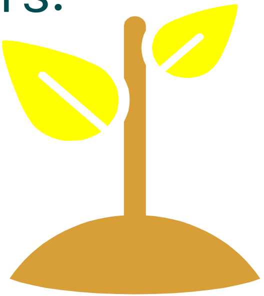
California Wildfire Damage Projection

Egan Bailey, Sean Gu, Ani Karenovna K. Douglas Nguyen, Amanda Jo Russell

Background

Wildfires in California are becoming increasingly-common due to a variety of factors:

- Climate change
- Drought Frequency & Seasonal Rains
- Human Causal Factors
- Historical Forest Mismanagement
- Increased population in vulnerable areas



Background (Cont.)

FEMA would benefit from a model projects the costs resulting from a fire disaster.

- Financial aid can then be properly allocated to the impacted communities.

Data Science Problem

Our Goal is to Predict the Dollar Amount of Damage done to specific counties within the state of California due to wildfires

Data Science Problem

Interaction between some of these variables have proven useful for making predictions:

1. Average size of the fire spread
2. Number of fire origins within a given area
3. Weather conditions
4. Variables representing relative county affluence
5. Percentages of wilderness/developed land in a given county

Data Science Process / Methodology

- Data Collection
- Cleaning data & EDA
- Feature Engineering
- Model Fitting & Tuning
- Evaluations

1.
Data Collection

Data Collection: Sources

- **Cal Fire: Redbooks (2010 - 2016)**
  - Number of Fires (sizes and causes)
  - Dollar Damage

- **Yelp**
  - Number of Campsites/RV Parks per County

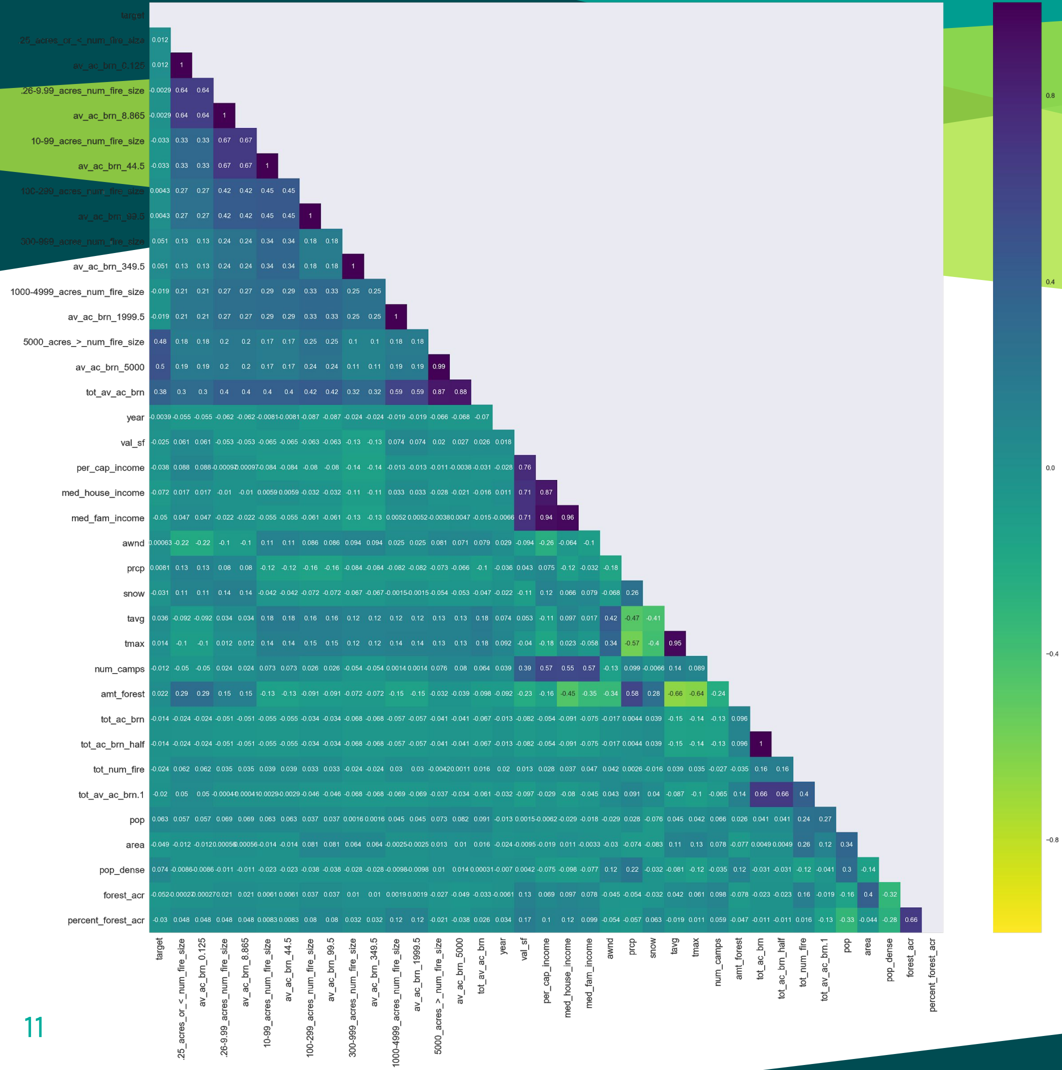
Data Collection: Sources

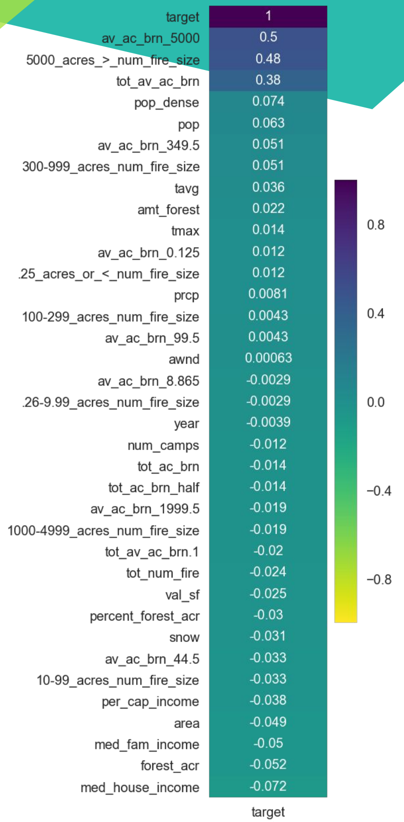
- National Oceanic and Atmospheric Administration (NOAA)
  - Weather Data
- National Census Bureau/Wikipedia
  - Population Data
- Employment Development Department
  - Workforce Data

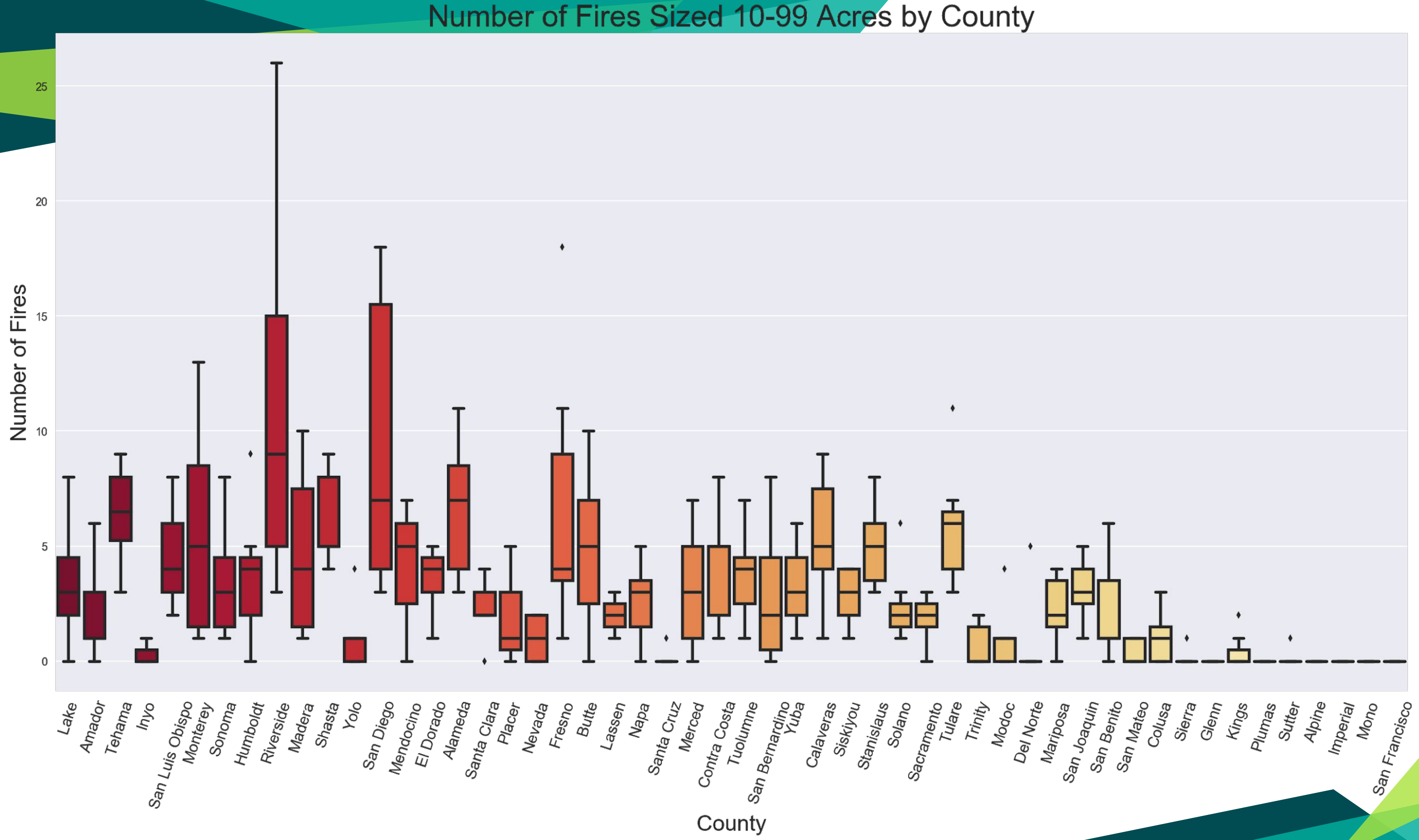
2.

Data Cleaning & EDA

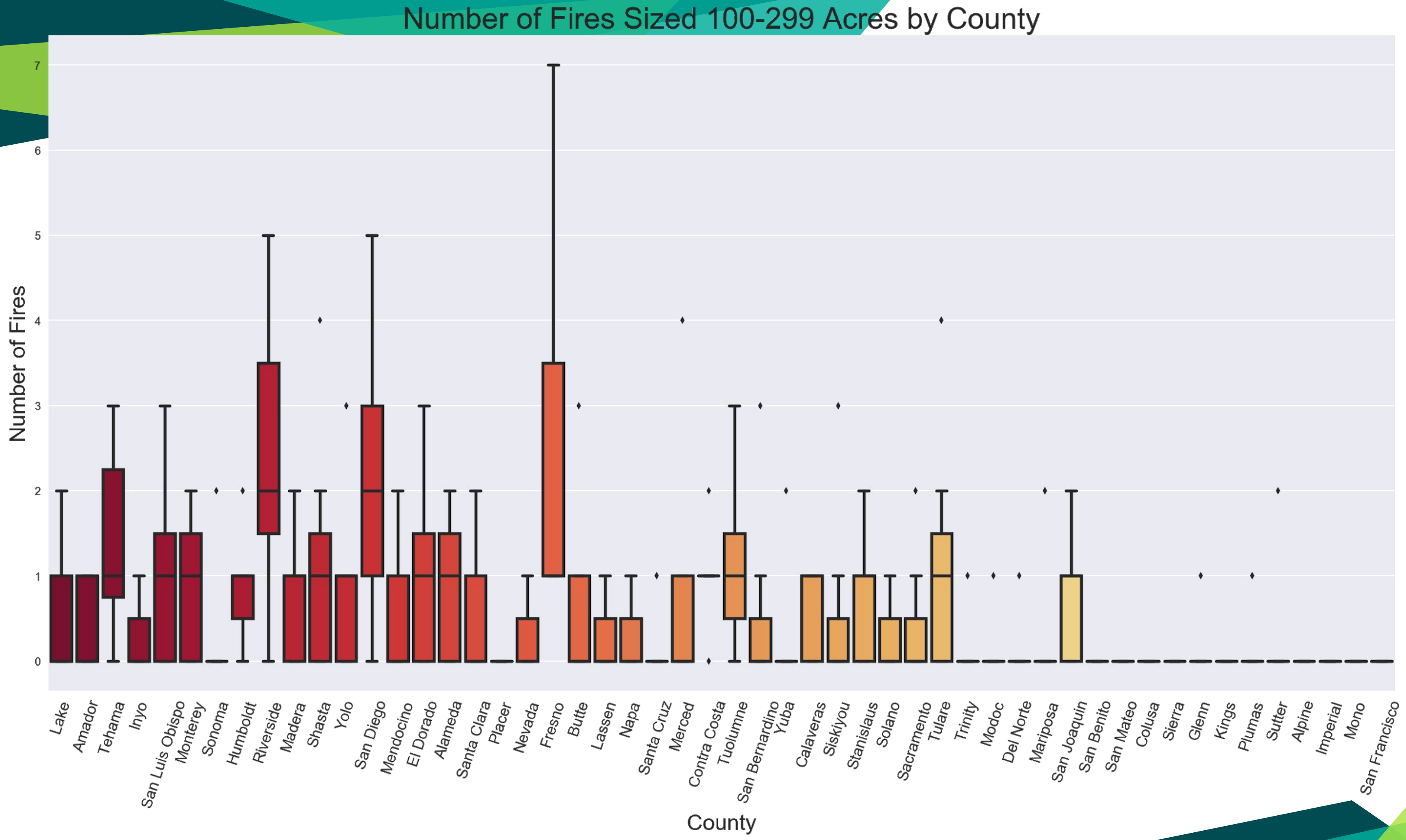
Correlation: Heatmaps



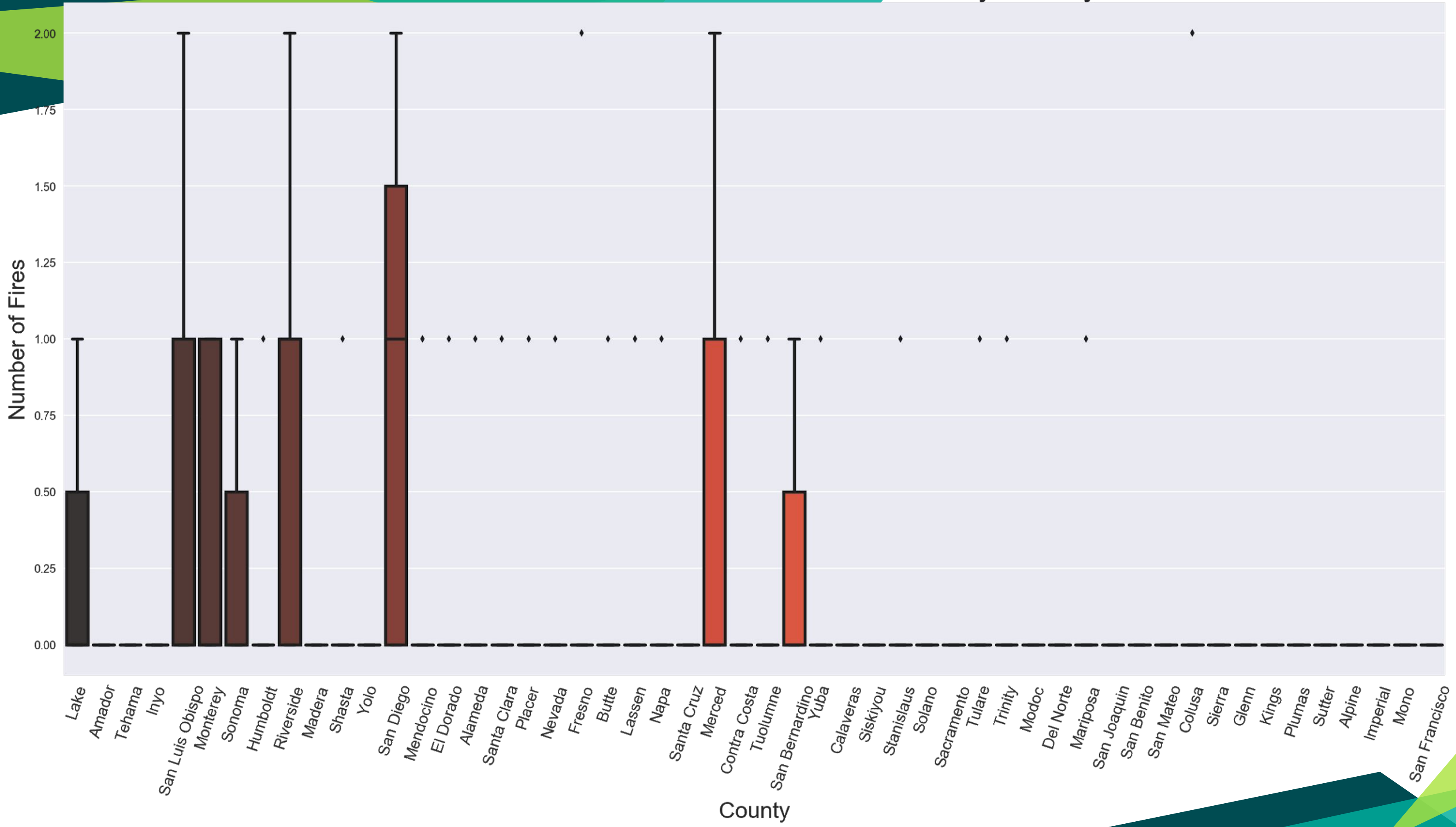


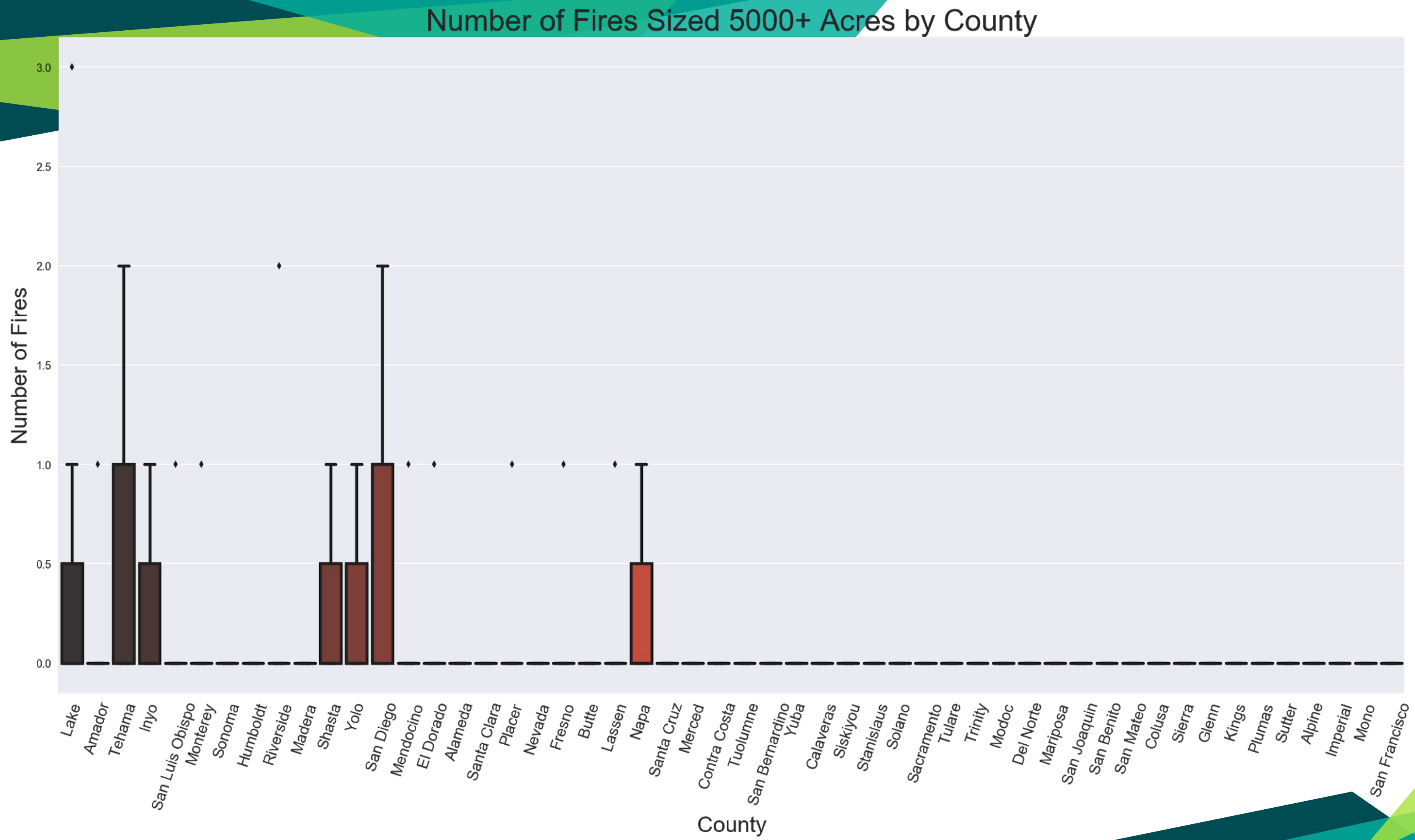


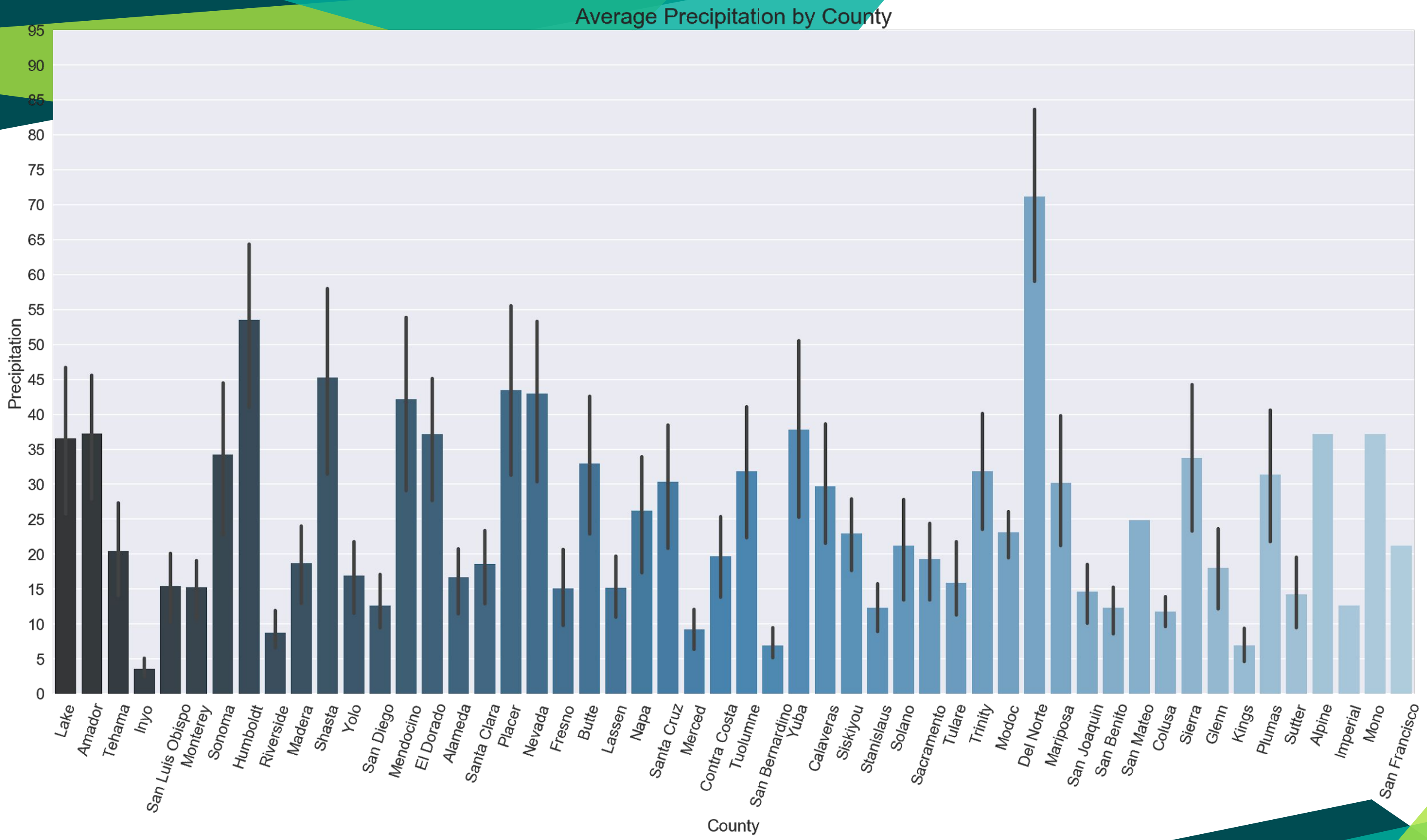
Number of Fires Sized 10-99 Acres by County
Number of Fires
0
5
10
15
20
25
Lake
Amador
Tehama
Inyo
San Luis Obispo
Monterey
Sonoma
Humboldt
Riverside
Madera
Shasta
Yolo
San Diego
Mendocino
El Dorado
Alameda
Santa Clara
Placer
Nevada
Fresno
Butte
Lassen
Napa
Santa Cruz
Merced
Contra Costa
Tuolumne
San Bernardino
Yuba
Calaveras
Siskiyou
Stanislaus
Solano
Sacramento
Tulare
Trinity
Modoc
Del Norte
Mariposa
San Joaquin
San Benito
San Mateo
Colusa
Sierra
Glenn
Kings
Plumas
Sutter
Alpine
Imperial
Mono
San Francisco
County

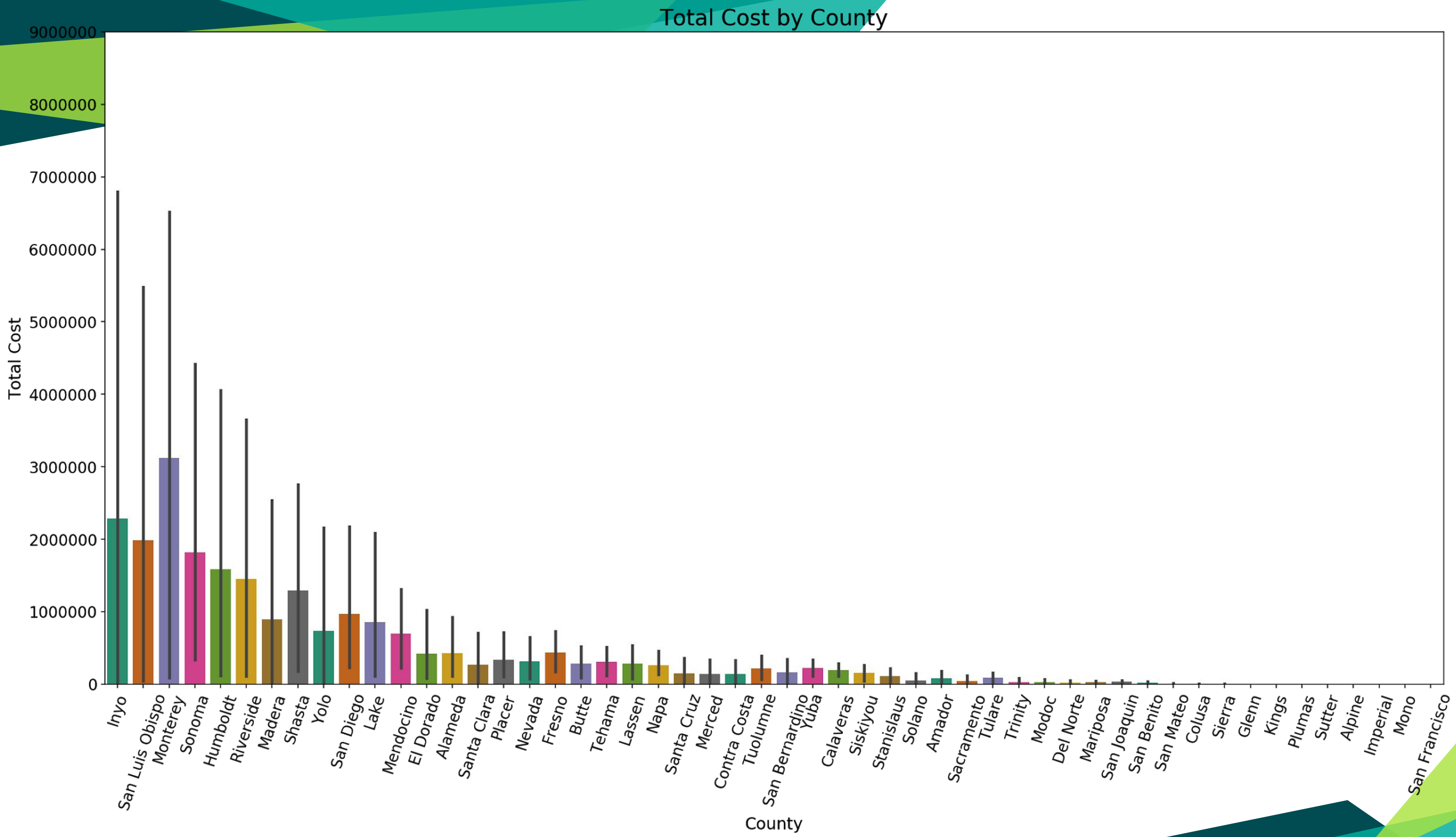

Number of Fires Sized 100-299 Acres by County
Number of Fires
0
1
2
3
4
5
6
7
Lake
Amador
Tehama
Inyo
San Luis Obispo
Monterey
Sonoma
Humboldt
Riverside
Madera
Shasta
Yolo
San Diego
Mendocino
El Dorado
Alameda
Santa Clara
Placer
Nevada
Fresno
Butte
Lassen
Napa
Santa Cruz
Merced
Contra Costa
Tuolumne
San Bernardino
Yuba
Calaveras
Siskiyou
Stanislaus
Solano
Sacramento
Tulare
Trinity
Modoc
Del Norte
Mariposa
San Joaquin
San Benito
San Mateo
Colusa
Sierra
Glenn
Kings
Plumas
Sutter
Alpine
Imperial
Mono
San Francisco
County

Number of Fires Sized 1000-4999 Acres by County




Number of Fires Sized 5000+ Acres by County
Number of Fires
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Lake
Amador
Tehama
Inyo
San Luis Obispo
Monterey
Sonoma
Humboldt
Riverside
Madera
Shasta
Yolo
San Diego
Mendocino
El Dorado
Alameda
Santa Clara
Placer
Nevada
Fresno
Butte
Lassen
Napa
Santa Cruz
Merced
Contra Costa
Tuolumne
San Bernardino
Yuba
Calaveras
Siskiyou
Stanislaus
Solano
Sacramento
Tulare
Trinity
Modoc
Del Norte
Mariposa
San Joaquin
San Benito
San Mateo
Colusa
Sierra
Glenn
Kings
Plumas
Sutter
Alpine
Imperial
Mono
San Francisco
County


Average Precipitation by County
Precipitation
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
Lake
Amador
Tehama
Inyo
San Luis Obispo
Monterey
Sonoma
Humboldt
Riverside
Madera
Shasta
Yolo
San Diego
Mendocino
El Dorado
Alameda
Santa Clara
Placer
Nevada
Fresno
Butte
Lassen
Napa
Santa Cruz
Merced
Contra Costa
Tuolumne
San Bernardino
Yuba
Calaveras
Siskiyou
Stanislaus
Solano
Sacramento
Tulare
Trinity
Modoc
Del Norte
Mariposa
San Joaquin
San Benito
San Mateo
Colusa
Sierra
Glenn
Kings
Plumas
Sutter
Alpine
Imperial
Mono
San Francisco
County


Total Cost by County
Total Cost
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Inyo
San Luis Obispo
Monterey
Sonoma
Humboldt
Riverside
Madera
Shasta
Yolo
San Diego
Lake
Mendocino
El Dorado
Alameda
Santa Clara
Placer
Nevada
Fresno
Butte
Tehama
Lassen
Napa
Santa Cruz
Merced
Contra Costa
Tuolumne
San Bernardino
Yuba
Calaveras
Siskiyou
Stanislaus
Solano
Amador
Sacramento
Tulare
Trinity
Modoc
Del Norte
Mariposa
San Joaquin
San Benito
San Mateo
Colusa
Sierra
Glenn
Kings
Plumas
Sutter
Alpine
Imperial
Mono
San Francisco
County

3.

Feature Engineering

4.

Model Fitting and Tuning

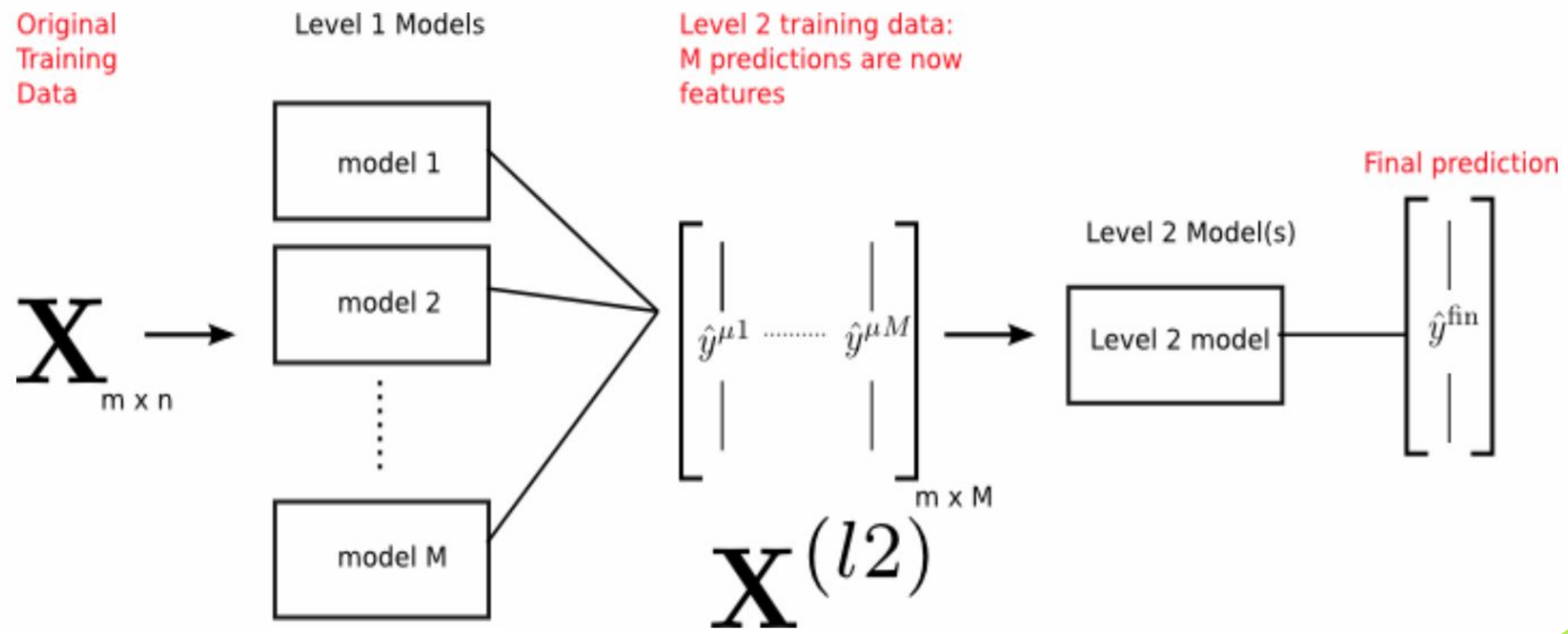
The Models: Regression

1. Multiple Linear Regression
2. Lasso
3. Ridge
4. KNN
5. SVR
6. Random Forest
7. Bagging
8. Extra Trees
9. Ada Boost
10. Gradient Boost

Each model was trained and fit in three different contexts:

1. Base Case - no tuning
2. Grid Search - tuning
3. Stacking - predictions from best gs models become the new features

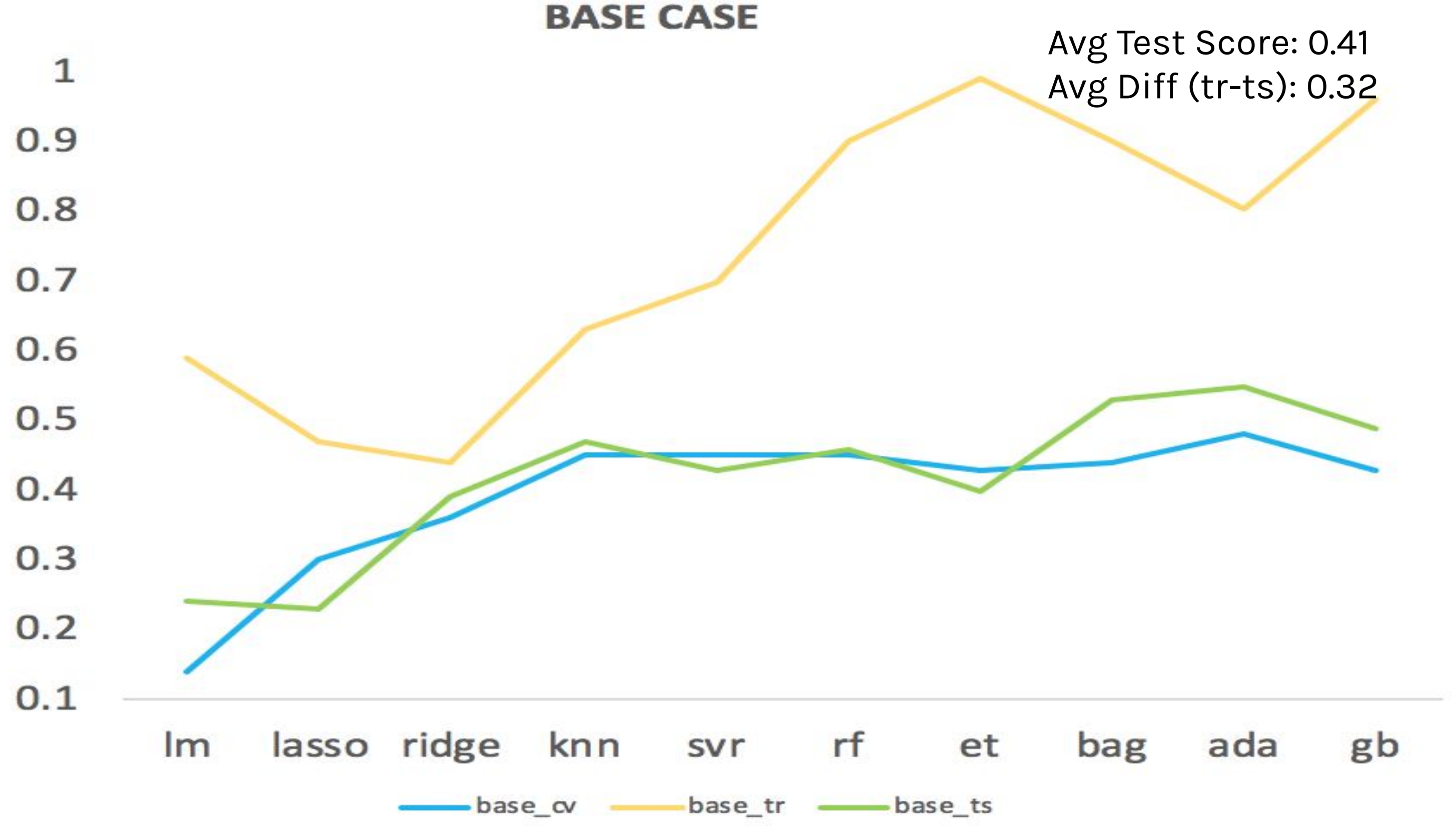
Stacking



Source: https://www.kdnuggets.com/2017/02/stacking-models-imropved-predictions.html

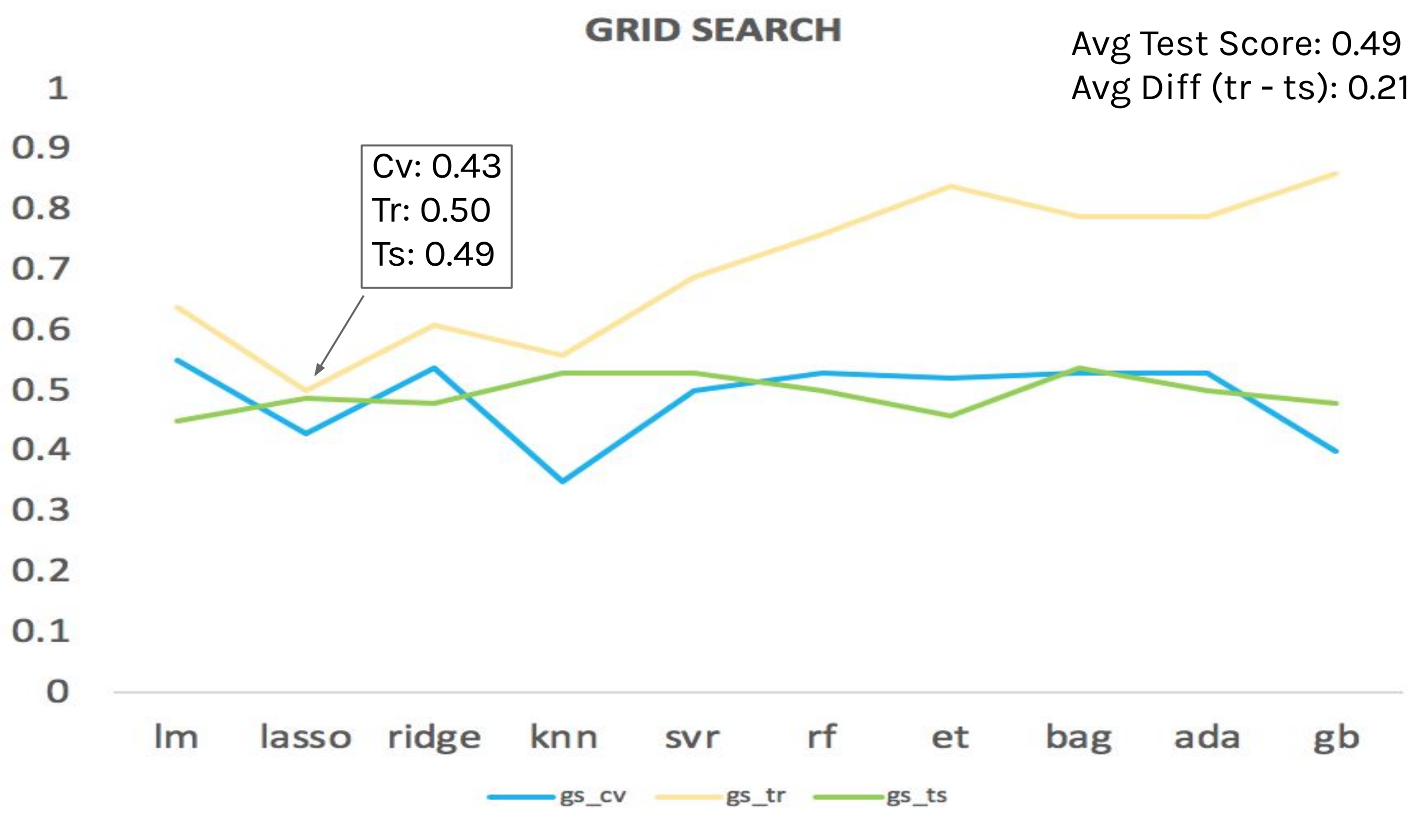
5.
Evaluation and Conclusions

Train Scores, Test Scores and CV Scores (R_squared) for Each Model



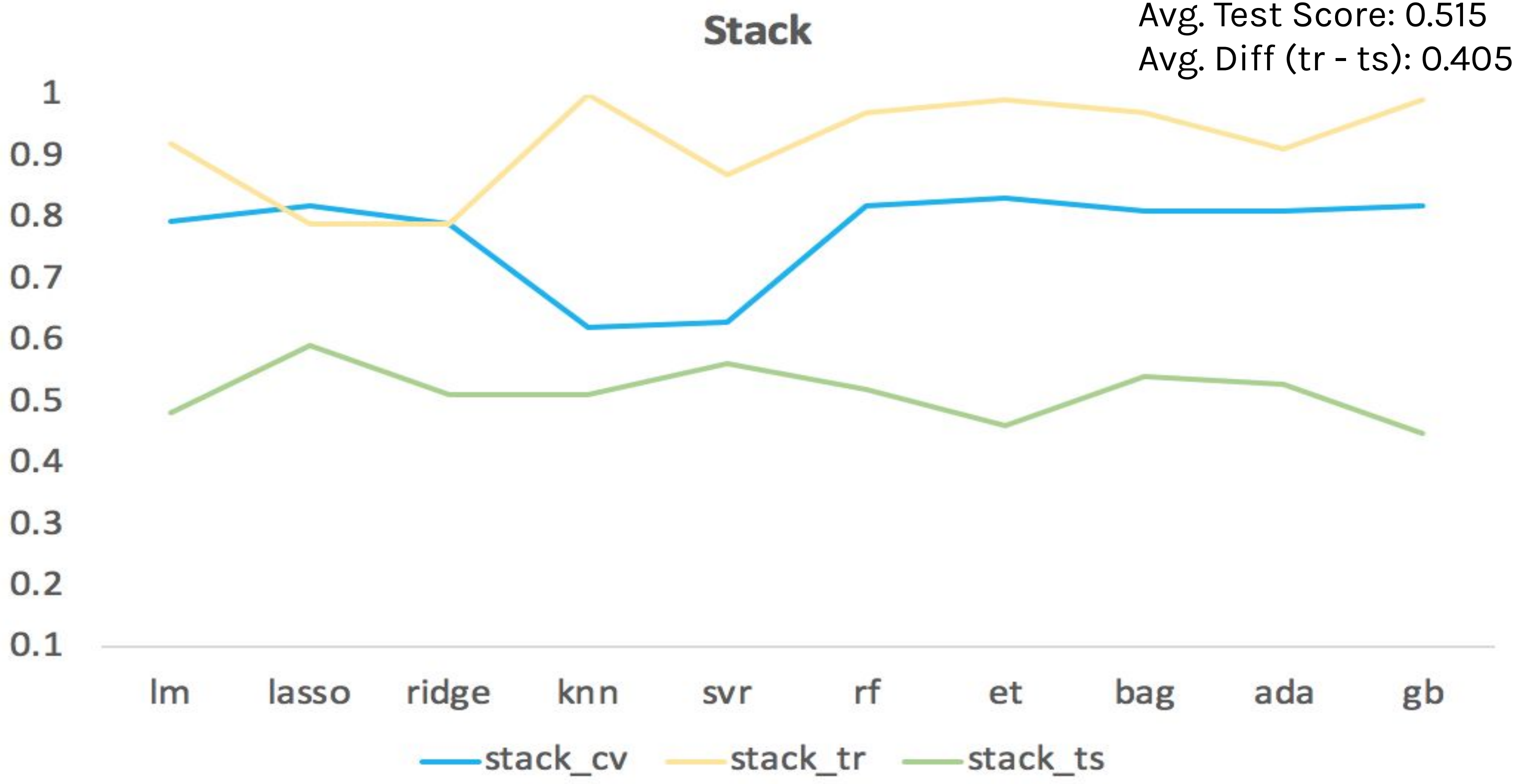
Avg Test Score: 0.41
Avg Diff (tr-ts): 0.32

Train Scores, Test Scores and CV Scores (R_squared) for Each Model



Train Scores, Test Scores and CV Scores (R_squared)for Each Model

Avg. Test Score: 0.515
Avg. Diff (tr - ts): 0.405



Demonstration of Our Solution

Interaction between some of these variables have proven useful for making predictions:

1. Average size of the fire spread
2. Number of fire origins within a given area
3. Weather conditions
4. Variables representing relative county affluence
5. Percentages of wilderness/developed land in a given county

How to Use the Model

- The best model is optimized for prediction, not for explainability
- Input: features
- Output: dollar amount damage

Limitations of the Model

- Data Collection Issues
  - Accuracy and completeness of raw data
- Missing values
  - Assumptions made while imputing missing values
- Overall complexity of variable space that are predictive of fire processes

Next Steps

- CalFire data that is more granular, so as to include detailed data on individual wildfires (e.g dollar amount damage, acres burned, etc).
- Data on the epicenter of fires to help determine spatial correlations.
- An examination of the impact of US Forest Service “controlled burns” and how it affects future fires.
- Insurance premium increases.
- Fire suppression/containment data.

Sources

- https://www.nytimes.com/2018/11/09/climate/why-california-fires.html
- https://www.kqed.org/science/1927354/controlled-burns-can-help-solve-californias-fire-problem-so-why-arent-there-more-of-them
- https://en.wikipedia.org/wiki/List_of_counties_in_California
- https://www.fire.ca.gov/fire_protection/fire_protection_fire_info_redbooks
- https://www.pewtrusts.org/en/research-and-analysis/reports/2018/06/19/what-we-dont-know-about-state-spending-on-natural-disasters-could-cost-us
- https://www.kdnuggets.com/2017/02/stacking-models-imropved-predictions.html
- https://www.codecademy.com/articles/seaborn-design-ii
- https://www.ncdc.noaa.gov/cdo-web/search

Thanks!

Any Questions?