UC San Diego

JACOBS SCHOOL OF ENGINEERING

University of California, San Diego
Data Science & Engineering

Data Science Capstone Project

Wildfire Data Analysis: Predicting Risk and Severity in San Diego County

Mike Gallaspy, Kevin Kannappan, Marty La Pierre, Sofean Maeouf & Sathish Masilamani

advised by
Ilkay Altintas de Callafon, Tom Corringhan, Dan Crawl & Mai Nguyen

June 5, 2020

Abstract

Major wildfires have grown in intensity and frequency not only in southern California, but also across the world, in the last few years. Property loss, damages and costs associated with these wildfires can be especially significant to residents, communities, public space, and the environment. Gas and electric companies have begun to proactively take down power grids as a preventative measure, a significant disruptor to everyday life. In this work, we analyze factors that affect wildfires and how changes in these factors over time affect wildfire behaviors and risks. Specifically, we took daily weather observations and associated fire measurements (indication of fire and the acres burned) in San Diego county over the course of the past 20 years, hypothesizing that changing weather conditions are the primary drivers of wildfires. We validate this theory by providing evidence that wildfire risk and severity are correlated with evapotranspiration (a measure of ground dryness), available vegetation, ambient temperature and wind speed. Using machine learning, we develop a robust model that detects the incidence and acres burned of a wildfire with a testing validation Recall score of 77%. Additionally, we pose a method to estimate the economic value of a region with associated "high" wildfire risks. Residents, firefighters, and public policy officials may leverage the results of this project for effective management and mitigation of wildfire risks.

1 Introduction and Question Formulation

Wildfires are large, relatively uncontrolled fires in areas of combustible vegetation typically occurring in rural or undeveloped areas. Even though wildfires are a natural occurrence, the growing concern is that wildfires have increased in both risk and severity in recent years. Unfortunately, there is some alarming evidence to support this notion. According to research by NOAA's National Centers for Environmental Information (NCEI), wildfire costs in California in 2018 were estimated to be $24 billion, a 33% increase from the previous record of $18 million set the previous year [1]. Furthermore, it was reported that more than 8.7 million acres burned across the U.S. during 2018, a measure significantly higher than the prior 10-year average (2009-2018) of 6.8 million acres. Outside of the United States, in Australia in 2019, the series of large bushfires burned more than 26.4 million acres (roughly the size of South Korea) with varying cost estimates, conservative ones around $65 billion USD [2, 3]. Considering that world population size and associated land use continues to grow, there is significant reasoning to be concerned with the possibilities and associated costs of major wildfires. The cost estimates of these fires, however, range wildly because of all the possible impacted factors to consider, such as: response (public funding), damage (residential, etc.), prevailing health (lung disease, etc.), and ecosystem (loss of habitat) costs. Hence, while the ignition of the fire is a relatively simple chemical reaction, both the reasoning for and associated outcomes of major wildfires are complex entities with considerable room for research.

Some of the latest research on California wildfire risk and severity in 2017 indicated that the relative significance of climate and weather is increasing [4]. They theorized that a wetter than average winter in 2016 allowed for increased vegetation growth, followed by an atypically warm summer which dried the excessive fuel, which culminated with dry Santa Ana wind conditions in the South. This sequence of conditions allowed for rapid and uncontrolled fire spread. The study conducted in [4] serves as the basis for the work in this paper: we focused on different weather observations at different temporal magnitudes (i.e. cumulative rainfall vs. recent average wind speed, for example) to analyze their effect on wildfire risk and severity. Even though we know that humans can contribute directly to the frequency of a fire through accidental (or intentional and negligent) ignition, evaluating the impact social factors on fire risk is beyond the scope of this project. Considering available resources to our team and affiliation with UCSD, we decided to limit our focus on wildfires that occurred within San Diego County and their associated weather conditions.

In this project, we used the data on weather and fires in San Diego County believing the results may generalize to other regions as well. We intend to use this data to answer the following problems:

1. How have the characteristics of wildfires changed over the past few years?
2. How can wildfire risk be quantified?
3. What are the trends in wildfire risk factors over time?

The practical application of our research would be to develop a model of fire "risk" (probability of ignition) and severity (measured in acres burned). Part of the challenge of this project is to identify, gather, and integrate relevant data (i.e. weather observations and fire incidence data). The project includes a substantial exploratory analysis component to discover the important characteristics and trends in the data.

1.1 Related Work

Existing models of fire risk and severity are already being deployed for a variety of use cases, including by the advisors on this project. At UCSD's Super Computer Center, for example, researchers have developed an application called Firemap [5]. Firemap executes rapid simulations of a physical model based on fires that are predicted to have rapid spread. Additionally, Firemap provides historical data on previous fires and their weather conditions in addition to future forecasts.

By collaborating with San Diego Gas & Electric (SDG&E) on this project, we learned that they have existing physical models, classifiers, and forecasts that they use to measure risk of fire ignition and spread. Much of their input data is highly sophisticated, including satellite vegetation data with resolution up to three meters. A lot of their research focuses around discrete grids that fit their service locations with the primary objective being determining whether or not to shut down portions of their grid in the event of a fire.

In addition to [4], these resources have been tremendously valuable for validating our approach and research. We believe that our model has additive value in the sense that we are providing an approach for developing fire risk and severity classifiers for public data sets (which we will cover in later sections). While previous research has focused on using proprietary data, our focus has been to provide an easily reproducible approach that can be incorporated to existing methods or applied to alternate locations.

2 Team Roles and Responsibilities

Assigned roles were the following, however many roles were flexed during the course of this project:

- Budget Manager: Sofean Maeouf
- Project Manager: Mike Gallaspy
- Project Coordinator: Kevin Kannappan
- Report Manager: Marty La Pierre
- Record Keeper: Sathish Masilamani

3 Data Acquisition

Initial data acquisition aspects of the project were challenging. At the onset of the project, in the absence of empirical research, we acquired as many related datasets as we could without a tangible direction to how we planned to integrate them or which features would drive our future predictions. In this phase of initial exploratory data analysis (EDA), much of our work involved determining the value of the information in a dataset and theorizing whether the information quality was good enough to be merged with other datasets.

3.1 Data Sources

We examined the following 9 datasets for subsequent analysis on fire risk and severity. Considering we were constrained by cost, data quality, data completeness and the technologies at our disposal, we had to be very specific in how we selected datasets. Notably, since our research led us to collect weather observations and generate suitable features for future analysis tasks, we needed to collect ample amounts of weather data (spatially, temporally, and type). Additionally, the weather data needed to be able to be properly merged with the

fire incidence data - which was predominately in GIS files (asymmetric geometries). From there, that limited the available technologies that we could explore for our subsequent research as manipulation and joining with GIS geometries are niche tasks. Lastly, a measure of economic data to quantify the significance of fires can help contextualize our results and provide value for decision-makers with resource constraints. Hence, in order to properly address our set of problems posed - leveraging different data sources with their own technologies was paramount to our work.

3.1.1 Weather Data

MesoWest: Synoptics Mesonet API is a database of real-time and historical surface-based weather observations from over 75,000 public and private observing stations accessible with a REST API. We initially identified this as a dataset of interest since it had been used in previous projects related to wildfire analysis, most notably [5]. However Synoptic has since started charging for Mesonet usage, and although it offers a small amount of free usage we determined it was not practical due to the large number of anticipated requests.

- Data Format
  - Not applicable
- Data Size
  - Not applicable
- Records
  - Not applicable
- Used in Final Pipeline
  - **No**

NOAA ISD: The Integrated Surface Database (ISD) consists of global hourly and synoptic observations compiled from numerous sources into a single common ASCII format and common data model. ISD was developed as a joint activity within Asheville's Federal Climate Complex. NCEI, with U.S. Air Force and Navy partners, began the effort in 1998 with the assistance of external funding from several sources. ISD integrates data from over 100 original data sources. Again because the data is freely available and very comprehensive, NOAA ISD is appealing, however due to the low spatial density, we opted to primarily use gridMET.

- Data Format
  - Not applicable
- Data Size

  - Not applicable

- Records
  - Not applicable
- Used in Final Pipeline
  - **No**

GridMET: gridMET is a dataset of daily high-spatial resolution (1/24th degree) surface meteorological data covering the contiguous US from 1979 to present, updated daily [6]. Due to its completeness and availability, this project has adopted gridMET as a primary data source. Not all of the data are observations since the density of weather stations is actually much lower than gridMETs resolution, so various interpolation methods are adopted. According to its maintainers:

> Validation of the...meteorological data, using an extensive network of automated weather stations across the western United States, showed skill comparable to that derived from interpolation using station observations, suggesting it can serve as suitable surrogate for landscapescale ecological modelling across vast unmonitored areas of the United States.
>
> [6]

- Data Format
  - Converted download from netCDF data files to Parquet
- Data Size
  - 1.2 GB
- Records
  - 10.7 million rows
- Used in Final Pipeline
  - **Yes**

3.1.2 Fire Data

GeoMAC: The Geospatial Multi-Agency Coordination or GeoMAC, is an internet-based mapping application originally designed for fire managers to access online maps of current fire locations and perimeters in the United States [7]. GeoMAC data is evidently one of the sources of the historical fire perimeter set that we received from an advisor (Dan Crawl). While GeoMAC data is publicly available, additional preprocessing and compiling was performed by our advisor already so we used that data.

CalFire: From FRAP the data set shows historical fires from as far back as 1850. The dataset contains attributes that describe the fire and the approximate radius. CalFire data is evidently one of the sources of the historical fire perimeter set that we received from an advisor (Dan Crawl) [8].

NFIRS: As a result of legislation mandating collection of national data on fires, the National Fire Incident Reporting System (NFIRS) was established in 1976 via a pilot program [9]. The U.S. Fire Administration (USFA), a component of the Department of Homeland Security (DHS), developed NFIRS as a means of assessing the nature and scope of the fire problem in the United States. NFIRS data is evidently one of the sources of the historical fire perimeter set that we received from an advisor (Dan Crawl). Most of the value of NFIRS, however, is the detailed economic valuations of each incident. Yet, economic data was not included in the merged perimeters dataset as a lot of data quality issues exist within the data.

As indicated, our adviser merged these three datasets for a more robust fire incidence and perimeter dataset. The resulting dataset was in GIS format that could be read into GeoPandas dataframes in Python, or another GIS mapping tool of choice.

- Data Format
  - Generated GIS shapefiles, which can be read using Python GeoPandas library
- Data Size
  - 1.7 GB
- Records
  - 92.5K rows
- Used in Final Pipeline
  - **Yes**

3.1.3 Other Data

HPWREN: The High Performance Wireless Research and Education Network (HPWREN), a University of California San Diego partnership project led by the San Diego Supercomputer Center and the Scripps Institution of Oceanography's Institute of Geophysics

and Planetary Physics, supports Internet-data applications in the research, education, and public safety realms. HPWREN functions as a collaborative, Internet-connected cyberinfrastructure. The project supports a high-bandwidth wireless backbone and access data network in San Diego, Riverside, and Imperial counties in areas that are typically not well-served by other technologies to reach the Internet. Data gathered includes weather observations and real-time video.

- Data Format
  - Not applicable
- Data Size
  - Not applicable
- Records
  - Not applicable
- Used in Final Pipeline
  - **No**

Landfire: The LF Program provides 20+ national geo-spatial layers (e.g. vegetation, fuel, disturbance, etc.), databases, and ecological models that are available to the public for the US and insular areas.

- Data Format
  - Not applicable
- Data Size
  - Not applicable
- Records
  - Not applicable
- Used in Final Pipeline
  - **No**

NSI: The US Army Corps of Engineers's National Structure Inventory (NSI) provides detailed information on the structures (and their types) in a given area. NSI improves upon the structure inventory data available through FEMA's HAZUS database. NSI was primarily developed as a tool to estimate regional flood risks, however it has been used as a resource for teams in the early iterations of city planning work. For the purposes of this work, only structures in San Diego County were considered.

- Data Format
  - GIS shapefiles, which can be read using Python GeoPandas library
- Data Size
  - 0.5 GB
- Records
  - 1.7 million rows
- Used in Final Pipeline
  - **Yes**

3.2 Data Collection

There were three primary components of data extraction for our workflow: extracting the fire incidence and geometry GIS files, extracting the weather data, and downloading some economic GIS data. For the fire data, since this was compiled by our adviser by merging a few datasets, we were able to access the data directly via Google Drive. Similarly, the economic data was provided by an adviser and available to us to download in a Google Drive. We relied on gridMET for our weather data and wrote a Python script to properly extract and filter weather data for grids in San Diego County.

Given the size of the weather data, we elected to upload all data sources to Amazon Web Services (AWS) Simple Storage Service (S3). Our reasoning for using S3 was primarily based on available resources to our team, as a computational budget was provided. As indicated, some of the datasets were in GIS format. Hence, we either converted data formats to shapefiles or to CSVs (after data merging transpired). Our primary tool for blending the data was the GeoPandas library which allowed for GIS joins with geometries & points - the primary format of the fire data. In the case of our weather grids, we had to convert the available latitude and longitude coordinates to geometries so that the data could be merged.

3.3 Data Pipelines

The data pipeline implemented for our analysis involved three primary sections: source data acquisition, upload and blending on AWS, and data analysis. We used Python and Spark as our primary languages, with a particular emphasis on GIS-related packages in Python. Data acquisition involved the processing of the fire incidence shapefiles provided by our adviser, accessing economic structure data from our adviser, and writing a script to download weather data directly from the gridMET site. From there, all data was uploaded separately to AWS S3 where the data could be merged using computing tools such as EMR.

Spark and Python were used for exploratory data analysis (EDA), data processing, modeling, performance assessment and data visualization. The workflow used in our work is documented visually below:

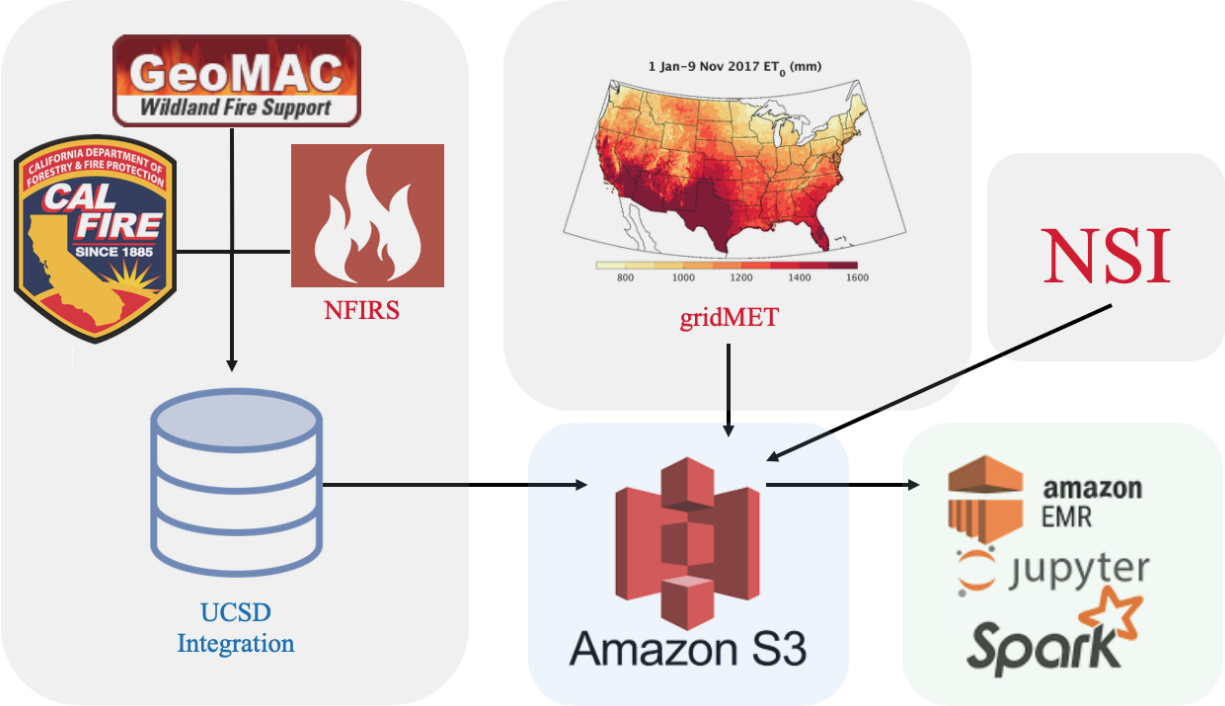


Figure 1: Data Pipeline

4 Data Preparation

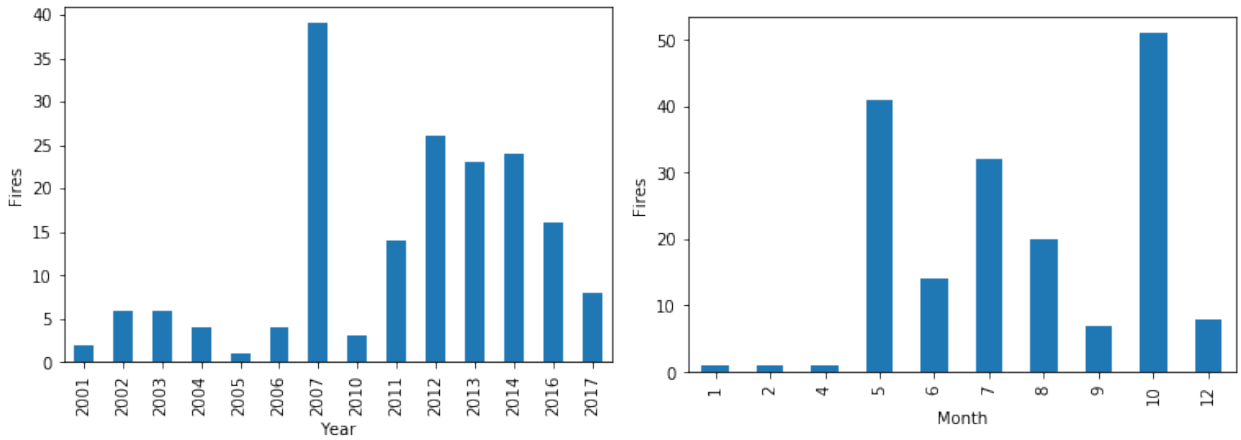
Our work was unique in that it was up to the team to research and determine which datasets would be valuable for analyzing wildfire risk and severity. As noted in the prior section, some of those datasets did not suit our criteria for use in our final pipeline for a variety of reasons specific to each dataset (data quality, cost of use, etc.). Hence, we will only focus on the three input datasets that we used in our final pipeline: the integrated fire incidence and geometry GIS files from our adviser, the NSI GIS files also from our adviser, and the downloaded gridMET weather data. To avoid issues with temporal data degradation, we only considered fires and related weather observations in the past 20 years - which luckily happened to align with the latest version of standardized fire reporting practices that were introduced in 1999 [9]. Each dataset had its own set of data quality issues, data transformations, and pre-processing methods.

4.1 Initial Analysis

4.1.1 UCSD Fire Data

The integrated Fire Data from UCSD had roughly 93K records of fires in the United States, with around 6K in San Diego County. When we filtered for fires in the year 2000 or later, the amount of fires drops to 600. Anecdotally, that number seemed too high. Assuming an even distribution of fires per year, that would mean that there were 30 fires per year. While we knew that there were significant fire seasons developing late summer/early fall, we still felt this number was too extreme. Upon further analysis, we found that there were a significant number of duplicates in the dataset - likely due from merging multiple fire perimeter sources. In order to create a 1:1 fire record per row ratio, we deduplicated fires assuming that the fire name, date, and acres burned could serve as a combination unique identifier. The last issue is that fires could burn multiple days, and while that is important to note for severity in a temporal sense, we thought the final acres burned would be satisfactory for capturing severity. Hence, we modified the deduplication critera to name, min date, and acres burned. Further manual inspection of the data confirmed that this set was unique.

The first thing we found is that fires were not evenly distributed throughout time in San Diego, with some years having notable peaks in fire incidence. Additionally, fires do not usually occur in all months as the summer months and fall have more fires. Peak fire occurrence is in October, which aligns with Santa Ana wind conditions: strong down slope winds that blow through the mountain passes in Southern California [4]. Naturally, both of these conclusions validated existing theories as some weather conditions may be more severe in a given year and it is highly doubtful that major wildfires may occur in the winter or spring (periods known to be cooler and with heavier rainfall). We can see these observations in the plot below:



(a) Yearly Distribution

(b) Monthly Distribution

Figure 2: Temporal Distributions of Fires in SD County

In addition, the distribution of the severity of the fires is quite large, with the most severe fire (Cedar Fire in 2003) burning over 270K acres or around 423 square miles. Leveraging the geometry GIS data provided that indicates the fire perimeter, we were able to look at the spatial patterns of fires:

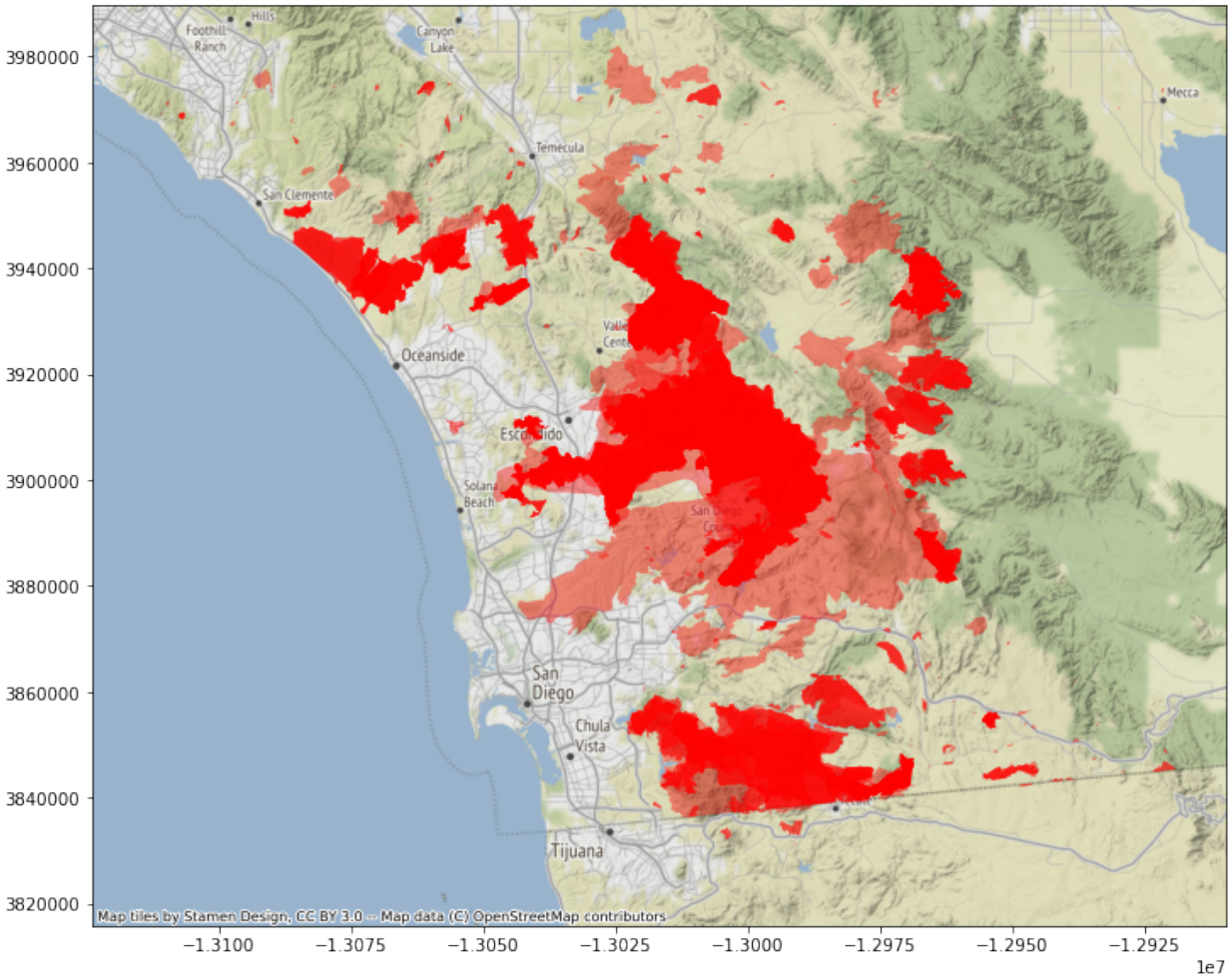


Figure 3: Heatmap of Fire Severity in SD County. Overlapping perimeters are shaded with increased opacity

The first conclusion is that major wildfires are indeed *wild* in that they do not occur in major populated areas (with the exception of near and around Escondido). The subtle piece to note is that there are no points within the red region signifying the origin (ignition) point of the fire. This is a notable omission and is unfortunately intentional. The dataset provided to us did not have the ingition point of the fire. For future modeling tasks, when we are asked to correlate weather data with the risk of fire, we have no choice but to include the entire region that the fire burned as an intersection with the weather data. This posed

a problem of data leakage in modeling tasks that we will detail later in the work.

4.1.2 gridMET Weather Data

We obtained a subset of gridMET data covering San Diego county from the years 1999 to 2019. The date range was selected so that we could use weather information for a year preceding the earliest fires in our dataset. This included 890 quadrilaterals and about 6.8 million unique data points. The variables and their meaning are described in **Table 1** below. As discussed further in later sections of this report, the potential evapotranspiration, wind direction, and wind speed all turn out to be important predictors of fires for our fire occurrence models.

Table 1: gridMET Weather Data

Weather Observation	Description
precipitation amount mm	Daily precipitation, in mm
relative humidity %	Daily ratio of the partial pressure of water vapor to the equilibrium vapor pressure of water at a given temperature, in %
specific humidity kg/kg	Daily kg of water vapor in daily kg of air
surface downwelling flux W-m-2	Daily observations of wind that causes surface water to build up along a coastline and the surface water eventually sinks toward the bottom
wind speed m/s	Daily standard wind speed in m/s
max air temperature K	Daily maximum temperature, in K
min air temperature K	Daily minimum temperature, in K
dead fuel moisture 100hr %	Daily measurement of the the time it takes a fuel particle to reach 2/3's of its way to equilibrium with its local environment, 100 hours (burn easier, not as severe)
dead fuel moisture 1000hr %	Daily measurement of the the time it takes a fuel particle to reach 2/3's of its way to equilibrium with its local environment, 1000 hours (harder to burn, very severe)
energy release component	Daily number related to the available energy per unit area within the flaming front at the head of a fire
potential evapotranspiration mm	Daily amount of evaporation that would occur if a water source was available
mean vapor pressure deficit kPa	Daily difference between the amount of moisture in the air and how much moisture the air can hold when it is saturated. Essentially, a measure of cloud cover

The following figures explore several features of the weather conditions across several

years in our dataset, including very severe fire seasons (2003 and 2007) and relatively normal fire seasons. Beginning with a polar plot on wind speed and direction, we can see how consistent the wind is in San Diego County year over year. However, some months of the year display different wind conditions: notably the phenomenon known as Santa Ana winds which typically occur in October.

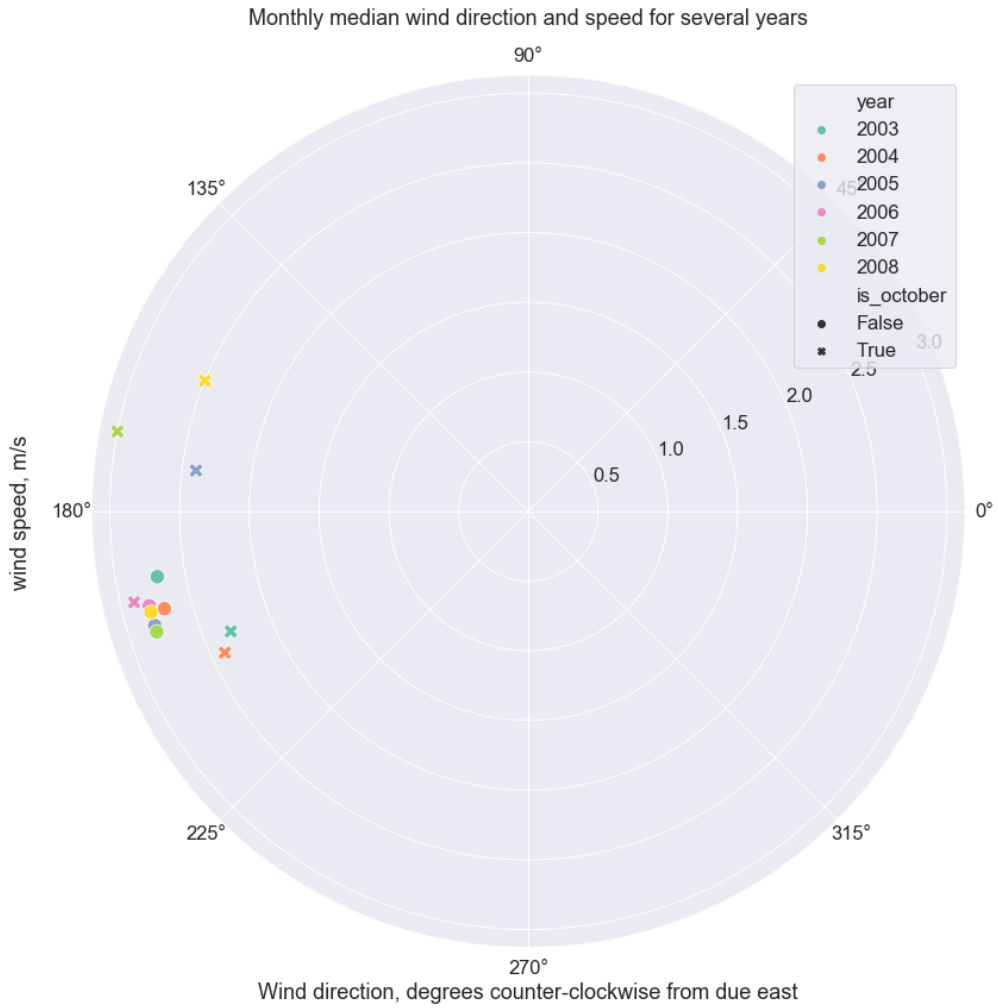
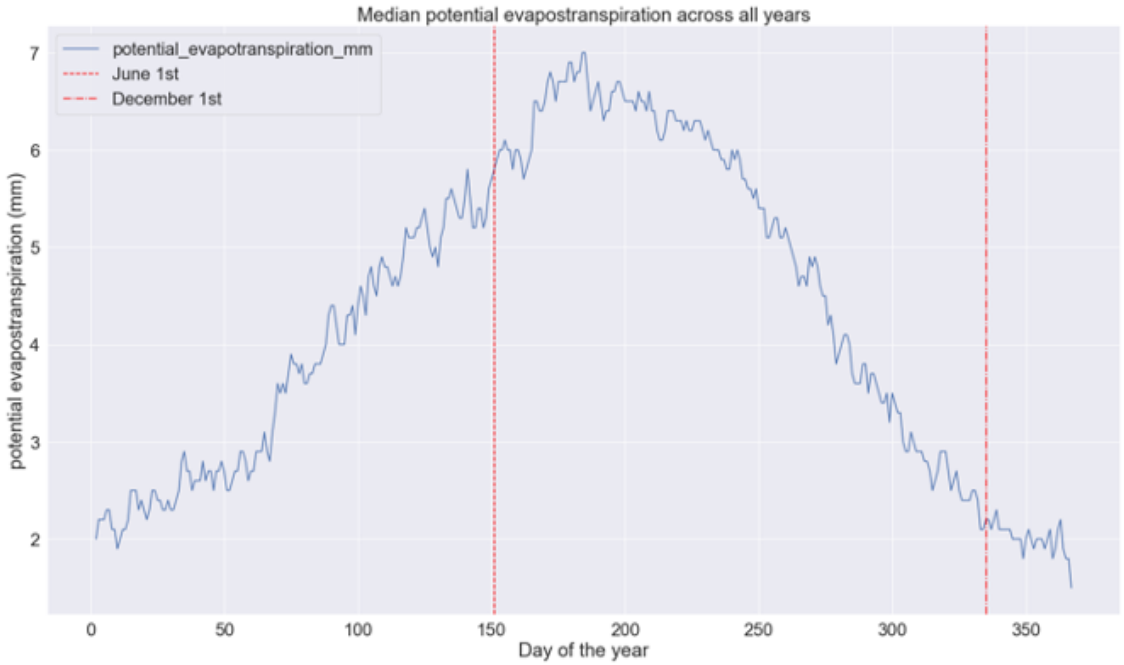
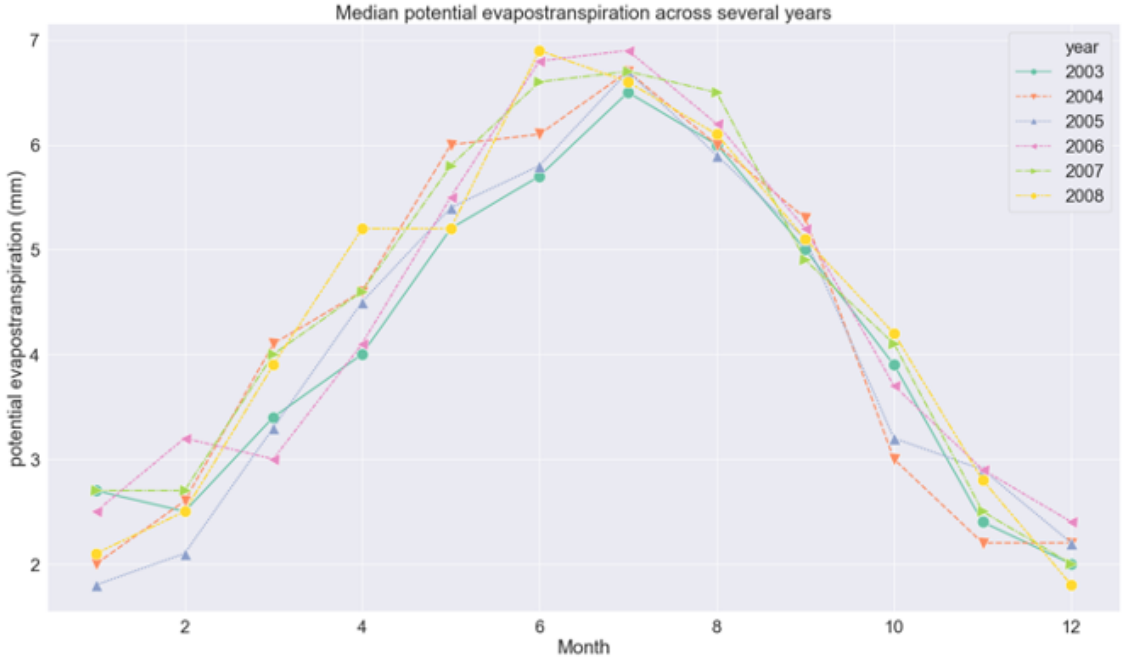


Figure 4: Monthly Median Wind Speed & Direction

Next, we look at the potential evapotranspiration in **Figure 5** on the following page, which is essentially a measure of ground moisture and potentially the most critical predictor of fire risk. Observing the plots below, we can determine that evapotranspiration is a seasonal measure - not one highly correlated with a supposed "fire-season", more details in subsequent sections. In the top plot, the red lines denote the fire season, and we see that potential evapotranspiration has considerable variation over this time period. In the bottom plot, we



(a) Median Yearly Potential Evapotranspiration



(b) Median Monthly Potential Evapotranspiration

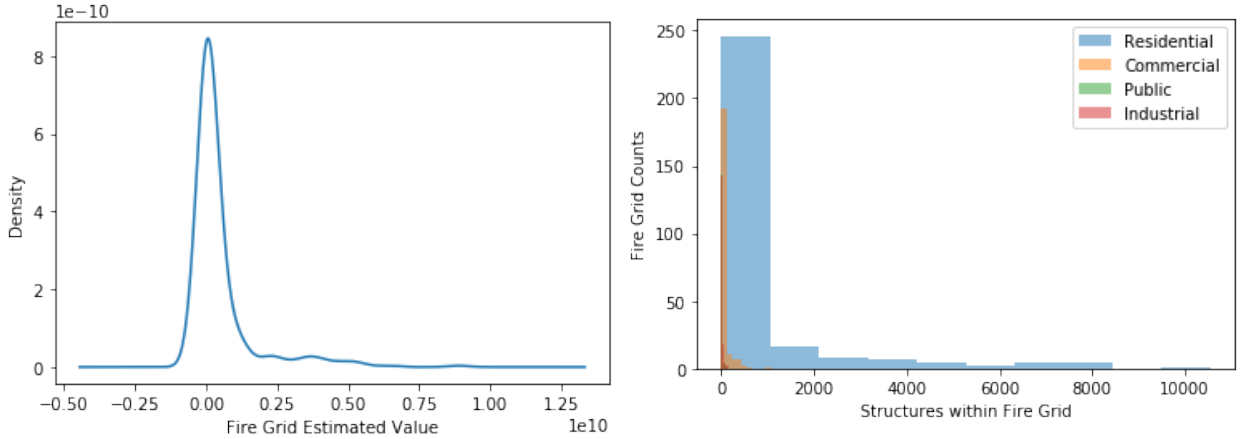
Figure 5: Temporal Views of Median Evapotranspiration

can see the year over year variation with the measure, which demonstrates how consistent it is at any point in the year.

Notably, it is difficult to discern what precisely is different about these variables under conditions of severe fire risk compared to mild or moderate fire risk, or indeed even to establish a trend across years. Part of the value of our analysis is in uncovering the importance of certain factors. Since fire risk can be extremely localized, it is important to consider local conditions, however this presents an analysis challenge as the number of locations to consider generally exceeds what is easily digestible for a single person. At the same time, aggregations that are more easily understood can mask important trends.

4.1.3 NSI Economic Data

After we produced the final merged gridMET and fire incidence data, performed necessary feature engineering and model building, our task was to contextualize the results of our model. Say, for example, a grid has high probability of fire risk, yet what does that mean for the potential impacts on the community? If for example, the risk is higher in one region with less economic value vs. the risk being slightly lower in a region with more economic value, can policy-makers make use of that data to make more informed decisions? Our belief is that resources can be allocated more efficiently with knowledge of the potential impacts to society.



(a) Density of Economic Value

(b) Distribution of Structure Type

Figure 6: Analyzing the Fire Grid Economic Value in SD County

The NSI data had roughly 874K records of structures within San Diego County. The data contained information on the estimated monetary value of structures, with valuations of the contents and vehicles. Furthermore, the data had information on the properties of the structure (foundation type, materials type, square footage, etc.). While some of these properties may indicate the relative flammability of the structure, we did not consider them as they were out of scope. Instead, we aggregated the different structure values and counts

for the grids where we had fire data. We visualize the economic value in our intersected fire grids on the previous page.

Immediately, we can determine in these wildfire intersection grids, that they are predominately empty or near empty in undeveloped areas without a direct economic impact on people. However, both graphs demonstrate long-tailed distributions, which indicate that there are some grids with significant economic value. Looking at the figure in (b), we see that there are grids with high quantities of residential structures.

4.2 Data Integration

In order to explore the predictive value of weather for fire occurrence and severity, we joined the gridMET and UCSD fire data leveraging spacial joins provided in the GeoPandas library. Each gridMET region comprises a quadrilateral area with sides extending 1/24th of a degree latitude and longitude to the north and east from a specified point. The integrated dataset augments gridMET by intersecting these regions with the historical fire perimeters in order to develop two new features indicating whether a fire occurred and the total acres burned. Each row in the integrated dataset is uniquely identified by the latitude and longitude of the southwest corner of the quadrilateral and a date (year, month, and day). A fire is considered to have occurred for a row if a historical fire perimeter intersected it by any amount on the date associated with the perimeter.

4.2.1 Duplicative Fire Grids

As a result of this joining scheme, a single fire incident consists of multiple rows when its perimeter extends over multiple quadrilaterals. Naively selecting rows at random to be included in the training or validation sets for future modeling tasks would result in a single fire incident being represented in both sets - which would inflate the perception of model robustness. In order to explore the effect of a single fire incident being included in both the training and validation sets, we developed two reduced datasets and trained various models for fire occurrence and severity on them. The reduced datasets were theoretically obtained as follows:

- For each day in the dataset, we randomly selected one record of fire occurrence as a representative for that day, discarding the others and randomly splitting the remaining occurrences into training and validation sets. Because none of the fires in our dataset span multiple days, this effectively ensured that a fire would appear exactly once in the training and validation sets, but reduced the total number of positive examples significantly to 414 rows.

- Since reducing the number of positive samples alone could conceivably affect the model results, we also prepared a dataset that randomly selected 414 positive examples, discarding the rest. This reduced dataset did not attempt to ensure that a single fire incident appeared exclusively in the training or validation set.

We concluded that separating all rows corresponding to a single fire exclusively into either the training or validation set was important. We expand on this in subsequent sections.

4.3 Feature Engineering

4.3.1 Time Lag Aggregation

Initial analysis of correlation between weather data and fire incidences suggests that gridMET meteorological data taken without further feature engineering is not highly predictive of fire occurrence. We expect historical information to be important - for example, whether the previous year had heavy rain may impact the following fire season due to increased dead fuel loads [4]. For that reason we explored the use of engineered features that aggregate past meteorological data over various time periods (e.g. 1 week or 1 month), and using various aggregation modalities (e.g. max, mean, or cumulative). Please refer back to **Table 1**, which indicates the raw weather measurements we used. As one may conclude from the list of attributes and the fact that the data is aggregated daily, we expect many of these data points to be highly correlated or even potential linear combinations of each other, while simultaneously being quite variable. That said, this dataset does appear to be as exhaustive as possible to the task at hand.

Our integrated dataset had these daily weather observations in San Diego reported in standardized grids since 1999, which amounted to roughly 6M rows of information. Considering that we have substantial class-imbalance in our dataset (think how many days we do not experience a major fire), we wanted to generate ample amounts of features that may assist in predicting such a small class (fire existed). Noted above, trailing weather conditions on a year to year basis may prove valuable for a robust classifier. Although our target class is major fires from 2000 onward, some of the lagged (cumulative or otherwise) features should be engineered based on 1999 data - hence expanding our training set to a prior year, without any additional labels of fires.

Based on empirical research, we figured more recent air data (humidity, temperature, wind, vapor pressure deficit, etc.) and more longitudinal ground data (precipitation, fuel moisture, etc.) may prove to be robust information towards predicting wildfires. For example, the amount rainfall on a given day is likely not informative of a fire in San Diego County (i.e. it does not rain here often - fire or not), however, the cumulative sum of rainfall over the course of a prior year or a month before a fire could indicate a rise in vegetative fuel for a fire - which could be especially valuable for fire prediction. By experimenting with different lag functions, we realized that lag-aggregating across our dataset was considerably costly. As one may imagine, generating weekly means/maxes by the form of serially repeating aggregations by 7 rows in a large dataset was quite substantial.

Through both parallelizing and leveraging more resources for computation, we were able to successfuly time lag aggregate our original feature set. We decided to look at both different forms of aggregation (mean, max, min, sum and standard deviation) and different subsets of

data to aggregate on (1 week, 2 weeks, all the way to 1 year). We chose different calculations other than the mean to understand:

1. Are peaks or valleys in certain weather conditions important over certain time periods?
    (a) Does a lot of wind speed in a week before a fire paired with high air temperature indicate Santa Ana?
2. Does the variability of weather conditions matter?
    (a) If it is really cold, but then really hot - is the incidence of a fire more likely? Or would consistently hot conditions matter more?
3. Accumulation of weather - does a lot of a given weather attribute over time matter?
    (a) Ex: rainfall

Based on these different combinations of aggregations derived on the original raw feature set, we were able to use Spark to engineer 560 features with different amounts of variability (and interpretability) resulting in 574 total features on 6.5M rows of weather data. Now, our problem has turned into reducing some of the noise that our data engineering process has created - a common tool for doing so is Principal Component Analysis (PCA).

4.3.2 PCA

We used PCA with Spark to reduce the dimensionality of the dataset, and additionally we derived insight from the learned PCA model about which of the features are informative by examining the principal components themselves.

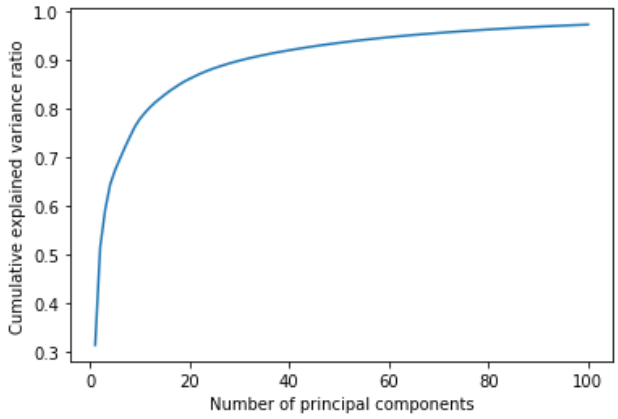
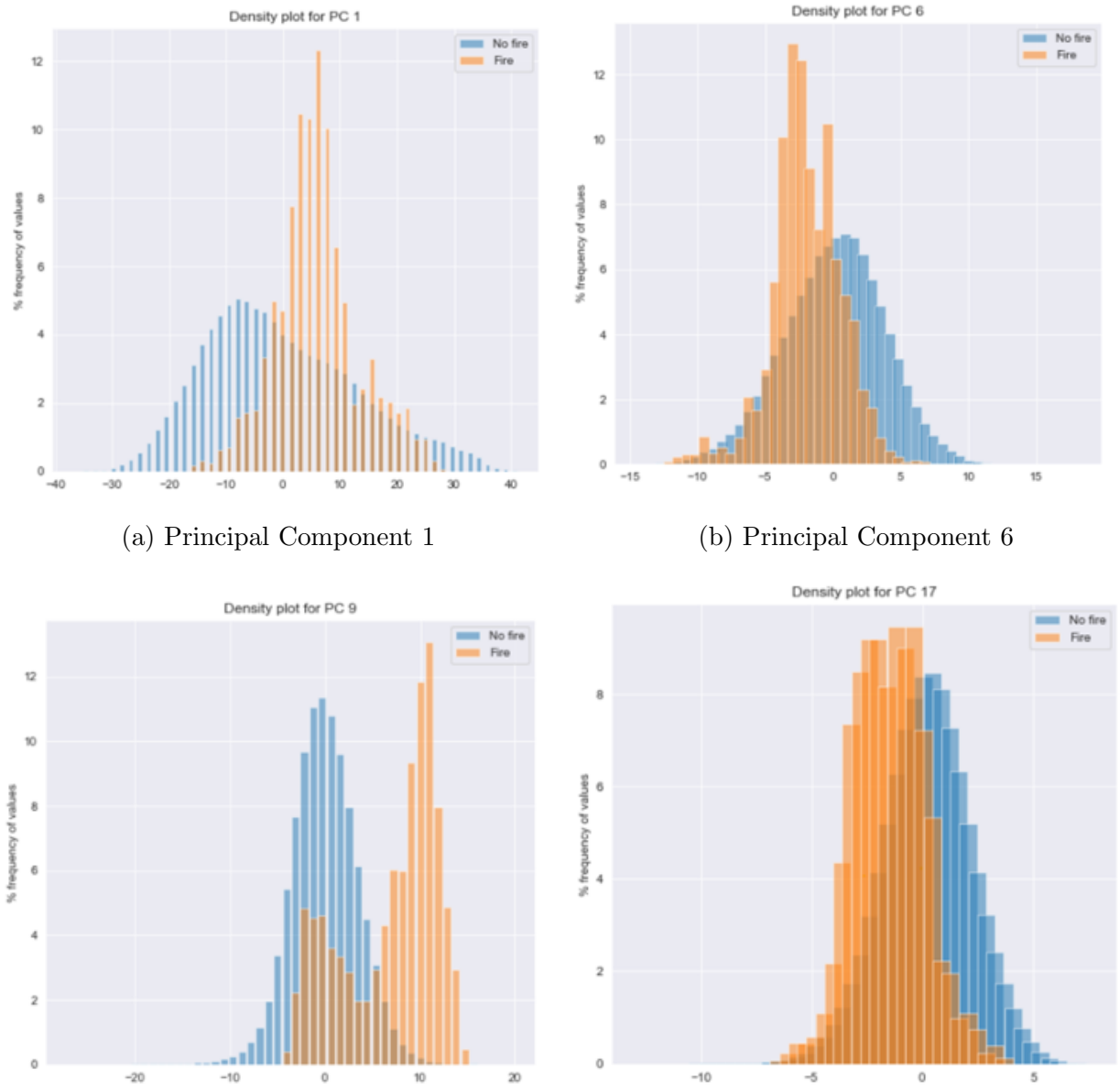


Figure 7: Analysis of PCA Results

Our analysis showed that many of the engineered features have little variance, and that only around 40 principal components are needed to explain more than 90% of the variance. Moreover the mean and cumulative aggregation modalities did not turn out to be important, whereas the min, max and standard deviation aggregation modalities did produce important features.



(a) Principal Component 1

(b) Principal Component 6

(c) Principal Component 9

(d) Principal Component 17

Figure 8: PCA Density Plots Along Target Dimension

Hence, we were able to transform 14 uninformative features with differing levels of vari-

ability into 40 features with maximum variance which are linear combinations of our original feature set.

The next task was to determine if the engineered features will be suitable for future modeling tasks, in this case predicting fire incidence. To do that, we plotted the values of the first few principal components across the target class dimension as a measure of potential model robustness. The more separation along the distribution of these eigenvectors, the better the subsequent model. The plots are available in **Figure 8** on the prior page. PC9 is an excellent candidate to be the most predictive feature in our new data set, with the remaining features as the next-best candidates. PC9 appears to have an almost bimodal distribution which would provide great separation for subsequent models to utilize.

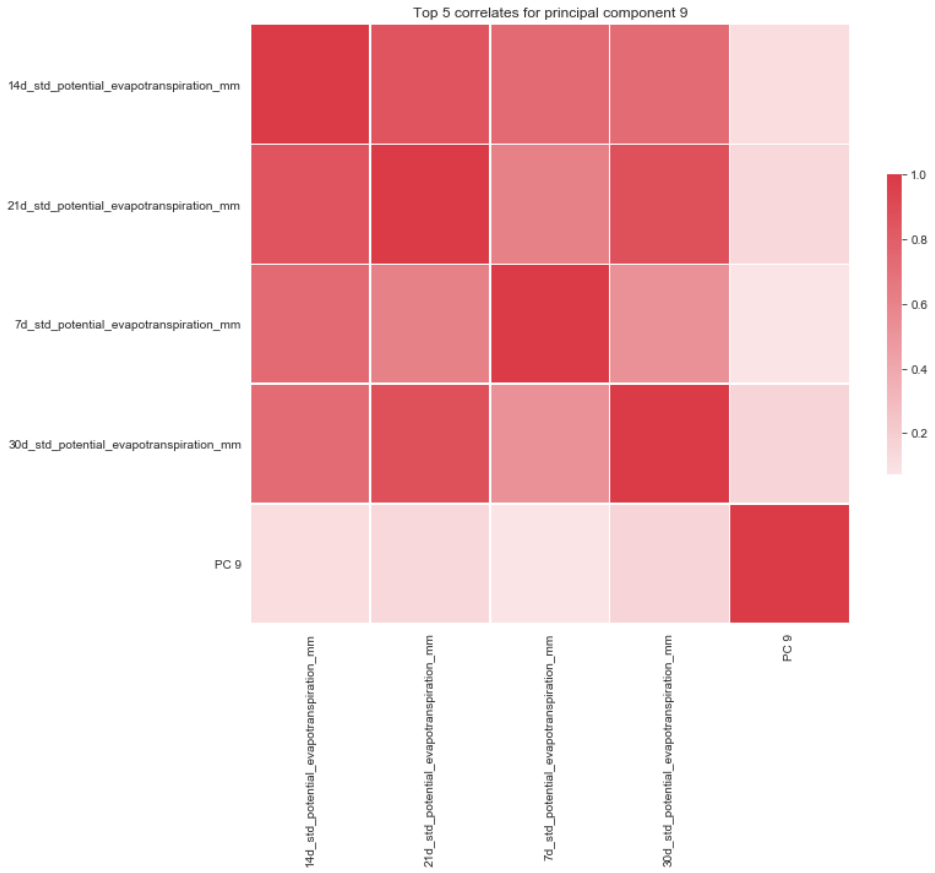


Figure 9: Correlation Matrix with PC9

While these principal components should be especially robust for predictive power, we know that interpretation of the model is equally important for policy-makers, so we also computed correlation matrices of principal components with the original features in order to help gain understanding of weather measurements that are indicative of fire risk. For

example, PC9 shows particularly strong separation between fire and non-fire days. **Figure 9** on the prior page shows that PC9 is strongly correlated to the standard deviation of potential evapostranspiration:

The benefit of the correlation matrix is that while we know PC9 is a linear combination of the original feature set, we can clearly see which of the original features it most strongly tracks with a familiar measure (correlation). By conducting this analysis, we are able to extrapolate what the important features are to predicting fire risk. We will explore these findings from our modeling and analysis work in more detail in subsequent sections.

5 Analysis Methods

After conducting an exhaustive EDA on our datasets, including the merged final and transformed dataset, our next step was to leverage this data to answer our aforementioned research questions. While analyzing the weather principal components and their correlations with fires provided considerable value for *interpretation* of their relationship, we felt that we may add value through some modeling tasks. Specifically, in order to better understand the driving factors of wildfires over time and to better understand the increased severity over time, we developed a series of machine learning models to predict both. We believed that by developing these models, we would gain a better understanding of the limitations of our data as well as their overall relevance in predicting these problems in the future.

Prior to explaining our modeling efforts below, it is worth noting that both modeling tasks were impacted by a considerable phenomenon known as "class-imbalance." In other words, the incidences in which there are not major wildfires far exceed the incidences where we do have them. In fact, while news coverage on wildfires may appear visceral at the time, we found that the incidence of wildfires in San Diego County is less than 1% over the past 20 years. Hence, for every 1000 records of weather data in our dataset, we expect slightly less than 1 fire to exist, on average. This severe class-imbalance likely impacted the robustness and furture predictability of all modeling efforts. In addition, as mentioned previously when integrating our dataset, we found that a singular fire may spread to multiple grids of weather data. When this occurred, we tended to overestimate the quality of our models since many records of fire incidence (which we initially believed to be unique) actually belonged to the same fire. This issue was corrected by arbitrarily selecting a random fire within each grid, which we chose as our approach since we did not have a record of an ignition point.

5.1 Predicting Risk: Logistic Regression

Leveraging the two reduced dataset approach outlined in the **Data Integration** section, we trained a Logistic Regression model for fire occurrence. The model results for certain choices of class weights are shown in **Table 2** on the following page. The fact that the model trained on the second reduced dataset shows superior precision and recall for some class

weights, and comparable performance for others, suggests that the degraded performance on the first reduced dataset cannot be attributed to a smaller number of samples alone. We concluded that it was important to ensure that when a fire consists of multiple rows, that it is not represented in both the training and validation sets in order to avoid possible training-validation contamination.

Table 2: Logistic Regression Tuning

Class Weight Ratio	Dataset 1 (unique) Precision & Recall	Dataset 2 (random) Precision & Recall
1000	0.37% precision, 16.9% recall	0.38% precision, 42.9% recall
10000	0.03% precision, 84.3% recall	0.04% precision, 75.7% recall
100000	0.01% precision, 97.6% recall	0.01% precision, 92.9% recall

5.2 Predicting Risk: LSTM

Long Short-Term Memory (LSTM) networks are a technique within the broader field of Deep Learning. Specifically, it is a variety of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies especially in sequence prediction problems and time series prediction. LSTM has feedback connections and is capable of processing the entire sequence of data, apart from single data points such as images. LSTM's have also been used in analyzing spatio-temporal datasets and have performed well. Knowing that there is a temporal aspect to our data (daily weather observations) and based on previous EDA tasks, the transformed time-lagged weather features have some relation with fire occurrence. We used input data as a sequence with length n days to predict fire risk on day $n+1$ to properly leverage LSTM as an approach. The advantage of using a Deep Learning model for our task, opposed to Logistic Regression, is that Deep Learning models excel in scenarios with large datasets - in which we had a fairly large dataset with $\approx$ 6M rows.

5.2.1 Applying LSTM

When applying a classification technique such as LSTM, there can be multiple different ways to tune the model to fit the task. First, independent models could be developed on a subset of the data (i.e. grid-specific models) compared to an aggregate model with a grid label. Based on the volume of data required to produce the model, we chose the latter. Second, with regards to tuning, one of the primary factors to consider is the so called "state" of the model. In a **Stateful LSTM**, the model will retain prior trained sequences prior to being dropped when the model no longer needs them. **Stateless LSTM** will only retain

information on the current trained sequence. While the former can be more robust, we implemented the latter based on the distribution of fires within the dataset (subsequent fires both temporally and spacially). Since the sequence length can be learned by the model as well, we decided to tune it through validation.

We implemented the stateless LSTM by creating unifrom sequences - they were comprised on n days and the $n+1$ day indicated the class label, fire or not. To cover the full dataset, the sequences were incremented by 1 day. This exacerbated the previous class imbalance problem, making class imbalance $\approx 0.0002\%$. To rectify the situation, we undersampled the sequences without fires and resampled the sequences with them. Thus, sequences were bifurcated intwo two classes: only unique non-fire sequences were taken and the fire sequences were created using every progressive day leading up to a fire. Fire occurrance information within the n days were dropped in the case of successive fires. Lastly, grids that never had a fire were dropped as well as some of the months that had low fire incidence.

5.2.2 LSTM Design

The input is array shaped (n,19) where n is the length of the sequence considered. We developed an LSTM model with 3 layers and the model was able to distinguish regions only after the third hidden layer, which created a non-linear boundary, was added. **Table 3** below summarizes the hyperparameters chosen:

Table 3: LSTM Hyperparameters

Hyperparemeter	Value Chosen
Sequence Length n	10 days
Input Layer	10 LSTM cells
Hidden Layer 1	10 Dense Neurons
Hidden Layer 2	10 Dense Neurons
Dropout Ratio	$> 20\%$
Loss Function	Binary Cross-Entropy
Class Weight 1	1:300
Activation Function	Inner Layers: Relu; Final Output Layer: SoftMax

Binary Cross Entropy with a skewed class weight allowed the model to focus on the target class and reduce its focus towards sequences without a fire. Typically, class weight is calculated by the inverse frequency of the sample occurrence in each class. By skewing the class weight, however, we could effectively control the recall (fraction of total relevant instances retrieved) vs. precision (fraction of relevant instances retrieved) ratio. Below, we can preview some of the results from hyperparameter tuning, note the scores are provided for both classes:

Seq Length	I/P Nodes	Hidden Nodes	Dropout	Clas weight	Loss Funtion	alpha	gamma	F1 Score	MCC Score	CM
10	40	20	0.3	None	binary_crossentropy	0	0	0.8954	0.8008	[[91346 21] [149 323]]
10	40	20	0.3	{0: 0.5, 1: 78.62}	binary_focal_loss(alpha=alpha,gamma=gamma)	0.15	2	0.8801	0.7683	[[91328 39] [158 314]]
10	40	20	0.3	{0: 1, 1: 2.5}	binary_focal_loss(alpha=alpha,gamma=gamma)	0.15	2	0.9059	0.8266	[[91367 0] [149 323]]
10	40	20	0.3	{0: 1, 1: 100}	binary_crossentropy	0	0	0.6083	0.3013	[[89308 2059] [145 327]]
15	40	20	0.3	{0: 1, 1: 2.5}	binary_focal_loss(alpha=alpha,gamma=gamma)	0.25	2	0.8919	0.7978	[[60819 12] [162 318]]
15	40	20	0.3	{0: 1, 1: 2.5}	binary_focal_loss(alpha=alpha,gamma=gamma)	0.65	2	0.8895	0.7961	[[60825 6] [169 311]]
25	40	20	0.3	None	binary_focal_loss(alpha=alpha,gamma=gamma)	0.25	2	0.8896	0.7882	[[36373 29] [152 324]]
45	40	20	0.3	{0: 1, 1: 200}	binary_focal_loss(alpha=alpha,gamma=gamma)	0.65	2	0.9003	0.8063	[[20092 34] [131 340]]

Figure 10: LSTM Training

5.3 Predicting Severity: Random Forest

Severity with regard to acres burned is a continuous value and therefore a regression task. However, with such severe class-imbalance, generating a robust enough model that can regress on a value with considerable variance was not a task we decided to perform. Instead, we re-purposed predicting severity to a classification task where we binned fires according to different thresholds of acreage burned. The logical question that follows is: how should we appropriately bin the fires? According to [10], they define three categories of fires: small (up to 5000 acres burned), medium (5000 to 150000 acres burned) and large (more than 150000 acres burned). Yet, looking at data in San Diego County, large fires were especially rare. Hence, in our collaboration efforts with SDG&E, they shared their binning criteria which we implemented in our subsequent modeling tasks: small (less than 20 acres), medium (20 to 250 acres), and large (250+).

5.3.1 Applying Random Forest

Training a model on a multiclass target class limits the applicable models. In the beginning, we attempted to use Random Forest, Xgboost, Logistic Regression, and Support Vector Machines. After validating the results on a held out dataset, we determined that Random Forest appeared to be the most robust to our data. From there, we manually created training, validation, and test datasets to control for potential fire duplicates (as noted previously). Based on the variability present in the data, we chose 2004 to be the validation dataset used to tune hyperparameters. Furthermore, class imbalance persisted with our multiclass fire bins. In the figure below, we note how the distribution of classes varies year to year. This imbalance made modeling on these labels a considerably challenging task.

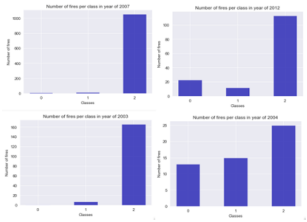


Figure 11: Multiclass Distribution by Year

6 Findings and Reporting

Here we present the findings from our various modeling and analysis tasks. We chose to present the results of all three methods, seeing as each could be equally valuable approaches towards building on the work in the future. Predicting severity could follow a similar multiclass approach and predicting occurrence can be achieved with either Logistic Regression or LSTM, dependent on data volume and computational resources. In addition, we have chosen to develop a dashboard that visualizes our predictions against actual outcomes for the Logistic Regression predictions. Furthermore, the dashboard also allows for exploration of the economic structure values for the grids that had fires in them.

6.1 Logistic Regression

For the fire occurrence classification problem, we trained separate logistic regression models for every year in our dataset to produce predictions that simulate working on unseen data. The model for a year is trained on every other year. The result is a daily map of fire occurrence probabilities over the San Diego region. The large number of false positive predictions means the models are not useful for predicting actual fire occurrence, but do serve as an indicator of whether local conditions could support a wildfire, an interpretation that is supported by experts at SDG&E.

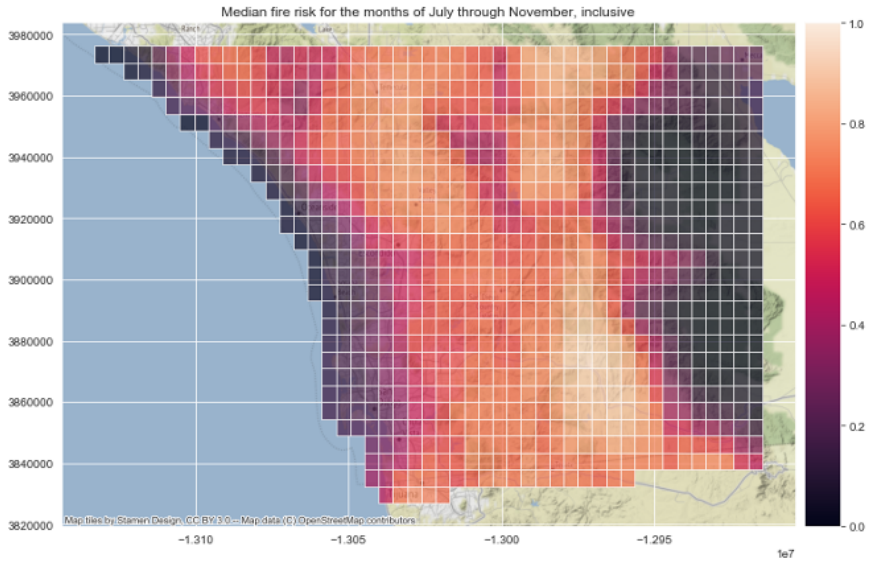
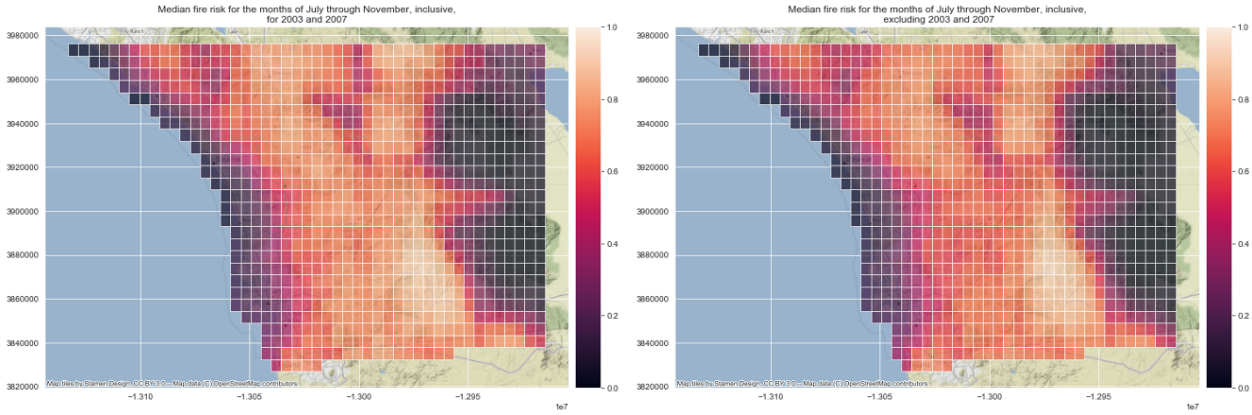


Figure 12: Median Fire Risk for Months Jul-Nov in San Diego County

By aggregating over time periods of interest, general trends can be observed immediately from these visualizations (see Figure 12 above), namely that fire risk generally follows a corridor inland between the coast and the mountains, and trails off further inland still.

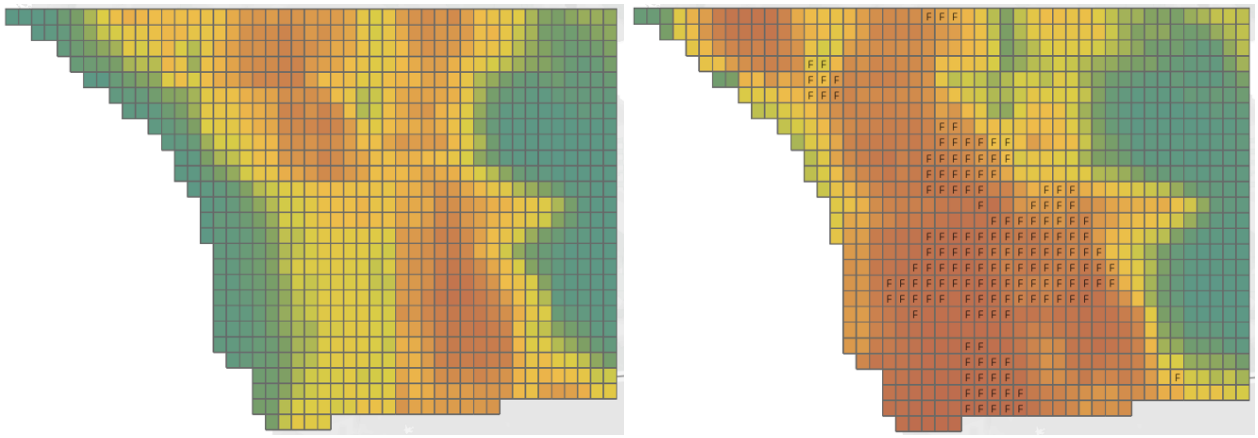
Some areas consistently have low or high fire risks, and such information could in principle be used to strategically deploy resources to combat fires.

Similarly one can start to answer questions such as, for very severe fire years, does the fire risk look different? **Figure 13** remarkably shows that the median fire risk for the months July through November for 2003 and 2007, which are both considered extraordinarily severe fire seasons, is actually very consistent with the median fire risk in the same span of months for all other years. Note however that while the median risk for the whole season is similar, the median fire risk for individual months can vary considerably. **Figure 14** shows that October 2003 has atypically higher risk in substantial portions of the county.



(a) Mean Risk in October In 2003 & 2007 (b) Median Risk Jul-Nov Except 2003

Figure 13: Mean Risk in October Except 2003 & 2007



(a) October 2001 (b) October 2003

Figure 14: Mean Risk in October 2001 vs. 2003

There remains a variety of interesting questions that if answered may improve the utility

of our risk measure, or inform the design of a more robust risk measure, but which we did not have the time to explore in our project. For example the San Diego fire season is generally considered to start in June, however we found that the 2007 fire season was prefaced with very high fire risk in the months preceding June. By comparison, the year 2010 shows very little fire risk in the same time period. Can the average risk over preceding months be used to predict the overall severity of a season? We have developed the means to answer such questions, but lacked the time to explore them fully in the course of our project. Exploring the data more will certainly suggest other hypotheses as well.

6.2 LSTM

The LSTM model predicting fire occurrence was deployed on the historical weather data in San Diego County. For performance evaluation, the data is split into 20 years of training data and 1 year of test data. The model is asked to predict fire risk for every day throughout the year, repeated for every test year with the remaining 20 held out for training. Performance metrics are provided summarizing these tasks in the figure below:

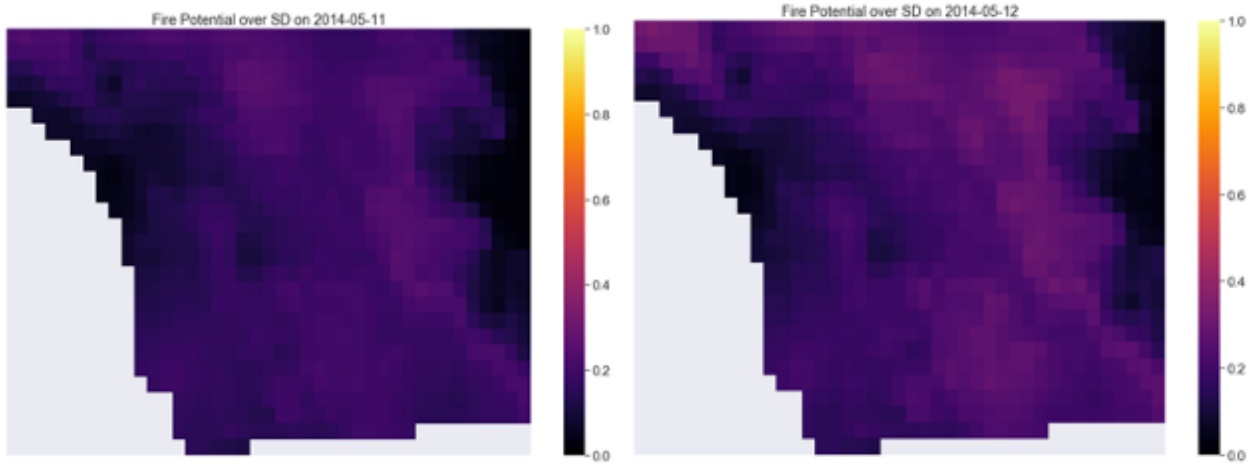
Year	F1_score	F2_score	MCC	Precision	Recall	missed_fire	Warned Fire Count
1999	0	0	0	0.00%	0.00%		0
2000	0	0	0	0.00%	0.00%		0
2001	0	0	-0.0025	0.00%	0.00%	bell	2
2002	0.00215	0.00536	0.02303	0.11%	68.00%	gavilan,nate	16
2003	0.01138	0.02792	0.06788	0.57%	89.66%		14
2004	0.00107	0.00266	0.013	0.05%	54.72%	merald,el mo	27
2005	0.00147	0.00367	0.02039	0.07%	78.95%		36
2006	0.00192	0.00478	0.0247	0.10%	86.30%	gunn 2	19
2007	0.03107	0.07373	0.1029	1.58%	87.28%		28
2008	0.00072	0.00179	0.01496	0.04%	85.71%		11
2009	0.00024	0.00059	0.00716	0.01%	80.00%	grande ic	9
2010	0.00094	0.00234	0.01734	0.05%	83.87%		16
2011	0.00239	0.00596	0.02928	0.12%	92.77%		14
2012	0.00554	0.01373	0.04388	0.28%	85.14%	cottonwood	27
2013	0.00262	0.00651	0.02756	0.13%	77.27%	vail	17
2014	0.00429	0.01065	0.03722	0.21%	82.84%	gun	19
2015	0.00033	0.00082	0.00887	0.02%	73.33%		7
2016	0.0015	0.00374	0.02431	0.07%	97.87%		4
2017	0.00119	0.00296	0.01929	0.06%	82.50%		8
2018	0.00023	0.00056	0.00497	0.01%	41.67%		8
2019	0	0	0	0.00%	0.00%		0

Figure 15: High Recall Rate with Low Precision in LSTM

There are scenarios where fires occur on subsequent days in addition to the first days within the sequence. Lower recall rates may be attributed to predicting the spread of the fire after initial ignition, something in which the addition of a physical model may help. Abstracting that out, the model was evaluated if it could predict fire occurrence on at least one grid a day or before the actual fire occurrence. In this isolated scenario, the LSTM has an average testing Recall rate of 96%, with most years being 100%. Completely missed fires were typically small fires.

There is considerable opportunity to boost precision with other datapoints. However, it is worth noting that the model was heavily tuned towards not missing a fire. The model does properly identify both the "fire-season" or the temporal aspect of fire risk and the more inland spatial risk. These results are similar to the results derived from the Logistic Regression model.

We take an instance demonstrating how the model predicts a fire in sequential format in **Figure 16** below. Beginning with the sequence concluding on May 11th, with the image displaying the risk generated from the preceeding 10 days of weather data, subsequent sequences gradually increase the risk probability. Grids marked with an F denote instances of reported fires, in this case it is the Tomahawk Fire - estimated to cost roughly $29.5 million.



(a) LSTM Tomahawk 5/11

(b) LSTM Tomahawk 5/12

(c) LSTM Tomahawk 5/13

(d) LSTM Tomahawk 5/14

Figure 16: LSTM Model Predicting Tomahawk Fire Sequence

Demonstrated above, the probability of fire risk increased throughout different grids in

San Diego County, as weather conditions built up to be conducive for ignition. Although, only a few of those regions actually had fire instances, indicating if ignition was present, most of these regions could have been burned. We strongly believe that with more information on ignition points and merging that data into this existing work would create both a more accurate and more precise fire occurrence risk index.

6.3 Random Forest

A Random Forest classifier was implemented across all years to predict the multiclass binned fire severity. Aligning with prior statements, the model misses mostly small fires which could be attributed due to class imbalance. **Figure 17** summarizes the model's performance on held out test data in each year:

	Random forest Classifier			MLP Classifier			Fire Samples		
Year	F1_score	Recall large fires	Precision large fires	F1_score	Recall large fires	Precision large fires	Small	Medium	Large
1999	0%	0%	0%	0%	0%	0%	0	0	0
2000	0%	0%	0%	0%	0%	0%	0	0	0
2001	82%	87%	93%	41%	33%	100%	0	2	15
2002	79%	88%	85%	77%	100%	77%	0	17	58
2003	**97%**	**99%**	**98%**	**97%**	**99%**	**98%**	**1**	**7**	**166**
2004	N/A	N/A	0%	N/A	N/A	N/A	N/A	N/A	N/A
2005	25%	100%	15%	44%	60%	26%	15	32	10
2006	71%	94%	76%	41%	47%	68%	10	10	52
2007	**98%**	**100%**	**99%**	**98%**	**100%**	**99%**	**11**	**12**	**1054**
2008	43%	0%	0%	29%	0%	0%	1	27	0
2009	53%	100%	29%	67%	100%	40%	3	10	7
2010	45%	64%	47%	61%	93%	59%	3	14	14
2011	80%	100%	79%	83%	88%	90%	0	19	64
2012	76%	94%	82%	76%	100%	76%	23	12	113
2013	85%	99%	86%	70%	75%	92%	5	11	72
2014	**91%**	**100%**	**92%**	**90%**	**100%**	**90%**	**4**	**9**	**121**
2015	33%	100%	31%	27%	100%	27%	1	10	4
2016	98%	100%	98%	79%	78%	100%	0	2	45
2017	85%	100%	85%	70%	70%	92%	0	7	33
2018	42%	60%	38%	58%	100%	50%	1	6	5
2019	0%	0%	0%	0%	0%	0%	0	0	0
Average	**70%**	**87%**	**67%**	**65%**	**79%**	**69%**			

Figure 17: Random Forest Model Validation Statistics

The Random Forest has variable accuracy, yet the overall mean is 68%, which outperforms the alternative NN MLP classifier. It is evident that besides high-class imblance issues, weather data is not enough to model the severity of fire with suitable accuracy. However, there is still enough value in this model for it to be usable over no other information.

Taking a look at a sample year, 2012 for instance, it has a multiclass distribution of 23, 11, and 113 respectively (small-large). The model achieves an accuracy of 75% with fairly good results. Most of the small fires are mislabeled as large fires and large fires are mislabeled as medium fires. Considering large fires create the most economic problems and damage to communities, the recall on large fires is adequate. Below, we have plotted a confusion matrix describing performance in 2012:

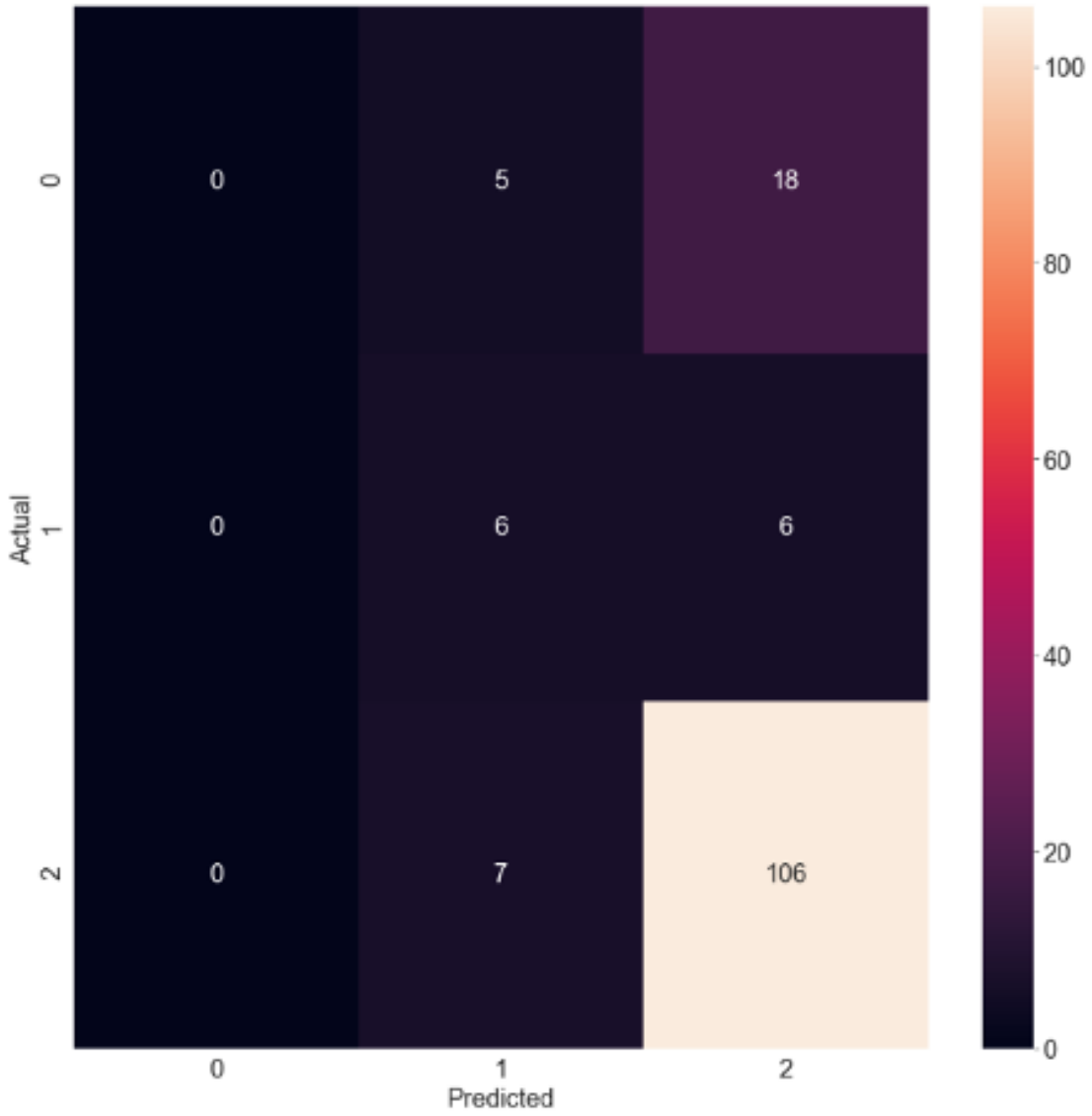


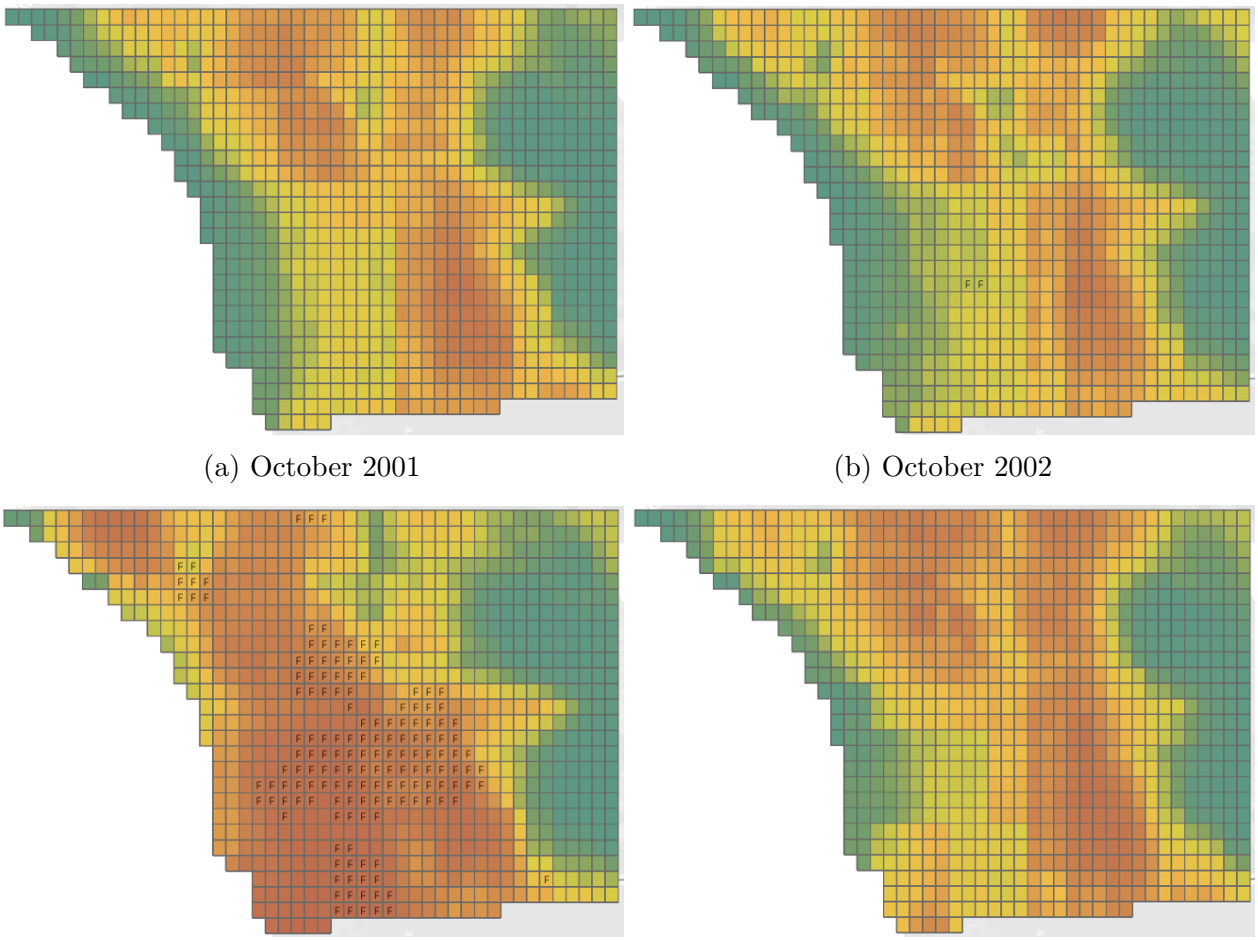
Figure 18: Multiclass Confusion Matrix with Fire Encoding 0-Small, 1-Medium, 2-Large

6.4 UCSD Wildfire Dashboard

The reporting dashboard that we developed to adequately visualize our results can be found here: UCSD Wildfire Dashboard.

Using this analysis broad spatial and temporal trends can be explored. Leveraging the temporal filter on the right-hand side of the first tab, we can determine that eastern San Diego county has several regions with consistently small risk, and the coastal regions generally also have low risk. The risk is generally higher in an inland corridor that roughly starts west of Interstate 5. The relative riskiness of consecutive years can easily be explored in the figure

below, where we can see the anomalously high risk for October 2003 in the soutwestern corner of the county compared to previous and subsequent years. It's also possible to observe the evolution of risk across a year and identify years with relatively high risks versus years with relatively low risk.



(a) October 2001

(b) October 2002

(c) October 2003

(d) October 2004

Figure 19: Mean Risk in October for Select Years

The second tab in the visualization product explores the economic value of grid regions that contained fires in the dataset. The total value as measured by the structure index can be displayed. Grid regions that did not contain fires are omitted in order to emphasize the impact of wildfires. Notably, we see that the region surrounding Escondido poses the most historical risk.

6.5 Other Fire Occurrence Factors

Clearly, based on the results from our three modeling sections, the weather alone is not sufficient to predict fire occurrence. However the results show that indeed other factors are in fact incredibly important for fire occurrence. If weather conditions played an outsized role, then we would expect to see higher estimated fire risks for more severe fire seasons, however in fact we see over the time scale of months that estimated fire risk is relatively uniform from year to year (and severity is mostly uniform predicting large fires). It is also possible that the uniformity of estimated risk is a quirk of our particular models. A fire occurrence model with greater precision and similar recall could clarify the impact of the model in estimated risk. Our discussions with fire experts suggest that other factors, such as the number of campers visiting backcountry areas or the state of electrical grid infrastructure, are very important.

In addition, geographical, vegetation, and social aspects of potential population density could be valuable to input into these models. These inputs could be used to augment the training data and potentially boost Precision without sacrificing Recall. Physical models could be layered on our fire incidence and severity models to create simulations of spread.

7 Solution Architecture, Performance and Evaluation

7.1 Measuring Performance

We evaluated the performance of our models by recall and precision. The reason we chose these metrics instead of another (such as accuracy) is due to the nature of fire occurrence. Fires may be ignited by non-weather phenomena, for example by a human with malicious intent or by accident, and since we didnt incorporate a dataset that could account for those factors, we didnt expect to accurately predict occurrence. However, we did expect that an occurrence model trained on weather could at least indicate conditions where a fire may be supported and serve as a measure of fire risk. For this reason, we determined a successful model would be one that could over-predict fires, but did not miss a fire, which is why recall is an important metric. At the same time, fewer false positives is clearly desirable so we chose to use precision as a secondary measure in order to select against models that always predict fires.

7.2 Model Scalability

In order to deal with models with compute requirements beyond what our laptops could handle, we moved to AWS and scaled both horizontally (with more machines) and vertically (with machines that individually had more resources). With our LSTM model scaled vertically, moving from a personal machine to an EC2 instance on AWS which had 4 cores and 16GB of memory. With our logistic regression model, it was a bit more work to have it scale horizontally. We changed frameworks from sklearn to pyspark, and then moved the model to AWS. Using AWS EMR, we were able to train and predict with the model. The main

reason for the change was due to the memory requirements of our feature engineering, which was beyond what our personal computers were capable of handling.

7.2.1 Scaling the Pipeline

We chose to limit our analysis to San Diego county initially since the volume of raw data for that region could fit in the main memory of a typical consumer laptop. Even still, some of the computations we performed had memory and compute requirements that exceeded the capabilities of typical consumer laptops, and required a distributed computing approach. We extended our data ingestion pipeline to all of California, with some modifications intended to reduce the compute requirements, as a proof of concept for a larger analysis. We did not extend our modeling efforts to all of California as our budget couldnt support it.

With our original pipeline we used geopandas to spatially join historical fire perimeters with gridMET quadrilaterals as described in another section. This join requires computing the intersection of irregularly shaped polygons, which turned out to be a bottleneck in our pipeline. For the proof of concept, we decided to approximate the geometries of entities in our dataset using geohashes instead of latitude-longitude pairs. Geohashing takes a latitude-longitude pair, and converts it into a variable-length string encoding a rectangular region encompassing the point – a more precise area can be obtained by computing more characters in the string representation. We expected that computing intersections using geohashes would require fewer compute resources compared to a geometric approach. With the newly joined dataset, we recreated our feature engineering in Spark SQL using window functions. Even still, the feature engineering demanded considerable compute resources on AWS, and we decided not to proceed with further analysis due to the expected cost exceeding our budget.

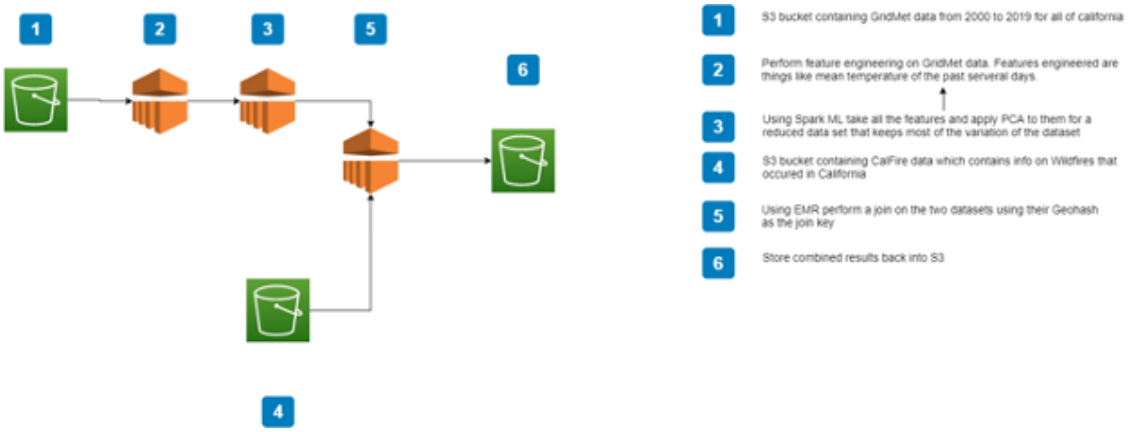


Figure 20: Alternate Scaled Pipeline Approach

7.3 Budget Management

With the two thousand dollar budget, initially we tried to be as conservative as possible. At first we tried a few experiments in order to familiarize ourselves with AWS and all of its resources. For our initial model building and EDA, we were working with a subset of gridMET that only contained San Diego weather data, and for most of our analysis this dataset could be contained on our personal machines. Eventually some of the models needed more memory and compute power, so we moved over to EMR, which could provide clusters of machines large enough to train the models. For the LSTM model we used a single m5.xlarge instance, and the logistic regression model we used a cluster of five m5.xlarge instances. A m5.xlarge has 4 cores for compute and 16 GB of memory which costs about $0.192 per hour. For the proof-of-concept extension of our pipeline to all of California, we saw a large jump in spending due to the increased compute and memory requirements. With this effort r5.16xlarge was spun instead of the m5.xlarge due to memory issues and issues specifically with AWS limitations. With the larger machines spun up and data being moved across AWS, we incurred some heavy costs of roughly $500 for the data transfer, and around $800 for the larger machines. Shortly after this we neared the budget, but also neared the end of the project.

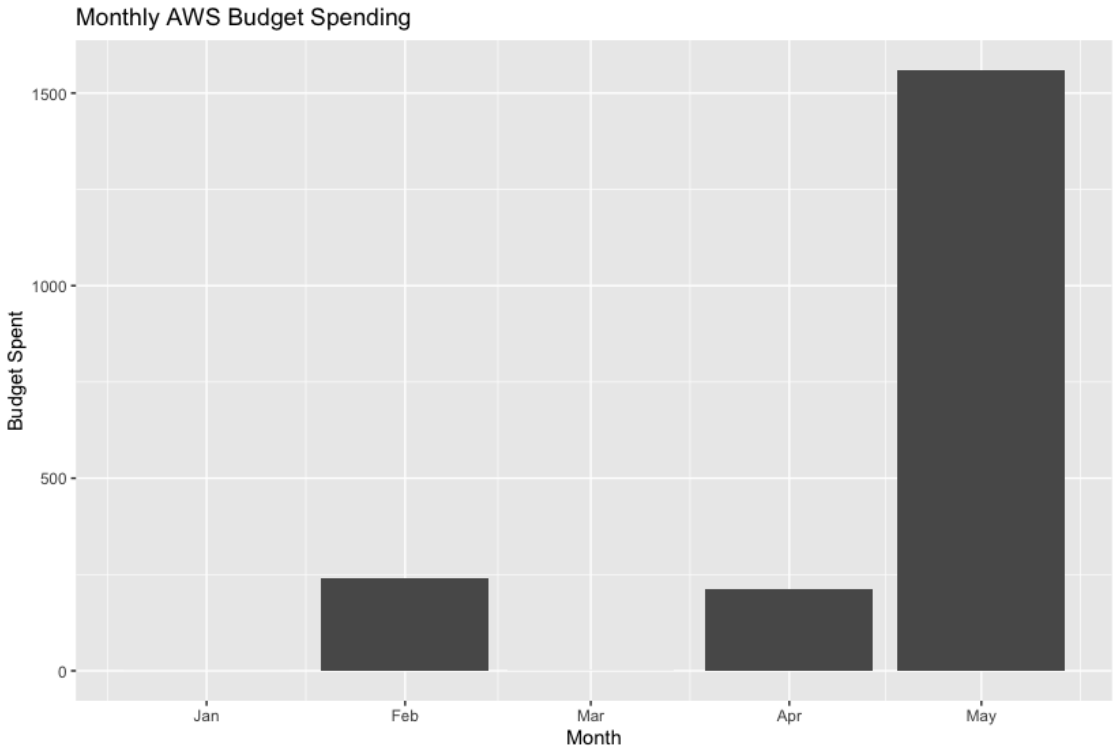


Figure 21: Significant Resources Used for Modeling Tasks

8 Conclusion

We have created models for predicting wildfire risk and severity that have considerable value to the public, researchers, firefighters, and policy-makers. Moreover, we believe we have adequately researched not only wildfire characteristics, but also have presented the relationship between changing weather conditions and wildfire risk/severity. Our product is unique in that it was built by publicly accessible datasets using techniques that can be applied to any region or any time frame and its output is available for public exploration. While we believe there is ample room for improvement with our models, notably with the inclusion of more datasets, we feel we have laid a successful foundation of work that can be built upon. Our approach has identified that there is a reasonable way to identify if a region is conducive for major wildfires.

In collaborating with our advisors and SDG&E, we learned that natural fire ignition causes are very rare - with the most *common* being lightning. Hence, with human causes, the ignition can be random with any number of possibilities. Predicting risk, therefore, is a layered problem: is it an attempt to predict ignition or rather is it predicting conditions conducive for ignition? The latter seems to be the more feasible approach which may slightly alter the course of future work. Specifically, there may be a bigger emphasis in acquiring geographical, vegetation, or population density datasets that may help to highlight these conditions.

Lastly, it is our belief that wildfire research is considerably limited by the availability of quality fire datsets. Through discussions with firefighters, we learned that data collection practices with NFIRS are inconsistent - largely leading to missing or improperly standardized data. Participation on entering data on such fires, considered to be the largest database of fires in the United States, remains optional. Many above-average entries provided have not determined an ignition point, or an estimated one. By improving data collection practices, not only would the tasks of predicting risk and severity would become easier, but also would gaining a better understanding of the associated costs for each fire. So even though these models may be improved by future work by incorporating more datasets, we believe more resources should be dedicated to increasing quality data collection.

A Link to UCSD Library Archive for Reproducibility

Gallaspy, Michael; Kannappan, Kevin; La Pierre, Martin; Maeouf, Sofean; Masilamani, Sathish; Altintas, Ilkay; Nguyen, Mai; Crawl, Dan; Corringham, Tom (2020). Wildfire Data Analysis: Predicting Risk and Severity in San Diego County. In Data Science & Engineering Master of Advanced Study (DSE MAS) Capstone Projects. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0G15ZB3.

References

[1] Smith, Adam B. "2018's Billion Dollar Disasters in Context: NOAA Climate.gov." 2018's Billion Dollar Disasters in Context — NOAA Climate.gov, 7 Feb. 2019, www.climate.gov/news-features/blogs/beyond-data/2018s-billion-dollar-disasters-context.

[2] Read, Paul, and Richard Denniss. "With Costs Approaching $100 Billion, the Fires Are Australia's Costliest Natural Disaster." The Conversation, Monash University, 21 Jan. 2020, theconversation.com/with-costs-approaching-100-billion-the-fires-are-australias-costliest-natural-disaster-129433.

[3] Quiggin, John. "Opinion: Australia Is Promising $2 Billion for the Fires. I Estimate Recovery Will Cost $100 Billion." CNN, Cable News Network, 10 Jan. 2020, edition.cnn.com/2020/01/10/perspectives/australia-fires-cost/index.html.

[4] E. Tourigny, J. Bedia, O. Bellprat, L. Caron, F. Doblas-Reyes, *An observational study of the extreme wildfire events of California in 2017 : quantifying the relative importance of climate and weather* EGU General Assembly Conference Abstracts (2008) 9545.

[5] Altintas I., Block J., de Callafon R., Crawl D., Cowart C., Gupta A., Nguyen M., Braun H.W., Schulze J., Gollner M., Trouve A., Smarr L., Towards an Integrated Cyberinfrastructure for Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience of Wildfires. In Proceeedings of the Workshop on Dynamic Data-Driven Application Systems (DDDAS) at the 15th International Conference on Computational Science (ICCS 2015).

[6] Abatzoglou, J. T. (2013), Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol., 33: 121131, www.climatologylab.org/gridmet.html.

[7] Geospatial Multi-Agency Coordination. GeoMAC incident database. https://www.geomac.gov/

[8] California Department of Forestry and Fire Protection (CAL FIRE). CAL FIRE incident database. https://www.fire.ca.gov/

[9] Federal Emergency Management Agency. "National Fire Incident Reporting System 5.0: Complete Reference Guide." Homeland Security Digital Library, United States. Federal Emergency Management Agency, 31 Dec. 2007, www.usfa.fema.gov/data/.

[10] Hoover, Katie, and Laura A Hanson. Wildfire Statistics. Congressional Research Service Reports, Congressional Research Service, 3 Oct. 2019, fas.org/sgp/crs/.