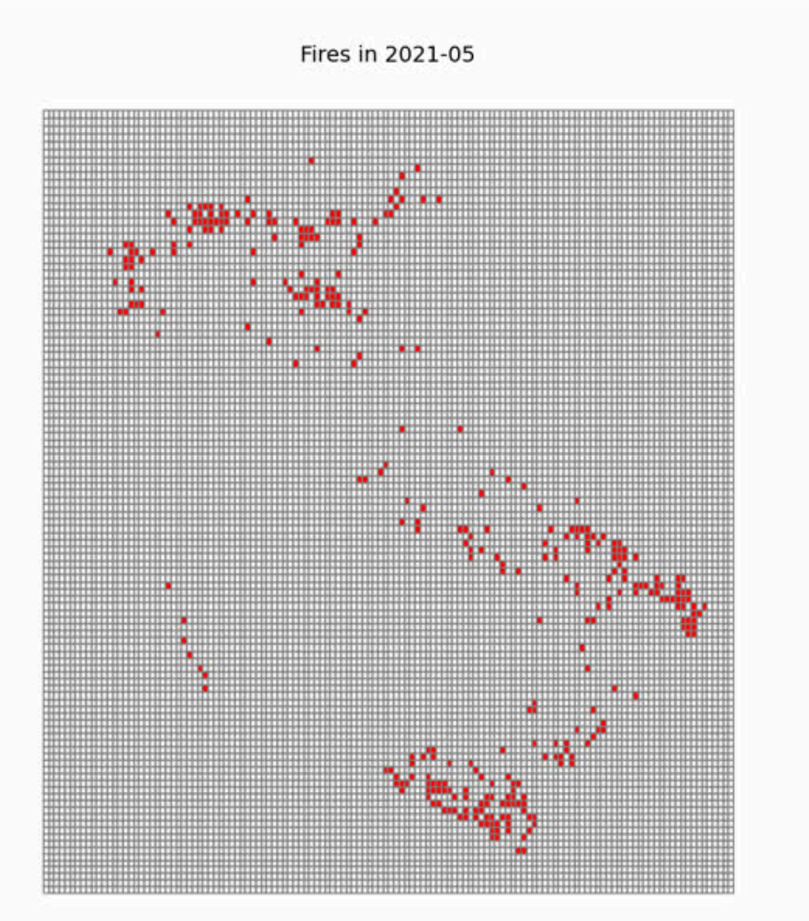
# Wildfire Prediction Model

**Goal**: Building from scratch a minimal model that reasonably predicts wildfires in Italy.

*Sather tower at UC Berkeley during so-called “orange days”*

# Data and Features




## 1 Raw Data

Satellite data from NASA FIRMS, NASA POWER (APIs)

## 2 Cleaning

- Data → GeoData
- Spatial and Temporal Aggregation

## 3 Feature Engineering

**Data Split**
Training (2021-03 → 2023-08)
Validation (2023-09 → 2024-02)
Test (2024-03 → 2025-03)
**Features**
Temperature
Precipitations
Lagged Features

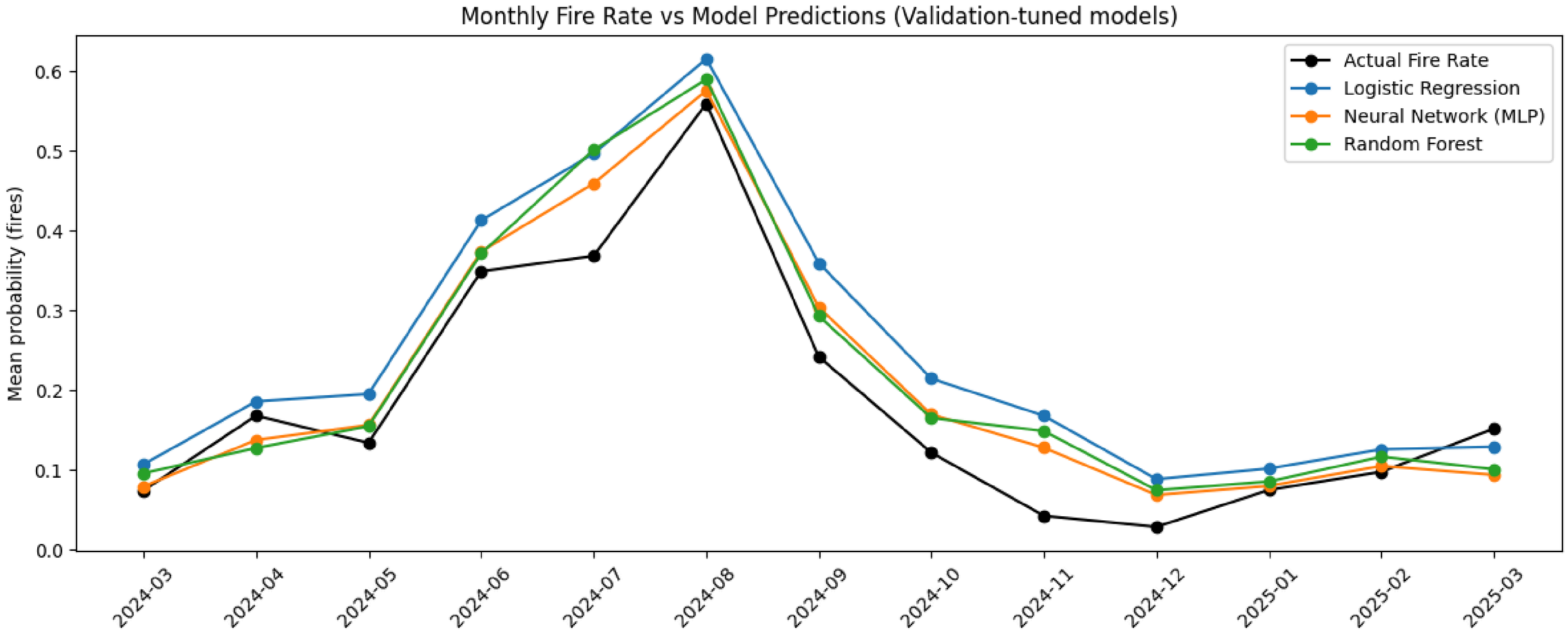
# Sample of 8 random rows from the dataset:

| | cell_id | month | temp2m_mean_c | precip_sum_mm | fire_last_month | lag2 | fire_occurred | data_split |
|---|---|---|---|---|---|---|---|---|
| 3448 | 1513 | 2024-10 | 12.2 | 3.9 | 0 | 0 | 0 | test |
| 45093 | 4370 | 2022-04 | 12.0 | 1.4 | 0 | 0 | 0 | train |
| 70402 | 5804 | 2023-04 | 12.3 | 1.0 | 1 | 2 | 1 | train |
| 98712 | 6995 | 2021-11 | 11.3 | 4.7 | 0 | 0 | 0 | train |
| 120540 | 7829 | 2023-08 | 26.5 | 0.4 | 1 | 2 | 1 | train |
| 208218 | 11817 | 2021-05 | 18.1 | 0.2 | 0 | 0 | 1 | train |
| 69188 | 5716 | 2023-11 | -1.2 | 2.9 | 0 | 0 | 0 | validation |
| 74547 | 6007 | 2023-12 | 8.9 | 0.6 | 0 | 0 | 0 | validation |

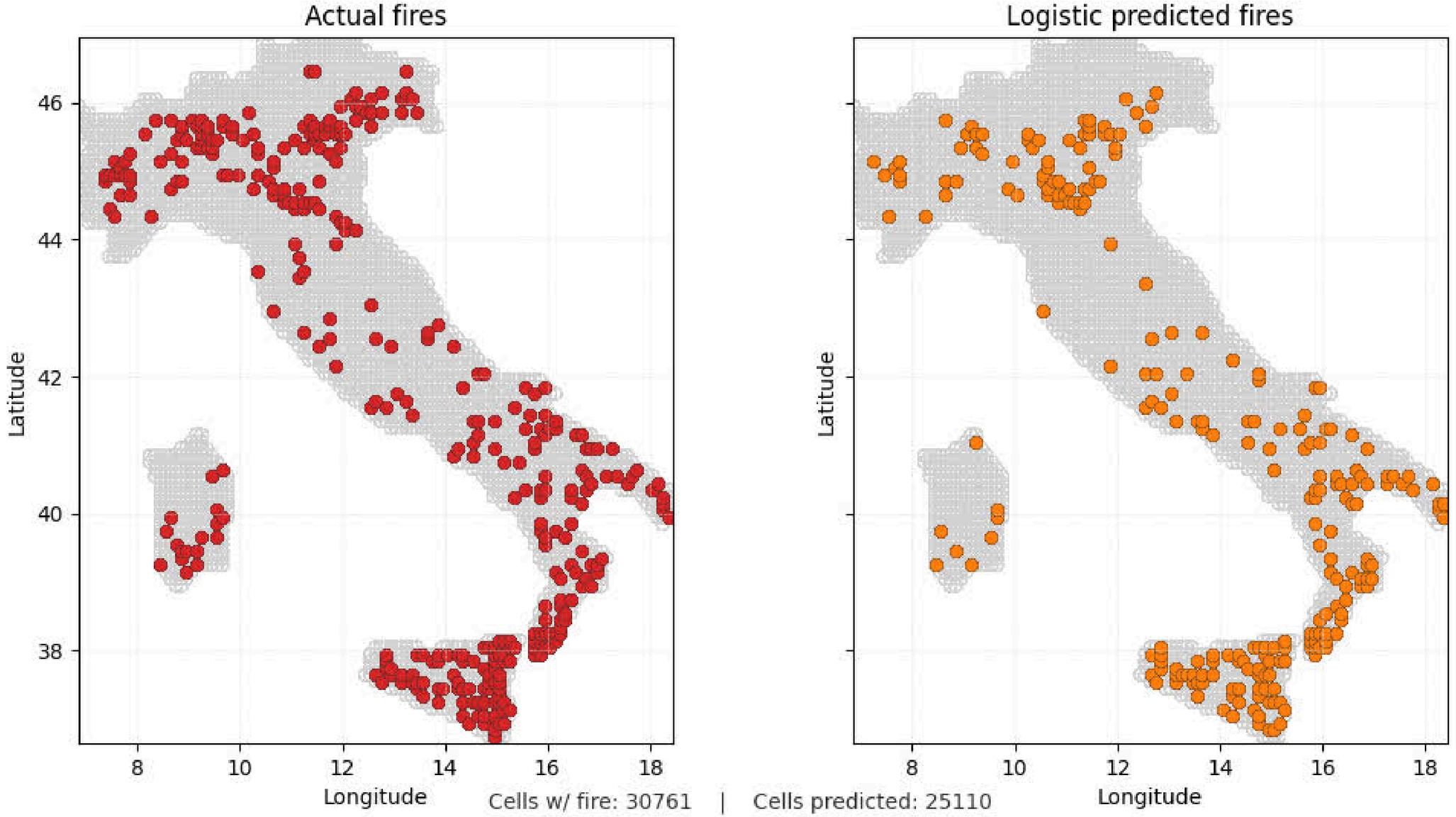
**Target variable to predict**

# Building Models

We trained three scikit-learn pipelines on the monthly cell grid: a class-weighted **logistic regression**, a two-layer **MLPClassifier**, and a 300-tree balanced **random forest** after StandardScaler preprocessing. (Geo)**Pandas** was used for feature prep while **scikit-learn** handled pipelines, metrics, and auto threshold tuning maximising F1 scores.



Month: 2024-04




Cells w/ fire: 30761 | Cells predicted: 25110

# Result & Performances

| model | split | threshold | accuracy | precision | recall | f1 |
|---|---|---|---|---|---|---|
| Logistic Regression | test | 0.61 | 0.911 | 0.804 | 0.764 | 0.783 |
| Neural Network (MLP) | test | 0.54 | 0.910 | 0.814 | 0.742 | 0.776 |
| Random Forest | test | 0.42 | 0.883 | 0.706 | 0.759 | 0.731 |
| Logistic Regression | validation | 0.61 | 0.919 | 0.846 | 0.880 | 0.863 |
| Neural Network (MLP) | validation | 0.54 | 0.920 | 0.859 | 0.866 | 0.863 |
| Random Forest | validation | 0.42 | 0.905 | 0.837 | 0.834 | 0.836 |

---

Even with just four coarse features (temperature, rain, and short lags), the models hit 91 % accuracy and **balanced precision/recall**, already flagging most risky cells. That makes this pipeline a **credible first draft for an early-warning tool** which is ‘lightweight’ and fully scalable yet actionable.