Wildfire Prediction Model

by

Aditya Verma

Bogdan Shmat

Ryan Sha

Sam Zhang

Data collection and initial preprocessing.

Dataset Description

The dataset we used is an SQLite database consisting of a record of wildfires in the United States from 1992 to 2015. This dataset is crucial for the Fire Program Analysis (FPA) system and contains over 1.88 million geo-referenced records, reflecting a total of 140 million acres burned. Key information in the dataset includes identification codes, fire names, the date and time of fire discovery, the cause of the fire, containment details, and geographical data such as latitude and longitude. Data sources include a mix of federal, state, and local fire organizations, ensuring a robust and representative collection of fire incidents. This dataset has undergone transformations to align with the National Wildfire Coordinating Group (NWCG) standards, and redundant records have been identified and removed to enhance the accuracy and reliability of the data.

Data Pre-processing

The data pre-processing for this project involved several steps to ensure the data was clean and ready for analysis. Initially, the size of the downloaded database file was verified to ensure completeness. Tools such as geopandas and pycountry were installed to handle geographic and country data respectively. The primary database connection was established using sqlite3 to access and query the data. A cursory check of the database tables confirmed the successful loading and integrity of the dataset. For analytical purposes, external datasets like naturalearth_lowres were loaded using geopandas to facilitate the visualization of geographic data. The entire pre-processing phase set a solid foundation for subsequent data exploration and machine learning tasks, such as modeling to predict wildfire characteristics based on historical data. This preparatory work ensures that the data analysis is grounded on clean and well-structured data, enabling more accurate and insightful outcomes from the AI model.

Model training and validation.

Model Training and Validation

Model Setup

For this project, logistic regression was chosen due to its simplicity and interpretability, which are crucial for a proof-of-concept model. The model is tasked with predicting the likelihood of large wildfires, defined as categories 'G' or 'H' in the FIRE_SIZE_CLASS data.

Model Training

The model was trained on the preprocessed training dataset X_train_preprocessed, with the binary target y_train_binary, where '1' represents large fires and '0' represents all other categories. The logistic regression was implemented using LogisticRegression from scikit-learn, with an increased max_iter parameter to ensure convergence.

Model Validation

The validation of the model was performed on the test dataset X_test_preprocessed. The model's ability to predict large fires was assessed by examining the probability outputs for the positive class (large fires). These probabilities were then converted to percentages to provide a more interpretable risk assessment metric.

Performance evaluation and result visualization

Performance Evaluation and Result Visualization

Evaluation Metrics

The performance of the logistic regression model was evaluated using the following metrics:

Accuracy: Measures the overall correctness of the model.

Precision and Recall: Specifically for the positive class, these metrics help understand the model's performance in terms of false positives and false negatives.

Area Under the ROC Curve (AUC-ROC): Provides an aggregate measure of performance across all possible classification thresholds.

Visualization of Results

Visualizations were created to aid in the interpretation of the model's predictions:

Histograms of Numerical Features: Before and after preprocessing to show the effect of scaling and imputation.

Probability Distributions: Histograms showing the distribution of predicted probabilities for large fires, helping to visualize the model's confidence in its predictions.

Expanded Visualizations Section

Fire Size Distribution

A bar chart displays the distribution of wildfires by size class, giving insight into the prevalence of different fire sizes.

Large Fires by Cause

This chart identifies the main causes of large wildfires, helping pinpoint major contributing factors.

Temporal Trends in Fire Size

A scatter plot with a regression line shows changes in total fire size over the years, highlighting trends in wildfire severity.

Analysis of Fire Causes Over Time

Line graphs depict the proportion of total wildfire size attributed to lightning and miscellaneous causes, tracking changes over time and identifying prevalent ignition sources.

These visualizations aim to enhance understanding of wildfire dynamics for better management and prevention strategies.

Risk Assessment Output

The model's output included a detailed risk assessment for each county (where available), showing the percentage likelihood of large fires. This output is crucial for stakeholders aiming to prioritize areas for fire prevention and readiness measures.

```
FIPS_NAME  Large_Fire_Risk_Percentage
0              NaN                 0.090092
1              NaN                 0.339467
2     Umatilla County                  0.204310
3              NaN                 0.229836
...            ...                     ...
46556    Lassen County                   0.154336
46557            NaN                 0.264602
46558            NaN                 0.073144
46559     Perry County                   0.060230
```

High-Risk Areas

Counties with a large fire risk percentage exceeding a threshold (e.g., 70%) were highlighted as high-risk areas, enabling targeted intervention strategies.

```
FIPS_NAME  Large_Fire_Risk_Percentage
8114   Siskiyou County                93.371351
8731     Valley County               80.302084
18130    Albany County                88.571394
27180   Trinity County               96.573803
```

Summary

The logistic regression model developed in this project provides a foundational approach for predicting large wildfire risks using historical and environmental data. The model's interpretability and the detailed risk assessments it generates are valuable tools for wildfire management strategies. Further improvements could include experimenting with more complex models and incorporating additional features such as real-time weather data to enhance prediction accuracy.

Data source:
https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires

colabs

https://colab.research.google.com/drive/1n7JAbTtVNvM3nhg63SAPtHeZN1fxUreh?usp=sharing

https://colab.research.google.com/drive/1SYAOTZYNv-dbG0MDsuAnXbL3DCLda2Af?usp=sharing

Github to video/presentation slides:\

https://github.com/ryanbat360/CPTS440