California Wildfire County Predictions

Rachel Chiang

1. Introduction

In 2020 alone, wildfires burned 4.2 million acres of land, and 31 people died [1]. Responses to fires were strained; there were not enough resources and crews to take care of the fires. They were forced to request off-duty firefighters to return to work. Aside from direct destruction, fires indirectly affect other aspects of civilian life, including health and business. Evacuations and poor air quality can prevent people from maintaining and using their facilities for extended periods of time. For example, due to the Caldor Fire, a restaurant in South Lake Tahoe lost $10,000 to $13,000 in perishables, even though there was no structural damage to the city, and overall, the city estimated its loss to be over $50 million [2]. Insurance companies, local governments, or fire control may want to allocate resources or policies or deploy different awareness and mitigation strategies more specifically depending on each region's unique risk. This project aims to find counties that have fires and how many fires they may have each year.

2. Datasets

The project begins with looking at two datasets, a wildfire dataset and an environmental conditions dataset, which were both found on Kaggle but use information scraped from California government websites.

The wildfire dataset contains 1636 wildfire incidents with 40 columns in the years 2013-2020, scraped from the California Department of Forestry and Fire Protection (CAL FIRE) website [3]. The environmental conditions dataset contains about 128000 rows with 19 columns, and has records in 2017-2020, from the California Irrigation Management Information System (CIMIS) from the California Department of Water Resources [4]. Further station information was taken from CIMIS to expand the climate dataset with more location data. Some more data was collected for the project, including those about counties from Wikipedia [5] which itself sourced the population information from the 2022 census and the area information from the National Association of Counties (NACo).

3. Data Wrangling

At this step, data was collected and defined and underwent a preliminary cleaning. The wildfire dataset had a lot of missing values, and much of the information did not appear to be standardized, but at least some important numerical features were mostly provided, such as the names, location information, number of acres burned, and dates. Some information that would have been nice to know include the fuel type, cause of fire, and more robust location data; there is indeed a column for fuel type, but a vast majority of the datapoints did not have any information on this. There were some duplicate fires because fires that were in multiple counties often had their own entries, but the unique IDs column helped to clean them up.

On the other hand, the environmental conditions dataset was almost completely filled in and clean, and all of the ranges and few outliers looked acceptable. However, this dataset did not have much information on the location, as it only had the station ID and the CIMIS region, which do not relate very intuitively or directly with the wildfire dataset; the wildfire dataset used primarily county names and geographic coordinates. Furthermore, there were some stations that were omitted though they exist in the CIMIS databases. The reason for this is not known. When taken in relation with the years covered in the wildfire dataset, the environmental conditions dataset does not have information from 2013-2016.

For data descriptions, please view Appendix A.

4. Exploratory Data Analysis

A. Fire Data

The fire data was pruned and cleaned, and if a feature had more than 20% missing values, it was dropped. Other missing values that were deemed potentially important at the time were cross-referenced with online sources or filled by imputation.

Larger wildfire incidents' dates could be fixed by checking their values in the Redbooks, which are written and released by the California Department of Forestry and Fire Protection. For the bad dates of smaller fires, most of them were approximated by considering all three of the incident's dates (Started, Extinguished, and Updated) and occasionally the canonical URL (which corresponds to an incident's Started date), since in most cases, only one of the three were erroneous. Finally, if both failed, the yet remaining small fires were simply assigned to have lasted only one day.

The dataset contained some bad latitude and longitude coordinates, and an investigation revealed that in most cases, this was generally a repeat input, a swap of the two coordinates, a typo, or an arbitrary input. About 160 bad coordinates for latitude and longitude were found using a generous rectangular range of California's boundaries. There was only one coordinate that skipped this basic filter that was revealed in a plot of the coordinates. The bad coordinates were simply reassigned to match the center coordinates of the fire's county [6].

A new feature was created from existing features: the ActiveDays feature represents the number of days that the fire burned, or a subtraction of the Started date from the Extinguished date, which revealed that there were many unreliable Extinguished dates. There were a lot of dates that appeared to be default dates. Namely, some fires were extinguished in the following January, even tiny fires that may have started in early summer. The correct date was imputed primarily using medians of subsets of fires, since the mean can be very volatile because a few outlier fires can burn substantially longer than most. The other two features added to the fire DataFrame were the year and month of the started date.

B. Environmental Conditions Data

Counties were added to the environmental conditions (climate) dataset. There were not too many missing values, so they were imputed using the mean values within the same month from either

the same station in a different year or from the same county if that station always had missing values for every year and month combination. The climate data was also aggregated in months per county for further exploration.

C. County Data

County characteristics were scraped from the "List of counties in California" Wikipedia page. Two features were basically added to the existing county data: the population in 2022 and the area in square miles. We lack information about the environments of these fire incidents, but the size and population density may hint at environmental characteristics. Wildfires need fuel, which means that they would prefer places with a lot of things to burn, such as forests and chaparrals, so a larger size and lower population density may be favored by fires, though houses can serve as fuel as well. Other features were aggregated and engineered for the counties, such as population density, fire proportions, and monthly statistics.

D. Visualizations

The number of acres burned and days active have extreme outliers, and a vast majority of the fires were small and short. Summer and autumn months (June to October) experienced the most fires (as seen in Appendix B), though there were a few outliers in December and February, and the two years that had the most fires were 2017 and 2018.

When all fires are compared to the environmental conditions, there does not seem to be strong correlations between them, but when looking at only the larger fires, there seems to be a bit more correlation with climate; however, it is still not particularly strong. Certain climate conditions may spread fires more and other climate conditions may obstruct response teams. Perhaps small fires may more often be started by random or unnatural causes that make them more unrelated to environmental conditions.

To illustrate the fires' general locations, a quick scatterplot (Appendix C) can be generated using their geographic coordinates, but a heatmap (Fig. 1) can more clearly depict intensities and frequencies of the wildfires. The maps clearly show specific areas where wildfires more commonly and intensely occur. These are often mountainous regions with a lot of forests and chaparrals. Sunnier, hotter regions such as those along the southern part of California also seem to experience a lot of wildfires.

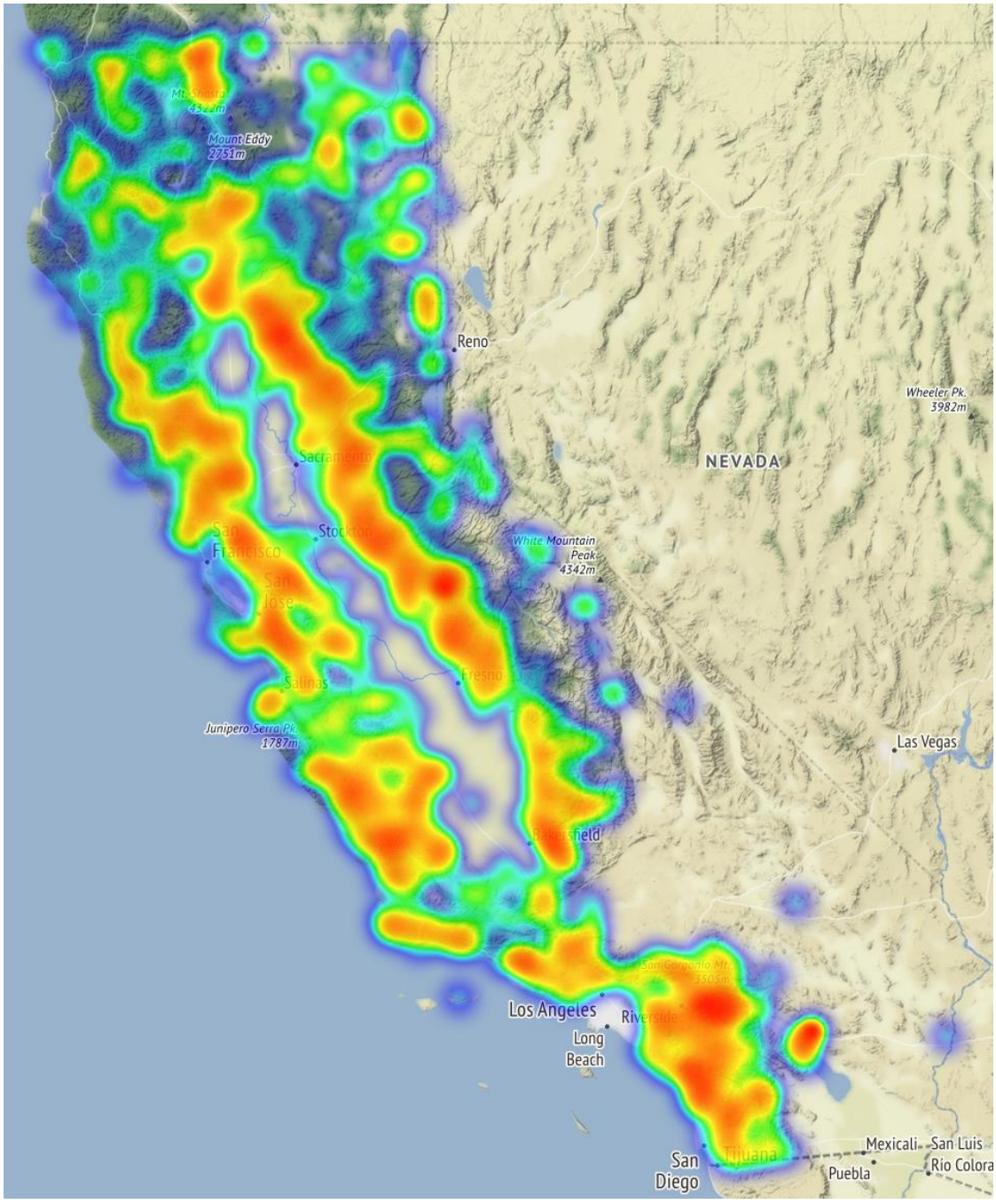


Fig. 1. Heatmap of wildfires in 2013-2020.

While we do not have particularly robust or detailed geographic data, we do have some basic county information. The numbers and sizes of fires tend to be slightly higher with larger counties, and counties with lower populations tend to have larger and more numerous fires. Although the relationship between the size of a county and the number of fires in the county (Fig. 2) is not extraordinarily strong, it may suffice for the scope of this project.

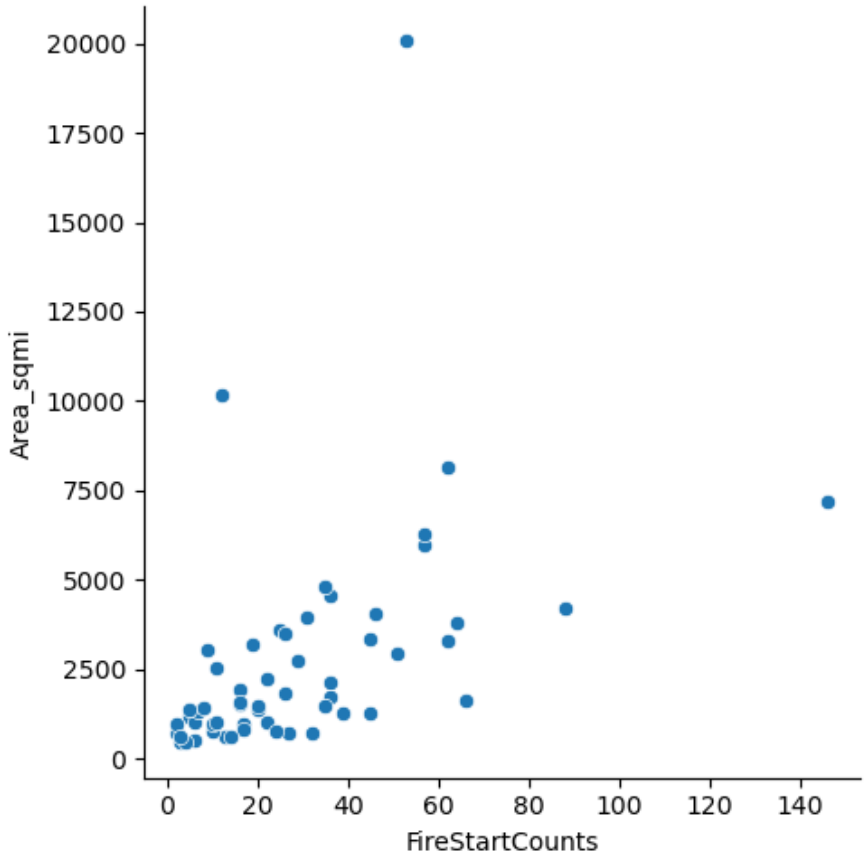


Fig. 2. Area in square miles of a county versus the number of fires in the county.

Even so, it should be mentioned that using counties to aggregate the data is not specific enough in representing its local area because counties have irregular boundaries and diverse terrains. The combined latitude and longitude are probably more important in generalizing the environment of the fires. However, aggregation by county is perhaps not wholly irrelevant because larger counties tend to encompass one or more areas that may frequently have fires, so a county's size may be able to abstractly describe these biomes somewhat. By using county-based information though, climate information loses some value because the climate can vary drastically within a county. For example, within San Bernardino County, certain areas like Chino or Redlands tend to be hotter throughout the year, yet it can be snowing at higher altitudes, such as in the San Bernardino Mountains. There is also perhaps another useful point about generalizing by county: policies, economics, and resources may be divided by county, so maybe we still should care about the countywide statistics.

5. Modeling

To prepare the data, fires were aggregated per county per year, and zeroes were filled in appropriately for counties and years that did not have fires. Each county was also dummy encoded, and Alameda was dropped. For the training and testing split, the data were split into two partitions of 290 and 116 data points, which is a test split of about 28.6%, with the training

set containing the first five years 2013-2017 and the testing set containing the latter two years 2018-2019.

Five preliminary models were quickly performed and evaluated: LinearRegression, Ridge, RandomForestRegressor, GradientBoostingRegressor, and AdaBoostRegressor. Some of the metrics are displayed in the table of Appendix D. The linear model performed decently with an R-squared of 0.57, and the closer that the Ridge model's alpha was to zero, the better it performed. Although the three tree-based algorithms were tested with arbitrary hyperparameters, they performed decently enough, so the Random Forest and Gradient Boosting models were moved forward to hyperparameter tuning using GridSearchCV.

The hyperparameters selected for the Random Forest regressor were the number of max features (max_features) of 0.2 and the number of estimators (n_estimators) of 297. The hyperparameters selected for the Gradient Boosting regressor were the learning rate (learning_rate) of 0.4, max depth (max_depth) of 2, and the number of estimators (n_estimators) of 12.

Cross-validation of the two models found that the gradient boosting regressor's R-squared fell into a generally better range, but the medians of both were comparable. When scored using the testing set—entirely unseen data, the random forest regressor slightly outperformed the gradient boosting regressor. This could perhaps be explained by the fact that the random forest regressor puts at least some weight on every county and is more exhaustive in its evaluations, whereas the gradient boosting model disregards most of the counties and puts much more weight on just the area (size) of the county. The tuned random forest model also performs slightly better than the first linear regression model.

Model	R^2	RMSE	MAE
Random Forest	0.575	3.324	2.221
Gradient Boosting	0.490	3.641	2.506

Table 1. Random forest and gradient boosting regressors and their R-squared (R^2), root mean squared error (RMSE), and mean absolute error (MAE).

In terms of prediction, the random forest model would be the better choice. However, the gradient boosting model was about 20 and 40 times faster in learning and prediction computation speed, respectively. Still, there is only a finite number of counties, and the period is yearly, so having a longer computation time may not be too problematic.

6. Conclusion and Future Work

The random forest model predicts the number of fires in each county per year. This can perhaps be used by groups like CAL FIRE for resource allocation and preparation of response teams or equipment to high-risk areas, or it could be used by insurance companies to help understand which counties are likely to experience high risk in a year.

Wildfires are complex and difficult to predict in both their beginnings and their degrees of destruction. There is still a lot of work that could be done, and this model is far from perfect. Using merely county information and historical counts of fires is rather broad, whereas fires likely rely on more granular, local information. Other ideas and considerations came to mind

while working on this project. Rather than per county yearly fire predictions, perhaps we could develop a model that produces a hierarchy of fire risk for counties, or the risk of spread of fire could be predicted or visualized.

Having more data about the environment could be useful in understanding the likelihood and behavior of wildfires. Perhaps green areas in satellite images could help determine dense areas of natural fuels, such as forests. In a similar vein, oxygen (or, conversely, carbon dioxide) levels in an area may help describe fuel potential for fires. Adding or identifying the elevation or aspect of a landscape could also be useful in understanding a wildfire's behavior. People's photographs of terrain coupled with image recognition could also help understand extremely specific conditions about the density of flammable objects.

On the other hand, we began with a lot of data, and much of it went unused, but some of it could be explored further. Using geographic coordinates may be better than just using county data because counties can be too diverse, or using some other way to organize and divide the areas of California may be more helpful in generalizing the risk of an area. Maybe a similar prediction to this project could be performed, but instead of using a period of a year, it used periods of a month or season. Looking at the climate data with the fires may also yield more insights.

References

[1] J. Cart, "California's 2020 fire siege: wildfires by the numbers," *Cal Matters*, July 2021. <https://calmatters.org/economy/2021/10/california-wildfires-economic-impact/>

[2] G. Gedye, "How much do wildfires really cost California's economy?" *Cal Matters*, October 2021. <https://calmatters.org/environment/2021/07/california-fires-2020/>

[3] ARES, "California Wildfires (2013-2020)," Kaggle, 2020. <https://www.kaggle.com/datasets/ananthu017/california-wildfire-incidents-20132020>

[4] C. Zaloumis, "California Environmental Conditions Dataset," Kaggle, 2020. <https://www.kaggle.com/datasets/chelseazaloumis/cimis-dataset-with-fire-target>

[5] "List of counties in California," *Wikipedia*. <https://simple.wikipedia.org/wiki/List_of_counties_in_California>

[6] "US County Boundaries," *opendatasoft*. <https://public.opendatasoft.com/explore/dataset/us-county-boundaries/table/>

Appendices

Appendix A

Original Wildfire and Environmental Conditions Datasets

Original wildfire dataset in fire_df

Column	Data Type	Description
AcresBurned	float64	Acres of land affected by wildfires
Active	bool	Is the fire active or contained?
AdminUnit	object	Administrative unit
AirTankers	float64	Resources assigned
ArchiveYear	int64	Year the data was archived
CalFireIncident	bool	Is the incident categorized as a CalFire incident?
CanonicalUrl	object	Substring of URL for information source
ConditionStatement	object	Descriptive status updates
ControlStatement	object	Information about current road closures and threats
Counties	object	Name of county origin
CountyIds	object	List of county IDs involved
CrewsInvolved	float64	Resources assigned
Dozers	float64	Resources assigned
Engines	float64	Resources assigned
Extinguished	object	Date the fire was extinguished
Fatalities	float64	Fatality count
Featured	bool	(Unknown but was 98% False)
Final	bool	(Unknown but was almost 100% True)
FuelType	object	Fuel type of the fire
Helicopters	float64	Resources assigned
Injuries	float64	Count of injured personnel
Latitude	float64	Latitude of the incident
Location	object	Description of the location
Longitude	float64	Longitude of the incident
MajorIncident	bool	Is the fire considered major or not?
Name	object	Name of the wildfire
PercentContained	float64	What percent of the fire is contained?
PersonnelInvolved	float64	Resources assigned
Public	bool	(Unknown but was 100% True)
SearchDescription	object	“Description” meta content in HTML head
SearchKeywords	object	“Keywords” meta content in HTML head
Started	object	Date the fire started
Status	object	Status of the fire
StructuresDamaged	float64	Count of structures damaged
StructuresDestroyed	float64	Count of structures destroyed
StructuresEvacuated	float64	Count of structures evacuated

StructuresThreatened	float64	Count of structures threatened
UniqueId	object	Unique ID for the wildfire incident
Updated	object	Last date of update
WaterTenders	float64	Resources assigned

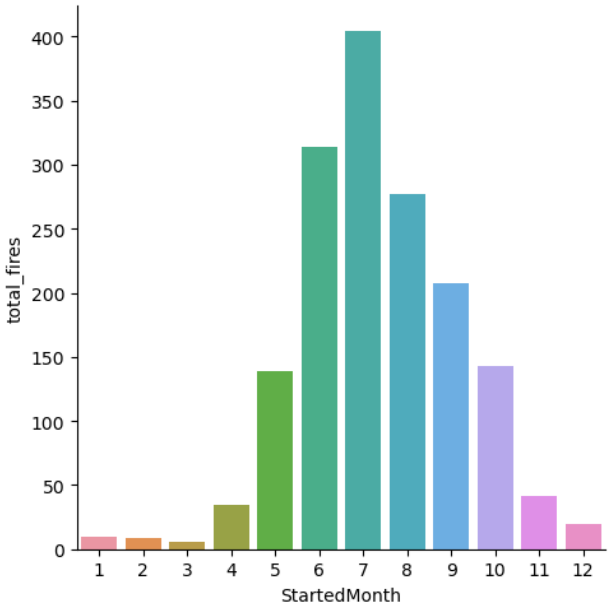
Original environmental conditions dataset in climate_df

Column	Data Type	Description
Stn Id	int64	ID of the station
Stn Name	object	Name of the station
CIMIS Region	object	Region of the station
Date	object	Date of the measurements
ETo (in)	float64	Reference evapotranspiration
Precip (in)	float64	Precipitation
Sol Rad (Ly/day)	float64	Average solar radiation
Avg Vap Pres (mBars)	float64	Average vapor pressure
Max Air Temp (F)	float64	Maximum air temperature
Min Air Temp (F)	float64	Minimum air temperature
Avg Air Temp (F)	float64	Average air temperature
Max Rel Hum (%)	float64	Maximum relative humidity
Min Rel Hum (%)	float64	Average relative humidity
Avg Rel Hum (%)	float64	Average air temperature
Dew Point (F)	float64	Dew point
Avg Wind Speed (mph)	float64	Average wind speed
Wind Run (miles)	float64	Wind run
Avg Soil Temp (F)	float64	Average soil temperature
Target	int64	Geography or weather station of interest

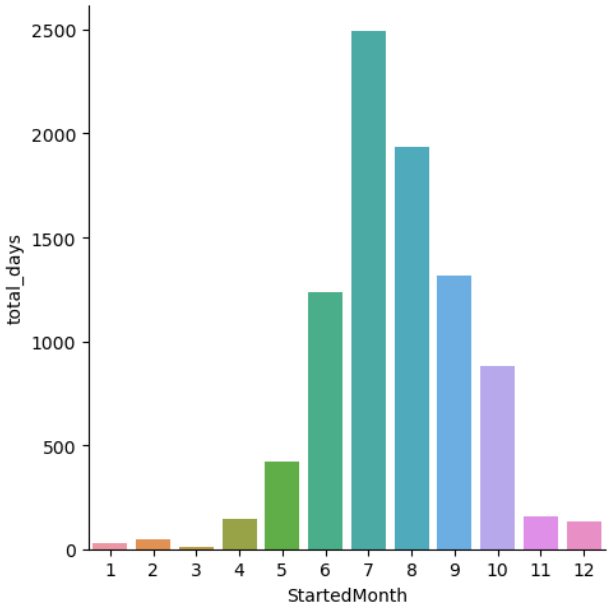
Appendix B

Monthly Fire Figures

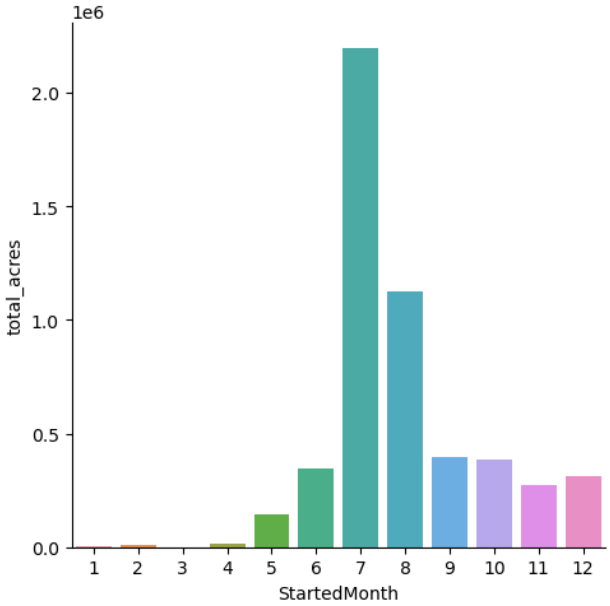
Number of fires started per month



Lengths of fires in days started in a month



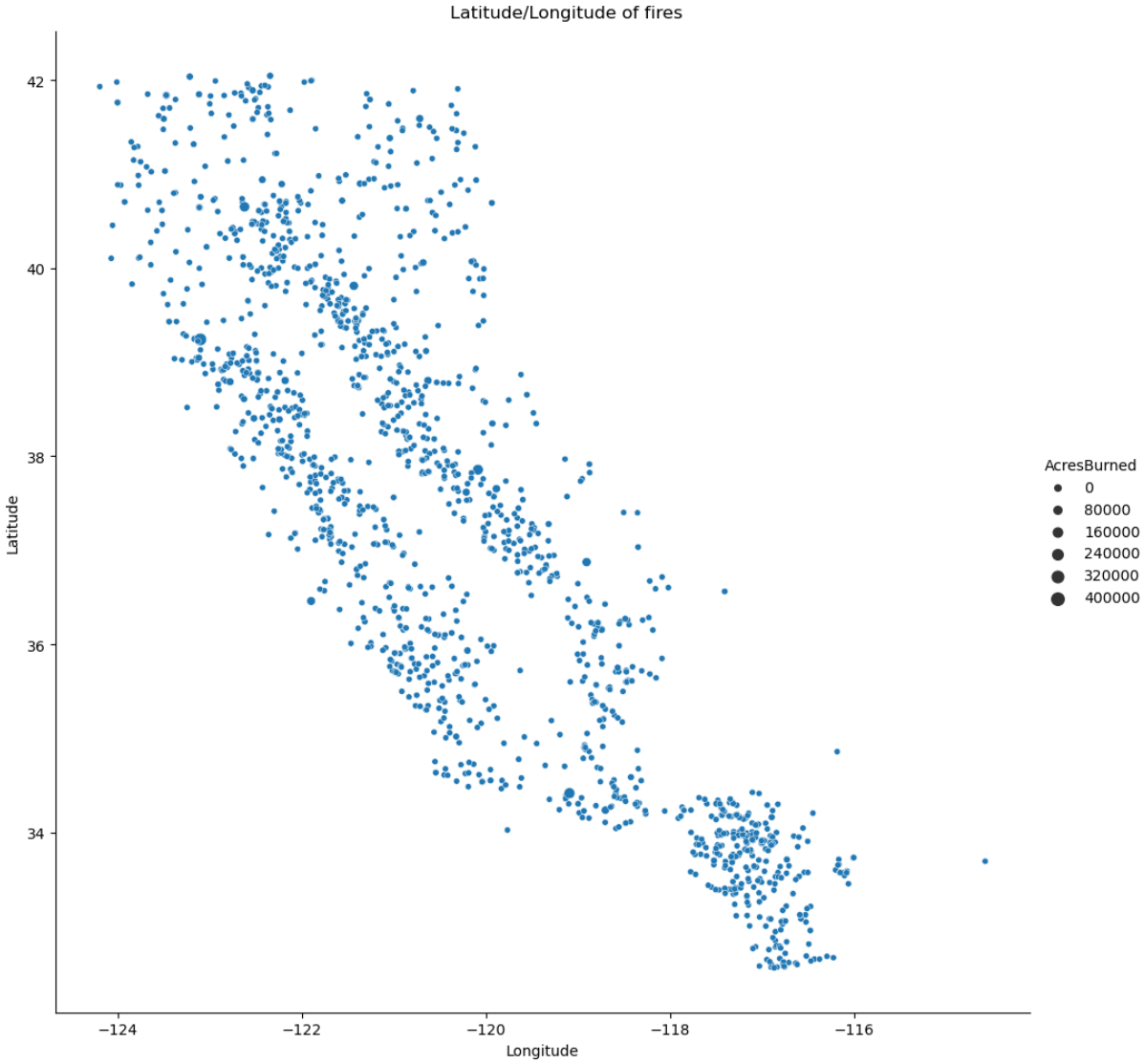
Total acres burned of fires started in a month



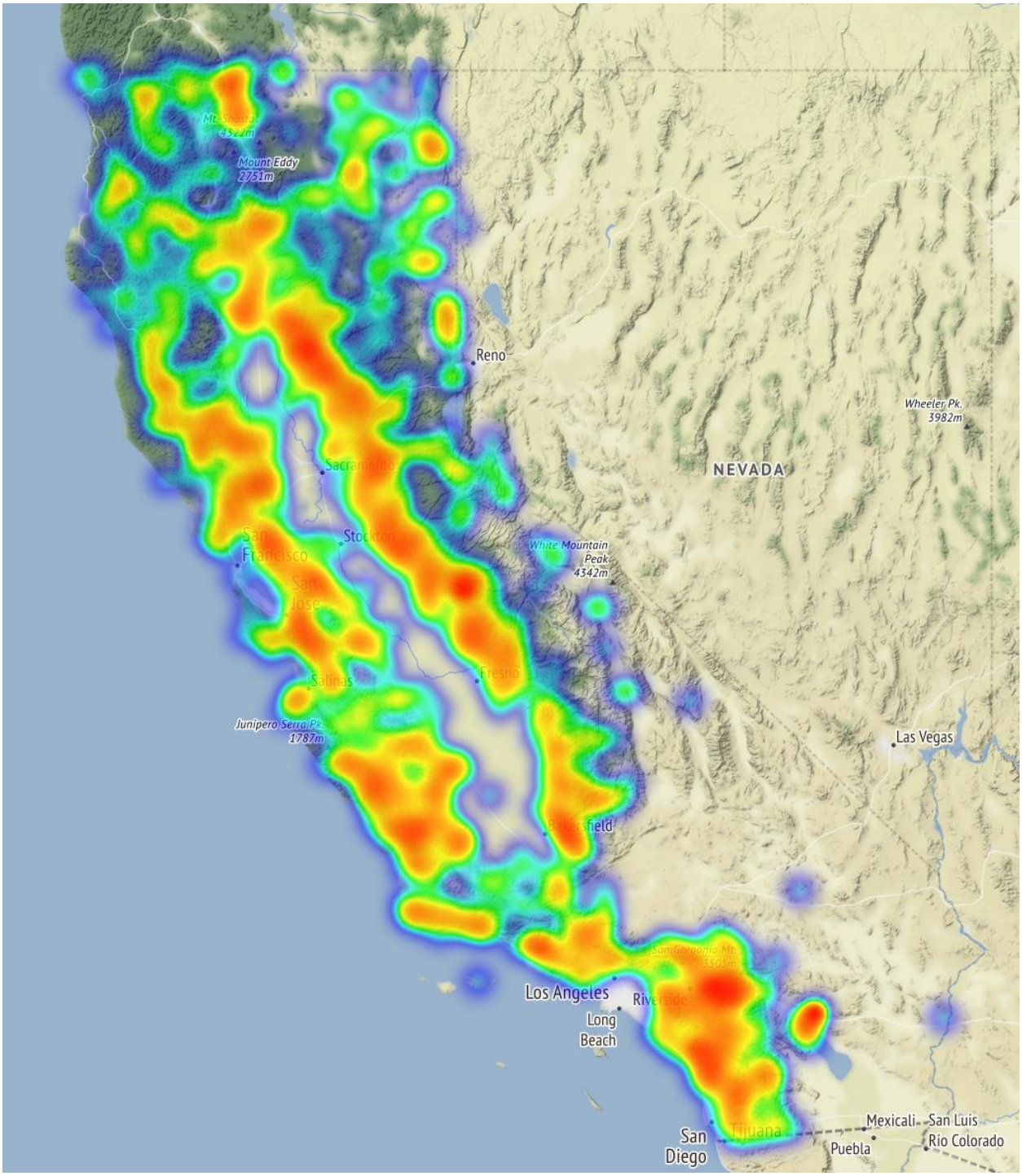
Appendix C

California Wildfire Maps for 2013-2020

Scatter plot of latitude and longitude of fires



Heatmap of all fires



Appendix D

Preliminary Model Metrics

Model	R^2	RMSE	MAE
Linear	0.574	3.328	2.224
Ridge α=0.01	0.573	3.330	2.226
Ridge α=0.1	0.568	3.349	2.236
Ridge α=1.0	0.528	3.503	2.311
Ridge α=10.0	0.292	4.290	2.849
Random Forest	0.379	4.017	2.831
Gradient Boosting	0.455	3.763	2.589
Ada Boost	0.520	3.533	2.608