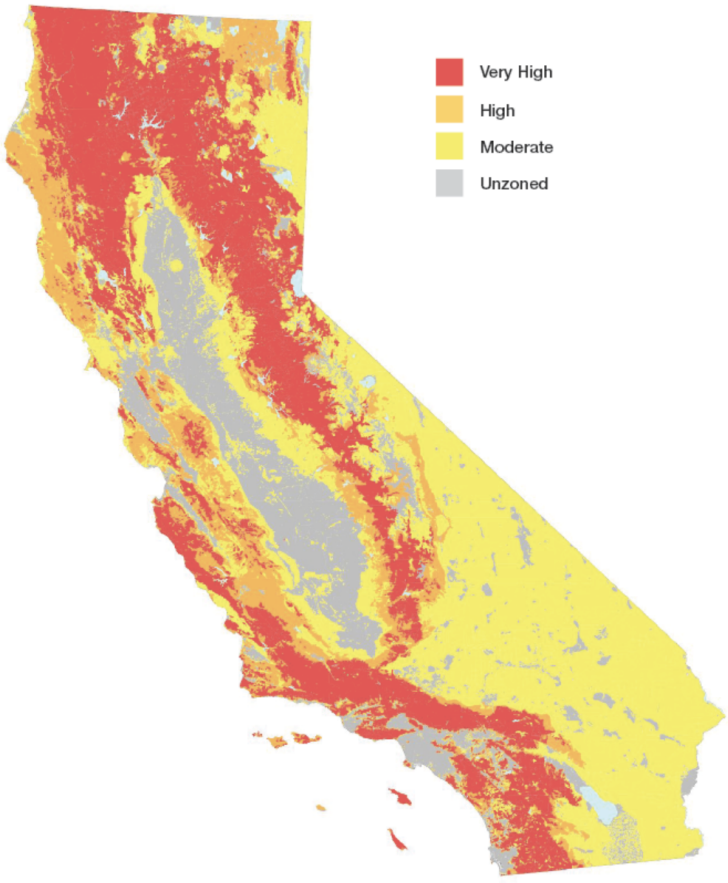
California Wildfire Prediction: Project Writeup

Mikaela McLean, Jovoney Morton, Emily Wong

BAX 452 - Machine Learning



Executive Summary

Wildfires pose a growing and increasingly severe threat to the state of California, driven by rising global temperatures, prolonged drought conditions, and expanding urban development into fire-prone areas. This project aims to predict the probability of a wildfire occurring through machine learning. California wildfire and weather data from 1984-2025 was used to create an XG-Boost model focused on recall and accuracy that can successfully predict wildfire occurrences. With this model, resource allocation can be optimized to better prepare Californians for wildfires and mitigate its effects on the economy, environment, and human health.

Background

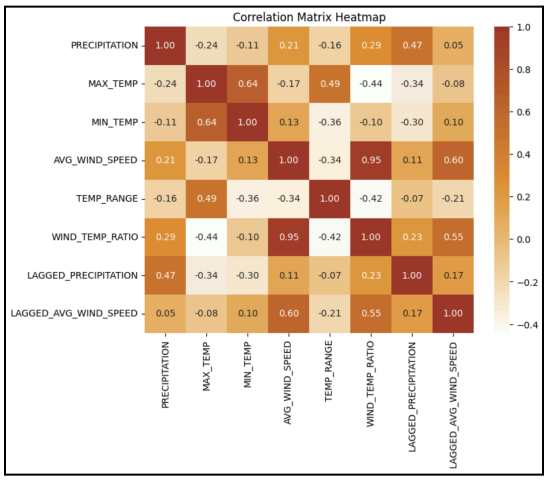
Wildfires in California have far-reaching environmental and economic consequences, creating an urgent need for improved wildfire prediction and mitigation strategies. According to the California Ocean Protection Council, wildfires severely impact coastal and marine ecosystems by introducing carbon pollution, which disrupts marine life, degrades habitats, and threatens marine industries. In addition to environmental harm, wildfires impose significant economic burdens. The UCLA Anderson Forecast estimates that the recent wildfires in Los Angeles alone have caused property and capital losses ranging from $95 billion to $164 billion, alongside a projected $4.6 billion decline in GDP for 2025.

These compounding risks underscore the necessity for advanced wildfire prediction systems, which can help businesses and policymakers develop proactive risk management strategies. By being able to predict wildfire risk, industries can better protect assets, minimize economic disruption, and contribute to environmental conservation in the face of escalating wildfire threats.

EDA & Data Processing

The dataset utilized for the classification model contains 14 features comprising over 14,000 rows of California weather data from 1984 to 2025 including temperature, wind speed, precipitation, and time. Some additional features were lagged variables and weather ratios. The target is a binary variable indicating whether a fire started on a particular day. Examining the data through EDA, the dataset

contained few null values which were removed from the dataset. In addition to non-normal and non-linear relationships, a few features displayed high correlation with one or more features, seen below in the correlation heatmap, which may affect model performance and interpretation (this is addressed further during model optimization).



Insights from EDA were used to preprocess the data. Further delving into the target variable, there was a class imbalance, where no-fire instances outnumbered fire-instances by a 2:1 ratio. To balance the data, oversampling was implemented via SMOTE to ensure that the model has equal instances of positive and negative occurrences. To ensure all features contribute equally to learning, standardization using StandardScaler was applied to the numeric features to optimize model performance. In addition, categorical features such as season were transformed into numerical encoding for the model to use.

Model Selection

Wildfires are influenced by complex interactions between multiple factors, and the goal of the model is to predict whether a fire will occur or not, which calls for a binary classification model. There are many possible models that can be implemented, ranging from logistic regression, decision-trees, to neural networks. Logistic regression models are useful if wildfire occurrences and its predictors follow a normal, linear relationship. Neural networks should be considered if wildfire occurrences followed a strong sequential pattern, where past fires directly influenced future ones. However, wildfires are generally influenced by real-time weather conditions, which follow complex and nonlinear relationships.

Decision-tree methods capture complex data very well, and are better suited for smaller datasets and interpretability, so tree-based ensemble methods Random Forest and XG-Boost were chosen to model the wildfire predictors. Through initial comparisons between both models, Random Forest showed slightly better performance than XG-Boost. However, XGBoost scored slightly better than Random Forest in cross-validation, indicating that XG-Boost fits the data better. Combined with the fact that XG-Boost models can be more extensively fine-tuned compared to Random Forests, the XG-Boost model was chosen and optimized to fit the data.

Model Evaluation and Optimization

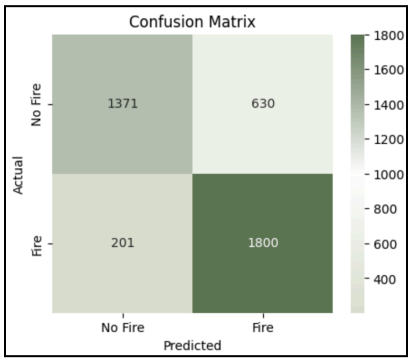
Due to the vast and destructive effects of wildfires, it is crucial that the model accurately predicts a wildfire occurrence when there is one. To prevent missed-fires (false-negative classifications), the XG-Boost model was optimized to achieve the high recall and accuracy score while maintaining high prediction performance.

The first step to improve the model was to perform additional feature selection and engineering. From the EDA, two features (average wind-to-temperature ratio and max temperature) had high correlation with other features. To mitigate the multicollinearity, the two features were removed from the model. In addition, weather is complex, and different aspects of weather can have interaction effects. So, based on domain knowledge, interaction between a subset of features was included into the model.

The next step to improve model performance was to determine the decision threshold for the model. In order to prioritize reducing missed-fires (false-negatives), an increase in recall is needed at the expense of precision, which is related to false-alarms (false-positives). Plotting a precision-recall graph showed that the optimal threshold should be around 0.65, but to maximize recall while still considering precision, the default threshold 0.5 was maintained in the model.

After feature selection and decision thresholds were determined, the next step is to fine-tune the hyperparameters. Using RandomCV search, various hyperparameters including penalization weights, tree size, and subset sizes were iterated to find the best hyperparameter values to optimize recall, with

secondary emphasis on accuracy and precision. After fine-tuning, the prediction model saw a 7.98% improvement in recall, which resulted in a 39.82% decrease in missed-fires. However, due to the precision-recall tradeoff, the increase in recall also caused precision to decrease by 7.62%.

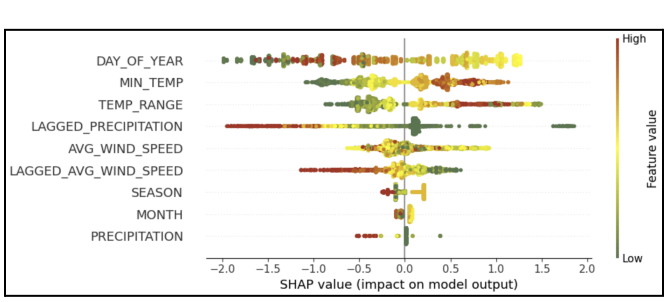


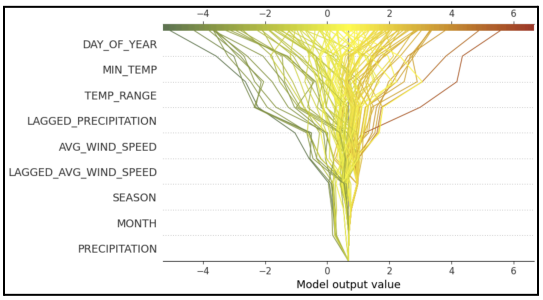
	precision	recall	f1-score	support
False	0.87	0.69	0.77	2001
True	0.74	0.90	0.81	2001
accuracy			0.79	4002
macro avg	0.81	0.79	0.79	4002
weighted avg	0.81	0.79	0.79	4002

Looking at the confusion matrix and classification report for the data, the model predicts wildfire occurrences extremely well, with a 10.04% missed-fire rate. But, the model struggles to predict negative occurrences, with a 31.48% false-alarm rate. However, for the scope of our model, this is acceptable due to the critical nature of missing fires compared to false alarms when it concerns business or policy decisions for wildfire mitigation.

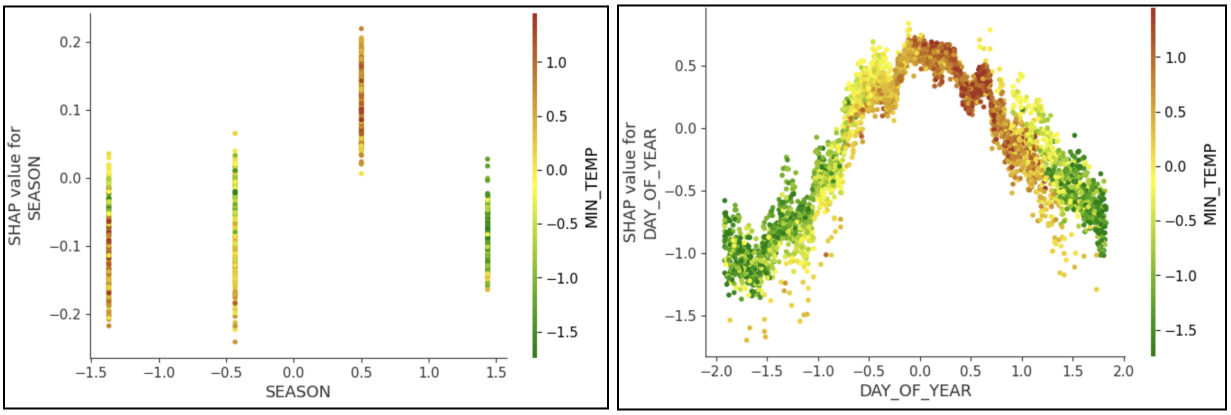
Understanding Model Decision-Making

To understand the XG-Boost model and how its predictors impact its decision-making for wildfire predictions, SHapley Additive exPlanations (SHAP) was utilized to visualize and understand feature importance. The summary plot (left) shows the impact of lower or higher feature values on the model through shape and color, where denser shapes and color determine whether a feature value impacted negative or positive fire predictions. The decision plot (right) shows how each feature affects the model's decisions, and the magnitude each feature has on negative (red) or positive (green) predictions.





For a deeper understanding of how different features are related to each other and their impacts on model decisions, SHAP plots for features were included as well. On the left, season and minimum temperature were plotted, and the color and SHAP values indicate that summer and fall (encoded) have a high impact on positive wildfire predictions. On the right, the day of the year and minimum temperature were plotted, revealing that lower temperatures during the early and late days of the year heavily influence negative predictions, while higher minimum temperatures in middle to late-middle portions of the year heavily impact positive predictions. These insights highlight key wildfire predictors and reinforce the practical applications of predictive modeling in risk management and policy development.



Recommendations

Using the wildfire prediction model in conjunction with weather forecasts as input data allows for the prediction of wildfire occurrences, which enables predictive wildfire modeling with broad applications across industries, government agencies, and urban planning efforts. Proactive planning and decision-making before these fires occur can significantly reduce loss of life and property damage during wildfire events. For wildfire resource allocation, emergency services and firefighters can proactively prepare and deploy resources to high-risk areas to reduce response times. Funding can be strategically directed toward these high-risk areas as well. Government agencies can implement stricter fire safety regulations in high-risk areas while enhancing evacuation planning and fire-resistant measures. To assess property damage and loss, insurance companies can use these predictions to adjust premiums and risk

assessments for properties in fire-prone areas. Aligning insurance rates with fire risk exposure protects insurers from unexpected losses and also increases financial protection for the insured.

Environmental conservation can be enhanced by incorporating wildfire risk assessments into carbon credit and conservation programs that focus on targeted fire prevention strategies - such as managing vegetation and enforcing controlled burns. This lowers the devastation caused by wildfires and can in turn mitigate pollution, environmental degradation, and deaths that accompanies these catastrophes. In addition, water and farming are critical resources for California, and lowering the destructive potential of wildfires helps preserve these resources and the ecosystems surrounding them.

These benefits could be further amplified if real-time prediction became viable, enabling more immediate and effective wildfire response. To enhance the model for the potential of such predictions, additional features such as topographical data and human activity data can be incorporated for the model to better understand factors that contribute to wildfire risk. This can improve performance and potentially scale the model to predict immediate wildfire risk, allowing for improved wildfire emergency response time and resources.

Conclusion

To conclude, devastating effects of wildfires seen in the recent LA wildfires can be minimized by effective predictive modeling for wildfire occurrences. This model helps inform businesses, policymakers, and community leaders to leverage predictive wildfire modeling and make informed decisions that minimize economic disruption, enhance public safety, and promote long-term climate resilience. By leveraging predictive modeling techniques like the one outlined in this report, businesses and policymakers can further shift from reactive wildfire management to proactive prevention - reducing economic losses, protecting communities, saving lives, and preserving natural ecosystems. As wildfires continue to pose a growing threat, integrating data-driven models into decision-making processes will be essential for building a more resilient and sustainable future.

Citations

California Ocean Protection Council. (2025, February). From ashes to action: Wildfire impacts on California's coast and ocean health. California Ocean Protection Council. https://opc.ca.gov/2025/02/from-ashes-to-action-wildfire-impacts-on-californias-coast-and-ocean-health/

UCLA Anderson Forecast. (n.d.). *Economic impact of Los Angeles wildfires*. UCLA Anderson School of Management.https://www.anderson.ucla.edu/about/centers/ucla-anderson-forecast/economic-impact-los-angeles-wildfires

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