DS 5500 – Final Report (Half-Semester Project)

Title: Exploring Wildfires: Can Only *We* Prevent Forest Fires?

Authors: Gopalika Sharma and Surya Menon

GitHub: https://github.com/gsharma14/DS-5500-California-Wildfires

Summary:

The increasing number of forest fires throughout the world is one the biggest threats to our lives and the environment. Climate change has produced higher temperatures and drier conditions, which have resulted in longer and harsher fire seasons, particularly in the western United States (1). While forest fires are natural processes that can benefit the environment—by clearing heavy brush, releasing nutrients into the soil, and facilitating seed germination for tree and plant species—these fires are becoming larger and harder to manage. As a result, we have observed increasingly catastrophic and devastating consequences, including the destruction of property, wildlife, and human lives (2).

In fact, during this year's wildfire season, which can be seen in Figure 1, California has experienced over 8,000 fire incidents across the state, with over 4 million acres burned, thousands of structures damaged or destroyed, and 31 fatalities as of October 26, 2020 (3). The costly interactions between humans and wildfires underline the importance in understanding the relationship between them, especially in the face of changing climate conditions and expanding human communities. The uncertainty about the size, location, and number of future wildfires establishes an intriguing analytical and predictive modeling challenge.

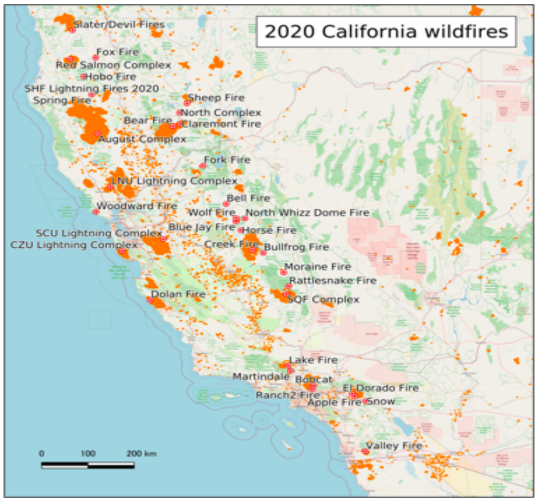


Figure 1: Forest Fire Spread in California, 2020

This capstone project aimed to research various aspects of forest fires in order to discover relevant spatial patterns and predictive elements of fire incidents that governments—such as the state of California—can use to develop strategies to manage the impact of future fires. Multiple datasets were used to achieve these goals, including a geospatial dataset for historic wildfires in the United States (US) between 1992-2015 and a dataset recording the forest fire and weather conditions (e.g., temperature, wind, humidity, etc.) for a set of fires located in northeast Portugal (4,5).

The spatial distribution of wildfires, specifically in California, was explored to find interesting patterns and trends over time with regards to the causes and severity of fire incidents. To better assess the overall impact of wildfires across the state, these fire occurrences were also evaluated alongside other spatial factors. We then created a dashboard to explore the spatial and historical patterns of California fires, including how the size and cause of fires vary across the state and over time; effectively visualizing this information can help the state in better analyzing wildfires and determining where additional intervention may be necessary in the future to minimize damage.

Additionally, we developed machine learning models to predict fire size (regression) and cause (classification) and identified the most significant features in the data for these modeling tasks. Better anticipating the fire severity or the possible cause can be instrumental in reducing future destruction in a community by ensuring resources are in place for safety, management, and faster containment. Both the Portugal and US fires (subset on California) datasets were used to conduct the regression task; these datasets have different features related to fire incidents (weather versus spatial factors), and we used the Portugal model results to inform which factors were most relevant in predicting size and augmented the California dataset (with climate data) accordingly. The California dataset was then also trained with both multi-class and binary classification to predict fire cause.

From our spatial analysis we observed that the severity of fire incidents has increased over time between 1992 and 2015, and particularly lightning-caused wildfires. Pockets of northern California and have faced persistent high frequencies of wildfires (exacerbated by climate change) that require closer monitoring and more aid to prevent considerable damage. Through predictive modeling, we found that weather-related factors were particularly relevant when looking to anticipate both fire size and cause; climate data that is specific to the actual fire incidents are likely needed to refine these predictions going forward.

Methods:

Spatial Analysis

To explore the spatial and temporal aspects of wildfires, the California subset of the US historic wildfires dataset was analyzed for insightful patterns that could be utilized by the state of California to improve fire response policy. Most of the analysis was conducted with ArcGIS Pro, a geographical information system (GIS) application that performs spatial mapping and geoprocessing; there is also an ArcGIS Python API, but due to issues with rendering maps and computation time, we primarily relied on ArcGIS Pro (6).

Since the dataset had available spatial coordinates (e.g., longitude and latitude), we easily mapped the data on top of geospatial data for California counties, and quickly visualized the spatial distribution of wildfires across the state (7). We applied the symbology (visualizing fire features by color, size, etc.) and filtering tools to find how wildfire sizes (in acres) and specific causes (e.g., lightning) varied across different counties and during specific years.

With the time-related tools in ArcGIS Pro and the *fire year* feature in the data, we created a time lapse animation to visualize how wildfires in the state evolved from 1992 and 2015. To establish some of the spatial and time patterns more definitively, emerging hot spot analysis (EHSA) was conducted to identify nascent and persistent hot spots (i.e., areas with high counts of wildfire incidents) in the data (8). Hot spot analysis creates a grid for the designated spatial area (e.g., state of California) and calculates the Getis-Ord-Gi* statistic (which is a measure of local spatial association) for each bin (or cluster) in the grid. The associated p-value indicates whether a cluster is a statistically significant hotspot; if a bin has a high number of fire incidents and is surrounded by other bins with similarly high values, then it is a significant hot spot in the data (9). EHSA additionally utilizes a time feature (e.g., *fire year*) to establish if the hot spot is persistent (has existed across most of the available time range) or has recently emerged (10).

To investigate the relationship between wildfires and other existing and relevant factors, we mapped additional spatial data alongside the 2015 fire incidents in the state. Location data was acquired for both California fire stations and healthcare facilities, and to visualize these points more easily alongside the wildfire data, we conducted a spatial join to aggregate the fire incidents by county and represented these counts with a choropleth map underneath the station and hospital points (11,12). We also briefly looked at other potentially relevant spatial data alongside our wildfire data, including population density by county, major roads across the state, and electric substation locations (13,14,15). Additionally, to support later modeling work, we utilized the proximity analysis tools in ArcGIS (specifically the distance measurement tool between different point features) to calculate the distance (in meters) to the nearest fire station for each fire incident; this metric could represent human accessibility to a fire, which may be an important determinant in the ability to contain a wildfire efficiently (16).

Finally, with the Plotly Dash framework, we created a dashboard that includes interactive visualizations of several aspects of the wildfires that can be explored over time and spatially (17). The kepler.gl package, which creates interactive multi-layer maps, was used to build a spatial mapping of California wildfires with other spatial factors (e.g., population, fire station/hospital locations). The resulting map contained a user interface with a lot of flexibility: turning off/on certain spatial data files, changing the color or size of different spatial points or polygons (or symbolizing by a particular feature), filtering the dataset, and incorporating additional spatial data to support further analysis. We embedded this map as an HTML file into the dashboard (18). We then created a bar plot of the wildfire frequencies by cause and a histogram of the distribution of the *fire size* feature (measured in acres) that included dropdowns to filter by year and county (19). Finally, a Bootstrap components library was added to apply consistent formatting and styling (20).

Modeling

We first focused on modeling fire size and identifying the relevant features in this prediction (21). For the initial regression model, we utilized the Portugal fires dataset, which only contained 517 observations, to predict *burn area* (measured in hectares); this dataset, which is described in Table 1 of the Appendix, had some location data but mostly contained the weather conditions associated with each individual fire, including humidity and temperature. Exploratory data analysis (EDA) revealed *temperature* as most strongly correlated with the target feature. Feature encoding of the time variables (e.g., *month* and *day*) and standard scaling were implemented to prepare the data for modeling. We fit multiple regression models on this dataset including Linear Regression, Random Forest (RF), Support Vector Machine (SVM), XGBoost Ensemble (XGB), and a simple dense Neural Network (22,23). Due to the dataset size, we conducted cross-validation to assess model performance and reported the mean absolute error (MAE).

Modeling the Portugal dataset confirmed the value of weather data for fire size prediction, so we acquired similar metrics for the US wildfires dataset, specifically the subset of California fires. The original data had many spatial and temporal features related to fire incidents but other than *fire cause* and *fire size*, lacked other details regarding the fire itself. To enhance the dataset, we requested yearly averages of temperature, precipitation, and wind speed for each county in California between 2000-2015 from the National Oceanic and Atmospheric Administration (NOAA) (24). We received 59 individual datasets that were merged with the original dataset after some initial cleaning and preprocessing. This process resulted in a dataset, described in Table 2 of the Appendix, with about 50,000 observations.

The nearest fire station distance metric calculated from our proximity analysis during spatial processing was also included, and a *burn time* metric was extracted by subtracting the discovery and containment dates in the original data; both measures explore human ability to contain wildfires. To prepare this data for modeling, we imputed missing weather data with median values, encoded nominal features (e.g., *fire cause*), dropped irrelevant and redundant features (such as state codes), and performed feature scaling. The data was then split into 80% training data and 20% test data.

To predict *fire size* (in acres) with the California data, similar to the Portugal data modeling, we fit multiple regression models, including: Linear Regression, RF, SVM, and XGB (25). Cross-validation and randomized grid search were employed to tune the hyperparameters for these models, and the test MAE values were reported.

In predicting *fire cause* with the California dataset, since there were 13 distinct cause types in the dataset, a few variations of this task were explored. First, we attempted multi-class prediction for individual fire causes with a RF classifier; we generally found low accuracy with these models, but identified lightning as the best predicted class, as seen in Figure J of the Appendix. We next incorporated additional labels to the dataset to create larger general cause classifications; specifically, we modified the task to predict four classes (natural, accidental, malicious, other) and fit a variety of classifiers: AdaBoost Ensemble with Decision Trees, Gaussian and Multinomial Naïve Bayes, K-Neighbors, Gradient Boosting, and RF (26).

Finally, due to the inclusion of climate data as well as the high number of lightning-caused fires in the dataset, as seen in Figure 2, we ran binary classification to specifically predict lightning-caused wildfires (27). We encoded the data to identify lightning and non-lightning fires and then balanced the data through under-sampling of the non-lightning data. Logistic Regression, RF, and XGB models were trained, again utilizing cross-validation for hyperparameter tuning. The test accuracy, precision, and recall were calculated to evaluate the final models.

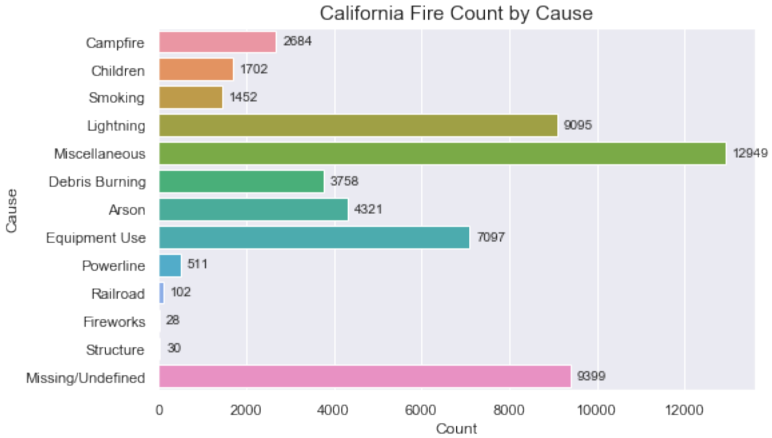


Figure 2: California Fire Counts by Cause

Results:

Spatial Analysis

When exploring the historical trends of wildfires in California between 1992-2015 with the time lapse visual, which can be accessed in Figure A of the Appendix, we found that there was an increase in the number of fire incidents, particularly after 2000. Additionally, from 2008 onwards, many larger fires (in terms of acres) occurred in northern California, indicating a high-risk area for devasting fires for which the state needs to develop more efficient containment strategies. We also observed the continued presence of wildfires in the Sacramento valley area across most of the available time span, which is likely due to the climate conditions (i.e., dry summer season) being conducive to fire activity; since climate change will exacerbate these effects, this area of the state also needs to be managed carefully.

The EHSA results in Figure 3 indicate that there were persistent hot spots of wildfire activity in the lower north and middle California areas (specifically the Sacramento and San Joaquin valleys) across all 24 years of data. Interestingly, we also found intensifying hot spot clusters surrounding this area; in the last couple of years of analysis, the spread and frequency of fires in this area had not been well contained and continued to expand. As these high concentration clusters grow, it is important for the state to ensure the surrounding community is adequately prepared to handle these wildfires on a regular basis.

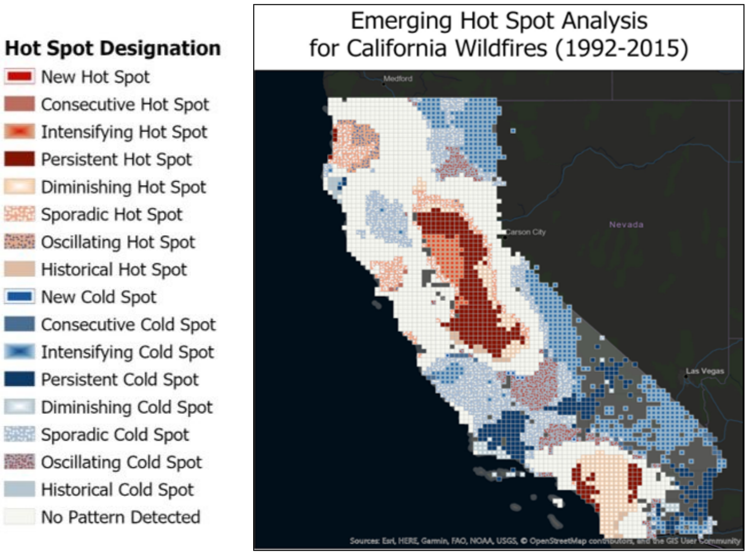


Figure 3: Emerging Hot Spot Analysis for California fires, 1992-2015

Since wildfire activity has increased in recent years due to changing climate conditions, we also examined how climate-caused fires (e.g., lightning) have evolved. Figure B in the Appendix looks at the size of lightning wildfires between 1992 and 2015 and shows a rise in incidents of severe (larger) lightning-caused fires over time, especially in northern California. If the state monitors weather conditions in areas with historically severe fires, they can possibly anticipate lightning fires and ensure proper precautions are taken to safely evacuate and minimize destruction.

Figure 4 visualizes 2015 wildfire counts alongside fire station and hospital locations in the state. While fire stations are well-distributed across the state as well as in counties with higher concentrations of fire incidents, healthcare resources are not as evenly distributed; parts northern California with high fire counts (particularly Shasta and Humboldt counties) have fewer healthcare options. Especially in areas with continued wildfire activity, having proper facilities in place for treatment of burns and general respiratory ailments is an important consideration for these communities.

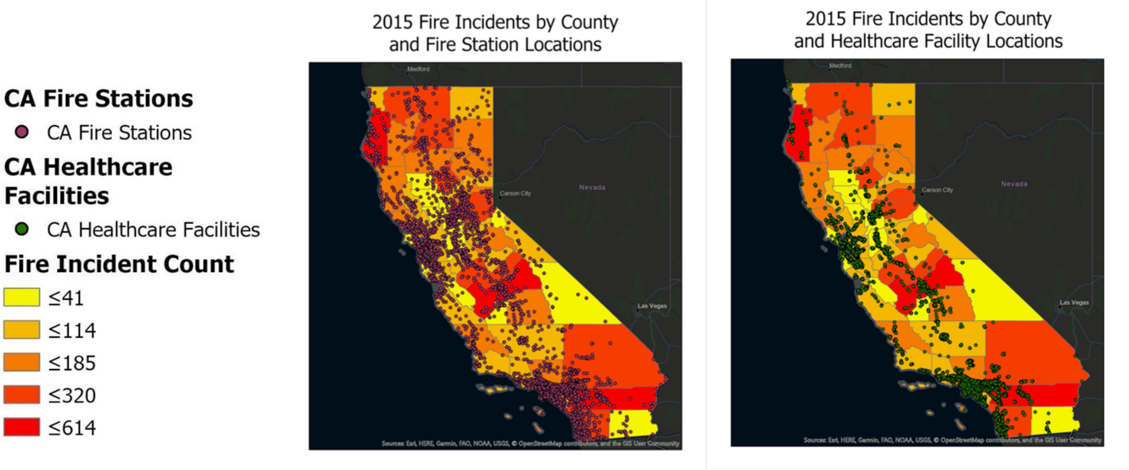
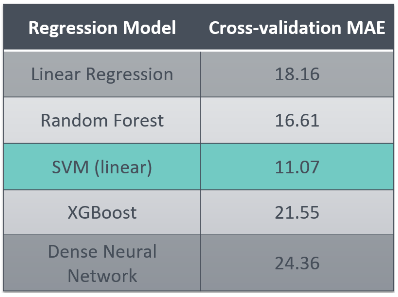

CA Fire Stations
CA Fire Stations
CA Healthcare Facilities
CA Healthcare Facilities
Fire Incident Count
≤41
≤114
≤185
≤320
≤614
2015 Fire Incidents by County and Fire Station Locations
2015 Fire Incidents by County and Healthcare Facility Locations
Medford
Carson City
Nevada
Las Vegas
Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Figure 4: 2015 Fire Incident by County and Distribution of Fire Stations/Healthcare Facilities

Modeling

In predicting *burn area* with the Portugal dataset, after performing hyperparameter tuning, the SVM (linear kernel) was the best performing model, as seen in Figure 5, with a cross-validated MAE of

11.07. Figure 6 shows the coefficient weights for this model, and we see that time features (such as *month*) were particularly impactful to the prediction, as were the temperature and humidity metrics. This regression task verified that weather-related features are relevant for this prediction and validated our decision to expand the California data.



Regression Model	Cross-validation MAE
Linear Regression	18.16
Random Forest	16.61
SVM (linear)	11.07
XGBoost	21.55
Dense Neural Network	24.36

Figure 5: Portugal Fire Size Regression Metrics

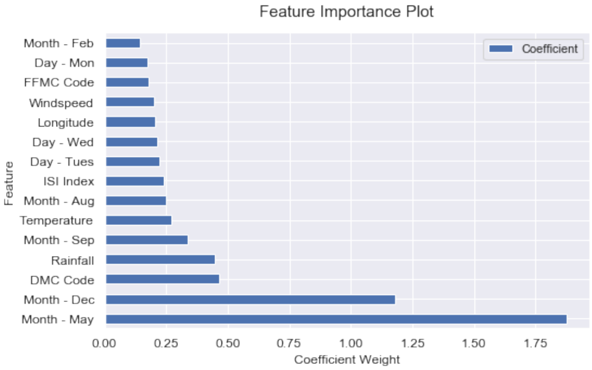
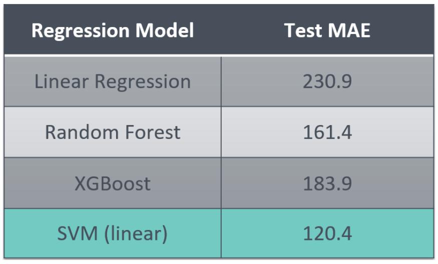


Figure 6: Portugal SVM model Feature Importance

For the *fire size* regression with the California data, even with hyperparameter tuning, we only achieved minimal performance improvement and experienced some overfitting issues; ultimately, the SVM (linear kernel) produced the lowest test MAE value, as seen in Figure 7, and had the least overfitting. The coefficient weights plot in Figure 8 confirms that some of the top contributing features were still weather-related (e.g., extreme temperature and wind speed). One plausible reason for overfitting is that the California dataset had many instances of very small fires, which can be seen in Figure G of the Appendix, making the model harder to generalize to larger examples. Additionally, while the Portugal modeling confirmed that climate data is relevant to this prediction, more detailed weather features may be necessary to improve performance; the Portugal weather data was recorded at the time of the fire, but due to availability and time constraints, we utilized yearly averages to supplement the California dataset. Incorporating more time-specific features to the dataset will most likely improve the model's ability to determine size more accurately.



Regression Model	Test MAE
Linear Regression	230.9
Random Forest	161.4
XGBoost	183.9
SVM (linear)	120.4

Figure 7: California Fire Size Regression Metrics

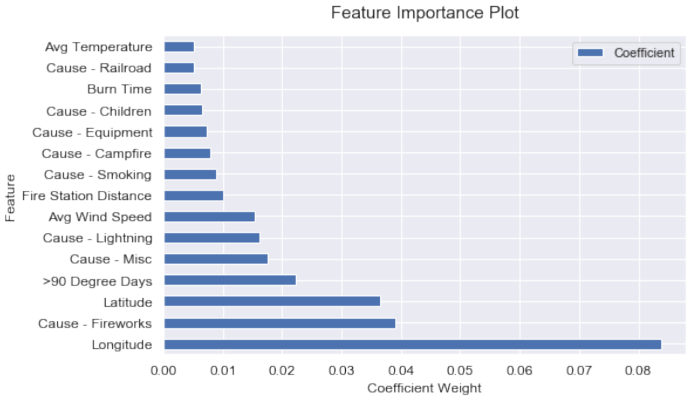
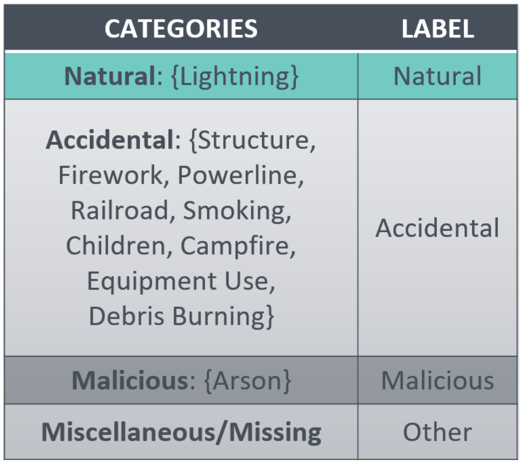


Figure 8: California SVM model Feature Importance

For multi-class prediction of *fire cause* using the new cause labels, the RF model produced the best predictions. While general model accuracy was relatively low (about 60%), Figure 9 shows that the natural/lightning class had the best performance and highest predictive potential (78.8% correct classifications), which coheres with the inclusion of weather features in the data.



CATEGORIES	LABEL
Natural: {Lightning}	Natural
Accidental: {Structure, Firework, Powerline, Railroad, Smoking, Children, Campfire, Equipment Use, Debris Burning}	Accidental
Malicious: {Arson}	Malicious
Miscellaneous/Missing	Other

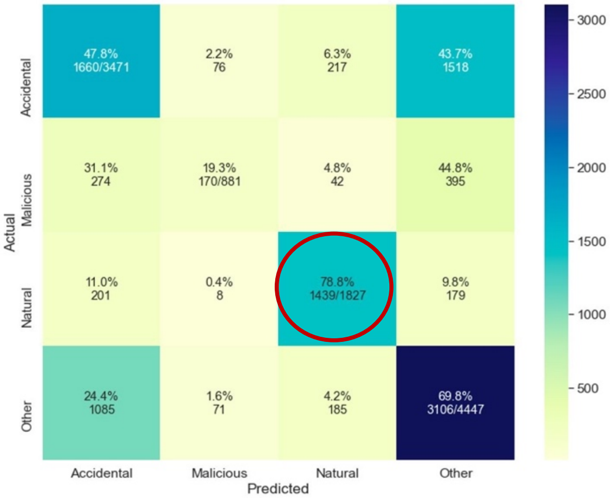
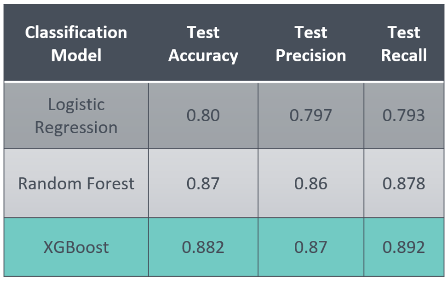


Figure 9: Confusion Matrix for Fire Cause – Labeled Classification (RF)

Finally, when conducting binary classification to predict lightning-caused fires, the XGB model had optimal performance, with 88.2% test accuracy and similar precision and recall metrics, as seen in Figure 10. When looking at the top features with Gini importance in Figure 11, we see that temperature was the most influential weather feature. Additionally, the *burn time* and nearest fire station distance metrics also had higher Gini values, which indicates that human ability to contain a fire may be a key factor to consider when assessing lightning-caused fires. However, more research is needed to better understand how these features relate to wildfire types and can be refined for future modeling.



Classification Model	Test Accuracy	Test Precision	Test Recall
Logistic Regression	0.80	0.797	0.793
Random Forest	0.87	0.86	0.878
XGBoost	0.882	0.87	0.892

Figure 10: Fire Cause (Binary) Metrics

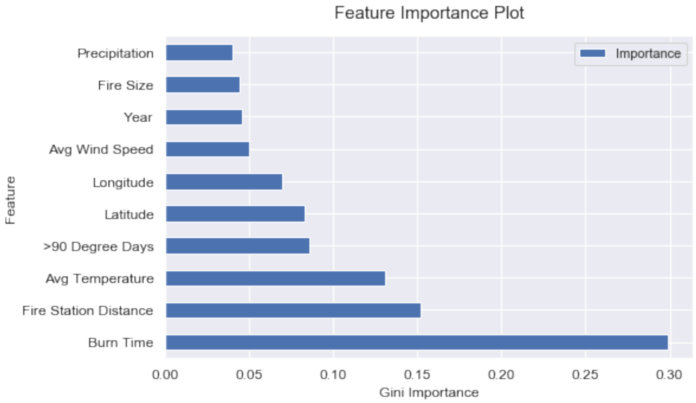


Figure 11: California XGB classifier Feature Importance

Discussion:

In this project, we researched the spatial and time-relevant patterns of historic wildfires in California and identified some of the critical features when predicting fire size and cause. Our spatial analysis showed that parts of northern California face a persistent risk of wildfires, that has been exacerbated by climate change; these areas need more efficient strategies to handle the increased frequency and severity of these fires, as well as additional resources (especially healthcare facilities) to ensure that the communities remain healthy as they face the health consequences of sustained exposure to these fires (28). Additionally, we produced a dashboard resource that the state of California can use to better visualize the distribution of wildfires across the state and gauge preparedness in different counties, in order to determine appropriate resource allocation moving forward.

In exploring prediction of fire size and cause, we created the best machine learning models with the available data. From these models, we identified that weather-related features (such as temperature and wind speed) were particularly relevant when exploring both fire severity and cause. Although more granular climate data is needed to improve predictive power, these findings provide initial factors that the state can monitor to forecast fires of specific types or sizes and a pathway for additional factors to research moving forward; the created features related to fire containment (including *burn time* and nearest fire station distance) could also be investigated further to build models focused on fire prevention.

Overall, we were able to successfully complete our project goals and met our milestones on time; we adjusted the order of some tasks but due to the buffer periods we built into our schedule, we still managed to meet our deadlines. One of the early challenges we faced was waiting to receive weather datasets from NOAA to enhance our California data, and the subsequent cleaning and merging of 59 separate datasets. Additionally, building an interactive dashboard took some time since neither of us had previous experience with this process. However, by exploring multiple frameworks to build a dashboard and embed interactive maps, we were able to create a functional user interface.

To improve this project in the future, we could incorporate more current wildfire data into both our spatial analysis and modeling to see how the recent extreme climate conditions may impact trends in frequency and severity. Additionally, including more detailed climate data (e.g., monthly county averages versus yearly numbers) could improve the California fire size regression predictions. It would also be interesting to develop models for multiple states in the US that have recently experienced increased wildfire activity (e.g., Alaska or Colorado) to see if there are similarities or differences in the relevant factors that are impacting wildfire behavior in these areas, such as weather, distribution of facilities, etc. Finally, some current research has incorporated satellite imagery into deep learning models (e.g., to find smoke patterns, evaluate before/after conditions), which is a promising avenue in developing more effective wildfire prediction models (29).

In general, the findings in this research can be useful in the future for tracking and analyzing fire conditions and developing systems to better manage the increased frequency of wildfires. We hope that our work can ultimately assist in minimizing the damage that wildfires have inflicted in communities across the world.

Statement of contributions:

Both team members contributed to all aspects of the project (e.g., spatial analysis, EDA and pre-processing, dashboard creation, regression and classification modeling) due to the small group size and a mutual interest in strengthening different skills. To complete this project, the team members made the following contributions:

1. *Gopalika Sharma* – performed initial spatial analysis on the distribution of California wildfires, strengthened the California dataset by requesting and merging relevant weather data (NOAA), conducted EDA on both the Portugal and California datasets, contributed towards designing and creating the dashboard, and performed predictive modeling for both prediction tasks, fire size and cause for both the Portugal and California datasets.

2. *Surya Menon* – performed spatial analysis on the California wildfires dataset (including EHSA, proximity analysis), contributed to data pre-processing of the augmented California data, conducted EDA on both the Portugal and California datasets, created the kepler.gl interactive map and worked on creating/implementing the dashboard, and conducted predictive modeling for both the Portugal and California predictions (regression and classification).

References:

1. Gray, E. (2019, September 19). Satellite Data Record Shows Climate Change's Impact on Fires. NASA - Global Climate Change: Vital Signs of the Planet. https://climate.nasa.gov/news/2912/satellite-data-record-shows-climate-changes-impact-on-fires/
2. California Department of Fish and Wildlife. (n.d.). Science: Wildfire Impacts. CA.gov. Retrieved 2020, from https://wildlife.ca.gov/Science-Institute/Wildfire-Impacts
3. Center for Disaster Philanthropy (CPP). (2020, October 23). 2020 North American Wildfire Season. Center for Disaster Philanthropy. https://disasterphilanthropy.org/disaster/2020-california-wildfires/
4. US Forest Service. (2020, April 29). National Interagency Fire Occurrence 1992-2015 (Feature Layer). https://enterprisecontentnew-usfs.hub.arcgis.com/datasets/e4d020cb51304d5194860d4464da7ba7_0/data?geometry=61.662%2C-2.200%2C54.279%2C76.163
5. UCI. (n.d.). UCI Machine Learning Repository: Forest Fires Data Set. UCI Machine Learning Repository. Retrieved 2020, from https://archive.ics.uci.edu/ml/datasets/Forest+Fires
6. Law, M., & Collins, A. (2019). Getting to Know ArcGIS Pro (Second ed.). Esri Press.
7. CA.gov. (2019, October 23). CA Geographic Boundaries - California Open Data. California Open Data Portal. https://data.ca.gov/dataset/ca-geographic-boundaries
8. Gates, S. (2017, August 15). Emerging Hot Spot Analysis: Finding Patterns over Space and Time. Azavea. https://www.azavea.com/blog/2017/08/15/emerging-hot-spot-spatial-statistics/
9. Bambrick, G. (2016, January 22). What is Hotspot Analysis? Geospatiality. https://glenbambrick.com/2016/01/21/what-is-hotspot-analysis/
10. ESRI. (n.d.). How Emerging Hot Spot Analysis works—ArcGIS Pro | Documentation. ArcGIS. Retrieved 2020, from https://pro.arcgis.com/en/pro-app/tool-reference/space-time-pattern-mining/learnmoreemerging.htm
11. CA Open Data - Department of Public Health. (2020, October 18). CDPH Licensing and Certification Healthcare Facilities. California State Geoportal. https://gis.data.ca.gov/datasets/CDPHDATA::cdph-licensing-and-certification-healthcare-facilities
12. Homeland Infrastructure Foundation-Level Data (HIFLD). (2020, September 11). Fire Stations. HIFLD Open Data. https://hifld-geoplatform.opendata.arcgis.com/datasets/fire-stations?geometry=-130.326%2C33.769%2C-104.442%2C39.917
13. ESRI. (n.d.). Living Atlas of the World | ArcGIS. ArcGIS. Retrieved October 2020, from https://livingatlas.arcgis.com/en/home/
14. US Census Bureau. (2019, August 15). TIGER/Line Shapefile, 2018, state, California, Primary and Secondary Roads State-based Shapefile - Data.gov. Data.Gov. https://catalog.data.gov/dataset/tiger-line-shapefile-2018-state-california-primary-and-secondary-roads-state-based-shapefile
15. CA.gov. (2020, August 11). California Electric Substation. California State Geoportal. https://gis.data.ca.gov/datasets/CAEnergy::california-electric-substation
16. ESRI. (n.d.-c). Proximity analysis—Help | ArcGIS for Desktop. ArcGIS for Desktop. Retrieved 2020, from https://desktop.arcgis.com/en/arcmap/10.3/analyze/commonly-used-tools/proximity-analysis.htm
17. plotly. (n.d.). Part 2. Layout | Dash for Python Documentation | Plotly. Plotly | Dash. Retrieved 2020, from https://dash.plotly.com/layout
18. kepler.gl. (2020). kepler.gl for Jupyter User Guide. https://docs.kepler.gl/docs/keplergl-jupyter
19. Moffitt, C. (2017, October 9). Creating Interactive Visualizations with Plotly's Dash Framework. Practical Business Python. https://pbpython.com/plotly-dash-intro.html

20. Dash Bootstrap Components. (n.d.). Quickstart - dbc docs. Retrieved 2020, from https://dash-bootstrap-components.opensource.faculty.ai/docs/
21. Stanford-Moore, A., & Moore, B. (2019). Wildfire Burn Area Prediction. Stanford University - CS229. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26582553.pdf
22. Jain, T. (2019, October 13). Forest Fire Prediction with the help of multiple regression models. Medium. https://medium.com/@tanmayjain84/forest-fire-prediction-with-the-help-of-multiple-regression-models-to-get-the-best-accurate-model-13f8446e4737
23. Nerd, R. (2018, April 9). Natural Calamity | Classifying Forest Fire Damage - Deep Learning. Medium. https://medium.com/machine-learning-bootcamp/natural-calamity-classifying-forest-fire-damage-c4139acfc009
24. National Oceanic and Atmospheric Administration (NOAA). (n.d.). Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Climate Data Online (CDO). Retrieved October 2020, from https://www.ncdc.noaa.gov/cdo-web/search
25. Ursner, M. E. (2020, June 19). Taming Of 1.88 Million Wildfires. Medium. https://medium.com/@martin_47009/taiming-1-88-million-wildfires-e2595c43b769
26. O'Neill, P. (2017, December 13). Predict the causes of wildfires using Python. Kaggle. https://www.kaggle.com/edhirif/predict-the-causes-of-wildfires-using-python
27. ArcGIS. (n.d.). Historical Wildfire Analysis | ArcGIS for Developers. ArcGIS API for Python. Retrieved 2020, from https://developers.arcgis.com/python/sample-notebooks/historical-wildfire-analysis/
28. lung.org Editorial Staff. (2016, January 2). How Wildfires Affect Our Health. American Lung Association. https://www.lung.org/blog/how-wildfires-affect-health
29. Coldewey, D. (2020, May 29). TechCrunch is now a part of Verizon Media. TechCrunch. https://techcrunch.com/2020/05/29/as-wildfire-season-approaches-ai-could-pinpoint-risky-regions-using-satellite-imagery/

Appendix:

Dashboard Resources

Code to run dashboard locally: https://github.com/gsharma14/DS-5500-California-Wildfires

kepler.gl HTML map: https://kepler.gl/demo/map?mapUrl=https://dl.dropboxusercontent.com/s/q1ijfisgc0prk1f/keplergl_q3m03h.json

Additional Spatial Analysis

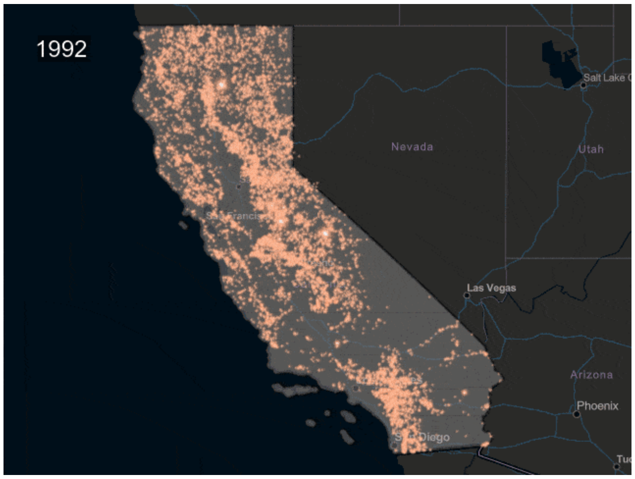


Figure A: Time Lapse of California Wildfires, graduated by Size (Acres), 1992-2015*

*Graphics Interchange Format (GIF) file available in Spatial Analysis folder of GitHub repository

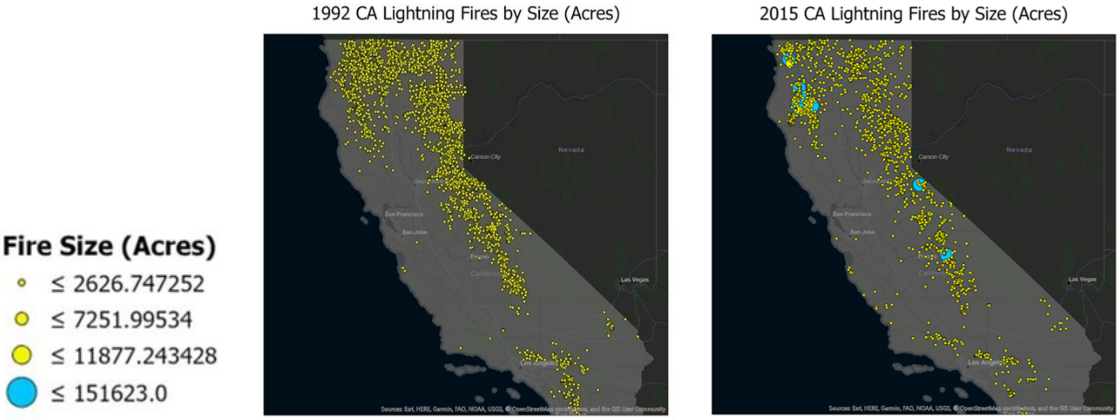
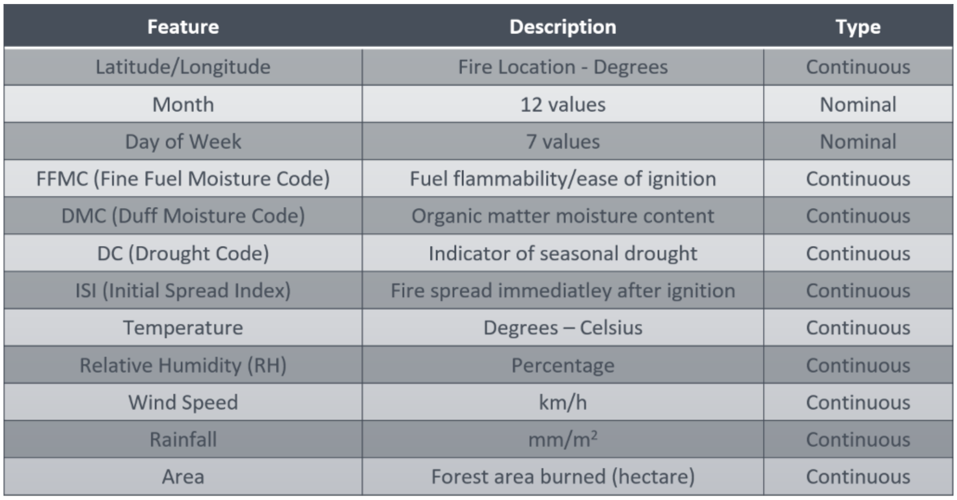


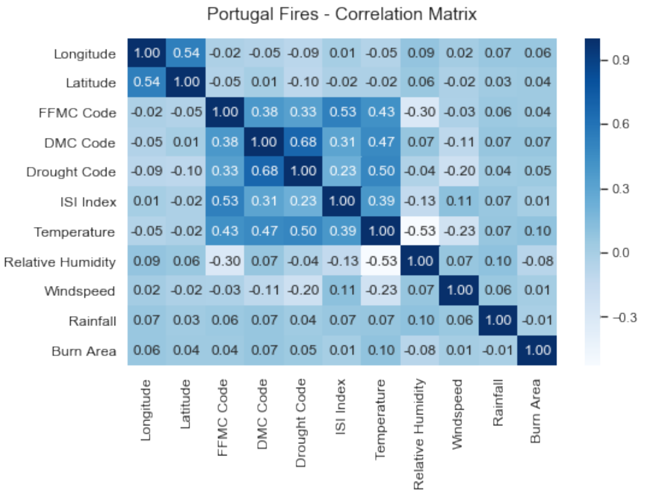
Figure B: Change in California climate-related fires by size from 1992 and 2015

Portugal Exploratory Data Analysis



Feature	Description	Type
Latitude/Longitude	Fire Location - Degrees	Continuous
Month	12 values	Nominal
Day of Week	7 values	Nominal
FFMC (Fine Fuel Moisture Code)	Fuel flammability/ease of ignition	Continuous
DMC (Duff Moisture Code)	Organic matter moisture content	Continuous
DC (Drought Code)	Indicator of seasonal drought	Continuous
ISI (Initial Spread Index)	Fire spread immediatley after ignition	Continuous
Temperature	Degrees – Celsius	Continuous
Relative Humidity (RH)	Percentage	Continuous
Wind Speed	km/h	Continuous
Rainfall	mm/m²	Continuous
Area	Forest area burned (hectare)	Continuous

Table 1: Portugal Forest Fires Dataset Features



	Longitude	Latitude	FFMC Code	DMC Code	Drought Code	ISI Index	Temperature	Relative Humidity	Windspeed	Rainfall	Burn Area
Longitude	1.00	0.54	-0.02	-0.05	-0.09	0.01	-0.05	0.09	0.02	0.07	0.06
Latitude	0.54	1.00	-0.05	0.01	-0.10	-0.02	-0.02	0.06	-0.02	0.03	0.04
FFMC Code	-0.02	-0.05	1.00	0.38	0.33	0.53	0.43	-0.30	-0.03	0.06	0.04
DMC Code	-0.05	0.01	0.38	1.00	0.68	0.31	0.47	0.07	-0.11	0.07	0.07
Drought Code	-0.09	-0.10	0.33	0.68	1.00	0.23	0.50	-0.04	-0.20	0.04	0.05
ISI Index	0.01	-0.02	0.53	0.31	0.23	1.00	0.39	-0.13	0.11	0.07	0.01
Temperature	-0.05	-0.02	0.43	0.47	0.50	0.39	1.00	-0.53	-0.23	0.07	0.10
Relative Humidity	0.09	0.06	-0.30	0.07	-0.04	-0.13	-0.53	1.00	0.07	0.10	-0.08
Windspeed	0.02	-0.02	-0.03	-0.11	-0.20	0.11	-0.23	0.07	1.00	0.06	0.01
Rainfall	0.07	0.03	0.06	0.07	0.04	0.07	0.07	0.10	0.06	1.00	-0.01
Burn Area	0.06	0.04	0.04	0.07	0.05	0.01	0.10	-0.08	0.01	-0.01	1.00

Figure C: Portugal Dataset Correlation Matrix

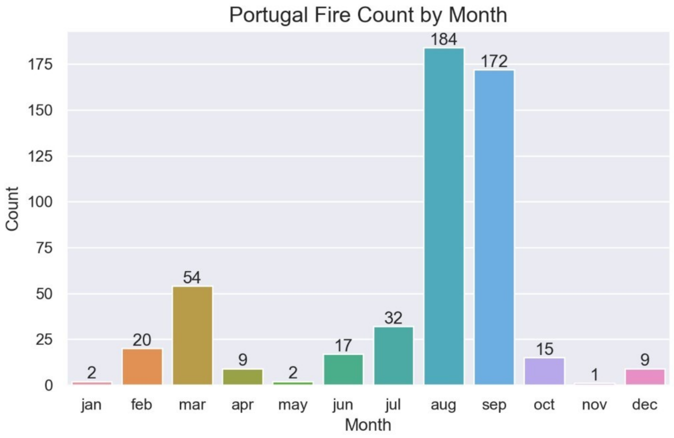
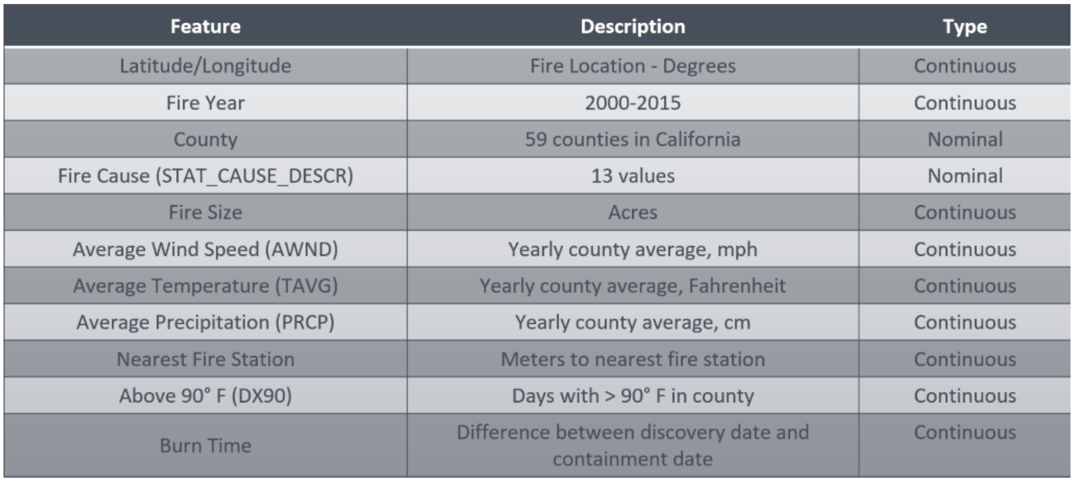


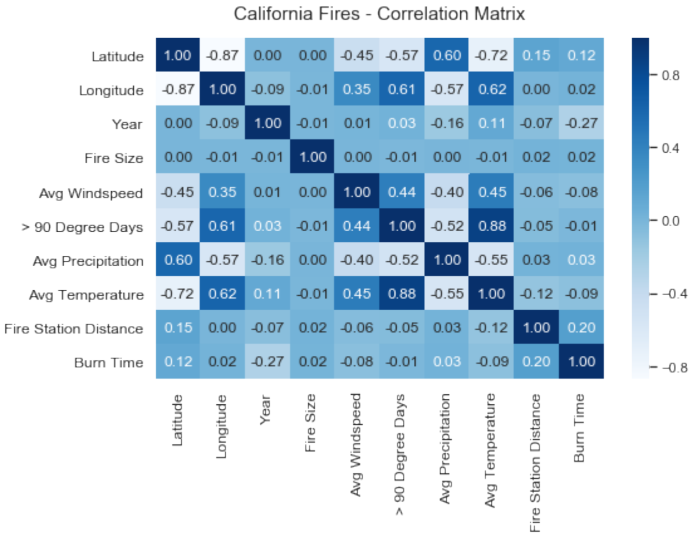
Figure D: Portugal Fire Counts by Month

California Exploratory Data Analysis



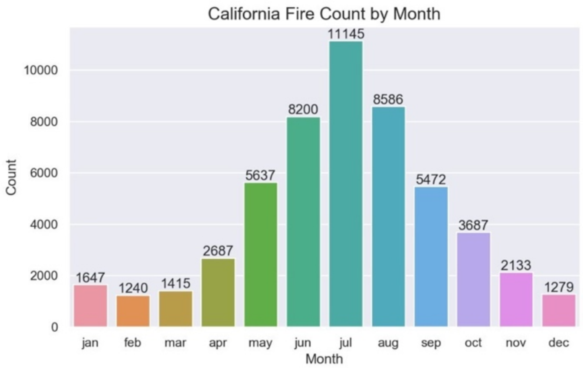
Feature	Description	Type
Latitude/Longitude	Fire Location - Degrees	Continuous
Fire Year	2000-2015	Continuous
County	59 counties in California	Nominal
Fire Cause (STAT_CAUSE_DESCR)	13 values	Nominal
Fire Size	Acres	Continuous
Average Wind Speed (AWND)	Yearly county average, mph	Continuous
Average Temperature (TAVG)	Yearly county average, Fahrenheit	Continuous
Average Precipitation (PRCP)	Yearly county average, cm	Continuous
Nearest Fire Station	Meters to nearest fire station	Continuous
Above 90° F (DX90)	Days with > 90° F in county	Continuous
Burn Time	Difference between discovery date and containment date	Continuous

Table 2: California Forest Fires Dataset Features



	Latitude	Longitude	Year	Fire Size	Avg Windspeed	> 90 Degree Days	Avg Precipitation	Avg Temperature	Fire Station Distance	Burn Time
Latitude	1.00	-0.87	0.00	0.00	-0.45	-0.57	0.60	-0.72	0.15	0.12
Longitude	-0.87	1.00	-0.09	-0.01	0.35	0.61	-0.57	0.62	0.00	0.02
Year	0.00	-0.09	1.00	-0.01	0.01	0.03	-0.16	0.11	-0.07	-0.27
Fire Size	0.00	-0.01	-0.01	1.00	0.00	-0.01	0.00	-0.01	0.02	0.02
Avg Windspeed	-0.45	0.35	0.01	0.00	1.00	0.44	-0.40	0.45	-0.06	-0.08
> 90 Degree Days	-0.57	0.61	0.03	-0.01	0.44	1.00	-0.52	0.88	-0.05	-0.01
Avg Precipitation	0.60	-0.57	-0.16	0.00	-0.40	-0.52	1.00	-0.55	0.03	0.03
Avg Temperature	-0.72	0.62	0.11	-0.01	0.45	0.88	-0.55	1.00	-0.12	-0.09
Fire Station Distance	0.15	0.00	-0.07	0.02	-0.06	-0.05	0.03	-0.12	1.00	0.20
Burn Time	0.12	0.02	-0.27	0.02	-0.08	-0.01	0.03	-0.09	0.20	1.00

Figure E: California Dataset Correlation Matrix



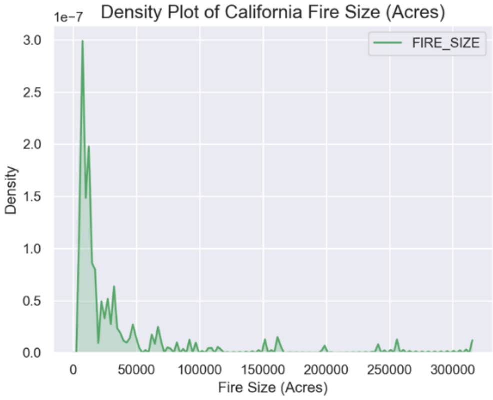


Figure F: California Fire Counts by Month

Figure G: California Fire Size Density Plot

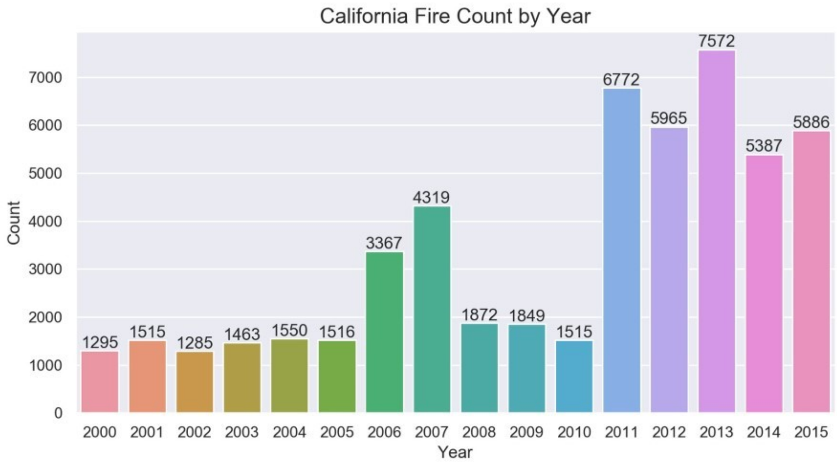


Figure H: California Fire Counts by Year

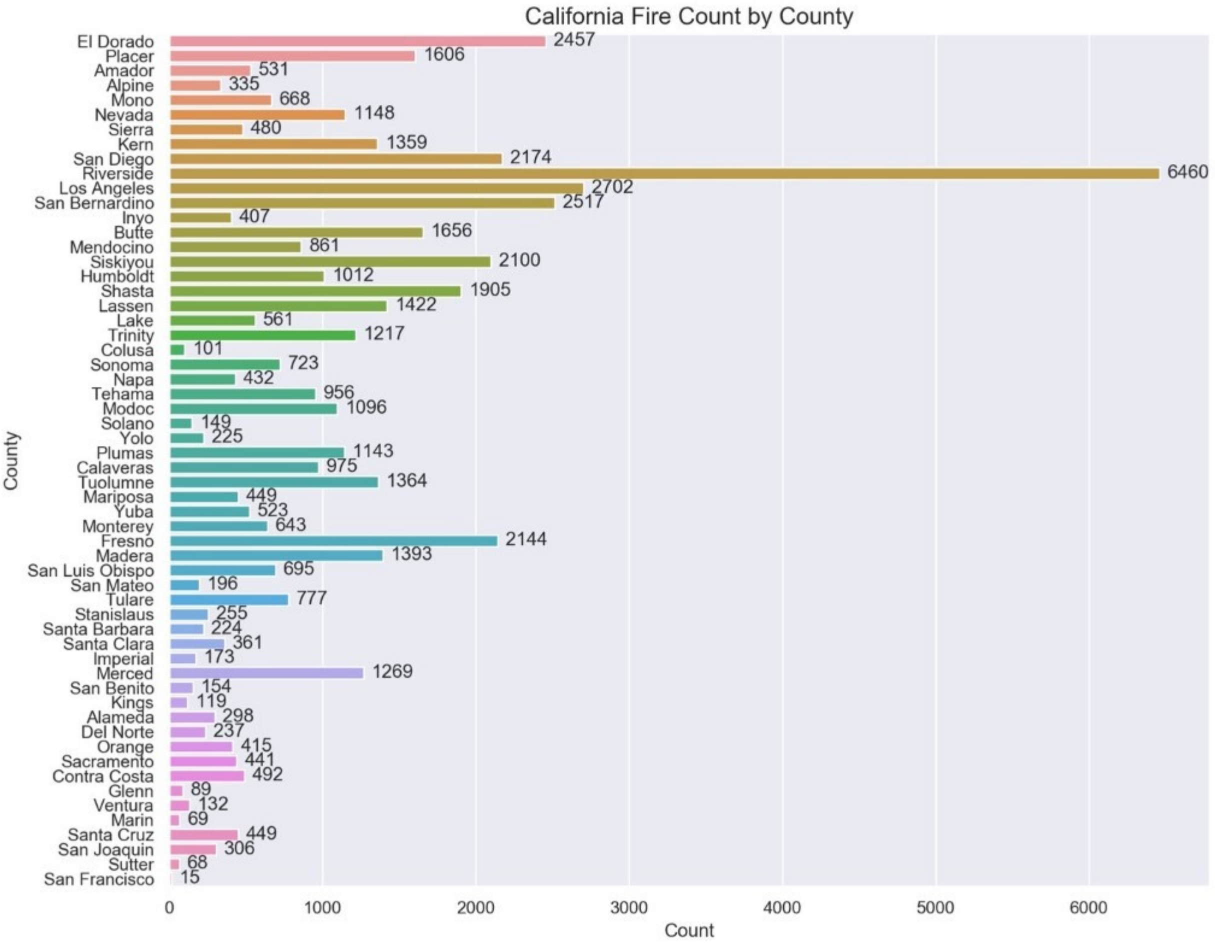


Figure I: California Fire Counts by County

California Fire Cause Classification

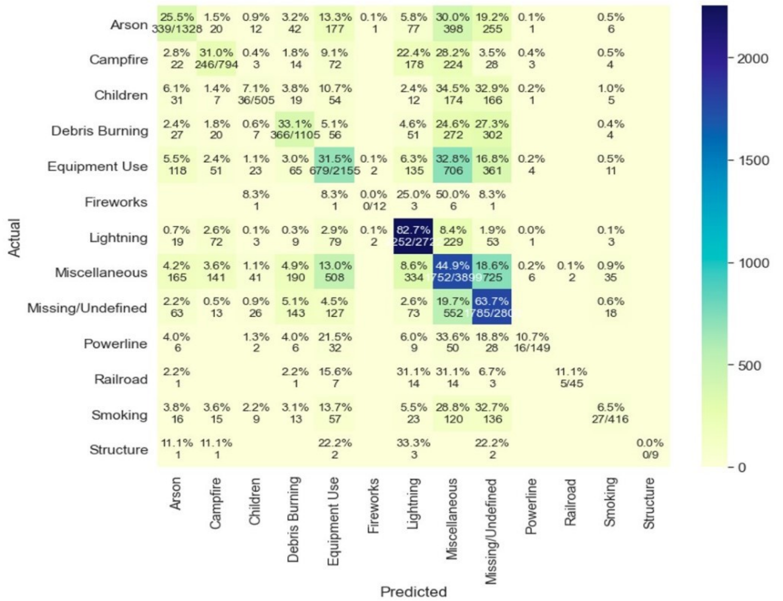


Figure J: Confusion Matrix for Fire Cause – Unlabeled Classification (RF)

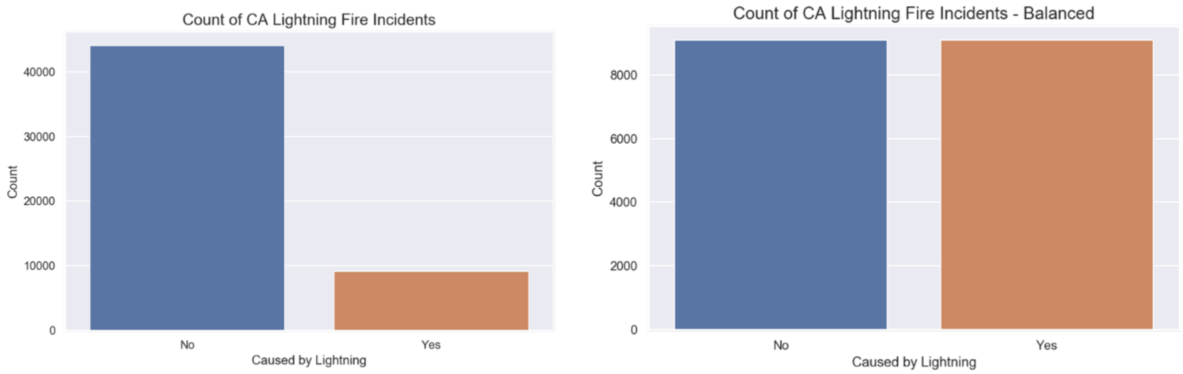
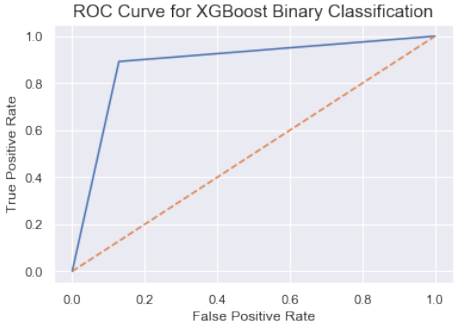


Figure K: Balanced and Unbalanced Lightning Fires



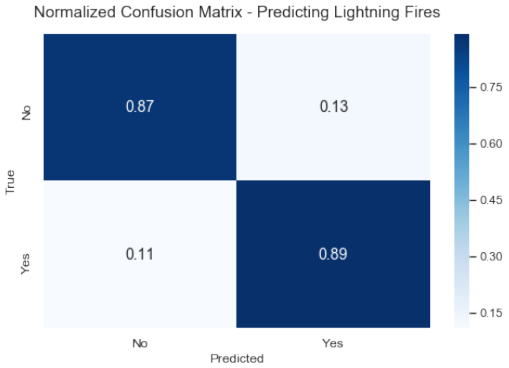


Figure L: XGB ROC Curve – Binary XGB

Figure M: Normalized Confusion Matrix, Binary XGB