Predicting US Wildfires

Ana Cahet
Rachel Star

INTRO

Datasets

Main - Wildfires, location, dates, duration, size, cause, etc

2.3 Million US Wildfires (1992-2020) 6th Edition

Spatial wildfire occurrence data for the United States, 1992-2020

https://www.kaggle.com/datasets/behroozsohrabi/us-wildfire-records-6th-edition?select=data.csv

Complementary - Weather and topography info

Predict Droughts using Weather & Soil Data

Predicting continental US drought levels using meteorological & soil data.

https://www.kaggle.com/datasets/cdminix/us-drought-meteorological-data

Final - Working file

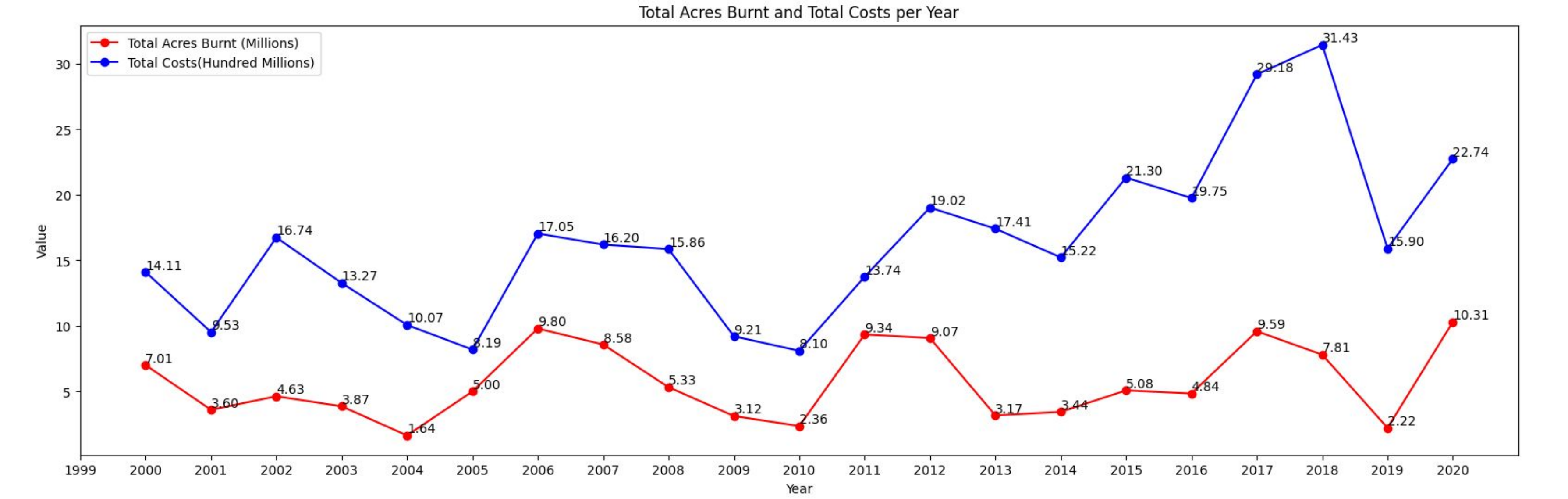
Merged and cleaned dataset	
Year range	2000 - 2020
Wildfires count	1.7 million
Location	Continental US (48 states)

Problem

Wildfires impact

- Human: health, lives, property
- Wildlife: ecosystems, biodiversity

- Costs:
  - Avg. Acres per year: 5.7 million
  - Avg. Spent per year: US$ 1.6 billion



https://www.kaggle.com/datasets/dylanfox10/federal-wildfires-acres-cost-temp-19852020

Objectives

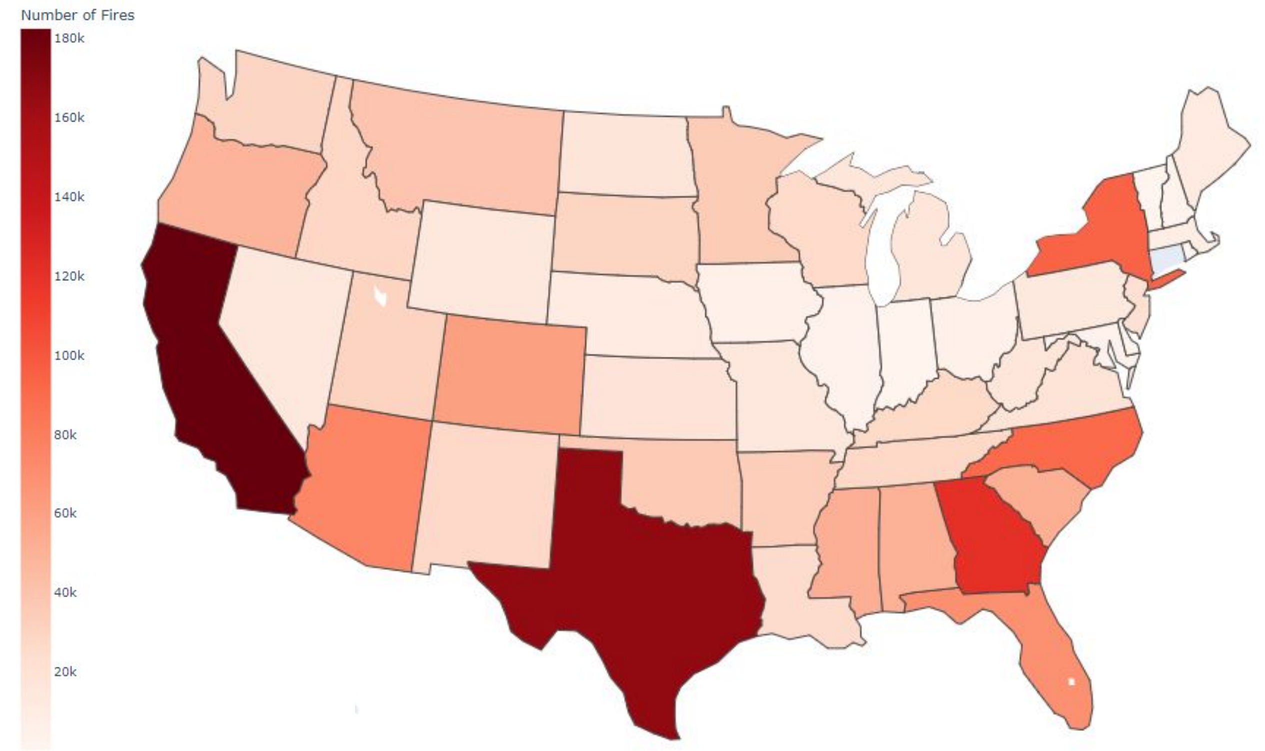
- Understand trends of wildfires and climate in the US

- Use machine learning/deep learning models to predict the CAUSE of the wildfires:
    - Lead to more prevention opportunities, saving money, lives & environment
    - Help solving undetermined causes (27%)

OVERVIEW

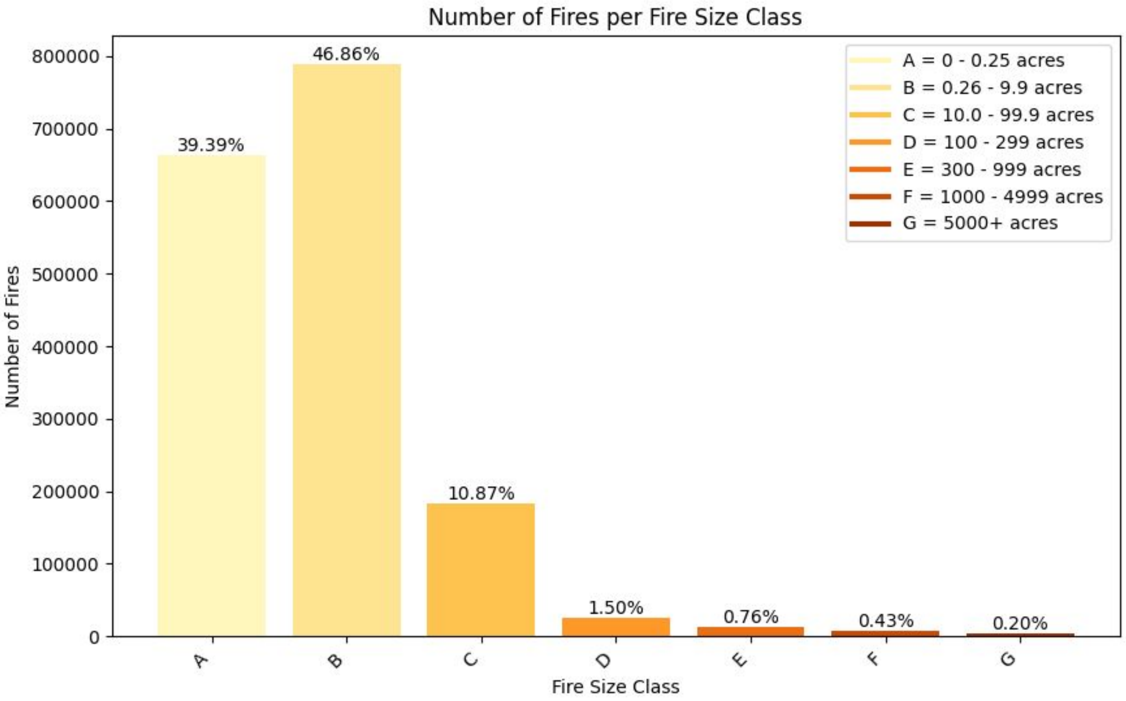
Wildfires x State

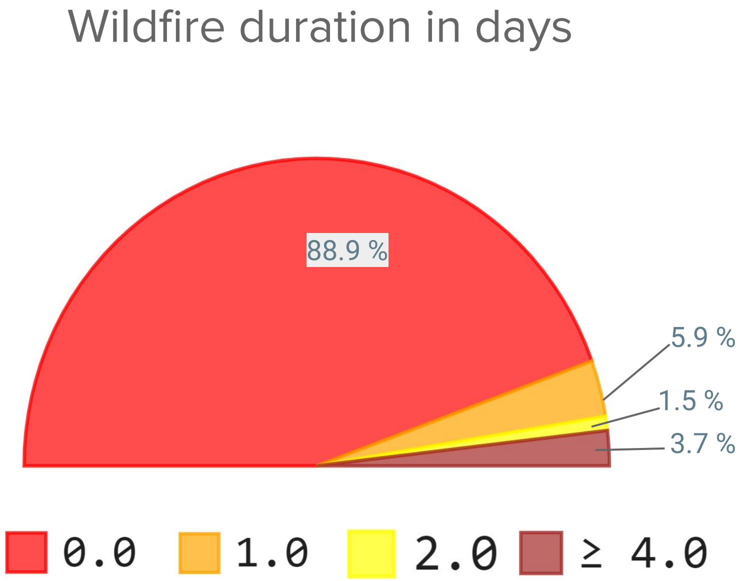
1. California - 182k
2. Texas - 167k
3. Georgia - 121k
4. New York - 94k
5. North Carolina - 91k



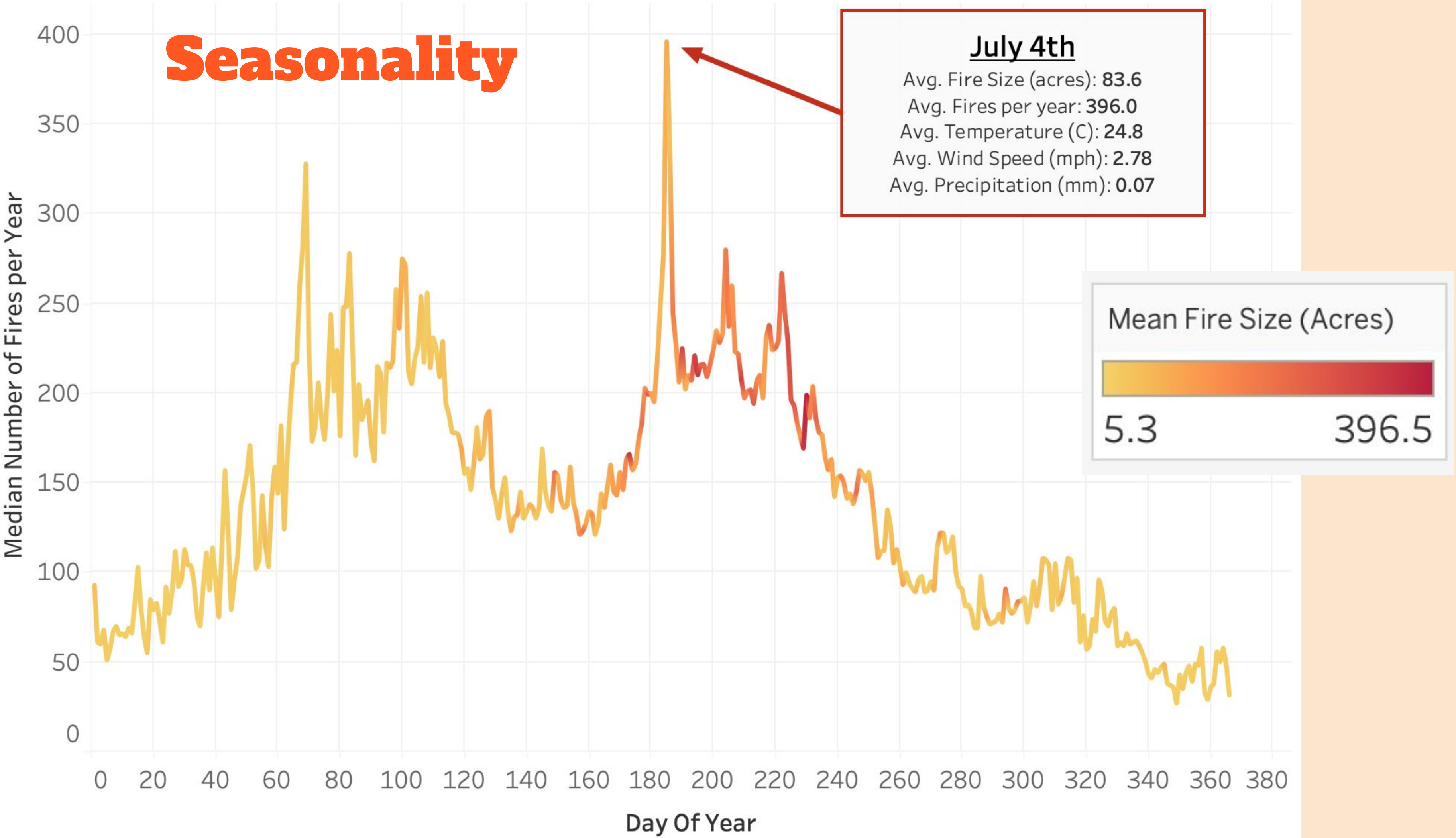
Size & Duration

- Majority of fires:
  - “Small” - 86,25% < 10 acres
  - Put out the same day (≒ 90%)





Seasonality



July 4th

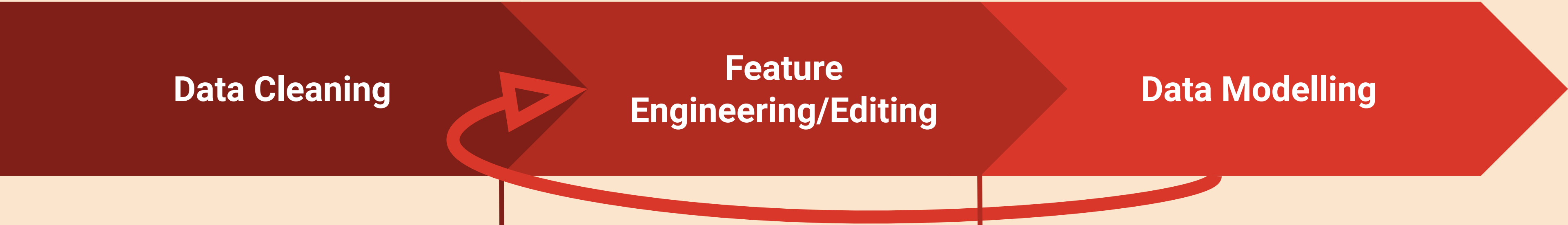
Avg. Fire Size (acres): **83.6**

Avg. Fires per year: **396.0**

Avg. Temperature (C): **24.8**

Avg. Wind Speed (mph): **2.78**

Avg. Precipitation (mm): **0.07**



Data Cleaning

- Removing unnecessary columns
- Joined datasets, removed fires missing climate data
- Cleaned categorical variables
- Filled in NA’s based on other data

Feature Engineering/Editing

- Encoded and ‘categorised’ necessary categorical features
- Combined categorical groups based on confusion matrices created by models
- Refined imbalanced dataset using under- and over- sampling methods.

Data Modelling

We ran and evaluated various models, using them to refine the dataset and determine the best model:

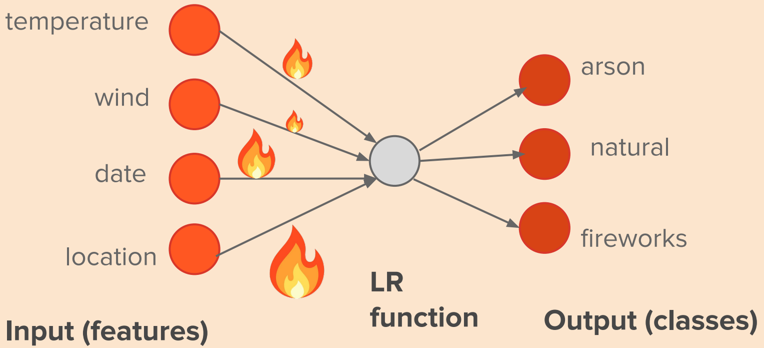
- Logistic regression
- Decision Tree
- Neural Network
- Random Forest

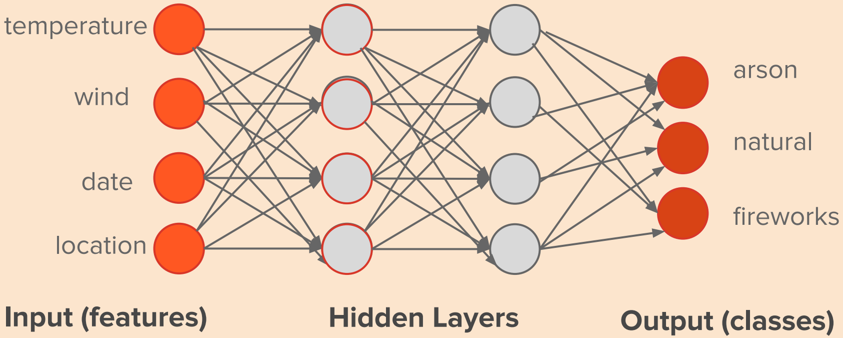
MODELS

Logistic Regression & Neural Networks

- A type of algorithm utilized for **prediction** by assuming a **linear relationship** between the input features (columns) and the output classes
- It assigns **weights (coefficients)** to the different **features**, enabling a clear understanding of their impact on the prediction
- Considered a **white-box baseline model,** logistic regression serves as an excellent starting point due to its **simplicity in interpretation**. It establishes a benchmark for evaluating more sophisticated models

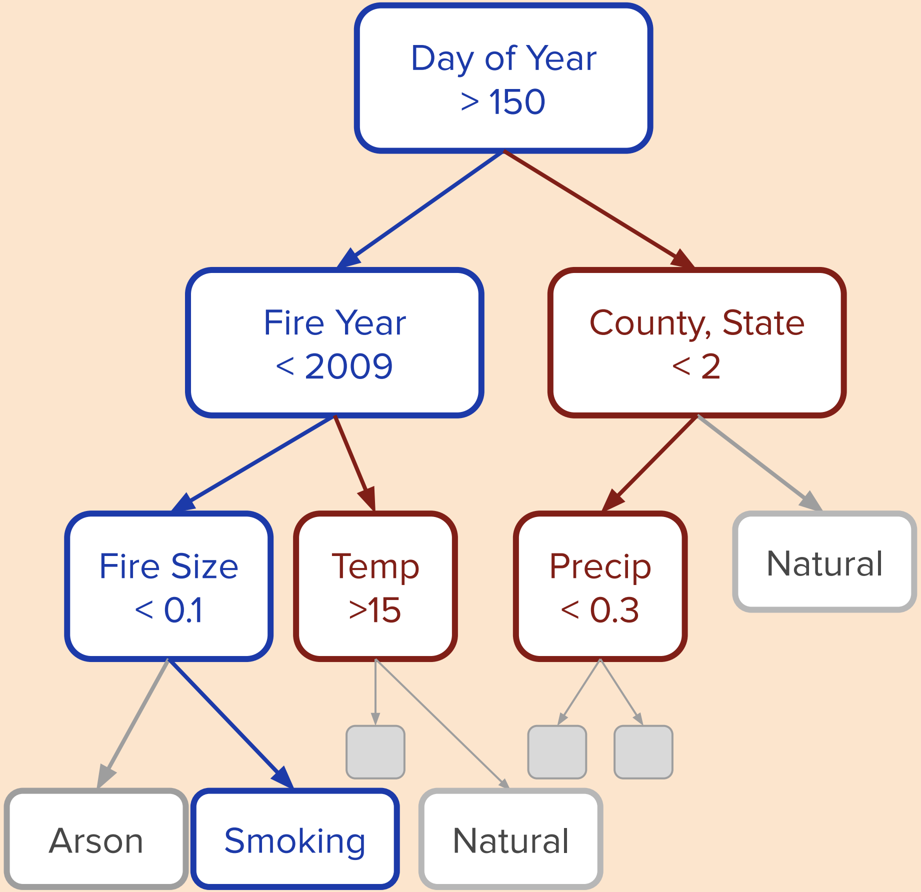
- Inspired by the functioning of the **human brain**, a **neural network** comprises interconnected **nodes (neurons)** organized in layers
- As a **deep learning model**, neural networks are more **complex**, also assigning **weights** to input features and utilizing various **activation functions**
- Logistic regression can be conceptualized as a single-layer neural network. Logistic functions are often used as activation functions in neural network hidden layers





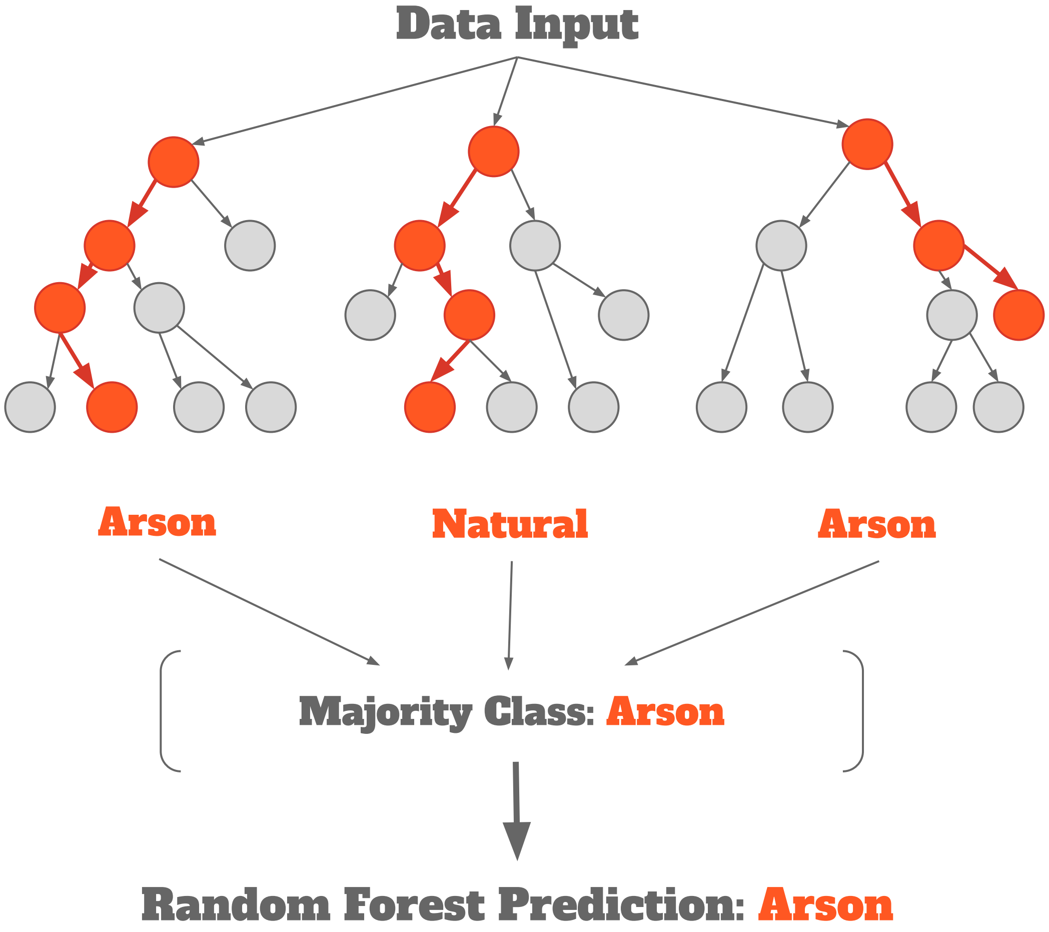
Decision Tree

- Whitebox model
- Model splits data based on conditions for each ‘node’, until a classification can be made
- First, calculates the ‘root node’
- Then, refines values, orders, and levels



Random Forest

- An ‘ensemble’ of decision trees
- Each tree is built using a different random sample of the dataset
- Each tree may predict a different class for a data point
- The ‘majority class’ prediction will be returned as the random forest output



Data Input

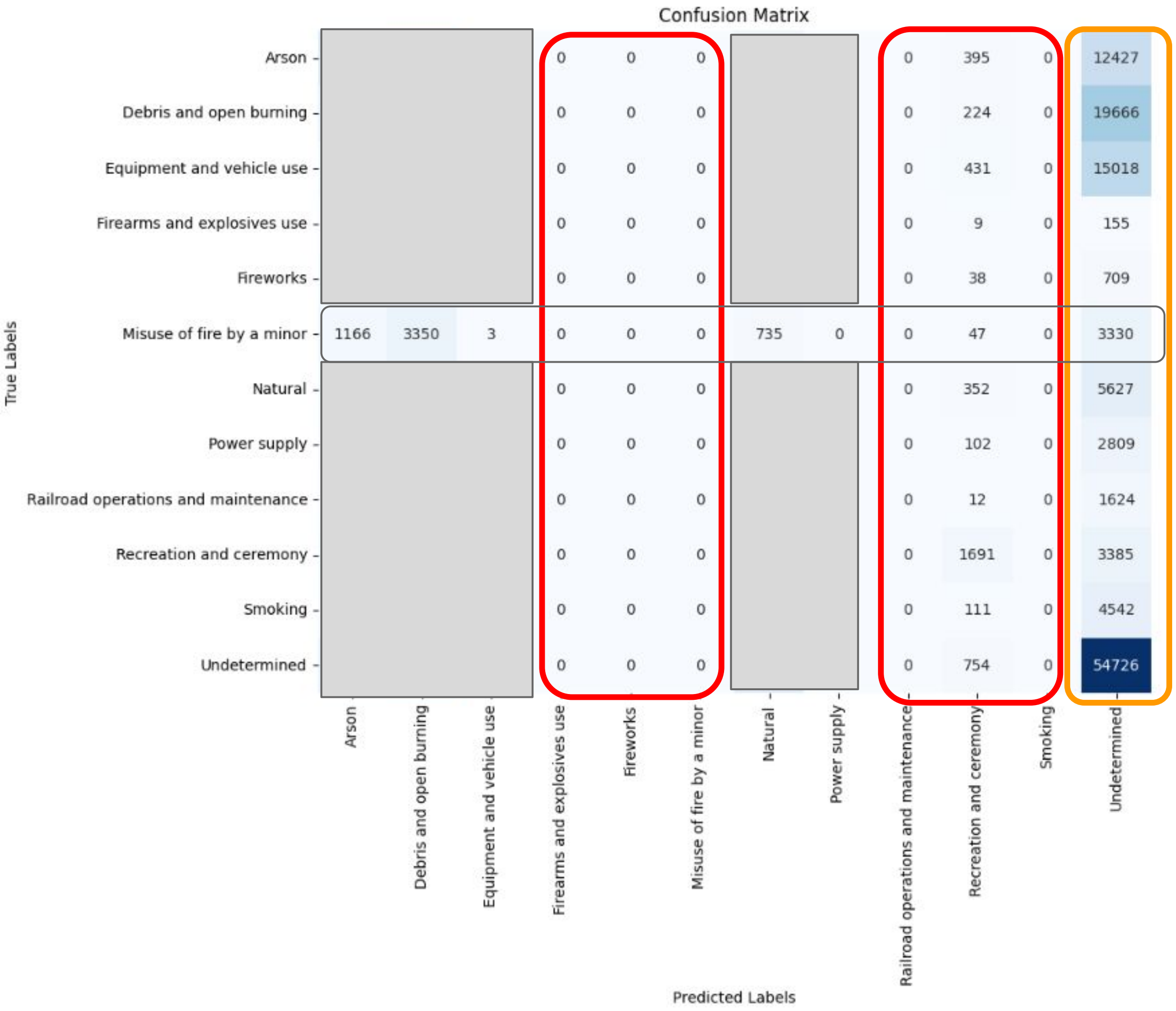
Arson

Natural

Arson

Majority Class: Arson

Random Forest Prediction: Arson


Confusion Matrix
True Labels
Predicted Labels

True Labels \ Predicted Labels	Arson	Debris and open burning	Equipment and vehicle use	Firearms and explosives use	Fireworks	Misuse of fire by a minor	Natural	Power supply	Railroad operations and maintenance	Recreation and ceremony	Smoking	Undetermined
Arson				0	0	0			0	395	0	12427
Debris and open burning				0	0	0			0	224	0	19666
Equipment and vehicle use				0	0	0			0	431	0	15018
Firearms and explosives use				0	0	0			0	9	0	155
Fireworks				0	0	0			0	38	0	709
Misuse of fire by a minor	1166	3350	3	0	0	0	735	0	0	47	0	3330
Natural				0	0	0			0	352	0	5627
Power supply				0	0	0			0	102	0	2809
Railroad operations and maintenance				0	0	0			0	12	0	1624
Recreation and ceremony				0	0	0			0	1691	0	3385
Smoking				0	0	0			0	111	0	4542
Undetermined				0	0	0			0	754	0	54726

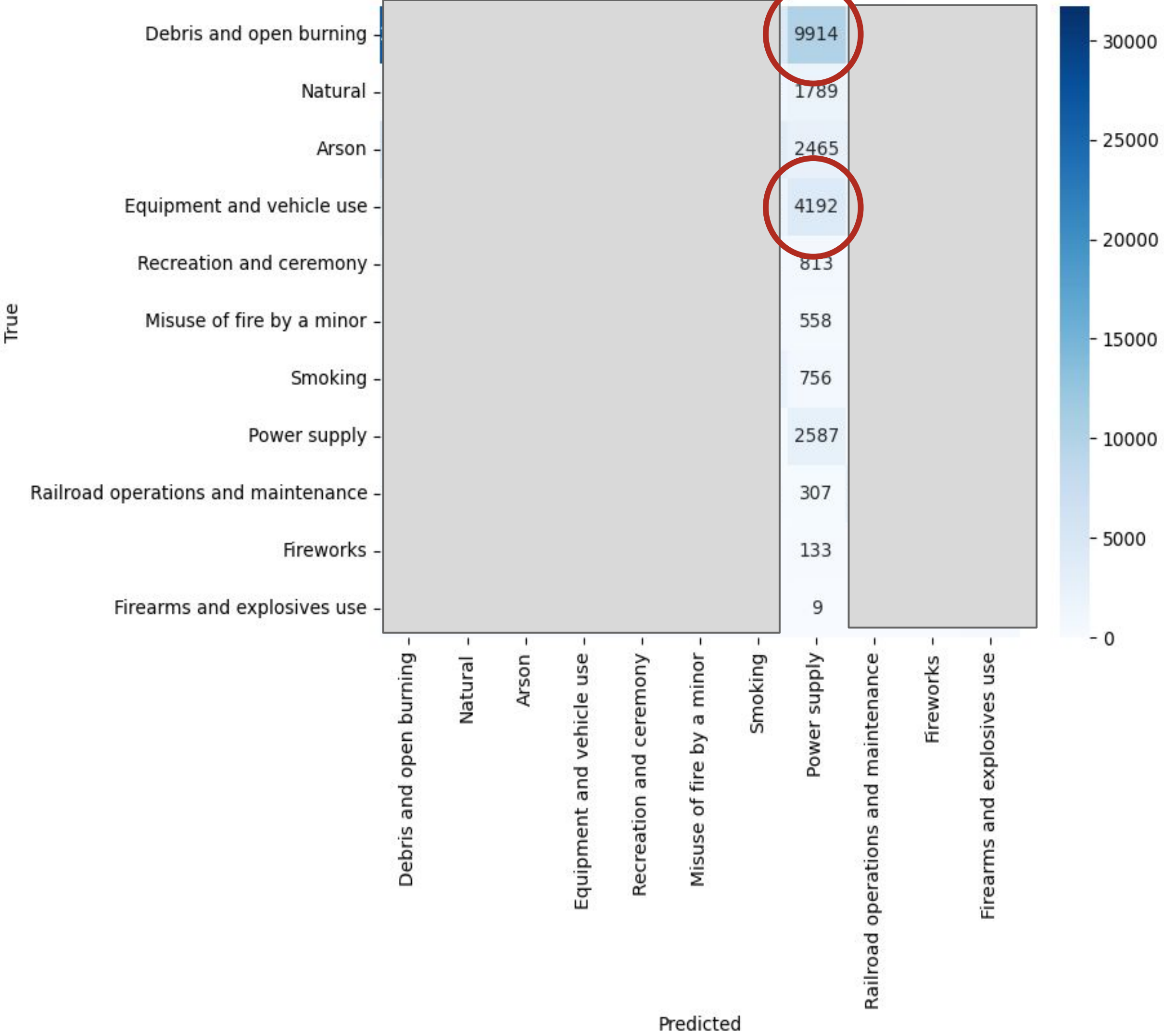
Confusion Matrix

- Neural network/logistic regression - 1st attempt

- True vs Predicted

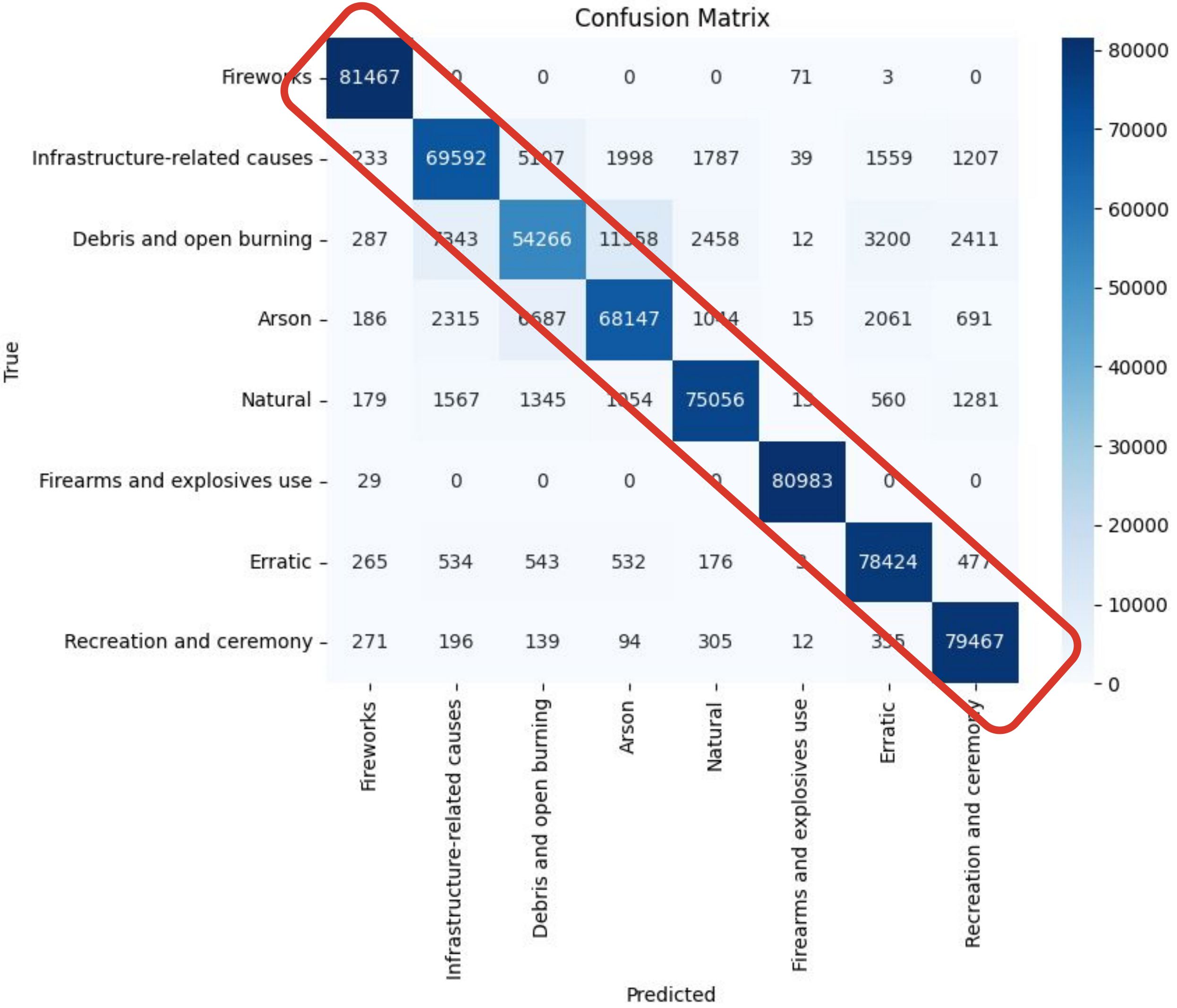
- Was not able to predict: ‘Firearms’, ‘Fireworks’, ‘Minor’, ‘Smoking’…

- Most of predictions mapped to ‘Undetermined’



Confusion Matrix

- Balanced Random Forest
- Identify where categories are being mis-labelled
- Allows us to adjust categories by combining them
- Leads to improved model performance



	Fireworks	Infrastructure-related causes	Debris and open burning	Arson	Natural	Firearms and explosives use	Erratic	Recreation and ceremony
Fireworks	81467	0	0	0	0	71	3	0
Infrastructure-related causes	233	69592	5107	1998	1787	39	1559	1207
Debris and open burning	287	7343	54266	11358	2458	12	3200	2411
Arson	186	2315	6687	68147	1044	15	2061	691
Natural	179	1567	1345	1054	75056	13	560	1281
Firearms and explosives use	29	0	0	0	0	80983	0	0
Erratic	265	534	543	532	176	3	78424	477
Recreation and ceremony	271	196	139	94	305	12	355	79467

Confusion Matrix

- Random Forest model with Over-sampling

- Clearly shows better classification

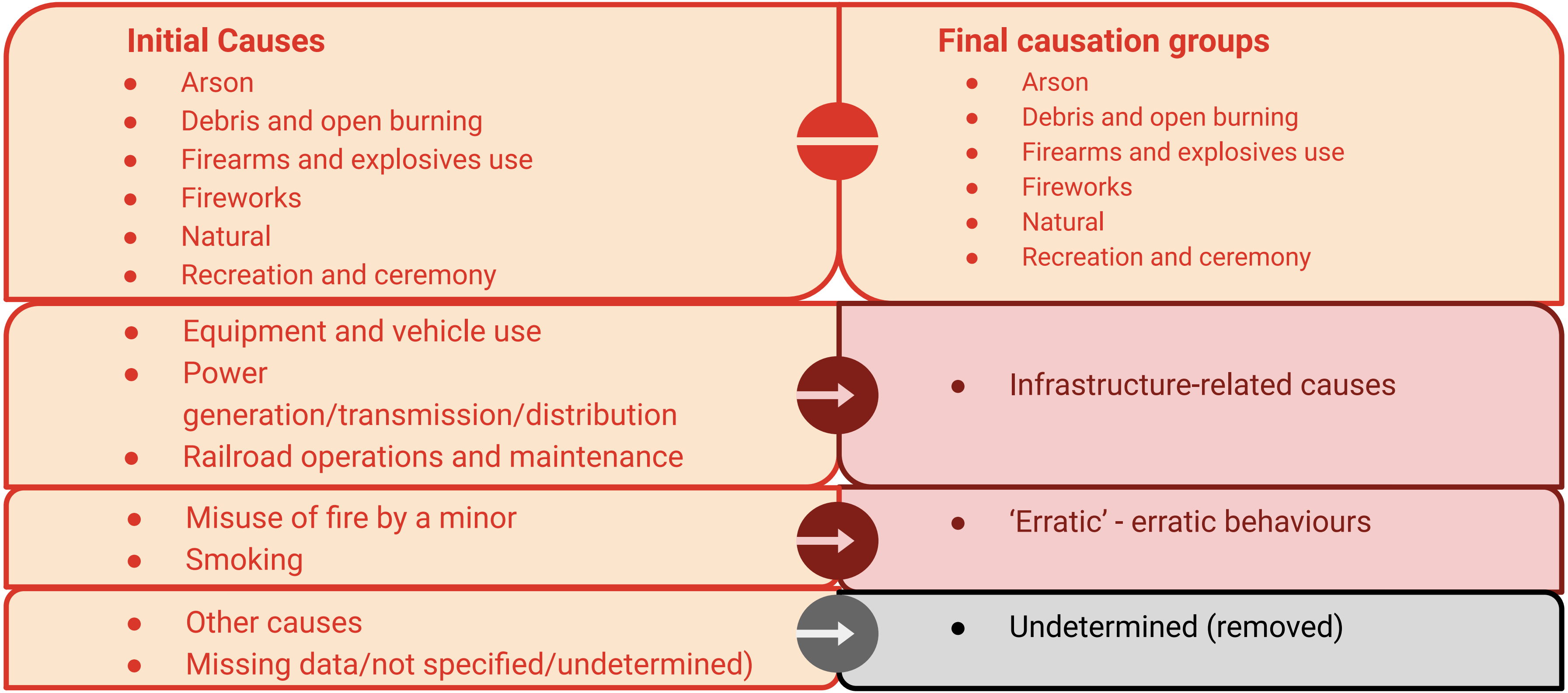
Models

Model	F1-Score
Logistic Regression	0.37
Decision Tree, with ‘oversampling’	**0.87**
Neural Network	0.57
Balanced Random Forest, with ‘oversampling’	0.64
Random Forest	0.59
Random Forest, with ‘oversampling’	**0.90**

An evaluation metric which combines measures of **precision** (how well the model can identify an instance of a class) and **recall** (how well the model can identify all instances of a class).

A method to balance sample size of categories, by creating false data for categories of smaller sample size.

Target variable: NWCG General Cause



Initial Causes	Final causation groups
• Arson • Debris and open burning • Firearms and explosives use • Fireworks • Natural • Recreation and ceremony	• Arson • Debris and open burning • Firearms and explosives use • Fireworks • Natural • Recreation and ceremony
• Equipment and vehicle use • Power generation/transmission/distribution • Railroad operations and maintenance	• Infrastructure-related causes
• Misuse of fire by a minor • Smoking	• ‘Erratic’ - erratic behaviours
• Other causes • Missing data/not specified/undetermined)	• Undetermined (removed)

Random Forest Model Evaluation

Class	f1-score
Arson	0.83
Debris and open burning	0.73
Erratic behaviour	0.94
Firearms and explosives use	0.99
Fireworks	0.99
Infrastructure-related causes	0.85
Natural	0.93
Recreation and ceremony	0.96

How well each class can be predicted

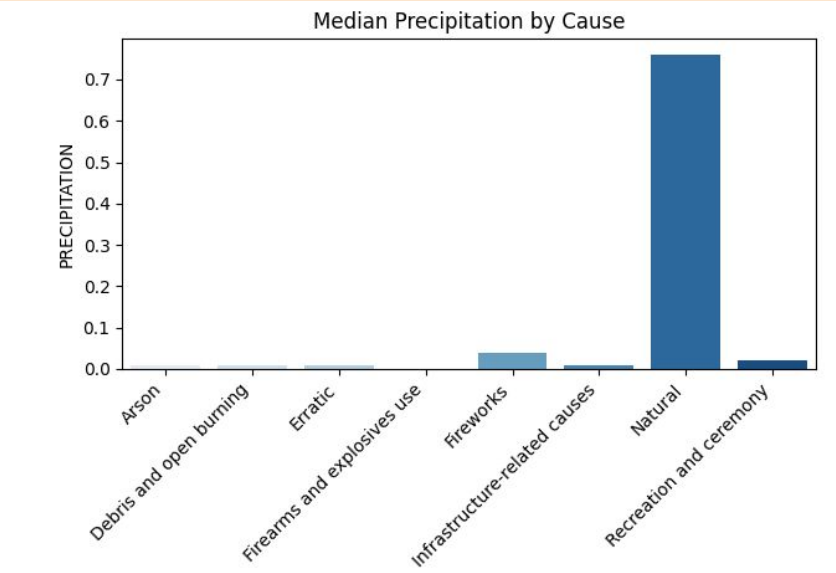
How important the data features are in determining the right class

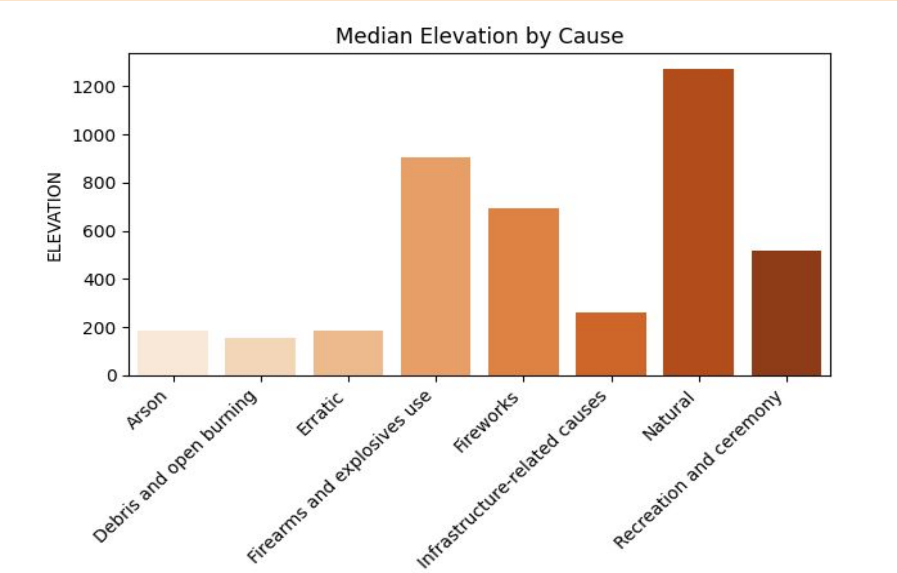
Feature	Importance
Day of Year	0.14
Elevation	0.13
County, State	0.12
Temperature	0.11
Wind Speed	0.10
Fire Size	0.09
Precipitation	0.08
Fire Year	0.08
State	0.08
Day of Week	0.05
Duration	0.02

CLASSIFICATION INSIGHTS

Precipitation & Elevation

- Human-related fires: drier weather
- Natural fires: higher elevation

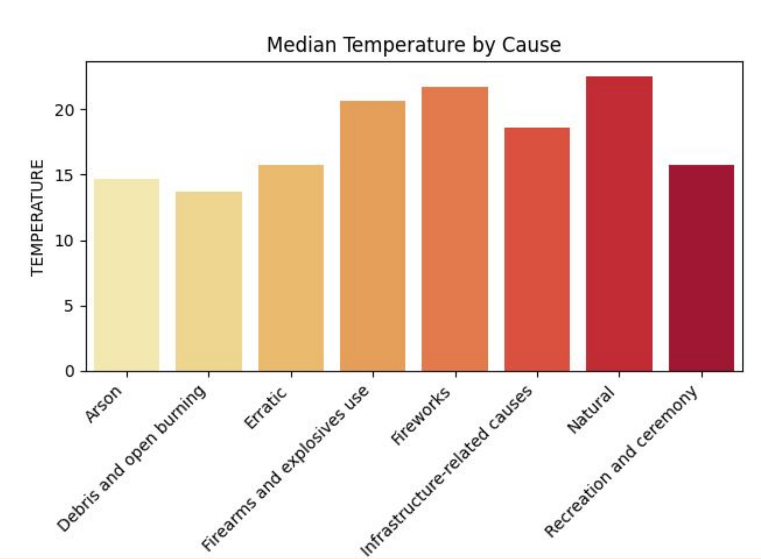


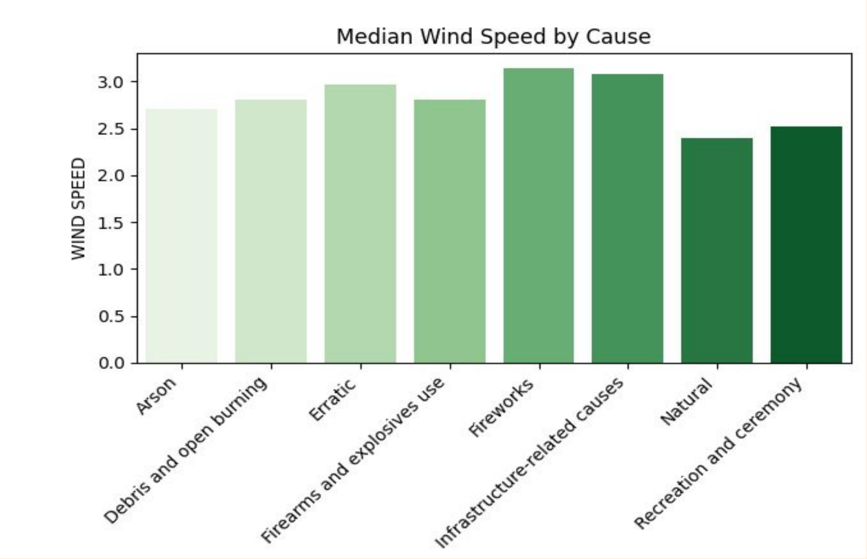


General

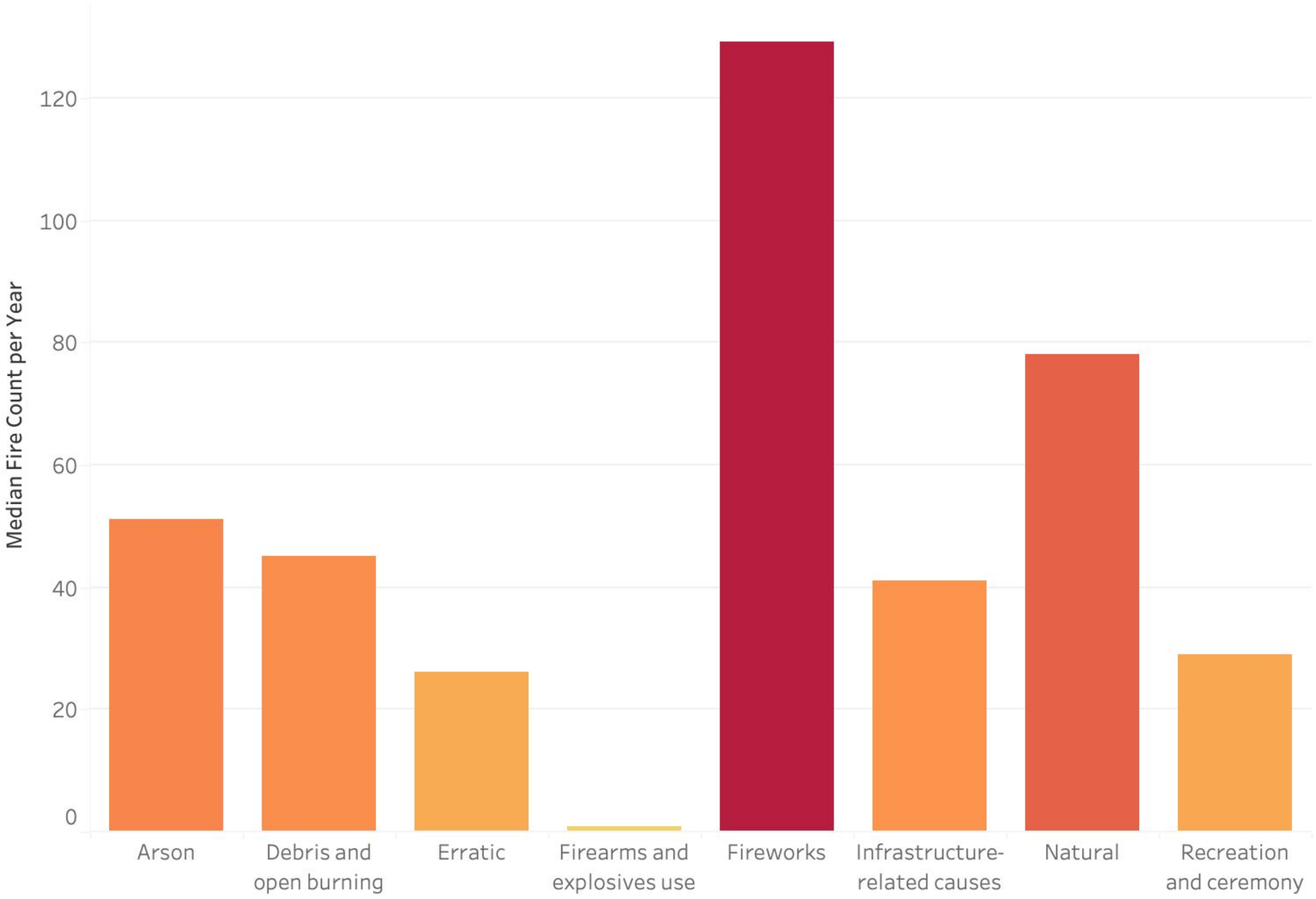
Temperature & Wind

- Human-related fires: higher wind speed
- Natural fires & Fireworks: higher temperature





Fire Causes on 4th July



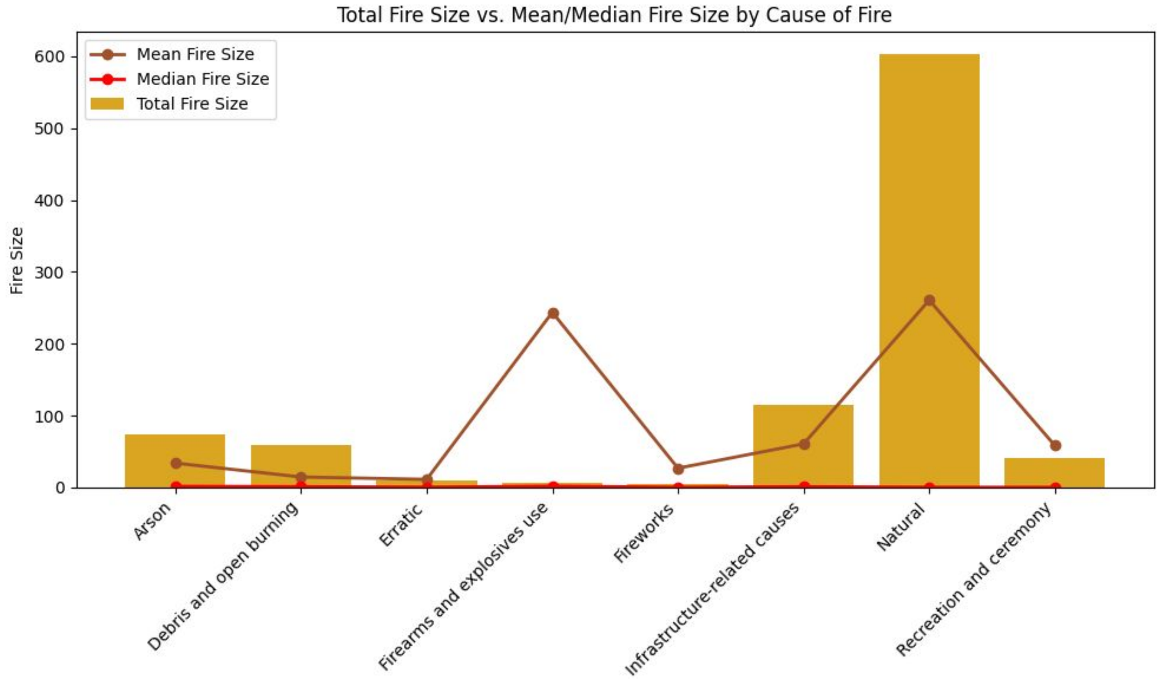
Specific

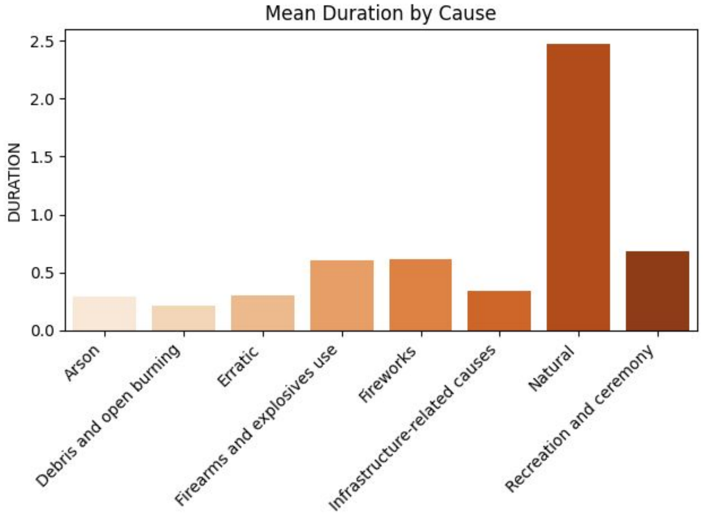
4th of July... Fireworks!

General

Size & Duration

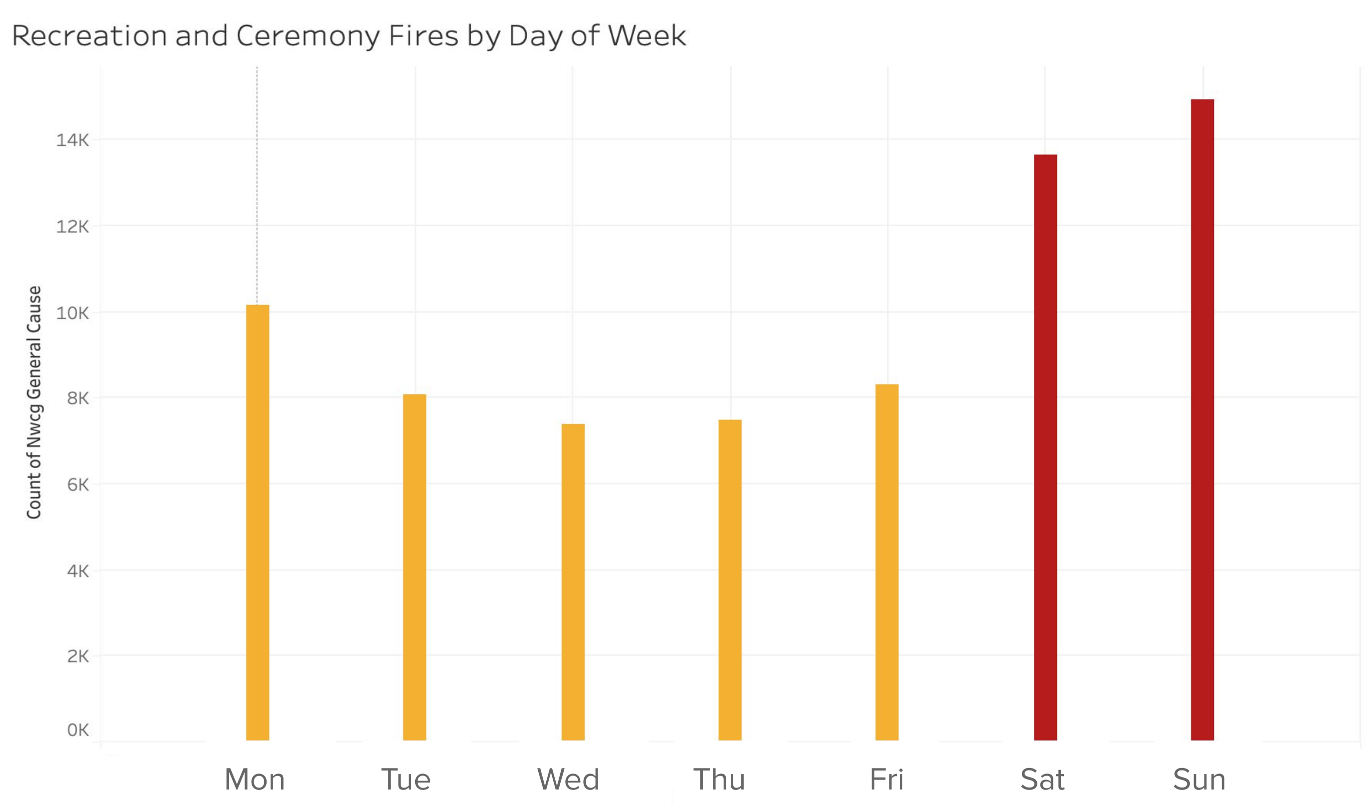
- Firearms & Explosives tend to generate bigger fires
- However, in total, Natural fires burnt more acres and, on average, are harder(take longer) to put out





Specific

Recreation & Ceremony



- Recreational fires exhibit twice the frequency of occurrence between Sunday (highest - 14k) and Wednesday (lowest - 7k)
- In contrast, other fire causes do not demonstrate such notable variations between different days of the week

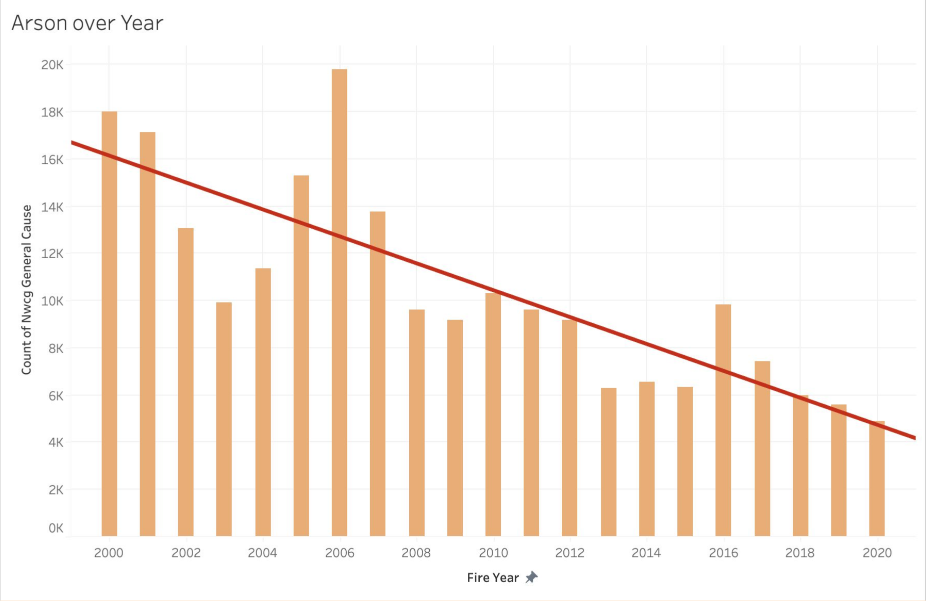
Specific

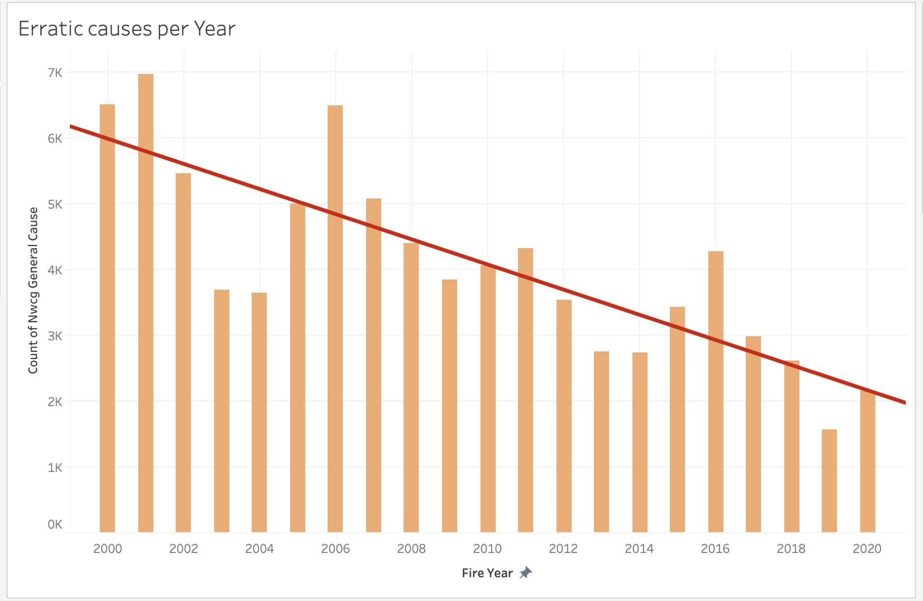
Arson & Erratic Over Years

Significant decline in the occurrences of two causes:

Arson: Potentially attributed to its classification as a felony, punishable by imprisonment and fines.

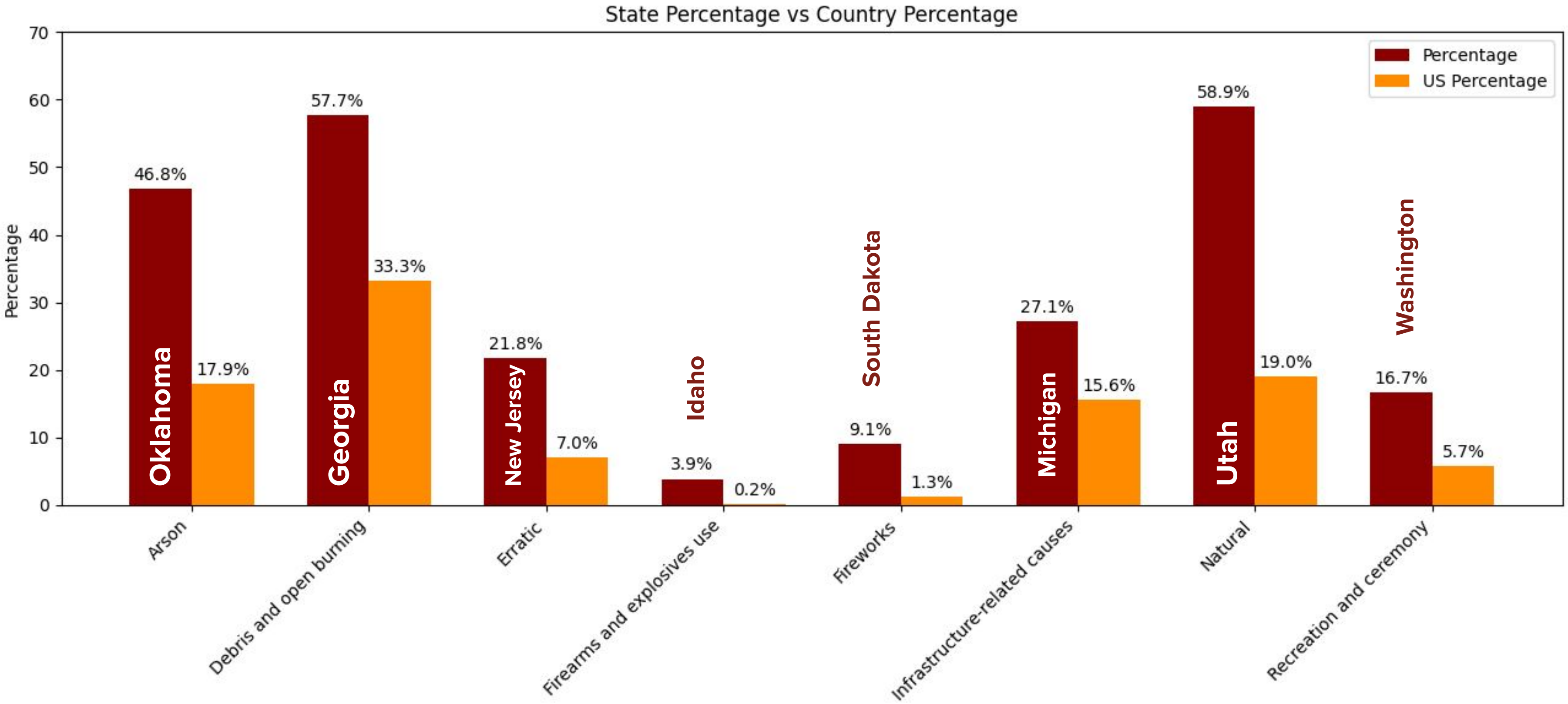
Erratic (Smoking + Minor): Potentially influenced by a decrease in the number of people smoking.



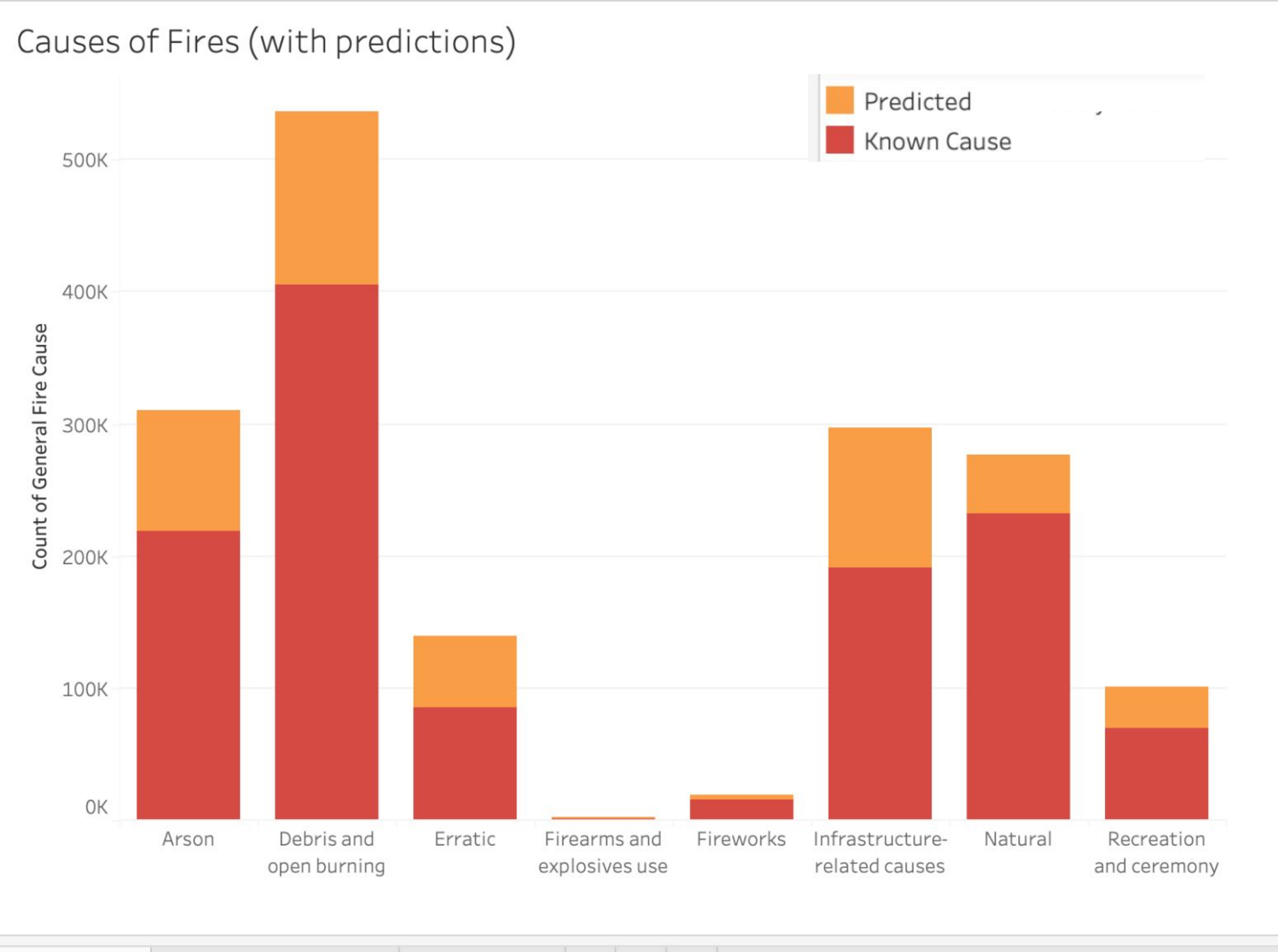


Specific

State-wise info



APPLICATION & PROSPECTS



Example of model use

- Predicting the ‘undetermined’ data removed earlier

- Good distribution of these predictions

Conclusions

- Random Forest, with over-sampling to balance the dataset, is the best model to predict wildfire causes in the US
- Understanding causes will save money, wildlife, lives, and the environment

Future Work

- Application to other locations/worldwide
- Further refinement to hyperparameters of Random Forest model
- Link more datasets, e.g. wildlife populations

Thank You!

Ana Cahet

www.linkedin.com/in/anacahet

Rachel Star

www.linkedin.com/in/rachelstar/