**Wildfire Occurrences**

Gary Chen, Seth Dogbey, Richard Romero, Matthew Smith, Daniel Stalica

IST-707: Group 1

**Introduction**

Wildfires are a form of uncontrolled natural disaster that have the potential to cause great devastation to communities across the United States. The goal of this project was to utilize historical wildfire data to identify the attributes of a fire that cause them to be the most destructive. This will allow for recommendations to limit fire occurrences and the devastation caused.

The data utilized in this project was published by Karen Short through the United States Department of Agriculture in 2021. The dataset is stored in an open database containing 2,166,753 records with 37 attributes (defined in appendix). Each record in this dataset represents an individual fire that occurred in the United States between 1992 and 2018.

Through various data mining techniques our group developed analytical models to answer the following questions:

* What seasons generate the highest volume of wildfires?
* Which causes are more common with each season?
* What factors play a role in the size of predicting wildfires?
* Is there any correlation between the cause of the fire and the destruction level?
* Do certain causes lead to longer containment periods?

Answering these questions (and any others that arises during EDA) allowed us identified if there were a correlation between attributes and use further machine learning algorithms to develop predictive models to predict causes and severity levels of wildfires.

**Data Prep**

Data preparation involved reading Access database into Alteryx to do some trimming of the dataset. Filtering of the dataset was done to get key attributes that included fire sizes greater than 1 acre and absence of nulls. The dataset was then trimmed into R and additional attributes to be used were generated. These additional attributes included resolution days, size of bin, resolution bin and season bin.SIZE\_BIN <- cut(fires$FIRE\_SIZE, breaks = c(0,25,50,100,200,400,800,1600,3200,6400,Inf), labels = c("1-25","25-50","50-100","100-200","200-400","400-800","800-1600","1600-3200","3200-6400",">6400"))fires$SIZE\_BIN <- SIZE\_BIN *# Bin Resolution Days*RESOLUTION\_BIN <- cut(fires$RESOLUTION\_DAYS, breaks = c(-1,0,7,14,28,56,112,224,448,896,Inf), labels = c("< 1","1-7","7-14","14-28","28-56","56-112","112-224","224-448","448-896",">896"))fires$RESOLUTION\_BIN <- RESOLUTION\_BIN *# Bin Season*SEASON <- cut(fires$DISCOVERY\_DOY, breaks = c(0,79,162,265,355,365), labels = c("Winter", "Spring", "Summer", "Fall", "Winter"))fires$SEASON <- SEASON

**Decision Tree Models**

One of the models that we are using to help find ways to prevent future fires is Decision Trees. The data set is still exceptionally large, so we trimmed the data set down further for Decision Tree modeling. The shuffle\_fire variable will be stored every 10th row in our data set and that sampling is split into training and testing data. We will be using the Trees to understand fire seasonality.

*#choosing the variables to evaluated for the Seasonality Decision Tree*myVars=c("NWCG\_CAUSE\_CLASSIFICATION", "NWCG\_GENERAL\_CAUSE", "SIZE\_BIN", "RESOLUTION\_BIN", "SEASON")

*#Sample a smaller section of our data set*shuffle\_fire <- fires[c(TRUE,rep(FALSE,9)),]D\_fire\_tree <- shuffle\_fire[myVars]

*# Set control cross validation of 5*control\_value <- trainControl(method ='cv',number = 5) *# Use accuracy Metric as comparison*metric\_value <- "Accuracy"

This will leave about 21 thousand records for our Decision Tree analysis. We used Cause Classification, General Cause, Size Bin, and Resolution Bin columns to predict Season. This will tell us which of those predictor values are most associated with each season. For both of our Decision Trees, Cross validation with 5 folds and accuracy measurement is used in the method.

**Seasonality Decision Tree**

*# Decision Tree*set.seed(5)d\_tree\_model <- train(SEASON ~ ., data = D\_fire\_train, method='rpart', metric=metric\_value, trControl=control\_value, na.action=na.exclude)

print(d\_tree\_model)

## CART ## ## 21164 samples## 4 predictor## 4 classes: 'Winter', 'Spring', 'Summer', 'Fall' ## ## No pre-processing## Resampling: Cross-Validated (5 fold) ## Summary of sample sizes: 16932, 16931, 16931, 16931, 16931 ## Resampling results across tuning parameters:## ## cp Accuracy Kappa ## 0.02516200 0.4177370 0.19543522## 0.04563629 0.3953887 0.17077544## 0.09637391 0.3397739 0.06145058## ## Accuracy was used to select the optimal model using the largest value.## The final value used for the model was cp = 0.025162.

Our Season Decision Tree is pruned with the best cp of 0.025162 and has an accuracy of 41.77%. This is not a high accuracy rate, but it will give us an idea of how the other attributes are set to predict which season the fire is in.

tree <- rpart(SEASON ~., data = D\_fire\_train, cp = 0.02612712)rpart.plot(tree)

Diagram

Description automatically generated

The nodes and rules of our seasonality tree show the causes of fire associated with each season. In the Winter, fires are usually caused by Arson, Incendiarism, Open Burning, or by a Minor. In the Spring, its causes are Undetermined, Power Generation, Transmission, Railroad Operation, Maintenance, Recreation, Ceremony, or Smoking. In the Summer, there are more Natural files, or fires caused by Fireworks, Firearms, Explosives, and vehicle equipment. Fall has no significant predictors, and it may be a slow season for fires.

**Decision Tree for Resolution Bin Sizes**

The next Decision Tree we modeled is using Fire Size Bin and Season to predict how many days it will take to put out the fires. The Resolution tree is set to binary splits for our modeling. This tree should give us an idea of the magnitude and how long it takes to put out a fire for each season. This tree is more accurate than our Seasonality tree with ~80% accuracy.

*#Decision Tree for Resolution Bin sizes*myVars3=c("SIZE\_BIN", "RESOLUTION\_BIN", "SEASON")R\_fire\_train\_bin <- D\_fire\_train[myVars3]set.seed(5)rbin\_tree\_model <- train(RESOLUTION\_BIN ~ ., data = R\_fire\_train\_bin, method='rpart', metric=metric\_value, trControl=control\_value, na.action=na.exclude)

print(rbin\_tree\_model)

## CART ## ## 21164 samples## 2 predictor## 10 classes: '< 1', '1-7', '7-14', '14-28', '28-56', '56-112', '112-224', '224-448', '448-896', '>896' ## ## No pre-processing## Resampling: Cross-Validated (5 fold) ## Summary of sample sizes: 16930, 16929, 16930, 16935, 16932 ## Resampling results across tuning parameters:## ## cp Accuracy Kappa ## 0.001765778 0.8100558 0.2120436## 0.001936660 0.8098668 0.2104966## 0.007366902 0.8079296 0.1779931## ## Accuracy was used to select the optimal model using the largest value.## The final value used for the model was cp = 0.001765778.

rbin\_tree <- rpart(RESOLUTION\_BIN ~ ., data = R\_fire\_train\_bin, method = "class",na.action=na.exclude) *#, cp = 0.001800991)*rpart.plot(rbin\_tree) *#, box.palette = "blue")*

A screenshot of a computer

Description automatically generated with low confidence

The rules associated with the Resolution Tree show not only the Seasonality sizes of fires but also how long it takes to put out a fire. If the fire is between 1-25 acres, it usually takes less than a day to put out the fire. If the fire is between 50-100 or 100-200 acres, and if it was either Winter, Spring, or Fall, it will usually take less than a day to put out the fire. For the same 50-100 and 100-200 acres in the Summer, it will take 1-7 days to get the fire under control. There was also a connection in 200–400-acre fires happening more frequently in the summertime. It will take 1-7 days to put it out. If the fire spreads more than 400 acres, it will take 1-7 days across all seasons.

**Naïve Bayes Modeling**

*# Creating Predictive Models and seeing which one is the best.  
# Naive Bayes Model  
# First using the attributes I think that would be most useful*myVarsNB = c('SEASON', 'STATE', 'NWCG\_CAUSE\_CLASSIFICATION', 'FIRE\_YEAR', 'SIZE\_BIN', 'RESOLUTION\_DAYS', 'DISCOVERY\_TIME', 'NWCG\_GENERAL\_CAUSE','SOURCE\_REPORTING\_UNIT\_NAME')DirtyNBFire = firesDirtyNBFire = DirtyNBFire[myVarsNB]sum(!complete.cases(DirtyNBFire))

## [1] 18490

NBFire <- DirtyNBFire[complete.cases(DirtyNBFire), ]set.seed(123) *#Output is Fire Class size so changing it to a factor*str(NBFire)

## 'data.frame': 263712 obs. of 9 variables:## $ SEASON : Factor w/ 4 levels "Winter","Spring",..: 4 4 4 3 3 1 1 1 2 2 ...## $ STATE : chr "CA" "CA" "CA" "NM" ...## $ NWCG\_CAUSE\_CLASSIFICATION : chr "Human" "Human" "Human" "Natural" ...## $ FIRE\_YEAR : int 2004 2004 2004 2004 2004 2004 2005 2005 2005 2005 ...## $ SIZE\_BIN : Factor w/ 10 levels "1-25","25-50",..: 1 10 10 1 1 1 3 4 1 1 ...## $ RESOLUTION\_DAYS : num 0 15 4 5 0 0 1 1 0 0 ...## $ DISCOVERY\_TIME : int 1200 1415 1618 1712 1405 1145 2200 1520 1625 1343 ...## $ NWCG\_GENERAL\_CAUSE : chr "Recreation and ceremony" "Equipment and vehicle use" "Power generation/transmission/distribution" "Natural" ...## $ SOURCE\_REPORTING\_UNIT\_NAME: chr "Eldorado National Forest" "Eldorado National Forest" "Eldorado National Forest" "Lincoln National Forest" ...

**NB Model**

Naive Bayes: This algorithm is one of the easiest algorithms because it believes that all the variables are naive (not correlated with each other). The reason for the specific model we decided to go towards the naive bayes model instead of the other to one see the base accuracy of our dataset and understand which attributes have a significant importance when predicting the size of the fire.

In this specific model, our outcome variable is the fire size bin we created when preparing the data to create models. The attributes that we decided to go best when creating this naive bayes model were Resolution Days, The Human Causes, Season, State, Discovery Time, the classification of the cause and the Year the fire started.

To get the model running it was important to change all the variables to factor in the carret package to run the algorithm. After doing so, I made sure to play around with tuning the algorithm and changing the number of times the algorithm folds. The 3 folds proved to be the best out of the three and is presented in the report. The NB algorithm ran in approximately 3 minutes and 2 seconds which is extremely fast for over 200,000+ data points.

***## Creating the Naive Bayes Model*** *#subsetting the training and testing data*indxTrain <- createDataPartition(y = NBFire$SIZE\_BIN,p = 0.75,list =FALSE) <- NBFire[NBFiretesting <- NBFire[-indxTrain,]  
  
prop.table(table(NBFiretraining$SIZE\_BIN)) \* 100

*##*   
*## 1-25 25-50 50-100 100-200 200-400 400-800 800-*1600

ModelPerformance <- varImp  
NB3)plot(ModelPerformance)

Text

Description automatically generated with medium confidence

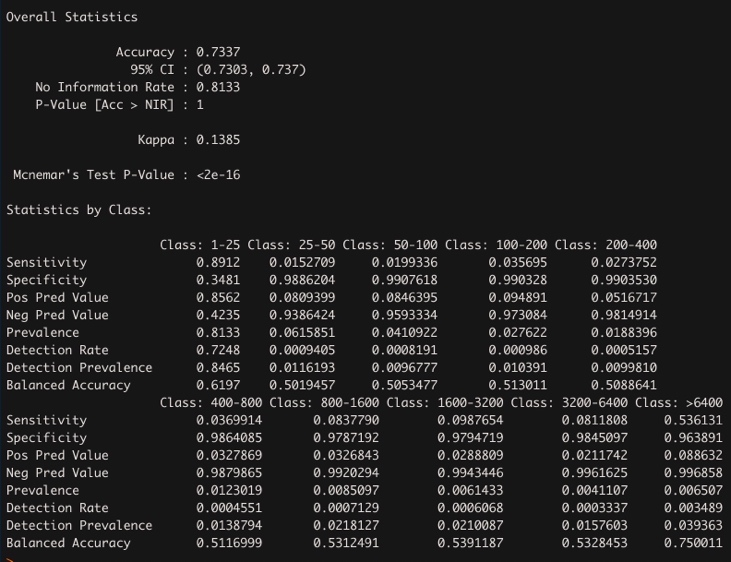
The results showed that both the true positives and true negatives for each fire acre bin size is around 73+ % accuracy for many of the bins. This means at a glance; the model is not particularly strong because the no information rate is higher than the total accuracy of the model. But there is valuable information in understanding what variables within the model have the most impact and what variables are doing extraordinarily little to impact the accuracy of the model.

The no information rate for this model is alarmingly high. This made me search what and conducting the varimp command in r allows you to see what had the highest impact in the model. The top two impactful variables were resolution days, Human Causes and Seasonality. Resolution days accounted for ~35%, Human Cause around ~14% and Season ~4%. This helps for future work with this model is to incorporate different data and left join them onto variables or that can have the same impact if we compare the ability

***### Plotting***  
cm <- confusionMatrix(factor(NBPredict3), factor(NBFiretesting$SIZE\_BIN), dnn = c("Prediction", "Reference"))  
  
plt <- as.data.frame(cm$table)  
plt$Prediction <- factor(plt$Prediction, levels=rev(levels(plt$Prediction)))  
  
ggplot(plt, aes(Prediction,Reference, fill= Freq)) +  
 geom\_tile() + geom\_text(aes(label=Freq)) +  
 scale\_fill\_gradient(low="green", high="#009194") +  
 labs(x = "Reference",y = "Prediction") +  
 scale\_x\_discrete(labels=c(">6400","3200-6400","1600-3200","800-1600","400-800","200-400","100-200","50-100","25-50","1-25"))

A picture containing calendar

Description automatically generated



This graph shows how many errors the best of the three models I constructed. The model with the highest frequency of true positives happens to be the first bin 1-25 which makes logical sense because of how many fires happen within a small radius in comparison to the bigger ones. Just because a bin has a lot of observations and true positive percentage does not mean that it is more accurate. If you look at acres of land greater than 6400 than you see it has the highest accuracy overall, but it has the fewest observations within the dataset so maybe if we had more data points, we could see the true accuracy in comparison to class 1-25.

scale\_y\_discrete(labels=c("1-25","25-50","50-100","100-200","200-400","400-800","800-1600","1600-3200","3200-6400",">6400"))

**Random Forest**

***### Running RF Model to compare to Naive Bayes Results***  
  
RFModel3 <- train(SIZE\_BIN ~ . ,   
 data = RFFiretraining[1:10000, ], *# Use the train data frame as the training data*  
 method = 'rf',*# Use the 'random forest' algorithm*  
 trControl = trainControl(method = 'cv', *# Use cross-validation*  
 number = 3)) *# Use 3 folds for cross-validation*  
  
RFModel5 <- train(SIZE\_BIN ~ . ,   
 data = RFFiretraining[1:10000, ], *# Use the train data frame as the training data*  
 method = 'rf',*# Use the 'random forest' algorithm*  
 trControl = trainControl(method = 'cv', *# Use cross-validation*  
 number = 5)) *# Use 5 folds for cross-validation*

Running the random forest to compare if this model is better than the previously run naïve bayes model was the next problem at hand. First it was important to change some of the variables back into their numerical form because when running a decision tree or random forest in R you cannot exceed a certain number of factors in the model. Then, after attempting to run a random forest with the entire sample dataset we originally had, I clearly noted that the model would take way too many hours to give me a concrete answer. Then I decided to cut the data into a smaller sample size.

After cutting the sample size to 10,000 data observations, I wanted to run the carret package because when you run the model, it runs with three different tree options depending on the number of folds you put into the model. There were three different numbers of tree options when running the algorithm with varying accuracy rates. I noticed that the creation of two trees and using them to predict the bins happens to be the best accuracy providing a 77% accuracy rate.

I would conclude that the random forest is slightly better than the naïve bayes model, but the random forest model ended up taking at least 12 and a half minutes to run compared to running three-minute models with naïve bayes which can be easily tuned and is only a few percentages lower in accuracy. If you’re looking for accuracy, random forest beats it but speed the naïve model dusts the random forest.

**Assocation Rule Minning**

*# generate rules*  
rules <- apriori(fires, parameter = list(supp = 0.1, conf = 0.9, maxlen = 3))

## Warning in apriori(fires, parameter = list(supp = 0.1, conf = 0.9, maxlen = 3)):  
## Mining stopped (maxlen reached). Only patterns up to a length of 3 returned!

*#Plotting rules*  
plot(rules)

Chart

Description automatically generated

*#Subsetting some relevant columns for association rule mining; general cause and fire size\_bin*  
firesAssoc<-fires[,c(24,39)]  
   
*# generate rules*  
rulesAssoc <- apriori(firesAssoc, parameter = list(supp = 0.001, conf = 0.8, maxlen = 3))

*#Viewing the rules generated*  
inspect(rulesAssoc)

## lhs rhs support confidence coverage lift count  
## [1] {} => {SIZE\_BIN=1-25} 0.813644127 0.8136441 1.000000000 1.0000000 229612  
## [2] {NWCG\_GENERAL\_CAUSE=Other causes} => {SIZE\_BIN=1-25} 0.003905004 0.8718354 0.004479061 1.0715194 1102  
## [3] {NWCG\_GENERAL\_CAUSE=Railroad operations and maintenance} => {SIZE\_BIN=1-25} 0.007891510 0.8309701 0.009496743 1.0212943 2227  
## [4] {NWCG\_GENERAL\_CAUSE=Misuse of fire by a minor} => {SIZE\_BIN=1-25} 0.015379055 0.9475983 0.016229509 1.1646348 4340  
## [5] {NWCG\_GENERAL\_CAUSE=Smoking} => {SIZE\_BIN=1-25} 0.015527884 0.9055590 0.017147292 1.1129669 4382  
## [6] {NWCG\_GENERAL\_CAUSE=Power generation/transmission/distribution} => {SIZE\_BIN=1-25} 0.017104769 0.8129000 0.021041665 0.9990854 4827  
## [7] {NWCG\_GENERAL\_CAUSE=Recreation and ceremony} => {SIZE\_BIN=1-25} 0.031356971 0.8623075 0.036364023 1.0598092 8849  
## [8] {NWCG\_GENERAL\_CAUSE=Equipment and vehicle use} => {SIZE\_BIN=1-25} 0.067210721 0.8170501 0.082260225 1.0041860 18967  
## [9] {NWCG\_GENERAL\_CAUSE=Missing data/not specified/undetermined} => {SIZE\_BIN=1-25} 0.146434115 0.8041097 0.182107143 0.9882818 41324  
## [10] {NWCG\_GENERAL\_CAUSE=Debris and open burning} => {SIZE\_BIN=1-25} 0.266582094 0.8880049 0.300203400 1.0913923 75230

The aim of the association rule mining, using the apriori algorithm, was to determine whether there was a correlation between the cause of the wildfires and the size of the fires (discretized into bins). Initial parameters setting, confidence of 0.9 and support of 0.1, with a rule length of 3 returned a total of 2861 rules using relevant attributes from the data. These rules had a few of them above a lift of 2.5.

A subset of two attributes, cause of fires and fire sizes, was done and the association rules using a confidence of 0.8 and a support of 0.001 was also executed. The parameters were reduced because initial results did not produce enough rules to draw any meaningful conclusions.

The resultant association rules did not give any indication of correlation since all right-hand side had the same attribute value; that is “SIZE\_BIN=1-25". Since we wanted to see the association of cause of fires to the fire size, it is difficult to determine a correlation between them because all left-hand side values were associated with the same right-hand value.

**Clustering**

*#Cut down size of the data*  
Wildfires <- FireDataV2  
WildfiresFinal <- Wildfires[1:2000,]  
*#Fire Cause Labels*  
firecause.labels = WildfiresFinal$NWCG\_GENERAL\_CAUSE  
table(firecause.labels)  
firecause.data <- data.frame(WildfiresFinal$DISCOVERY\_TIME, WildfiresFinal$CONT\_TIME, WildfiresFinal$FIRE\_SIZE)  
*#Calculating how many clusters needed*  
*#With the elbow method. The sum of squares.*  
fviz\_nbclust(Wildfire.scale, kmeans, method = "wss") +  
 labs(subtitle = "Elbow Method")

Chart, line chart

Description automatically generatedChart

Description automatically generated

*#Kmeans Clustering*  
km.fires <- kmeans(Wildfire.scale, centers = 8, nstart = 100)  
*#print(km.fires)*

*#Visualize the clustering results*  
kmcluster.fires <- km.fires$cluster  
rownames(Wildfire.scale) <- paste(WildfiresFinal$NWCG\_GENERAL\_CAUSE, 1:dim(WildfiresFinal)[1], sep = "-")  
*#Wildfire.scale*  
fviz\_cluster(list(data = Wildfire.scale, cluster = kmcluster.fires))

*#Look at clusters as a crosstab or datatable*  
*#Cut down size of the data*  
Wildfires <- FireDataV2  
WildfiresFinal <- Wildfires[1:2000,]  
  
*#Create a containment time variable*  
WildfiresFinal$Resolvetime <- WildfiresFinal$CONT\_DOY -   
 WildfiresFinal$DISCOVERY\_DOY

Some descriptive analysis was done to see results that showed an average discovery time of 14 hours and 56 minutes, average fire size was 262 acres and average containment time was 15 hours 51 minutes. This preceded the question of whether we can use unsupervised learning to cluster these factors based on fire cause. The Kmeans algorithm was employed with a parameter setting of 8 centers. The was an assumption made for the number of clusters due to lack of enough domain knowledge.

The results produced an 82.7% measure of the total variance in the data set that is explained by the clustering. No one cause was truly contained in a cluster and therefore unable to attribute fire size, containment time or discovery time to a specific fire cause.

**Conclusion and Recommendations**

Mapping the decision trees gave us a lot of insight on Seasonality. The most interesting Seasonality cause association is Spring with Power Generation and Maintenance. It seems like there are various infrastructure breakdowns in the Winter that can cause some of these fires in the Spring. There should be better weather proofing and maintenance of our infrastructure to avoid some of these fires in the Spring. It is intuitive that the Summer will have more severe and longer fires. It is good that we have an accurate model to confirm this. We should avoid having a longer summer season to combat fewer intensive fires. This can be done by combating Climate Change.

The area that we were able to draw the most meaningful conclusions were based on seasonality. This makes intuitive sense as the weather is a key variable in wildfires. Human Causes are shown to have the second largest impact on the predictability of the fire sizes. Humans cause 82% of all wildfires in the dataset having Debris and Open Burning making up 30% of this cause of wildfire.

The Association Rule mining, which was the closest model needed to determine correlation between cause of fires and the size of fire, did not produce any correlation between the two variables selected. This might be due to the absence of any correlation between those two variables.

No one cause was truly contained in a cluster when the clustering technique, using the Kmeans algorithm, was employed. We were therefore unable to attribute fire size, containment time or discovery time to specific fire cause. Domain knowledge is seen as a key factor here especially using the Kmeans algorithm and therefore expertise in wildfires is recommended for modeling and predictions.

In predicting fire size, the random forest was seen to be slightly better than the naïve bayes model, but it ended up taking more time to run compared to naïve bayes, which can be easily tuned and is only a few percentages lower in accuracy.

**Data Citation**:

Short, Karen C. 2021. Spatial wildfire occurrence data for the United States, 1992-2018 [FPA\_FOD\_20210617]. 5th Edition. Fort Collins, CO: Forest Service Research Data Archive. <https://doi.org/10.2737/RDS-2013-0009.5>

**Data Link**:

<https://www.fs.usda.gov/rds/archive/Catalog/RDS-2013-0009.5>