Wildfire Risk Prediction System

A Machine Learning Approach to Assess Wildfire Risks Based on Geographic, Temporal, and Environmental Features

Problem Statement

Wildfires pose a significant threat to ecosystems, human safety, and infrastructure, particularly in regions prone to high temperatures and prolonged droughts. This report presents the development and deployment of a Wildfire Risk Prediction System that leverages machine learning techniques to predict the likelihood of high or low wildfire risks in specific regions.

The system utilizes historical wildfire data and various factors such as location, fire cause, and environmental features to provide real-time risk predictions. The model is integrated into an interactive web application, enhancing wildfire preparedness and resource allocation.

Project Overview

The primary objective of this project is to develop a Supervised Learning Binary Classification Model that can predict wildfire risk, categorized as either Low or High for different regions in the United States based on geographic, temporal, and environmental features. This project leverages the comprehensive FPA-FOD (Fire Program Analysis - Fire Occurrence Dataset), which contains 1.88 million wildfire records from 1992 to 2015, to build a machine learning-based **Wildfire Risk Prediction System**. The model is integrated into an interactive Streamlit application that allows users to input key parameters, including state, city, and cause of fire, to receive real-time risk predictions. By providing actionable insights, this tool aims to enhance wildfire preparedness, optimize resource deployment, and support data-driven policy-making.

Project Workflow

1. **Data Exploration & Preprocessing:**
    a. Extracted relevant attributes from SQLite table Fires. Cleaned, transformed, and merged data to build a refined modeling dataset.
2. **Exploratory Data Analysis (EDA):**
    a. Identified spatio-temporal trends, fire size distributions, and cause-based insights.
3. **Feature Engineering:**
    a. Converted DOY to month, simplified fire causes, categorized fire sizes, and created new features for better predictive power.
4. **Model Development:**
    a. Trained multiple classification models including Logistic Regression, Gradient Boosting, and AdaBoost with RandomizedSearchCV for hyperparameter tuning.
5. **Binary Risk Classifier:**
    a. Designed a risk prediction framework that classifies fire instances into ‘High Risk’ vs. ‘Low Risk’ zones based on historical trends and environmental inputs.
6. **End-to-End Integration:**
    a. Developed a lightweight application that accepts state, city, and fire cause as inputs and outputs wildfire risk using the trained model.

This project demonstrates the potential of historical data-driven decision-making in supporting wildfire management, resource allocation, and early warning systems.

By surfacing interpretable and actionable insights, the model serves as a foundational step toward smarter wildfire preparedness.

Data Exploration & Preprocessing

The primary data source for this project is the Fire Program Analysis - Fire Occurrence Database (FPA FOD), which contains over 1.88 million geo-referenced wildfire records in the United States from 1992 to 2015.

This is an overview of some of the important columns in the Fires table, which holds key information on each wildfire incident.

Column Name	Description
FIRE_YEAR	The year the fire was discovered or confirmed to exist.
STAT_CAUSE_DESCR	Description of the fire's statistical cause.
FIRE_SIZE	Estimated acres burned by the fire.
FIRE_SIZE_CLASS	Classification of fire size based on acres burned (e.g., A: 0.25 acres, B: 0.26-9.9 acres, C: 10-99.9 acres).
LATITUDE	Latitude (NAD83) of the fire's location.
LONGITUDE	Longitude (NAD83) of the fire's location.
OWNER_DESCR	Description of the primary landowner or entity.
STATE	The two-letter state code for where the fire occurred.
FIRE_CODE	Code used within the interagency wildland fire community to track suppression costs.
FIRE_NAME	The name of the fire incident.
DISCOVERY_DATE	The exact date the fire was discovered.
DISCOVERY_DOY	Day of the year when the fire was discovered.
DISCOVERY_TIME	Time of day when the fire was first discovered.

Key steps for data preprocessing include:

- Extracted relevant attributes from the Fires table.
- Dropped columns with excessive missing values (>90% null values).
- Dropped duplicates and kept only selected features.
- Kept only selected columns/features
- Dropped rows with null values in selected columns
- Added month of year column from DISCOVERY_DATE.

Exploratory Data Analysis (EDA)

In this step, we perform an in-depth exploratory data analysis to uncover key patterns, trends, and insights from the wildfire dataset. The goal of this analysis is to develop a thorough understanding of the factors

contributing to wildfire occurrences, their distribution across time and geography, and the characteristics of different fire classes.

The EDA process is structured to examine the data across the following dimensions:

- **Temporal Trends:** Analyzing how wildfire frequency and intensity vary over the years across US states to identify recurring patterns.
- **Geographical Distribution:** Mapping fire occurrences across different states and regions to detect high-risk zones and regional disparities.
- **Cause Analysis:** Understanding the distribution of wildfire causes to highlight predominant ignition sources.
- **Class Imbalance Exploration:** Evaluating the distribution of fire size classes to assess imbalance, which has implications for model training.
- **Feature Correlation:** Investigating relationships between numerical features to identify potential multicollinearity and influential predictors.

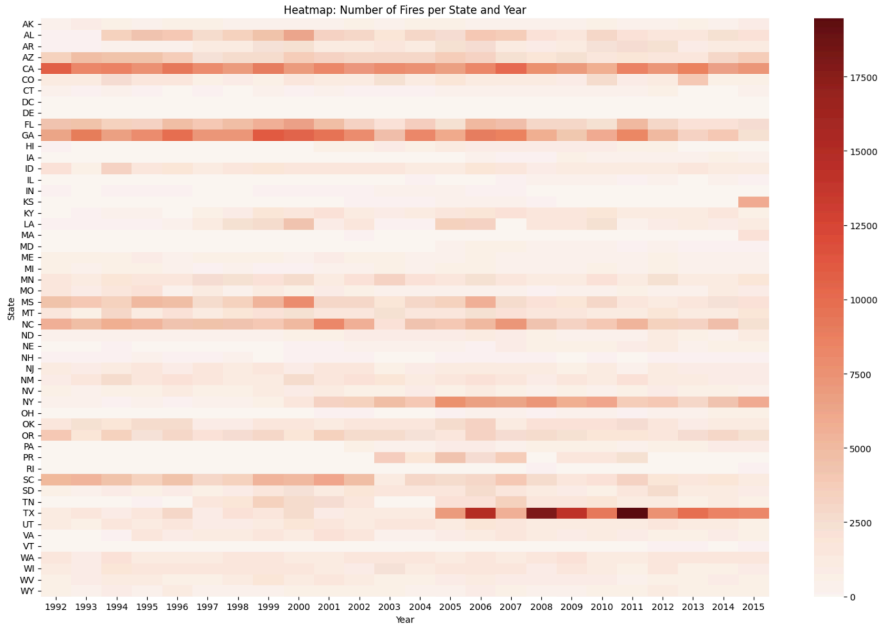
Wildfire Frequency & Intensity Across States and Time

Where and when are wildfires most common?

In this section, we analyze the frequency and intensity of wildfires across different states and years. This analysis helps us understand the geographical and temporal patterns of wildfires.

Heatmap of Fire Counts by State and Year: I used a heatmap to visualize the number of wildfires occurring in each state over the years. This heatmap allows us to identify regions with frequent wildfire occurrences and track changes in the intensity of wildfires over time.

- **Groupby Operation:** The data is grouped by STATE and FIRE_YEAR to count the number of wildfires in each state for each year.
- **Visualization:** A heatmap is generated using Seaborn's heatmap function, where the x-axis represents the years, the y-axis represents states, and the intensity of color represents the number of fires. The color gradient with darker shades indicates higher numbers of wildfires.



Geographic Patterns

Are wildfires concentrated in specific regions?

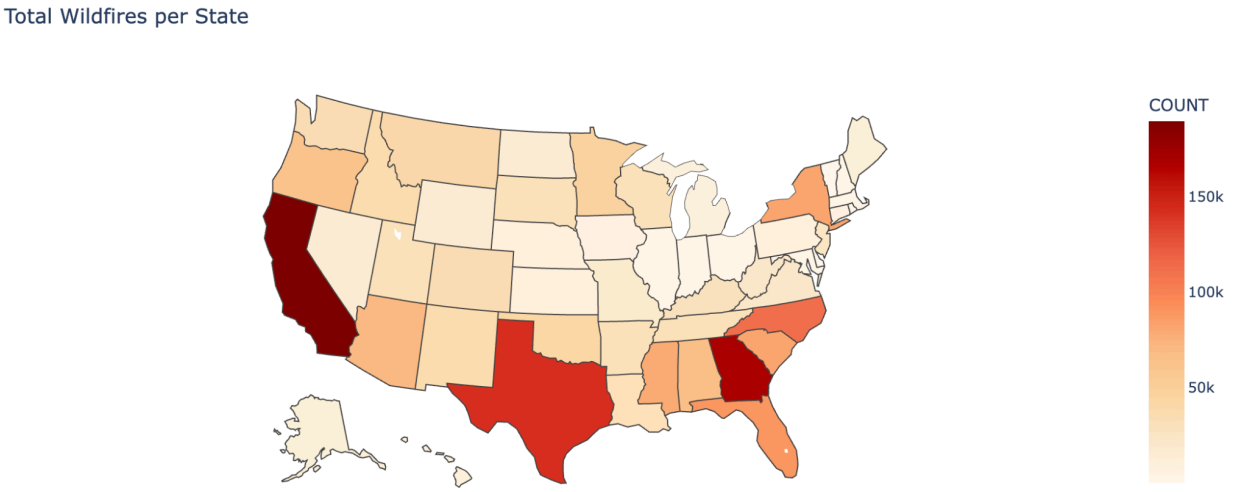
This section focuses on visualizing the spatial distribution of wildfires across the United States to identify regional hotspots. Understanding where wildfires occur most frequently can inform strategic resource allocation and policy planning.

These geospatial visualizations reveal strong regional patterns in wildfire activity, particularly in western and southeastern states, and help establish geographic baselines for further modeling and risk prediction.

Total Fires per State Map:

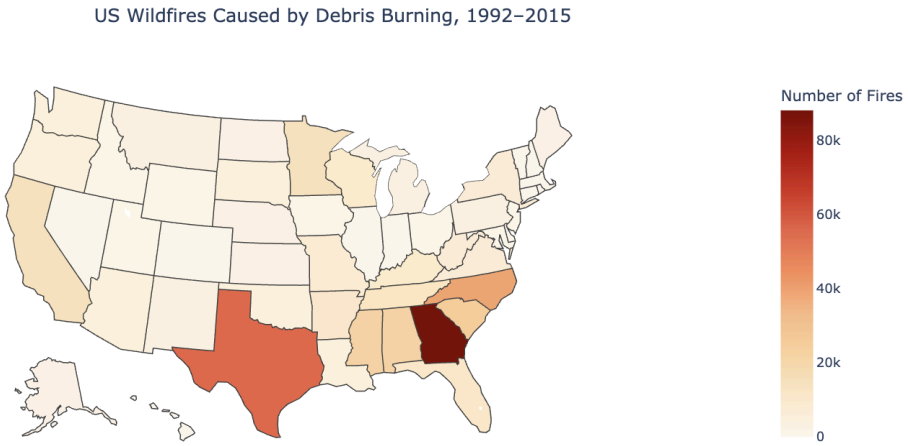
I have plotted a choropleth map representing the total number of wildfires per state from 1992 to 2015. This provides a high-level overview of fire-prone regions.

- The map uses a color gradient to represent the number of fires, with darker shades indicating a higher number of incidents.
- The plot is generated using Plotly's interactive choropleth function and scoped to the United States for clarity.



Cause-Specific Choropleth Map:

Further to analyze how specific causes contribute to wildfire distributions geographically, I filtered the data by a selected cause (Debris Burning in this case which is also one of the biggest causes of wildfire in the US). I then aggregated the number of wildfire incidents by state and plotted a choropleth map. This visual highlights states where debris burning has historically led to the most fire incidents, helping identify cause-specific risk zones.



What Causes the Fires?

Most common and most damaging causes

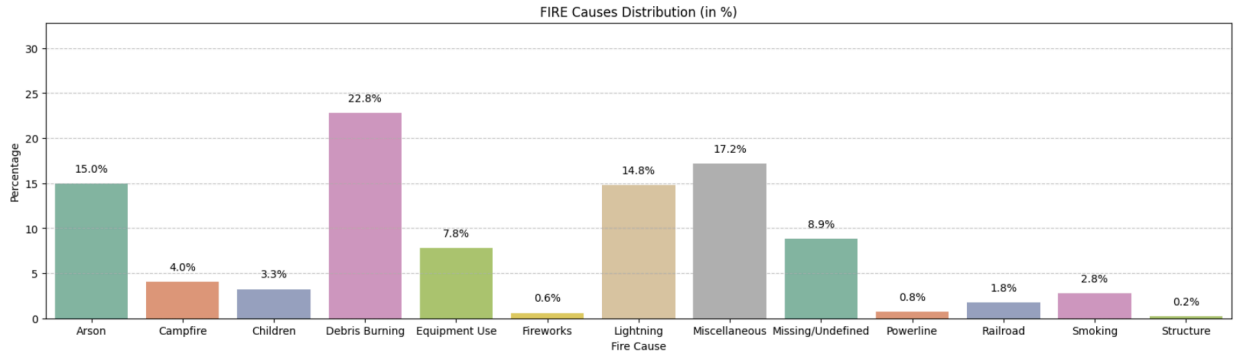
This section explores the underlying causes of wildfires and their relationship to fire risk levels. By analyzing the distribution of fire causes and their associated fire sizes, we gain insights into which causes are most prevalent and which are most damaging.

Cause Distribution Analysis:

I calculated the percentage share of each fire cause category by normalizing their occurrence counts. This breakdown was visualized using a horizontal bar plot where:

- Each bar represents a fire cause category (e.g., Lightning, Debris Burning, Campfire, etc).
- The height of the bar indicates its proportion relative to all recorded wildfires.
- Percentage labels above each bar offer a quick quantitative reference.

This visualization helps in identifying the leading contributors to wildfire incidents over the years.

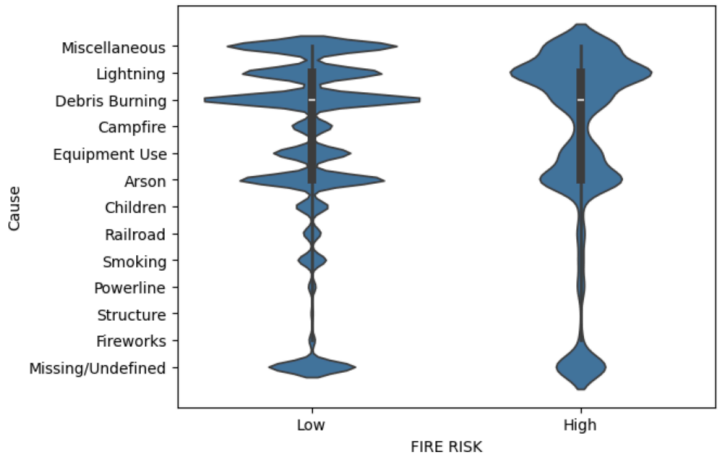


Cause vs. Fire Size (Violin Plot):

To assess the severity of wildfires caused by different origins, I used a violin plot to show the distribution of fire causes against the binary fire risk classification (Low vs. High).

- Each vertical "violin" represents the distribution of a specific cause across the two fire risk levels.
- Wider areas indicate higher density of fires at that risk level for the respective cause.

This plot helps distinguish causes that tend to result in more severe wildfires, aiding targeted prevention strategies.



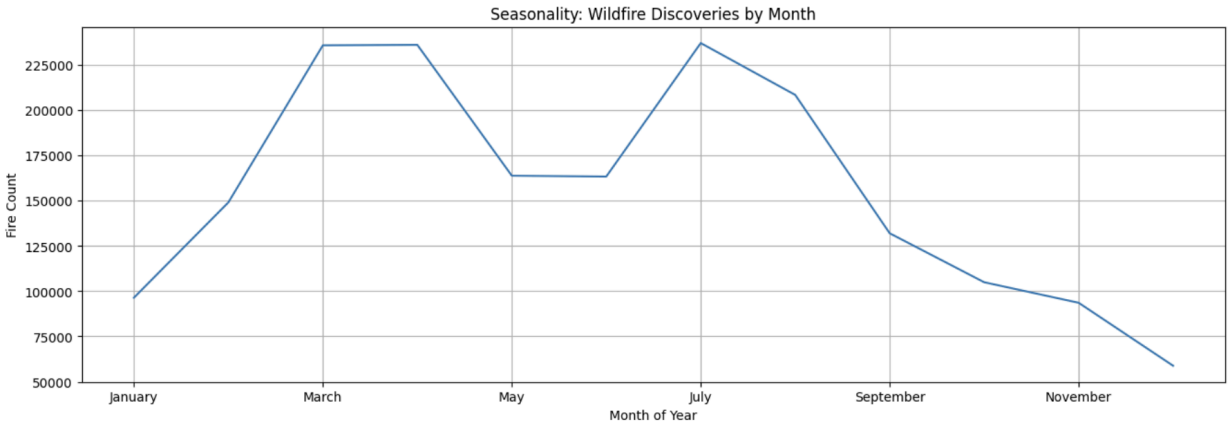
Seasonality of Fires

Do fires spike in specific months?

In this section, we investigate the seasonal trends of wildfire discoveries to identify months with higher wildfire activity. Understanding temporal seasonality is crucial for proactive resource allocation and fire prevention strategies.

- **Month-wise Grouping:** I grouped the data by the month in which each fire was discovered. Months were explicitly ordered from January to December to maintain temporal consistency in the visualization.
- **Trend Analysis:** The resulting monthly counts of wildfires were plotted to observe seasonal fluctuations.
- **Visualization:** A line plot presents the monthly distribution of wildfires, allowing us to visually identify peak fire months and analyze seasonal trends over the years.

This analysis provides valuable insights into the time periods most susceptible to wildfires, which can aid agencies in targeted prevention and mitigation planning.



Class Imbalance in Target

Are fire size classes balanced?

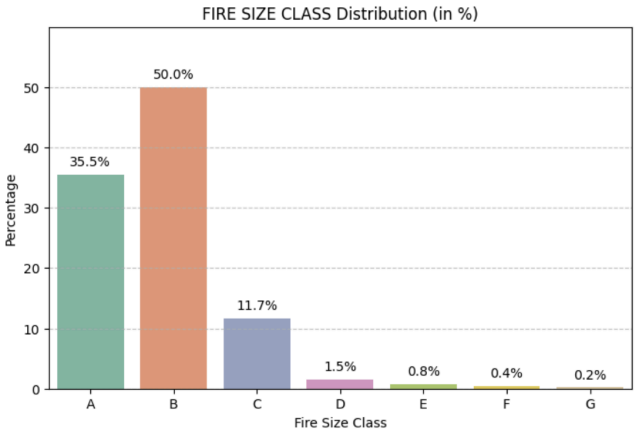
In this section, we analyze the distribution of the FIRE_SIZE_CLASS variable, our original target for classification, to assess whether the data is balanced across fire size categories.

Distribution Analysis

To begin, I plotted the percentage distribution of each fire size class (A–G):

The FIRE_SIZE_CLASS field categorizes wildfires based on their fire size covering acres of land:

- Class A: < 0.25 acres
- Class B: 0.25 – 9.9 acres
- Class C: 10 – 99.9 acres
- Class D: 100 – 299.9 acres
- Class E: 300 – 999.9 acres
- Class F: 1000 – 4999.9 acres
- Class G: 5000+ acres



The resulting bar plot clearly shows a significant class imbalance:

- **Classes A, B, and C** together represent the vast majority of records, accounting for over 90% of all wildfires.
- **Classes D through G** are severely underrepresented, making up less than 10% combined.

This skewed distribution presents challenges for training machine learning models:

- Multi-class classifiers tend to favor majority classes, leading to poor predictive performance on rare but critical large fires.
- The minority classes (D–G), while fewer in number, represent the most dangerous and resource-intensive incidents.

How to balance the target?

To address the imbalance and improve model interpretability and reliability, let's reframe the classification problem as a binary task:

- **Low Risk** (Class A, B, C): Represents smaller, more frequent fires.
- **High Risk** (Class D, E, F, G): Captures larger, more impactful fires.

This transformation is critical for the following reasons:

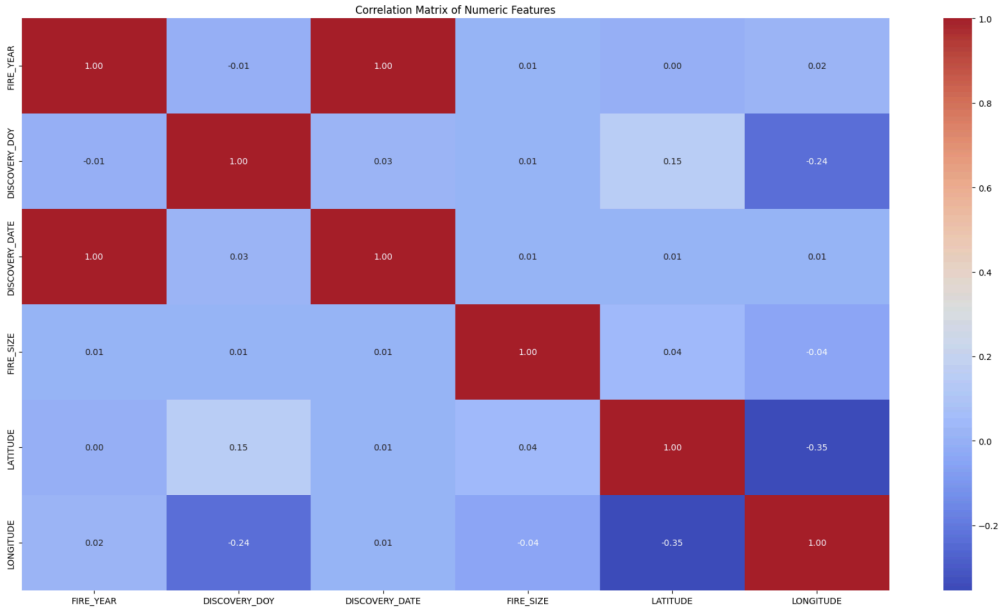
- **Better Class Balance:** Combining categories helps reduce the severity of imbalance, making the classification task more manageable for ML algorithms.
- **Operational Relevance:** Emergency services, forestry departments, and disaster management teams are primarily concerned with early identification of high-risk fires. This binary perspective aligns better with real-world response priorities.
- **Improved Model Performance:** Binary classification often leads to higher precision and recall on the minority class, especially when using metrics like F1-score tailored for imbalanced datasets.
- **Simplified Decision-Making:** A binary risk label provides a clearer decision signal for deploying mitigation efforts.

This final transformation lays the foundation for a focused and effective binary classification model aimed at predicting the likelihood of a wildfire being high-risk, which is the core objective of this project.

Feature Correlation

Which features are most related to fire size?

In this section, we explore the relationships between different features and fire size to understand which factors most strongly influence the size of wildfires. Understanding these relationships is essential for building more accurate predictive models, as it helps us identify key features that significantly impact fire size. Feature correlation analysis aids in selecting the most relevant features for our model, while also providing insights into how we can better allocate resources and improve fire prediction systems.

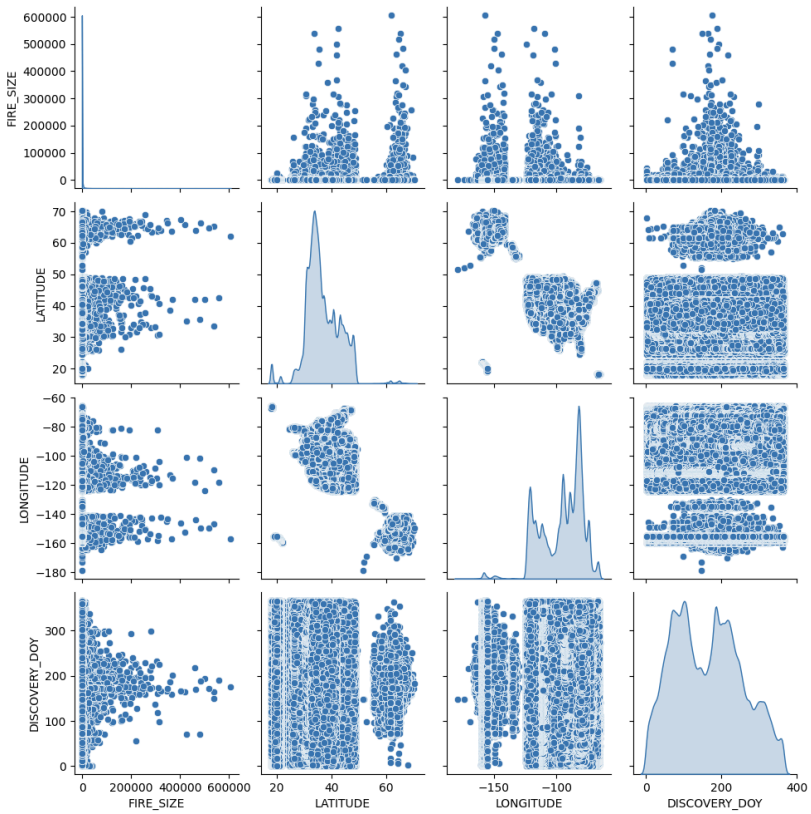


Correlation Heatmap Insights:

The correlation matrix provides insights into how different features related to fire data are interrelated. The color gradient interpretation shows blue representing negative correlations and red representing positive correlations, with darker shades indicating stronger correlations.

- FIRE_YEAR has strong positive correlation with DISCOVERY_DATE (1.00), indicating trends over the years. They're functionally redundant from a modeling perspective and one can be dropped without any loss of information.
- DISCOVERY_DOY has strong positive correlation with LATITUDE suggesting a geospatial seasonality in fire behavior.
- LATITUDE and LONGITUDE have a strong negative correlation (-0.35), indicating a geographical trend where higher latitudes correspond to lower longitudes.
- FIRE_SIZE has moderate positive correlation with LATITUDE implies that as latitude increases, the likelihood of larger fire sizes also increases.

The correlation analysis reveals both temporal and spatial patterns in wildfire behavior across the U.S. The moderate positive correlations between FIRE_YEAR and discovery-related features suggest evolving wildfire trends over time, possibly influenced by climate shifts, land management practices, or reporting accuracy. The strong negative correlation between LATITUDE and LONGITUDE highlights distinct geographical clustering, while the positive correlation between LATITUDE and DISCOVERY_DOY points to seasonal geospatial dynamics, fires in higher latitudes tend to be discovered later in the year. This aligns with regional characteristics like northern areas such as the Northwest and Northern Rockies typically contain denser forests and more combustible biomass, resulting in larger, harder-to-control fires. In contrast, southern regions often experience faster-moving, lower-intensity fires in grasslands or scrub environments. These insights underscore the importance of incorporating both spatial and temporal features in predictive wildfire models to capture region-specific fire behavior and improve early warning systems.



Pair Plot Insights:

The diagonal histograms in the pair plot provide valuable insights into the distribution of key wildfire features. FIRE_SIZE exhibits a heavily right-skewed distribution, with the majority of fires being relatively small and only a few extreme outliers reaching sizes up to 600,000 acres. LATITUDE and LONGITUDE distributions reflect the broad geographic spread of fire incidents across the U.S., with most events occurring between 30°–50° N latitude and -130° to -110° W longitude, consistent with wildfire-prone regions like the Western U.S. DISCOVERY_DOY spans nearly the entire calendar year, indicating that wildfires can ignite year-round, though denser clusters in certain parts of the plot may hint at seasonal peaks. Together, these distributions reinforce both the spatial breadth and temporal persistence of wildfire activity.

The off-diagonal scatter plots illustrate the pairwise relationships between key wildfire variables. The plots between FIRE_SIZE and each of the spatial and temporal features, LATITUDE, LONGITUDE, and DISCOVERY_DOY, reveal widely scattered data points, indicating no strong linear correlations. This suggests that fire size alone cannot be directly predicted based on geographic coordinates or discovery timing, at least not in a simple linear fashion. However, the LATITUDE vs. LONGITUDE plot displays clear clustering, which points to distinct geographic regions where wildfires are more frequent, likely aligned with fire-prone landscapes such as the Western U.S. Meanwhile, both LATITUDE vs. DISCOVERY_DOY and LONGITUDE vs. DISCOVERY_DOY show minor clustering without strong trends, implying that while there may be regional or seasonal groupings, neither coordinate strongly dictates the timing of fire discovery. These patterns underscore the complexity of wildfire behavior and the need for non-linear modeling techniques to capture hidden relationships.

Based on the correlation and pair plot analyses, LATITUDE, and LONGITUDE emerge as meaningful predictors, capturing temporal and spatial dimensions of wildfire behavior. Additionally, incorporating STATE and STAT_CAUSE_DESCR as categorical predictors can further enhance the model by capturing regional and causative patterns in wildfire occurrences, providing a more comprehensive view of wildfire risk across the U.S. While their individual correlations with FIRE_SIZE are not strong, they likely contribute to more complex, non-linear interactions that can enhance model performance when combined. Including these features, alongside enriched environmental data, would be strategic for building a robust and geographically-aware wildfire prediction model.

Model Building, Training, & Evaluation

To improve the predictive power of the model, several transformations were applied to the dataset:

- **Temporal Features**: Converted the day of the year (DOY) into a more interpretable format, such as month and season.
- **Categorization**: Simplified fire causes into broader categories to reduce noise in predictions.
- **Geographical Encoding**: Encoded state and location-related variables to ensure consistency across the dataset.
- **Normalization**: Scaled latitude and longitude values to bring them into a comparable range.

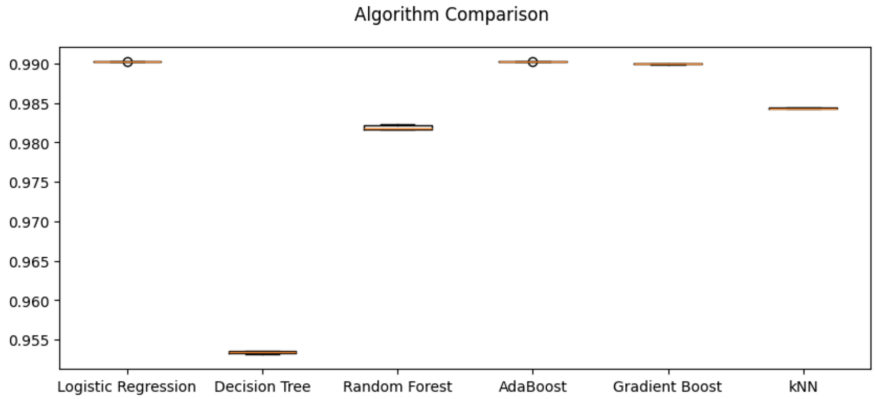
These feature engineering steps helped in making the dataset more suitable for model training and improved its generalization capability.

Model Development

Multiple machine learning models were tested, including:

- Logistic Regression
- Decision Trees
- Random Forest
- AdaBoost
- Gradient Boosting
- kNN

These models were evaluated based on accuracy, precision, recall, and F1-score.



A boxplot is plotted to visualize the performance distribution of all models based on their cross-validation scores.

Along the x-axis, it presents the names of the algorithms being evaluated: Logistic Regression, Decision Tree, Random Forest, AdaBoost, Gradient Boost, and k-Nearest Neighbors (kNN). The y-axis displays the cross-validation scores, spanning a range between lowest 0.955 and highest 0.990.

From the plots, Logistic Regression demonstrates a tight performance range. In contrast, the Decision Tree model has a visibly wider spread, highlighting its susceptibility to variance across different folds. AdaBoost, and Gradient Boost all show high median performance with minimal spread, suggesting not only strong predictive power but also robustness. kNN, on the other hand, exhibits greater variability in performance, which may imply sensitivity to the dataset's distribution or the choice of k.

Overall, the box plot serves as an effective diagnostic tool for comparing model stability and effectiveness. While models like Logistic Regression, AdaBoost, and Gradient Boost emerge as top contenders due to their high and consistent performance, others like Decision Tree and kNN highlight trade-offs between flexibility and reliability.

Based on the performance analysis, **Logistic Regression**, **AdaBoost**, and **Gradient Boost** have been shortlisted for further hyperparameter tuning to identify the most optimal model for wildfire risk predictions.

Hyperparameter Tuning for Shortlisted Models

To improve model performance, hyperparameter tuning is performed on three shortlisted classifiers: Logistic Regression, AdaBoost, and Gradient Boosting. This step uses RandomizedSearchCV with a custom scoring function to efficiently explore combinations of parameters.

- **Logistic Regression:** The C parameter (regularization strength) is tuned.
- **AdaBoost:** The number of estimators and learning rate are optimized, using decision trees as base learners.
- **Gradient Boosting:** Key parameters like number of estimators, learning rate, and tree depth are tuned.

Each model is evaluated using 3-fold cross-validation, and the best parameters along with their corresponding scores are stored for further use in final training and evaluation.

Building model with best parameters

To assess the effectiveness of the tuned shortlisted models, we evaluate its performance on both training and test datasets using key classification metrics: Accuracy, Precision, Recall, and F1 Score. These metrics help provide a balanced view of model performance.

Performance metrics are computed using a custom helper function - evaluate_classification_model, offering a structured and consistent evaluation across datasets. This step ensures the tuned model generalizes well and maintains robust predictive accuracy.

Model Performance comparison and choosing the final model

This section consolidates performance metrics (Accuracy, Precision, Recall, and F1-score) for the three shortlisted and tuned models: Logistic Regression, AdaBoost, and Gradient Boost. The performance is evaluated separately for both the training and test datasets. By comparing these metrics side-by-side, we can make a data-driven decision on which model performs best overall and should be selected as the final model for deployment or further use.

Training performance comparison:

	Logistic Regression Tuned	AdaBoost Tuned	Gradient Boost Tuned
Accuracy	0.971281	0.971281	0.971310
Recall	1.000000	1.000000	0.999977
Precision	0.971281	0.971281	0.971331
F1	0.985431	0.985431	0.985446

Testing performance comparison:

	Logistic Regression Tuned	AdaBoost Tuned	Gradient Boost Tuned
Accuracy	0.971126	0.971126	0.971154
Recall	1.000000	1.000000	0.999963
Precision	0.971126	0.971126	0.971187
F1	0.985352	0.985352	0.985365

Based on the test performance results, the **Gradient Boost** Tuned model emerges as the most suitable choice for final deployment. While its accuracy (0.971154) is only marginally higher than that of Logistic Regression and AdaBoost, it consistently outperforms the others across critical metrics:

- Precision is the highest among all models, indicating better control over false positives.
- F1 Score, which balances precision and recall, is also marginally superior, highlighting its robustness across both dimensions.
- Recall remains nearly perfect (0.999963), only slightly below the others, which is a negligible difference given the gains in precision and F1.

Overall, Gradient Boost Tuned offers the most balanced performance profile. Its ability to maintain high precision without sacrificing recall makes it the most reliable model, especially in contexts where minimizing false positives is as crucial as detecting all true positives.

Final Model Evaluation

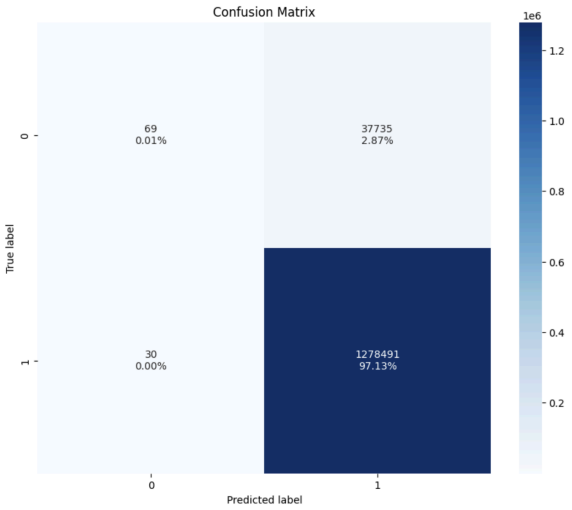
The trained models were evaluated using a **confusion matrix**. Additionally, the models were assessed on:

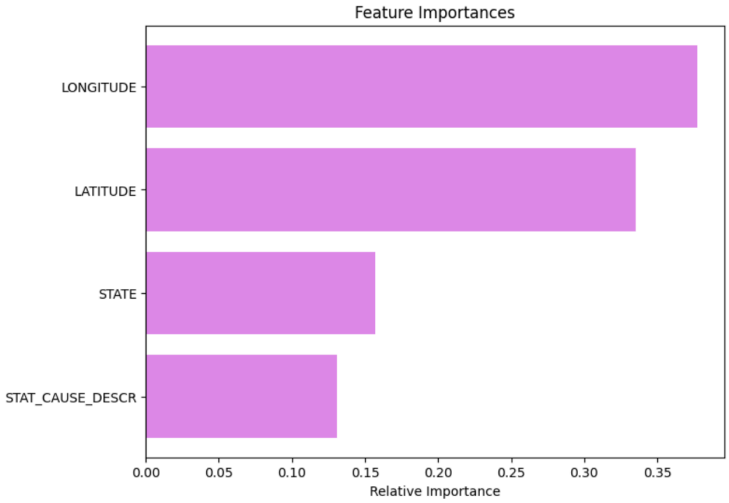
- **Precision**: The proportion of predicted high-risk areas that were indeed high-risk.
- **Recall**: The proportion of true high-risk areas that were correctly predicted.
- **F1-Score**: The balance between precision and recall for high-risk predictions.

The final model achieved an **accuracy of 0.97** for high-risk predictions, demonstrating its effectiveness in identifying regions at risk of wildfires.

```
evaluate_classification_model(final_model, X_test, y_test)
```

	Accuracy	Recall	Precision	F1
0	0.971069	0.999693	0.971352	0.985319





The above bar chart displays features having maximum importance for making accurate low or high risk predictions.

Application Deployment

The predictive model was integrated into a web-based application built with **Streamlit**. The app allows users to input a state, city, and fire cause to predict wildfire risk levels in the selected region.

The model and associated components are saved as .pkl files, allowing for seamless integration with the Streamlit app for efficient and consistent prediction serving.

It loads pre-trained machine learning models and encoders, preprocesses user inputs, predicts wildfire risk using the trained model, and displays the predicted location on an interactive Folium map. The application allows users to select a state and city, choose a likely cause of fire, and receive real-time risk predictions, enhancing wildfire preparedness and resource allocation.

Streamlit code is available in wildfire_app.py and configurations in config.py.

The application can be run using command: **streamlit run wildfire_app.py** which will be hosted on localhost.

Key features of the application:

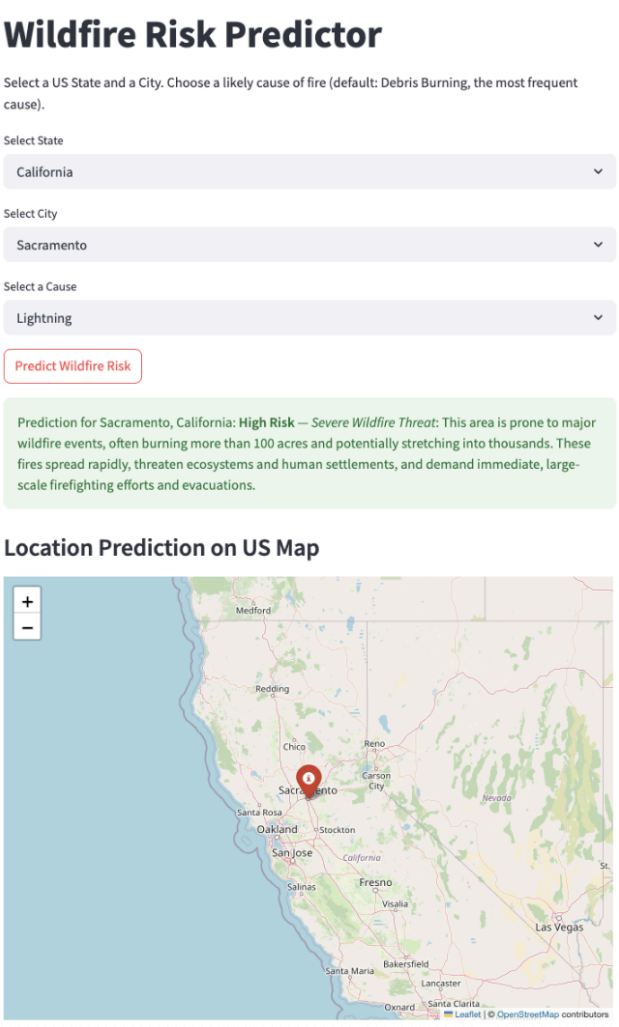
- **Input Interface**: Users select a state, city, and fire cause from dropdown menus.
- **Prediction**: On button click, the app uses the model to predict whether the area is at high or low risk of a wildfire.
- **Map Visualization**: The app displays a folium map with a marker indicating the selected location and its predicted wildfire risk level.
- **Result Display**: The predicted wildfire risk (high/low) is shown to the user.

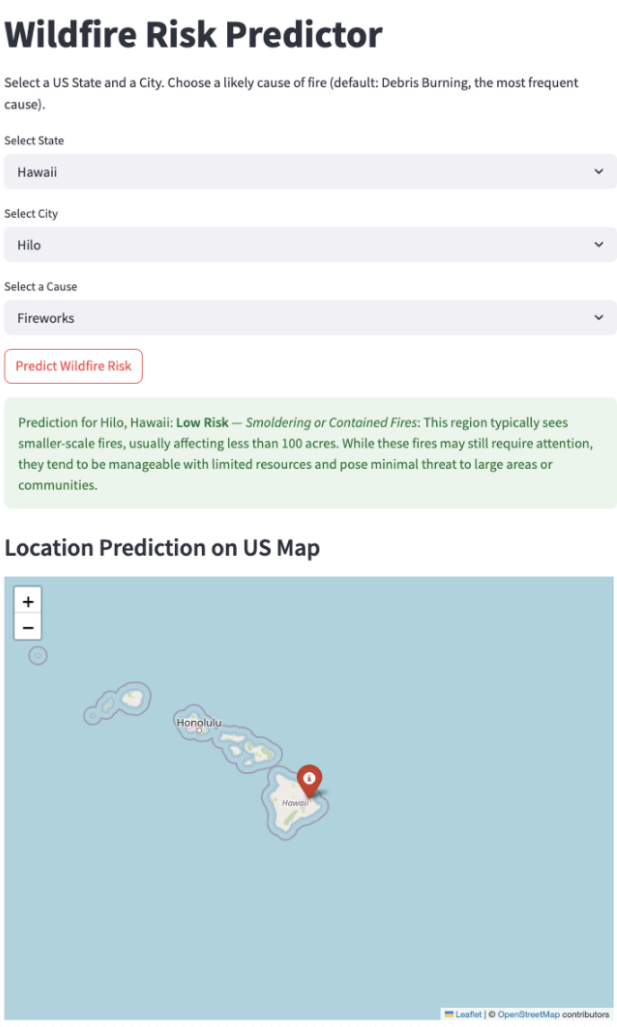
Results

The model's predictions were tested across several geographic regions with different fire causes. For instance, regions with frequent **Debris Burning** had a higher chance of being classified as high-risk, which aligns with historical patterns.

The app provided actionable insights, helping users quickly assess risk levels, which is critical for wildfire preparedness and resource allocation.

The Streamlit application output snapshots are shown below:





Conclusion

This project has successfully developed a machine learning-based Wildfire Risk Prediction System aimed at supporting proactive wildfire mitigation efforts across the United States.

A comprehensive machine learning approach was presented to predict wildfire risk using the historical FPA-FOD (Fire Program Analysis - Fire Occurrence Dataset). By analyzing 1.88 million wildfire records spanning 24 years, built a robust and interpretable framework that classifies fire incidents into **Low Risk** or **High Risk** zones based on geospatial, temporal, and environmental features.

After experimenting with and evaluating multiple classification models, including Logistic Regression, Decision Trees, k-NN, and ensemble methods, we selected **Gradient Boost** as the final model. It demonstrated strong performance in balancing precision and recall for the high-risk class, which is critical in minimizing false negatives in wildfire risk prediction. The model was further optimized using RandomizedSearchCV to address class imbalance and tune hyperparameters effectively.

A lightweight end-to-end application was also built to productionize the final model. This application accepts user inputs such as state, city, and fire cause, and returns a binary wildfire risk classification,

“Low” or “High”, making it operationally viable for integration into early warning systems or public safety dashboards.

Modeling outcomes suggest that using this ML system enables a significantly improved risk awareness layer, compared to no historical trend-based system. It offers the potential to guide resource allocation, pre-position firefighting units, and issue targeted alerts in regions with elevated wildfire risk, ultimately contributing to reduced economic loss, improved public safety, and ecosystem preservation.

The most impactful predictors for wildfire risk classification were found to be:

- LATITUDE and LONGITUDE
- STATE
- STAT_CAUSE_DESCR

These insights can inform not only better fire prevention strategies but also help federal and state agencies prioritize data collection, especially in terms of geolocation accuracy, cause attribution, and temporal data granularity.

By surfacing actionable patterns from two decades of wildfire history, this project lays the foundation for data-driven wildfire resilience strategies, and underscores the value of combining geospatial intelligence with scalable AI solutions.

References

- **Dataset Citation:** Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPA_FOD_20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4
- Kaggle Reference: https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires/data
- Streamlit Documentation: https://docs.streamlit.io/
- Scikit-learn Documentation: https://scikit-learn.org/
- Folium Documentation: https://python-visualization.github.io/folium/

Appendix

- Project is uploaded on github at https://github.com/snehapatil1/wildfire-risk-prediction-ml/
- ML code is in wildfire_app.ipynb.
- Streamlit app code is in wildfire_app.py, with configurations in config.py.
- Model and encoder artifacts are saved as .pkl for deployment.
- This report is included in the repository.
- A video demo is available on YouTube; the link is provided in the repo.