Predicting the Danger of Particulate Matter Pollution from Wildfires Using Classification Models

William Lai[1]

[1]The Bishop's School

ABSTRACT

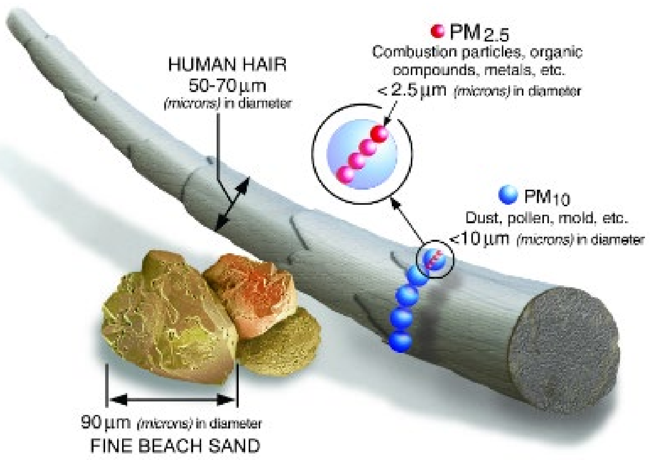
Due to a mix of climate change and California's mega-drought, California's wildfire seasons have overall gotten progressively longer, more destructive, and more expensive. In 2020 alone, around 9,900 wildfires burned about 4.3 million acres, costing the state over $12 billion. (Kerlin, 2022) Larger and more numerous wildfires pollute billions of harmful particles into the atmosphere, including PM2.5. This study aims to use features of wildfire and other factors to predict whether a wildfire pollutes enough PM2.5 particles to be detrimental to human health. The 8 features used in the model are the acres burned, the length in days of the fire, available green space within a 15-mile radius of the fire, the highest population density within a 15-mile radius of the fire, electricity usage, median income, temperature, and precipitation. A Gradient Boosting Classifier (GBC) was applied to the dataset to predict whether a wildfire's emissions necessitated an evacuation. The GBC results achieved a high accuracy of 0.931 as well as a great Area Under the Curve (AUC) of 0.911. By far the most important feature in the GBC is Length, with a feature importance score of 0.109 +/- 0.009.

Introduction

Wildfires are uncontrolled fires that burn vegetation in the wild. They occur in forests, grasslands, savannas, and other ecosystems. There are many types of wildfires, including ground fires, surface fires, and crown fires. Ground fires ignite in soils with a lot of vegetation, like plant roots. Surface fires burn dead or dry plants on the surface. Crown fires burn the leaves and the canopies of trees and bushes. The severity of wildfires depends on the wind, high temperature, little rainfall, and other environmental factors (National Geographic Society, 2019)

Miller, R. (2020, October 29). Climate change is central to California's wildfires. Scientific American. Retrieved June 28, 2022

Not only are there the physical cost of the wildfires, but millions of California residents are also exposed to harmful levels of air pollution from wildfires, with particulate matter (PM) particles being the worst. PM particles are made up of a multitude of tiny particles and liquid droplets. Some examples are acids, organic chemicals, metals, and soil/dust particles. Within PM, there are classifications based on size, with the important ones being PM2.5 (μm) and PM10. PM10 is primarily caused by the combustion of gasoline, oils, etc while PM2.5 is caused mainly by fires. PM smaller than 10 micrometers is detrimental to the health of humans because the particles can pass through our throat and mouth into our lungs. When in the lung, it either goes into our bloodstream affecting other organs or inhibits the function of the lungs and heart. (Environmental Protection Agency, n.d.) This study focuses on PM2.5 particles since the particle matter is so small that it gets trapped deeper in the human body, affecting more critical systems and making them harder to remove. Thus it is imperative to study possible ways to reduce fire damage and protect the health of Californians and their structures.


PM2.5
Combustion particles, organic compounds, metals, etc.
< 2.5 µm (microns) in diameter
HUMAN HAIR
50-70µm
(microns) in diameter
PM10
Dust, pollen, mold, etc.
<10 µm (microns) in diameter
90 µm (microns) in diameter
FINE BEACH SAND

Environmental Protection Agency. (n.d.). EPA. Retrieved June 28, 2022.

The main goal of this study is to predict whether the severity of PM2.5 emissions was high enough that the surrounding population would have to evacuate. Machine learning techniques were used to explore the relationship between PM2.5 emissions and the attributes of wildfires and their surroundings.

Methods

3.1 Features in the model

The 8 features used in the model were the acres burned by the fire, the length in days of the fire, the amount of vegetation within a 15 mile radius of the fire, the maximum population density of cities within 15 mile radius of the fire, the electricity usage of counties, the median income of counties, the temperature of the days before, and the amount of precipitation in the days leading up to the fire. Acres burned, represented in the model as AcresBurned, was selected to be a feature due to the belief that the more acres burned, the more PM2.5 emissions there would be due to a larger amount of land to burn. Length, represented by Length, was a feature because the longer a fire burns, the more PM2.5 it pollutes into the atmosphere. Vegetation, represented by available_green, was selected to be a feature due to the belief that the more plants there are, the more the fire can combust, resulting in more PM2.5 emissions. Population, represented by nearby_maxdensity, was a feature because the more people that are living in a certain area, the more fire stations are needed to support the neighborhood, resulting in more firefighters putting out a potential fire, reducing PM2.5 emissions. Electrical usage, represented by dratio, was selected to be a feature because sparks from power lines are one of the biggest igniters of wildfires. The more electricity used, the higher the probability of starting a wildfire, increasing PM2.5 emissions. Median income, represented by income, was a feature due to the theory that places with a higher income would have a quicker response to a wildfire that threatened the county. Counties with higher median incomes would have more funding for their fire department, potentially resulting in quicker responses and faster extinguishment. Temperature, represented by temp, was selected to be a feature because the higher the temperature, the easier it is for a fire to spread and the harder it is to extinguish it. A larger fire will emit more PM2.5 particles. Precipitation, represented by precip, was a feature because the less rain, the drier the conditions. One of the driving factors of fire intensity is the dry conditions that make fires more destructive and harder to put out. More intense fires will result in more PM2.5 emissions.

A. California Wildfires

The starting point for the main dataset was a dataset off of Kaggle called ‘California wildfires (2013-2020)’. This dataset was chosen because California has one of the most numerous and largest fires. This dataset provided all the fires from 2013-2020 in all counties of California with features such AcresBurned, Date_Started, Date_Ended, longitude, and latitude. By counting the number of days between Date_Started and Date_Ended, the feature Length was created.

B. Filters for Wildfires

Three filters were added to this dataset, with the first filter limiting the length of the fire to be greater than 0 days and less than 100. In the dataset, more than half of the fires had lengths greater than 100 days without large acres burned. Including these extraneous fires would have skewed the data. The 0 day restriction was added because there weren’t any increases in the PM2.5 measurements. After applying the length filter, the number of fires went from a total of 1636 to 728 fires.

The second filter was getting rid of fires under 50 acres. Looking at the PM2.5 data in the timeframe of the fire, there were no noticeable changes in the air quality measurements. Including these fires would have lessened the potential positive correlations between acres burned and the concentration of PM2.5 particles. This filter reduced the number of fires from 728 to 523.

The last filter was removing certain counties. When getting the air quality data, some counties (Alpine, Amador, Lassen, Modoc, Sierra, Tuolumne, and Yuba) didn’t have any data at all. Including these fires would have made it impossible to calculate any correlation between acres burned and air quality. Adding this filter decreased the number of fires from 523 to 474 fires.

C. PM2.5

From California Environmental Protection Agency Air Resources Board’s Air Quality Query tool, PM2.5 measurements were taken between the date started and ended for all fires. Three different ways to quantify the PM2.5 data into a single number were calculated from those measurements. The first way was PM2.5_integrated, or PM2.5_int. PM2.5 integrated is just the sum of PM2.5 measurements subtracted by the median value of the year. This was done to increase the disparity between small and large fires, making it easier to predict the severity of PM2.5 emissions. The mean wasn’t selected because if the change in air quality from the wildfires were massive, the baseline value would be skewed to a higher value. With median, even if there were a lot of outliers, there would be no skewed data. The second way was PM2.5_max. This is just the maximum concentration of PM2.5 within the timeframe of the wildfire. The larger the fire, the higher the PM2.5_max value. The third way was PM2.5_1day. This is just the PM2.5 measurement one day after the start of the fire. This was considered because the higher the PM2.5 measurement is at the start, the higher the chance that this will become a large fire that pollutes more PM2.5 particles. To represent

PM2.5 emissions in binary, the feature to be predicted, dangerous, was created. Humans that are exposed to PM2.5 levels above 50 µg/m^3 for long periods are at a heightened risk for cardiovascular disease. We define the target dangerous to have a value of 0 for fires that have a PM2.5_max value from 0 up to 50 µg/m^3 (i.e., are not dangerous) while a 1 represents all fires with a PM2.5_max value of greater than 50 µg/m^3 (i.e., are dangerous).

D. *Green Space and Population*

The vegetation data and population data used in the model were from the USDA Forest Service's Urban Forest Data. Since most of the fires didn't start within cities, the vegetation (nearby_greenspace) and population (nearby_pop) data attached to each fire were the sums of the data from all cities within a 15 mile radius from the longitude and latitude of each fire. A 15 mile radius was chosen because this was the potential amount of vegetation that could be burned as well as the population that would be in immediate danger. Since the USDA dataset didn't have longitude and latitude for the cities, CDPs, and towns, another dataset was needed. The dataset used was one from simplemaps of cities in the US. The final cities, CDPs, and towns considered were those that were in both USDA and simplemap datasets.

The feature available_green was chosen over nearby_green because including areas that couldn't be burned would add unnecessary noise to the dataset. An alternate feature representing population is nearby_maxdensity. It was calculated by finding the maximum density out of all the nearby cities to the fire. This was included in the model instead of nearby_pop because the model had a higher Area Under the Curve (AUC) with nearby_maxdensity.

E. *Electricity Usage*

Electricity Usage wasn't initially considered to be a feature of the model, but after reading through the paper "Data-driven wildfire risk prediction in Northern California", it was added as a feature. Electricity usage data used was from a dataset from Kaggle called California- electricity consumption by county. The variable used to quantify the data into one number is called dratio. Dratio is the difference between the electricity usage from 2019 subtracted by the usage from 1999 divided by the usage from 2019. For each fire, the feature dratio was added by the county in which the fire was located.

F. *Median Income*

The median income dataset used in the model was taken from KidsData. Each value represents the average median income for each county from 2014 to 2018. For each fire, the feature income was added by the county in which the fire was located and the year the fire occurred.

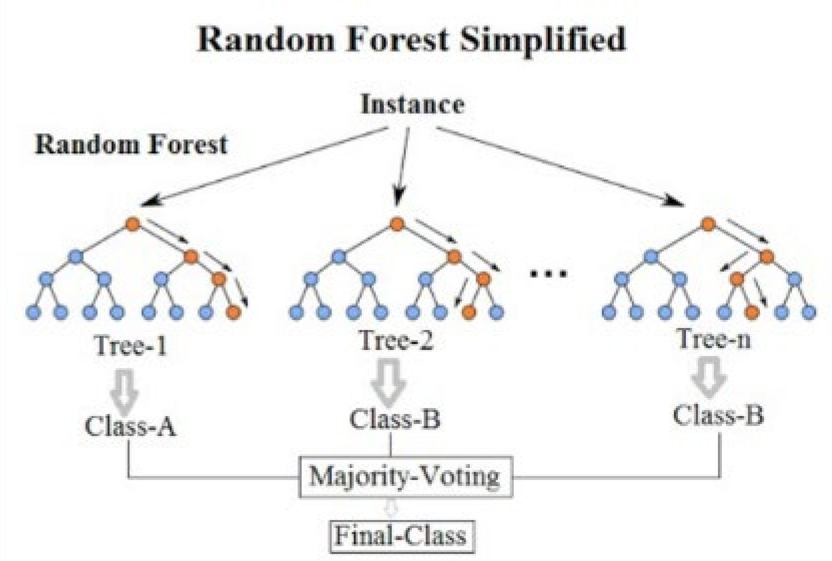
G. *Temperature and Precipitation*

The temperature and precipitation datasets were from the NOAA's National Centers for Environmental Information. The dataset contained the monthly averages of temperature and precipitation for each county of all 50 states from 1895 to 2022. Only the data from Californian counties from the years 2013-2022 were taken. Since the temperature before the fire determines fire creation, the initial plan was to get the average temperature 30 days before the start date. Since the dataset only contained monthly averages, fires that occurred before the 15th day took the previous month's temperature and precipitation averages while the fires that occurred on or after the 15th took the current month's temperature and precipitation averages. The temperature and precipitation values were added to each fire based on the county and start date.

3.2 Algorithm

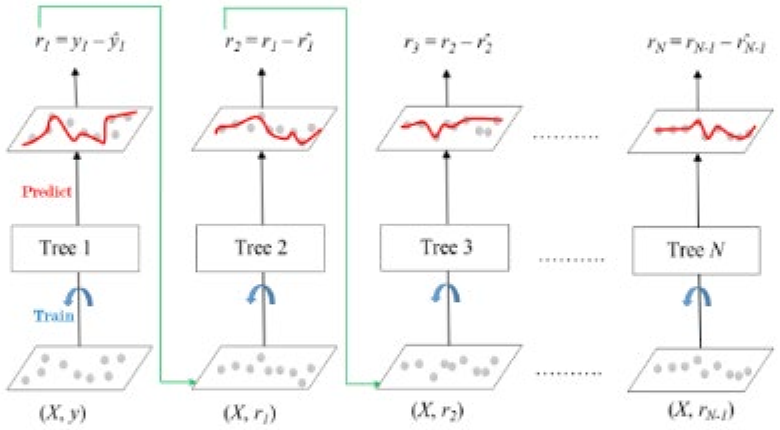
Ensemble methods

In machine learning, ensemble methods are a technique used to balance bias and variance. Bias is the difference between the actual value and the value predicted by the model. Variance represents a model's sensitivity to changes in data. A model with low bias usually has high variance while a model with high bias usually has low variance. Ensemble methods combine multiple models to better balance bias and variance. Two ways of combining multiple models are bagging and boosting. (Chong, J. 2021)



Wikimedia Foundation. (2022, June 20). *Random Forest*. Wikipedia. Retrieved June 28, 2022

The bagging method creates several independent parallel decision trees[1] that use different subsets of the training data. Once it creates the trees, it uses the mode of the predictions from all the trees to determine a final result, which reduces error. An example of the bagging method is a Random Forest Model (RFM). RFMs should be used when looking for the significance of predictors, a quick benchmark model, and with imperfect data. (Chong, J. 2021).



nikki2398@nikki2398. (2020, September 2). *ML - gradient boosting*. GeeksforGeeks.

[1] A decision tree is like a flow chart; e.g., to classify fruit it might ask is the fruit round? Then, is it red? etc.

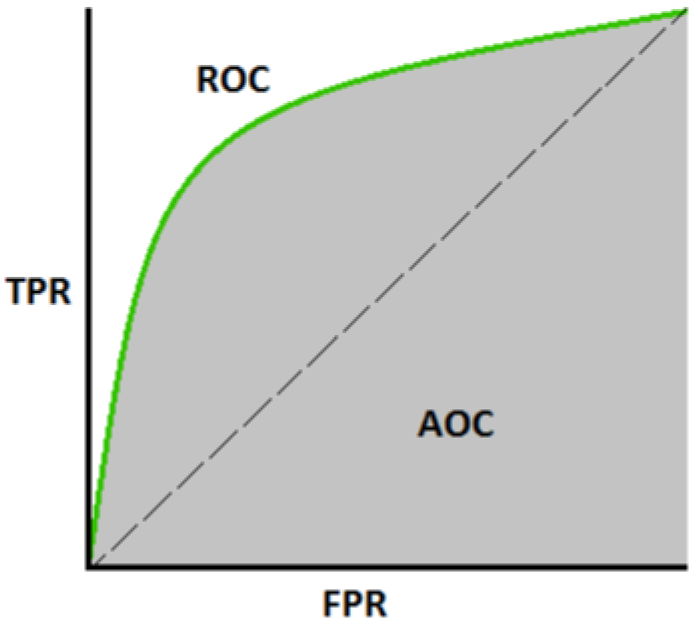
The boosting method creates multiple sequential models that focus on the mistakes of the previous one and improves upon it to create a better model and reduce error. The key differences between boosting and bagging is that boosting methods aren't parallel but are sequential and aren't independent but fully dependent. An example of a boosting method is a Gradient Boosting Model (GBM). Like a RFM, GBMs should be used when looking for the significance of predictors. But unlike RFMs, GBMs should be used when prediction time is important because building sequential trees is time-consuming. (Chong, J. 2021)

A. *Model tuning and performance metrics*

For both algorithms, hyperparameters were adjusted to reduce overfitting. With a dataset of only 474 rows, any ensemble method would just overfit with default hyperparameters. To reduce overfitting, the hyperparameters of n_estimators, max_depth, and min_sample_leafs were adjusted. N_estimators is the number of boosting stages in the model. Normally more boosting stages result in better performance. However, to limit overfitting, a smaller number of stages was needed. To balance the number of estimators and performance, 20 estimators were used. Max_depth limits the number of nodes per tree. Since the best performance resulting from max_depth depends on how each input variable interacts with each other, different values were tested. A max_depth of 5 was chosen due to a high AUC. Min_sample_leafs are the minimum number of samples per leaf node. The higher the min_sample_leafs, the smoother the model is, or the more extreme values are removed. Since most of the extreme values were important, a smaller min_sample_leafs would benefit the model, resulting in a value of 5.

Accuracy is the percentage of correct classifications by the model out of all cases, i.e. True Positives plus True Negatives divided by the sum of True Positives, True Negatives, False Positives, and False Negatives. Area Under the Curve (AUC) is a way to measure the performance of a classification model on all possible classification thresholds. It is the probability of choosing a random positive example over a random negative example. The higher the AUC, the better it is at distinguishing between the two. (Bhandari, A. 2020).

The classification model provides a probability that a fire is dangerous, which is converted into a yes/no classification based on a threshold that we allow the algorithm to determine automatically. In deciding which model to use (Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC)), simplicity and performance were the key factors. The GBC was selected because it had a higher AUC out of the two.



Narkhede, S. (2021, June 15). Understanding AUC - roc curve. Medium

Results

4.1 Performance and Feature Importance

Figure 1.

Accuracy and AUC of the GBC on Training and Test data

	Accuracy	Area Under the Curve
Training	0.993	1.000
Test	0.931	0.911

After applying the GBC to the training data, its accuracy was 0.993 and its AUC was 1.0. When the model was applied to the test data, its accuracy was 0.931 and its AUC was 0.911. (Figure 1) This means that the model is slightly overfitting, resulting in slightly biased predictions.

Figure 2.

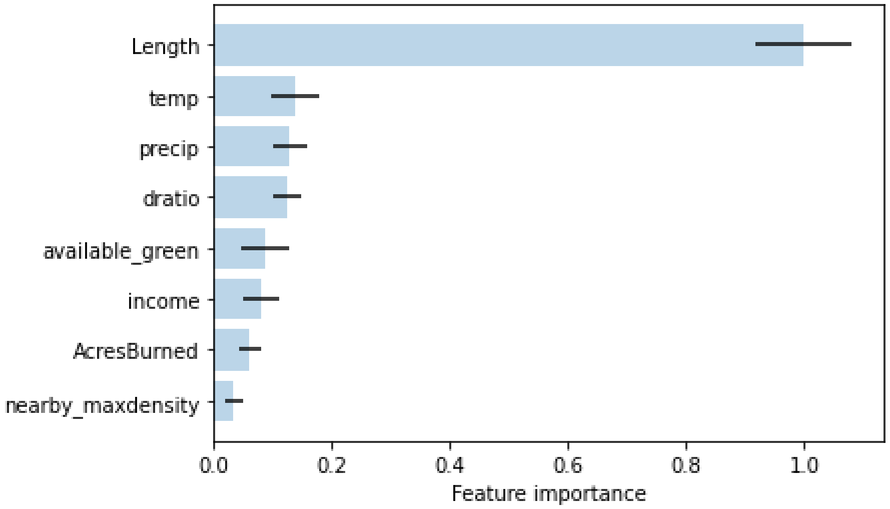
Confusion Matrix for GBC on Test Data

True Positive *Correctly classified dangerous* 11 5.851%	**False Positive** *Incorrectly classified dangerous* 6 3.191%
False Negative *Incorrectly classified not dangerous* 7 3.723%	**True Negative** *Correctly classified not dangerous* 164 87.234%

After applying the GBC to the test data, there are 164 true negatives, 7 false negatives, 6 false positives, and 11 true positives. This means that the model isn't leaning toward either false negatives or false positives, resulting in a good balance between safety and risk-taking.

Figure 3.

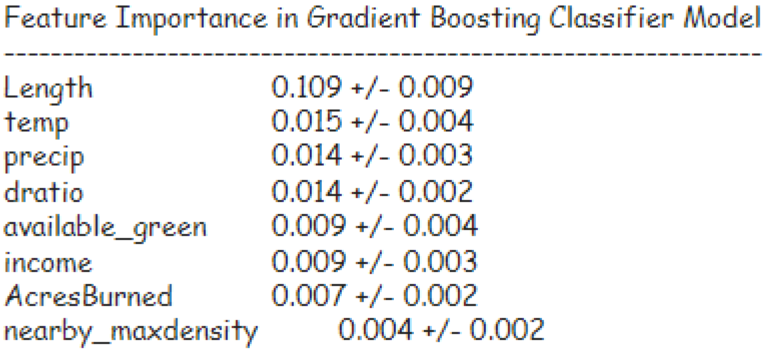
Bar Graph of Relative Feature Importance from GBC



Compared to the rest of the features, Length is the most important by far. (Figure 3)

Figure 4.

Statistically Important features in GBC

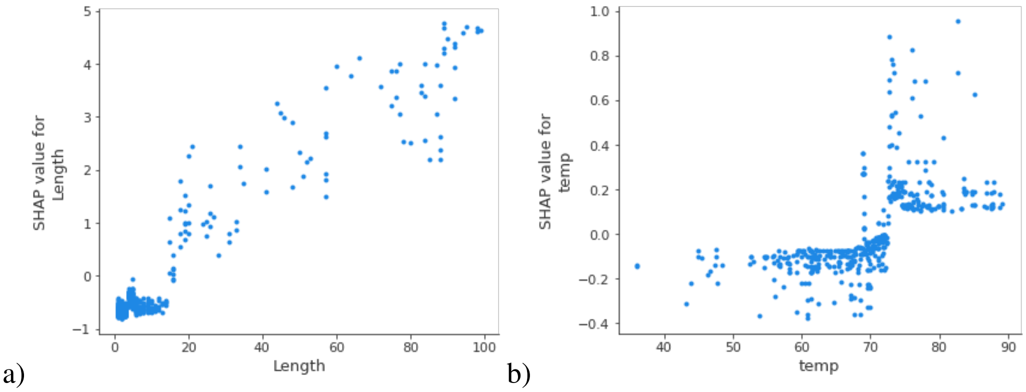


```
Feature Importance in Gradient Boosting Classifier Model
------------------------------------------------------------------
Length                 0.109 +/- 0.009
temp                   0.015 +/- 0.004
precip                 0.014 +/- 0.003
dratio                 0.014 +/- 0.002
available_green        0.009 +/- 0.004
income                 0.009 +/- 0.003
AcresBurned            0.007 +/- 0.002
nearby_maxdensity           0.004 +/- 0.002
```

In calculating the importance of each feature, all features are statistically significant, with all means being greater than twice the standard deviation. (Figure 4) These scores also support Figure 3, with Length's score of 0.109 +/- 0.009 being significantly greater than temp's score of 0.015 +/- 0.004. (Figure 4)

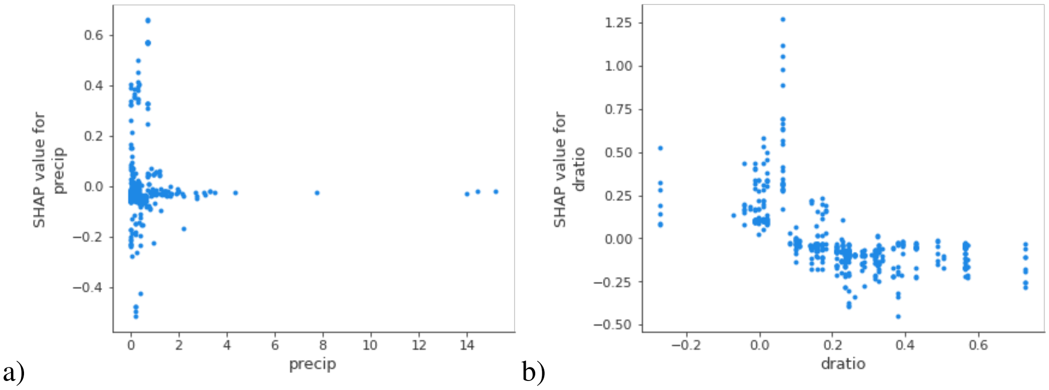
4.2 Partial Dependence Plots

Figure 5.
Partial Dependence plots of Length (a) and temp (b)



The partial dependence plot for *Length* (a) supports the positive correlation between length of the fire and PM2.5 emissions. Supporting the feature importance graph, the partial dependence plot of temp (b) shows that the model predicts there is an overall positive relationship between temperature and PM2.5 emissions.

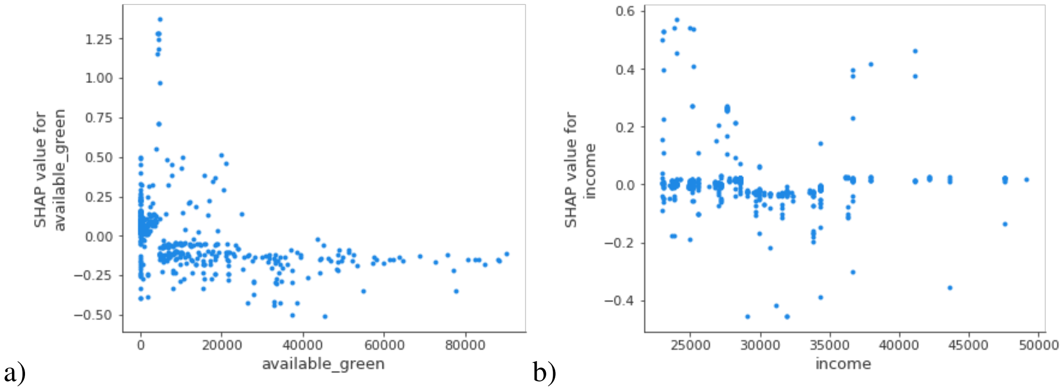
Figure 6.
Partial Dependence Plots of precip (a) and dratio (b)



The partial dependence plot of precip (a) shows an overall negative relationship between precipitation and PM2.5 emissions. Similarly the partial dependence plot for dratio (b) shows an overall negative relationship between electricity usage and PM.5 emissions.

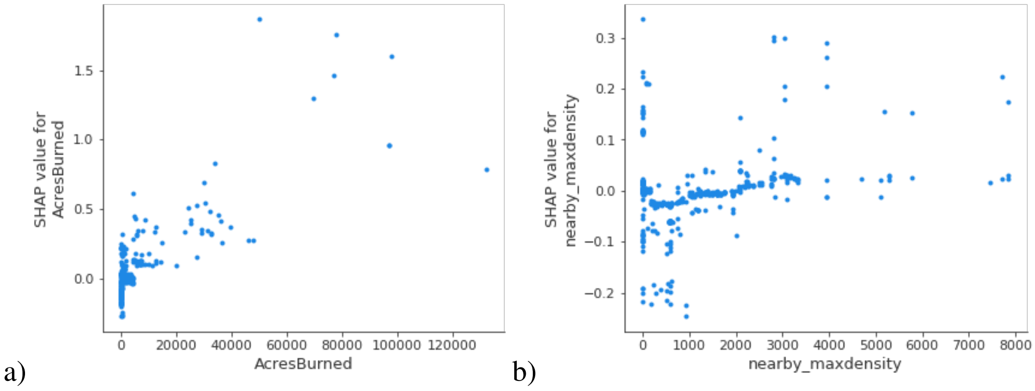
Figure 7.

Partial Dependence Plots of available_green, and income



In addition to the feature importance graph showing available_green isn't particularly important, the partial dependence plot (a) shows that the model predicts that there is no strong pattern in the relationship between the hectares of available green space and PM2.5 emissions. This is also the case with the partial dependence plot of income (b) where there is no strong pattern.

Figure 8.

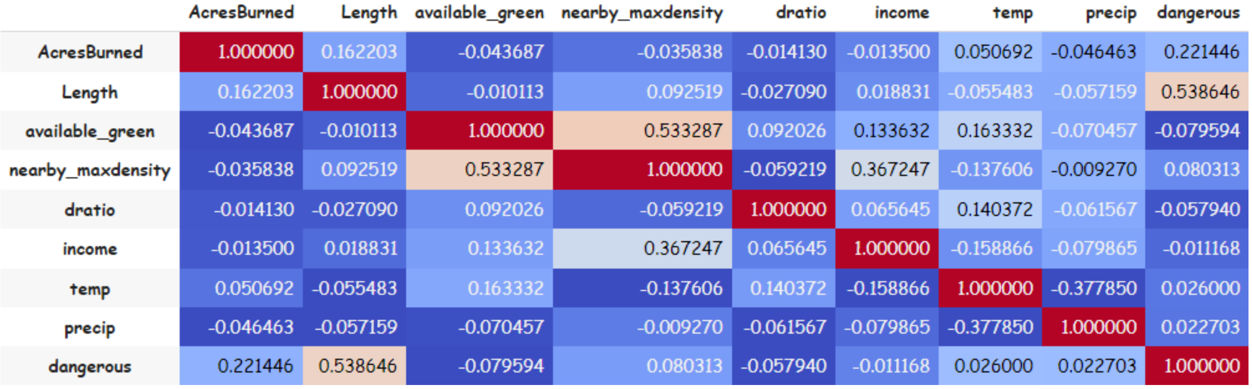


Even though AcresBurned didn't have a high feature importance score, the partial dependence plot (a) shows that the model predicts a positive relationship between the acres burned and PM2.5 emissions. The partial dependence plot for nearby_maxdensity (b) shows that the model predicts there is an overall slight positive relationship. Partial Dependence Plots of AcresBurned and nearby_maxdensity

4.3 Correlations

Figure 9.

Correlation between Features Heat Map using a Linear Relationship

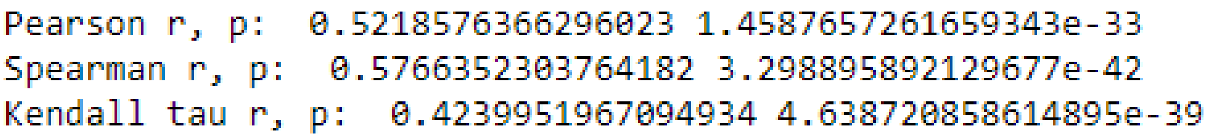


	AcresBurned	Length	available_green	nearby_maxdensity	dratio	income	temp	precip	dangerous
AcresBurned	1.000000	0.162203	-0.043687	-0.035838	-0.014130	-0.013500	0.050692	-0.046463	0.221446
Length	0.162203	1.000000	-0.010113	0.092519	-0.027090	0.018831	-0.055483	-0.057159	0.538646
available_green	-0.043687	-0.010113	1.000000	0.533287	0.092026	0.133632	0.163332	-0.070457	-0.079594
nearby_maxdensity	-0.035838	0.092519	0.533287	1.000000	-0.059219	0.367247	-0.137606	-0.009270	0.080313
dratio	-0.014130	-0.027090	0.092026	-0.059219	1.000000	0.065645	0.140372	-0.061567	-0.057940
income	-0.013500	0.018831	0.133632	0.367247	0.065645	1.000000	-0.158866	-0.079865	-0.011168
temp	0.050692	-0.055483	0.163332	-0.137606	0.140372	-0.158866	1.000000	-0.377850	0.026000
precip	-0.046463	-0.057159	-0.070457	-0.009270	-0.061567	-0.079865	-0.377850	1.000000	0.022703
dangerous	0.221446	0.538646	-0.079594	0.080313	-0.057940	-0.011168	0.026000	0.022703	1.000000

In addition to the GBC, linear relationships between independent variables shows that Length correlates with dangerous significantly more than the other features. (Figure 8) Other notable correlations are between nearby_maxdensity and available_green, as well as nearby_maxdensity and income. These correlations are relatively strong, with values of 0.533 and 0.367 respectively. They make sense, as large cities would have a high density, as well as a lot of green space within and around it, and cities with higher densities, usually have higher median incomes.

Figure 10.

Pearson, Spearman, and Kendall tau Correlation Coefficient for Length vs PM2.5 max.



```
Pearson r, p:  0.5218576366296023 1.4587657261659343e-33
Spearman r, p:  0.5766352303764182 3.298895892129677e-42
Kendall tau r, p:  0.4239951967094934 4.638720858614895e-39
```

With the Pearson Correlation Coefficient being 0.522 (Figure 10), there is a significant positive correlation between Length and the severity of PM2.5 emissions. With the Spearman Correlation Coefficient being 0.577 (Figure 10), there is a positive association between the two. The Kendall tau Correlation Coefficient of 0.424 (Figure 10) supports the positive association found with the Spearman Correlation Coefficient. All p values are near 0 indicating that the correlations are highly significant.

Discussion

It was hypothesized that AcresBurned, Length, available_green, nearby_maxdensity, dratio, income, temp, and precip would predict whether the level of PM2.5 emissions would jeopardize the health of humans. The results of a GBC provide an overall high predictive power, with an accuracy of 0.931 and an AUC of 0.911.

(Figure 1) This result is plausible since all the features included in the model have a logical connection to wildfire creation.

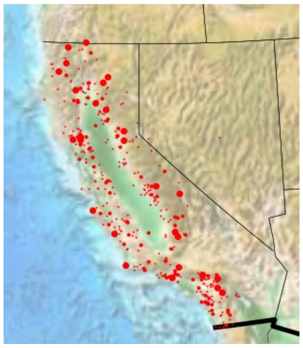
The most important feature in the model is Length with a feature importance value of 0.109 +/- ±0.009 (Figure 4). Not only was Length important in the GBC, but it was also important with linear relationships, as it had a correlation value of 0.538. (Figure 10) This makes sense because the longer a fire burns, the higher amounts of PM2.5 it will pollute. While the feature importance scores of the other features weren't as high as Length, all the features had statistically significant scores, making them of secondary importance but still helpful in predicting the severity of PM2.5 emissions. The partial dependence plots support this conclusion, as only the plot for Length shows a strong relationship.

Several limitations of the datasets used may explain why some of the features that were expected to be strongly predictive did not have high relative importance in the model. The only dataset available for green space was those within city limits of California cities, which is not necessarily the amount of vegetation at the site of the fire. The electricity usage dataset used to calculate dratio only had yearly averages, which isn't ideal since electricity usage varies by month, and this variation could explain why dratio doesn't have a high feature importance score. Temperature and precipitation datasets also suffer from a similar limitation because instead of having daily values, the dataset used only had monthly averages. The way the monthly temperature and precipitation averages were included may have addressed this limitation to a certain degree, but there still is plenty of noise within the dataset. The PM2.5 measurements weren't necessarily accurate because of the distance between air quality reading stations and the location of the fire. Environmental features could have introduced additional noise into the data, such as the wind blowing the smoke away from the stations, producing an inaccurate reading.

Conclusion

This paper addresses the problem of predicting whether the severity of PM2.5 emissions after a wildfire is detrimental to human health. The hypothesis is that acres burned, length in days, green space near the fire, population density close to the fire, electricity usage, income, temperature, and precipitation would be crucial to predicting PM2.5 emissions. The results validate the hypothesis. The GBC had high accuracy as well as a decent AUC. Out of all the features, length had the highest feature importance. The lower feature importances for the other features could be from the noise in the dataset.

A continuation of this research project would be to get more precise datasets and add new features. Instead of relying on air quality reading stations not near fires, collecting PM 2.5 measurements at the actual fire site would be more precise. If possible, determining the wind speed would be a feature that could potentially increase the accuracy and AUC of the model. After plotting the fires on a topographical map of California, the fires all seemed to be located in the forests of the mountains and not in the farmland of Central Valley. This raises the potential of having land type as a new feature in a future model.



As mentioned before, the main goal of this study was to predict how dangerous a wildfire is based on PM2.5 emissions. The most important feature in the model is Length, which in real life is not an ideal way to predict if the PM2.5 threshold will be exceeded because length is not a known value until after the fire. One remedy would be to change what Length represents. Instead of being the duration of the fire in days, it perhaps could be the number of days it takes to exceed the PM2.5 emission threshold, or the length of the fire so far.

References

Air Quality Data (PST) query tool. California Environmental Protection Agency Air Resources Board. (n.d.). Retrieved February 28, 2022, from https://www.arb.ca.gov/aqmis2/aqdselect.php

Ares. (2020, February 9). *California wildfires (2013-2020)*. Kaggle. Retrieved February 28, 2022, from https://www.kaggle.com/ananthu017/california-wildfire-incidents-20132020

Bhandari, A. (2020, June 16). *AUC-Roc Curve in machine learning clearly explained*. Analytics Vidhya. (2022, June 14). Retrieved June 28, 2022, from https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#:~:text=The%20Area%20Under%20the%20Curve,the%20positive%20and%20negative%20classes

Brownlee, J. (2020, August 14). *Linear regression for machine learning*. Machine Learning Mastery. Retrieved March 4, 2022, from https://machinelearningmastery.com/linear-regression-for-machine-learning/#:~:text=Linear%20regression%20is%20a%20linear,the%20input%20variables%20(x).

Chong, J. (2021, August 29). *Battle of the ensemble-random forest vs gradient boosting*. Medium. Retrieved March 3, 2022, from https://towardsdatascience.com/battle-of-the-ensemble-random-forest-vs-gradient-boosting-6fbfed14cb7

Environmental Protection Agency. (n.d.). EPA. Retrieved June 28, 2022, from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

Environmental Protection Agency. (n.d.). *What is particulate matter? | urban environmental program in New England*. EPA. Retrieved June 28, 2022, from https://www3.epa.gov/region1/eco/uep/particulatematter.html#:~:text=%22Particulate%20matter%2C%22%20also%20known,and%20soil%20or%20dust%20particles

Google. (n.d.). *Classification: Roc curve and AUC | machine learning crash course | google developers*. Google. Retrieved June 28, 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Kambhampati, D. D. (2020, October 5). *California- electricity consumption by county*. Kaggle. Retrieved March 2, 2022, from https://www.kaggle.com/devkambhampati/california-electricity-consumption-by-county/version/1

Kerlin, K. E. (2022, May 4). *California's 2020 wildfire season*. UC Davis. Retrieved June 25, 2022, from https://www.ucdavis.edu/climate/news/californias-2020-wildfire-season-numbers

Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A. K., Liu, Q., Chiao, S., & Gao, J. (2021, January 13). *Data-driven wildfire risk prediction in Northern California*. MDPI. Retrieved February 27, 2022, from https://www.mdpi.com/2073-4433/12/1/109/htm

Median family income, by family type (regions of 10,000 residents or more). Kidsdata.org. (n.d.). Retrieved March 2, 2022, from https://www.kidsdata.org/topic/545/income-family-type-10k/table

Miller, R. (2020, October 29). *Climate change is central to California's wildfires*. Scientific American. Retrieved June 28, 2022, from https://www.scientificamerican.com/article/climate-change-is-central-to-californias-wildfires/

Narkhede, S. (2021, June 15). *Understanding AUC - roc curve*. Medium. Retrieved July 23, 2022, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

National Geographic Society. (2019, July 15). *Wildfires*. National Geographic Society. Retrieved March 2, 2022, from https://www.nationalgeographic.org/encyclopedia/wildfires/

nikki2398@nikki2398. (2020, September 2). *ML - gradient boosting*. GeeksforGeeks. Retrieved June 28, 2022, from https://www.geeksforgeeks.org/ml-gradient-boosting/

Sklearn.ensemble.gradientboostingclassifier. scikit. (n.d.). Retrieved June 28, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

Sklearn.ensemble.gradientboostingregressor. scikit. (n.d.). Retrieved February 27, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

Sklearn.ensemble.randomforestregressor. scikit. (n.d.). Retrieved February 27, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Sklearn.ensemble.gradientboostingclassifier. scikit. (n.d.). Retrieved June 28, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

United States Cities Database. simplemaps. (n.d.). Retrieved March 2, 2022, from https://simplemaps.com/data/us-cities

Urban Forest Data for California. State Urban Forest Data: California. (n.d.). Retrieved February 28, 2022, from https://www.nrs.fs.fed.us/data/urban/state/?state=CA

Wikimedia Foundation. (2022, June 20). *Random Forest*. Wikipedia. Retrieved June 28, 2022, from https://en.wikipedia.org/wiki/Random_forest