WILDFIRE SIZE PREDICTOR

Swathi Munikoti

Image Credit: [Forest Fire - WallpaperBat]

PROBLEM STATEMENT

Build a model to predict wildfire size in California using historical, weather, and geospatial data.

AUDIENCE

Fire Fighters 🧑‍🚒

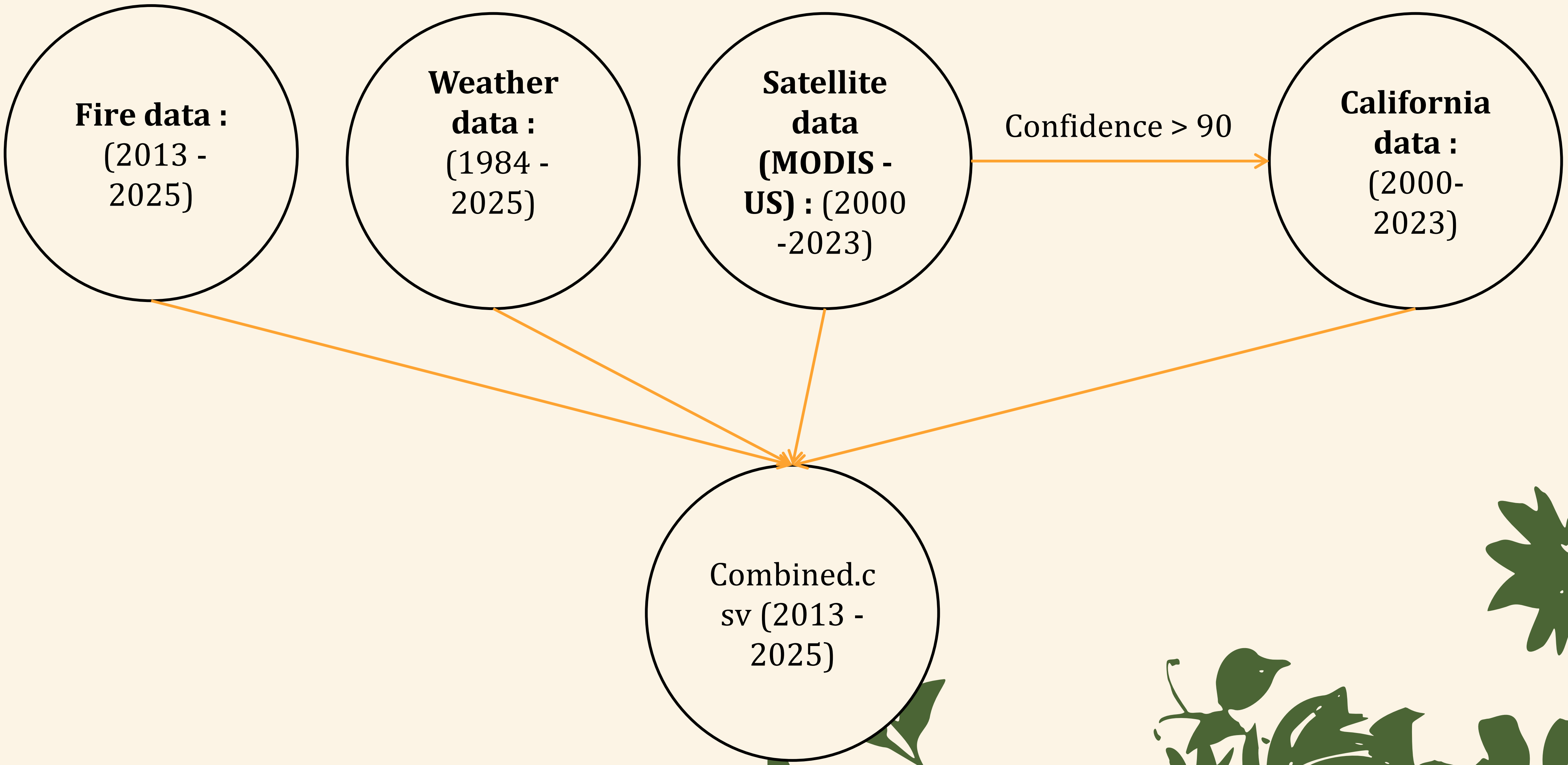
DATA SOURCES

- 🌤️ **Weather and Environmental Data (**Zenodo**)**
https://zenodo.org/records/14712845 - Thanks to **Andranique**

- 🔥 **Wildfire Incident Records** (CAL FIRE)
https://www.fire.ca.gov/incidents - Thanks to **Matt**

- 🛰️ **MODIS Satellite Fire Data** (NASA FIRMS)
https://firms.modaps.eosdis.nasa.gov/country/

GATHERING DATA

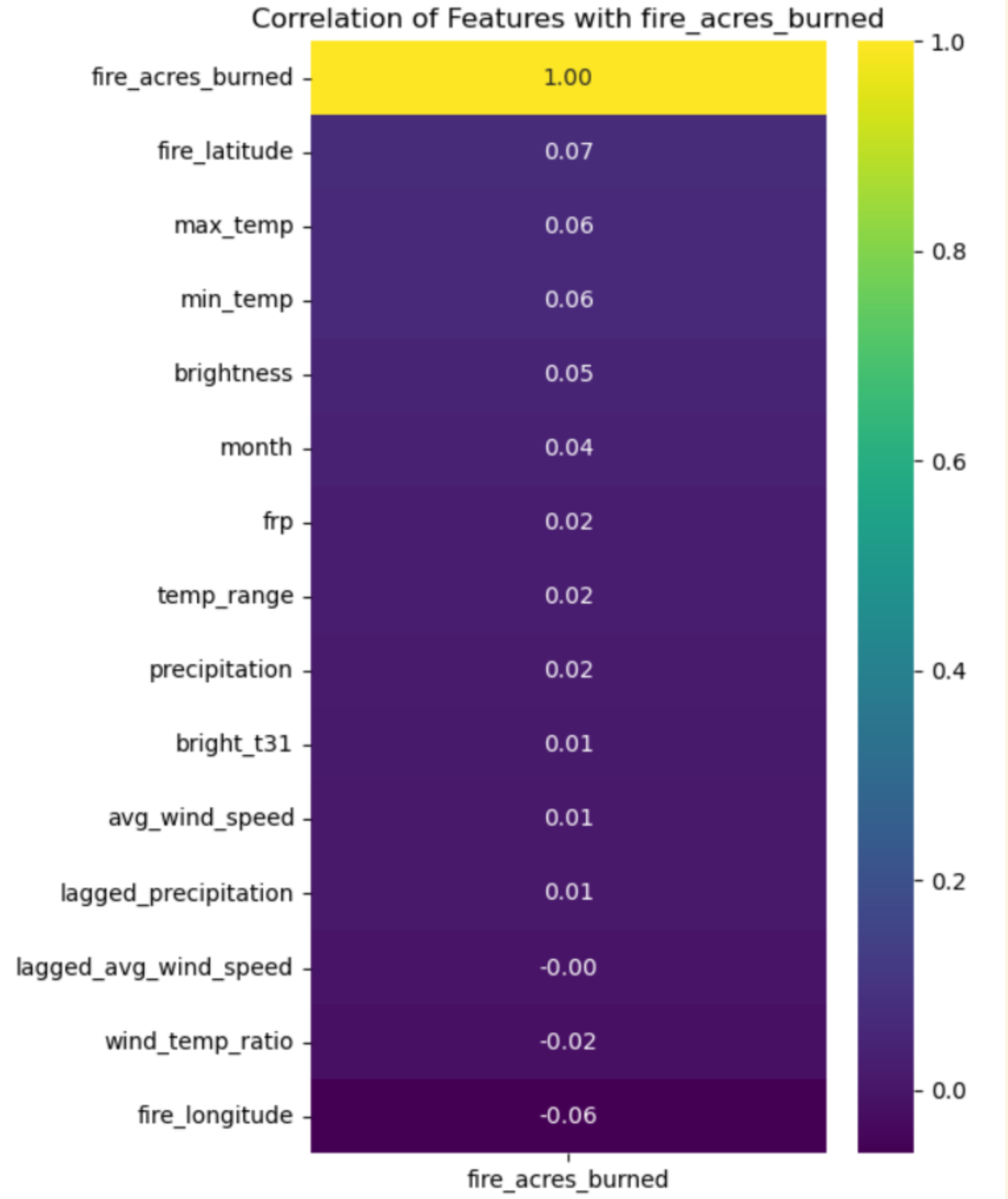


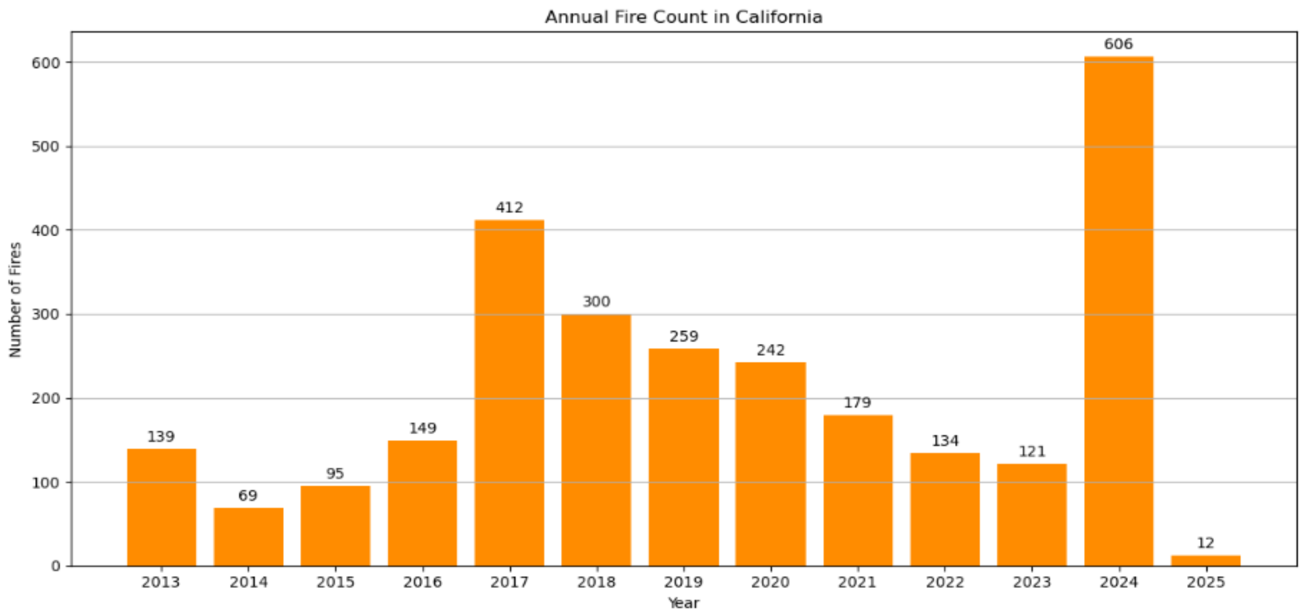
CLEANING DATA

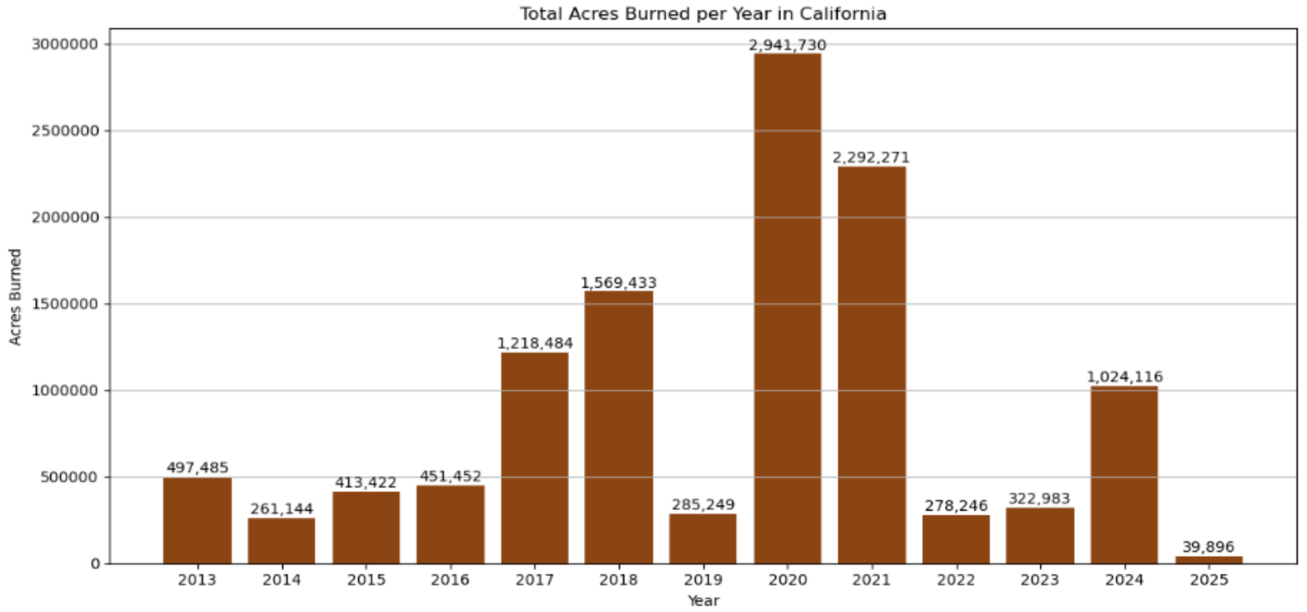
- **Coordinate Validation**: Ensured latitude (32.5–42.0) and longitude (–124.5 to –114.0) fall within California; removed invalid entries (e.g., 5487.0, -1191414610.0).
- **County Mapping:** Used California_Counties.geojson to assign each wildfire to the correct county and removed out-of-state records.
- **Missing Data Handling**: Filled missing values (especially for 2024–2025) using iterative imputation.
- **Result:** Reduced from **2,817 to 2,706** clean and valid fire records.

EXPLORATORY DATA ANALYSIS

Fire size may depend on nonlinear relationships.
Most features exhibit a correlation coefficient below 0.1, indicating weak or no linear relationship with the target variable, fire_acres_burned.







1. **Number of Fires**

•Peaked in **2017**, dipped after **2018**, with a spike in **2024** .

2. **Total Acres Burned**

•**2020** saw the most damage: **2.9M+ acres**, despite fewer fires.
•High-damage years: **2018**, **2020**, **2021**.

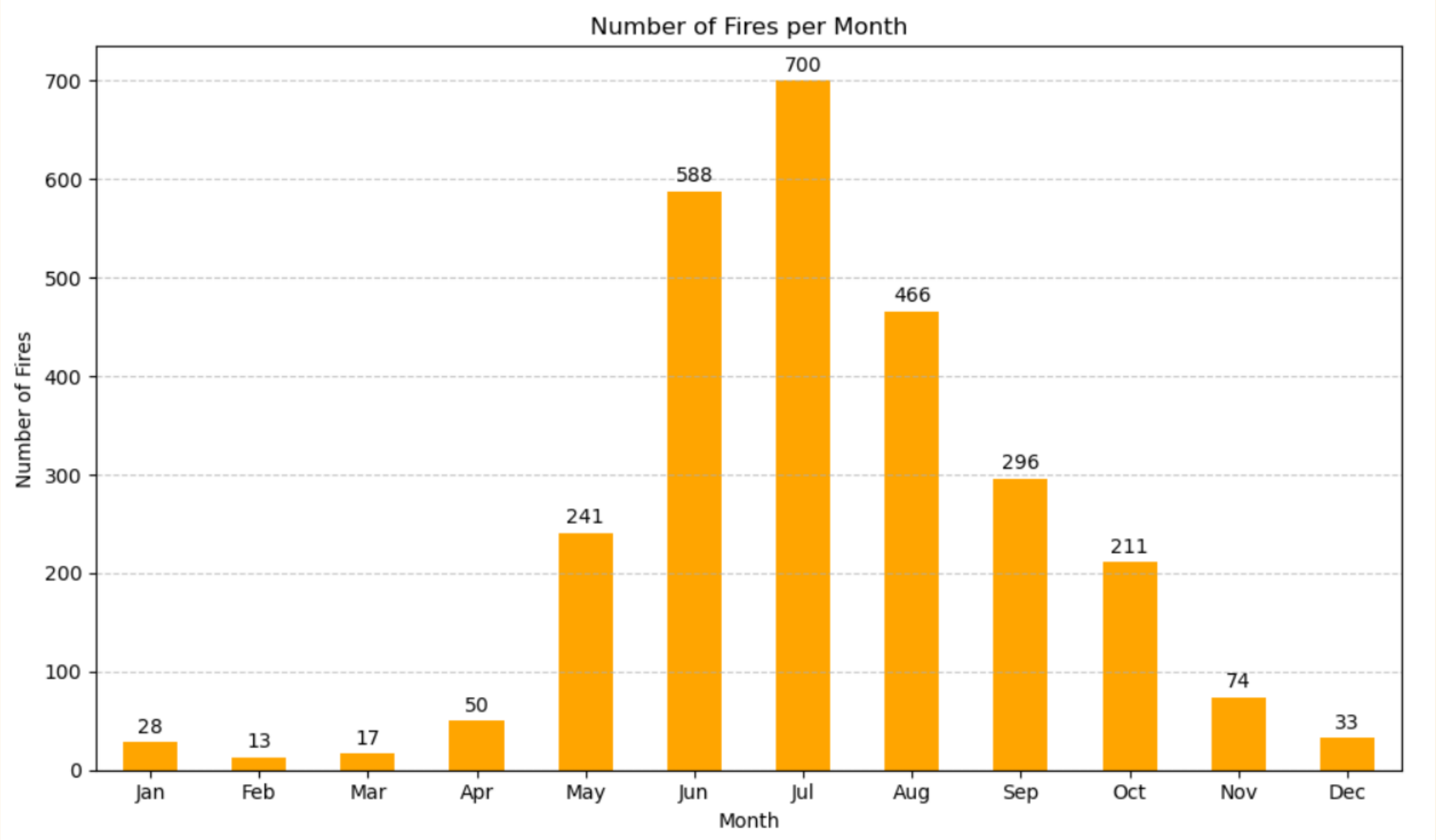
Key Takeaways

•**More fires ≠ more destruction** (e.g., 2020).
•Focus should be on **fire scale**, not just count.

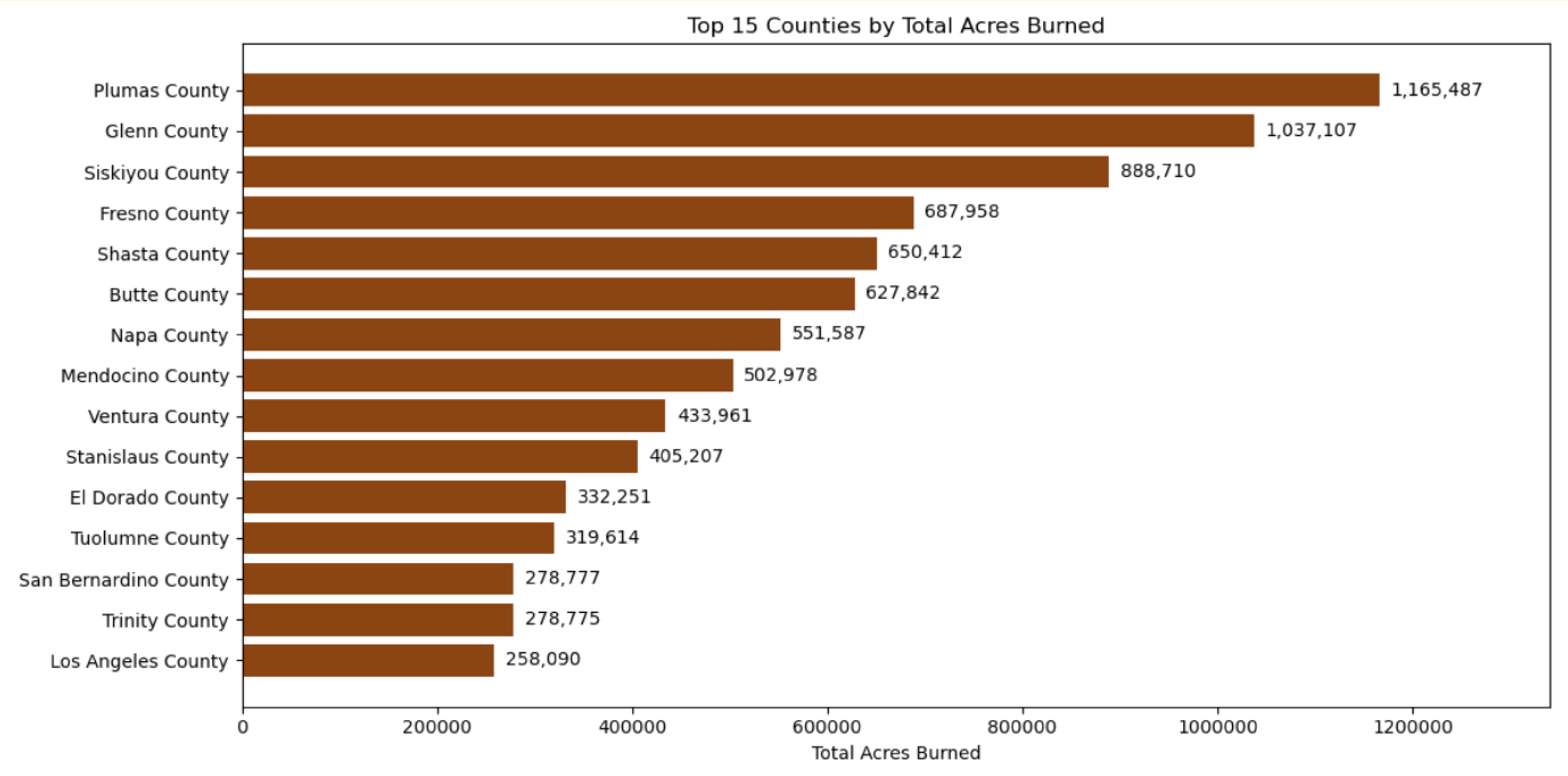
•**Summer is the most dangerous fire season** in terms of scale and frequency of large wildfires.

•**Winter and Spring are relatively safer**, with fewer and smaller incidents.

•These seasonal insights are valuable for **fire preparedness planning**, resource allocation, and modeling fire risk over time.

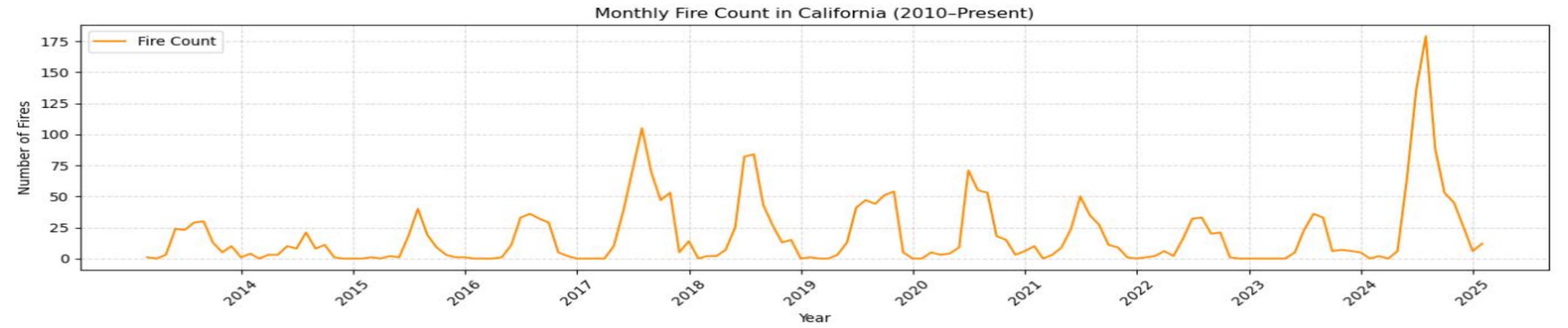


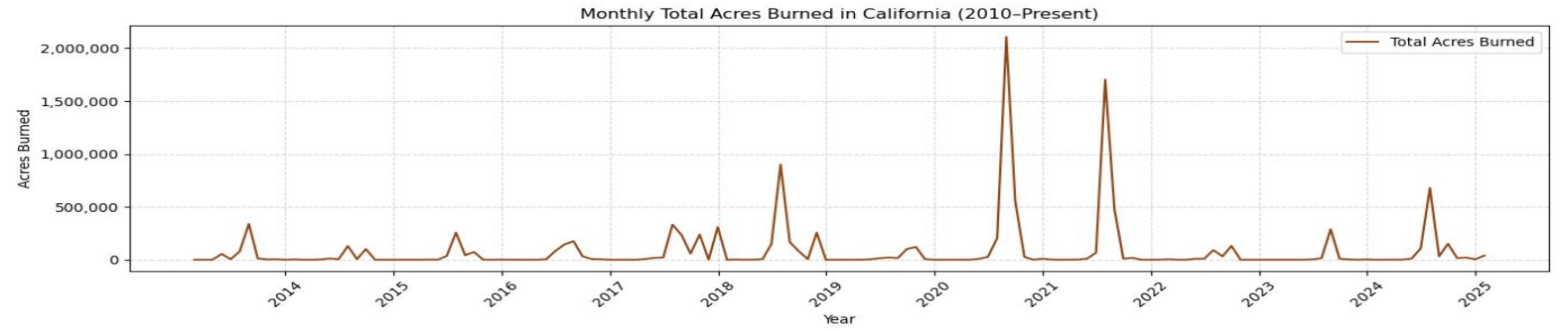
Counties such as **Plumas, Glenn** consistently recorded high total acreage burned, making them geographic hotspots for large-scale fires.



•**Seasonality is clear and recurring in the monthly fire count**: fire activity peaks consistently between **summer and early fall** (typically June–October). Notable **fire count spikes** occurred around: Late **2017**, Mid **2018**, **2024.**

•Acreage burned shows **fewer but more dramatic spikes**, indicating the presence of **large, high-impact fires, Late 2020** – the highest burn month (>2 million acres) and **Mid-2021.**





MODELLING

🎯 **Target Variable:** acres_burned — continuous (regression task)
🧭 **Spatial:** latitude, longitude, (Feature Engineered) , top 10 counties (One Hot Encoded)
📅 **Temporal:** fire_created_date, month,
🌧️ **Weather:** precipitation, max_temp, avg_wind_speed, temp_range, lagged_precipitation, lagged_avg_wind_speed
🔥 **Fire behavior**: brightness, frp

	Fire Size Range	Number of Fires
0	0–10 acres	103
1	11–100 acres	1370
2	101–1,000 acres	807
3	1,001–10,000 acres	295
4	10,001–100,000 acres	112
5	100,001–1,000,000 acres	18
6	Over 1,000,000 acres	1

- Convert lat/long into Cartesian coordinates:

```python
x = cos(lat) * cos(lon)
y = cos(lat) * sin(lon)
z = sin(lat)
```

- Helps models capture spatial relationships more smoothly than raw lat/long.

84% of number of fires is less than 1000 acres.

RANDOM FOREST BEST PERFORMING MODEL WITH MAE OF 109 ACRES

1. On average, my model's predictions are off by about 109 acres.

2. So the average prediction error (109 acres) is 109 ÷ 1000 = 10.9% of the maximum possible fire size in this dataset.

3. It generalizes well to unseen small fires.

4. All models outperformed the baseline, meaning they learned patterns from the input features.

5. Random Forest showed the best performance, reducing MAE by ~ 29 acres (~21% improvement).

Model	Test MAE (acres)
Baseline	137.67
Linear Regression	112.46 (105.03)
Ridge Regression	112.18
Lasso Regression	111.71
Random Forest	**108.95**
XGBoost	114.52

<u>DEMO</u>

RECOMMENDATIONS

Use the model at the earliest signs of a fire to quickly estimate potential fire size and guide rapid, informed decision-making.

Counties like Plumas and Glenn, which frequently experience large fires, should be equipped with additional resources and rapid response teams.

Allocate extra crews and equipment between June and September, when most large-scale fires occur.

Display model outputs on real-time maps to help agencies visualize hotspots and coordinate responses more effectively.

Incorporate additional variables like wind speed, drought levels, and vegetation density to enhance prediction accuracy over time.

NEXT STEPS

- Add vegetation and fuel data to improve how the model predicts fire spread (e.g., type of plants, how dry they are).
- Include elevation and terrain data to understand how landscape affects fire movement.
- Fix and connect the NOAA API to pull in live weather data for fire locations.
- Use real-time fire data from FIRMS to enable live alerts and updates.
- Enhance forecasting with time-based models to spot seasonal or climate-driven fire trends.
- Show predictions on detailed maps (like ArcGIS) with county lines, population, and key infrastructure.

THANK YOU

Temp	Wind	Humidity
70	5	20
75	NaN	30
NaN	10	25

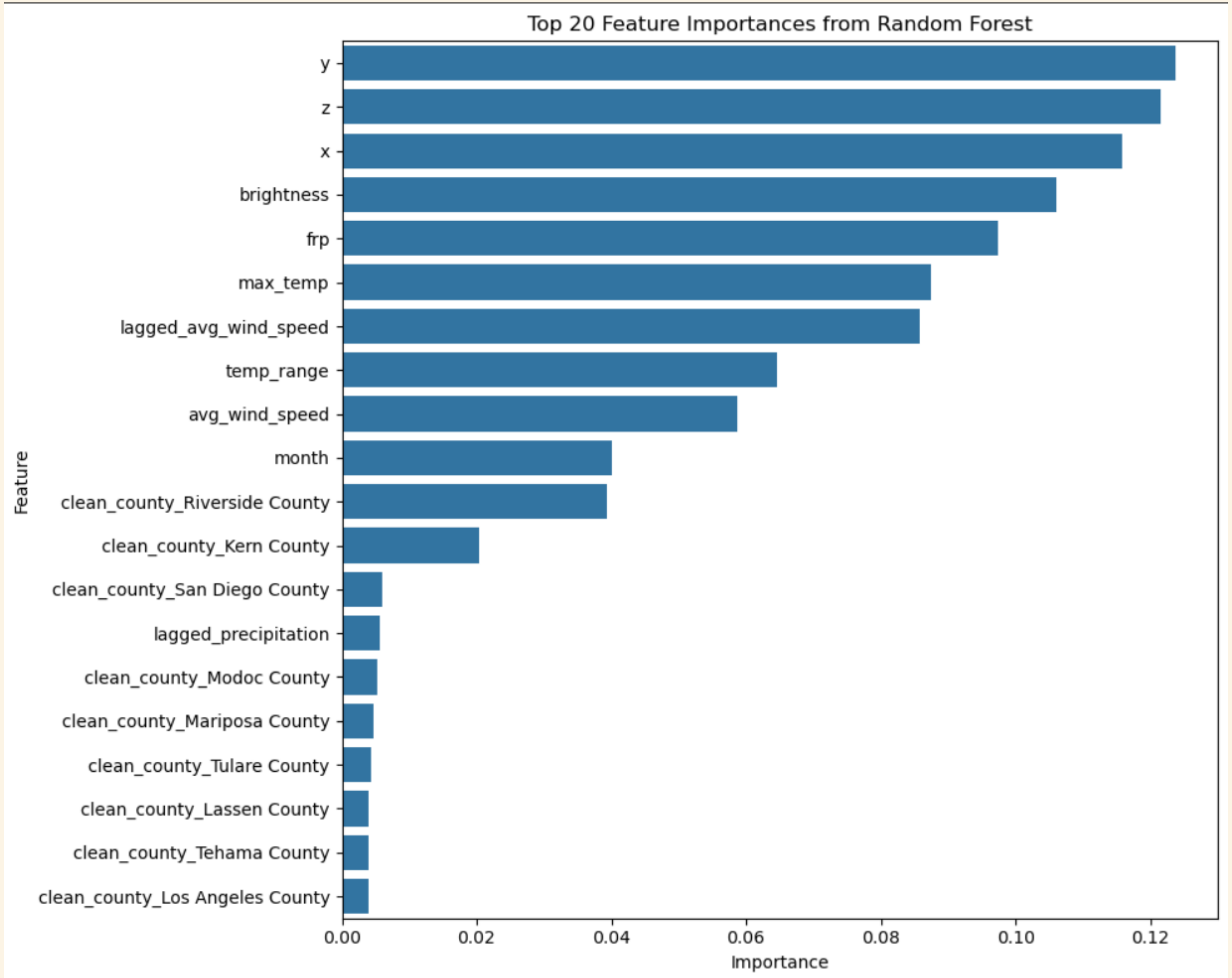
First, fill NaNs with simple values.

Then, build a model: predict Wind using Temp and Humidity.

Use that to update the missing Wind.

Do the same to predict missing Temp.

Repeat until predictions stabilize. (More accurate than Mean/ median imputation.


Top 20 Feature Importances from Random Forest
y
z
x
brightness
frp
max_temp
lagged_avg_wind_speed
temp_range
avg_wind_speed
month
clean_county_Riverside County
clean_county_Kern County
clean_county_San Diego County
lagged_precipitation
clean_county_Modoc County
clean_county_Mariposa County
clean_county_Tulare County
clean_county_Lassen County
clean_county_Tehama County
clean_county_Los Angeles County
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance

```json
{
  "type": "Feature",
  "properties": {
    "name": "California",
    "population": 39500000
  },
  "geometry": {
    "type": "Polygon",
    "coordinates": [
      [
        [-124.4096, 40.0000],
        [-114.1312, 32.0000],
        ...
      ]
    ]
  }
}
```

Folium is a Python library used to create **interactive maps** using **Leaflet.js**, a leading JavaScript mapping library — all without needing to write any JavaScript.

🗺️ What Can Folium Do?

It lets you:

- Plot **geographic data** (like lat/lon points)
- Overlay **GeoJSON** data (e.g., county boundaries, fire perimeters)
- Add **heatmaps**, **markers**, **popups**, **tiles**, and **layers**
- Easily visualize data from **Pandas**, **GeoPandas**, or even shapefiles