

# PREDICTING CALIFORNIA WILDFIRE BURN ACREAGE CLASS

Annie Miller, Aryana Far, Sarah Zhang

# TABLE OF CONTENTS

**01**

Question Formulation

**02**

Preprocessing

**03**

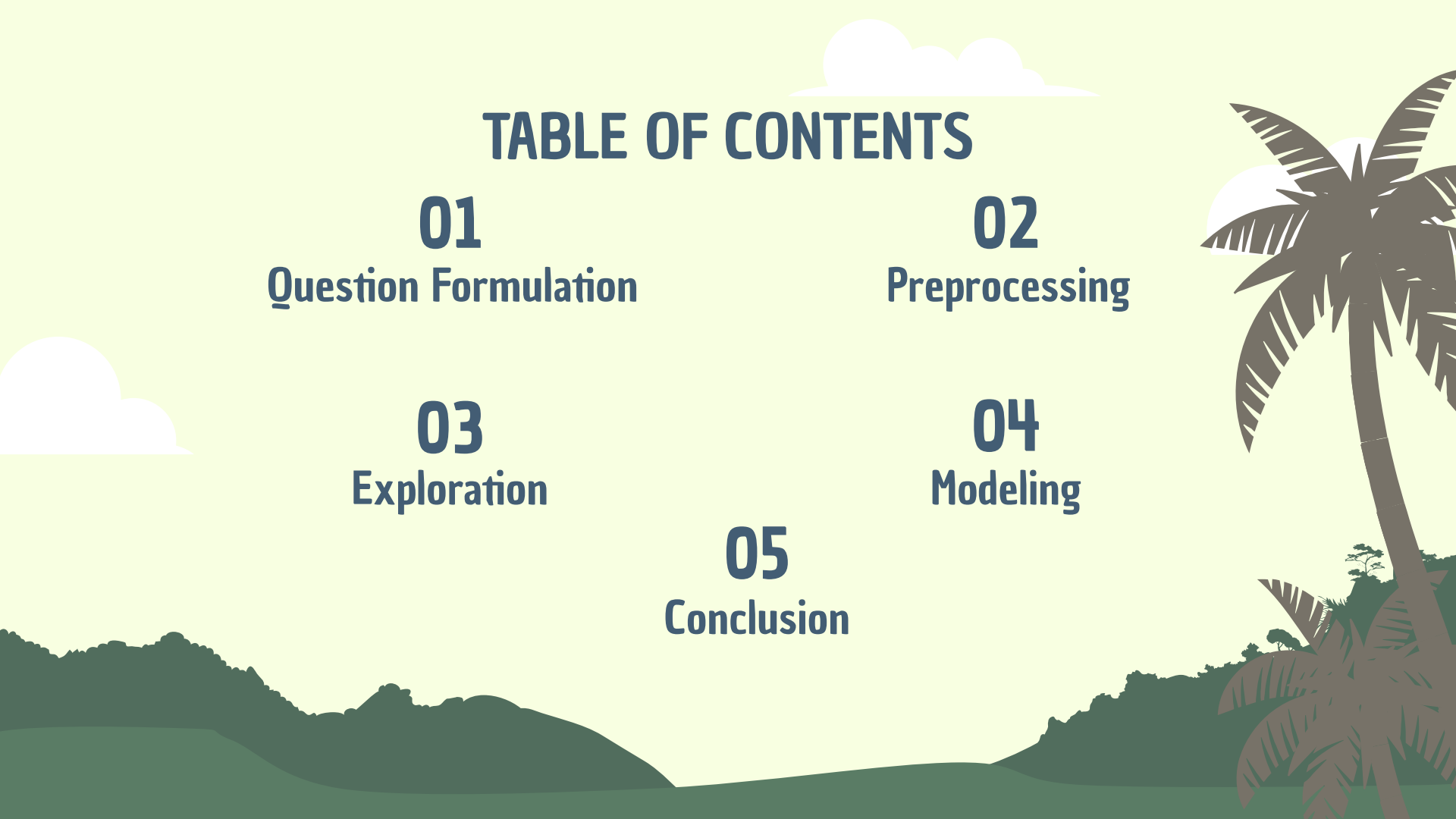
Exploration

**04**

Modeling

**05**

Conclusion



# 01 Question Formulation



# Backstory

Wildfires affect humans, animals, agriculture, air quality, and so much more...

Wildfires grow in size and increase in frequency each year.

California has been a major site for large and quickly-spreading wildfires.

Thus, this issue continues to be more and more pressing.



# Motivation

## Problem:

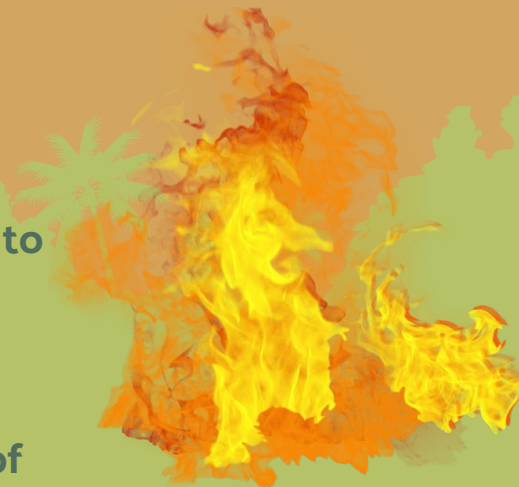
How can we prepare ourselves to address any given wildfire with the proper measures? Current prediction relies upon remote sensing, which is expensive, uncertain, and difficult to interpret.

## Proposed Solution:

As some areas don't have capacity for remote sensing, we propose building a data-driven model to predict the burn acreage class of wildfires.

## Goal:

Thus, we aim to use data to predict the severity of California wildfire acreage in order to better equip mitigation efforts and better prepare counties for fire season.





# 02 Data Preprocessing

# The Data



# Wildfire Data

## Granularity:

1636 rows; one row for each recorded wildfire in California occurring 2013-2020.

## Relevant Features:

- 'AcresBurned': *integer*, # acres burned # dropped rows with NaNs
- 'ArchiveYear': *integer*, year that wildfire occurred
- 'County': *string*, California county in which the wildfire occurred → *categorical*, one-hot-encoding # RegEx
- 'Extinguished': *datetime*, the date the fire was extinguished → *integer*, day of year # dropped rows with NaNs
- 'Started': *datetime*, the date the fire started → *integer*, day of year out of 365
- 'CalFireIncident': *boolean*, whether or not the wildfire was handled by CalFire
- 'MajorIncident': *boolean*, whether or not the wildfire was a Major Incident

## Added Features

- 'BurnDuration': *integer*, how many days the fire lasted # 'Extinguished' - 'Started'
- 'BurnAcreageClass': *categorical*, according to 'AcresBurned', is this a class A, B, C, . . . , G fire?



# Imputing Extinguished NaNs



Suspected relationship between the length of a fire (BurnDuration) and its burn acreage (AcresBurned)

- Fire start date (Started) + length of a fire (BurnDuration) = extinguish date (Extinguished)

Goal: impute NaN Extinguished dates using a model based on AcresBurned

- Linear regression:
  - X: AcresBurned
  - y: BurnDuration

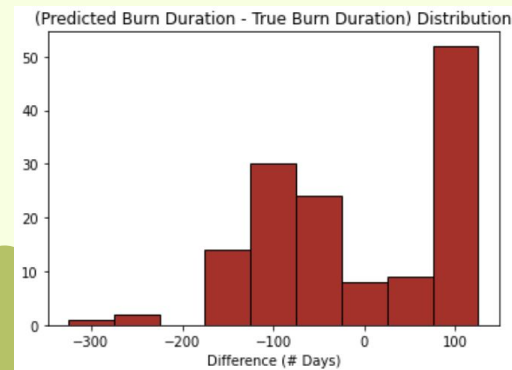
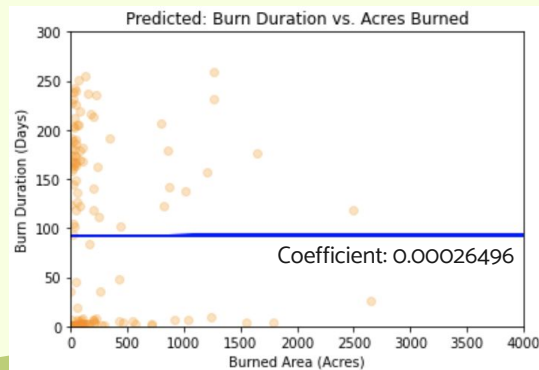
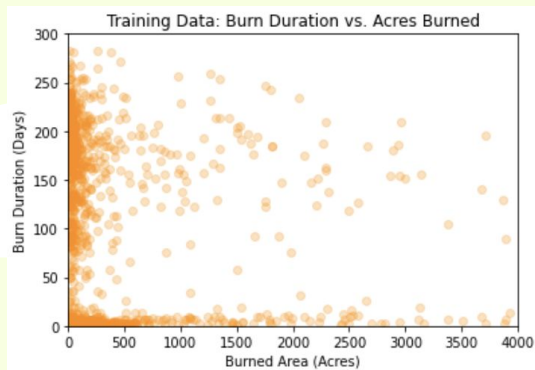
# Imputing Extinguished NaNs

No relationship between burn duration and acres burned

- Regression coefficient close to 0
- Vast differences between predicted and actual burn durations

Drop data containing NaN Extinguished values

- 59 rows dropped
- 1636 rows total (before dropping Extinguished NaNs)



# 'BurnAcreageClass' - Target

We labeled our data into different classes by burn acreage.

**Class A** - 1/4 acre or less

**Class B** - more than 1/4 acre, but less than 10 acres

**Class C** - 10 acres or more, but less than 100 acres

**Class D** - 100 acres or more, but less than 300 acres

**Class E** - 300 acres or more, but less than 1,000 acres

**Class F** - 1,000 acres or more, but less than 5,000 acres

**Class G** - 5,000 acres or more.

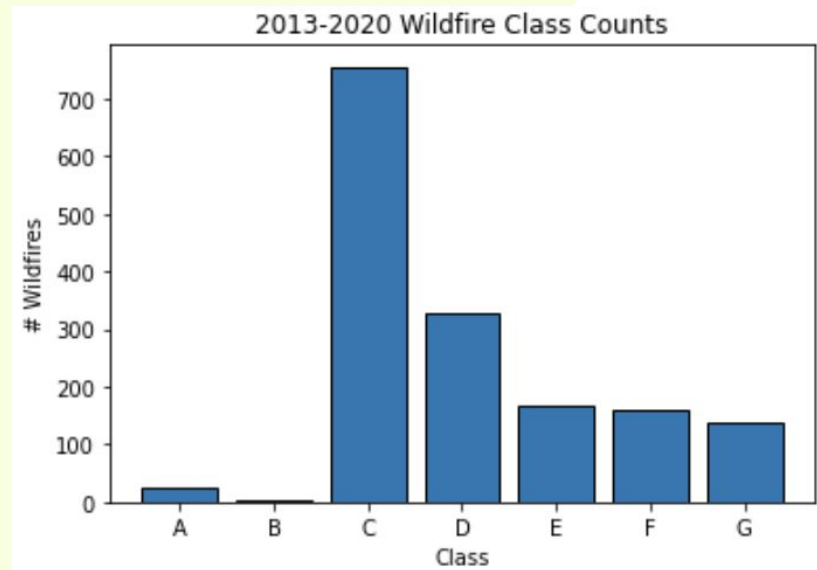
This class became our target for our predictive modeling to come!

# 'BurnAcreageClass' Cleaning

Low number of **Class A** and **Class B** recorded fires

- CalFire classifies wildfires as having burned  $\geq 10$  acres
- **Class A:** 0 to 0.25 acres
- **Class B:** 0.25 to 10 acres

Drop these classes from the dataset (27 rows)



# External Data on Land, Population, and Climate



## 5 Total External Datasets...

### Land Data

- **'SquareMileage'**: *integer*, square mileage of county
- **'County'**: *string*, California county # standardized to match wildfire data via RegEx

### Population Data

- **'AveragePopulation'**: *integer*, average population between years 2013-2020 of county
- **'County'**: *string*, California county # standardized to match wildfire data via RegEx

### Climate Data

- 3 DataFrames
    - **Temperature**
    - **Wind**
    - **Relative Humidity**
  - For each DataFrame: 7 columns for months January to October, 1 column for county, 1 column for year. This data was obtained from the EPA, and cleaned to have monthly averages of each county for each year.
    - “The climate of the preceding winter and spring can explain over 50% of the year-to-year variability and overall trend in summer fire activity“
- 

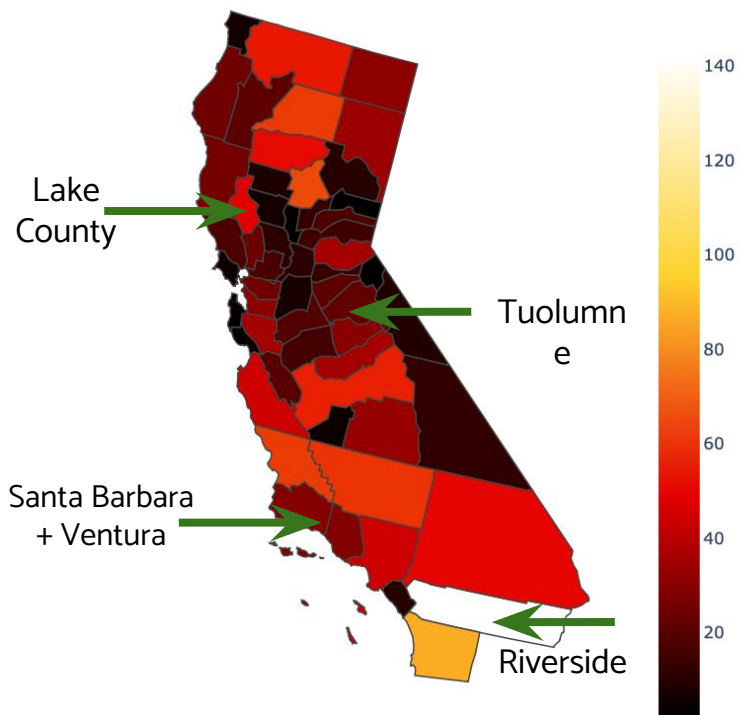
03

# Exploratory Data Analysis

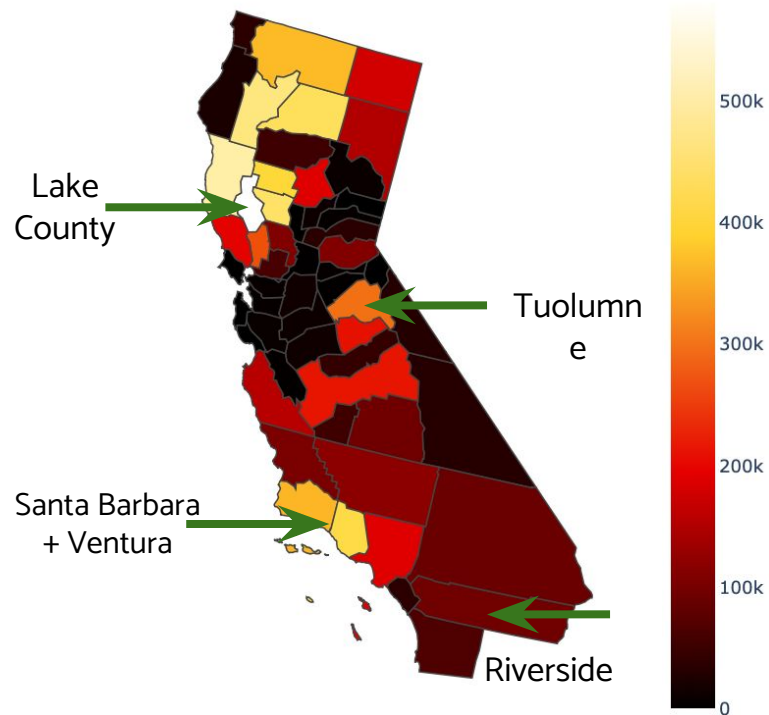


# California Wildfire Heatmaps

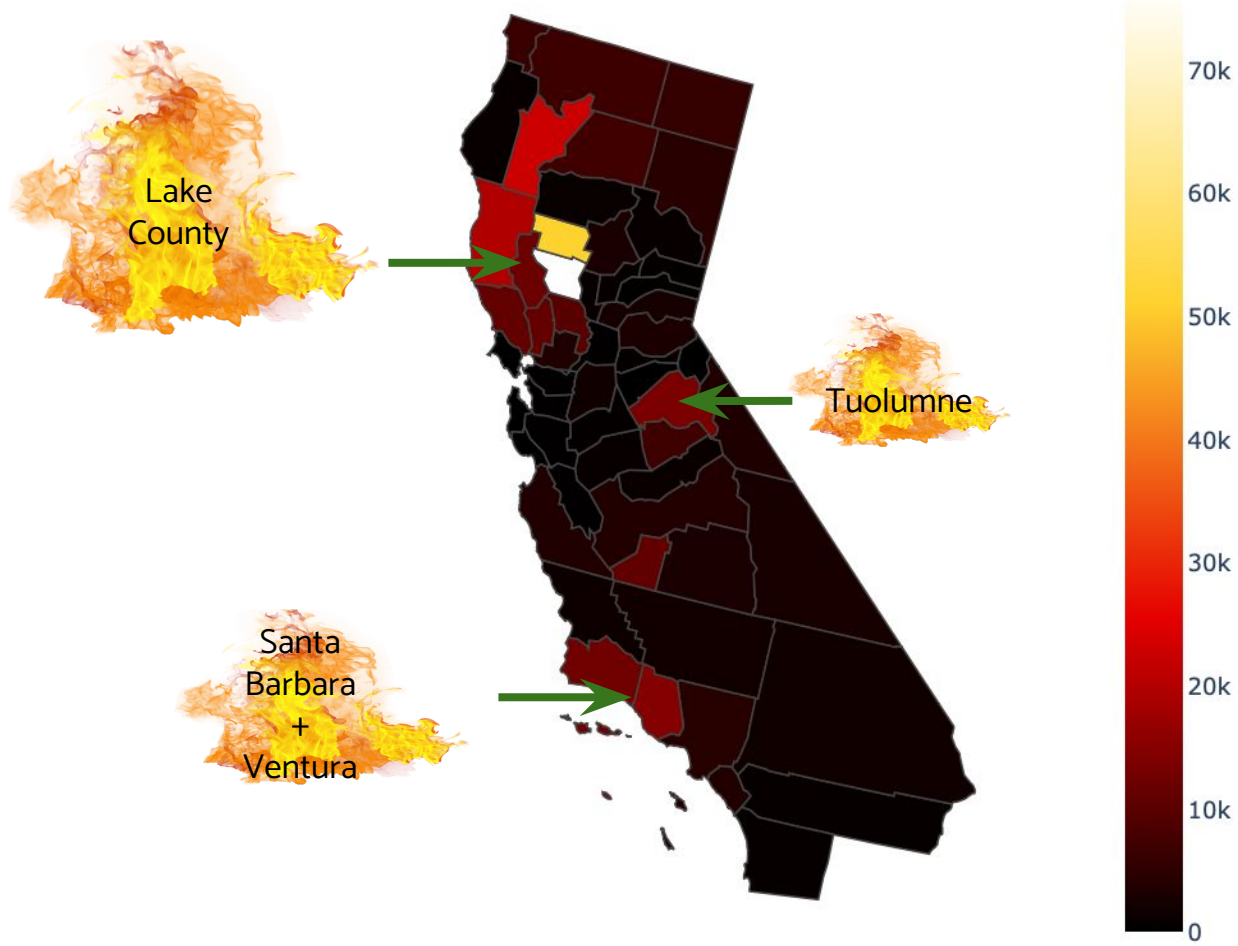
Total Fire Frequency from  
2013-2020 per County



Total Acres Burned 2013-2020



# Average Acres Burned Per Fire Per County



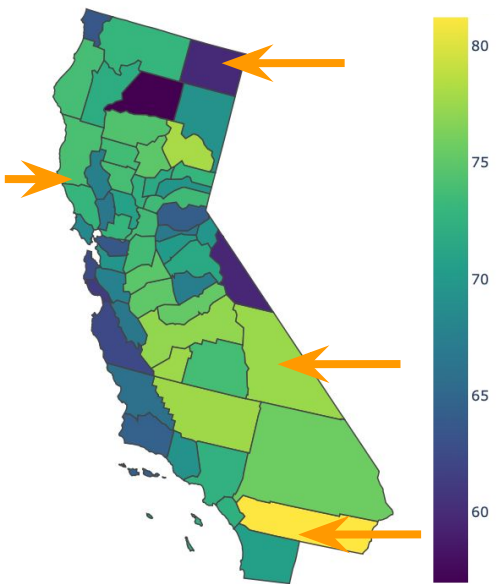


# Average population Per County

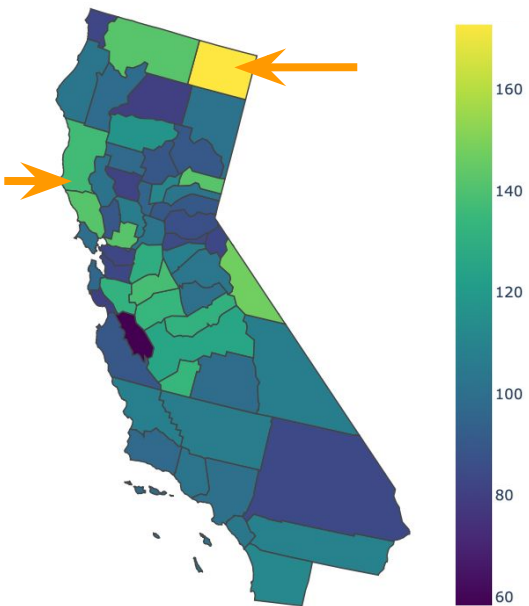


# California Climate Heatmaps

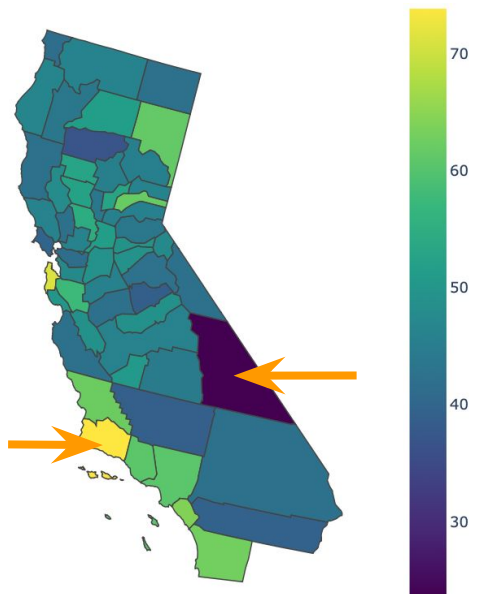
Average Summer Temperature  
from 2013-2020 per County



Average Summer Wind Speed  
from 2013-2020 per County



Average Summer Relative  
Humidity from 2013-2020 per  
County



04

# Modeling





# Model Selection + Accuracy



For our model we decided to use a random forest to predict burn acreage class.

## Model Tuning:

- Training & Test Split
    - We initially started with a train test split of 80/20 and after some tuning we decided to change the train/test split to 70/30
    - Since the representation of acreage burn classes in the data is not equal and the C class is heavily overrepresented in the data set we decided to use a stratified train test split to have the training and test set have the same percentage of each class label
  - After tuning our model we found that the best parameters for the random forest classifier were:
    - Max depth - 8
  - Model Train Accuracy: 0.758
  - Model Test Accuracy: 0.501
  - Percent Accuracy if only predicted class C: 0.488
- 
- 

# Test Prediction Accuracy

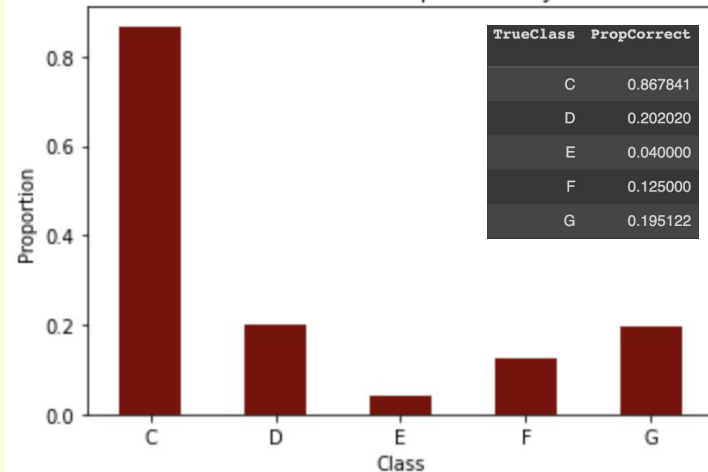
**Class C** (10 to 100 acres burned)

- Highest proportion of correct predictions
- Predicted most frequently (over other classes)
- Most represented in the training dataset

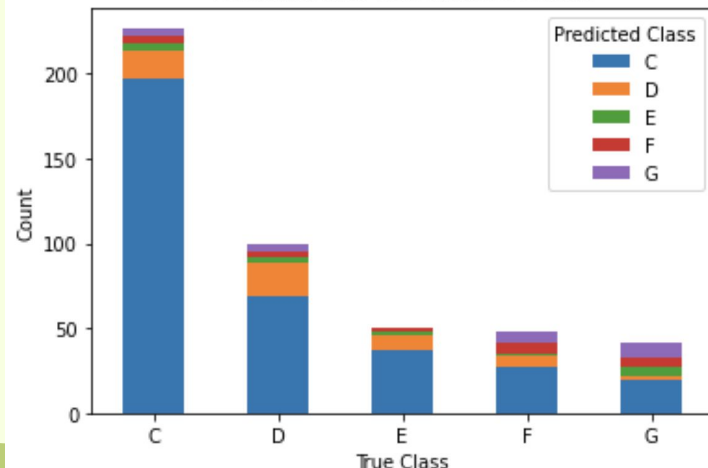
**Class E** (300 to 1000 acres burned)

- Lowest proportion of correct predictions

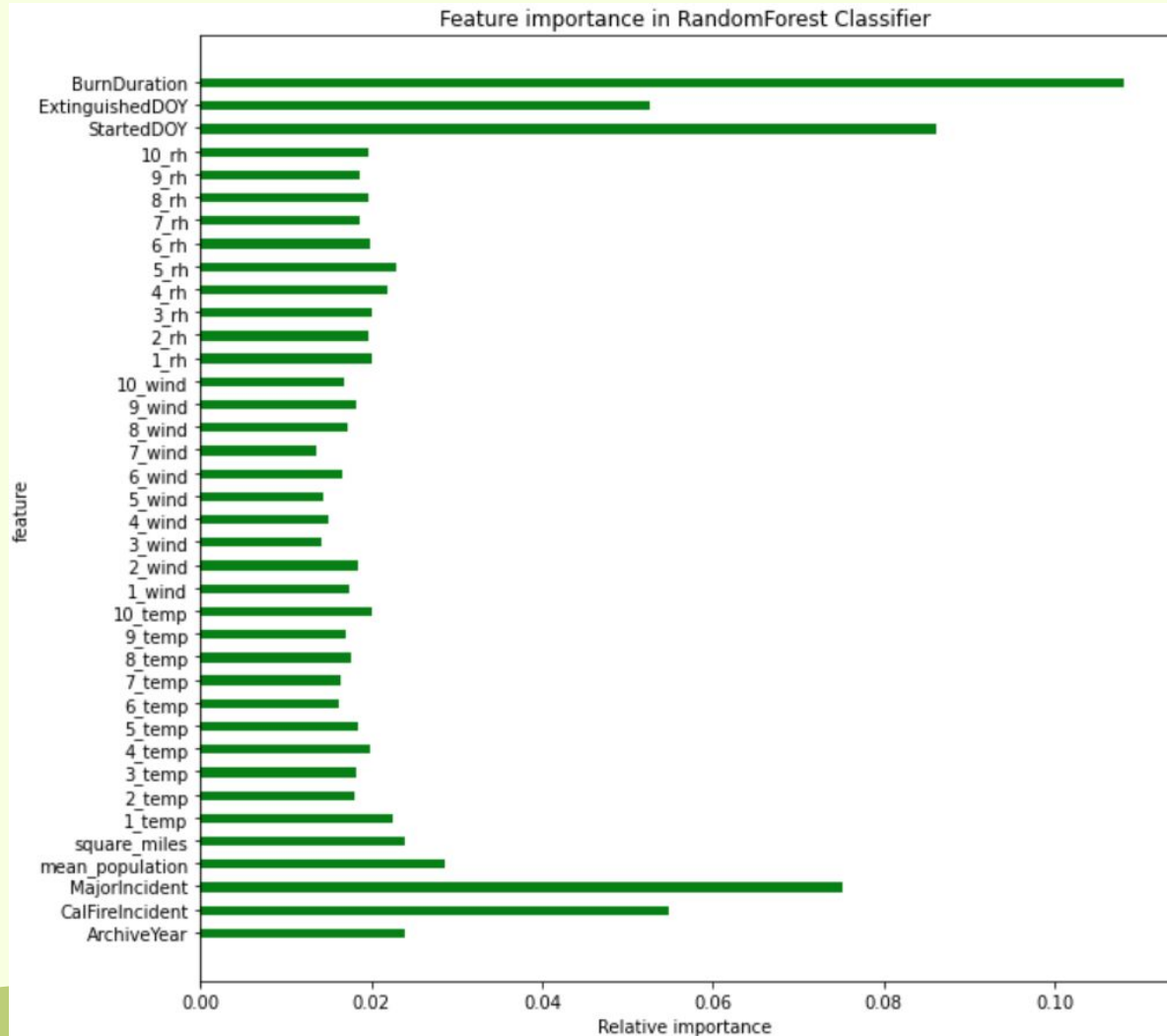
Correct Prediction Proportions, By Class



Predicted and True Wildfire Classes



# Model Feature Importance





05

# Conclusion

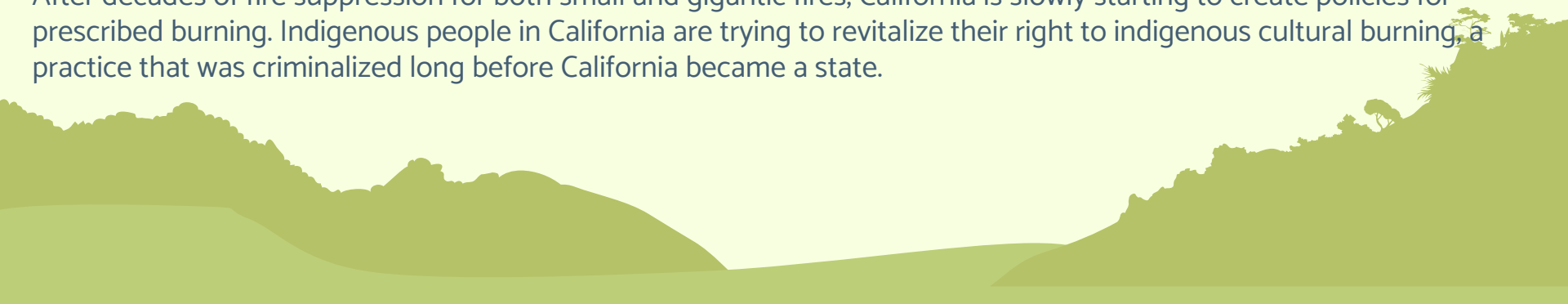
# Prescribed Burns + Policy Implications in California

For millennia, indigenous peoples used small intentional fires to renew local food, medicinal and cultural resources, create habitat for animals, and reduce the risk of larger, more dangerous wildfires.

The US government outlawed the process for a century before recognizing its value and the importance it plays in defending against large wildfires and its importance for the land.

The disregard for traditional ecological knowledge and the importance of prescribed burning in favor of a view that demonized all types of burning left the landscapes of California and other western states with an abundance of vegetation on the ground that was easily flammable. As the climate crisis creates hotter, drier, more volatile weather, that fuel has helped drive larger wildfires faster and further across the west.

After decades of fire suppression for both small and gigantic fires, California is slowly starting to create policies for prescribed burning. Indigenous people in California are trying to revitalize their right to indigenous cultural burning, a practice that was criminalized long before California became a state.





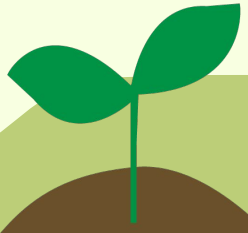
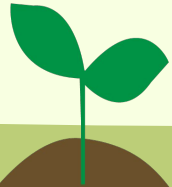


# Model Applicability

This model could be really useful for counties as well as policy making and the distribution of resources preceding the fire season. Fire preparedness is one of the most important factors in mitigating the intensity of the fire season and if the model predicts larger classes of fires for a county this could be used to try to employ better management techniques and have strong emergency preparedness protocols in place for that year.

# Future Model Improvements

- Adding columns for the central Latitude and Longitude of each county to capture any similarities between that county and surrounding counties.
- Improving spatial aspect of our model through spatial interpolation of missing values – Ordinary Least Squares or Geographically Weighted Regression.
- Autoregressive modeling to better incorporate date-time temporality – adding temporal aspect to features, changing granularity of data.
- Adding information on vegetation index and precipitation – time-related and access limitations.
- Adding information on prescribed burning – socio-political limitations.



# Works Cited



<https://www.theguardian.com/us-news/2019/nov/21/wildfire-prescribed-burns-california-native-americans>

<https://www.fire.ca.gov/incidents/>

<https://news.ucar.edu/132845/scientists-develop-method-seasonal-prediction-western-wildfires#:~:text=The%20new%20method%2C%20detailed%20in,States%20during%20the%20following%20summer>

<https://iopscience.iop.org/article/10.1088/1748-9326/ac6886/meta#erlac6886s2>

<https://blog.breezometer.com/how-do-wildfires-start-can-we-predict-them/>

<https://www.eurekalert.org/news-releases/618840>

<https://catalog.extension.oregonstate.edu/em9222/html>

