Scott Babcock

Western Governors University

9/7/2025

Executive Summary

Forest Fire Classification Using Random Forest

**A1, Statement of Problem & Hypothesis**

In the Western United States, wildfires are a growing problem that have a far-reaching impact. Wildfires have a devastating impact on citizens and communities and have the potential to damage critical infrastructure and ecosystem services. The U.S. Department of the Interior reported that as of 2023, the cost of wildfires had elevated to $424 billion annually (U.S. Department of the Interior, 2023).

The analysis aimed to predict when a fire could start based on weather conditions. Understanding how these variables influence ignition probability is critical for resource allocation, early warning systems, and risk mitigation. The belief is that the weather features will predict when a fire ignition occurs, thus establishing the following research question, null and alternative hypotheses:

**Research Question:** Can environmental conditions be used to accurately predict when a wildfire will occur?

**Null Hypothesis (H0):** Weather features have no predictive power regarding the timing of fire ignition events.

**Alternative Hypothesis (H1):** Weather features have predictive power regarding the timing of fire ignition events.

**A2, Data-Analysis Process**

The data for the project was hosted on Zenodo, and can be found at this location <https://zenodo.org/records/14712845>. The data was downloaded in CSV format with a file size of 1,482KB. There were 14,988 rows and 14 columns. The dataset contains daily weather conditions in California from 1984 to 2025, indicating whether a fire ignition occurred on a given day.

The data was checked for duplicate records and missing values. No duplicate records were found. There were twelve missing values, which represented 0.08% of the data. Records from 2025 were dropped from the dataset because they contained incomplete data. Records from 2024 were ultimately dropped as well because no fire occurrences were listed, which was inaccurate. The remaining missing values were estimated using linear interpolation. This choice was ideal for ordered data because it would preserve trends.

The data was manipulated to get features into the correct data types, including re-establishing the target variable as a factor with two clear levels: Fire and No Fire. An additional feature was created to capture a human element in fire ignition, resulting in a six-day true/false buffer around a holiday. Additional lagging weather indicators for maximum and minimum average temperatures were created to potentially account for conditions leading up to a fire ignition.

Two Random Forest models were created, one that contained data for the entire training timeframe (1984 to 2017) and one that contained training data from 2004 to 2017, when it was determined that fires became more prevalent, as indicated by a t-test. Both models were validated on the same unseen data from 2018 to 2023. Prior to modeling, it had been determined that a class imbalance existed. There were 10,005 no fire occurrences compared to 4,971 fire occurrences, representing a 67%/33% proportion, respectively. The class imbalance meant it was critical to focus on the true positive rate, or sensitivity, as opposed to overall accuracy, given the nature of the research question.

**A3, Outline of Findings**

The Random Forest Classifier with the full training timeframe produced an overall accuracy of 76.4% and, more importantly, a sensitivity rate of 72.6% indicating that the model was learning patterns that predict fire occurrence. These results were significantly better than the no information or random chance rate of 54.7%.

A screenshot of a computer

AI-generated content may be incorrect.

A confidence interval was created for ROC-AUC, which determined whether to accept or reject the null hypothesis. If the lower bound of the confidence interval exceeded 0.5, the analysis would reject H0. The Random Forest Classifier achieved a ROC-AUC score of 0.828 (95% CI: 0.8103-0.845 (DeLong)), significantly above the 0.5 threshold. Therefore, the null hypothesis that weather features have no predictive power was rejected, and the alternative hypothesis was supported, indicating that weather conditions predict fire ignition events better than chance.

**A screenshot of a computer

AI-generated content may be incorrect.**

**A4, Limitations of Techniques and Tools**

A Random Forest Classifier was used in the analysis, which effectively ran 6,000 decision trees. The primary limiting factor is time and memory consumption. The model took considerable time and resources to train. To implement on a larger scale, improvements would be needed to improve the speed without sacrificing predictive power.

Another limitation of the analysis was the vague description of where the data was from geographically. The authors indicated that the data came from weather observations and wildfire data in California. However, there was no indication of where the fires occurred within the state. Having coordinates would help improve the dataset, or if the data represented averages for the entire state, it should be specified.

**A5, Summary of Proposed Actions**

The analysis findings implied that environmental factors such as minimum and maximum temperature and lagging 7-day temperature averages contain meaningful predictive signals, and that machine learning approaches could leverage these signals to anticipate wildfire risk. The recommended course of action would be to integrate weather-based predictive models into wildfire risk monitoring systems, allowing resource managers to better allocate prevention and response efforts during periods of elevated risk. Another approach would be to build more localized models. Local models could include more fine-tuned variations in predictors that may get obscured when looking at larger areas, such as if the data represented all of California. For future study, one approach would be to incorporate additional predictors such as vegetation indices, moisture levels, or human activity data to improve predictive accuracy.

**A6, Expected Benefits of Study**

The cost of wildfires is increasing, and as of 2023, it was estimated to be approximately $424 billion annually. Incorporating weather-based predictive models into wildfire risk monitoring systems would allow resource managers to be proactive instead of reactive in response to wildfires. If early detection and prevention led to even a 5% reduction in expense, the savings would amount to over $21 billion annually. Further, this would better protect homes and communities. It is estimated that the 2025 Los Angeles fire alone destroyed over 15,000 structures (Pequeno, 2025). Improvements in detection and prevention will also have an impact on wildlife. Smoke from the Colorado wildfires in 2020 has been linked to the mass avian die-off in 2020, where an estimated one million birds perished (Hardt, 2023). Implementing the learnings of this study would be a statistically supported strategy to reduce wildfire-related losses and protect both human infrastructure and wildlife.

**Sources**

Hardt, B. (2023, July 21). *How does wildfire smoke affect wildlife*? National Wildlife Federation. <https://blog.nwf.org/2023/07/how-does-wildfire-smoke-affect-wildlife/>

Pequeno IV, A. (2025, January 21). *California fires: Here’s the data behind the historic blazes that have burned through 40,000 acres*. Forbes. <https://www.forbes.com/sites/antoniopequenoiv/2025/01/21/california-fires-heres-the-data-behind-the-historic-blazes-that-have-burned-through-40000-acres/>

U.S. Department of the Interior, Office of Policy Analysis. (2023, May 25). *Increasing damages from wildfires warrant investment in wildland fire management* [Brief]. U.S. Department of the Interior. <https://www.doi.gov/sites/doi.gov/files/ppa-brief-wildland-fire-econ-review-2023-05-25.pdf>

Yavas, C. E., Kadlec, C., Kim, J., & Chen, L. (2025). *California Weather and Fire Prediction Dataset (1984–2025) with Engineered Features* [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.14712845>