Fundamentals of Data Science: Prediction, Inference, Causality

MS&E 226

Predict the size of wildfires
Part 2

Author:
Thibaut Badoual
Melvin Juwono

Fall 2020

1 Prediction

Taking our final OLS model for the regression task in Part 1 of the project, and applying it to the test set, we got a **RMSE value of 7061**. This error falls slightly below the error we received when the model was applied to our training set (RMSE = 7898). As the errors are of the same magnitude, this could potentially be due to the element of randomness, or the dataset not having enough data. Note that despite having a total of 18,000+ data entries, the values for some variables are not uniformly distributed. For example, certain causes of fire such as debris burning, arson, and lightning, show up substantially more frequently than others such as fireworks and powerline accidents (shown in Figure 4). For the less frequently observed covariates, the lack of data may be affecting our train and test set errors.

Next, taking the model we used for the classification task in Part 1, we see an increase in accuracy when applying the test set relative to results from the train set. This follows the trend of the test set doing better than the train set as observed with our final OLS model. However, the measurement we are most concerned with is sensitivity, which decreased slightly when the test set was applied (77.0%) compared to when the train set was applied (77.7%). We believe the explanation of our observations is similar to what we observed with the regression task; possibly due to the element of randomness or the dataset not having sufficient data.

2 Inference

2.1 Statistical Significance

We chose to focus on our final OLS model (for the regression task) for this part of the project. Applying the model on the training data, we were able to obtain the regression table as shown in Figure 1. The coefficients we believe have **statistical significance at the 95% level** are:

- **stat_cause_descrCampfire** (Campfires as the cause of fire)
- **stat_cause_descrEquipment** Use (Equipment use as the cause of fire)
- **stat_cause_descrLightning** (Lightning as the cause of fire)
- **longitude** (Longitude)
- **stateFL** (State of Florida as the origin of the fire)
- **Prec_pre_30** (Precipitation in the area 30 days prior to the fire)
- **Temp_pre_30** (Temperature of area 30 days prior to fire)
- **Wind_pre_30:Hum_pre_30** (Humidity in the area 30 days prior to the fire given Wind in the area 30 days prior to the fire)
- **Wind_pre_7:Hum_pre_7** (Humidity in the area 7 days prior to the fire given Wind in the area 7 days prior to the fire)
- **Temp_pre_30:Hum_pre_30** (Humidity in the area 30 days prior to the fire given Temperature in the area 30 days prior to the fire)

The coefficients mentioned above have p-values below 0.05, which means that the probability of observing a coefficient value as extreme as the one observed if the null hypothesis were true is below 5%. With the t-test, we reject the null hypothesis. Note that in this case, the null hypothesis is the coefficient being equal to 0 (i.e. the covariate is not correlated with fire size).

In other words, assuming that the cause of fire was either campfires, equipment use, or lightning, it is unlikely that the fire size is not correlated with the cause of fire. Similarly, it is unlikely that we observe a coefficient as extreme as the coefficients for longitude, precipitation 30 days prior, and others noted, if these coefficients were actually 0.

We believe in most of the results that indicate statistical significance. Often it is difficult to predict when a fire will start due to an accident caused by campfires or equipment use, and lightning can also be unpredictable, so this may affect response times and hence fire size. It is also possible that longitude and meteorological data affect fire size as these factors can contribute to heat needed to start and grow a fire. However, we are also wary that vegetation in the area of the fires is not statistically significant. The same can be said for season (or month the fire started in). For example, we hypothesized that areas highly dense with vegetation will fuel fires to boost fire size or fires during the summer may be greater in size than during the winter because of variations in heat. That being said, with regards to seasons, results may be affected by strong correlations (collinearity) with meteorological data.

When the model is fit on the testing data, we observe numerous changes regarding which coefficients are significant as shown in the regression table (Figure 2). Amongst those we previously determined to be statistically significant, only stat_cause_descrLightning (lightning as the cause of fire) remains significant. Differences may be due to the element of randomness affecting p-values in the regression table and indicating statistical significance by chance - a systematic issue in multiple hypothesis testing that comes from allowing a 5% rate of false positives. For other covariates, this effect may be exacerbated by a lack of data, such as campfires as the cause of fire (note the relatively low number of fires started by campfires in our dataset shown in Figure 4).

2.2 Confidence Intervals

Using **bootstrap** to estimate confidence intervals, we were able to compare the confidence intervals with results from the standard regression output (Table 1). Overall, the two sets of confidence intervals were within the same order of magnitude suggesting that for many of the covariates, **there is enough data such that bootstrap resampling has little effect**. However, particularly prevalent in covariates such as stat_cause_descrCampfire (campfire as the cause of fire) where there is a low number of data points, the confidence interval from bootstrap can differ significantly. An additional point to note, is that the confidence intervals for many covariates are large relative to the estimate. We believe this is primarily due to collinearity.

As our chosen model does not include all covariates we had available, we constructed a regression table using a model that includes all covariates. We observe many similarities in the covariates that are statistically significant such as equipment use and lightning as the cause of fire, as well as longitude and some meteorological data. Changes include more states where the fire originated from being statistically significant, in addition to winter season and remoteness also being statistically significant. We believe the only notable change is remoteness being statistically significant. Early in the project, we dropped remoteness as a variable because of the ambiguity around how the remoteness metric was constructed. However, this may be an example of omitted variable bias and an indication that we should not have dropped this variable.

We believe that some level of **collinearity** is impacting our results. Specifically certain pairs of meteorological data are highly correlated. For example, humidity and precipitation are highly correlated; the area's humidity signals the likelihood precipitation is present. Furthermore, longitude and seasonality are correlated with meteorological data. For example, as longitude changes from the middle of North America to the coasts, wind and precipitation changes. And as seasons change, temperature, humidity, precipitation, and wind are all correlated.

2.3 Multiple hypothesis testing

We also believe **multiple hypothesis testing** is impacting our inference. Including the various causes of fires, the top ten states selected, and interaction terms, there are a total of 44 covariates being tested. As we use a cutoff of 5%, we allow a 5% rate of false positives which applies across all the hypothesis tests carried out. Hence, we expect this to affect our inference, noting that some covariates interpreted as statistically significant will be falsely interpreted. Applying the Bonferroni correction, we reconsider the statistically significant covariates mentioned above and only declare it statistically significant if the p-value is below 0.05p, where p is the number of hypotheses tests being conducted. Doing so, leaves the following as statistically significant:

- **stat_cause_descrLightning** (Lightning as the cause of fire)
- **Wind_pre_30:Hum_pre_30** (Humidity in the area 30 days prior to the fire given Wind in the area 30 days prior to the fire)
- **Temp_pre_30:Hum_pre_30** (Humidity in the area 30 days prior to the fire given Temperature in the area 30 days prior to the fire)

It is worth noting that the Bonferroni correction might be too conservative. As we would prefer to be on the safe side, we may not necessarily want to decrease the false positive rate in favor of an increase in the false negative rate.

For the relationships we found to be statistically significant, we would not be willing to interpret them as causal relationships as we lack information on alternative actions. Taking "lightning as the cause of fire" as an example, we do not have data on situations where lightning was present but did not start the fire and therefore cannot compare outcomes to determine causation. Furthermore, some covariates are highly correlated with one another and so, as they are not independent, cannot be interpreted causally. Sticking to our example, lightning is correlated with precipitation and hence not completely independent.

Lastly, it is also possible that we have omitted variable bias that is affecting our results. One variable we mentioned earlier is remoteness. Other variables we considered include the amount of resources deployed to fight the fire, distance of fire from sensitive sites that may fuel the fire, and more. These variables may be critical to explaining fire size or may be critical in supporting other covariates to explain fire size. Without these relevant variables, results would be biased.

3 Discussion

We expect our models to be used as a **decision support tool**, providing guidance regarding estimated fire size so that the necessary parties can be informed and prepare / act in a manner to minimize fire damage. A graphic of the top 20 largest California Wildfires (Figure 5), indicate this problem is more prevalent than ever. We think that the models will be used primarily for prediction to help firefighters respond appropriately to a fire; a simple user interface can be developed that allows a user to enter known parameters, and output estimated fire size. That being said, our models could be used for inference as city planners may be interested in understanding which factors to focus on when attempting to mitigate large fires. One of the more salient pitfalls of using our models is the impact it has on response times. A fast response time is important when containing fires, and to use our models, time is needed to collect all the relevant data and input them into the models to get a prediction.

The models will need to be updated every couple of years as there are variables whose relationship with fire size may evolve over time. We are less concerned about the meteorological covariates as we believe any changes in their relationships with fire size will happen over a very long period of time, however, covariates such as cause of fire and state the fire originated in, might see changes in their relationship with fire size over a shorter period of time. For example, camp-sites in a particular state may have updated rules and regulations that reduce the potential size of a fire started by campfires.

A choice we made in our data analysis that we would want users to know is that we dropped the remoteness metric, which was not a standardized measure (it was created by the author of our dataset). The purpose of the remoteness metric was to provide a sense of how close the nearest city was from the fire, which could be correlated with firefighter response times or amount of resources deployed to contain the fire. Our models dropping this column is important to note as this means the models assume that all the fires were fought in the same manner. The reality is quite different but we would need additional data to provide meaningful insights.

If we were able to change the data collection process, we would choose to collect a **"mean deployed" variable**, which will be a measure of the average amount of resources (e.g. firefighters) deployed to fight the fire. We can also add information regarding distances between the fire and sensitive sites (e.g. nuclear power plants, paper mills, etc.), major traffic routes, and other factors that could add fuel to a fire.

On the other hand, if we were to attack the same dataset again, we would reduce the number of categories per categorical covariate. We attempted to do this with the number of states we considered when we only considered the top ten states that had the most wildfires recorded, yet we failed to consider the number of categories in the cause of fire for example. As shown in Figure 4, there are some causes of fire with too few data points for us to extract reasonable insights from. We would also focus more on determining which covariates are highly correlated with one another and attempt to **mitigate collinearity**.

4 Appendix

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-8264.23	4376.67	-1.888	0.059013	.
stat_cause_descrCampfire	1662.60	540.35	3.077	0.002096	**
stat_cause_descrChildren	-361.89	470.77	-0.769	0.442073	
`stat_cause_descrDebris Burning`	-207.15	206.48	-1.003	0.315759	
`stat_cause_descrEquipment Use`	-678.52	300.95	-2.255	0.024172	*
stat_cause_descrFireworks	-1076.23	2336.69	-0.461	0.645109	
stat_cause_descrLightning	1443.32	331.65	4.352	1.36e-05	***
stat_cause_descrMiscellaneous	108.90	256.86	0.424	0.671607	
`stat_cause_descrMissing/Undefined`	670.61	314.16	2.135	0.032809	*
stat_cause_descrPowerline	-449.89	721.19	-0.624	0.532759	
stat_cause_descrRailroad	-484.94	530.13	-0.915	0.360333	
stat_cause_descrSmoking	-316.48	512.88	-0.617	0.537195	
stat_cause_descrStructure	-567.67	2427.11	-0.234	0.815077	
longitude	-101.37	46.88	-2.162	0.030617	*
stateAR	-747.05	497.73	-1.501	0.133397	
stateCA	-190.76	1513.35	-0.126	0.899692	
stateFL	1069.78	439.13	2.436	0.014857	*
stateGA	439.01	347.80	1.262	0.206880	
stateMS	-348.41	341.75	-1.019	0.307991	
stateNC	157.85	572.72	0.276	0.782855	
stateOK	-1110.64	582.70	-1.906	0.056669	.
stateSC	529.74	475.31	1.115	0.265078	
stateTX	-768.23	560.34	-1.371	0.170393	
`VegetationOpen Shrubland`	-20.71	1053.21	-0.020	0.984309	
`VegetationPolar Desert/Rock/Ice`	551.99	1028.03	0.537	0.591318	
`VegetationSecondary Tropical Evergreen Broadleaf Forest`	-230.89	1038.06	-0.222	0.823983	
seasonspring	-306.29	263.83	-1.161	0.245692	
seasonsummer	-92.44	300.53	-0.308	0.758386	
seasonwinter	-471.24	255.88	-1.842	0.065546	.
Prec_pre_30	-360.83	156.01	-2.313	0.020744	*
Prec_pre_15	179.31	182.83	0.981	0.326714	
Prec_pre_7	152.54	118.90	1.283	0.199516	
Temp_pre_30	-1154.62	398.35	-2.899	0.003755	**
Temp_pre_15	646.33	504.27	1.282	0.199961	
Temp_pre_7	282.70	305.68	0.925	0.355083	
Wind_pre_30	255.67	243.07	1.052	0.292900	
Wind_pre_15	-187.38	294.34	-0.637	0.524396	
Wind_pre_7	277.10	180.24	1.537	0.124216	
Hum_pre_30	63.23	218.36	0.290	0.772155	
Hum_pre_15	-336.95	272.90	-1.235	0.216962	
Hum_pre_7	-92.58	177.79	-0.521	0.602551	
`Wind_pre_30:Hum_pre_30`	-568.52	141.34	-4.022	5.79e-05	***
`Wind_pre_15:Hum_pre_15`	338.11	172.97	1.955	0.050628	.
`Wind_pre_7:Hum_pre_7`	-304.08	117.16	-2.595	0.009459	**
`Temp_pre_30:Hum_pre_30`	641.10	174.85	3.667	0.000247	***

Figure 1: Regression table when model is fit on train data

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-7288.93	7512.86	-0.970	0.3320	
stat_cause_descrCampfire	1021.14	891.65	1.145	0.2522	
stat_cause_descrChildren	-218.43	856.15	-0.255	0.7986	
`stat_cause_descrDebris Burning`	-144.58	345.00	-0.419	0.6752	
`stat_cause_descrEquipment Use`	-163.49	515.35	-0.317	0.7511	
stat_cause_descrFireworks	-288.01	2919.37	-0.099	0.9214	
stat_cause_descrLightning	3574.28	567.30	6.301	3.33e-10	*
stat_cause_descrMiscellaneous	411.18	433.27	0.949	0.3427	
`stat_cause_descrMissing/Undefined`	578.45	557.51	1.038	0.2995	
stat_cause_descrPowerline	2256.95	1094.55	2.062	0.0393	*
stat_cause_descrRailroad	35.66	926.08	0.039	0.9693	
stat_cause_descrSmoking	292.78	838.14	0.349	0.7269	
stat_cause_descrStructure	148.12	3559.75	0.042	0.9668	
longitude	-60.71	80.41	-0.755	0.4502	
stateAR	-364.36	849.12	-0.429	0.6679	
stateCA	1556.79	2594.65	0.600	0.5485	
stateFL	-44.48	754.67	-0.059	0.9530	
stateGA	577.83	594.05	0.973	0.3308	
stateMS	-186.52	579.26	-0.322	0.7475	
stateNC	418.48	1000.08	0.418	0.6756	
stateOK	59.21	978.14	0.061	0.9517	
stateSC	149.95	790.46	0.190	0.8496	
stateTX	-409.14	958.26	-0.427	0.6694	
`VegetationOpen Shrubland`	2850.97	1766.38	1.614	0.1066	
`VegetationPolar Desert/Rock/Ice`	2442.79	1730.77	1.411	0.1582	
`VegetationSecondary Tropical Evergreen Broadleaf Forest`	2600.77	1740.38	1.494	0.1352	
seasonspring	-555.28	447.93	-1.240	0.2152	
seasonsummer	-775.54	514.61	-1.507	0.1319	
seasonwinter	-759.97	440.35	-1.726	0.0845	.
Prec_pre_30	-341.61	260.81	-1.310	0.1904	
Prec_pre_15	138.01	307.45	0.449	0.6535	
Prec_pre_7	63.51	201.46	0.315	0.7526	
Temp_pre_30	-805.40	671.03	-1.200	0.2301	
Temp_pre_15	-206.85	836.80	-0.247	0.8048	
Temp_pre_7	949.84	511.53	1.857	0.0634	.
Wind_pre_30	610.18	416.17	1.466	0.1427	
Wind_pre_15	-685.02	497.94	-1.376	0.1690	
Wind_pre_7	291.79	298.92	0.976	0.3291	
Hum_pre_30	-188.13	372.41	-0.505	0.6135	
Hum_pre_15	218.09	462.86	0.471	0.6375	
Hum_pre_7	-303.49	298.59	-1.016	0.3095	
`Wind_pre_30:Hum_pre_30`	-187.12	243.59	-0.768	0.4424	
`Wind_pre_15:Hum_pre_15`	357.16	281.58	1.268	0.2047	
`Wind_pre_7:Hum_pre_7`	-248.60	160.57	-1.548	0.1216	
`Temp_pre_30:Hum_pre_30`	93.18	309.48	0.301	0.7634	

Figure 2: Regression table when model is fit on test data

Table 1: Table comparing results from bootstrap and standard regression output

METHOD USED	OLS		BOOSTRAP	
confidence interval	Estimate	Std Error	Estimate	Std Error
(Intercept)	-2001.83	4542.31	-2001.82960	4959.26287
stat_cause_descrCampfire	1970.44	520.90	1970.43882	1021.25650
stat_cause_descrChildren	-262.89	472.58	-262.88680	155.42038
stat_cause_descrDebris Burning	-216.47	215.62	-216.47301	96.32420
stat_cause_descrEquipment Use	-643.94	315.21	-643.93878	188.52233
stat_cause_descrFireworks	96.43	3292.67	96.43141	284.22971
stat_cause_descrLightning	2236.48	348.87	2236.48320	706.09524
stat_cause_descrMiscellaneous	265.96	260.72	265.95709	261.84143
stat_cause_descrMissing/Undefined	851.08	355.73	851.08227	662.05123
stat_cause_descrPowerline	164.72	728.31	164.72078	553.66634
longitude	-37.77	48.39	-37.77109	46.90503
stateCA	1582.80	1549.36	1582.79817	1544.93068
stateFL	570.11	455.41	570.11443	387.41932
stateGA	398.46	362.27	398.46199	150.67840
stateMN	-1756.57	631.38	1756.57115	1214.98626
stateMS	-182.07	357.00	-182.07099	157.43426
stateNC	-742.88	627.37	-742.87554	1129.43885
stateNY	-1636.23	987.98	-1636.23465	1710.16140
stateSC	110.17	490.49	110.17411	424.64034
VegetationGrassland/Steppe	585.64	1156.92	585.64306	1721.71676
VegetationOpen Shrubland	-603.03	1058.78	-603.03306	2031.63420
VegetationPolar Desert/Rock/Ice	697.65	1028.16	697.64676	1710.61950
VegetationSecondary Tropical Evergreen Forest	-793.08	1044.24	-793.08032	1924.46333
VegetationTemperate Evergreen Forest	624.35	1316.09	624.34811	1734.52934
seasonspring	-564.54	269.79	-564.53786	325.24502
seasonsummer	-451.97	313.42	-451.96695	318.61904
seasonwinter	-568.54	270.70	-568.54336	212.34928
Prec_pre_30	-415.90	165.91	-415.89625	121.86367
Prec_pre_15	220.45	195.89	220.44805	99.31571
Prec_pre_7	130.08	126.66	130.08202	97.48515
Temp_pre_30	-1377.73	421.20	-1377.73394	638.33661
Temp_pre_15	757.89	525.72	757.89101	633.74844
Temp_pre_7	428.17	322.06	428.16602	237.42120
Wind_pre_30	383.81	244.39	383.81204	161.68659
Wind_pre_15	-269.90	293.96	-269.90150	192.31085
Wind_pre_7	291.29	178.13	291.28915	154.89199
Hum_pre_30	56.95	221.70	56.94793	217.68210
Hum_pre_15	-299.22	277.78	-299.21898	180.37782
Hum_pre_7	-97.27	181.13	-97.27094	191.57300
Wind_pre_30:Hum_pre_30	-586.51	143.35	586.50543	200.84426
Wind_pre_15:Hum_pre_15	337.56	172.41	337.56112	222.35057
Wind_pre_7:Hum_pre_7	-295.57	112.23	-295.56796	241.26485
Temp_pre_30:Hum_pre_30	610.09	184.95	610.08530	377.39846
Temp_pre_15:Hum_pre_15	-231.07	241.81	-231.07228	313.81128
Temp_pre_7:Hum_pre_7	-193.28	158.05	-193.28204	151.60024

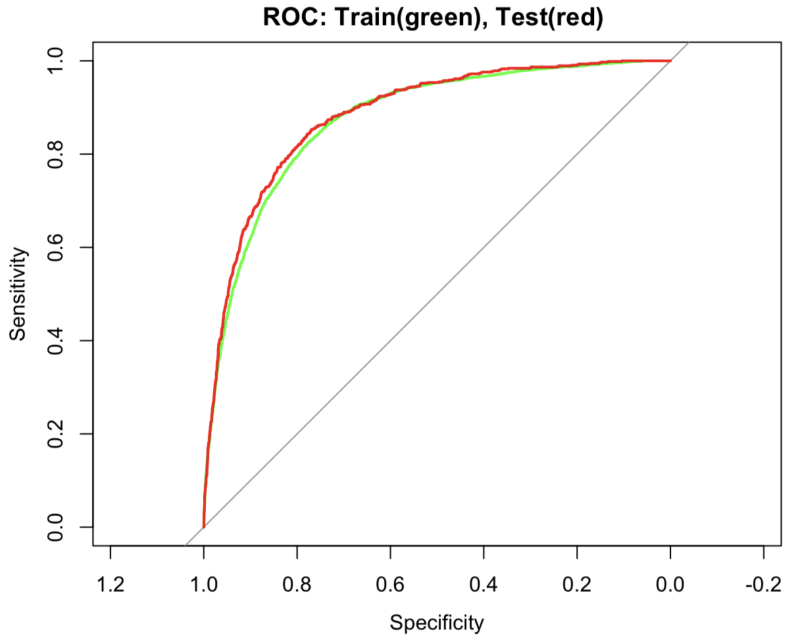

Figure 3: ROC curve

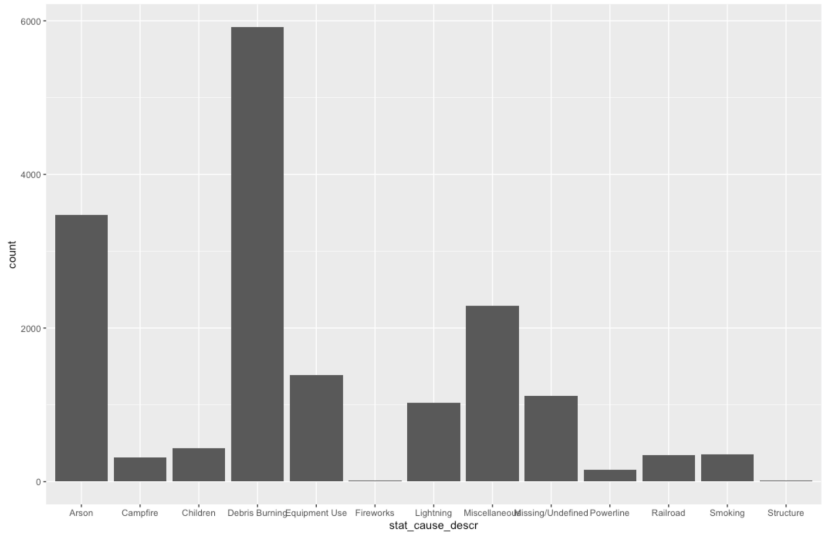

Figure 4: Count vs. Cause of Fire

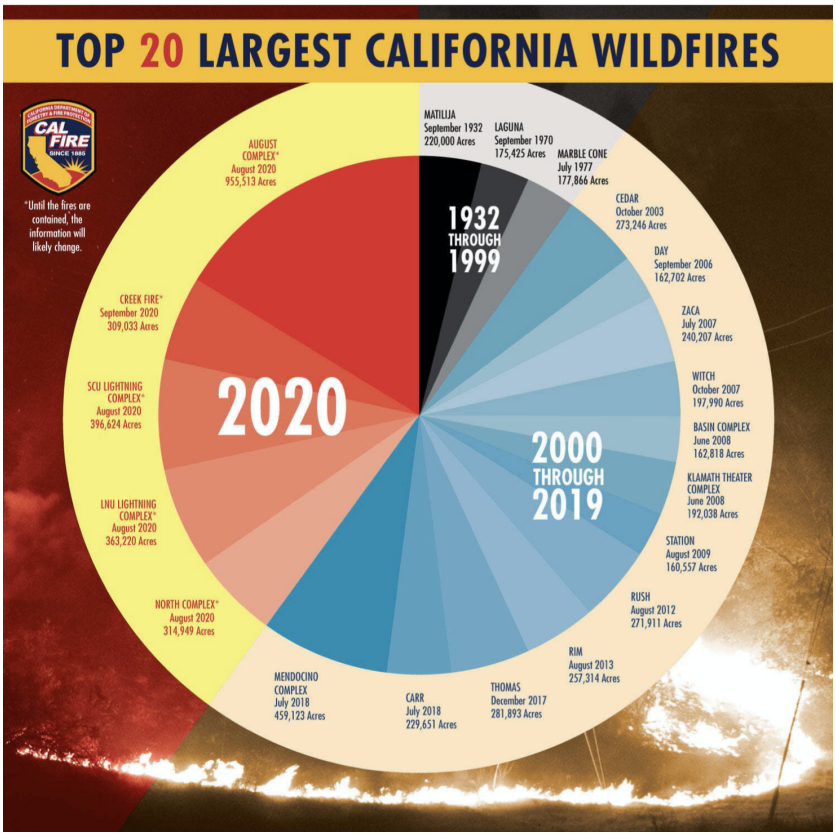

Figure 5: Top 20 Largest California Wildfires

5 Code

```r

# Install packages
install.packages("visdat")
install.packages("corrplot")
install.packages("GGally")
install.packages("gridExtra")
install.packages("naniar")
install.packages("cvTools")
install.packages("glmnet")
install.packages("ISLR")
install.packages("boostrap")

# Load packages
library("ggplot2")
library("knitr")
library("visdat")
library("corrplot")
library("GGally")
library("gridExtra")
library("naniar")
library(dplyr)
library("cvTools")
library(tidyverse)
library(caret)
library(randomForest)
library(kernlab)
library(rpart)
library(neuralnet)
library("ISLR")
library("glmnet")
library(ggplot2)
library(mltools)

# Loading Data
data <- read.csv("/Users/thibautbadoual/Desktop/Aut/MS&E_226/MS&E_226_Project/Fire/FW_Veg_Rem_Combined.csv")
# data <- read.csv("~/Docs/Stanford/MSE226/Project/archive/FW_Veg_Rem_Combined.csv")

#------ Data Cleaning ------------------------------------------------------------------#
# Remove unnecessary columns
data <- subset(data, select = -c(X, Unnamed..0, fire_name, fire_mag,
                                 cont_clean_date, disc_date_final, cont_date_final,
                                 putout_time, disc_date_pre, disc_pre_month,
                                 wstation_usaf, dstation_m, wstation_wban,
                                 wstation_byear, wstation_eyear, weather_file,
                                 Prec_cont, Hum_cont, Wind_cont, Temp_cont,
                                 disc_clean_date, remoteness, disc_pre_year))

# Select only the top ten states
top_ten_states <- group_by(data, state) %>%
  summarise(count = n())
top_ten_states <- top_ten_states[order(-top_ten_states$count),]
top_ten_states <- top_ten_states$state[1:10]
data = filter(data, data$state %in% top_ten_states)

# Change vegetation type
data$Vegetation <- as.character(data$Vegetation)
data$Vegetation[which(data$Vegetation == "4")] <- "Temperate Evergreen Needleleaf Forest"
data$Vegetation[which(data$Vegetation == "9")] <- "Grassland/Steppe"
data$Vegetation[which(data$Vegetation == "12")] <- "Open Shrubland"
data$Vegetation[which(data$Vegetation == "14")] <- "Desert"
```

```r
data$Vegetation[which(data$Vegetation == "15")] <- "Polar Desert/Rock/Ice"
data$Vegetation[which(data$Vegetation == "16")] <- "Secondary Tropical Evergreen
    Broadleaf Forest"

# Remove duplicate rows
data <- data[!duplicated(data), ]

# Drop missing values for vegetation data
data = filter(data, data$Vegetation != 0)

# Drop missing values for meteorological data
data = filter(data, data$Prec_pre_30 != -1.00000)
data = filter(data, data$Hum_pre_30 != 0)
data = filter(data, data$Hum_pre_15 != 0)
data = filter(data, data$Hum_pre_7 != 0)

# Add column with binary response variable (fire size class above C = 1; C or below =
    0)
data_0 = filter(data, data$fire_size_class %in% c("A", "B", "C"))
data_1 = filter(data, data$fire_size_class %in% c("D", "E", "F", "G"))
data_0$binary_class <- 0
data_1$binary_class <- 1
data <- rbind(data_0, data_1)

#----- Data Overview
    ----------------------------------------------------------------------#
str(data) # Information on columns
summary(data) # Summary of columns
head(data) # Other info on dataset
glimpse(data)
dim(data)

#----- Data Transformations
    ----------------------------------------------------------------------#
# Add column with the seasons instead of the months
data_summer = filter(data, data$discovery_month %in% c("Aug", "Jul", "Sep"))
data_fall = filter(data, data$discovery_month %in% c("Nov", "Oct", "Dec"))
data_winter = filter(data, data$discovery_month %in% c("Feb", "Mar", "Jan"))
data_spring = filter(data, data$discovery_month %in% c("Apr", "May", "Jun"))

data_summer$season <- "summer"
data_fall$season <- "fall"
data_winter$season <- "winter"
data_spring$season <- "spring"
data_transformed <- rbind(data_summer, data_fall, data_winter, data_spring)

data_transformed$discovery_month <- NULL
data_transformed$latitude <- NULL

data_transformed$Prec_pre_30 <- log(data_transformed$Prec_pre_30 + 1)
data_transformed$Prec_pre_15 <- log(data_transformed$Prec_pre_15 + 1)
data_transformed$Prec_pre_7 <- log(data_transformed$Prec_pre_7 + 1)
data_transformed$Temp_pre_30 <- log(data_transformed$Temp_pre_30 + 373.15)
data_transformed$Temp_pre_15 <- log(data_transformed$Temp_pre_15 + 373.15)
data_transformed$Temp_pre_7 <- log(data_transformed$Temp_pre_7 + 373.15)

continuous_vars <- c("Prec_pre_30", "Prec_pre_15", "Prec_pre_7",
                     "Temp_pre_30", "Temp_pre_15", "Temp_pre_7",
                     "Wind_pre_30", "Wind_pre_15", "Wind_pre_7",
                     "Hum_pre_30", "Hum_pre_15", "Hum_pre_7")
factor_vars <- c("fire_size", "fire_size_class", "stat_cause_descr",
                 "longitude", "state", "Vegetation", "binary_class", "season")

data_normalized <- cbind(data_transformed[factor_vars], scale(data_transformed[
    continuous_vars]))
```

```r

#------ Regression Predictions
    ------------------------------------------------------------------------#

# Set the train et test dataset
set.seed(1)

n = nrow(data);
idx = sample(n, 0.8*n)

data_regression <- data
data_regression_normalized <- data_normalized

data_regression$binary_class <- NULL
data_regression$fire_size_class <- NULL
data_regression_normalized$binary_class <- NULL
data_regression_normalized$fire_size_class <- NULL

train_regression = data_regression[idx,];
train_regression_normalized = data_regression_normalized[idx,];
dim(train_regression)

test_regression = data_regression[-idx,];
test_regression_normalized = data_regression_normalized[-idx,];
dim(test_regression)

#correlation
ggcorr(data_regression,
       method = c("all.obs","spearman"),
       nbreaks = 4, palette = 'RdBu', label = TRUE,
       name = "spearman correlation coeff.(rho)",
       hjust = 0.8, angle = -70, size = 3) +
  ggtitle("Spearman Correlation coefficient Matrix")

glimpse(train_regression)
glimpse(test_regression)

#cross validation
ctrl_lm <- trainControl(method = "cv", number = 10)

#baseline model
lmCVFit <- train(fire_size~., data = train_regression_normalized, method = "lm",
    trControl = ctrl_lm, metric = "Rsquared")
print(lmCVFit)
summary(lmCVFit)

#final model
#train dataset
lmCVFit <- train(fire_size~. + Hum_pre_30:Wind_pre_30 + Hum_pre_15:Wind_pre_15 + Hum_
    pre_7:Wind_pre_7
                 + Hum_pre_30:Temp_pre_30 + Hum_pre_15:Temp_pre_15 + Hum_pre_7:Temp_
                     pre_7
                 , data = train_regression_normalized, method = "lm",
                 trControl = ctrl_lm, metric = "Rsquared")
print(lmCVFit)
summary(lmCVFit)

predictions <- predict(lmCVFit, newdata = test_regression_normalized, interval = "
    confidence")
rmse_test = rmse(test_regression_normalized$fire_size, predictions)

#test dataset
lmCVFit <- train(fire_size~. + Hum_pre_30:Wind_pre_30 + Hum_pre_15:Wind_pre_15 + Hum_
    pre_7:Wind_pre_7
```

```
                    + Hum_pre_30:Temp_pre_30 + Hum_pre_15:Temp_pre_15 + Hum_pre_7:Temp_
                        pre_7
                    , data = test_regression_normalized, method = "lm",
                    trControl = ctrl_lm, metric = "Rsquared")
print(lmCVFit)
summary(lmCVFit)

#------ Classification Predictions
    ------------------------------------------------------------------------------#

data_predictions2 <- data

data_predictions2$fire_size <- NULL
data_predictions2$fire_size_class <- NULL

train2 = data_predictions2[idx,];
dim(train2)

test2 = data_predictions2[-idx,];
dim(test2)

#Baseline model / Train
fm.logisticR <- glm(binary_class ~ ., data=train2, family=binomial(link="logit"))
predicted_train <- predict(fm.logisticR, newdata = train2, type = "response")  #
    predicted scores

#Baseline model / Test
fm.logisticR <- glm(binary_class ~ ., data=test2, family=binomial(link="logit"))
predicted_test <- predict(fm.logisticR, newdata = test2, type = "response")  #
    predicted scores

# ROC curves
library(pROC)
ROC_train <- roc(train2$binary_class, predicted_train)
ROC_test <- roc(test2$binary_class, predicted_test)

# Review ROC objects
threshold_train <- coords(ROC_train, "best", "threshold")
threshold_test <- coords(ROC_test, "best", "threshold")

# Area Under Curve (AUC) for each ROC curve (higher -> better)
ROC_train_auc <- auc(ROC_train)
ROC_test_auc <- auc(ROC_test)

# plot ROC curves
plot(ROC_train, col = "green", main = "ROC: Train(green), Test(red) ")

points(ROC_test, col="red", pch="*")
lines(ROC_test, col="red")

# print the performance of each model
paste("Accuracy % Train Dataset: ", mean(train2$binary_class == round(predicted_train,
     digits = 0)))
paste("Accuracy % Test Dataset: ", mean(test2$binary_class == round(predicted_test,
    digits = 0)))
paste("Area under curve Train Dataset: ", ROC_train_auc)
paste("Area under curve Test Dataset: ", ROC_test_auc)
paste("Sensitivity Train Dataset: ", threshold_train[2])
paste("Sensitivity Test Dataset: ", threshold_test[2])


#------ Bootstrap
    ------------------------------------------------------------------------------#

library("boot")
```

```
237
238 #cross validation
239 ctrl_lm <- trainControl(method = "cv", number = 10)
240
241 n_rep <- 5000
242
243 coeff.boot = function(data, indices){
244   fm <- lm(data = data[indices,], fire_size~. + Hum_pre_30:Wind_pre_30 + Hum_pre_15:
        Wind_pre_15 + Hum_pre_7:Wind_pre_7
245                    + Hum_pre_30:Temp_pre_30 + Hum_pre_15:Temp_pre_15 + Hum_pre_7:Temp_
                           pre_7)
246   return(coef(fm))
247 }
248
249 boot.out = boot(train_regression_normalized, coeff.boot, n_rep)
250 boot.out
251
252 boot.ci(boot.out, type="norm", index=1)
253 boot.ci(boot.out, type="norm", index=2)
254 boot.ci(boot.out, type="norm", index=3)
255
256 boot.out$t0
257
258 plot(boot.out, index = 1)
```