Wildfires in the United States

Samantha Maillie and Ariel Sim

Link to GitHub repository:
https://github.com/UCDSTA208/2018-SimMaillie/tree/master

Background

Every year wildfires cost billions, which is hardly comparable to the lives lost from wildfires, (3). Utilizing an SQL database containing information on 1.88 million wildfires nationwide, we want to investigate if cause of fire can be predicted by size and location of the fire. Why is this important? Once fire blazes spread and grow to a certain point, keeping them under control becomes very difficult and poses high risks. Our best bet, therefore, is to prevent wildfires from starting at all if possible. As such, deeper understanding of the relationship between cause, size, and location of fire can help fine tune policies designed to prevent wildfires, which in turn offer better insight on wildfire prevention.

Our plan is to do some preliminary analysis to determine which classifications algorithms best suit our problem. We will then implement the relevant procedures and compare them using numerous methods including k-fold cross validation and confusion matrices.

Why this dataset? We wanted a dataset that could offer insight into a problem we are familiar with. This dataset found at the link below is large, informative and not excessively worked over like many of the datasets easily accessible online. This offers us the opportunity to discover new conclusions from the dataset as opposed to essentially replicating results.

Data Source and Description

This dataset was originally designed for the use of the Fire Protection Agency, (FPA). It contains 1.88 million data points with information on wildfires nationwide between the years 1992 and 2015, (8). The goal of this dataset's publication on Kaggle is to investigate the relationship between the size, location, and cause of wildfires, (8). Our project aims to tackle this problem.

The database in considerably large and can be downloaded via this link:
https://www.kaggle.com/rtatman/188-million-us-wildfires

Data Information and Selection

The variables provided contain a lot of excess information. The list of original variables can be found on the Kaggle site linked above, (8). We also have a lengthier discussion regarding the variables in the associated GitHub repository listed at the top of this paper.

These excess of variables are not all predictors but identifying indicators for the authorities to file the proper paperwork regarding the fire. As such, these variables are not candidates for predictors and will not be included in the analysis. Fitting the model with all remaining predictors would still be problematic because there are a lot redundancies and overlap in the variables. One obvious example of variable overlap is “FIRE_SIZE” and “FIRE_SIZE_CLASS”, which clearly measures the same thing in two different ways. The updated list of variables that we deemed as potential predictors can be found in the GitHub repository under “WildfirePredictors.pdf”.

After an initial exploratory analysis, we chose to utilize latitude and longitude over the state classification as our location parameter in our model. Our decision was largely motivated by the fact that latitude and longitude would capture the same information we would get from the state variables without having a large number of dummy variables that would come with a categorical variable with 50 levels to represent each state. This in turn led to a more parsimonious model with better results. In addition to latitude and longitude, we chose to use the continuous variable “FIRE_SIZE” over “FIRE _SIZE_CLASS” because grouping into classes we risked losing vital information to predicting the cause of the fire.

Model Selection

This problem falls into the category of multiclass predictive modeling. We have a variety of model types to choose from each with their own advantages and disadvantages. Our choice was to select the best model by fitting the data to numerous algorithms capable of multiclass prediction. Then following the utilization of numerous model validation techniques we selected our choice for the most appropriate model. We tested seven different model types and found that ensemble extra trees classification, random forest classification, and decision tree classification had the most favorable results. The remaining quadratic discriminant analysis, logistic classification, ridge classifier and Bayesian classification methods proved to be less capable of prediction given our variables of interest. We will discuss the three most favorable methods here. A lengthier discussion can be found in the potential models document located in the aforementioned GitHub repository.

Decision Trees

We start with the discussion of decision trees because they are the foundation for the next two methods and it is vital that the reader understands this method first. The decision tree algorithm has gained popularity for its ability to fit complex datasets, (1). One can think of a flow chart where as the model takes a data point through it follows a series of binary questions until it results in a selected class, (1). There is a large chance for overfitting/bias when using this method, (1). Some methods use techniques to control over fitting such as in our next two models, one of which ended up with the leading predictive capability, (13).

Random Forest

This is one of the most powerful machine learning methods, (1). Think of random forests as developing multiple trees which all make a recommendation for a particular class selection for a data point, (1). Then, the class with the most votes is what the data point is where the data point is assigned, (1). A more statistical explanation is simply this is a bagging method made for decision trees, (1).

Top Pick: Ensemble Extra Trees

This is a particular type of random forest method, (1). Utilizing the ensemble method of either extra trees or random forest attempts to add some additional randomness within the steps of developing the individual trees to reduce the variance. Of course, this can become an issue since it increases bias. The use of averaging can help control the amount of over-fitting, (13).

Model Validation and Results

Prior to model selection, we split the data into a training and test set with a 75/25 percent ratio before testing our training set on different learning algorithms to see which method had the most powerful prediction. We then used k-folds cross validation to account for potential bias that might have occured during the initial splitting of the data. For example, we noticed in our exploratory analysis that the dataset contained several fire sizes with high severity but low frequencies which skewed the data greatly to the right. If the initial split somehow captured more of these fire sizes than the test set did, our results would be skewed and bias as well.

Comparing the prediction results across all methods, we found that the Ensemble Extra Trees algorithm fits the data best for wildfire cause and produced a prediction rate of 44.43%. The k-folds cross validation verifies this result as well, with the average prediction rate to be 44.2%. Although results for Random Forests ranked higher at a prediction rate of 44.8% for the initial test set, k-folds cross validation showed an overall lower rate of 35% average accuracy rate for this model. Plotted heat maps to check and compare model performances also verified that Ensemble Extra Trees performed best, as the highest heat are on the diagonal. As such, we stand by our top pick as the Ensemble Extra Trees.

Conclusions

We ultimately decided that the ensemble extra trees classification algorithm was the best choice for this particular dataset in predicting cause of fire. Throughout our various measures of accuracy and validation this model continued to be the best choice. However, the highest level of predictive capability we were able to attain was 45%, which leaves a lot of variation unexplained. One potential cause behind this is the dataset uses two classes that are quite ambiguous: miscellaneous and missing/undefined. These two classes alone account for 490,528 data points. We believe this omission in data points resulted in some decrease in accuracy since they make up about 26% of the dataset. Since these algorithms' performance depend on well-defined classes, lacking the true fire causes of a large part of this dataset greatly affects the accuracy of our model. However, we still feel that decision-tree based algorithms are best suited to handle this dataset and that if we update the model with a new dataset containing more well-defined classes and variables, the results would improve.

Future Research

Although cause of fire is an interesting variable to predict as a measure of wildfire prevention, another interesting variable to predict is fire size. Should a fire still begin even with prevention policies in place, which is not an unusual case given that one of the main cause of fire in our analysis are debris from campfires, being able to estimate the fire size can serve as a second measure of prevention. If a fire is predicted to be significantly devastating once started, fire prevention authorities can make the appropriate preparations for a quick response to curb the fire from spreading.

For cases where predicting fire size is the main focus, including a time variable (like day, month, and year) would likely result in higher precision than our model due to the presupposition that wildfires tend to have seasonalities. For instance, if California experienced dry seasons in 2012-2014 that led to more brushfires than previous years, the likelihood of a drought is high (as is the case for California from 2012-2017) and the model might suggest this pattern to carry on into 2015. This added information can increase risk prediction and further improve prevention measures.

We understand that our model lacks this prediction ability and that a spatio-temporal regression model is often used in this type of fire prediction/analysis, (9). However, we did not include this method in our project since such models were not covered over the course of our program, and we reserve this as a future research opportunity.

Sources

1. Géron, Aurélien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. OReilly, 2018.
2. "Precison Recall" http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
3. .Press, Associated. “The Costs to Fight the Deadly Wildfires in the West Are Spiraling out of Control.” Business Insider, Business Insider, 15 Oct. 2017, www.businessinsider.com/ap-us-states-struggle-to-pay-spiraling-cost-of-fighting-fires-2017-10.
4. “Sklearn.tree.DecisionTreeClassifier.” *Scikit-Learn*, scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html.
5. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
6. “Sklearn.linear_model.LogisticRegression.” *Http://Scikit-Learn.org/*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.
7. “Sklearn.linear_model.RidgeClassifier.” *Scikit-Learn*, scikit-learn.org/stable/modules/generated/sklearn.linear_model.RidgeClassifier.html.
8. Tatman, Rachael. “1.88 Million US Wildfires | Kaggle.” Countries of the World | Kaggle, 13 Sept. 2017, www.kaggle.com/rtatman/188-million-us-wildfires.
9.Taylor, S. W., et al. “Wildfire Prediction to Inform Fire Management: Statistical Science Challenges.” Statistical Science, vol. 28, no. 4, 2013, pp. 586–615., doi:10.1214/13-sts451.
10. “1.2. Linear and Quadratic Discriminant Analysis.” *Scikit-Learn* , scikit-learn.org/stable/modules/lda_qda.html.
11. “1.9. Naive Bayes¶.” *Scikit-Learn*, scikit-learn.org/stable/modules/naive_bayes.html.
12. “3.2.4.3.1. Sklearn.ensemble.RandomForestClassifier.” *Scikit-Learn*, scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
13. “3.2.4.3.3. Sklearn.ensemble.ExtraTreesClassifier.” *Scikit-Learn*, scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html.