ST309 Group Project

Machine Learning for County-Level Wildfire Risk Prediction in California Through Optimised Classification Models

Table of Contents

1. Introduction

In California, the 9 largest wildfires on record have occurred in the last 7 years. The destruction of the LA fires in January 2025 has underscored the critical need for prediction to minimise wildfire impacts. Evidence suggests that human-caused climate change has driven a 172% increase in area burned compared to just natural forces in California, with a 320% increase from 1996 to 2021 alone (Turco et al., 2023). Several attempts have been made to use Machine Learning to inform fire detection and prediction systems that could mitigate these escalating damages (Walters, 2022).

Wang et al. (2021) acknowledge that Machine Learning has enabled "significant improvements in wildfire predictions", highlighting the importance of data analysis in reducing wildfire impact. However, they caution that the "limited interpretability" of these models has inhibited advances in the understanding of wildfires. Previous ML-based wildfire models have limitations that reduce their policy utility:

1. Lack of local granularity or interpretability: Models often predict on a national, state, or grid level, limiting utility for local-decision makers who "ensure constant monitoring of the natural landscape" (Moghim and Mehrabi, 2024).
2. Fire size and danger interpretability: Typically, models classify wildfires as binary ("Fire" vs "No Fire") or employ a regression model that outputs a continuous, larger range of acreage with a lower uncertainty due to imbalance and high variability rather than providing an informative threshold of acreage burnt, which would be useful for policy-decisions. (Walters, 2022)
3. Forecasting timeframes: Lack an informative temporal scale that makes it difficult for developing a fire management and prevention system.

To address these problems, we compare machine learning techniques to predict the acres burned in the next month by county, and across counties, in California. Addressing these problems, the local governance is dealt with as we divide our prediction by county level which can help resource allocation. We classify the fires into "Small" and "Large" by the log transformed median fire size that helps to suggest which months may be of higher risk and address the imbalance toward smaller fires in the data. We predict on a monthly level rather than a daily forecast as this is noisier. While this model design may lose predictive power compared to modelling at a California, "Fire" vs "No Fire", or daily level, it proves more interpretable for resource allocation.

Contrary to the wildfire prediction literature, we opted not to use the "CALFIRE" dataset in favour of a dataset originally set up to support the National Fire Program Analysis (FPA) system (CAL FIRE, 2024), (Tatman, 2019). As well as having more data points, our dataset provides information related to fire causes. However, previous analysis of the FPA dataset had only been used to predict the cause of wildfires as it lacked relevant features to predict fire size or location. To predict these variables, we found it necessary to link this dataset to the drivers of wildfires: temperature and dryness (Turco et al., 2023). We incorporated each county's rainfall, maximum temperature, and minimum temperature into our analysis from the Daymet dataset (ORNL DAAC, 2024). Krause et al. (2013) critique the wildfire literature for focusing on temperature and precipitation, keeping other factors like ignition agents constant (such as lightning where a small number of fires are responsible for most of the burned area). Our combined dataset enables us to address this issue.

Through exploratory analysis we find that California is highest in number of fires and second highest in total acres burned, reaffirming our decision to focus on the state. Following our EDA, we address wildfire risk through comparing results of various ML classification models. We find that the Maximum Temperature is the most important feature, but our analysis finds little evidence of fire size based on the cause.

The goal of this paper is to attain the optimal balance between data interpretability and prediction. As a result, we also develop an adaptable optimisation between accuracy and catching larger wildfires. We measure the effectiveness of this through Recall, informing resource allocation based on a county's probability of larger wildfires in a month.

2. Data Description and Preparation

To formulate a policy-informing model for driving features of predicting wildfires by month in California, we integrate two datasets: the Fire Perimeter and Attributes Dataset and NASA's Daymet V4 climate dataset.

The FPA dataset holds 1.88 million records of fires between 1992 and 2015, and contains information on acres burned, cause, discovery date and time, containment date and time, and fire county and coordinates. To efficiently load it in, we split the dataset into 6 CSVs to abide by the Moodle 100MB upload limit. Investigation of the California data found that only 80,000 out of 180,000 had county information. To solve this, we assigned missing counties using spatial join with data from US data.gov website to GeoJSON format using the sf library and converted them to the WGS84 coordinate system. The bounding box in the sf package specifies the minimum and maximum coordinates that encompass all the points in each spatial dataset. Employing this method means we now only lose 0.16% of the wildfire data when sorting by county. Similarly, to create the fire density heatmap of Yosemite in section 3.4, we converted the park's tract and boundary data using the GeoJSON format. The dplyr package was used for data filtering to only include fires within Yosemite. Lastly, the heat map fire density layer was created using ggplot2's stat_density2d_filled() function.

Additionally, the FPA dataset recorded dates in Julian format which required conversion to Gregorian calendar to be interpretable. We set the day count to 0 and then counted the days in the respective years to assort the data into the Gregorian calendar, extracting a discover month column so we could analyse seasonality. Furthermore, to make the data more interpretable, we aggregated the total fire size grouping by the county, year and month. Given the large amount of missing Containment Time data, we chose not to consider this as a feature to preserve predictive power. Other than time, investigation reveals that the other columns do not have missing values.

The Daymet dataset contains coordinate based daily climate summaries for the whole of North America from 1950, which we sorted to match the FPA data. Given that the raw data exceeds 4GB, we pre-processed and aggregated the data prior to uploading it to Moodle. Using the coordinates of each recording, we repeated the GeoJSON process to create county-aggregated weather data for California and its 58 counties. We created monthly variables for Total Precipitation, Average Maximum Temperature, and Average Minimum Temperature. This merged dataset captures the high resolution of the daily weather records to enhance the model prediction. Due to the high imbalance of the data towards small fires and anomalous large fires, we log transformed our data to reduce outlier influence and ensure a less skewed distribution to increase classification effectiveness – further analysis on this is in Figure 6. Further analysis of this imbalance is seen in the data explanatory section.

The features used in Section 4 are consistent across our prediction models for easier performance comparison. The models make predictions for next month using:

- Total precipitation (current)
- Maximum temperature (current)
- Minimum temperature (current)
- Rolling mean precipitation (moving average using current and previous 2 months)
- Rolling mean maximum temperature (moving average using current and previous 2 months)
- County average fires (uses mean of log-transformed acres burned for each county)
- Monthly Historical Average (uses mean of log-transformed acres burned for each month)

3. Exploratory Data Analysis

3.1 U.S. Analysis

Figure 1: Number of Fires and Total Acres Burned by State

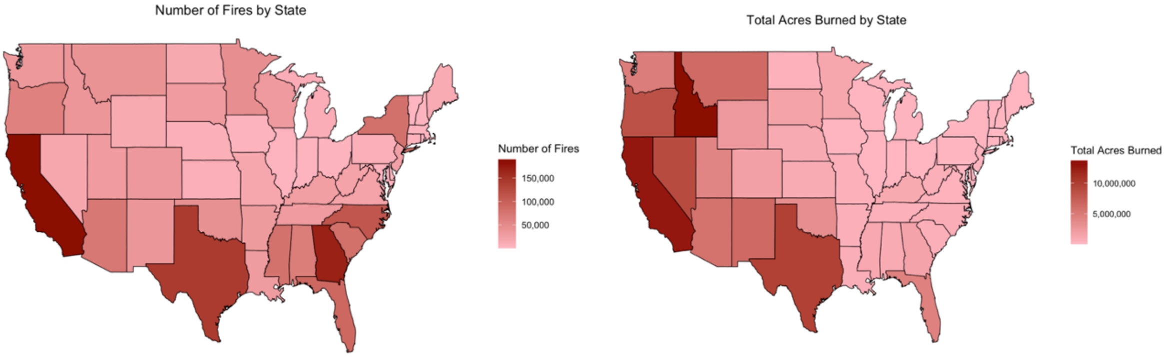
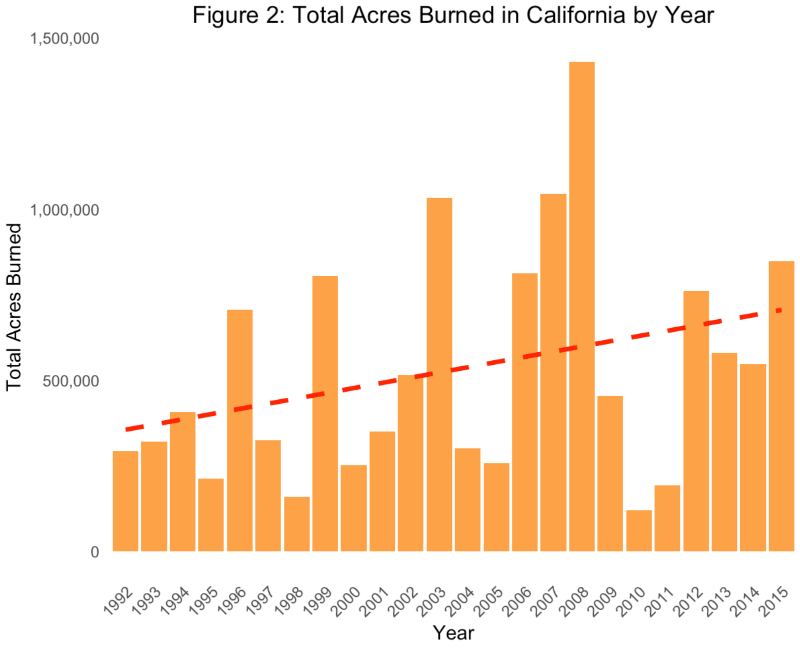


Figure 1 shows the number of wildfires and total acres burned across the U.S. from 1992-2015. Notably, California has the highest number of fires and the second highest in total acres burned. This highlights its susceptibility to large-scale wildfires and the critical need for effective fire management and mitigation strategies. The L.A. fires in January 2025 estimated to have cost up to $164 billion, with projections of broader economic slowdown resulting in a total potential $275 billion loss to the U.S. economy, making it potentially the costliest natural disaster in U.S. history (Li and Yu, 2025). This is why California was chosen as the primary focus of this project. Predicting wildfires at a more localised level within California provides greater practical relevance for wildfire management.

3.2 California Analysis



From Figure 2, we can see an upward trend in total acres burnt, confirming the relevance of our study and the evidence on the increasing damage of wildfires over time (Hantson et al., 2014). Most prominently, the years 2003, 2007, and 2008 experienced over one million acres burned. Five out of the last 9 years since 2015 have also surpassed 1 million acres burned, with 2020 reaching 4.4 million, putting in context the sudden severity of wildfires in the state.

Figure 3: Total Acres Burned in California per Discovery Month

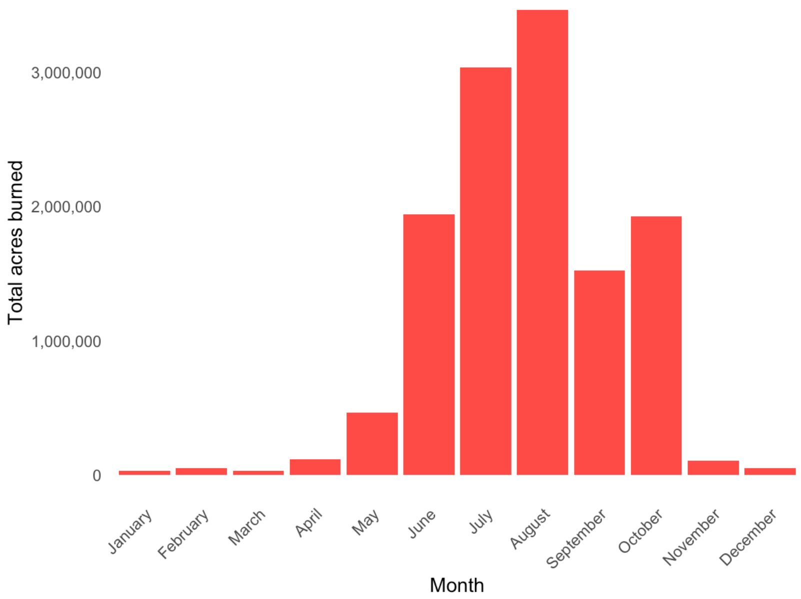


Figure 3 highlights a clear seasonal pattern, with wildfire activity peaking during the summer months, correlating with average monthly temperatures. This seasonal variation highlights the importance of temperature and climate conditions in wildfire occurrence, supporting their inclusion in our model for targeted prevention.

Figure 4: Fire Incidents in California by Cause

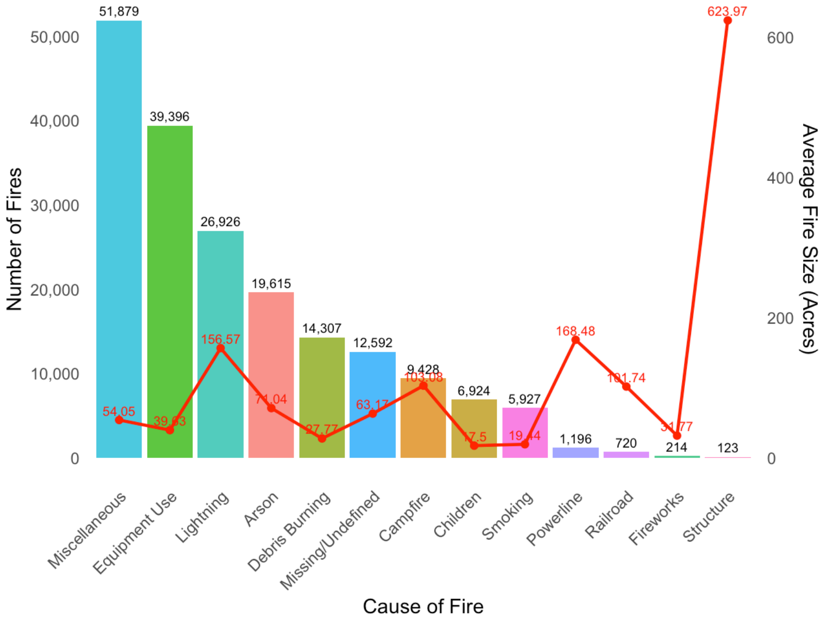
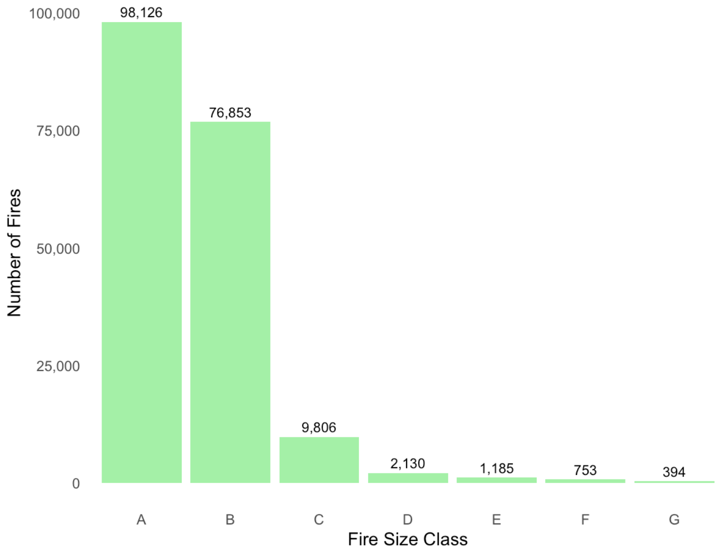


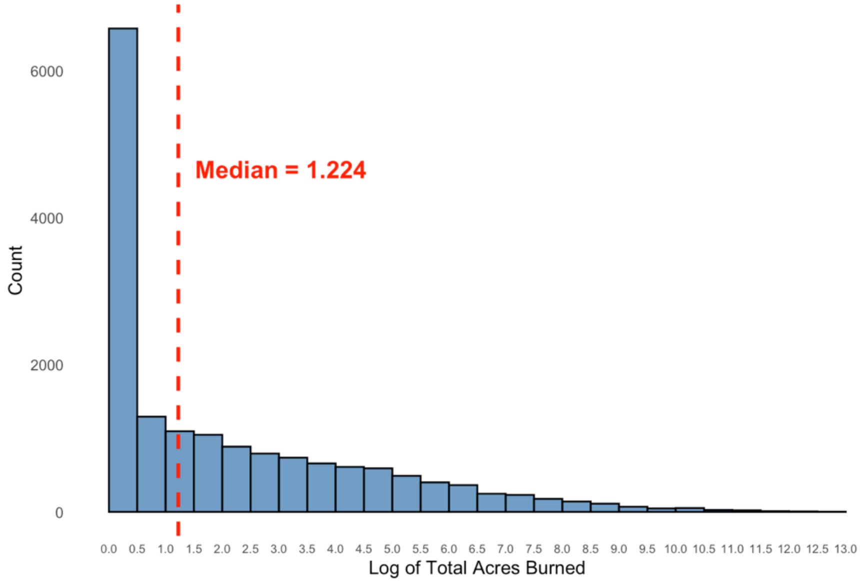
Figure 4 presents the number of wildfire incidents in California by cause (bars) and their corresponding average fire size (red line). The most frequent causes of wildfires are Miscellaneous and Equipment Use, though their average fire sizes are relatively moderate. Lightning ranks third in frequency but stands out with a significantly higher average fire size of 156 acres per fire. Krause et al. (2014) outline how global flash rate is expected to increase with climate change, so should be explored in our model analysis. Structure fires are the least common but have disproportionately high average fire size at 624 acres per fire, hence are not as relevant as lightning but this should still be noted for emergency response procedures.

Figure 5: Number of Fires in California by Fire Size Class



In Figure 5, the fire size classes range from Class A (smallest) to Class G (largest fires). Most of the recorded wildfires fall within Class A and B, indicating that most fires are small. As the size classification increases, the frequency of wildfires decreases. Despite lower occurrences, large wildfires contribute disproportionately to the total acreage burned, making them a critical concern for wildfire management.

Figure 6: Histogram of Log-Transformed Total Acres Burned



From Figure 6, we see that after log transformation the distribution is still strongly skewed, with 25% of the values being 0. This presents difficulties for prediction in Section 4 with less variance. Despite this, there are months when the acres burned is high, reaching up to 300000 acres.

3.3 Daymet Weather Data Analysis

Figure 7: California Choropleths on precipitation, maximum temperature, and minimum temperature

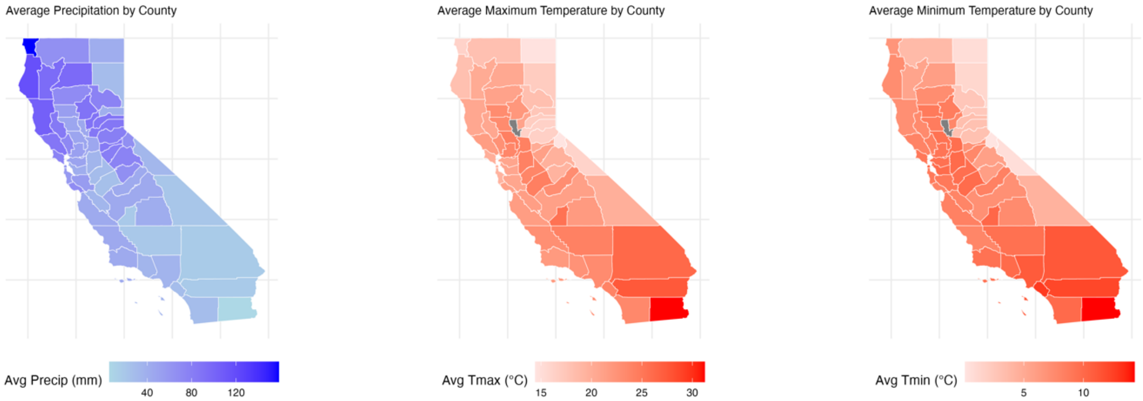
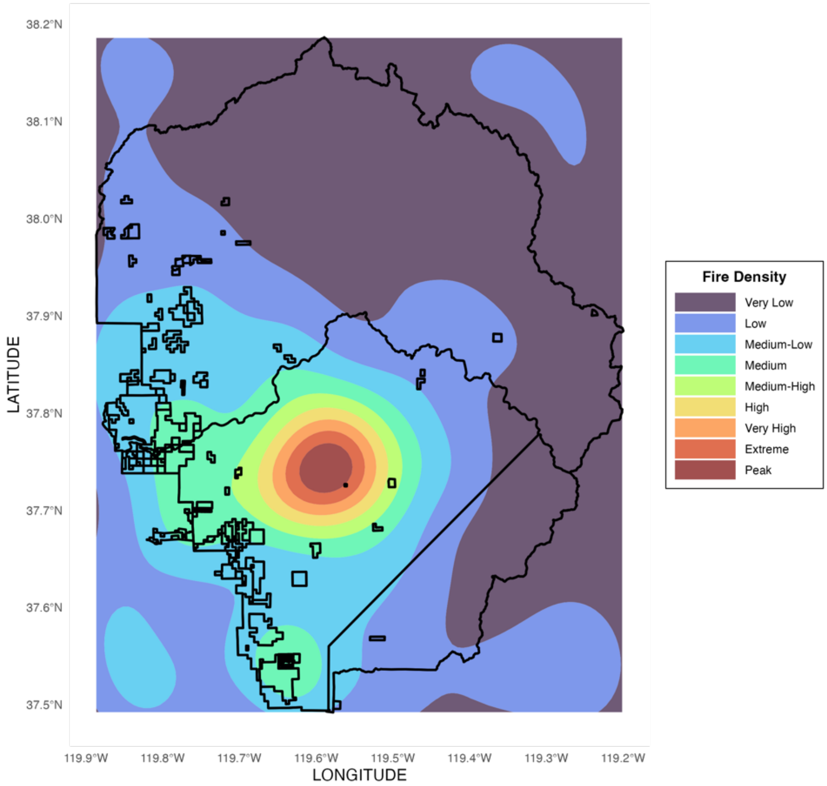

Average Precipitation by County
Average Maximum Temperature by County
Average Minimum Temperature by County
Avg Precip (mm)
40
80
120
Avg Tmax (°C)
15
20
25
30
Avg Tmin (°C)
5
10

Figure 7 illustrates choropleth maps representing the average total rainfall, average maximum temperature and average minimum temperature over the 24-year period by each county in California. Rainfall is higher and the temperature is lower in the northwest of the California and vice versa in the southeast of the state.

3.4 Yosemite National Park Case Study Analysis

Figure 8: Wildfire Density in Yosemite National Park (1992-2015)

Areas of higher fire occurrence shown in warmer colors



Yosemite's wildfire density plot reveals that a significant concentration of wildfires occurs in the central-western part of the park, between White Wolf and Yosemite Creek. Looking at a vegetation distribution plot and a map of the area (shown in Appendix 3 and 4), the high-density area in orange and red corresponds largely to the Lower Montane Forest. This area may have higher density levels due to lower elevation, drier conditions, more frequent lightning strikes, and human causes. Interestingly, the Lower Montane Forests have adapted to frequent but low-intensity fires. Hence, this area of the plot doesn't necessarily reflect severe fire damage but smaller and more frequent fires due to Ponderosa pines having evolved tolerance with thick fire-resistant bark and self-pruning lower branches that help prevent fires from climbing the canopy (Owen, 2020).Meanwhile the lower fire density (in grey) of the Eastern Subalpine Forest and Alpine Meadow zones likely experience less fire density due to their higher elevation, leading to more snow cover, sparser vegetation, and cooler and more moist conditions.

4. Classification Prediction of Wildfires in California and its Counties

4.1 Methodology

In this section we analyse the results of four different model types: Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting Machines. We investigate if we can successfully predict wildfires on a county-level and global level in California. We chose to predict wildfires for the 58 Californian counties individually to get interpretable results for local management decisions.

We selected a classification approach instead of regression due to the difficulties of attempting precise estimation on our heavily skewed data, shown in Figure 6. Classification's practical use here is still strong as it can give guidance on how to pre-position resources, allocate budget, and issue warnings. We use our decision tree model as an example of the ineffectiveness of regression prediction for wildfires, as it gave a very high RMSE of 9,774 and an MAE of 870 – as shown in Appendix 4.

Specifically, our models are predicting whether next month's total acres burned in each individual county will be higher or lower than the median of our log transformed training data's total acres burned. This median is based on the whole global California data, from 1992-2015. We chose to use the log transformed target variable median given that it helps us capture the wide range in our fire size data, given that for example, the difference between a 1-acre and 10-acre wildfire is arguably more significant and important than the difference between a 1000-acre and 1010-acre wildfire. The median was chosen given its resistance to outliers and because it allows for balanced classes.

In formula terms, our models are predicting:

$$\log(A_{t+1} + 1) > \mathrm{median}(\log(A_i + 1) : i \in [1, N])$$

Where:

- A_{t+1} is the total acres burned in the next month
- A_i represents each historical monthly total acres burned observation

- N is the total number of observations across all counties and months
- Adding 1 before taking the log handles cases where $\mathrm{A_i} = 0$
- The set $\mathrm{A_i}$ includes all historical fire data from 1992-2015 across all California counties

In section 4.2 we show the global performance metrics, while section 4.3 shows the county-level performance and the global comparison of ROC curves. The global models learn from all California data to predict whether each county will experience above-median acres burned in the next month. Instead of outputting a singular statewide total, they use the full California data to understand the effects of temperature and rainfall on risk anywhere in California, which counties have higher fire risk, and how historical patterns to inform county-level predictions.

All models use an 80-20 time-series split of training and testing data. This leaves 12,882 and 3,362 months of wildfires for the global training and testing data, respectively. The temporal sub setting means that we can reduce overfitting and underfitting and is more realistic to practical applications where only historic information is available. The proportion of classes above or below the median is 48% above and 52% below for the training data, and 51% above and 49% below for the testing data.

The performance metrics used for each model are Accuracy, Precision, Recall, F1 Score, and the Area under the ROC curve (AUC). The formulas used for these metrics are below, where TP, TN, FP, FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ or\ Sensitivity = \frac{TP}{TP + FN}$$

$$F1\,Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Feature importance calculations:

- Decision tree: uses reduction in Gini impurity/information gain when a feature is used for splitting nodes, and measures how well each feature separates classes
- Random forest: calculates mean decrease in Gini by averaging the impurity reduction across the whole tree ensemble, providing more stable importance scores than singular trees
- Logistic regression: uses z-scores to measures statistical significance of feature effects
- Generalised boosting model / Gradient boosting machine: calculates gain metric showing how each feature improves accuracy when used in splits, accounting also for feature interactions

4.2 Results

4.2.1 Logistic Regression

Logistic Regression (LR) is a “simple, fast and popular” classification method. It is used for binary classification problems (Moghim and Mehrabi, 2024), modelling the relationship between predictor variables and the probability of an outcome. We constructed an LR as our baseline for model performance comparison. Analysing the statistical significance of features and the Akaike Information Criterion (AIC), we employed backward elimination to determine whether more variables was improving wildfire prediction or adding unnecessary complexity.

Performing our feature ablation after the log transformation improved the rainfall significance to the 5% significance level, showing the importance of reducing data imbalance. The previous month area burned, and the month variable were found to not be significant, and the county risk level showed NA, raising concerns of perfect collinearity with the acres burned. We removed these variables in three stages to see if they individually improved our wildfire prediction. This process saw no significant change in the AIC (which balances model fit with complexity to avoid overfitting) nor the accuracy or AUC. We kept the more simplified model for our analysis to maintain model simplicity.

Table 1: Global Logistic Regression Confusion Matrix

Prediction/Actual	Less than Median	Greater than Median
Less than Median	1197	480
Greater than Median	456	1229

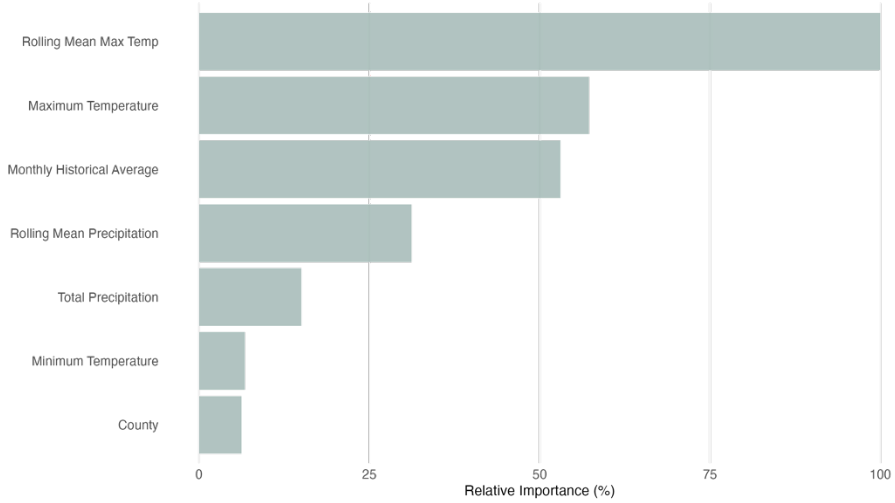
Table 2: Global Logistic Regression Performance Metrics

Versions	Accuracy	Precision	Recall	F1 Score	AUC
All Features	0.725	0.735	0.718	0.726	0.794
Reduced Features	0.722	0.729	0.719	0.724	0.793
Reduced and Fire Cause	0.711	0.687	0.73	0.708	0.776

We also examined the hypothesis that the global lightning rate is expected to respond to climate change as it is correlated with temperature short-term and our explanatory findings that lightning ignited fires cause greater damage (Krause et al., 2013). Feature-engineering the “fire cause” column into a binary variable of “Lightning” or “Human-Cause” did not change the LR results. As a result, we do not consider the cause feature further in our model analysis.

Figure 9: Global Logistic Regression Model Feature Importance

Relative importance of features in predicting next month's fire activity



Running this reduced feature model, we found that the rolling monthly maximum temperature was the most important feature in the global model, almost two times more relevant than the maximum temperature and monthly historical temperature average features. Precipitation played a small but not insignificant role. Interestingly, minimum temperature was a less important feature for prediction. Adding the climate variables to the dataset proved useful for prediction and management. Next, we look to improve predictiveness through advanced modelling.

4.2.2 Decision Tree

The decision tree was implementing using R's rpart package, which creates trees based on recursively partitioning the data based on the most important features. We used a complexity parameter of 0.001 to control growth, minimum splits of 20 observations, and terminal nodes of at least 10 observations, to ensure statistical reliability. We used the information gain criteria for splitting decisions because it works well given the continuous nature of our weather variables.

Table 3: Global Decision Tree Confusion Matrix

Prediction/Actual	Less than Median	Greater than Median
Less than Median	1188	489
Greater than Median	466	1219

Table 4: Global Decision Tree Performance

Versions	Accuracy	Precision	Recall	F1 Score	AUC
All Features	0.716	0.723	0.714	0.719	0.769

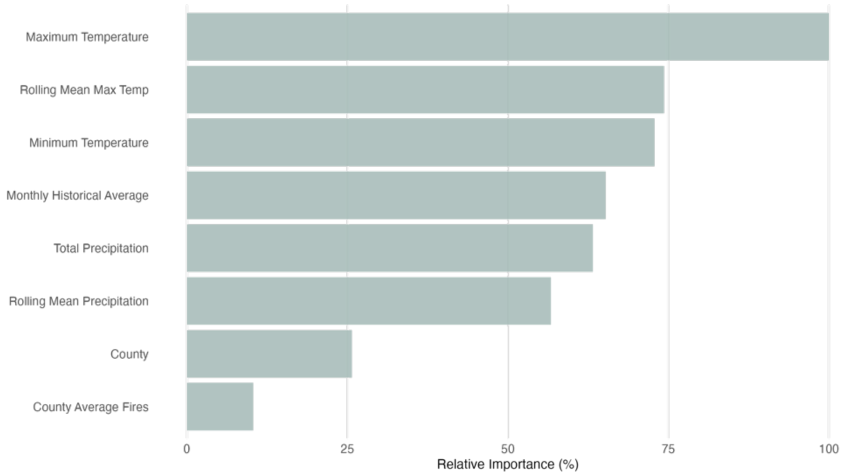


Figure 10: Global Decision Tree Model Feature Importance

Relative importance of features in predicting next month's fire activity

The model performed reasonably well, as seen by the performance metrics indicated above. Looking at the plot for feature importance, it is evident that maximum temperature and its rolling mean are the most influential features, and the minimum temperature has become a more relevant feature.

4.2.3 Random Forest

The random forest model was created using R's randomForest package, using an ensemble of 500 decision trees. Each tree is built on a random bootstrapped sample of the data and considers a random features subset of 3 at each split. To reduce class imbalance, we used a sample size parameter that takes the minimum of the classes sample size for both the classes. Each of our 500 trees make binary prediction and then for each new observation, the majority vote is taken as the final prediction. The random forest helped reduced overfitting compared to the previous decision trees given the improved adaptability of the model to new testing data.

Table 5: Global Random Forest Confusion Matrix

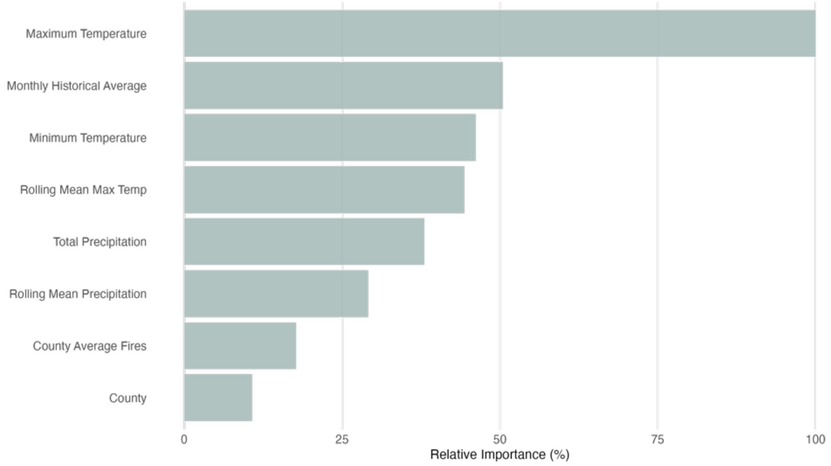
Prediction/Actual	Less than Median	Greater than Median
Less than Median	1208	458
Greater than Median	445	1251

Table 6: Global Random Forest Performance Metrics

Versions	Accuracy	Precision	Recall	F1 Score	AUC
All Features	0.725	0.735	0.718	0.726	0.801

Figure 11: Global Random Forest Model Feature Importance

Relative importance of features in predicting next month's fire activity



As mentioned at the beginning of section 4.2.3, the random forest was superior to the decision tree by a notable margin with an AUC of 0.801 compared to 0.769 for the decision tree. Compared to the decision tree we can see that the random forest placed higher feature importance on monthly historical average, suggesting that the random forest's ensemble approach may capture long-term patterns, especially considering it performed better in all metrics.

4.2.4 Generalised Boosted Model (GBM)

In the lectures we covered the GBM model using R's gbm package. Having compared the results of the simpler gbm package versus a more complex model utilising the XGBoost (eXtreme Gradient Boosting) package, we found the latter's performance was superior, likely due to its regularisation on leaf weights which helps prevent overfitting. The XGBoost model uses the logarithmic loss function, which punishes incorrect predictions more heavily when the predicted probability is far from the actual label. We used a standard learning rate of 0.1, a default maximum tree depth of 6, a small minimum child node weight of 1, and both a row and column sampling ratio of 0.8 to prevent overfitting while retaining its robustness. Unlike previously shown models, XGBoost's GBM builds trees additively using gradient descent to minimise the loss function (logarithmic in this case).

Table 7: Global Generalised Boosted Model Confusion Matrix

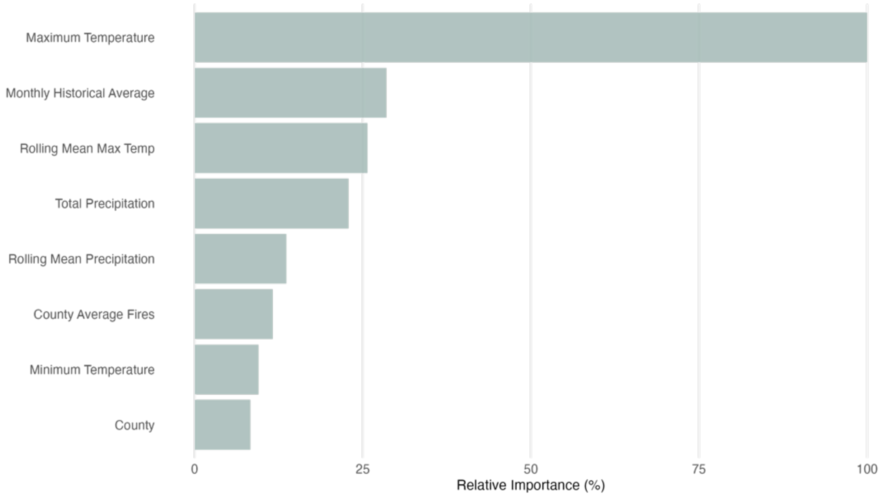
Prediction/Actual	Less than Median	Greater than Median
Less than Median	1214	457
Greater than Median	440	1252

Table 8: Global Generalised Boosted Model Performance Metrics

Versions	Accuracy	Precision	Recall	F1 Score	AUC
All Features	0.733	0.740	0.733	0.736	0.809

Figure 12: Global GBM Model Feature Importance

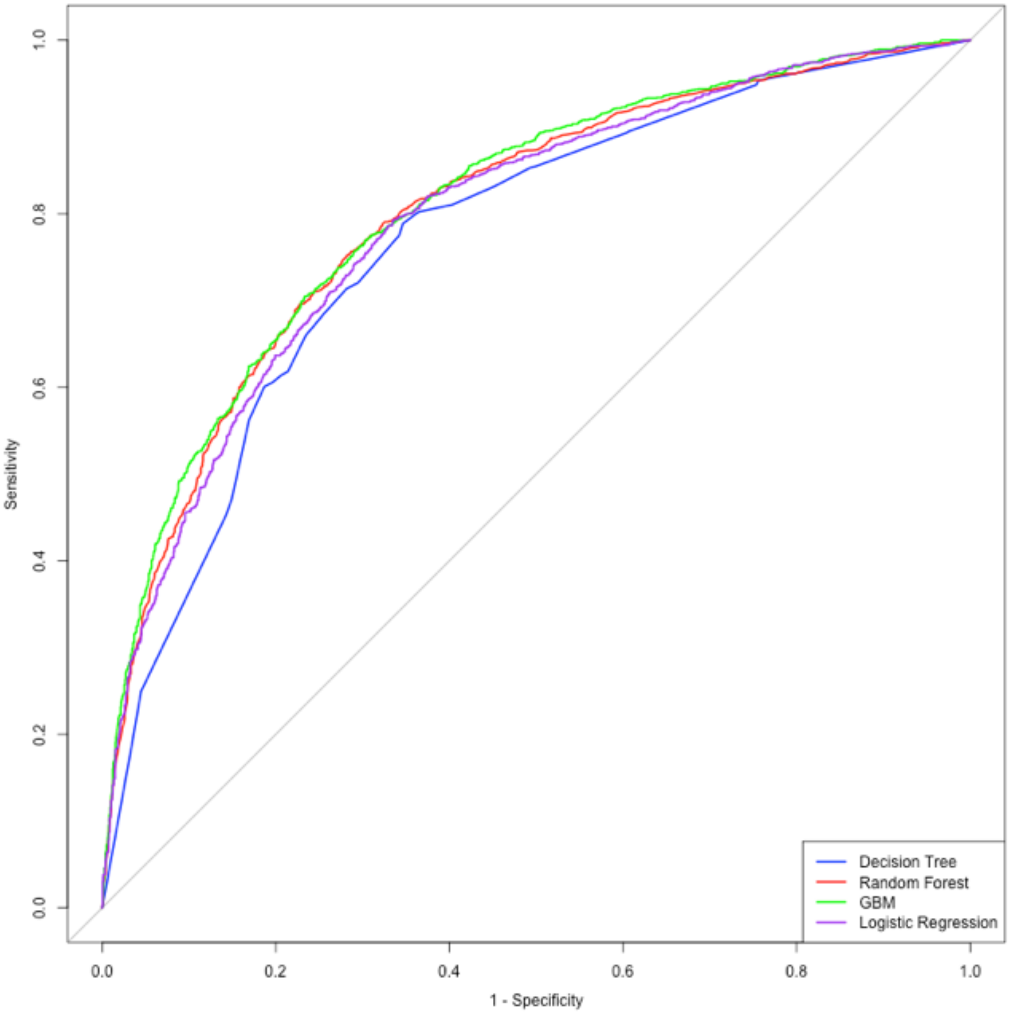
Relative importance of features in predicting next month's fire activity



The GBM model achieved strong metric performance, likely due to better capturing non-linear relationships, typical in wildfire data, and its iterative learning process which corrects previous errors. Yet again, maximum temperature is dominant in feature importance, however there is a much larger drop off in feature importance for our other variables, compared to prior models. Most notably, minimum temperature has now become one of the least important features in the GBM model compared to much higher importance for the decision tree and random forest. This suggests that the GBM determined that maximum temperature is the strongest predictor of burned acreage. The decrease in minimum temperature importance makes sense given the link of higher temperatures to fires as opposed to lower temperatures.

4.3 Comparing Prediction Models and County-level Prediction Analysis

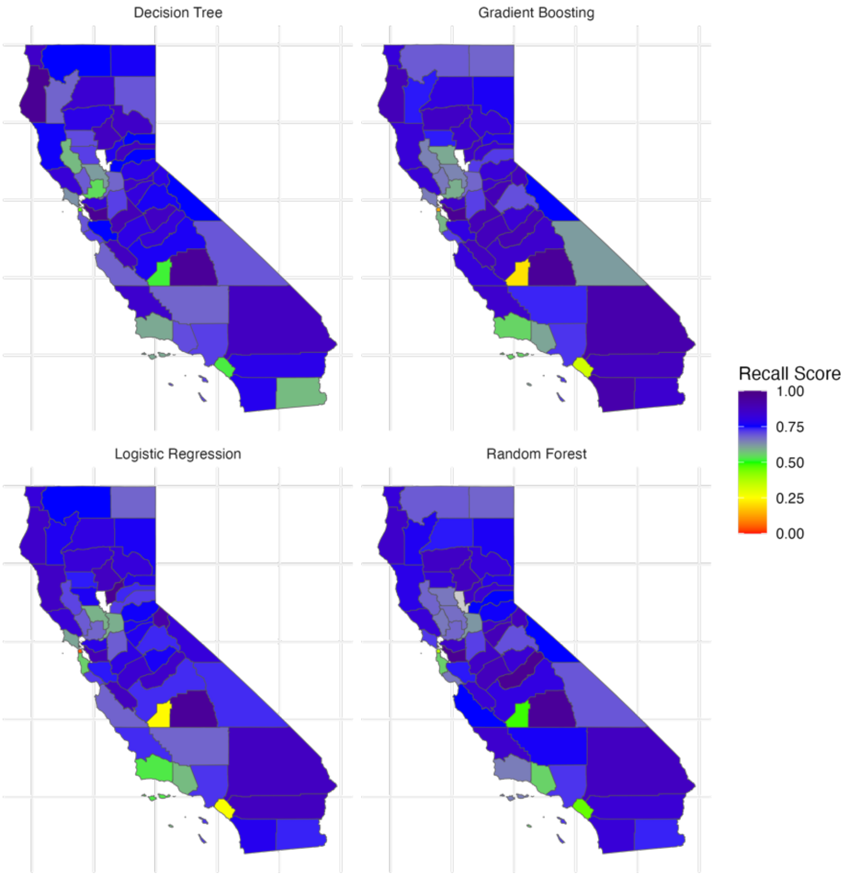
Figure 13: Global Model ROC Curves



The Receiver Operating Characteristic (ROC) curves provide valuable insights into the comparative performance of our 4 prediction models. All our models perform significantly better than random chance, illustrated by the diagonal line. The GBM and random forest show slightly better performance than the decision tree and logistic regression model, particularly in the lower false positive rate of around 0-0.2. Thee GBM and random forest reduce false alarms while maintaining recall, which could help prevent resource misallocations. The fact that all our models have similar AUC value ranges of about 0.76-0.81 indicates robust predictive capability across various classification thresholds. Furthermore, this suggests that our feature engineering was effective and that they provided a stable signal. The fact that the simpler models like the LR performed similarly to more advanced models like the XGBoost GBM suggest we are not overfitting.

Figure 14: County-Level Model Performance (Recall)

Performance across Decision Tree, Random Forest, GBM, and Logistic Regression



Recall is used as the comparison metrics for the county-level performance of our models due to its importance for wildfire classification prediction. The cost of false negatives is higher than the cost of false positives as it is better to have resources ready and not necessarily need them, than not have resources ready but need them. The increased cost of false negatives is that it may lead to insufficient resource allocation, delayed evacuations, and loss of life or property.

Analysing the county-level performance, we can see that the Northern and Southern California regions show relatively high and consistent recall scores across models. Our previously better performing models (GBM and LR) performed identifiably worse in Kings County, whereas our simpler models seem to have better captured the effects here. We found that for the training and testing data for Kings, 98% of the observations are below the total acres burned median. This is because the model uses the global median, even for county-level analysis, resulting in this imbalance. The differing performance highlighted that the LR and GBM are potentially over reliant on probabilistic predictions and the extreme imbalance likely biased them towards the majority class. Meanwhile, the found random forest and decision tree models performed better in Kings County as the decision tree makes decisions based on feature splits rather than probabilities, and the random forest's approach can give more priority to minority classes with its bootstrap sampling.

Coastal counties such as San Francisco show high variability between models, again due to an extreme class imbalance. Sutter County is grey in our choropleth because there have been no fires in this county in our dataset. An important county to highlight is L.A., given the January 2025 wildfires. Analysing the county-level results we find the recall score to be consistent with a range of 0.714-0.724 for all 4 models, likely due to there being less noise and more data points.

Our models were relatively consistent on many different counties, but some models struggled more when dealing with imbalanced classification testing data. This leads us on to Section 4.4 where we optimise for what we consider the most important metric in this context, Recall.

4.4 Optimising for Recall in Wildfire Analysis

As outlined above, we prioritised reducing false negatives over reducing false positives. To address this, we created a cost function to weight the misclassification of larger fires as a greater loss. The cost matrix counted a missed large fire as –100, a missed small fire as -5, a correct large fire prediction as +10, and a correct small fire prediction as +1. We applied this to the GBM, as the most effective prediction model.

Table 9: Global GBM Confusion Optimal Threshold (alpha=0.12)

Prediction/Actual	Less than Median	Greater than Median
Less than Median	408	88
Greater than Median	1246	1620

Note. See appendix: extreme GBM had a FN of 65 (23 less than this optimal model)

Table 10: Global Generalised Boosted Model Performance Metrics

Versions	Accuracy	Precision	Recall	F1 Score	AUC
Original GBM	0.725	0.735	0.718	0.726	0.794
Optimal GBM	0.675	0.624	0.910	0.740	0.807
Extreme GBM	0.892	0.192	0.522	0.281	0.866

Note. Extreme model created a new classification of large fires at the 95th percentile (1229 acres)

The results achieved the intended result at the cost of accuracy due to more false positives. In Table 9, we reduced the false negative count 458 to 88, meaning we correctly predicted 370 more of the large fires than our previous model. In Table 10 our recall rate went up – reflecting how well the model predicted the large fires - important for infrastructure investment decisions. Our cost values were arbitrarily selected, meaning this model can flexibly be adapted based on county priorities. A more extreme solution for reducing false negatives it to create a third classification category of large fires (e.g. we tried fires above the 95th percentile size), however this comes at too great a cost to the false negative rate, reflected in the significantly reduce recall rate, reducing the model's usefulness for overall resource allocation.

5. Conclusion

From our analysis we have shown that our classification models perform strongly for wildfire prediction and can be leveraged by local fire management to properly allocate resources. Most notably, weather patterns are shown to be strong predictors for the total acres burned in counties, whereas historical data is treated as less important. This corresponds with the data referenced in our introduction regarding the abnormality of the 9 largest Californian wildfires occurring in the last 7 years, as well as the devastating effects of the LA wildfires in 2025. The practical problem of this field is how fire mitigation and protection plans should be addressed when wildfire causes consist of both man-made reasons and climate patterns, of which the latter is becoming increasingly unusual. The approach we took in this analysis was to focus on binary classification above a set threshold. This proved fruitful with respectably high-performance metrics across models and most importantly, very high recall scores when the cost matrix was optimised. Although our models fall short on exact accurate prediction of total acreage burned in respective counties, this was to be expected given the skewed, tailed distribution of wildfire data. Having said this, our EDA section provided useful insights into the causes and patterns of wildfires across the US, California and its Counties, and even Yosemite National Park. Notably, outside research on the vegetation and elevation of Yosemite Park helped the understanding of why certain areas were more prone to fire risk. We believe that further success could be found by analysing areas on a more granular level and using this to enhance the features of predictive ML models. A highlight from our research was the consistency in the various performance metrics of our models, indicating robustness and that our feature engineering was effective. The close performance of our simple and more complex models suggests less concern of overfitting.

We believe that data analysis and predictive ML models will continue to be integral in improving wildfire management in the U.S. by better informing budgets, policymakers, and fire management departments on the causes and patterns of wildfires. The strong AUC values of our model are valuable for practical use cases where the threshold for high/low fire risk may need to be adjusted subject to resource availability and risk tolerance. Further improvements to our EDA and ML models could be made through predominantly two channels: 1. Improving feature engineering further through the inclusion of more complex data and vegetation information. 2. Dealing with class imbalances on a county-level through stratified sampling or applying SMOTE (Synthetic Minority Oversampling Techniques).

6. Bibliography

Primary Data

ORNL DAAC (2024). *Daymet: Station-Level Inputs and Cross-Validation for North America, Version 4 R1*. Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics. Available at: https://daac.ornl.gov/daacdata/daymet/daymet_xval_V4R1/comp/Daymet_xval_V4R1.pdf [Accessed 3 Feb. 2025].

Tatman, R. (2019). 1.88 million US Wildfires: 24 years of geo-referenced wildfire records. Kaggle. Available at: https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires/data [Accessed 3 Feb. 2025].

Secondary Literature

Krause, A., Kloster, S., Wilkenskjeld, S., and Paeth, H. (2013). *The sensitivity of global wildfires to simulated past, present, and future lightning frequency*. *Journal of Geophysical Research: Biogeosciences*. Available at: https://doi.org/10.1002/2013JG002502 [Accessed 3 Feb. 2025].

Hantson, S., Pueyo, S. and Chuvieco, E. (2014). Global fire size distribution is driven by human impact and climate. Global Ecology and Biogeography, First published 26 September 2014. Available at: https://doi.org/10.1111/geb.12246 [Accessed 3 Feb. 2025].

Hernandez, K. and Hoskins, A.B. (2024). Machine learning algorithms applied to wildfire data in California's Central Valley. Trees, Forests and People, 15, 100516. Available at: https://doi.org/10.1016/j.tfp.2024.100516 [Accessed 3 Feb. 2025].

Li, Z. and Yu, W. (2025). Economic Impact of the Los Angeles Wildfires. UCLA Anderson Forecast. Available at: https://www.anderson.ucla.edu/about/centers/ucla-anderson-forecast/economic-impact-los-angeles-wildfires [Accessed 3 Feb. 2025].

Moghim, S. and Mehrabi, M. (2024). Wildfire assessment using machine learning algorithms in different regions. Fire Ecology, 20, Article 104. Available at: https://doi.org/10.1186/s42408-024-00335-2 [Accessed 3 Feb. 2025].

Owen, S.M. (2020). Tree regeneration following large wildfires in southwestern ponderosa pine forests. U.S. Forest Service Research and Development. Available at: https://research.fs.usda.gov/treesearch/60282 [Accessed 6 Feb. 2025].

Turco, M., Abatzoglou, J.T., Herrera, S., Cvijanovic, I., and others (2023). *Anthropogenic climate change impacts exacerbate summer forest fires in California*. *Proceedings of the National Academy of Sciences (PNAS)*, 120(25), e2213815120. Available at: https://doi.org/10.1073/pnas.2213815120 [Accessed 3 Feb. 2025].

Walters, M. (2022). *Predicting the Likelihood and Scale of Wildfires in California using Meteorological and Vegetation Data*. Graduate Theses and Dissertations. University of Arkansas. Available at: https://scholarworks.uark.edu/etd/4521 [Accessed 3 Feb. 2025].

Extra

California Department of Forestry and Fire Protection (CAL FIRE) (2024). *California Fire Perimeters (All)*. Fire and Resource Assessment Program (FRAP). Available at: https://gis.data.cnra.ca.gov/datasets/CALFIRE-Forestry::california-fire-perimeters-all [Accessed 3 Feb. 2025].

Geography Fieldwork (2023). Yosemite Tourism Attractions. Available at: https://geographyfieldwork.com/YosemiteTourismAttractions.htm [Accessed 4 Feb. 2025].

ResearchGate (2015). Vegetation zones of Yosemite National Park, California, USA. Available at: https://www.researchgate.net/figure/egetation-zones-of-Yosemite-National-Park-California-USA-The-elevation-range-in-the_fig1_223666915 [Accessed 5 Feb. 2025].

Documentation

Ande, E. (2023). *Making maps with R*. Available at: https://eriqande.github.io/rep-res-web/lectures/making-maps-with-R.html [Accessed 1 Feb. 2025].

Becker, R.A., Wilks, A.R., and Brown, L. (2023). *maps: Draw Geographical Maps*. CRAN. Available at: https://cran.r-project.org/web/packages/maps/index.html [Accessed 1 Feb. 2025].

Chen, T. and Guestrin, C. (2023). *xgboost: Extreme Gradient Boosting*. CRAN. Available at: https://cran.r-project.org/web/packages/xgboost/xgboost.pdf [Accessed 3 Feb. 2025].

Henry, L. and Wickham, H. (2023). *purrr: Functional Programming Tools*. CRAN. Available at: https://cran.r-project.org/web/packages/purrr/index.html [Accessed 3 Feb. 2025].

Pierce, D. (2023). *ncdf4: Interface to NetCDF (Version 4) Format*. CRAN. Available at: https://cran.r-project.org/web/packages/ncdf4/ncdf4.pdf [Accessed 28 Jan. 2025].

Pebesma, E. (2023). *Simple Features for R (sf)*. Available at: https://r-spatial.github.io/sf/ [Accessed 3 Feb. 2025].

Therneau, T. and Atkinson, B. (2023). *rpart: Recursive Partitioning and Regression Trees*. CRAN. Available at: https://cran.r-project.org/web/packages/rpart/rpart.pdf [Accessed 3 Feb. 2025].

Urbanek, S. (2023). *RSQLite: R Interface to SQLite*. CRAN. Available at: https://cran.r-project.org/web/packages/RSQLite/vignettes/RSQLite.html [Accessed 10 Jan. 2025].

Walker, K. (2021). tigris: Load Census TIGER/Line Shapefiles. RDocumentation. Available at: https://www.rdocumentation.org/packages/tigris/versions/2.1 [Accessed 2 Feb. 2025].

Appendix

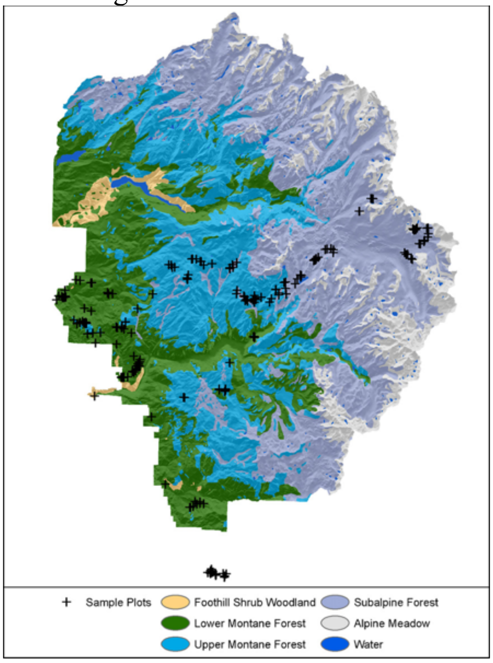
Appendix 1: Global GBM Confusion Extreme Threshold

Prediction/Actual	Less than Median	Greater than Median
Less than Median	2928	65
Greater than Median	298	71

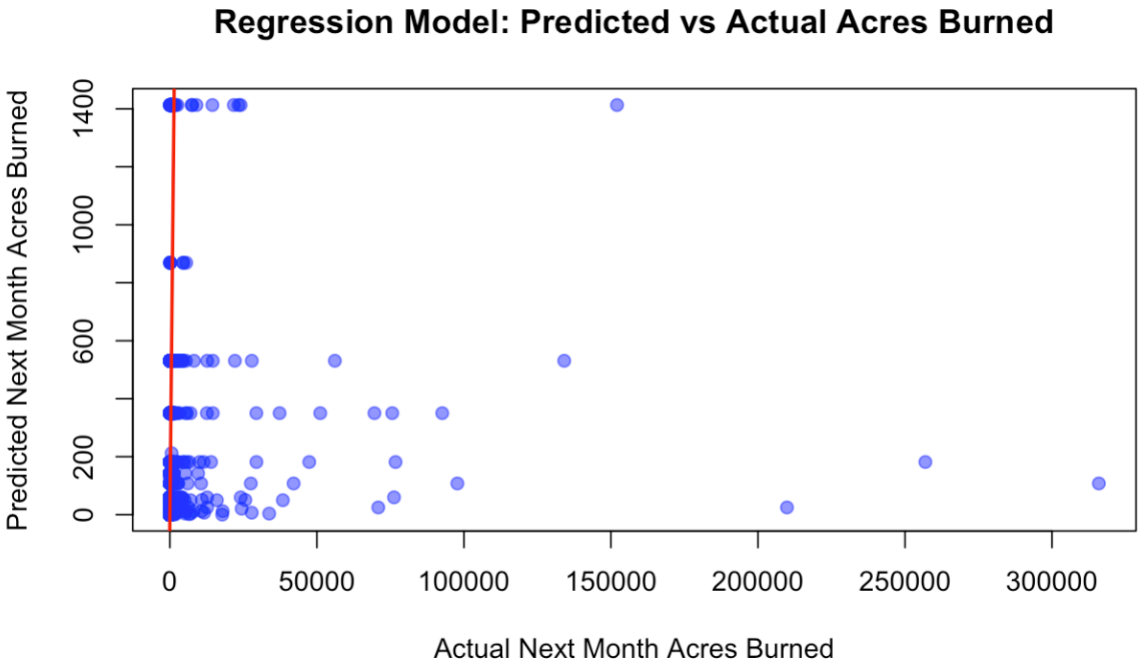
Appendix 2: Map of Yosemite National Park

STANISLAUS NATIONAL FOREST
5 miles
©GORP
TOIYABE NATIONAL FOREST
Pacific Crest Trail
STUBBLEFIELD CANYON
MATTERHORN CANYON
VIRGINIA CANYON
Lake Elanore
Hetch Hetchy Reservoir
GRAND CANYON OF THE TUOLUMNE RIVER
Hetch Hetchy Entrance
Tuolumne River
Tioga Pass Entrance
White Wolf
Tioga Rd.
Camp Mather
120
Tuolumne Meadows Visitor Center
Big Oak Flat Entrance
Yosemite Creek
John Muir Trail
Tuolumne Grove
Merced River
Big Oak Flat Rd.
Tamarack Flat
Merced Lake
140
Arch Rock Entrance
Badger Pass Ski Area
YOSEMITE NATIONAL PARK
SIERRA NATIONAL FOREST
Wawona
Mariposa Grove
South Entrance
Fish Camp
San Francisco
Los Angeles
395
167
108
120
6
140
41
5
99
395
14

Appendix 3: Yosemite National Park Vegetation Zones Plot



Appendix 4: Demonstration of the Poor Performance of Regression Prediction for our Decision Tree


Regression Model: Predicted vs Actual Acres Burned
Predicted Next Month Acres Burned
Actual Next Month Acres Burned