
LINEAR REGRESSION
CALIFORNIA WILDFIRES

BACK STORY

California Department of Forestry and Fire Protection (CalFire)

Acres burned Google Maps area tool
https://www.daftlogic.com/projects-google-maps-area-calculator-tool.htm#

BACK STORY

8k

wildfires

BACK STORY

8k

wildfires

5.8k

structures

BACK STORY

8k

wildfires

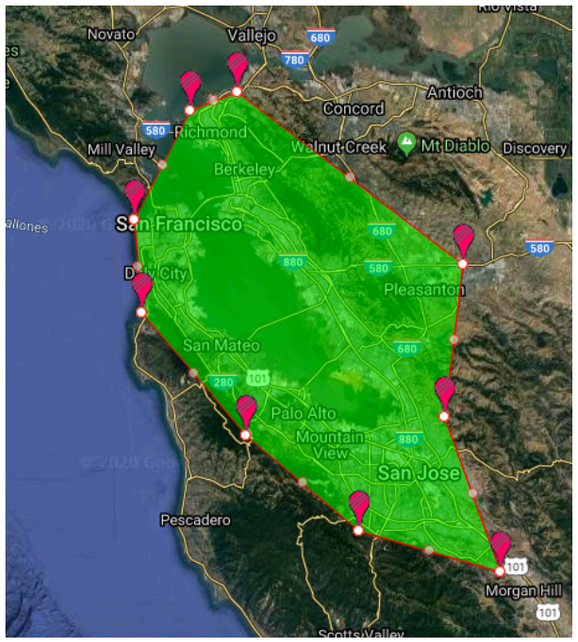
5.8k

structures

900k

acres burned

BACK STORY



https://www.daftlogic.com/

900k

acres burned

$400B

AccuWeather estimated the total damage and economic loss caused by wildfires cost California $400 billion in 2018 and $85 billion in 2017.
https://www.accuweather.com/en/weather-news/california-wildfires-will-cost-tens-of-billions-accuweather-estimates/612548

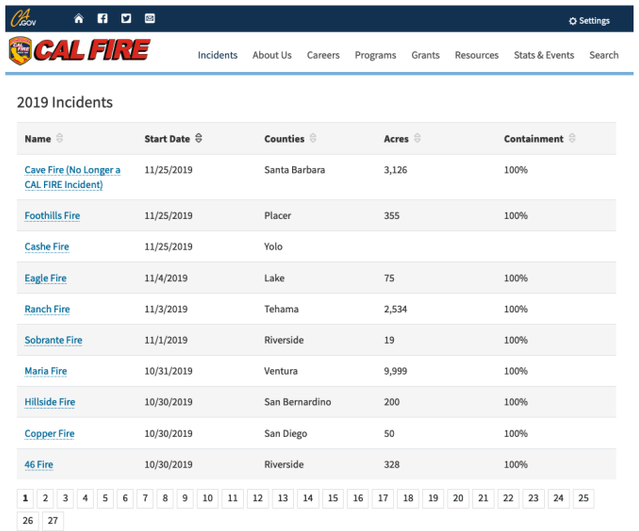
OBJECTIVE

Predict acres burned using linear correlation

Predict acres burned given a California wildfire by using linear regression to identify relationships between acres burned and **weather & time**

By accomplishing this, we can then look into how many resources might be needed to combat wildfires

METHODOLOGY



Using Python – my approach was to scrape all wildfire data available from Cal Fire (**last 7 years)**: acres burned, started, location by latitude/longitude
https://www.fire.ca.gov/

Using the Dark Sky API, I made requests get weather data: temperature, humidity, windspeed, etc.

My final dataset included about 1300 rows where each row represented a wildfire along with weather attributes

METHODOLOGY

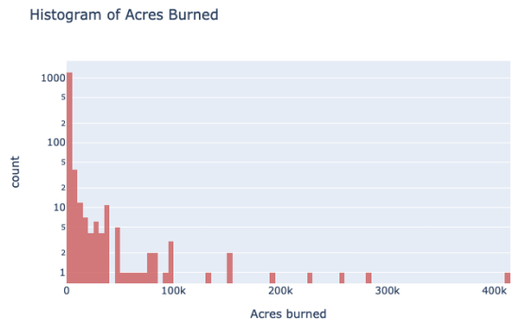
Dark Sky API

CAL FIRE

METHODOLOGY

METHODOLOGY

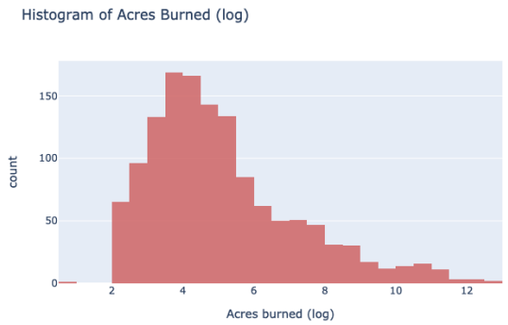
Transform target variable → acres burned



Acres burned is highly skewed right

METHODOLOGY

Transform target variable → acres burned



METHODOLOGY

1. Split my data into 80/20 train/test
2. Standardized the data so that all features are on the same scale
3. OLS Model – mainly to see summarized results (p-values)
    1. p-values helped identify features to drop
    2. R^2 after dropping features
4. Lasso Model
    1. Helped zero out coefficients
5. Prediction

METHODOLOGY

TRAIN/TEST

METHODOLOGY

TRAIN/TEST

STANDARDIZE

METHODOLOGY

TRAIN/TEST

STANDARDIZE

OLS

METHODOLOGY

TRAIN/TEST

STANDARDIZE

OLS

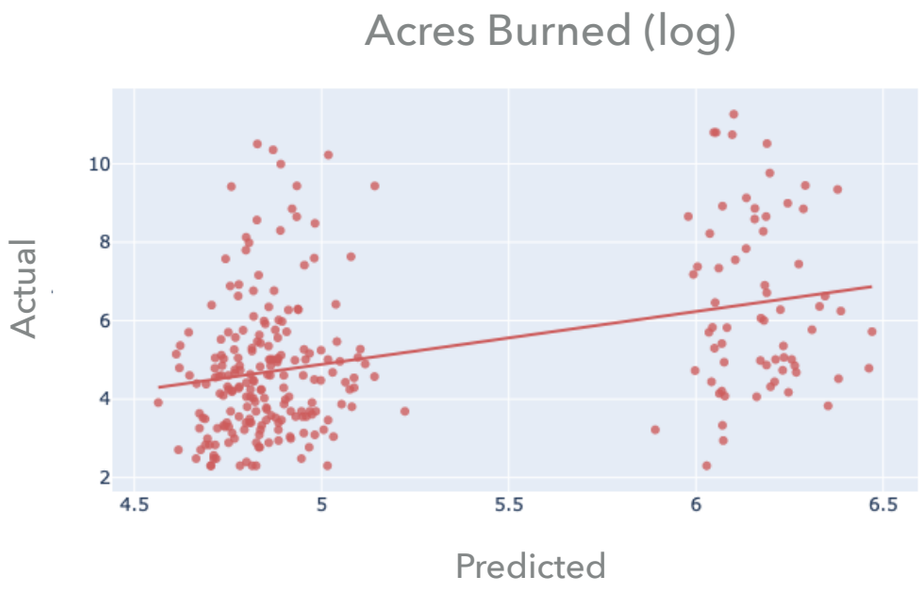
LASSO

METHODOLOGY

MODEL

RESULTS

R^2 score = 0.14



Ineffective model
Two clusters (bi-modal distribution?)

x = y_predict (lasso_model.predict(X_te)
y = y_test

FUTURE WORK

- Larger weather date range
- Categorical features
- Cross validation
- Model comparison

Need more data, particularly weather data to include a larger date range
Categorical features such as **season, location**

Quote
https://www.accuweather.com/en/weather-news/california-wildfires-will-cost-tens-of-billions-accuweather-estimates/612548

QUESTIONS

Questions:
Why did I only choose Lasso?
– mainly concerned with trimming features
– helped zero out coefficients

Number of features
– started with 13 and ended with 5