Wildfire Ignition Risk Prediction (Colorado)

Malik Ibrahim Ali Khan
M.S. in Data Science
Regis University, Denver, CO, USA
mibrahimalikhan001@regis.edu

Abstract

Wildfires are becoming a serious concern in Colorado because they affect communities, infrastructure, and the environment. Predicting the exact time and location of a wildfire ignition is very difficult, so this project focuses on ranking locations by wildfire ignition risk instead of predicting exact fire events.

This project combines wildfire ignition data from NASA's MODIS FIRMS system with daily climate data from the gridMET dataset. The final dataset includes grid-level observations across Colorado from 2018 to 2023. Exploratory data analysis showed that wildfire events are rare, seasonal, and more closely related to accumulated dryness than single-day weather conditions.

Several machine learning models were tested, including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. The models were trained on data from 2018 to 2022 and evaluated on unseen data from 2023. Since wildfire events are rare, the focus was placed on risk ranking rather than traditional accuracy.

The results suggest that climate-based features can help identify areas with higher wildfire ignition risk. This approach can support wildfire monitoring, planning, and decision-making.

I. Introduction/Background

Wildfires are a recurring problem in the western United States, including Colorado. Climate conditions such as high temperature, low precipitation, and seasonal dryness can increase the chance of wildfire ignition. Because of this, it is important to identify areas where wildfire risk may be higher.

Traditional wildfire risk methods often depend on past fire locations or static risk maps. However, fire risk changes over time because weather and climate conditions change. This project uses a data-driven approach to study wildfire risk using climate data and satellite fire detection data.

This project also connects to several important data science skills, including data preparation, feature engineering, exploratory analysis, machine learning, model evaluation, and reporting.

II. Problem Statement

The main problem addressed in this project is how to identify and rank locations in Colorado based on wildfire ignition risk using climate data. Since wildfire ignitions are rare and affected by many factors, predicting the exact location and timing of each fire is not realistic.

Instead, this project focuses on relative risk ranking. The goal is to find which locations are more likely to experience wildfire ignition compared to others. This is useful because even if exact prediction is difficult, risk ranking can still help with monitoring and resource planning.

III. Related Work

Previous wildfire research has shown that climate conditions, drought, fuel dryness, and seasonal patterns are important factors in wildfire activity. Many studies use climate variables and historical fire records to model fire occurrence or burned area.

However, wildfire prediction is difficult because fire events are rare and unevenly distributed. Traditional accuracy can be misleading in this type of problem because a model could predict "no fire" most of the time and still appear accurate. Because of this, this project focuses more on ranking high-risk areas rather than only measuring accuracy.

IV. Methodology/Approach

This project used historical wildfire and climate data to build machine learning models. Fire data was obtained from NASA's MODIS FIRMS dataset, and climate data was taken from the gridMET dataset. The analysis was completed using Python with libraries such as pandas, NumPy, scikit-learn, and XGBoost.

The data preparation process included filtering fire detections to Colorado, aligning fire detections with the gridMET spatial grid, and creating binary fire labels for each grid-cell day. The yearly datasets were then combined into one modeling dataset covering 2018 to 2023.

Feature engineering included lagged precipitation values, rolling precipitation sums, and day-of-year features to capture seasonal effects. These features were used because exploratory analysis showed that wildfire risk is related to accumulated dryness and seasonal patterns.

The models tested included Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. Logistic Regression was used as a baseline model, while the tree-based models were used to capture nonlinear relationships between climate variables and wildfire risk.

V. Data Analysis

Exploratory data analysis showed that wildfire ignition events are extremely rare in the dataset. Most grid-cell days had no fire event, which created a strong class imbalance problem. This made traditional accuracy less useful as a main evaluation measure.

The analysis also showed that wildfire ignitions are seasonal, with more activity during warmer and drier periods. Rolling precipitation and lagged precipitation were useful because they helped represent accumulated dryness over time. Temperature also played a role, but dryness-related features were more important for understanding wildfire risk.

For evaluation, the models were trained on data from 2018 to 2022 and tested on unseen data from 2023. This helped check whether the models could perform on future data rather than only fitting past patterns.

VI. Expected Outcomes

The expected outcome of this project is a practical method for ranking wildfire ignition risk across Colorado. The goal is not to predict every fire exactly, but to identify higher-risk locations more effectively than random selection.

The results are expected to show that climate variables can provide useful information for wildfire risk ranking. This type of approach could support wildfire monitoring, resource allocation, and early planning.

VII. Timeline

The project was completed during the academic term. The early stage focused on collecting and understanding the datasets. The middle stage focused on data preparation, exploratory analysis, and feature engineering. The final stage focused on model training, evaluation, visualization, and reporting.

VIII. Conclusion

This project shows that wildfire ignition risk can be studied using climate data and machine learning models. Since exact wildfire prediction is difficult, risk ranking provides a more practical way to identify areas that may need more attention.

The project also demonstrates how data science can be applied to an environmental problem. By combining fire detection data with climate features, the model can help rank areas by relative wildfire ignition risk and support better decision-making.

References

[1] J. T. Abatzoglou, "Development of gridded surface meteorological data for ecological applications and modelling," *International Journal of Climatology*, vol. 33, no. 1, pp. 121–131, 2013.

[2] NASA Earth Observing System Data and Information System (EOSDIS), "Fire Information for Resource Management System (FIRMS)," NASA, 2023. [Online]. Available : `https://firms.modaps.eosdis.nasa.gov/`

[3] L. Giglio, J. Descloitres, C. O. Justice, and Y. J. Kaufman, "An enhanced contextual fire detection algorithm for MODIS," *Remote Sensing of Environment*, vol. 87, no. 2–3, pp. 273–282, 2003.

[4] J. T. Abatzoglou and C. A. Kolden, "Relationships between climate and macroscale area burned in the western United States," *International Journal of Wildland Fire*, vol. 22, no. 7, pp. 1003–1020, 2013.

[5] J. T. Abatzoglou and A. P. Williams, "Impact of anthropogenic climate change on wildfire across western US forests," *Proceedings of the National Academy of Sciences*, vol. 113, no. 42, pp. 11770–11775, 2016.

[6] Z. Ben Bouallègue and D. S. Richardson, "On the ROC area of ensemble forecasts for rare events," *Weather and Forecasting*, vol. 37, no. 5, pp. 849–864, 2022.