CS 182 Final Project: Wildfire Risk Prediction & Response Optimization in California

Blake Bullwinkel, Teresa Datta, Kristen Grabarz

December 10, 2020

1 Introduction

Wildfires can be devastating events that cause severe environmental destruction as well as substantial loss of resources and human life. The most recent bushfire season in Australia drew global attention, resulting in 34 deaths, 46 million acres of land burned, and over 9.3 thousand buildings destroyed - netting $103 billion AUD in costs. More recently, the 2020 wildfire season in the western United States caused similar devastation: 8 million acres of land and 13.9 thousand buildings were destroyed, resulting in 37 deaths and $2.7 billion USD in costs. Further, millions of species have been affected by fires disrupting their habitats, and the long-term health impacts of smoke exposure are still being studied. In the coming years, climate change may stand to increase the frequency and severity of wildfire events.

Our project leveraged artificial intelligence and machine learning methods to explore tactics for mitigating the effects of wildfires. We took a two-pronged approach, first predicting the fire risk in California counties in each month of the year using tree-based models with historical fire incident and weather data. We then leveraged mixed integer programming (MIP) using Google's OR-Tools package to determine the optimal assignment of limited firefighter personnel resources based on estimated risk and financial limitations, taking into account both department-affiliated firefighters and inmate firefighters.

2 Background and Related Work

Applications of artificial intelligence in fire prediction and management have been previously explored. For example, Vasconvelos et al. utilized neural networks and meteorological data to predict fire risks in central Portugal [18]. Madaio et al. worked with the Atlanta Fire Rescue department to create a predictive risk model [14]. Donovan and Rideout use mixed integer programming techniques [11] to identify efficient wildfire management systems with budget constraints commonly faced by fire managers. However, the system and problem-solving approach elaborated in this report were designed by us.

Of course, the work of tackling wildfires is not simply accomplished by academic groups, but more importantly by on-the-ground firefighting organizations in California. At the first sign of a fire, a rapid response team is sent out by a local fire station for the "first attack." Here, the goals are to provide an estimate of the size and to try to contain the fire as quickly as possible before it grows out of control. If the fire is not controlled in the first phase, an ”extended attack” begins and an incident control team is set up to coordinate the wider strategy[12]. Both ”dry” and ”wet” firefighting methods are employed- preemptively burning areas in the path of the fire, removing vegetation via axes, plows, chainsaws, and bulldozers, and dousing with water and flame retardant- both on the ground and via aircrafts[16]. If necessary, reinforcement from other states, other countries, the national guard, or the U.S. military is brought in.

The importance of the initial attack cannot be overstated. The longer a fire burns, the more out of control it can become, and bringing in reinforcements is a last resort option. Since this initial attack is coordinated and completed by local crews, this provides valuable motivation to our work of optimizing the location of firefighters.

The California wildfire management system has also faced severe scrutiny for its use of prison labor [13] through their Conservation Camp system [2]. A significant portion of active firefighters in California are inmates who are paid less than $1/hour. One formerly incarcerated lead firefighting engineer was paid only 37 cents per hour, or $56/month [15], as compared to career firefighters in CA who are paid $4,500/month with benefits [1]. Inmate firefighters also experience higher rates of certain injuries [19] and do not have a path to becoming career firefighters once their sentence is complete.

The Conservation Camp program provides about 3 million hours of response to fires and emergencies, and their services save California taxpayers about $100 million [4]. We were inspired by the social good applications presented in class, and thus wanted to explore how undervalued this population was. In addition, due to the COVID-19 pandemic, many of the correctional facilities that housed these inmate firefighters were dedensified. The meant that there were a reduced number of inmate firefighters for a year with a historic number of acres burned. By comparing the 2020 occupancy number with the full capacity of these camps, we were able to craft a natural experiment to examine the effects of the pandemic on California's ability to fight fires and the unrecognized amount of value and exploitative nature of this program.

3 Problem Specification

In short, this project aimed to leverage artificial intelligence and machine learning techniques to address two critically important questions related to wildfires: first, how can we predict wildfire risk across counties and at different times of the year? Second, given these risk estimates, how should we allocate limited fire response resources?

We addressed the first question using tree-based and ensemble learning methods to make monthly fire risk classification predictions across California counties, where risk is determined by total acreage burned per month. To improve the accuracy of our predictions and address imbalances in our original data, we also leveraged minority oversampling techniques. We then tackled the second question by optimizing the assignment of firefighters across California over the course of the year, given constraints on the numbers of department-affiliated and inmate firefighters in each county and using our fire risk predictions as the basis of our objective function. Specifically, we converted each risk prediction to an average monetary value and minimized total cost after accounting for wildfire damages and the monthly salaries of department-affiliated and inmate firefighters.

4 Approach

4.1 Data Assembly & Processing

The California Department of Forestry and Fire Protection (CAL FIRE) maintains data on wildfire incidents, which we accessed via a JSON API [3]. This dataset provides information on the timing, location, and damage (acres burned) of each fire. We exclude active fires from the analysis because their impact is still in-progress, and only include observations that are attributed to a specific county in California, since we focused our analysis on county-level predictions and allocations.

Along with the wildfire incident data, we also included topographical features of the counties[9] and historical monthly weather and climate data[8]. We then aggregated the data to the county and month level such that each row in the table represented a county-month pair and contained information on the total number of acres burned, total number of fires, year, county's average elevation, average precipitation, temperature (minimum, average, and maximum), air pressure (minimum and maximum), population, land area, and water area for the past seventeen years, from 2003 to 2020. Given the relationship between weather elements and fire incidents, as noted by the National Weather Service's Red Flag Program[17], we were particularly interested in exploring meteorological elements that could signal fire risk. After cleaning and assembling our data, we used scikit-learn's method `train_test_split` to split the data into a train and test set for our tree-based prediction models, reserving 20% of the data as the test set on which we evaluated the out-of-sample prediction effectiveness of the model. The other 80% was used for training and tuning the model.

4.2 Decision Tree & Random Forest Ensembles

After cleaning and processing our data, we moved onto modeling wildfire risk. As seen in Algorithm 1, a decision tree is a commonly used supervised learning algorithm that continuously splits the data based on certain parameters.

Algorithm 1 Decision Tree Learning Algorithm

procedure DECISION-TREE-LEARNING($examples, attributes, default$) **returns** a decision tree
  if *examples* is empty **then return** *default*
  else if all *examples* have the same classification **then return** the classification
  else if *attributes* is empty **then return** MAJORITY-VALUE
  else
    best $\leftarrow$ CHOOSE-ATTRIBUTE(*attributes, examples*)
    tree $\leftarrow$ a new decision tree with root test *best*
    for each value v_i of *best* **do**
      $examples_i \leftarrow$ {elements of *examples* with $best = v_i$ }
      tree $\leftarrow$ DECISION-TREE-LEARNING($examples_i, attributes - best$,
      MAJORITY-VALUE(*examples*))
      Add a branch to *tree* with label $v_i$and subtree *subtree*
    end for
    return *tree*

Building upon this basic decision tree learning algorithm, ensemble methods are techniques that leverage multiple learning algorithms to obtain improved predictive performance. Random forest is an ensemble method that randomly samples n subsets within the dataset and performs a best-fit decision tree within each of those bootstrapped datasets. Through this approach, each decision tree has its own feature set and resulting predictions, and a simple majority of trees determines the classification output. Compared to simple decision trees, random forest provides an added benefit of reduced variance and overfitting, since it weights random samples with replacement equally within the dataset. To mitigate overfitting, we use random forest classification, leveraging sci-kit learn's `RandomForestClassifier` method.

Initially, fire damage was represented quantitatively in the data, with both number of fires and number of acres burned as continuous variables. To more effectively fit our use case of classifying risk level, we binned the acres burned into six groups loosely inspired[7] by the National Wildfire Coordinating Group's class system, resulting in a modified response variable that is one of seven classes.

In order to tune the model hyperparameters, we used out-of-bag error to identify a best-performing model. In short, OOB error is the average error for each training observation calculated using predictions from the trees that do not contain that training observation in their respective bootstrap sample. Specifically, we experimented with different values for `n_estimators` (the number of trees in the forest) and `min_samples_leaf` (the minimum number of samples required to be at a leaf node).

The resulting random forest model with the best OOB error had 117 estimators and a minimum of 1 sample required to be at a leaf node, resulting in a 30.8% accuracy on the test set. With random chance yielding an accuracy of 16.7%, this classification algorithm performed better than random, but we wanted to improve our predictions.

4.3 Tree-Based Modeling Refinements: Minority Oversampling

As noted above, our initial random forest model had relatively unspectacular performance, with a test set accuracy of just over 30% (see section 5.1). After digging further into the data, we realized that the categorical classes of acres burned are highly unbalanced, with two of the seven classes comprising very few observations, and mid-tier classes out-representing the highest and lowest tier classes.

To address this, we leveraged the Synthetic Minority Oversampling Technique (SMOTE)[10]. SMOTE is a statistical technique used to increase the number of cases in a dataset in a balanced way by generating new instances from existing minority cases. We used the Imbalanced-Learn (`imblearn`) package to accomplish this. As shown in Figure 1, applying SMOTE balanced the dataset by oversampling the underrepresented classes to match the frequency of the most populous class.

Using this balanced data, we then performed a similar random forest modeling approach as described above. After tuning the parameters through cross-validation and OOB error evaluation, the resulting model had 370 estimators and a minimum of 1 sample required to be at a leaf node, resulting in a 63% accuracy on the test set - over double that of the previous model and substantially greater than the accuracy expected from random chance (see section 5.1).

At this point, we had constructed a machine learning model to classify wildfire risk-level across California counties. In practice, understanding wildfire risk would be only the first step in combating the damage these events can cause. Next we used these risk classifications to inform firefighter resource assignment in combating fires across the state.

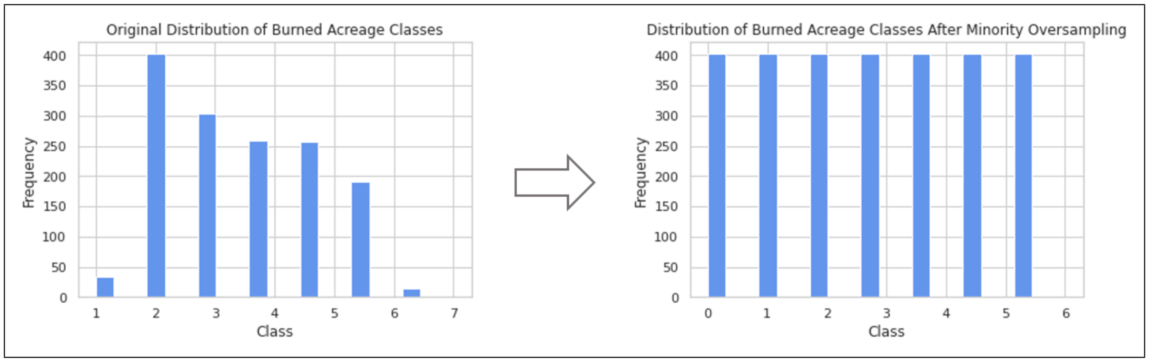

Figure 1: Class Distribution Before & After Minority Oversampling

4.4 MIP & Constrained Optimization

To explore the question of resource allocation for fighting wildfires, we focused on human capital: the number of firefighters who can be deployed to combat fires across California.

To more accurately inform our MIP implementation, we used a dataset containing fire departments registered to the National Fire Department Registry[5]. This resource contained information on each department's location and number of active firefighters, which we matched to the county level to correspond with our risk classification predictions. It's worth noting that the FEMA fire department registry website did not indicate specifically whether a firefighter can be affiliated with multiple departments. For the purposes of this project, we assumed that departments are mutually exclusive. We also found data on the inmate firefighter population in California via the California Department of Corrections and Rehabilitation, which includes information on the total number of inmates and the total inmate capacity at each of the 43 conservation or "fire" camps in the state [2]. Because of the COVID-19 pandemic, fire camps are currently operating at lower-than-typical capacity. With this in mind, we performed optimization on both actual and full inmate capacity.

Given that our risk classifications are on a monthly time-scale, we optimized the assignment of firefighters across the 58 CA counties in *each month* using the following MIP implementation.

Variables:

$$\begin{aligned}
c &\in \{1, 2, ..., 58\} : \text{the 58 California counties} \\
X_c &\in \{a_1, a_2, ..., a_c\} : \text{the number of inmate firefighters assigned to county } c \\
Y_c &\in \{b_1, b_2, ..., b_c\} : \text{the number of department-affiliated firefighters assigned to county } c \\
F_c &\in [0, 2 \cdot 10^9] : \text{the estimated cost (USD) of wildfire damages in county } c \\
\alpha &= 56 : \text{the average monthly cost (USD) of one inmate CA firefighter} \\
\beta &= 4500 : \text{the average monthly cost (USD) of one department-affiliated CA firefighter} \\
\gamma &= 12500 : \text{the estimated cost (USD) of wildfire damages saved per firefighter per month}
\end{aligned}$$

Note that a_c and b_c are the total number of inmate and department-affiliated firefighters, respectively, in county c. These values may be different for each month of the year. The estimated cost of wildfire damages F_c also changes for each month and is calculated from the monthly proportion of risk (from our tree-based model predictions) in county c, scaled by the total wildfire damages in California in 2020, which is estimated at $2 billion USD [6]. α, β, and γ were estimated from labor data as discussed in Section 2 and are assumed to be constant in each month and across all counties [1].

Constraints:

In addition to the constraints on the possible values of X_c and Y_c, we must have that:

$$F_c - \gamma(X_c + Y_c) \geq 0$$

for each county c. That is, the net cost due to wildfire damages and firefighter suppression efforts is at least zero (when the firefighters suppress fire damages completely).

Objective Function:

Finally, we sought to minimize the total cost across all California counties in a given month after accounting for the cost of suppression efforts and net wildfire damages, which is given by:

$$\begin{aligned} &\sum_{c=1}^{58} \alpha X_c + \beta Y_c + F_c - \gamma(X_c + Y_c) \\ =& \sum_{c=1}^{58} F_c + (\alpha - \gamma)X_c + (\beta - \gamma)Y_c \end{aligned}$$

5 Experiments & Results

5.1 Comparing Tree-Based Models

As can be seen in Table 1 and Figure 2, applying minority oversampling to our data resulted in substantially improved predictive accuracy. The plots below illustrate the OOB error rate resulting from various combinations of the hyperparameters `n_estimators` and `min_samples_leaf`. After extracting the pair of values with the lowest OOB error, we used these to train and test our final model. Interestingly, performing minority oversampling also made the difference in OOB error rate at different values of `min_samples_leaf` more stark.

	Test Accuracy
Random Forest: Unbalanced Model	0.31
Random Forest with Oversampling (SMOTE)	0.63

Table 1: Description of the results.

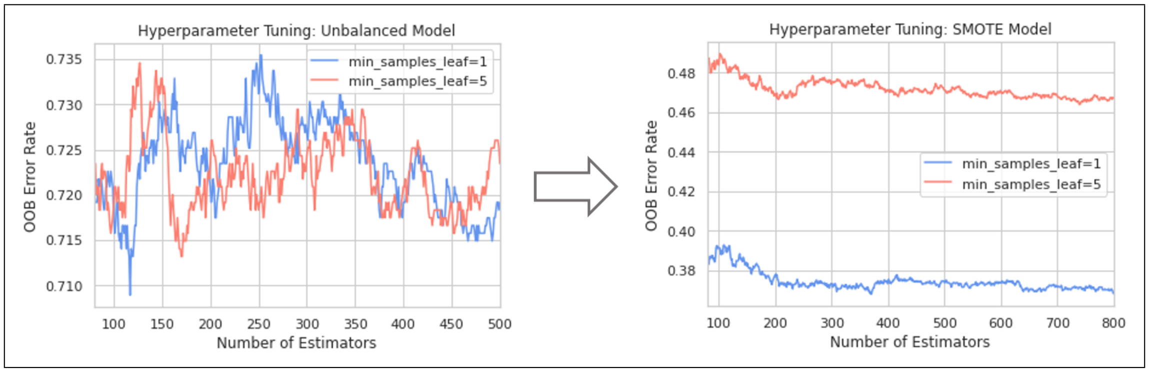

Figure 2: Hyperparameter Tuning of Unbalanced and SMOTE Models

We were also curious about which features were most important in determining splits in the decision trees. For random forests, each tree is slightly different due to the random sampling of predictors, which precluded us from generating a visual representation of the model. Instead, we plotted the feature importance of the top thirteen predictors in our final random forest model. As shown in Figure 3, feature importance is fairly consistent, with maximum and minimum humidity, year, and maximum temperature in the top four.

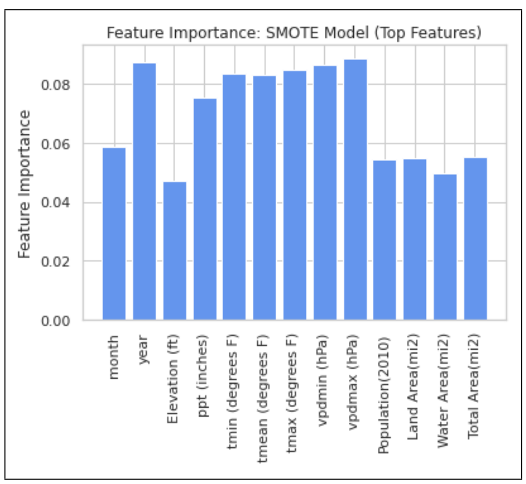

Figure 3: Feature Importance of SMOTE Model: Top Features

In a tangible use-case, one might imagine using this type of model to predict fire risk for future years given weather forecast data. Given our top four predictors and that weather conditions are predicted to get hotter and dryer under climate change, our model suggests that we can only expect the size of wildfires in California to increase in the coming years.

The improved model performance after minority oversampling is also evident in classification accuracy. As can be seen in Figure 4, the density of correctly-classified classes increases measurably after applying SMOTE, with classifications concentrated along the diagonal. Incorrectly classified observations are also frequently concentrated within one or two classes of the true value (as evidenced by less diagonal spread than the unbalanced model), which is advantageous for a use-case like this involving risk measurement for resource allocation. Additional measurements of model precision and recall can be found in Appendix C.

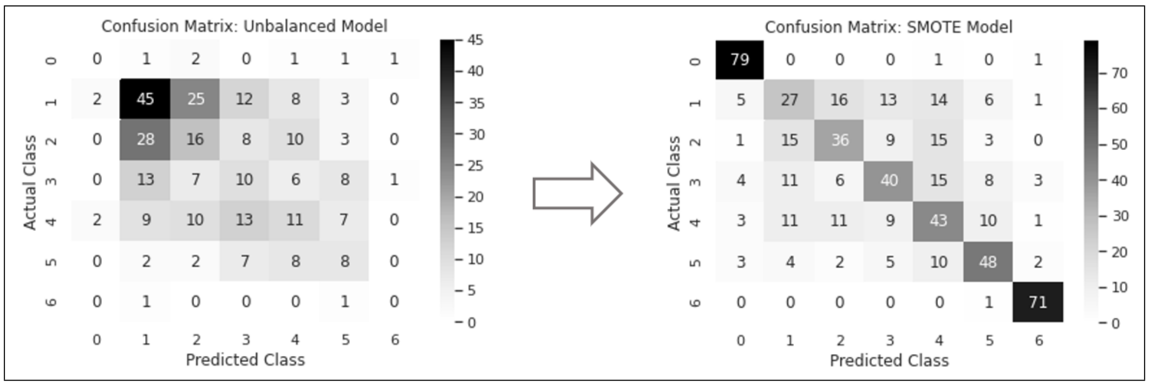

Figure 4: Confusion Matrices: Unbalanced vs. SMOTE Model

5.2 Constrained Optimization

Using our monthly wildfire risk classifications, along with the data and implementation outlined in Section 4.4, We produced an optimal (as defined by our objective function) assignment of inmate and department-affiliated firefighters across California counties in each month of the year. In order to better understand the unique impact of COVID-19 on firefighting resources, we performed the optimization on both the reduced inmate firefighter capacities this year and the full capacities.

A .csv file containing all assignments can be found in our project GitHub repository (optimized_firefighter_allocations.csv), and we have displayed the results for two counties in Figure 5. The plot on the top left shows the optimal number of firefighters to be assigned to Los Angeles in each month of 2020. As we would expect, the total number of firefighters assigned increases throughout the summer and peaks in August and September, when wildfire risk is at its peak. Also notice that the MIP solver assigns inmate firefighters before department-associated firefighters, as can be seen in January, April, and December, because the former is far less costly than the latter. When demand is high, inmate firefighters are assigned until capacity is reached, as indicated by the dotted red line.

The plot on the bottom left illustrates the same results when there is full inmate capacity. As before, the MIP solver assigns inmate firefighters first and until capacity is reached, which is higher than the reduced capacity due to COVID-19, thereby requiring fewer department-affiliated firefighters than before. To illustrate consistency of results, the plots on the right paint a similar picture for Humboldt county, which is located in the north of California.

Assuming the state of California seeks to minimize the cost of its wildfire response, our results also indicate an overreliance on inmate firefighters. By taking the difference between the value of our objective function using the reduced inmate capacities and its value at the full capacities, we estimated that an additional $26 million USD was lost this year due to the reduced numbers of available inmate firefighters alone. For workers who are paid only 37 cents an hour, this is a huge loss and a strong indication that the state of California relies heavily on them, without providing much in return.

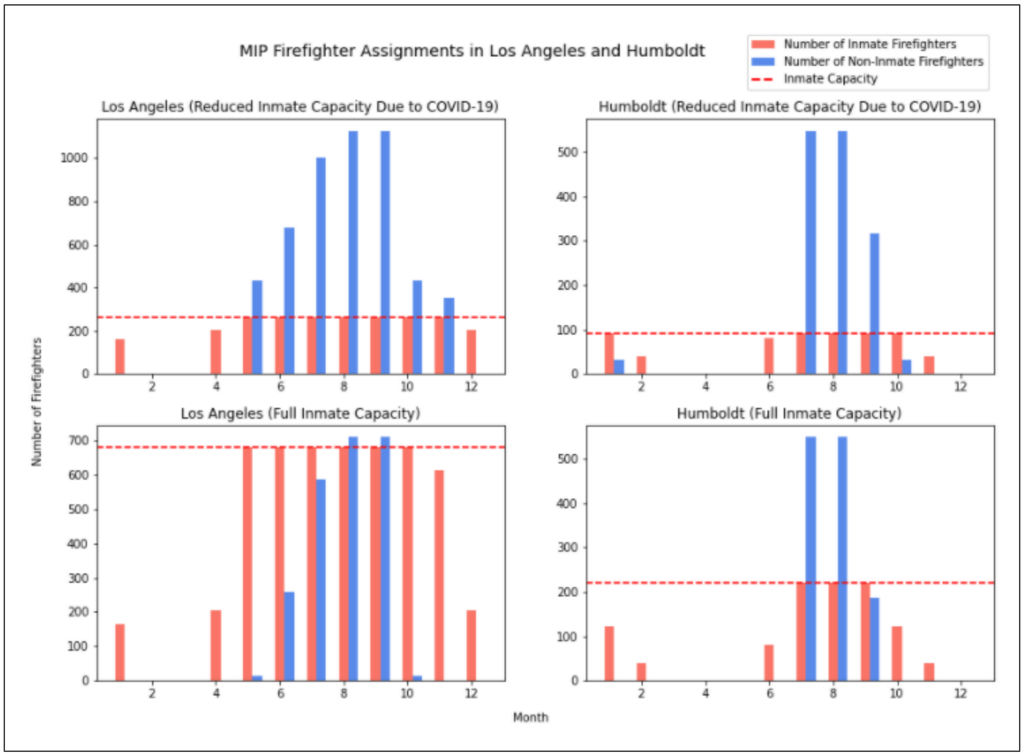

Figure 5: MIP Firefighter Assignments to LA and Humboldt Counties

6 Discussion

We predicted monthly wildfire risk in California counties using tree-based classification models. By converting these risk scores to cost estimates, we defined an objective function based on total cost in order to determine optimal assignments of department-affiliated and inmate firefighters using mixed integer programming.

The results of our random forest risk prediction model can be seen in Figure 6. The expected seasonal trends are immediately apparent - summer months are clearly associated with higher risk predictions across the state, and this seasonal pattern suggests that resources need not always be allocated at full capacity. In addition, the feature importance rankings of our final random forest model illustrated that weather variables, including temperature and humidity, were the best risk predictors. With climate change expected to impact both of these variables in the coming years, our model predicts that wildfire risk across California will only increase.

These maps also indicate which counties are more at risk than others, and our predictions provided the basis for our MIP objective function. The results of our optimized firefighter assignments spawned an interesting consideration regarding the treatment of disadvantaged groups. In particular, the low cost of inmate firefighters makes them more likely to be disproportionately put in harm's way. As illustrated by 5, the MIP solver assigned all the low-cost inmate firefighters until capacity was reached. By estimating the effect of the reduced inmate firefighter population on total cost, we found that an additional $26 million USD was lost this year, even though inmates are paid only 37 cents per hour.

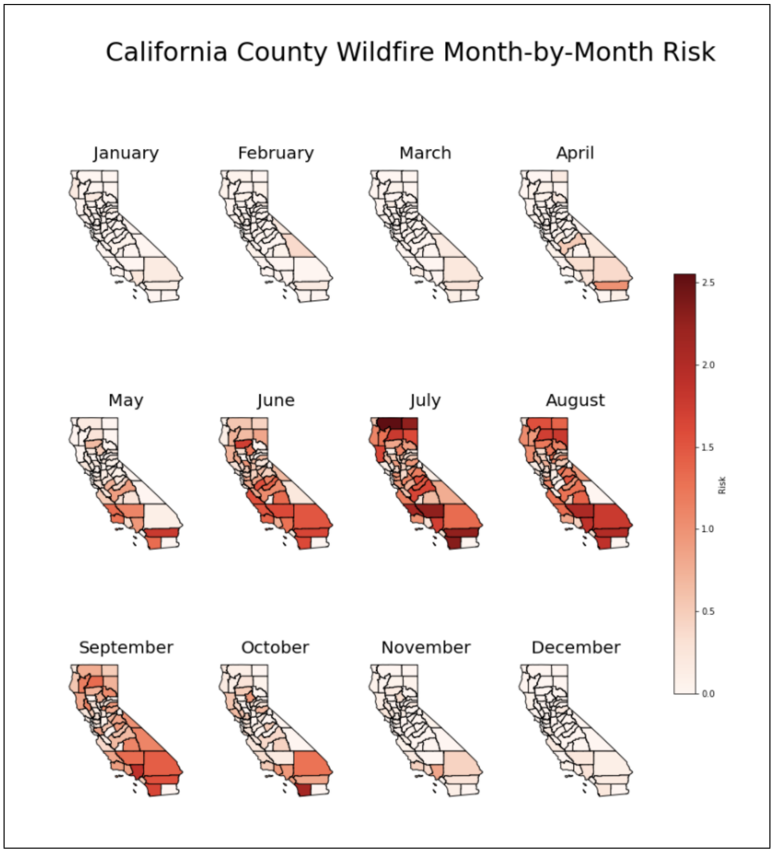

Figure 6: California County Wildfire Risk Map by Month

This exemplifies a broader concern regarding algorithmic approaches to resource allocation, particularly when human subjects are involved: disadvantaged groups (such as the inmate firefighter population) may have lower cost structures associated with them, resulting in disproportionate demand for their labor. There is also evidence to suggest that inmate firefighters are disproportionately harmed compared to civilian and professional firefighters: they are more than four times as likely to incur object-induced injuries and over eight times as likely to be injured after inhaling smoke and particulates in the air as compared to civilian firefighters. [19] To address these concerns, we could refine our approach by adding constraints designed to protect the inmate firefighter population, for example by limiting the proportion of fire camp residents that can be assigned or by reducing their maximum working hours. Future iterations of this product could also increase the accuracy of our predictions by leveraging sub-county, rather than county-level data. Given the size and geographic diversity of California counties, this approach might yield different and interesting results.

A System Description

Code and results can be found at the following GitHub Repo:

`https://github.com/teresadatta/CA-Wildfire-Risk-Prediction-and-Optimization`

Detailed descriptions are in the README.md. Feel free to contact our group if there are any questions or issues.

B Group Makeup

We operated very collaboratively and cohesively on this project and worked together to complete each of the above components. Some areas where we each contributed more:

Blake Bullwinkel: Constrained Optimization

Teresa Datta: Chloropleth maps, Data wrangling for Constrained Opimization

Kristen Grabarz: Minority Oversampling via SMOTE, Data Collecting

C Additional Plots & Charts

Balanced Model: Class Precision & Recall							
Class	0	1	2	3	4	5	6
Precision	84%	39%	53%	51%	54%	59%	85%
Recall	98%	34%	51%	48%	48%	64%	97%

Figure 7: SMOTE Random Forest Model: Class Precision & Recall

	% Accurate	% Misclassified by 1 Tier	% Misclassified by > 1 Tier	% Classified as Accurate or Within 1 Tier
Actual 0:	93%	1%	6%	94%
Actual 1:	37%	23%	40%	60%
Actual 2:	39%	35%	25%	75%
Actual 3:	46%	22%	32%	68%
Actual 4:	41%	23%	36%	64%
Actual 5:	61%	16%	23%	77%
Actual 6:	97%	1%	1%	99%

Figure 8: Balanced Model: Classification Accuracy Spectrum

References

[1] Bureau of labor statistics: Firefighters. *Bureau of Labor Statistics*.

[2] California dept. of corrections and rehabilitation: List of conservation camps. *https://www.cdcr.ca.gov/facility-locator/conservation-camps/camps/*.

[3] California wildfire data. *https://www.fire.ca.gov/incidents/*.

[4] California's inmate firefighters: 9 things to know. *abc10*.

[5] Federal emergency management agency national fire department registry. *https://apps.usfa.fema.gov/registry/download*.

[6] National large incident year-to-date report. *https://gacc.nifc.gov/*.

[7] Nwcg fire size classification. *https://www.nwcg.gov/term/glossary/size-class-of-fire*.

[8] Prism climate data. *https://prism.oregonstate.edu/recent/*.

[9] Us census bureau gazetteer files: Us county data. *https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html*.

[10] Bowyer Kevin W. Hall Lawrence O. Chawla, Nitesh V. and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[11] Geoffrey H. Donovan and Douglas B. Rideout. An integer programming model to optimize resource allocation for wildfire containment. *Forest Science*, 49(2):331–335, 2003.

[12] Gareth Evans. How do you fight extreme wildfires? *BBC*, Nov 2019.

[13] Thomas Fuller. Covid limits california's efforts to fight fires with prison labor. *NYTimes*, August 2020.

[14] Michael Madaio, Shang-Tse Chen, Oliver L. Haimson, Wenwen Zhang, Xiang Cheng, Matthew Hinds-Aldrich, Duen Horng Chau, and Bistra Dilkina. Firebird. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, August 2016.

[15] Amika Mota. I saved lives as an incarcerated firefighter. to california, i was just cheap labor. *The Guardian*, September 2020.

[16] John Schwartz. How to fight a wildfire: Grueling work and managing risks. *NYTimes*, Sept 2020.

[17] NOAA US Department of Commerce. Fire weather criteria, Mar 2015.

[18] M. Vasconcelos. Spatial prediction of fire ignition probabilities: Comparing logistic regression and neural networks. *Photogrammetric Engineering and Remote Sensing*, 67:73–82, 2001.

[19] Abby Vesoulis. Inmates fighting california wildfires more likely to be hurt. *Time*, Nov 2018.