Zoë Langhoff Weinstein

Professor Ming

Environmental Data Science

3 May, 2025

Predictive Model of Forest Fires Using Machine Learning Algorithms

Introduction

This project aims to predict whether a forest fire will start on a given day in California using machine learning. Wildfire prediction is difficult due to the variability of environmental factors. Accurate forecasting could mitigate damage and save lives. This research was inspired by the devastating Palisades Fire in Los Angeles, which burned over 23,000 acres and destroyed more than 6,500 structures [1]. Unlike other disasters, wildfires are highly dynamic and hard to prepare for, especially due to wind patterns that change fire behavior rapidly [2]. In addition to building predictive models, this study also investigates which environmental factors are most important in predicting fire onset.

Data

The dataset combines wildfire records from CAL FIRE and meteorological data from NOAA Climate Data Online, covering 40 years (1984–2024). Each row represents one day of the year, totaling just under 15,000 records. The target variable is a binary indicator of whether a fire started on that day. There are various independent variables used to predict the fires which include precipitation, maximum temperature, minimum temperature, average wind speed, temperature range, wind temperature ratio, month, precipitation in the last seven day, average wind speed in the last seven days, day of the year (from 1-365). The dataset includes engineered features to support better model performance [3]. The hypothesis is that temperature and

precipitation would be key predictors, and that fire days would show higher temperatures and wind speeds.

Methodology

To model fire occurrence, I've employed three supervised classification algorithms: Logistic Regression, Random Forest, and XGBoost. The dependent variable is fire occurrence (1 = fire, 0 = no fire). The data is split up into a train and test dataset, the train dataset is 1984-2016, and the test is 2017-2024 this is roughly a 80%-20% split. The purpose of the logistic model is to get a baseline for how the independent variables interact with dependent variables. The Random Forest and XGBoost models were selected for their strength on structured tabular data, ability to model nonlinear interactions, and built-in handling of feature scaling and interactions [4].

Results

Logistic Regression

The model achieved 72% precision for both fire and no-fire predictions. It correctly identified 86% of no-fire days but for true recall it achieved a 51%. The F-1 score tells us that predicting no fires is strong with 78%, but predicting fires does a moderate job, with 60% accuracy.

Random Forest

Random Forest produced similar metrics to the logistic model: 72% precisions, 86% recall for no-fire days, and 51% for fire days. F1-scores again favored no-fire predictions 78% for no fire and 60% for fire, indicating a tendency to under-predict fires.

XGBoost Classification

XGBoost slightly improved fire detection, achieving 73% precision and 53% recall for fire days. It maintained 86% recall for no-fire days, with F1-scores of 0.79 for no fire and 0.61 for fire making it the strongest performer overall. Prediction of fires is still a quite low percentage at 53%, just slightly better than a coin toss.

Discussion

All three models—Logistic Regression, Random Forest, and XGBoost—produced similar overall classification metrics, with notable differences in how they handle fire vs. no-fire predictions. The three models are strong at predicting when a fire is not going to occur, however they struggle with detecting actual fire events which can have some serious drawbacks. The XGBoost Classification model performs the best of the three models with a true recall of 53%. This means of the times a fire occurred, the XGBoost model was able to idsentify it just over half of the time. The relatively low recall for the fire class across models suggests that many fire cases are being missed which can be harmful because early fire detection is essential to prevent harm. All models would benefit from improvements such as adjusting class weights, resampling techniques, or threshold tuning to increase sensitivity to fire events. Because fires are not that common they do not appear often in the dataset. Overall, while the models demonstrate solid baseline performance, additional strategies are needed to address the imbalance and improve fire detection without sacrificing too many false positives.

References

[1] California Department of Forestry and Fire Protection. (2025, January 7). *Palisades Fire*. CAL FIRE. https://www.fire.ca.gov/incidents/2025/1/7/palisades-fire

[2] Liu, N., Lei, J., Gao, W., Chen, H., & Xie, X. (2021). Combustion dynamics of large-scale wildfires. *Proceedings of the Combustion Institute*, 38(4), 6415–6422. https://doi.org/10.1016/j.proci.2020.11.006

[3] Yavas, C. E.), Kadlec, C., Kim, J., Chen, L. (Jan 21 2025). California Weather and Fire Prediction Dataset (1984–2025) with Engineered Features. ZENODO. 10.5281/zenodo.14712845

[4] Gangani, P., Alyaseri, S., Hosseini, S., (March 28, 2025). *Machine Learning Approaches for Predicting Diabetes Onset: A Comparative Study of XGBoost, Random Forest, and Traditional Models*. TechRxiv. https://doi.org/10.36227/techrxiv.174317792.24675569/v1