Using Analytics for Wildfire Prediction and Management

15.072 Advanced Analytics Edge

Kiran Gite

Pei-Pei Kuo

Jorge Quintanilla

Rebecca Schubertrügmer

11/25/2020

1. Introduction

Wildfires in places like Australia and California in 2020 made headlines globally before COVID-19 became the new problem for this year. While wildfires are normal and can be beneficial to ecological systems, the unprecedented large, uncontrollable wildfires that have occurred in recent years can cause serious property damage, harm ecosystems, and negatively impact people's health [1,2]. We are motivated to use analytics in order to help government bodies manage their limited firefighting resources in a way that can reduce the devastating impact of wildfires.

We investigate two applications of analytics for wildfire management in California. First, we try to predict whether a fire will be large or small using weather data at the time of fire discovery. Previous literature has accomplished this task with moderate success [3]. Second, we try to estimate the expected time until the next fire in a county using past weather and fire data for that county. We show that using predictive analytics applied to wildfire management has the potential to save the state of California roughly $300 million in total fire damage costs in 1 year.

We obtained our data from two different data sets, both from the National Wildfire Coordinating Group (a US government agency). The first dataset contains data for daily weather observations in different locations in California [4]. The second dataset contains fire observations [5]. We have >30,000 observations of forest fires merged with weather data on the fire discovery date in the fire discovery location. More information and exploratory visualizations are in Appendices 1 and 2.

2. Analysis and Results: Classifying big and small fires

Our first goal was to predict fire size to help government officials make decisions about firefighting resource allocation. Regression to predict numerical fire size in acres did not give good accuracy (Appendix 3), so we pivoted to binary classification, trying to identify whether a fire is "big" or "small". From the data, we observed that all fires under 0.5 acres took less than 1 day to contain, and 90% of fires under 0.5 acres were contained within 5 hours (Appendix 2). Thus, we can safely assume that the end users of our model also will not be making predictions on fires under 0.5 acres, because these fires will likely be contained by the time the fire discovery information is communicated to the decision makers. We defined a small fire as ones that burned less than 2 acres, and defined big fires as ones that burned 2 acres or more. We chose 2 acres as a cutoff, as that is the size of about 12 average lots in California, which we anticipate would take time to contain [9]. With these definitions of "big" and "small" fires, we have 53% small fires and 47% big fires in our dataset. One issue with this split is that the range of fire size for big fires is rather large (2-300000 acres), but we are assuming each big fire gets the same intervention for cost calculation purposes. Therefore, we consider the cost of a big fire in this problem to be an average cost for our cost analysis later.

We tested different models to find good fire size classifications. We made predictions using Logistic Regression, CART, Random Forest and Boosted Tree models. All hyperparameters were tuned through cross-validation. We also tested different loss functions for our models, since the cost of a false negative (predicting a fire is small when it is actually big) is higher than the cost of a false positive in our wildfire management context. Hence, we would not approve a model which had a false negative rate higher than 25% (missing at most one of every 4 big fires). Taking all of this into consideration, we evaluated all of our models which aimed for high accuracy, high area under the curve (AUC), and low false negative rate.

Table 1 - Prediction Model Metrics

Model	Test Accuracy	Test AUC	Test False Negative Rate (rate of missing big fires)
Logistic Regression	0.5614	0.6161	0.1580
CART	0.5931	0.6210	0.1679
Random Forests	0.6498	0.6391	0.2470
Boosted Trees	**0.6564**	**0.6438**	**0.2216**

Different modeling techniques produced models that performed similarly. We ultimately decided to go with the Boosted Trees model, due to the best test accuracy and AUC. Our second choice was CART due to interpretability (results in Appendix 5). To build our boosting model, we used LightGBM, an efficient gradient boosting framework. We used grid search with cross validation to find the best parameter values for maximum tree depth, number of tree estimators, learning rate, and other hyperparameters (refer to Appendix 4 for details). When assessed on test data, the model's detection of big fires (labeled as positives in our dataset) has 66% precision and 78% recall, which means the model catches most big fires and also retains good precision. The false negative rate (1 – recall) is 22.2%, which indicates that it would miss 22.2% of the big fires and classify them as small. The model also gives an out-of-sample AUC of 0.64.

In order to understand which variables play an important role in the prediction model, we compared the number of times that each variable appears in each tree over the entire boosting model. We observe that (1) Discovery Time, (2) Average Wind Speed, (3) Discovery DOY, (4) Wind Azimuth, and (5) Atm Moisture are the top 5 important features in our model. The inclusion of Discovery Time is plausible because if a fire is discovered in the morning, it is likely that it was burning all night and thus may become a big fire. Average Wind Speed and Atmospheric Moisture are also explainable as more wind and less humidity seem to be ideal conditions for the development of large fires. Wind Azimuth (direction) is likely important because of directional winds like the Santa Ana winds that are known for spreading fires [13]. Finally, Discovery Day of Year can be explained by the fact that more fires occur in the summer than in the winter. There are 10 weather variables that appear in the 15 most important features. Therefore, the utilization of weather variables in our prediction model was clearly beneficial. Weather features should be included in any future predictive analysis.

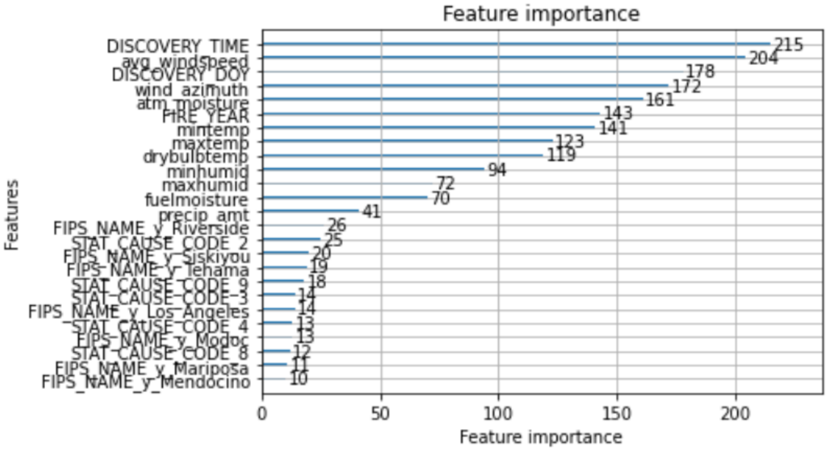


Figure 2 - Feature importance of boosted tree model

3. Analysis and Results: Survival analysis

In addition to predicting the size of a fire after it is discovered, we also perform a forward-looking analysis so that counties can prepare for fires before they happen. We choose to do this analysis at the county level because in California, individual counties control their own firefighting resources and can choose to sign contracts with statewide firefighting services if extra capacity is needed [6]. Survival analysis is ideal for predicting the risk of fires and the expected time until the next fire in a given county. We constructed our survival analysis dataset by splitting the data by county and calculating survival time to be number of days until the next fire happens in that county. We take the last fire's size and the county's weather data in the previous time interval to be the features for the time interval of interest. Full data details are in Appendix 5.

We use the Cox Proportional Hazard model to perform survival regression. The model was trained on data from 1993 to 2009 and tested on data from 2010 to 2015. The concordance index (an analog of AUC for survival models) on the test data is 0.64, which means the model is better than the baseline of assuming all entities have the same survival time [8]. With this model, we can make predictions for 1) the probability of fire on a day in a county and 2) the expected number of days until the next fire in that county. Examples of predictions for different input data values are in Appendix 5.

County decision makers can use this information to anticipate the probability of a fire happening in the next few days and allocate or contract appropriate firefighting resources. Counties can also use fire probabilities in combination with current fire risk measures like the Fire Potential Index (which uses vegetation greenness information) to advise residents on their behavior, like warning residents not to start campfires on days with high fire risk [7]. While it is difficult to numerically quantify the benefit that this information will provide, this analysis will provide valuable insights that can help county decision makers with the challenging problem of wildfire management.

4. Quantifying Impact

Using the predictions on our test set of whether a given fire develops into a big or small fire, we would like to quantify the financial impact this information can generate for the state of California. Take the perspective of a CA fire department manager who receives notice of a wildfire; they must decide how many firefighters and fire trucks to send to a fire site upon fire discovery. Let us consider the four possible scenarios that a fire department manager could encounter and their outcomes:

Table 3 – Wildfire scenarios

	Small fire	Big fire
Send resources for a small fire	The fire can be put out effectively.	The firemen have difficulties putting the fire out. This results in high costs, as the fire is able to spread and do more damage.
Send resources for a big fire	The additional resources sent to the fire site are wasted.	The fire can be put out effectively.

For the purpose of evaluating the financial impact of our modeling solution, we re-trained a gradient boosting model on fires from years 1993 to 2014 and tested the model on 2015 fires. We calculated a net cost of $1 million for the case when resources of a big fire were sent to a small fire (false positive), and a cost of $4 million for the case when resources of a small fire were sent to a big fire (false negative). Information on the calculations of these costs can be found in Appendix 7. The cost of the false negative aims to include not only the monetary cost, but also environmental and social costs of wildfires. In this context, we would like to keep the number of false negative cases as low as possible, while we allow for false positive cases to be more numerous.

Our estimations are broad approximations of false positive and false negative costs. In reality, the cost of sending unneeded resources to a fire would vary from county to county and would likely be based on fuel prices and labor costs in the time period of interest. The costs of missing a big fire would vary greatly based on the true fire size, since “big” fires vary from 2 – 300000 acres in our dataset. However, estimating these true costs is out of the scope of our project and would require more detailed study of the California wildfire management system. To assess the effectiveness of our approach, we compare the actions taken by fire department managers based on two different models. The first is the baseline that predicts all fires as small fires. The second is our gradient boosting model elaborated in section 2.

For the baseline model, additional costs are only incurred by the false negatives. The number of false negatives in the test set is exactly the number of big fires. The total additional cost for this model is:

$$\textit{number of FN} * \textit{cost of FN}$$

$$= 122 * \$4{,}000{,}000$$

$$= \$488{,}000{,}000$$

For our second model, the gradient boosting prediction model, the costs incurred by false negatives and false positives are the following:

$$\textit{number of FN} * \textit{cost of FN} + \textit{number of FP} * \textit{cost of FN}$$

$$= 19 * \$4{,}000{,}000 + 106 * \$1{,}000{,}000$$

$$= \$182{,}000{,}000$$

Purely from the perspective of whether to send more or fewer resources to a given fire, the model is able to save the fire department around $300 million. We realize that the assigned costs are very rough estimates, which is why the cost savings calculations will also be rough estimates. While we are unable to show a precise estimate of the monetary value we are generating through our model, this analysis more so shows the magnitude of the problem we are trying to solve. The potential savings generated through even a basic prediction model are magnified through the enormity of the damage that wildfires can create. The impact of the prediction model is therefore considerable and not only minimizes expenditure, but also helps prepare the fire departments to tackle big fires and mitigate their impact.

5. Conclusions and Future Work

There is clear evidence that making predictions about wildfire size and incidence based on weather data can provide significant benefits over a baseline model that treats all fires and counties the same. However, as with all natural phenomena, there are numerous additional factors that influence wildfire ignition and development, like vegetation data (which was missing from our dataset), and data on the fire-starting behavior of humans. It is notable that some of the top fire causes were human-related; arson, equipment use, debris burning, children, campfires, and smoking caused 37% of wildfires in California from 1992-2015. If we can incorporate information about human behavior and how it relates to fire incidence, we can recommend more immediate interventions to prevent fires from getting too big to handle. With more data, we will also be able to predict fire size and incidence more granularly and with higher accuracy, which is a task that will only increase in importance as wildfires continue to worsen.

Finally, the challenge of allocating resources to a fire is more complex than simply choosing how many firefighting units to send. Fire departments across the United States have extreme budget constraints and often do not have the financial means to appropriately respond to all fires. The next step for improving fire resource allocation would be to take these budget constraints into account to help fire departments maximize their very limited resources, which could be an application of prescriptive analytics. Analytics can help to mitigate the property destruction, health issues, and ecological damage that wildfires cause and help governments adapt to the changing global climate.

References

1. https://wildlife.ca.gov/Science-Institute/Wildfire-Impacts
2. https://www.epa.gov/air-research/wildland-fire-research-health-effects-research
3. http://efi.eng.uci.edu/papers/efg_212.pdf
4. https://fam.nwcg.gov/fam-web/weatherfirecd/fire_files.htm
5. https://www.kaggle.com/rtatman/188-million-us-wildfires
6. https://www.fire.ca.gov/about-us/
7. https://www.usgs.gov/ecosystems/lcsp/fire-danger-forecast/fire-potential-index-map
8. https://square.github.io/pysurvival/metrics/c_index.html
9. https://www.homeadvisor.com/r/average-yard-size-by-state/
10. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival6.html
11. https://www.nwcg.gov/sites/default/files/ibc-standard_rates.pdf
12. https://www.fire.ca.gov/media/10061/2015_redbook_final.pdf
13. https://www.sfchronicle.com/bayarea/article/Here-are-7-things-to-know-about-Santa-Ana-winds-12413155.php

Appendix 1: Data

Data source for **weather** observations: This data source gave us many individual fixed-width format files each containing daily weather data for an individual weather station, as well as a text file for each weather station containing information like latitude and longitude. To get the data in a usable format, we joined all weather station files vertically, identified weather station county using weather station latitude and longitude, grouped data by county and observation date, and finally aggregated weather data by taking averages for numeric weather columns and mode for categorical columns. We dropped columns that were empty or mostly empty, and then dropped all rows with missing values to get a complete weather dataset.

Features

Date: Date of weather measurement

DryBulbTemp: Air temperature measured by a thermometer freely exposed to the air. Fahrenheit.

Atm_moisture: Moisture in the air (Moisture type units).

Wind_azimuth: Wind direction in degrees (0-360).

Avg_windspeed: Average magnitude of wind speed.

Fuel_moisture: Fuel moisture.

Max_temp: Maximum temperature in Fahrenheit.

Min_temp: Minimum temperature in Fahrenheit.

Max_humid: Maximum relative humidity in the air in percentage.

Min_humid: Minimum relative humidity in the air in percentage.

Precip_duration: Precipitation duration in hours.

Precip_amt: Precipitation amount in inches.

Moisture_type: Classification for moisture type (1=Wet bulb temperature, 2=Relative Humidity, 3=Dewpoint).

Meas_type: Classification for measurement units type (everything is in US units).

County: County where the weather data comes from.

Wetflag: Boolean indicator for incidence of precipitation.

Snowflag: Boolean indicator for incidence of snowflakes precipitation.

FIPS_name_x: County name for the observation.

Data source for **fire** observations: NWCG dataset sourced from Kaggle

Data cleaning performed: from the original 1.88 million fires, we selected only California fires from the sqlite database, which resulted in 189,000+ rows, each representing one fire. The columns are below:

FIPS_code: County code.

FOD_id: Unique id number for fire occurrence.

Fire_year: Year of the fire incidence.

DATE: Date of discovery.

MONTH: Month of discovery.

DISCOVERY_DOY: Day of year when the fire was discovered.

DISCOVERY_TIME: Time of day when the fire was discovered.

STAT_CAUSE_CODE: Code for fire cause.

STAT_CAUSE_DESCR: Fire cause description (one of

CONTDATE: Date when fire got contained.

CONT_DOY: Day of year when fire got contained.

CONT_TIME: Time of day when the fire was contained.

FIRE_SIZE: Estimate of acres within the final perimeter of the fire.

FIRE_SIZE_CLASS: Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).

LATITUDE: Latitude of the fire.

LONGITUDE: Longitude of the fire.

FIPS_name_y: County name for the observation.

We took cleaned data from both sources and joined the files together on weather observation date = fire discovery date, and weather county = fire county. Due to significant missing data, this joining process dropped over 80% of the original fire observations and we are left with a 30,000+ row dataset. We have fires from the year 1997 to the year 2015. In the merging process, we went from 58 counties to 36 counties, as we had to drop all the counties with missing weather data. Because of this, we may have to proceed with caution when generalizing model results to counties other than the ones we have in our dataset.

One issue in the data is that we do not know exactly what interventions are taken on different fires, which could potentially impact the eventual size of a fire. For this problem, we will assume that we do not need to know interventions taken, although in a real situation we would need to consider this.

In addition, we expect significant multicollinearity between weather columns, so we will need to consider how to deal with this in our modeling. We also have one extreme outlier fire that burned over 300,000 acres, double the next largest fire size value, which we may need to remove from our dataset.

Appendix 2: Exploratory Visualizations

1. A map of fires in our dataset colored by fire size. It shows that some counties had more frequent and intensive fires, and that some counties are missing from our dataset. The fires that we have data on are concentrated in the north, middle and south area of California.

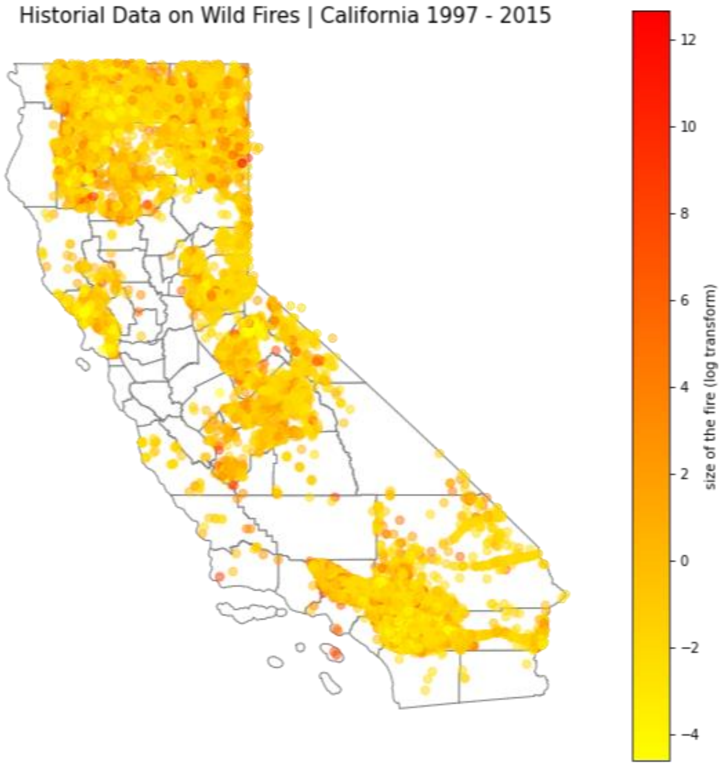

Historial Data on Wild Fires | California 1997 - 2015
12
10
8
6
4
2
0
−2
−4
size of the fire (log transform)

Figure A2.1 - Log Fire Size in California

2. A plot of yearly fire count and total acres burned over time. The factors follow each other closely, which is expected. However, we can pick out some years (2008, 2014) that have big increases in acres burned from the previous year but actually had fewer total fires, meaning that the fires that occurred were especially large. Clearly, fire size and count have both been increasing over the years, which adds urgency and importance to the problem.

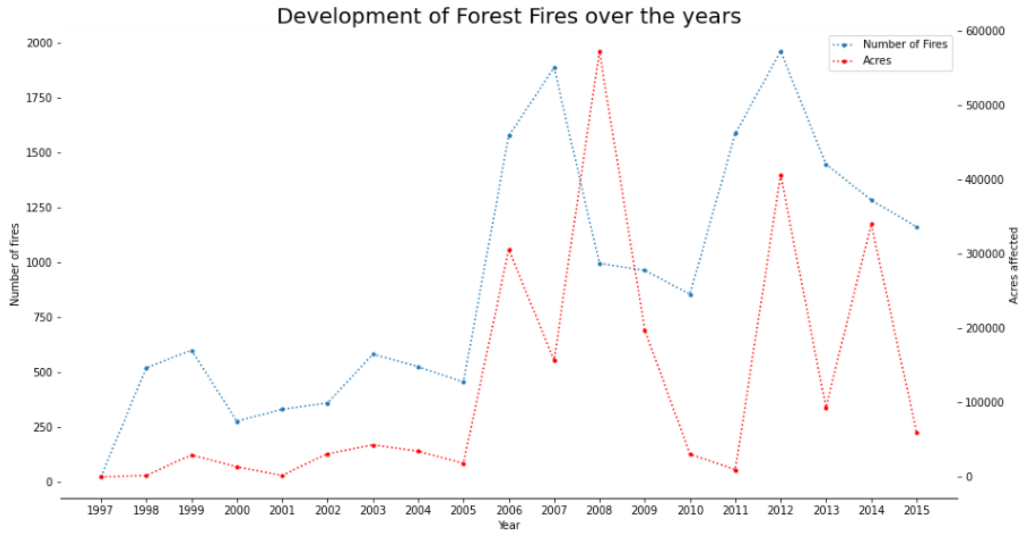


Figure A2.2 - Development of Forest Fires over the Years

3. From the plots below, we see log(fire size) on the y axis plotted against different weather variables on the x axis. We can see that temperature and fuel moisture have visible relationships with fire size, and the relationship is less clear for atmospheric moisture and wind speed. Overall, the data is very noisy, which will present challenges in the modelling process.

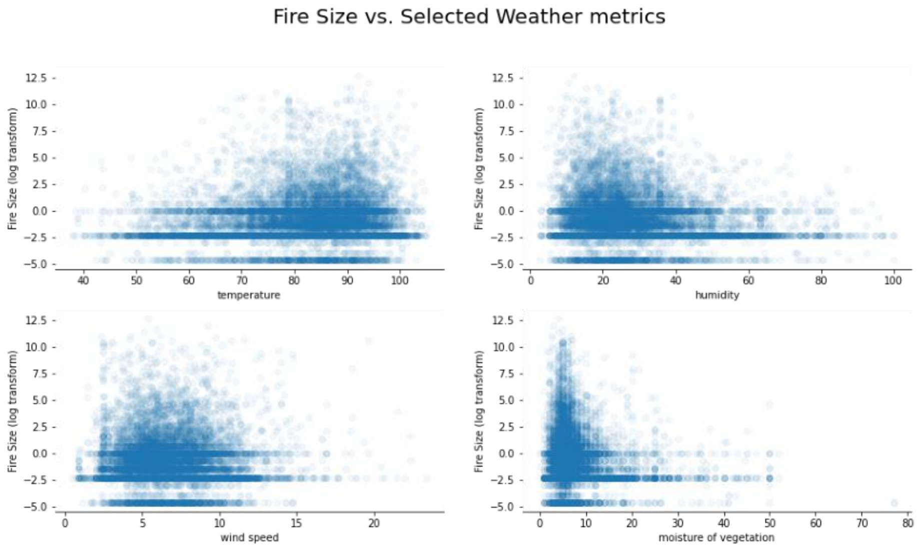


Figure A2.3 - Weather Variables vs Log Fire Size

4. Correlation matrix for all weather and fire features. It helps us to identify which features have high correlation that we need to pay attention to when doing feature engineering and selection. Due to the nature of weather data, there are many correlated factors like humidity and precipitation. There are also correlations among temperature measurements, and between latitude and wind speed.

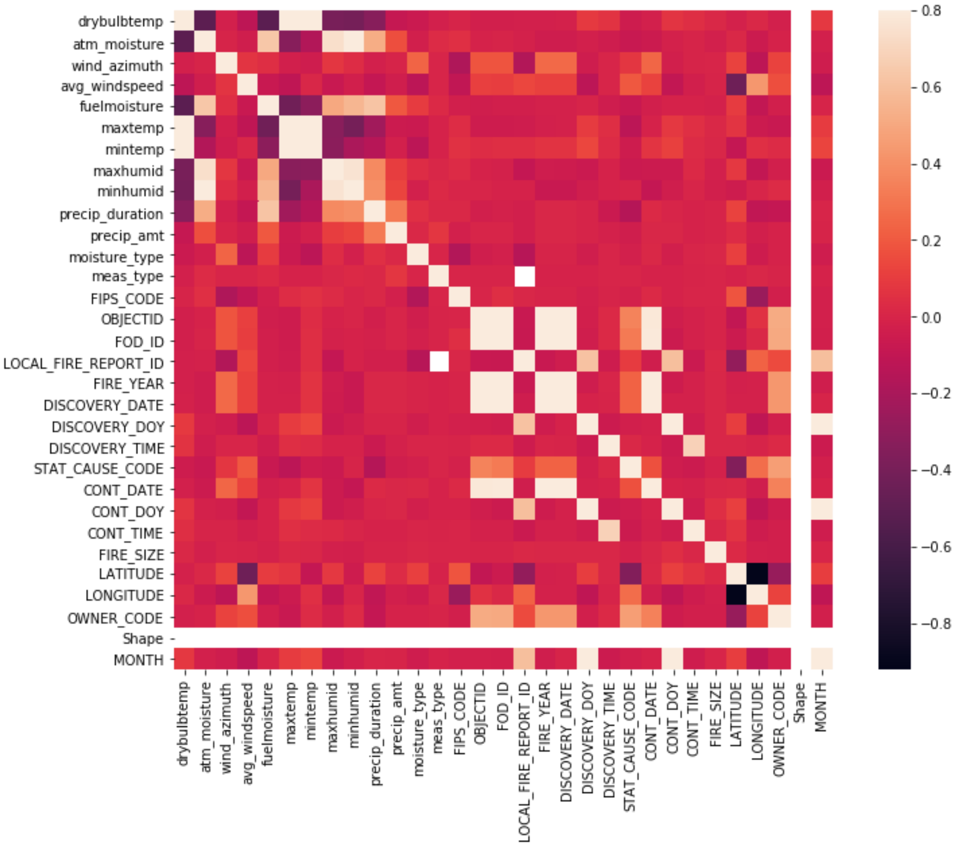


Figure A2.4 - Weather and Fire Data Correlation Matrix

5. The following graph shows average days until containment for each fire size class (A is the smallest class and G is the largest). We can see that fires in classes A and B take around one day or fewer to extinguish, while fires in class G can burn for a month on average and cause excessive damage/

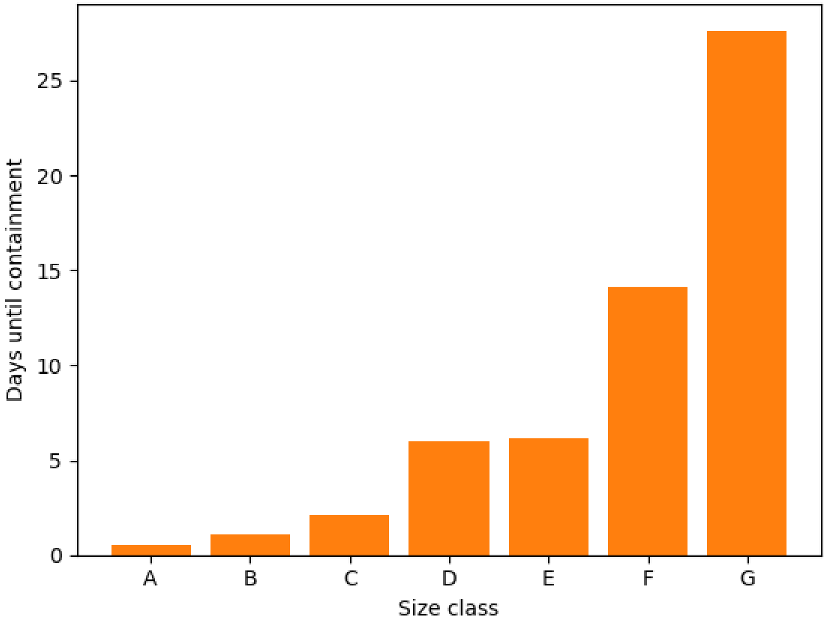


Figure A2.5 - Days until containment by Fire Size Class

6. The following table shows the distribution of fire size classes in our data. Extremely big fires are rare occurrences whereas extremely small fires are very common.

Table A2.1 - Proportion of Fires by Fire Size Class.

Class	Proportion
A	63.3122%
B	29.3620%
C	4.4370%
D	1.1446%
E	0.7823%
F	0.5930%
G	0.3688%

Appendix 3: Regression for fire size prediction results

Table A3.1 - Regression results for fire size prediction

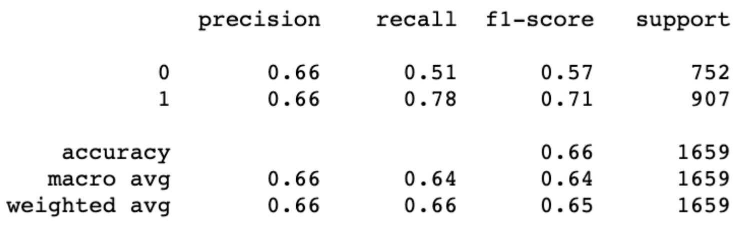
Model	Out-of-sample R^2
Ordinary least-squares linear regression	0.045
Lasso regression (w/ cross-validation)	0.046
Ridge regression (w/ cross-validation)	0.043

Appendix 4: Boosting model information

Table A4.1 - LightGBM hyperparameter settings

Parameters	Best value found through cross-validation
objective	binary
metric	binary_logloss, auc
learning_rate	0.1
max_depth	15
num_leaves	40
feature_fraction	0.8
min_child_samples	21
min_child_weight	0.001
bagging_fraction	0.9
bagging_freq	3
lambda_l1	0.6
lambda_l2	40
cat_smooth	0
num_iterations (number of trees)	200

Table A4.2 - Classification Report for Boosting model



```
              precision    recall  f1-score   support

           0       0.66      0.51      0.57       752
           1       0.66      0.78      0.71       907

    accuracy                           0.66      1659
   macro avg       0.66      0.64      0.64      1659
weighted avg       0.66      0.66      0.65      1659
```

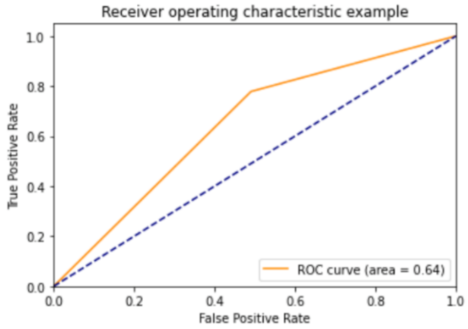


Figure A4.1 - ROC curve for Boosting model

Appendix 5: CART Model Results

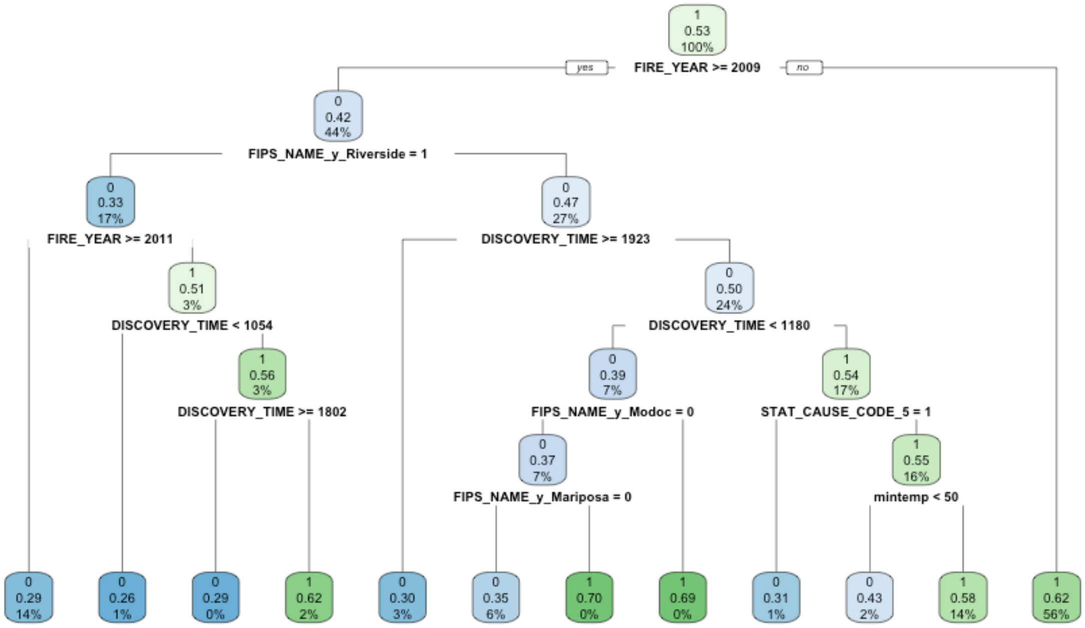


Figure A5.1 - CART Tree

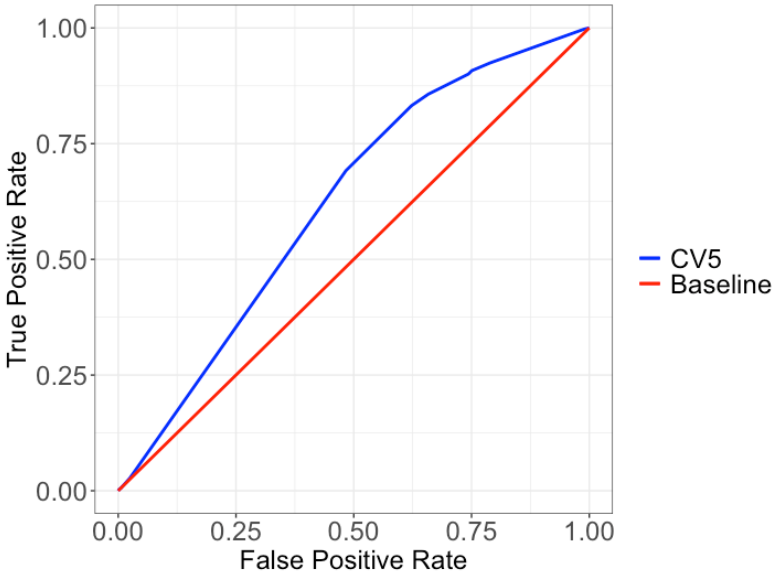


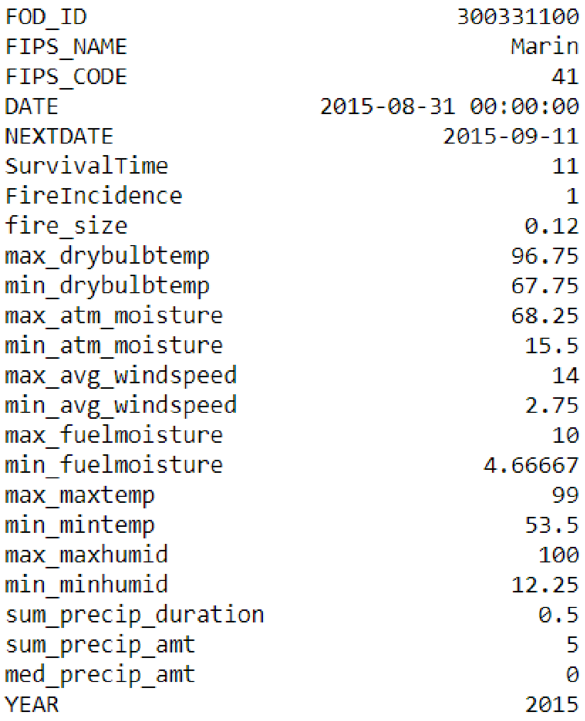
Figure A5.2 - ROC Curve for CART model

Appendix 6: Survival

Creating survival data

In Marin county, there was a fire on 8/31/2015. The most recent past fire in Marin county happened on 7/29/2015, and the next fire happened in 11 days, on 9/11/2015. The survival time for the 8/31/2015 Marin data point is 11. The weather features for the 8/31/2015 Marin data point are taken from the 7/30/2015 to 8/31/2015 time interval, summarized by taking the maximum, minimum, median, or sum on the time interval. Another feature will be the size of the 7/29/2015 fire. This is what this row of data would look like:

Table A6.1 - Data description for survival analysis.



```
FOD_ID                            300331100
FIPS_NAME                             Marin
FIPS_CODE                                41
DATE                    2015-08-31 00:00:00
NEXTDATE                         2015-09-11
SurvivalTime                             11
FireIncidence                             1
fire_size                              0.12
max_drybulbtemp                       96.75
min_drybulbtemp                       67.75
max_atm_moisture                      68.25
min_atm_moisture                       15.5
max_avg_windspeed                        14
min_avg_windspeed                      2.75
max_fuelmoisture                         10
min_fuelmoisture                    4.66667
max_maxtemp                              99
min_mintemp                            53.5
max_maxhumid                            100
min_minhumid                          12.25
sum_precip_duration                     0.5
sum_precip_amt                            5
med_precip_amt                            0
YEAR                                   2015
```

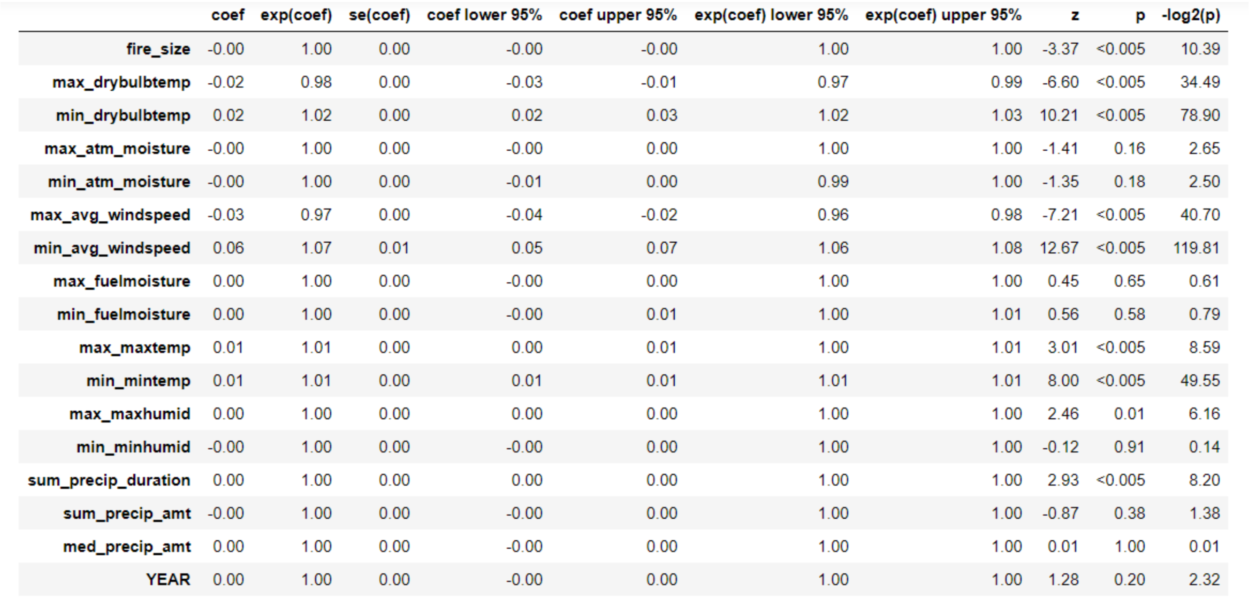
This process was applied to all fires in all counties for which there was sufficient weather data, dropping the first fire in each county because there is no past weather data and dropping the last fire in each county because it is a censored observation. The FIPS_NAME, FOD_ID, FIPS_CODE, and full date columns were dropped from the data before training the model.

We obtained a training dataset containing data from years 1993-2009 with 20867 rows, and a testing dataset containing data from years 2009-2015 with 5216 rows (80-20 training/testing split).

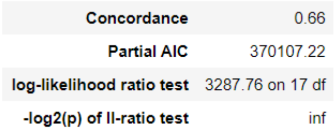
Model fitting

The results of the model are below:

Table A6.2 - Survival analysis results.



	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
fire_size	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	-3.37	<0.005	10.39
max_drybulbtemp	-0.02	0.98	0.00	-0.03	-0.01	0.97	0.99	-6.60	<0.005	34.49
min_drybulbtemp	0.02	1.02	0.00	0.02	0.03	1.02	1.03	10.21	<0.005	78.90
max_atm_moisture	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-1.41	0.16	2.65
min_atm_moisture	-0.00	1.00	0.00	-0.01	0.00	0.99	1.00	-1.35	0.18	2.50
max_avg_windspeed	-0.03	0.97	0.00	-0.04	-0.02	0.96	0.98	-7.21	<0.005	40.70
min_avg_windspeed	0.06	1.07	0.01	0.05	0.07	1.06	1.08	12.67	<0.005	119.81
max_fuelmoisture	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.45	0.65	0.61
min_fuelmoisture	0.00	1.00	0.00	-0.00	0.01	1.00	1.01	0.56	0.58	0.79
max_maxtemp	0.01	1.01	0.00	0.00	0.01	1.00	1.01	3.01	<0.005	8.59
min_mintemp	0.01	1.01	0.00	0.01	0.01	1.01	1.01	8.00	<0.005	49.55
max_maxhumid	0.00	1.00	0.00	0.00	0.00	1.00	1.00	2.46	0.01	6.16
min_minhumid	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.12	0.91	0.14
sum_precip_duration	0.00	1.00	0.00	0.00	0.00	1.00	1.00	2.93	<0.005	8.20
sum_precip_amt	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.87	0.38	1.38
med_precip_amt	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.01	1.00	0.01
YEAR	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	1.28	0.20	2.32



Concordance	0.66
Partial AIC	370107.22
log-likelihood ratio test	3287.76 on 17 df
-log2(p) of ll-ratio test	inf

The concordance index on the training data is 0.66 and the index on the testing data is 0.649, indicating that the model is better than the baseline estimation (which would have a concordance index of 0.5). Here is a plot of the coefficients (the log hazard ratios) for each feature and its confidence interval:

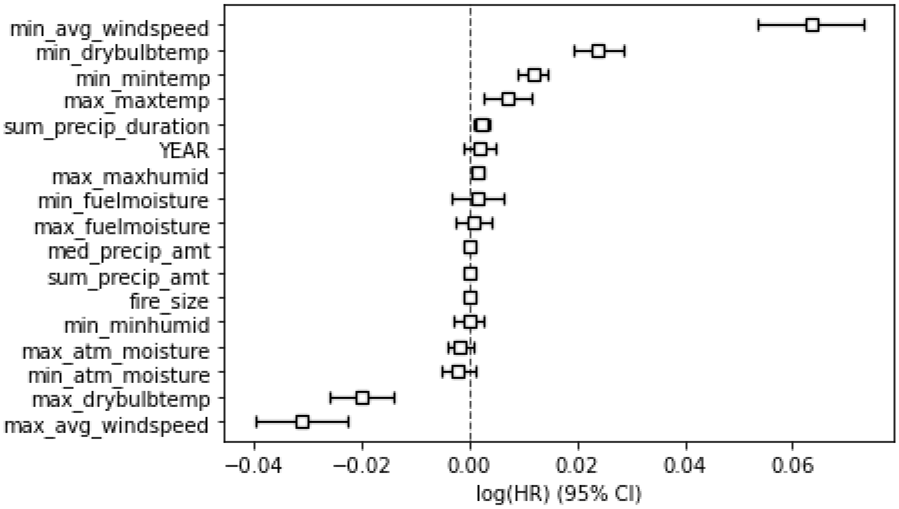


Figure A6.1 - Confidence intervals for survival analysis coefficients.

Interpretations of significant coefficients: “If the hazard ratio for a predictor is close to 1 then that predictor does not affect survival. If the hazard ratio is less than 1, then the predictor is protective (i.e., associated with improved survival) and if the hazard ratio is greater than 1, then the predictor is associated with increased risk (or decreased survival)”, from source 10 in the references section. The logs of the hazard ratios are plotted here, so we will be looking at whether the log hazard ratios are positive or negative.

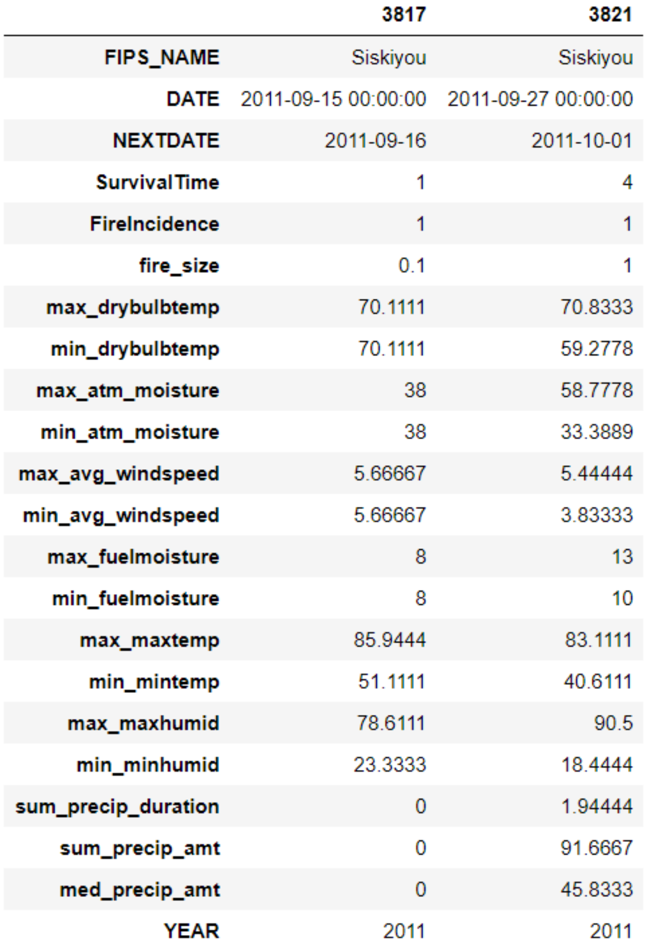
Minimum average windspeed, minimum dry bulb temperature, minimum mintemp, and maximum maxtemp have positive coefficients, which means that high values of these features are associated with low survival time. Low values of these features are associated with high survival time.

Maximum dry bulb temperature and maximum average windspeed have negative coefficients. The interpretation for this is slightly unclear; one explanation could be that maximum windspeeds tend to only occur for short durations of time, so a high maximum windspeed does not necessarily mean that an entire week was very windy. This is likely why high minimum average windspeed is predictive of low survival time.

Survival predictions

We will illustrate the prediction-making process for these two data rows (the true survival times will be removed before predictions are made):

Table A6.3 - Survival analysis predictions.



	3817	3821
FIPS_NAME	Siskiyou	Siskiyou
DATE	2011-09-15 00:00:00	2011-09-27 00:00:00
NEXTDATE	2011-09-16	2011-10-01
SurvivalTime	1	4
FireIncidence	1	1
fire_size	0.1	1
max_drybulbtemp	70.1111	70.8333
min_drybulbtemp	70.1111	59.2778
max_atm_moisture	38	58.7778
min_atm_moisture	38	33.3889
max_avg_windspeed	5.66667	5.44444
min_avg_windspeed	5.66667	3.83333
max_fuelmoisture	8	13
min_fuelmoisture	8	10
max_maxtemp	85.9444	83.1111
min_mintemp	51.1111	40.6111
max_maxhumid	78.6111	90.5
min_minhumid	23.3333	18.4444
sum_precip_duration	0	1.94444
sum_precip_amt	0	91.6667
med_precip_amt	0	45.8333
YEAR	2011	2011

The predicted survival functions for each of these rows:

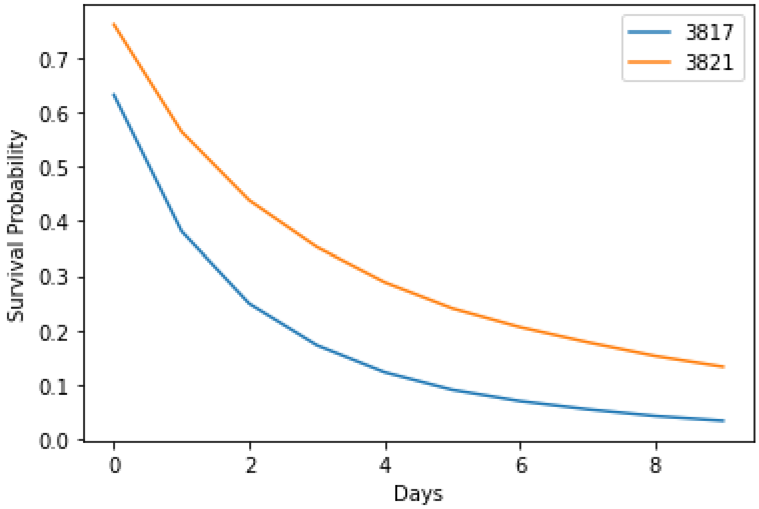
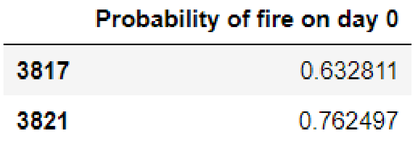


Figure A6.2 - Survival Curve.

From the plot, we can see that data point 3817 has a lower predicted rate of survival over 10 days compared to data point 3821. The expected number of days until next fire for data point 3817 is 1.788 and the expected number of days until next fire for data point 3821 is 5.361, which matches up with the real survival times for these points. These survival differences are likely because data point 3821 has lower minimum temperatures and windspeeds, higher maximum humidity, and a lot of precipitation.

County decision makers could use predictions of the survival function to come up with probabilities that there will not be a fire on the current day or the next day:

Table A6.4 – Predicted probabilities for a fire in the next day.



	Probability of fire on day 0
3817	0.632811
3821	0.762497

Based on this probability, they could advise residents on their behavior and prepare for any fires by allocating resources beforehand. For example, on 9/27/2011 in Siskiyou county, county officials could put a restriction on campfire burning in the county to prevent a campfire from turning into a wildfire. This survival analysis gives decision makers a forward-looking tool to complement our fire size prediction modeling, which makes predictions after a fire has already happened.

Appendix 7: Cost Estimation

True positive cost = Fire is predicted to be big and it is actually big = big fire firefighting costs + $0 damage	False positive cost = Fire is predicted to be big and it is actually small = big fire firefighting costs + opportunity cost
False negative cost = Fire is predicted as small and it is actually big = small fire firefighting costs + damage costs	True negative cost = Fire is predicted as small and it is actually small = small fire firefighting costs + $0 damage

From the NWCG's standard cost component documentation in 2016, the cost of using one unit each of all available firefighting crews and equipment (including trucks, ambulances, debris clearing equipment, helicopters, and support personnel for firefighters) for one day amounts to about $250,000 [11]. For the purposes of this project only, we assume that a small fire will require 1 unit of all available firefighting crews and equipment, and a big fire will require 4 units of all available firefighting crews and equipment on average. Therefore, the treatment for a fire predicted to be small would cost $250,000 and the treatment of a fire predicted to be big would cost $1M. These are conservative estimates since they are daily costs, but large fires will likely require many days of firefighting to control.

Cost estimation false negatives:

- To get true net false negative cost, we would normally take the cost of improperly treating big fires (big fires receiving small fires' resources) and subtract the cost of properly treating big fires (big fires receiving big fires' resources). However, we do not see a straightforward way of figuring out which fires were treated properly and which were treated improperly from historical data. We must make a significant assumption here that properly treated big fires and properly treated small fires incur *no* damage costs. This assumption does not hold true in reality, since fires that are properly intervened on can still cause damage.
- We need to quantify the additional damage costs incurred by big fires that are treated with small fires' resources initially.
- CA historical fire data indicates that the total cost of wildfire-related property damage was estimated to be $3.061 billion in 2015 [12].
- Technically, this figure contains damage amounts for big and small fires. However, small fires cause a very small amount of damage compared to the biggest wildfires. Therefore, we make the assumption for the purposes of this analysis that most of the $3.061 billion damage is generated by big fires.
- We use $3 billion as a conservative estimate of the *total* impact of big fires. The $3 billion figure does not include environmental and health costs associated with fires.
- In addition, we are estimating that $3 billion would be the total damage if appropriate interventions were not taken for *all* big fires, but at least some appropriate interventions were

likely taken. Therefore, we believe that $3 billion is a reasonable estimate for the total impact of improperly treated big fires in 1 year.

- Since we lost about 83% of wildfire observations while merging weather and wildfire data, we are only considering about 1/6 of all California wildfires.
- Our test set, comprised of fires in 2015, has 122 fires that were classified as big fires.
- We therefore estimated the damage caused by a big wildfire initially treated with a small fire's resources to be around $3 billion/ (6*122), which is roughly $4 million per fire.

Cost estimation false positives:

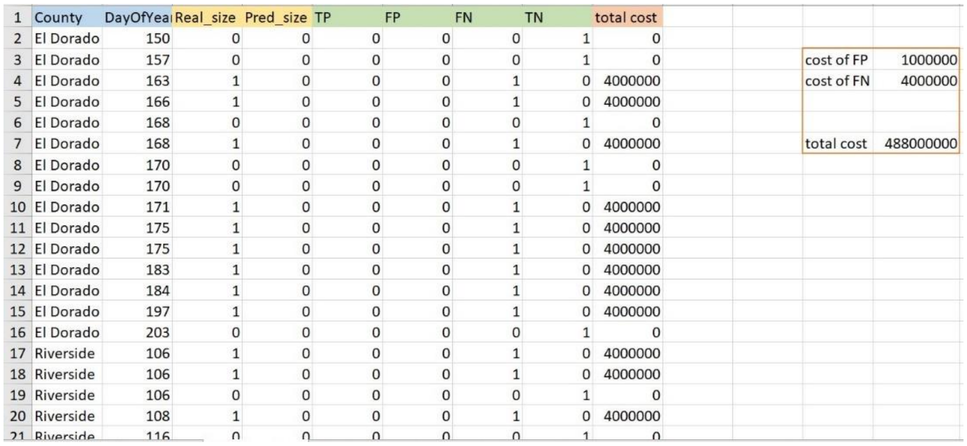
- From our resource cost calculations above, we estimate that it costs $1 million to pay for the resources required by a big fire.
- If the fire we are allocating resources for is actually a small fire, we also incur an opportunity cost, as those resources are wasted and cannot be re-allocated quickly to an actual big fire. We estimate this opportunity cost as $1M as well.

Summary of costs:

True positive cost = $1M + $0 damage	False positive cost = $1M + $1M opportunity cost
False negative cost = $250k + $4M damage	True negative cost = $250k + $0 damage

Summary of net costs:

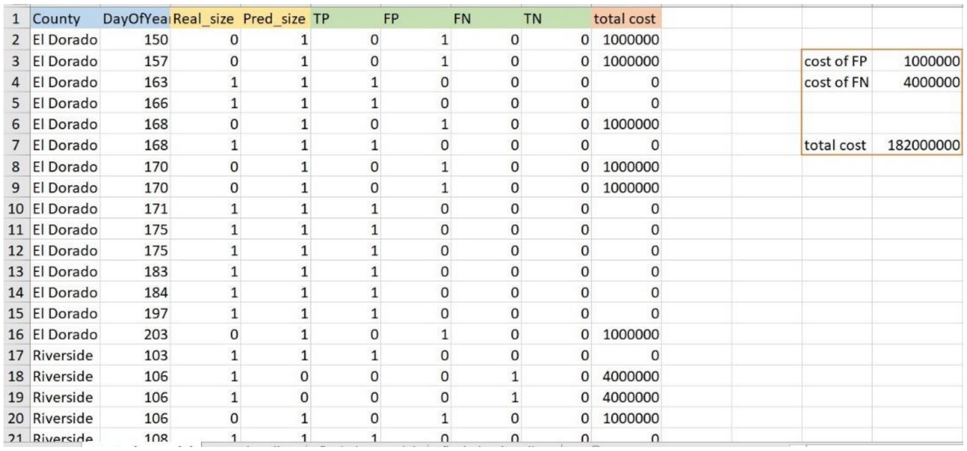
True positive net cost $0	False positive net cost $1 million
False negative net cost $4 million	True negative net cost $0



1	County	DayOfYea	Real_size	Pred_size	TP	FP	FN	TN	total cost		
2	El Dorado	150	0	0	0	0	0	1	0		
3	El Dorado	157	0	0	0	0	0	1	0	cost of FP	1000000
4	El Dorado	163	1	0	0	0	1	0	4000000	cost of FN	4000000
5	El Dorado	166	1	0	0	0	1	0	4000000		
6	El Dorado	168	0	0	0	0	0	1	0		
7	El Dorado	168	1	0	0	0	1	0	4000000	total cost	488000000
8	El Dorado	170	0	0	0	0	0	1	0		
9	El Dorado	170	0	0	0	0	0	1	0		
10	El Dorado	171	1	0	0	0	1	0	4000000		
11	El Dorado	175	1	0	0	0	1	0	4000000		
12	El Dorado	175	1	0	0	0	1	0	4000000		
13	El Dorado	183	1	0	0	0	1	0	4000000		
14	El Dorado	184	1	0	0	0	1	0	4000000		
15	El Dorado	197	1	0	0	0	1	0	4000000		
16	El Dorado	203	0	0	0	0	0	1	0		
17	Riverside	106	1	0	0	0	1	0	4000000		
18	Riverside	106	1	0	0	0	1	0	4000000		
19	Riverside	106	0	0	0	0	0	1	0		
20	Riverside	108	1	0	0	0	1	0	4000000		
21	Riverside	116	0	0	0	0	0	1	0		

Figure A7.1 - Cost estimation for baseline model.

Each row represents a fire. The table contrasts real size vs. predicted size, indicates whether a data point is a TP, FP, FN, or TN, and shows the associated cost with each case.



1	County	DayOfYea	Real_size	Pred_size	TP	FP	FN	TN	total cost		
2	El Dorado	150	0	1	0	1	0	0	1000000		
3	El Dorado	157	0	1	0	1	0	0	1000000	cost of FP	1000000
4	El Dorado	163	1	1	1	0	0	0	0	cost of FN	4000000
5	El Dorado	166	1	1	1	0	0	0	0		
6	El Dorado	168	0	1	0	1	0	0	1000000		
7	El Dorado	168	1	1	1	0	0	0	0	total cost	182000000
8	El Dorado	170	0	1	0	1	0	0	1000000		
9	El Dorado	170	0	1	0	1	0	0	1000000		
10	El Dorado	171	1	1	1	0	0	0	0		
11	El Dorado	175	1	1	1	0	0	0	0		
12	El Dorado	175	1	1	1	0	0	0	0		
13	El Dorado	183	1	1	1	0	0	0	0		
14	El Dorado	184	1	1	1	0	0	0	0		
15	El Dorado	197	1	1	1	0	0	0	0		
16	El Dorado	203	0	1	0	1	0	0	1000000		
17	Riverside	103	1	1	1	0	0	0	0		
18	Riverside	106	1	0	0	0	1	0	4000000		
19	Riverside	106	1	0	0	0	1	0	4000000		
20	Riverside	106	0	1	0	1	0	0	1000000		
21	Riverside	108	1	1	1	0	0	0	0		

Figure A7.2 - Cost estimation for GBM.