Data Science Final Project: Predicting Fire Risk in California

Anna Graves, Joshua Graves, & Emma Johnson 12/20/2019

Introduction and Motivation

In the last few years, stories of California's wildfires have dominated the news. Hellish photos and videos of flames racing across hills and highways captivated the nation's attention as its most populous state burned. Fires can cost billions of dollars directly and indirectly such as through the destruction of property, spending of tax dollars, and loss of human life. As such, predicting if a fire might start based on the conditions is incredibly important.

Main Objectives

1. What variables might influence the risk of fire?
2. Can we develop a model that can predict fire risk accurately by combining fire history, climate data (rain, temperature, etc.), and PG&E presence?
3. Does our model have different predictions about the fire risk by county?

4. Does the addition of PG&E presence help predict future fire risk?

Data Approach

In our model, we use a series of climate and population data to predict the presence of a fire. We began by compiling three separate datasets to construct our model: the history of fires in California from January 1, 2017 to December 31, 2018, daily climate data from all California counties from January 1, 2017 until December 13, 2019, and population data by county from the 2010 census. We used QGIS, a geographic information system software, to merge all the historical fires by county and extract the data. We then joined these three data sets by county to create a master data set and selected our predictor variables. Finally, we generated three additional variables that we believe might `increase fire risk.

First, we created a PG&E dummy variable "PGE" which tells us whether PG&E has a presence of at least 50% in a county (a binary yes/no variable). This is because if electrical lines are damaged, sparks can cause fires in the right conditions. PG&E is accused of not properly maintaining its equipment, which in turn caused California's most lethal fire, the Camp Fire (Ailworth and Brickley, 2019). On December 6, 2019, PG&E agreed to a $13.5 billion settlement with the victims of the fire (Ailworth and Brickley, 2019). This implies that PG&E must be somewhat responsible for causing these fires. Therefore, the presence of their services might be an important indicator of fire risk.

The next categorical variable we created was "fire threat," which assigns a threat level to each county based on the percentage of houses in that county at risk of fire. This classification is similar to a methodology we found from Verisk Analytics which used three factors to determine risk:

1. Proximity to forests, shrubs, and trees
2. Proximity to hilly or mountainous terrain
3. Degree of isolation

We used the percentages by county and sorted them into three different classes. The table below illustrates the classification of fire threat according to this methodology.

Classification	% of houses at fire risk
Low	0% to 20%
Medium	21% to 50%
High	51% +

Last, we created the binary variable "fire" that indicates whether or not there was a fire present in a county by month, which we use as our dependent variable.

The following two graphs below illustrate the characteristics of the past two years of fires in the state of California. Figure 1 displays the number of fires according to their cause. Interestingly, the main cause of fire is California remains unknown, which sparked our interest in how to predict fire risk. Additionally, 157 of the fires were started from equipment and power lines, which could be directly attributed to the negligence of power and electric companies in California like PG&E and Southern California Edison (SCE).

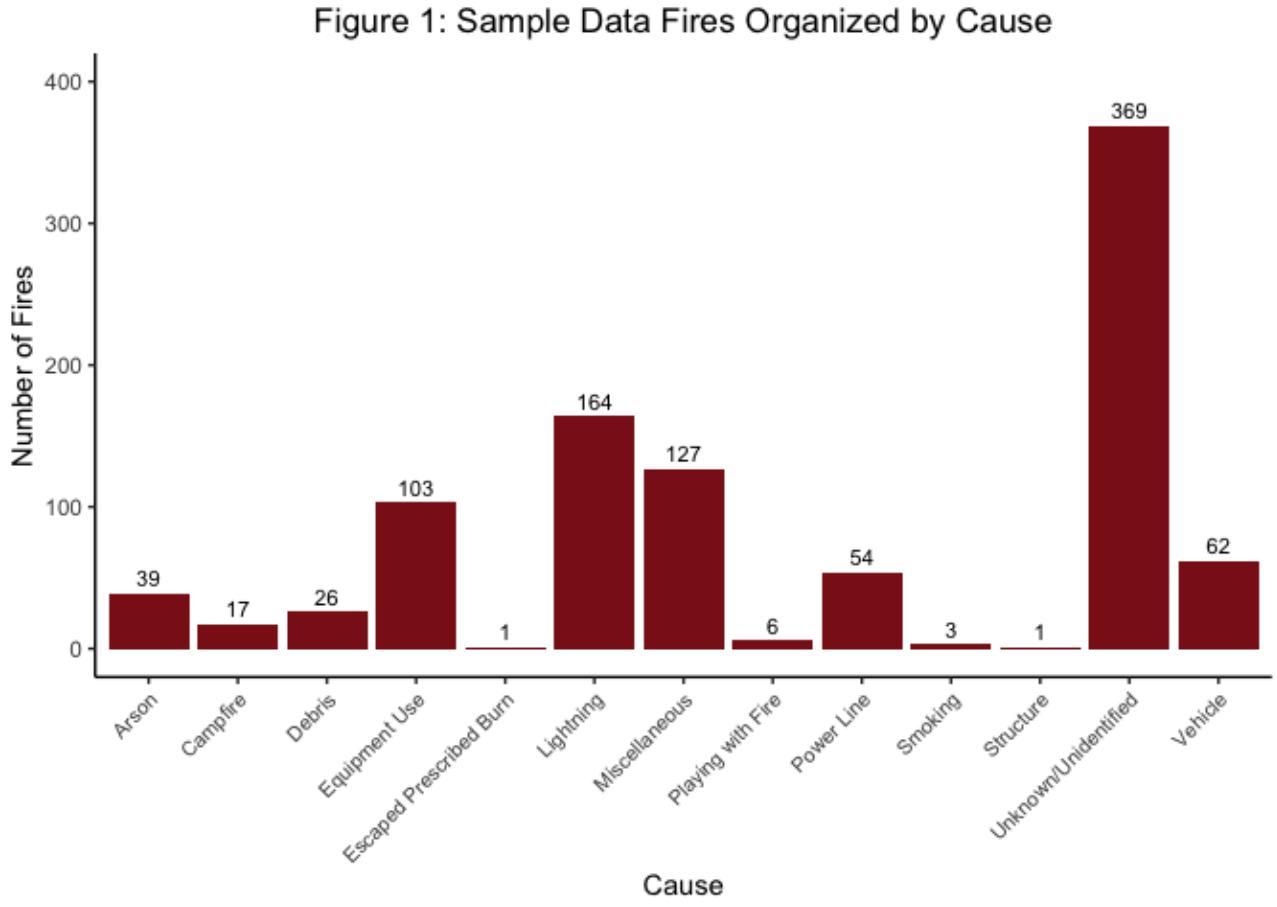
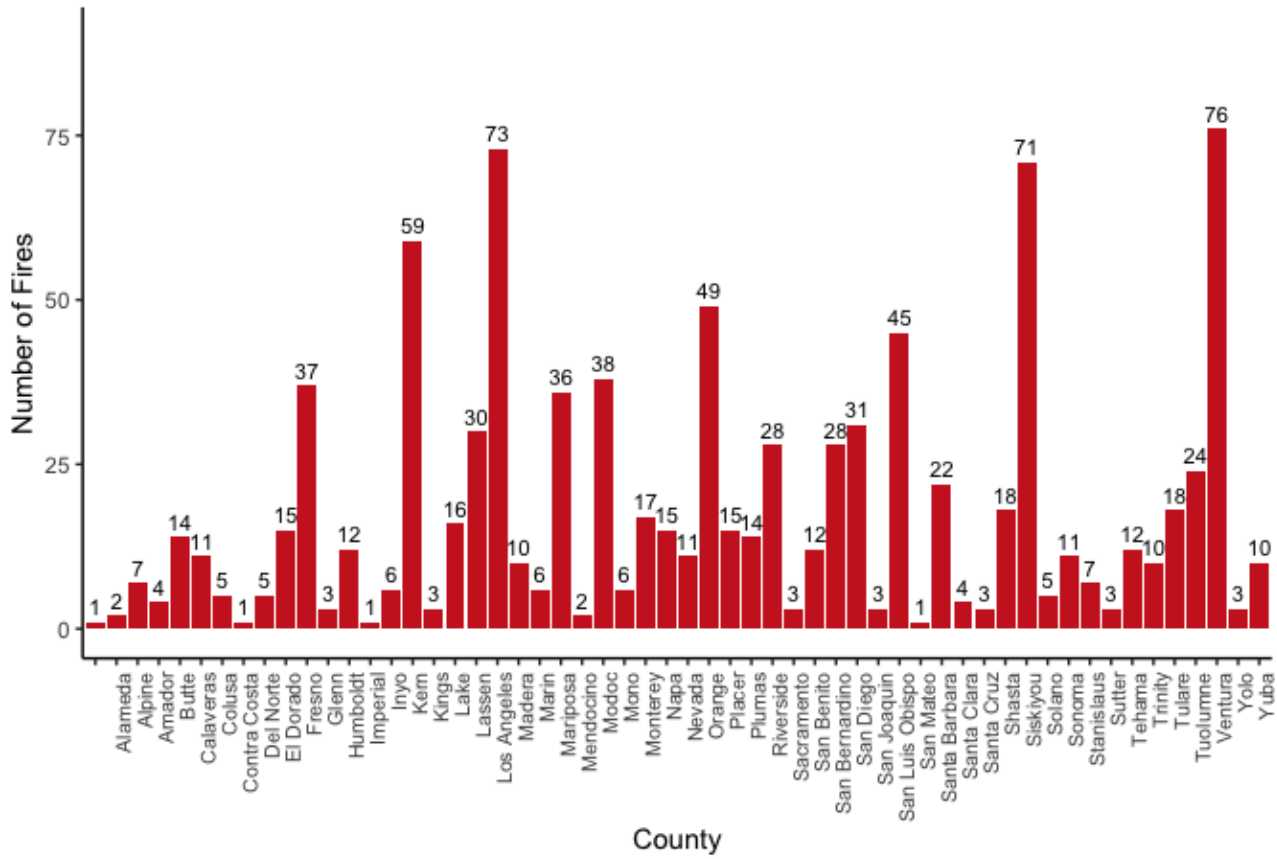


Figure 2 shows the number of fires by county over the past two years. This graph gives us a good indication of what counties in "The Golden State" have the most fires. Ventura County had the highest number of fires at 76, while Los Angeles County claimed the second highest at 73 fires. On the other hand, Figure 2 conveys which counties had the most severe fires. The most severe fires in 2017-2018 were the Ranch, Thomas, Carr, and Camp Fires in Colusa County, Santa Barbara County, Shasta County, and Butte County, respectively. These fires blazed between 150,000 and 400,000 acres and caused incredible harm to those living in those areas and to the environment.

Figure 2: Fires by California County (January 1, 2017 - December 31, 2018)

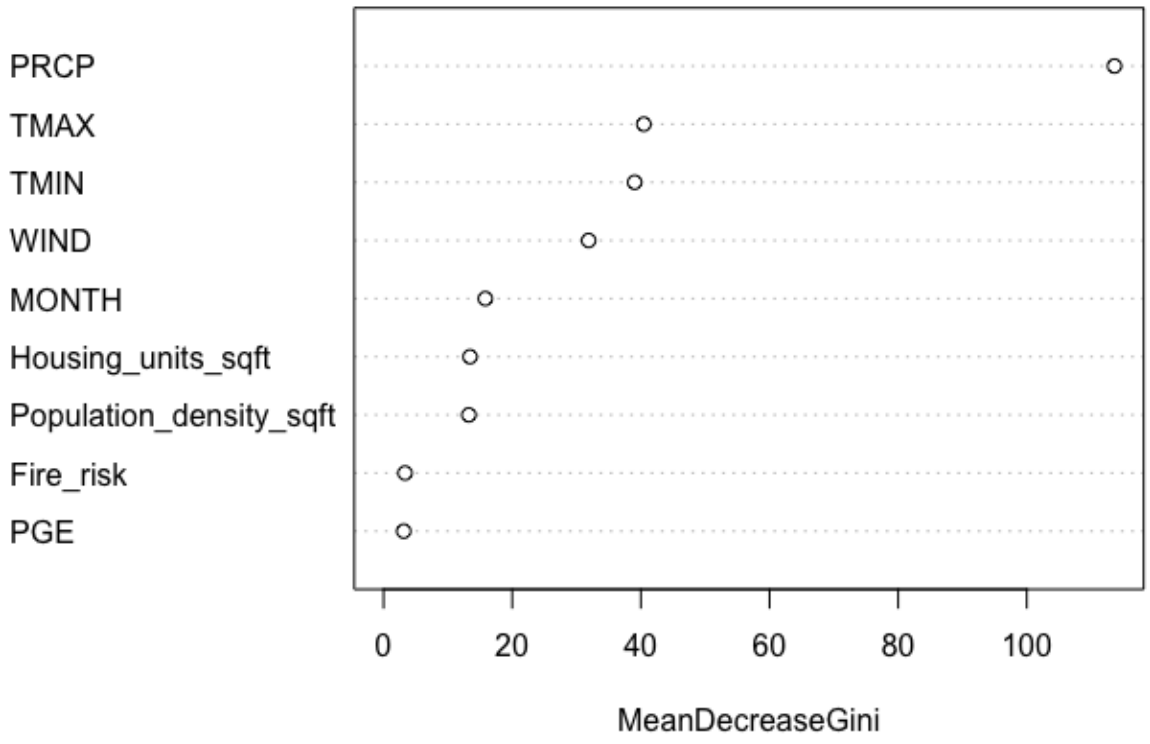


The final data set includes predictor variables by county: maximum temperature, minimum temperature, precipitation, wind, PG&E coverage, population density per square mile, housing density per square mile, and fire risk by county. Our dependent variable is a dummy variable indicating “fire” or “nofire.” We then took the monthly averages of each variable to consolidate the data set and make county level monthly data more accurate. The final train dataset included 644 observations, drastically smaller than our original daily dataset which included daily data for 58 counties over two years.

Analysis

In order to predict the presence of a future fire, we used the following random forest model: `train_rforest <- randomForest(fire ~ Fire_risk + MONTH + TMAX + TMIN + PRCP + WIND + Population_density_sqft + Housing_units_sqft + PGE, data=train, ntree=500, mtry=3)` . We first used variable importance function on this random forest model, which yielded the following results:

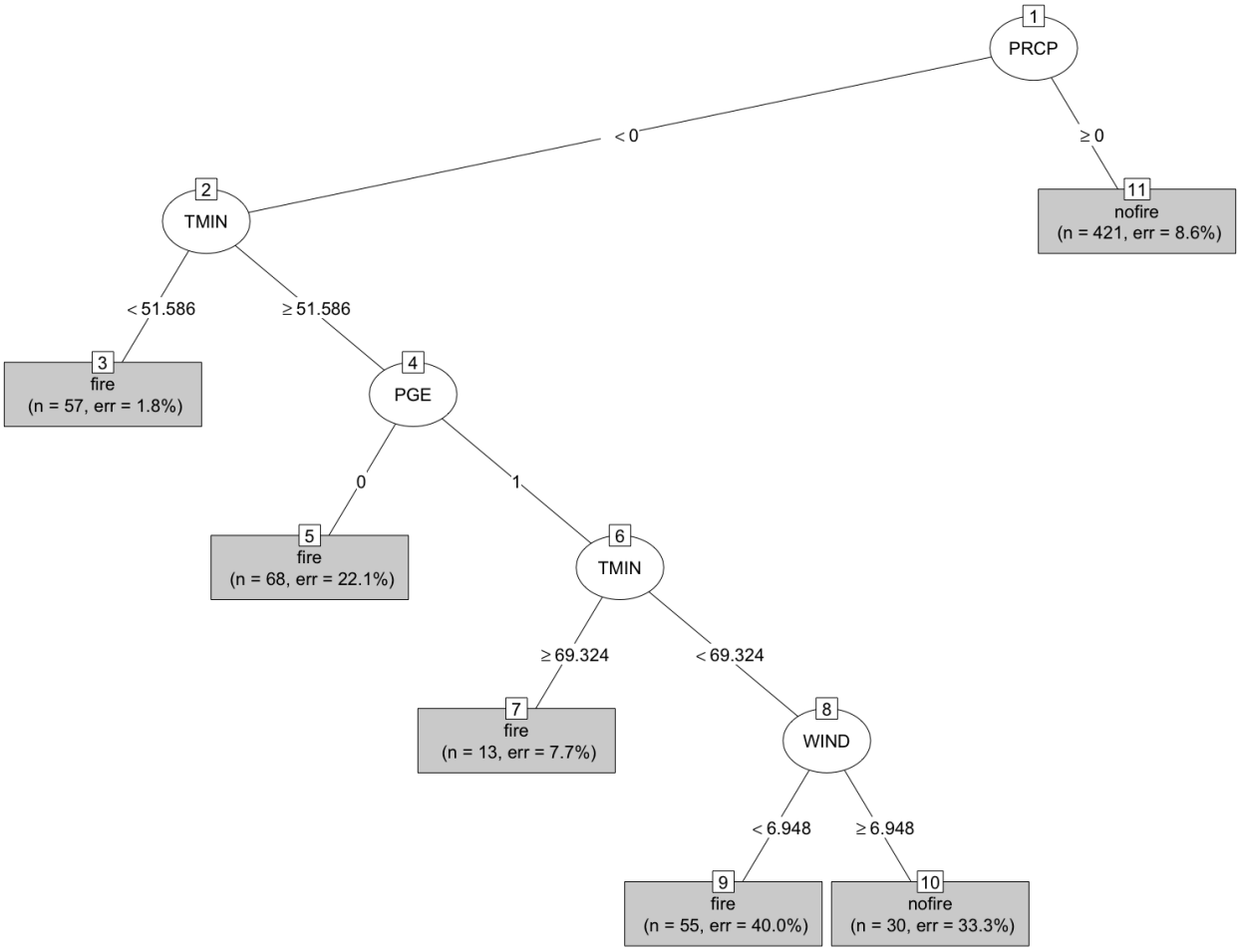
Figure 3: Importance of Variables in Random Forest Moc



The results from the importance table tell us that the presence of rain (precipitation) was the most important variable in predicting if there was a fire or not. This of course makes intuitive sense because the presence of water would likely extinguish fires. The next two most important variables in the decision tree were minimum and maximum temperature, while PG&E presence appeared to have the smallest importance. This implies that the accusation of PG&E being at fault may not be grounded.

To humor ourselves, we also ran a decision tree to compare outcomes with the importance table, which we found to be similar. Notably, PG&E does appear to be important in this model. However, we did not bootstrap this data and thus one should be wary of the accuracy of this model, and only consider it as “bonus” information.

Figure 4: Fire Risk Decision Tree



Findings

Our baseline random forest model, using data from 2017 and 2018, yielded an accuracy of 83.38%, precision of 73.13%, and recall of 73.5%. Next, we decided to split the data by PG&E presence to see if our accuracy would increase when we conditioned the data set on PG&E being present (filtering where PGE=1). We ran the same random forest model as above, but this time we did not control for PG&E. This model found accuracy of 82.71%, precision of 67.32%, and recall of 68.00%, all lower results than in our initial model.

In a similar fashion, we decided to split the data by fire risk (low, medium, and high) to see if conditioning the data on a single fire risk zone alone would increase the accuracy of our model. We ran the same random forest model separated by fire risk. The high fire risk zone yielded an accuracy of 74.62%, precision of 58.33%, and recall of 36.84%. The medium fire risk zone found an accuracy of 81.70%, precision of 72.55%, and recall of 72.55%. The low fire risk zone resulted in an accuracy of 83.03%, precision of 71.43%, and recall of 69.62%. These results were also all lower than our initial model, thus indicating that our baseline model will yield the most accurate results: our predictor variables of fire risk, month, maximum temperature, minimum temperature, precipitation, wind, population density, housing units, and PG&E presence predicted the presence of a fire correctly 83% of the time in 2017 and 2018.

Finally, we tested our model on a subset of our data from 2019. Because the fire data for 2019 is not yet published on the California Department of Forestry and Fire Protection (FARP), we cannot currently predict the accuracy, precision, and recall of our model in 2019. However, our model predicts that 6.4% of counties will have a fire in California in 2019.

Conclusion

In summary, we know that the "usual suspects" (e.g. fire risk, precipitation, month, wind, temperature, and population density) play a strong role in determining if there is a fire. Perhaps the most surprising (or unsurprising) result was that PG&E presence does in fact play a role in future fire risk, according to our model. In the context of economics, the fires could be considered a negative externality of production, which in turn means that stricter regulations or fines are justified for PG&E so that the company internalizes the true cost to society. The 2019 data for fires is unavailable, however the model can continue to be refined once the complete data set comes out. Fires can be considered a natural disaster and with the advent of "big data," this means that we can now use machine learning to predict other natural disasters. What sorts of indicators might be used to predict hurricanes or earthquakes? These sorts of disasters are incredibly costly to society, so the ability to forecast other likely events could have tremendous benefits to humankind. The possibilities are limited only by the data and our imagination.

References

Ailworth, Erin, and Peg Brickley. "PG&E Agrees to Pay $13.5 Billion in Settlement With Victims of California Wildfires." The Wall Street Journal, Dow Jones & Company, 6 Dec. 2019, www.wsj.com/articles/pg-e-agrees-to-pay-13-5-billion-in-settlement-with-victims-of-california-wildfires-11575691223?mod=searchresults&page=2&pos=1

Finch, Michael. “These California Counties Have the Highest Concentration of Homes Vulnerable to Wildfire.” The Sacramento Bee, 6 Aug. 2018, www.sacbee.com/news/california/fires/article216076320.html Preisler, Haiganoush K., and Anthony L. Westerling. “Statistical Model for Forecasting Monthly Large Wildfire Events in Western United States.” Journal of Applied Meteorology and Climatology, vol. 46, no. 7, 2007, pp. 1020–1030

“Size Class of Fire.” National Wildfire Coordinating Group, www.nwcg.gov/term/glossary/size-class-of-fire “2019 Verisk Wildfire Risk Analysis.” Verisk, 2019, www.verisk.com/insurance/campaigns/location-fireline-state-risk-report/

Data Sources

Fire boundary data: California Fire Data sets 2018. Updated 2019. https://hub.arcgis.com/datasets/405757064f1c478396922cb9ed0d189a?geometry=-124.349%2C36.416%2C-122.289%2C37.947&fbclid=IwAR0guvfKGYdLthT7jj08g-lOgHo6jKJJUF4kUN5kyvNMOPBYAc45MfDdsks

Fire boundary data: Fire Resource Assessment Program (FRAP). California Department of Forestry and Fire Protection. Updated 2019. https://frap.fire.ca.gov/mapping/gis-data/

Wind, temperature, and precipitation data: Climate Data Online. National Centers for Environmental Information, National Oceanic and Atmospheric Administration. Updated 2019. https://www.ncdc.noaa.gov/cdo-web/

Census data: California Census Data 2010. http://www.census-charts.com/Density/California.html