Predicting Number of Personnel to Deploy for Wildfire Containment

Olivia Samples

[1]*College of Computing and Technology, Lipscomb University Nashville, TN, USA*

ABSTRACT

Wildfire size, frequency, severity, and associated fatalities have surged at an alarming rate over the past 25 years, resulting in steep budget increases. According to USDA, the annual budget of U.S. Forest Service devoted to wildfires had more than tripled, jumping from 16% to 52% between 1995 and 2015 and it exceeded $2 billion in 2017. Under the current budget and capacity constraints, allocating correct amount of personnel and equipment to a fire timely is vital in suppresing fire, reducing costs, and saving lives. In this paper, we use gradient boosting decision trees to predict the number of personnel needed to effectively fight a wildfire. By combining the US wildland fire incident data and weather data, our model obtained a coefficient of determination (R^2) 77.78%, a significant improvement over historical solution. Our model can potentially be used to provide decision support for those units who fight wildfires.

KEYWORDS

Forest fires, wildfire management, firefighting resources management, weather impact, Highly Valuable Resources and Assets.

1. INTRODUCTION

Wildfire size, season length, and variability have increased at an alarming rate over the past 25 years.[8] These fires posed major threat to our air quality, in particular in western US,[16] lakes,[12] and residential land. In turn, the US budget for fighting wildfires tripled from 1995 to 2015 and exceeded $2 billion in cost in 2017.[19] As fire activity rises, the US forest service has had to increase firefighting staff and decrease land management staff. From 1998 to 2015, the fire staff adjusted from 5,700 employees to 12,000 employees, accounting for a significant portion of budget adjustments within that time frame.[18]

A recent study showed that fire management is accountable for at least half of fire suppression costs .[7] Hence, accurate forecasting of personnel needs is vital to improve the efficiency of U.S. Forest Service operations. A model that predicts the optimal number of personnel needed for fire suppression would potentially trim the firefighting budget, allowing for the focus to be redirected to land management and preventative efforts.

In this paper, we use gradient boosting decision trees to predict the correct number of personnel needed to effectively fight a wildfire. Our model paired with wildfire forecasting models would aid firefighting agencies by facilitating more efficient resource management and decision support, enabling those fighting the fires to do so in the most effective manner possible.

1.2 Related Work

An abundance of previous work has been attributed to the prediction of forest fires, using tools such as meteorological data, satellite data, and infrared smoke scanners.[3] While they are able to predict fire location and size, they do not provide insight on how to react in real-time. Historically, the USDA Forest service uses tools such as the National Fire Management Analysis System (NFMAS), the National Park service uses a model called FIREPRO, and even the California fire service uses its own system called the CFES-IAM model. Each fire service utilizes these tools as preparation insight towards optimizing resource allocation per annual fire season, incorporating *multiple* fire incidents.

Conversely, multiple models have been built in order to provide insight on *distinct* fire incidents. Donovan's Integer Programming model attempts to do this by predicting features such as time and cost of resource per individual fire.[4] Wei's Chance Constrained programming model optimizes a fire fighting perimeter for a particular fire incident.[17] Similarly, Haight's scenario based standard response model determines the correct number of dispatched engines to efficiently suppress individual fires.[5] However, all of these models fail to provide insight for deploying personnel, an attribute that all of these prior resources are dependent on.

Susana Martin-Fernandez was successful in incorporating the manual resource in her real-time model. After applying discrete simulation algorithms and Bayesian optimization methods, she marked the importance of providing a real-time model so that fire management teams are able to efficiently apply a model onto distinct fire incidents.[11]

In 2019, John Carr and Matt Lewis used a unique data set received from the US Department of Interior to answer the following question: What is the correct number of personnel needed to effectively fight a wildland fire in the U.S. without using unnecessary resources? Their exploratory data analysis revealed that most of the incidents were not destructive and were easily contained. Roughly 2.3% of the fire incidents burned more than 1 acre, requiring extensive resources and personnel ranging from a few to over 1,000.

Carr and Lewis were able to devise a model capable of explaining 61% of the data variation. This work had limitations, as it included highly correlated features, a basic training model with no hyper parameter tuning, and did not incorporate weather data as features. Their solution set is the basis for our project.[2]

2. DATA COLLECTION AND PREPROCESSING

2.1 Fire Incident Data Collection

IRWIN Observer is a read-only web application designed for viewing data that is being shared through the Integrated Reporting of Wildland-Fire Information (IRWIN) integration services. Access to this application is granted by the Chief Information Officer for the GeoPlatform ArcGIS software (https://irwin.doi.gov/observer). IRWIN Observer provides current and transactional views of incident and resource data being shared by partners within the wildland fire community. This data provides the location of existing fires, size, conditions and several other attributes that help classify fires.

We retrieved fire incident data from the US Department of Interior's IRWIN Observer database which details the Office of Wildland Fire's complete records of wildland fires within the country. The data represents all US wildland fires within a five year span between January 1st, 2015 and December 31st, 2019. The incident dataset includes 135 features that describe 139,859 fire incidents. Additionally, we pulled a resource dataset detailing 16 resource features associated with these incidents. The resources table includes information such as: number of personnel, trucks, planes, helicopters, boats, and agencies supporting the firefighting efforts. We will combine these datasets together using the fire incident's "Irwin ID".

2.2 Weather Data Collection

Visual Crossing Corporation is a technology leader when it comes to historical and forecasted weather data and weather APIs. Access to the historical weather API can be gained after creating an account here: https://www.visualcrossing.com/weather-api. We were able to bulk query weather data for our corresponding incidents based on the data and geo-location of each incident.

The resulting features include: Wind Direction, Temperature, Max Temperature, Min Temperature, Visibility, Windspeed, Heat Index, Cloud Cover, Precipitation, Precipitation Cover, Sea Level Pressure, Dew, Humidity, Wind Gust, Conditions, and Wind Chill. These new data features were added to the fire incident dataset for further modeling.

2.3 Data Preprocessing

In terms of data cleaning, we first replaced any strings that should be null, but were mis-labeled. We also dropped homogenous variables that had no variance or were 100% null. Of the remaining variables we selected the ones that we believed would be relevant and useful to this research and did not select the ones that we deemed irrelevant or highly correlated. Most of the 135 features were extremely sparse and required significant cleaning. Many had mixed data types that required 'hard-coded' transformations. Our final features were selected based on the following criterion: must be relevant, non-correlated or duplicating of other features, and decently populated.

We created new features by converting date-time entries into 'Year', 'Month', and 'Day' values. Afterwards, the 'datetime' column was dropped.

Lastly, two of the features that we included were the longitude and latitude coordinates of the wildfire locations. Because of this we needed to remove records that did not have both lat/long entries or a valid ID number to identify the incident. The ID number was a mandatory field because we needed it to join the resources table, allowing us to aggregate personnel, equipment, and other resources that were associated with each fire. The lot/long entries were a mandatory field because we needed them to join the weather data, providing further insight for the model. Minimal records were dropped to maintain as much data as possible.

Next, we incorporated the resources table and created a new aggregated database. The resources table included information such as: number of personnel, trucks, planes, helicopters, boats, and agencies supporting the firefighting efforts. All of this data was aggregated by grouping by each unique incident and summing the resources. This resulted in data such as the types of resources, their quantity, and the total combined personnel. For fires without resources, we filled these columns with zero.

Before running the model, we chose to drop all the resources variables, so as to avoid training the model on correlated features to personnel. For example, each engine had four personnel assigned. So, if there were three engines deployed to a fire, there would be 12 personnel. We also removed 'Final Acres', 'Days Not Contained', and 'Fire Code' as we recognized their future unavailability at time of model use. These features were not removed in John Carr and Matt Lewis' model.

Our initial final dataset without weather data had 33 features, 1 target, and 139, 858 records. We will use this dataset for exploratory data analysis. We will also train models on two datasets that do not contain any 'Total Personnel' = 0 (see Section 2.4), one without weather data and one with weather data. Comparison of these two will show the impact of weather data on our model. Our final dataset without weather data had 33 features, 1 target, and 5,852 records. Our final dataset with weather data had 51 features, 1 target, and 5,852 records. We explore and build a model on each of these three datasets.

2.4 Exploratory Data Analysis

After we had fully processed the data, cleaned it, and combined the resources table, we then conducted exploratory data analysis (EDA). We started by looking into the fire incident distribution based on Point of Origin (POO), and we found that the majority of incidents were on the west coast of the United States, mainly California.

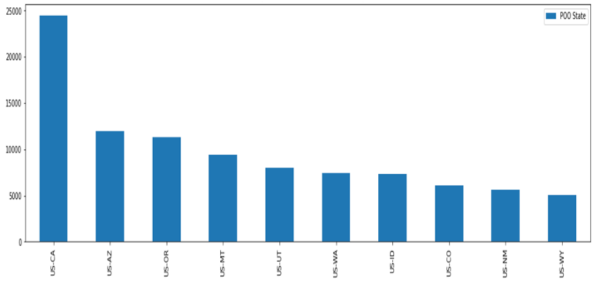

Figure 2: Histogram of Top Ten Represented States of Point of Origin

Additionally, we found that personnel are mostly required for fire incidents within the summer months, with a few exceptions in October, November. Both of these observations are consistent with general understanding of the United States' current fire threats.

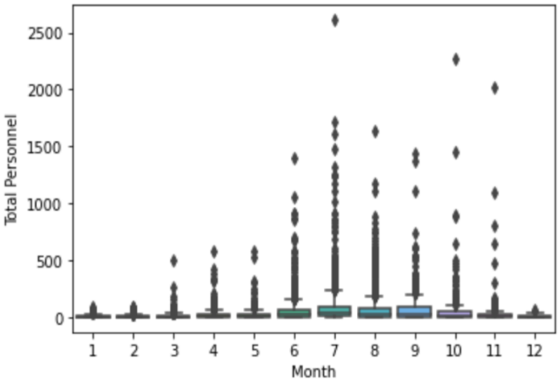

Figure 3: Box Plot of Total Personnel (Greater than Zero) based on Month in a Calendar Year. Displays spike in summer months.

What's also interesting, is that we found that the vast majority of incidents were easily contained and were not very destructive, in fact most did not need any personnel at all. The following pie chart displays this personnel distribution.

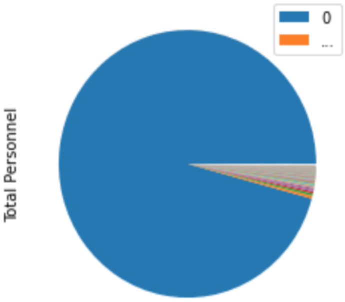

Figure 4: Pie Chart of Total Personnel Distribution *Including* Records where 'Total Personnel' is Equal to Zero.

There are relatively few, extremely destructive fires that cause the majority of the damage and are responsible for most of the cost. This is important because, according to the Department of Agriculture, the ten largest fires in 2014 cost more than $320 million dollars and that fire suppression is predicted to increase to nearly $1.8 billion by 2025. As suppression costs increase, a significant threat is presented to all of the services that support our national forest, fire-related or otherwise.

Therefore, it is more important to have a model that is accurate for these extremely destructive fires, the fires that actually need personnel. That is why we created the final two datasets that represent the feature distribution without records that contain 'Total Personnel' equal to 0. The models based off of this idea will provide a more useful, real-time solution for fire agencies when predicting personnel coverage in the circumstances they need those personnel.

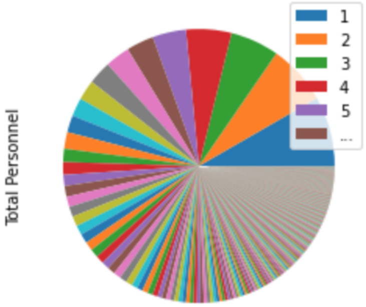

Figure 5: Pie Chart of Total Personnel Distribution *Excluding* Records where 'Total Personnel' is Equal to Zero.

4. METHODS

Rank Order Label Encoding (ROLE) – a form of target encoding – was used on the categorical variables intended to be included in the model. What this does is provide us with an average of the target variable for every unique value within each feature. For example, after applying ROE to the POO State feature, we had an average number of personnel for each of the states included in the data. The states were then ranked based on these averages, from one to fifty. The importance of this method is that it turns these categorical values into ordinal values, which we applied to the training set and then used the results – or averages – on the test data set to prevent data leakage. We can also include features with many unique values (greater than 10), where using other methods, such as one-hot-encoding, would increase the feature space by a significant amount.

ROLE combined with Label Encoding, converting categories into numeric form, prepares the categorical features for cross-validation and grid search via XGBoost, as the combination maintains the feature space and converts all data types to integers.

For the predictions we used scikitlearn's XGBoost as our model library for training. XGBoost is a distributed gradient boosting library that provides parallel tree boosting to machine learning algorithms and typically out-performs most machine learning models across a variety of applications. For further tuning, we used a k-fold cross-validation model with grid search for hyper parameter estimation. Grid search uses brute force to test optimal choice of parameter set with the model, while cross validation compares results on k-subsets of the data to retain best parameters. Combined, these tools allow hyper parameters to positively control the training process.

Our optimal parameter set for XGBoost included the maximum depth of tree, where a larger depth correlates to overfitting. We kept ours small. Also, the set included a minimum weight of child, where larger correlates to more conservative. We increased ours from the default of 1. Finally, the set included the number of cross-validation splits and number of jobs to run in parallel.

We trained this model using a 60/40 train/test split on the three datasets. For model interpretation, we used R^2 and Root Mean Squared Error (RMSE) for benchmarking performance, and XGBoost's *plot_importance* library for understanding feature importance. R^2, or the coefficient of determination, represents how close the model is to explaining the data variability. RMSE represents the +/- uncertainty of personnel representation that the model predicts.[5] Additionally, the *plot_importance* library utilizes an F score to represent the significance of each feature on the model, where the score represents how accurate each feature is in determining the target variable.

5. RESULTS

5.1 Dataset 1: Includes Total Personnel = 0, Excludes Weather

We found that, with our 5-fold cross-validation using a grid search tuning model on XGBoost with a max depth of 2 and minimum child weight of 3, we are able to obtain a RMSE of +/- 15.62 and an R^2 of 0.7534 on the test set, meaning that our model can explain 75.34% of the variation in the data in terms of number of personnel deployed to fight a fire. This performs better than the previous model that obtained an R^2 of 0.61.

Examining the most influential features, we see that 'N Agencies Supporting' and 'Estimated Cost to Date' are the most significant. 'Agencies Supporting' refers to the geographic area coordination center associated with the National Wildfire Coordinating Group.[13]

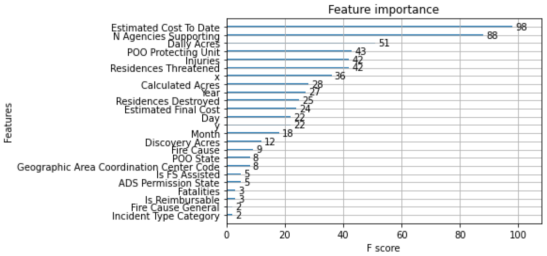

Figure 6: Feature Importance Plot of Dataset 2 Model Results where 'Estimated Cost to Date' Represents Most Important Features

5.2 Dataset 2: Excludes Total Personnel = 0, Excludes Weather

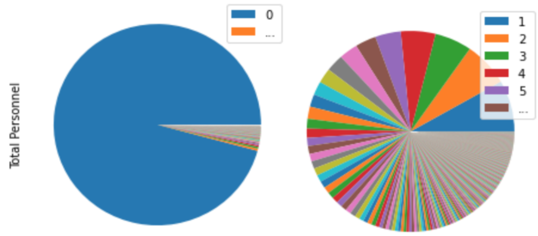

Figure 7: Pie Chart of Total Personnel Distribution *Including* (Left) and *Excluding (*Right*)* Records where 'Total Personnel' is Equal to Zero.

Recall the dataset created without 'Total Personnel' equal to 0 when considering the distribution of 'Total Personnel' in Figure 7. The model trained on the dataset of 5,852 incidents without any records where 'Total Personnel' equal to 0 resulted in an R^2 of 0.7757 on the test set and a RMSE of +/- 67.12, where the max depth and minimum child weight were both 3. This model already performs better than Dataset 1.

Some features increased or decreased their F score from Dataset 1 model to Dataset 2 model. For instance, 'Daily Acres' went from a 18 to 51 score. This could be because fire incidents that required personnel tended to have a more distributed number of daily acres. Also, 'Residences Destroyed' went from a 13 to 25 score. This could be more influential because personnel are particularly needed when residences are involved.

5.3 Dataset 3: Excludes Total Personnel = 0, Includes Weather

Finally, we will add weather data to this dataset of 5,852 incidents. In other words, we will consider the dataset that includes weather data and does not include any records where 'Total Personnel' is equal to 0. The model performs best with an R^2 of 0.7778 on the test set and a RMSE of +/- 66.80, where the max depth and minimum child weight were both 3.

Considering feature importance for this result, 'N Agencies Supporting' and 'Estimated Cost to Date' remain at the top, but they also now include weather features such as 'Wind Direction', 'Min Temperature', and 'Sea Level Pressure'. These features provided insight to the model that was not there previously.

6. DISCUSSION

These results provide an improved solution for predicting personnel required to distinguish wildfires in the United States. Previously, highly correlated features, features that would not be available at the time of model use, a basic model with no hyper parameter tuning, and no weather data were used to build a model with this solution in mind. Accounting and responding to all of these limitations, our results performed better than these historical solutions.

The results indicate that the model will be best used when firefighting agencies know they need at least one personnel. In other words, when they ask the question: *When we actually need personnel, how many do we need?* Furthermore, this question will be answered best when accounting for weather data. This inclusion of weather data is a massive win since no other models had utilized these features prior.

Even while this solution has provided an improved result from historical models, challenges exist for implementation. Firstly, a manageable API or any other deliverable source is not yet available. Also, model drift is inevitable, so resources need to be allocated for future model training. Lastly, social challenges include: deployment of the model may challenge the status quo and cost savings in fire management may not directly benefit the forest service, dimming the support of such implementation.

REFERENCES

[1] Canton-Thompson, J., Thompson, B., Gebert, K., Calkin, D., Donovan, G., & Jones, G. (2006). Factors affecting fire suppression costs as identified by incident management teams. Res. Note RMRS-RN-30. Fort Collins, CO: US Department of Agriculture, Forest Service, Rocky Mountain Research Station. 10 p., 30. https://www.fs.fed.us/rm/pubs/rmrs_rn030.pdf Retrieved on Feb. 10th, 2020.

[2] Carr, J. & Lewis, M.(2019). Predicting Number of Personnel to Deploy for Wildfire Containment. Research Symposium Lipscomb University.

[3] Cortez, P., & Morais, A. D. J. R. (2007). A data mining approach to predict forest fires using meteorological data. http://www3.dsi.uminho.pt/pcortez/fires.pdf. Retrieved on Feb. 10th, 2020.

[4] Donovan, G. H., & Rideout, D. B. (2003). An integer programming model to optimize resource allocation for wildfire containment. Forest Science, 49(2), 331-335.

[5] Grus, Joel (2019). Data Science from Scratch. First Principles with Python, 63-99.

[6] Haight, R. G., & Fried, J. S. (2007). Deploying wildland fire suppression resources with a scenario-based standard response model. INFOR: Information Systems and Operational Research, 45(1), 31-39.

[7] Ingalsbee, T. & Raja, U. (2015) The Rising Costs of Wildfire Suppression and the Case for Ecological Fire Use. The Ecological Importance of Mixed-Severity Fires: Nature's Phoenix. Chapter 12. Elsevier Inc.

[8] Jones, Matthew W., Smith, Adam., Betts, Richards., Canadell, J. G. (2020). Climate Change Increases the Risk of Wildfires. Rapid Response Review using ScienceBrief.org. Retreived from https://www.preventionweb.net/files/73797_wildfiresbriefingnote.pdf.

[9] Lee, Y., Fried, J. S., Albers, H. J., & Haight, R. G. (2013). Deploying initial attack resources for wildfire suppression: spatial coordination, budget constraints, and capacity constraints. Canadian Journal of Forest Research, 43(1), 56-65.

[10] Martell, D. L. (2015). A review of recent forest and wildland fire management decision support systems research. Current Forestry Reports, 1(2), 128-137.

[11] Martin-fernÁndez, S., Martínez-Falero, E., & Pérez-González, J. M. (2002). Optimization of the resources management in fighting wildfires. Environmental Management, 30(3), 352-364.

[12] McCullough, Ian., Cheruvelil, Kendra., Lapierre, Jean-Francois., Lottig, Noah., Moritz, Max. (2020). A fireside chat: large wildfires are a looming threat to US Lakes. et al.Earth and Space Science Open Archive ESSOAr.Retreived from https://search.proquest.com/docview/2463801277/fulltextPDF/6B8AD3BB3F4149D2PQ/1.

[13] National Wildfire Coordinating Group, & Nwcg. (2015, July 29). National Wildfire Coordinating Group (NWCG). Retrieved March 28, 2020, from https://www.nwcg.gov/sites/default/files/stds/standards/geographic-area-coordination-center_v1-0.htm.

[14] NFPA 1. (2018). Retrieved from https://www.nfpa.org/codes-and-standards/all-codes-and-standards/list- of-codes-and-standards/detail?code=1.

[15] North, M. P., Stephens, S. L., Collins, B. M., Agee, J. K., Aplet, G., Franklin, J. F., & Fule, P. Z. (2015). Reform forest fire management. Science, 349(6254), 1280-1281.

[16] Selimovic, Vanessa, "Air Quality and Climate Impacts of Western U.S. Wildfires" (2020). *Graduate Student Theses, Dissertations, & Professional Papers*. 11634. https://scholarworks.umt.edu/etd/11634.

[17] Wei, Y., Bevers, M., Belval, E., & Bird, B. (2015). A chance-constrained programming model to allocate wildfire initial attack resources for a fire season. Forest Science, 61(2), 278-288.

[18] United States Department of Agriculture(USDA). Forest Service. (2015). The rising cost of wildfire operations: Effects on the Forest Service's Non-Fire Work. https://www.fs.usda.gov/sites/default/files/2015-Rising-Cost-Wildfire-Operations.pdf. Retrieved on Feb. 10th, 2020.

[19] Reuters. Cost of fighting U.S. wildfires topped $2 billion in 2017 | Environment, 9/14/2017, https://www.reuters.com/article/us-usa-wildfires/cost-of-fighting-u-s-wildfires-topped-2-billion-in-2017-idUSKCN1BQ01F. Accessed 14 Jan 2021.

ACKNOWLEDGEMENTS

Trent Girard currently serves as a Fire Management officer for Cherokee National Forest. As a strong advocate for prescribed fires, Girard has aided in the optimization of wildland fire control throughout his entire career. Girard provided insight on current US wildland fire protocol and clear explanation of fire management terminology. (2020)