildfires are discovered every day.



The maximum number of wildfires discovered in a single day was on June 21, 2008, when 551 new wildfires were discovered across the State, 434 of them caused by lightning due to an anomalous outbreak of thunderstorms[11]:

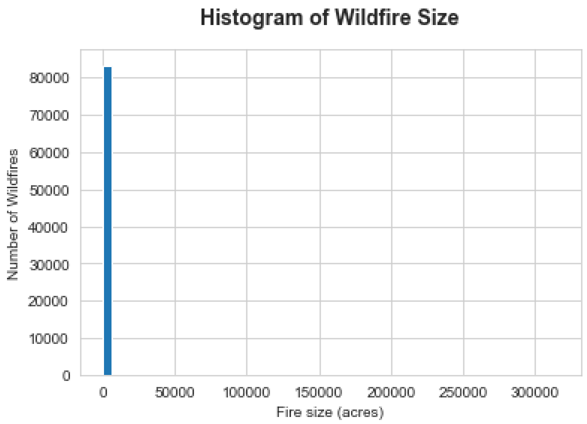


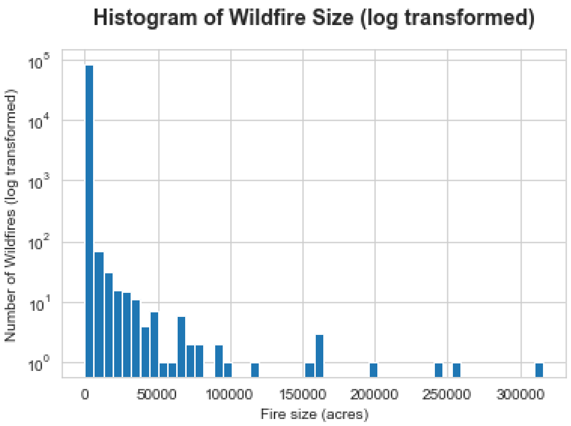
Distribution of Wildfire Size

As mentioned above, small wildfires are much more numerous than large wildfires. A histogram of wildfire size is so skewed that only one bin is visible. A log-transformed histogram is more

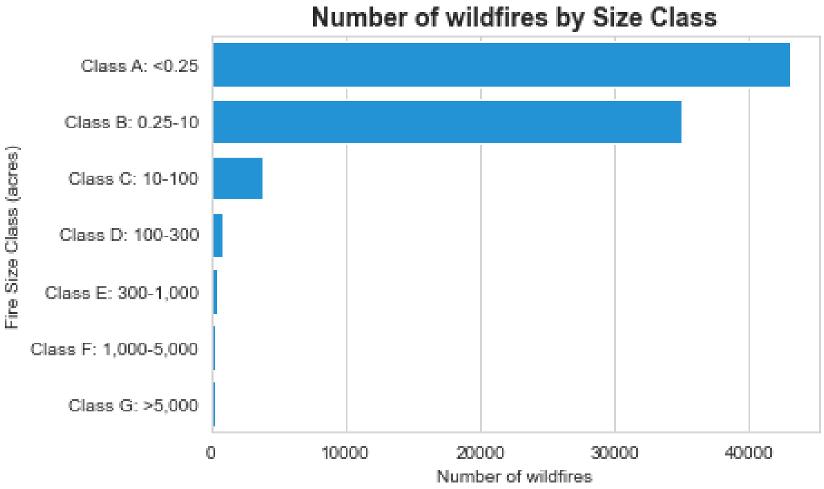
[11] https://www.weather.gov/media/wrh/online_publications/TAs/TA0814.pdf

legible and reveals a strong right-tailed distribution. This is caused by a) a large number of very small wildfires (62.71% of the wildfires are 1 acre or less) and b) a handful of extremely large wildfires (99% of wildfires are <400 acres but there are 9 wildfires that each burned more than 100,000 acres). The largest wildfire as of 2015 burned 315,579 acres (almost 500 square miles).





The National Wildfire Coordinating Group defines wildfires into size classes based on area burned[12] - these classifications are already assigned to each record in the wildfire dataset. The number of wildfires recorded in each class are summarized in the figure below.



The figure displays the severe class imbalance between small and large wildfires. While an ideal model would be able to predict the size class of a given wildfire, due to the low number of records of the larger size classes, a binary classification model will likely achieve better results and still be useful for wildfire prioritization. A wildfire that could burn more than 300 acres would definitely be of concern. The class imbalance can also be mitigated by filtering out very small wildfires (Class A, <0.25 acres), which make up a majority of the dataset.

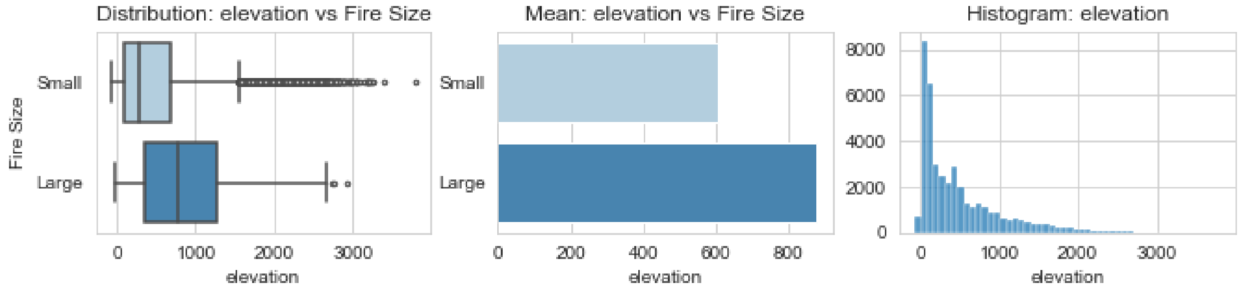
[12] https://www.nwcg.gov/term/glossary/size-class-of-fire

In other words, this project will seek to develop a binary classification algorithm with the goal of classifying wildfires >0.25 acres as “small” or “large”. Subsequent exploratory data analysis was performed on a subset of the full dataset (Class A, <0.25 acres is dropped).

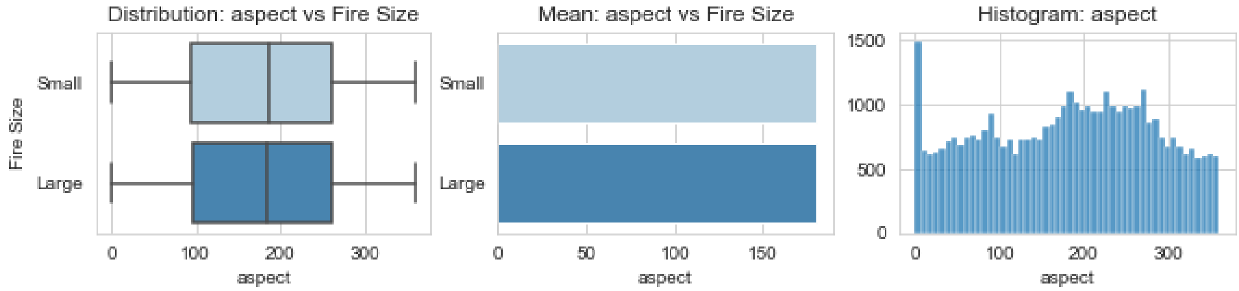
Wildfire Size Category vs. Continuous Explanatory Features

Prior to model development, the relationship between each explanatory feature and the binary target variable (small or large) was explored, as well as each feature’s distribution. For brevity, figures are only shown for a subset of features.

Elevation: The mean elevation for large wildfires is higher than for small wildfires (p<0.0001) even though wildfires with the highest elevations are small. This is likely because at very high elevations there is less vegetation to burn.

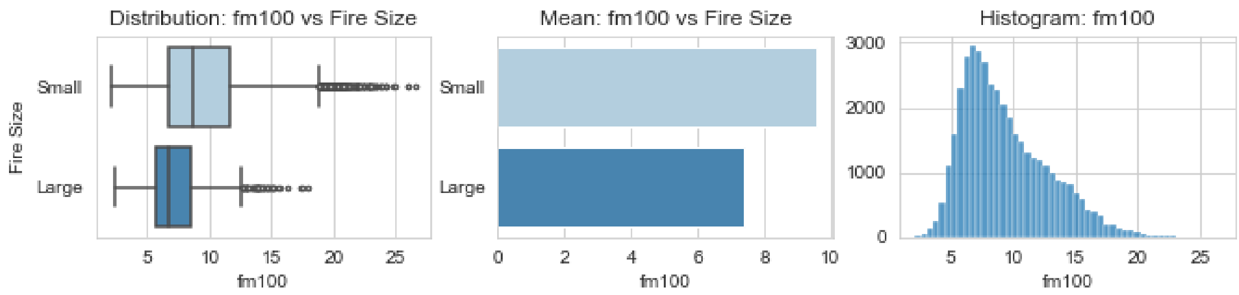


Aspect (degrees from North): Distributions and means of aspect for small and large fires are not significantly different (p=0.8398). This is initially surprising because wildfire risk is supposedly higher for south facing slopes than north facing slopes[13]. The lack of apparent difference is actually because aspect degrees are a cyclical variable (360 degrees and 0 degrees are both due North). Aspect and discovery day of the year (another cyclical variable) will be treated differently from the other continuous variables (see Modelling Approach).

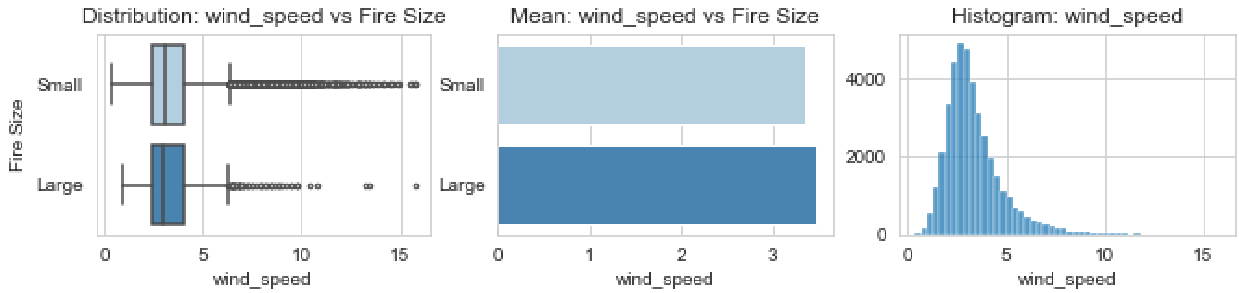


100-hour dead fuel moisture: Small wildfires tend to occur when there is higher 100-hour dead fuel moisture than larger wildfires (p<0.0001), which is to be expected.

[13] CoreLogic. 2021. 2019 Wildfire Risk Report. Accessed: https://storymaps.arcgis.com/stories/cb987be2818a4013a66977b6b3900444



Wind speed: this feature does not have as large a difference in means between small and large wildfires as expected, although the difference is statistically significant (p=0.0093). This may be because the wind speed is only for the day the wildfire is discovered, not for the days/weeks following that.

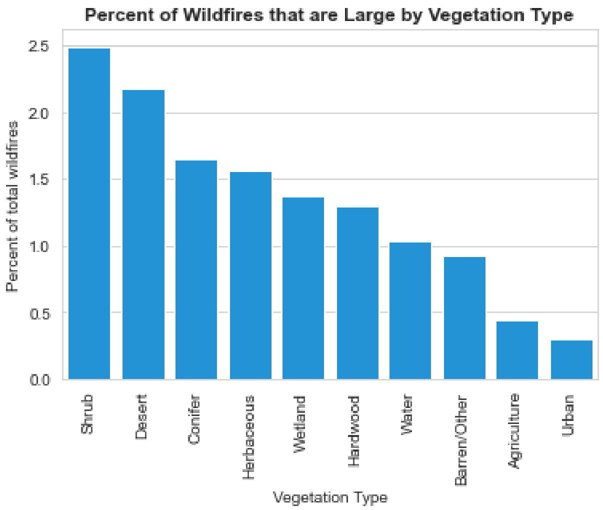


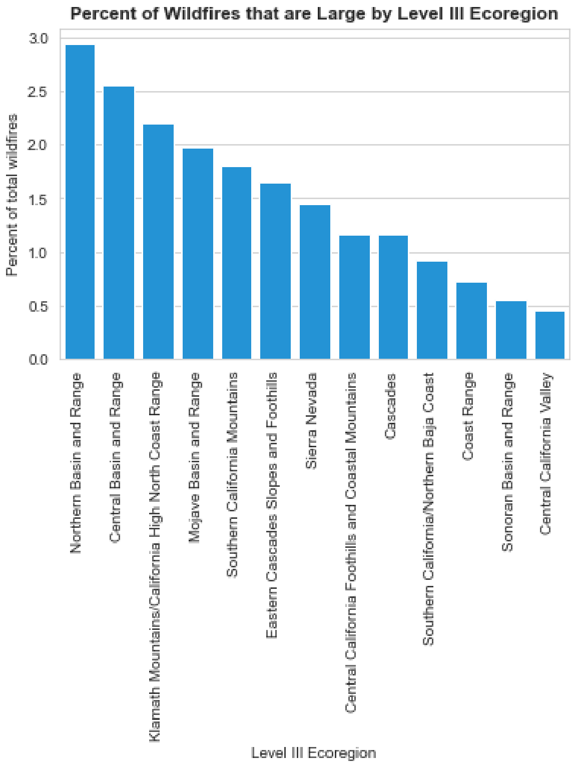
The table below summarizes the results of two-tailed t-tests for all continuous explanatory variables. In general, the means for most of the variables have a significant difference (with low p values) between for small and large wildfires, suggesting that these variables will have predictive power during modelling.

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				

Wildfire Size Category vs. Categorical Explanatory Features

Three categorical variables denote different vegetation types/ecoregions each wildfire is located within. To explore general trends between wildfire size and the categorical explanatory variables, the proportion of wildfires within each zone that are large was calculated and compared to other zones within that typology. The figures below summarize these results for vegetation type and EPA Level III Ecoregions.





Vegetation type: the vegetation types with the greatest portion of wildfires within them becoming large are shrub, dester, and conifer. The vegetation types with the lowest proportion are barren land, agriculture, and urban. On further inspection, presence of wildfires in the "water" vegetation type is due to the grid resolution of the vegetation type dataset - coordinate locations of some wildfires near rivers/lakes may be within a grid cell designated as water.

Level III Ecoregion: ecoregions with the greatest portion of wildfires within them becoming large are mostly mountainous/forested regions while those with the lowest proportion are the Sonoran Basin (very barren) and Central California Valley (agricultural area).

Modelling Approach[14]

Preprocessing/Feature Engineering

Prior to modelling, transformations were applied to the continuous explanatory variables in order to reduce skew and bring their distributions as close to normal as possible. For each continuous variable, the skew of the original data was measured and then compared with the data transformed by logn(x), log10(x), sqrt(x), cbrt(x), 1/x, x^2, and x^3. Each continuous feature was then transformed using the transformation that reduced skew the most.

A correlation matrix was then used to check for highly correlated variables. Latitude and Longitude were found to have strong intercorrelation, as did energy release component, 100-hour dead-fuel moisture, and 1000-hour dead-fuel moisture. For these sets of intercorrelated features, the feature with the strongest relationship with the target variable was kept and the others were dropped.

As mentioned previously, aspect (i.e. the cardinal direction a slope faces in degrees from north) and discovery day of the year are actually cyclical features in that their values “wrap around” - the highest values are close to the lowest values. In order for this nuance to be apparent to the models, these features were both transformed into dual harmonic variables that swing back and forth out of sync:

$$x=\sin(2\pi*\text{period}) \qquad y=\cos(2\pi*\text{period})$$

Period is the number of units in the full circle (i.e. 365 for days of the year and 360 for aspect degrees). This works due to the fact that the derivative of either sin or cos changes while the (x,y) position varies smoothly as it travels around the period’s circle.

Model Evaluation

F2 score was selected as the metric to evaluate model performance, although other metrics such as ROC AUC and recall were also calculated. While the F1 score is the harmonic mean of precision and recall[15], the F2 score calculates the harmonic mean with an additional coefficient that essentially weights recall higher than precision. The F2 score is applicable to this problem because it is important to correctly classify as many large wildfires as possible (without too many false positives), rather than maximizing the number of correct classifications overall.

The models were evaluated based on the mean F2 score resulting from 10-fold cross validation. Then they were trained on the full cross validation set and tested on a validation set in order to produce confusion matrices (see Model performance). The best performing model was that

[14] Preprocessing notebook:
https://github.com/allankapoor/wildfire_prediction/blob/master/Step3_Preprocessing.ipynb
Modeling notebook: https://github.com/allankapoor/wildfire_prediction/blob/master/Step4_Modeling.ipynb

[15] Precision = TruePositives / (TruePositives + FalsePositives)
Recall = TruePositives / (TruePositives + FalseNegatives)

which produced the highest mean F2 score. This model was then trained on the full training set and tested on the hold-out test set (see Results).

Several algorithms were tried, including: logistic regression, random forest, and two gradient boosting frameworks: XGBoost and LightGBM

Addressing Class Imbalance

As mentioned previously, the severe class imbalance between large and small wildfires was a major challenge. For each algorithm, two methods were tried to address this:

1) Using the algorithms' built-in class weighting, or
2) Synthetic Minority Oversampling Technique (SMOTE) oversampling + random undersampling. SMOTE works by selecting examples from the minority class that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line[16]. Random undersampling drops majority class records in order to reduce class imbalance. Both can be implemented in a machine learning pipeline via the package Imbalanced Learn.

Hyperparameter Tuning

For each algorithm, hyperparameters were initially tuned using RandomSearchCV in which 100 permutations of hyperparameters were tested from a grid. The ranges within the grid were determined based on a review of optimal ranges for imbalanced classification problems. The searches returned the hyperparameter grid with the highest mean F2 score based on 10-fold cross validation. For pipelines that included SMOTE oversampling/undersampling, various hyperparameters for those steps were also included in the hyperparameter grid. Note that because the sampling step is embedded in the pipeline, sampling is never formed on the test fold in each CV run.

Once each algorithm had been tuned with RandomSearchCV, the best performing algorithm was then tuned further using Optuna, a bayesian hyperparameter optimization framework (200 trials, 10-fold cross-validation).

Assessing Models

The performance of each model tested is presented in the table below. The columns to the left summarize the mean results of the 10-fold cross validation and the columns to the right display results when the model was trained on the full cross validation set and then tested on a single validation set. This is the same validation set that was used to produce confusion matrices and ROC/PRC plots for each model, but is not the hold-out test set that is used at the very end of test performance of the best model. The table is ordered based on CV mean F2 score.

[16] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

LightGBM (without SMOTE_ achieved the best results via RandomSearchCV and was then optimized with Optuna, yielding even higher performance.

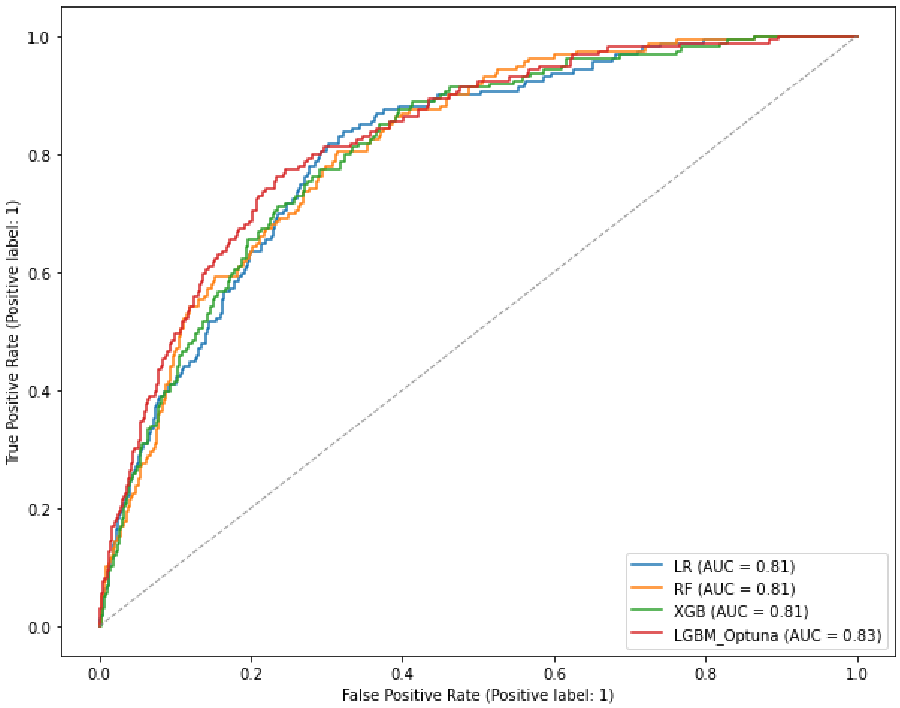
	Cross Validation Mean			**Validation Set**		
	f2	**recall**	**roc auc**	**f2**	**recall**	**roc auc**
LGBM (Optuna)	0.296	0.457	0.688	0.287	0.434	0.678
LightGBM	0.280	0.447	0.68	0.259	0.415	0.662
XGBoost	0.260	0.550	0.699	0.258	0.535	0.694
Random Forest	0.255	0.413	0.661	0.268	0.447	0.674
LightGBM w/ SMOTE	0.248	0.476	0.674	0.243	0.440	0.662
Random Forest w/ 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
Logistic Regression	0.232	0.748	0.731	0.243	0.780	0.749
Dummy Model	0.025	0.025	0.500	0.019	0.019	0.496

A few observations about model performance:

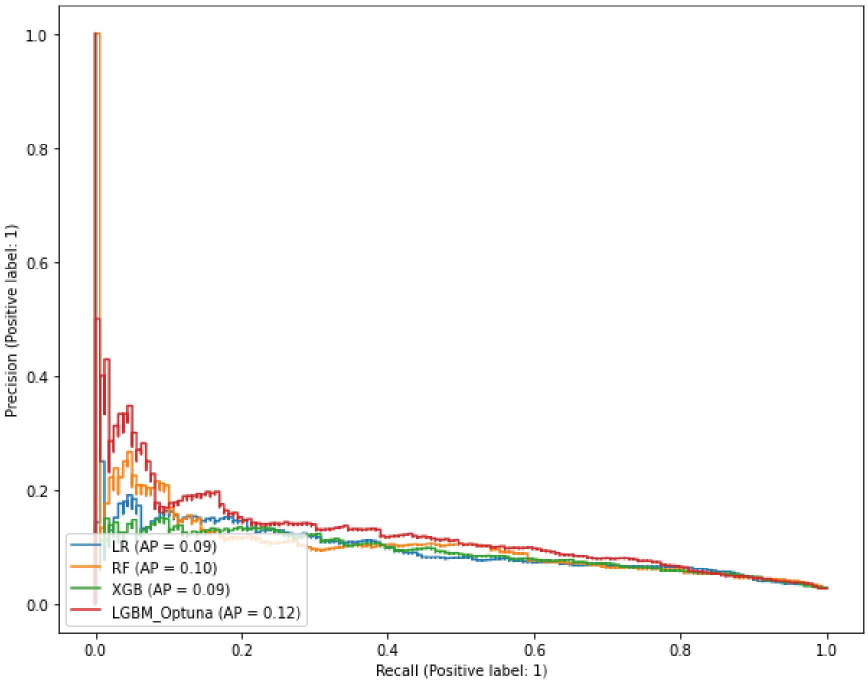
- All algorithms perform substantially better than the dummy model.
- Algorithms with higher F2-scores do not necessarily perform best for other evaluation metrics. For example, Logistic Regression produces the highest recall and ROC AUC, but the lowest F2 score. On further inspection this is because the Precision score for Logistic Regression is very low due to a high number of false positives.
- All ensemble algorithms performed worse when used in tandem with SMOTE oversampling/random undersampling. This was surprising considering that conventional wisdom is that SMOTE increases performance dramatically for imbalanced classification problems such as fraud detection. Poor performance for wildfire severity prediction is likely due to the geospatial nature of the problem - when synthetic samples are generated between real samples in the feature space, the process may generate samples that are physically impossible in the real world (i.e., a wildfire in a particular location with a vegetation type or elevation that is not actually present there). Training the model with these impossible samples included may be leading to increasingly incorrect classification on unseen test data.

The figures below illustrate the trade-off between various performance metrics. The ROC and Precision/Recall curve charts below display this trade-off. From the ROC Curve it is clear that while the LightGBM model performs best when the false positive rate is kept low, but if the false positive rate is allowed to increase, other models may produce a better true positive rate.

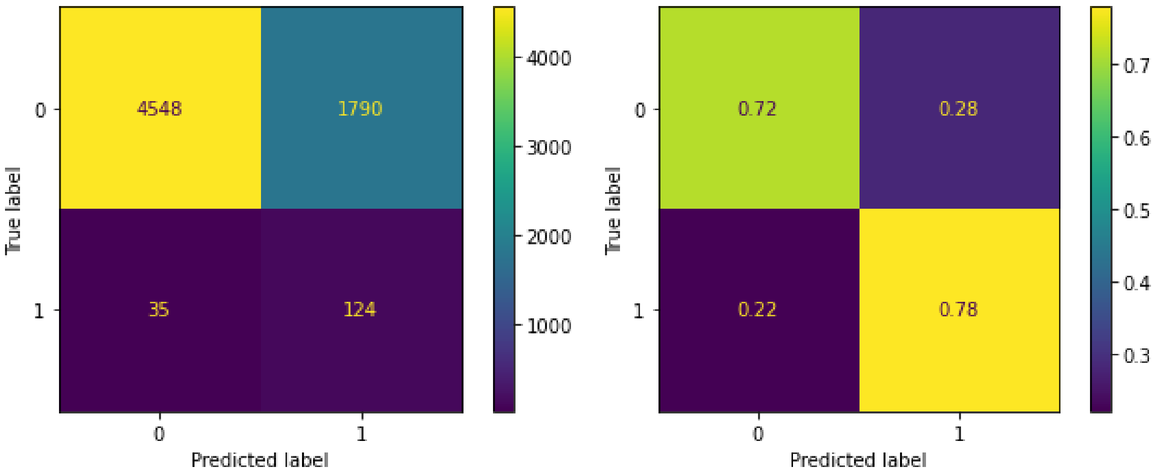
ROC curve comparison



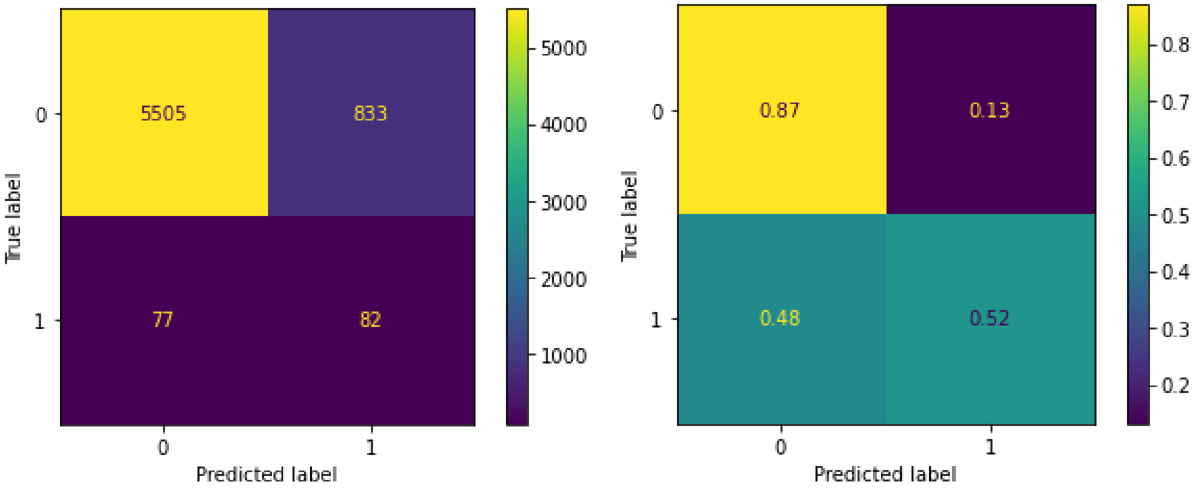
PR curve comparison



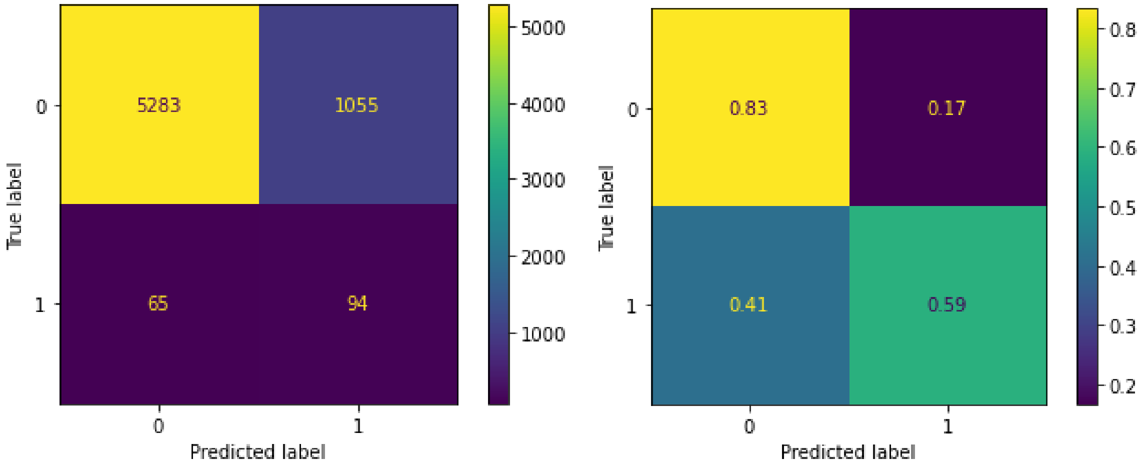
Logistic Regression - Validation Set (non-normalized and normalized)



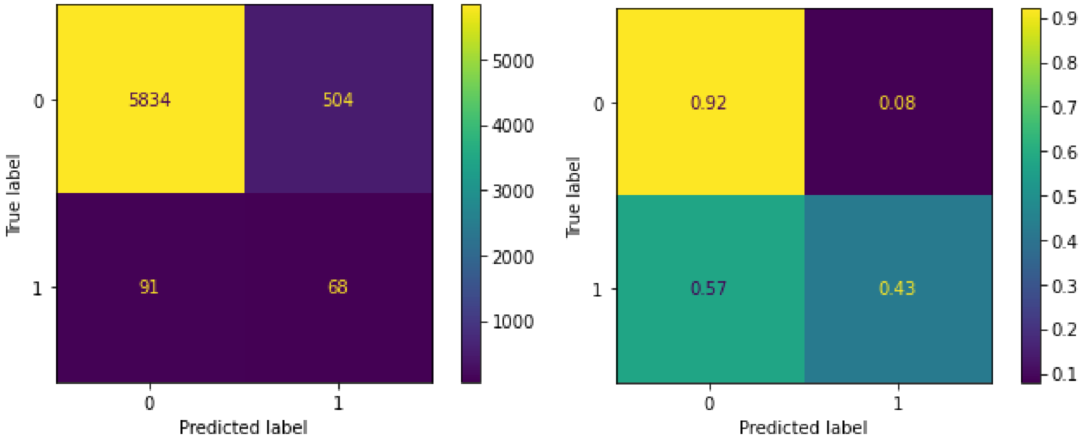
Random Forest - Validation Set (non-normalized and normalized)



XGBoost - Validation Set (non-normalized and normalized)



Light GBM (with Optuna) - Validation Set (non-normalized and normalized)



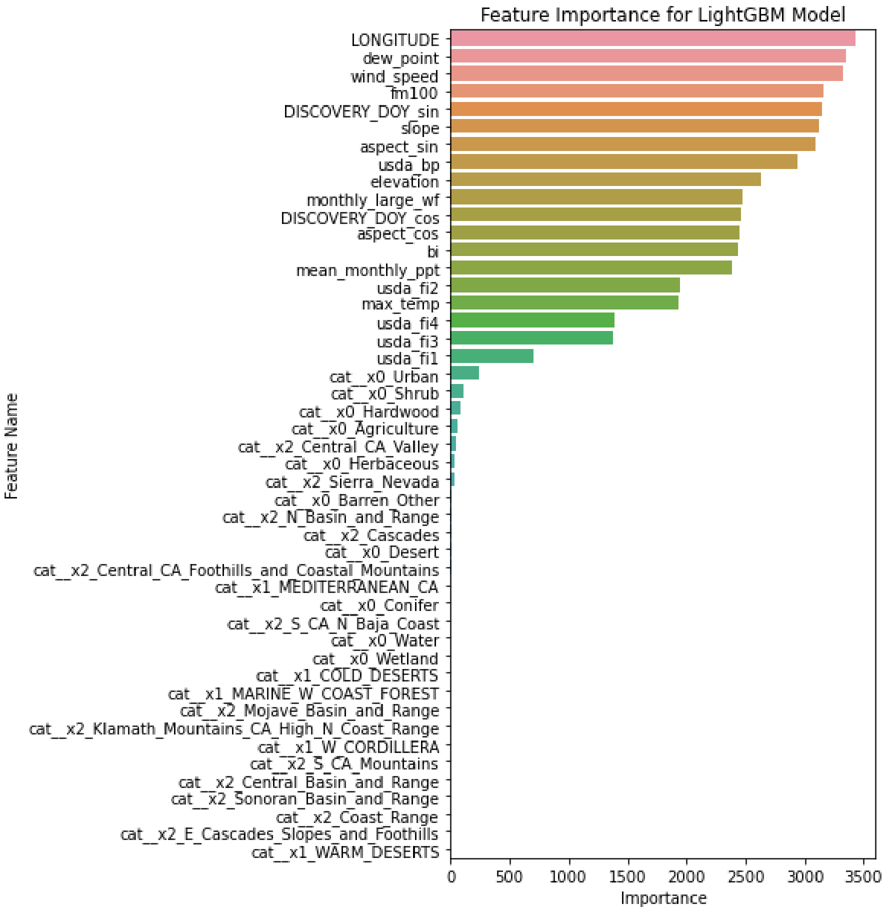
Final Model Performance

Out of the algorithms tested in the previous section, LightGBM without SMOTE, tuned with Optuna, resulted in the highest mean F2 score during cross validation. This is not the model that correctly identified the most large wildfires, it is the model that achieved the best balance between identifying a good amount of large wildfires without misclassifying too many small wildfires as large.

Final Model Performance	**f2**	**recall**	**roc auc**
Cross Validation Mean	0.296	0.457	0.688
Test Set	0.297	0.467	0.691

The chart and table below summarize optimized hyperparameters and the feature importances of the best model:

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



Based on domain knowledge, the feature importances of the final model make conceptual sense. Longtitude is a strong predictor because the westernmost portion of the state is the fire prone counties of Humboldt, Trinity, and Siksiyou and the easternmost portion of the state is mostly desert with little vegetation (inland San Bernardino, Riverside, and Imperial counties). Other strong predictors include dew point, wind speed, 100-hour fuel moisture, discovery day of the year, and aspect. Interestingly, it seems that none of the categorical variables have much predictive power.

Conclusion

Improvements and Next Steps

While the final model does have substantial predictive power, it could certainly be improved. The fact that the advanced gradient boosting models only improved incrementally over the logistic regression model suggests that the greatest opportunities for increasing prediction are improving the input data. There are several ways this could be done:

- The timeframes for the weather features calculated from Google Earth Engine could be revisited. In particular, the timeframe for precipitation (previous year) could be shortened.
- Additional features that address human activity/influence could be added. For example, distance from paved roads or distance from CALFIRE airports.
- The categorical vegetation type and ecoregion datasets did not end up having as strong of predictive power as anticipated. These could be replaced or supplemented with more granular quantitative datasets such as Normalized Difference Vegetation Index (NDVI) (for the days preceding each wildfire), canopy density, fuel load, etc.

The model may also be suffering from not having enough examples of the positive minority class to train on. This could be addressed by using updated data that extends to 2018[17] (rather than 2015). This updated dataset was unfortunately released after the feature extraction phase of this project was complete. Another possibility is to extend the start date back from 2005 to 2000, or as far as 1992. Early on in data wrangling, it was assumed that this was necessary (since wildfire severity has increased due to climate change). However, since many of the weather features are extracted based on date, this may actually not have been a problem. Assuming that the proportion of wildfires >300 acres has stayed roughly constant, expanding the timeframe from 2005-2015 (10 years) to 1992-2018 (26 years) could more than double the number of large wildfires available for the model to train on.

Finally, the class imbalance could also be addressed by reducing the scope of the model. The model could be limited to months in summer and early fall (when large wildfires actually occur) or to wildfire prone areas, rather than the entire state. This might ensure that the training data is more directly relevant to the desired use case for the model.

Using the Model

While this model was evaluated based on a hold-out test set split from a dataset of historic wildfires, the purpose of this model is to make predictions for future wildfires as they occur. Data for all the explanatory features could be extracted for a new wildfire on the fly given location coordinates and date (although as some PRISM and GRIDMET features have a 3 day lag, if the prediction is happening the same day a wildfire is discovered, the model would have to use conditions from 3 days before the wildfire was discovered.

The model could be put into production with a front-end interface where the user could indicate the location of a wildfire on an interactive browser-based map, enter the date, and then receive a prediction. For situations where many wildfires are occurring at once, a spatial file (shapefile, geojson, etc.) or table of wildfire locations could be uploaded in order for the model to make batch predictions. In both cases, the predictive algorithm could be used in tandem with existing deterministic/physical wildfire models.

[17] Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA_FOD_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.5