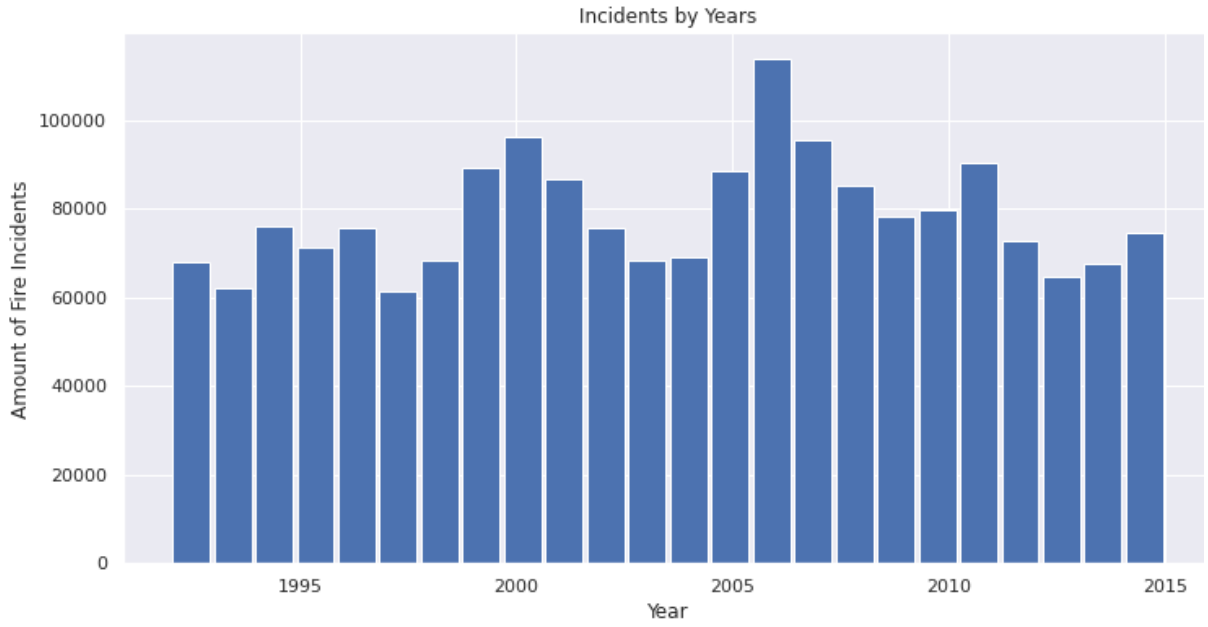
U.S. Wildfire Analysis Report

Q1: Have wildfires become more or less frequent over time?

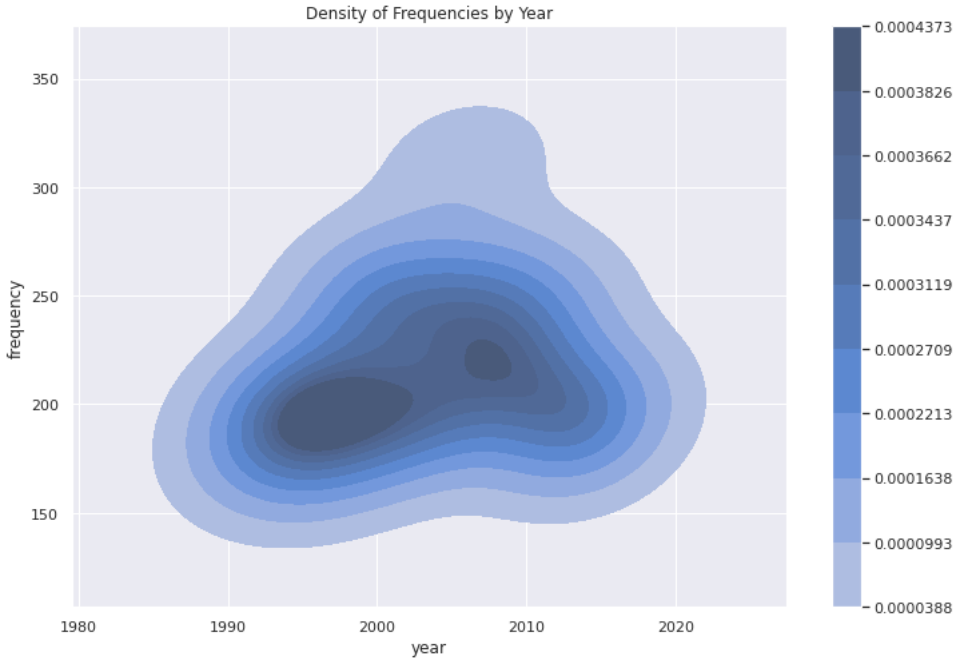
In 1992, the number of incidents are 67,975. The incidents first decreases in 1994, then increases back to 71,472 in 1995. In 1997, the number of incidents are minimum with 61,450. Then the incidents increases to 96,416 in 2000, after that increase, the numbers drop down to 68,261 in 2003. After the drop, the incidents creates a peak with a number of 114,001, then drops to 78,325 in 2009. The numbers increases to 90,552 in 2011, then rolls back to 64,780 in 2013. Finally, the numbers increases back to 74,491 in 2015.



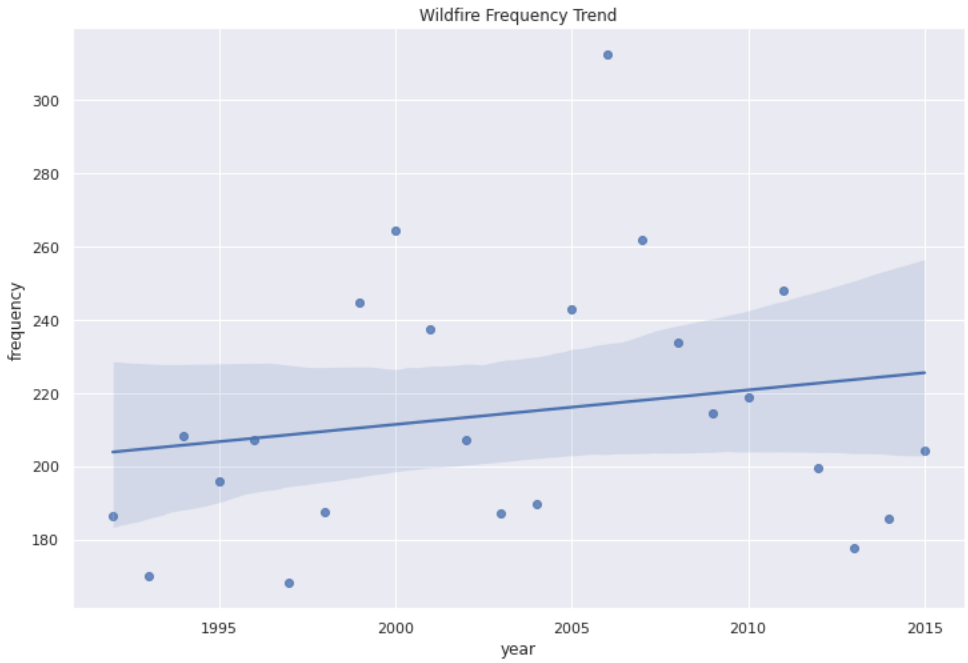
It is hard to determine if the wildfires are becoming more or less frequent from this histogram. To provide a clearer sight, number of incidents are given below in the table with respect to their matched years.

1992	1993	1994	1995	1996	1997	1998	1999
67975	61989	75955	71472	75574	61450	68370	89363
2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**
96416	86587	75656	68261	69279	88604	114004	95573
2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
85378	78325	79889	90552	72769	64780	67753	74491

In the graph below, a density estimation is given with the incident frequencies (num. of incidents / 365). The wildfires that are between 1992 and 2000 have more or less similar frequencies, this makes the timeline looking more dense than others in the graph. In 2006, the frequency of incidents rises exceptionally and makes the graph seem more sparse. We can see the minimums are slightly increases. To get a better understanding, we will examine the frequency trend.



The graph below shows a clearer sight that the frequency of the wildfires, from 1992 to 2015, is in an increasing trend, which proves ***the frequency of wildfires in U.S. is increasing over time***.



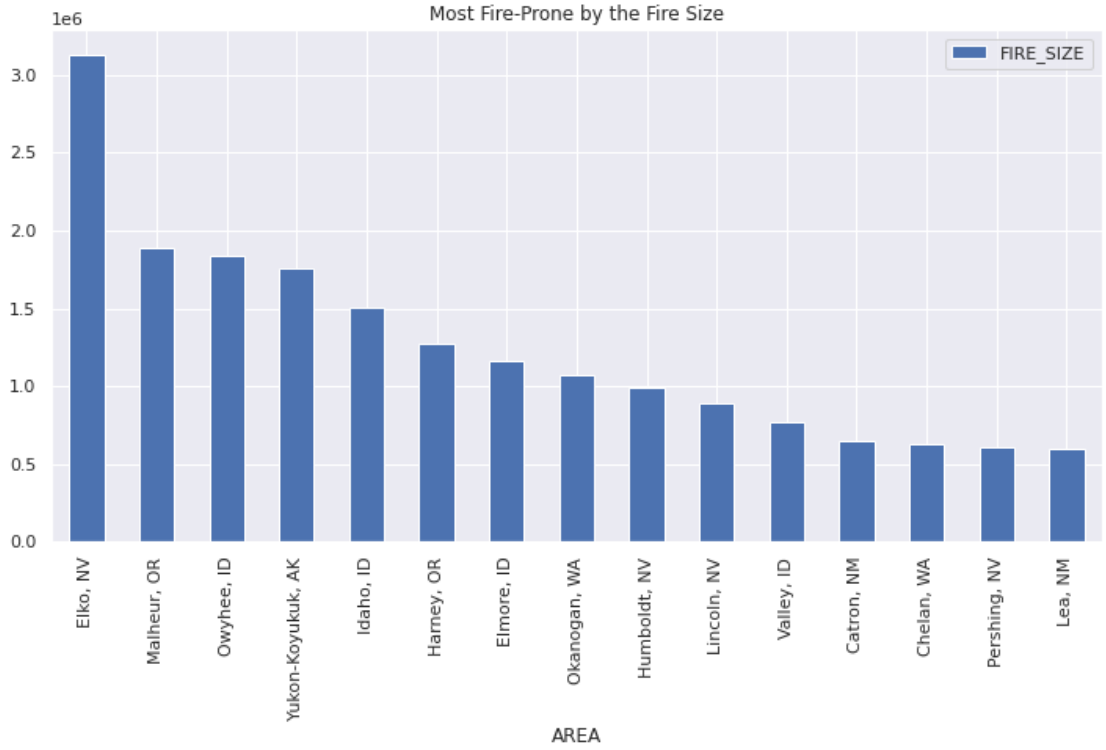
Q2: What counties are the most and least fire-prone?

Q2.1) Fire Size

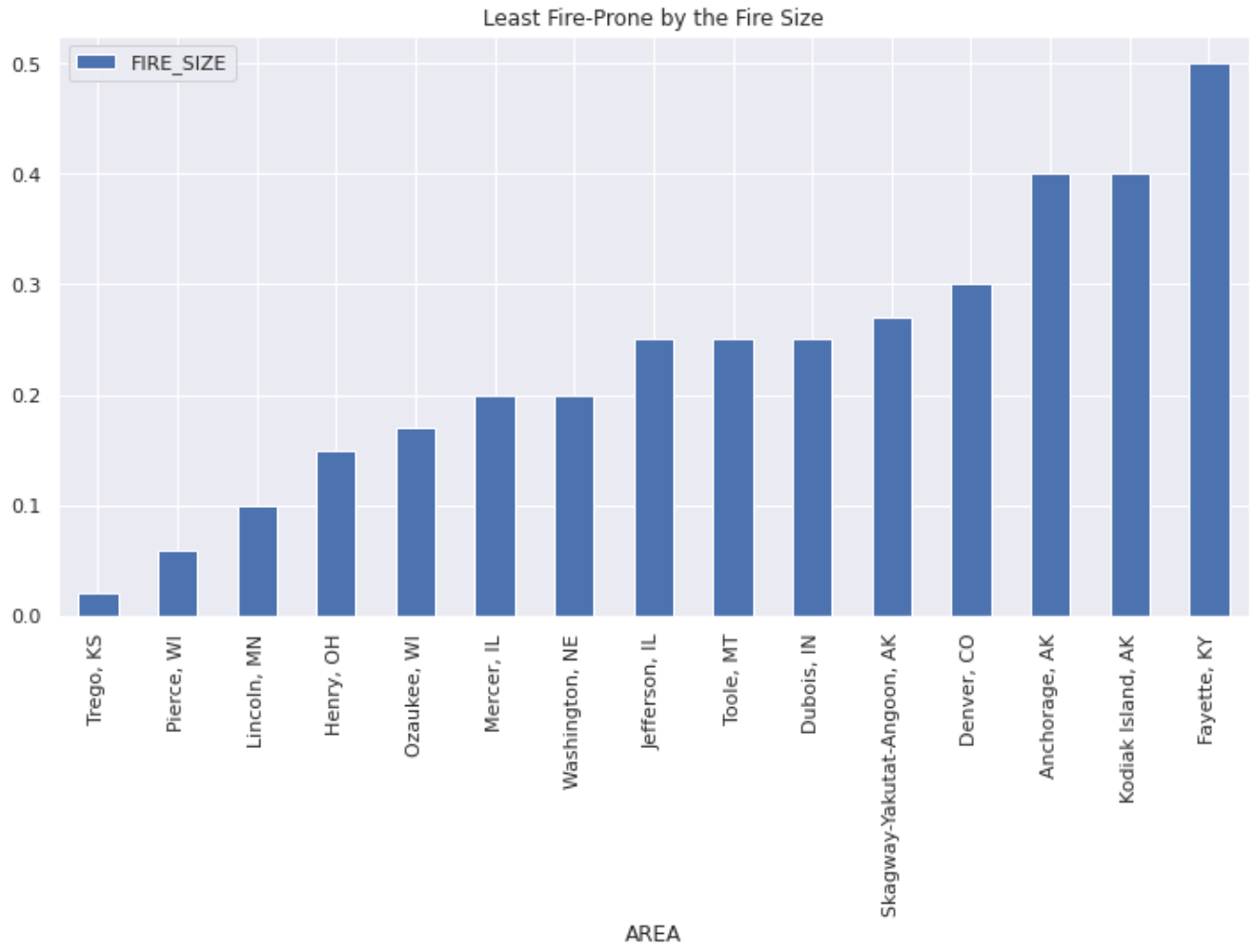
To understand which counties are the most and least fire-prone, we should also check the fire sizes as well as frequency of reports. First, we will look at the 15 counties with their number of total wildfire sizes by acres.

Elko, NV	**Malheur, OR**	**Owyhee, ID**	**Yukon-Koyukuk, AK**	**Idaho, ID**
3132882.440	1886590.550	1834768.900	1754427.690	1509590.250
Harney, OR	**Elmore, ID**	**Okanogan, WA**	**Humboldt, NV**	**Lincoln, NV**
1272853.200	1159911.300	1076506.830	992270.260	895149.400
Valley, ID	**Catron, NM**	**Chelan, WA**	**Pershing, NV**	**Lea, NM**
765921.830	644717.680	627131.590	610328.400	593666.819

In the graph below, the counties are sorted by the least fire sizes. With respect to the graph we can say ***Elko, Nevada is the most fire prone*** by the measure of fire sizes.



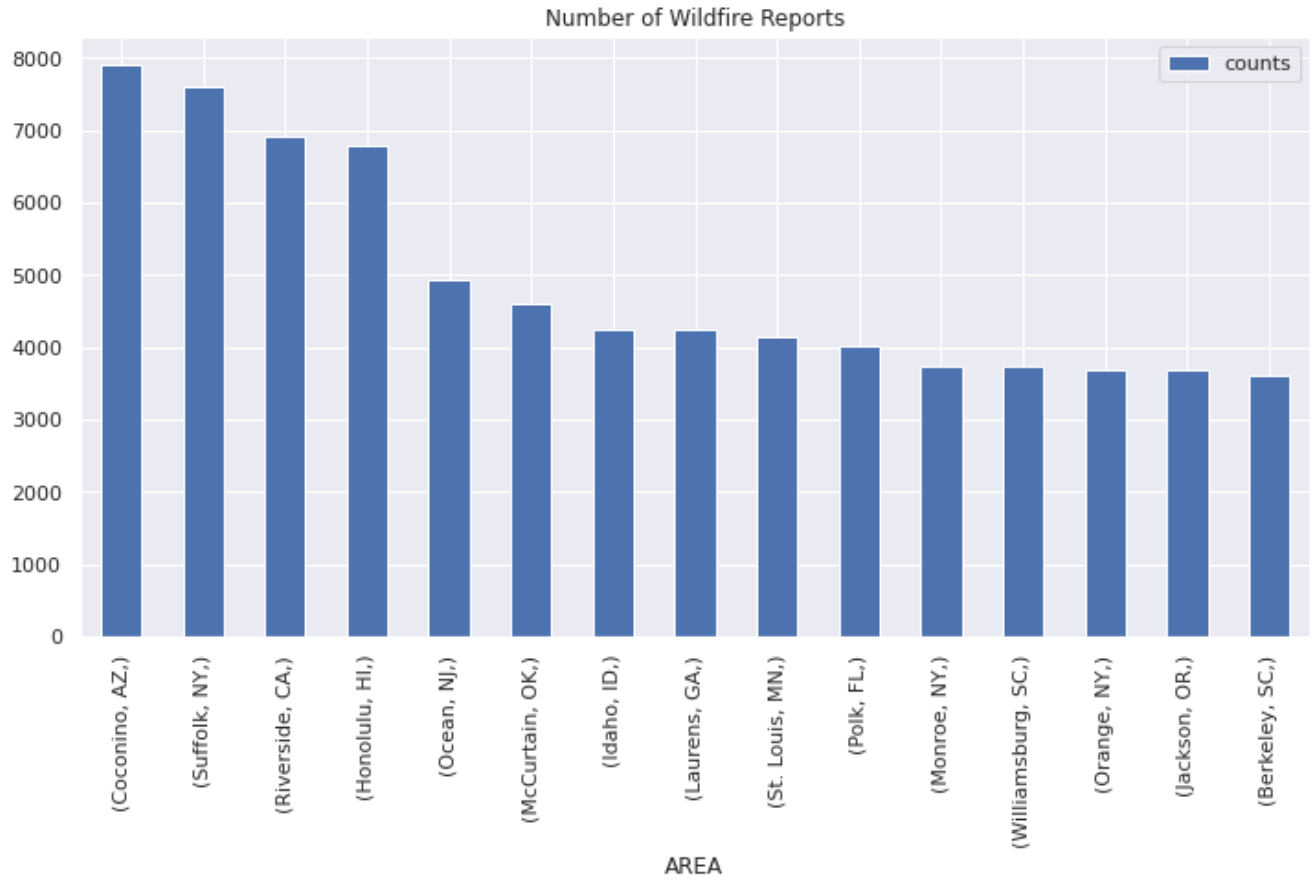
In the graph below, the counties are sorted by the least fire sizes. With respect to the graph we can say ***Trego, Kansas is the least fire prone*** by the measure of fire sizes.



Q2.1) Frequency

To understand which counties are the most and least fire-prone, we now will check the frequency reports. First, we will look at the 15 counties with their number of incident frequencies.

Coconino, AZ	**Suffolk, NY**	**Riverside, CA**	**Honolulu, HI**	**Ocean, NJ**
7900	7595	6925	6780	4941
McCurtain, OK	**Idaho, ID**	**Laurens, GA**	**St. Louis, MN**	**Polk, FL**
4593	4255	4246	4153	4013
Monroe, NY	**Williamsburg, SC**	**Orange, NY**	**Jackson, OR**	**Berkeley, SC**
3747	3732	3694	3675	3610



According to the graph given above, ***Coconino, Arizona is the most fire prone*** by the number of wildfire reports it have. After Coconino; Suffolk, Riverside and Honolulu are one of the most fire prone counties.

Least fire prone counties with respect to the frequency, they will be listed below <u>if they get lower than 2 reports in a year</u>.

Red Lake, MN | Alfalfa, OK | Craig, OK | Costilla, CO | Madison, OH | Ascension, LA | Concordia, LA | Hamilton, IL | Freeborn, MN | Adams, NE | Henry, OH | Faribault, MN | Hillsdale, MI | Pontotoc, OK | Milwaukee, WI | Daniels, MT | Cuyahoga, OH | Hampton City, VA | Gurabo Municipio, PR | Fayette, KY | Fairfax, VA | Kingfisher, OK | Garden, NE | Jefferson, IL | Washington, NE | Gratiot, MI | Lake, OH | Ozaukee, WI | Lafayette, WI | Owyhee, NV | Cass, IL | Rogers, OK | Kenosha, WI | Buena Vista City, VA | Dubois, IN | Robertson, KY | Canadian, OK | O'Brien, IA | Walworth, WI | Kodiak Island, AK | Douglas, CA | Caddo, OK | Kiowa, OK | Sheridan, MT | Calhoun, IL | Boyd, NE | Pepin, WI | Colonial Heights City, VA | Denver, CO | Platte, MO | Humboldt, IA | Pipestone, MN | Love, OK | Barber, OK | Garvin, OK | Terrebonne, LA | Richland, MT | Ellis, OK | Gibson, IN | Wichita, KS | Lipscomb, OK | Clay, AR | Richmond (city), VA | Lincoln, NE | Lincoln, MN | Berrien, MI | Wells, ND | Todd, SD | Blaine, NE | Toole, MT | Glacier County, MT |

Q3: Given the size, location and date, can you predict the cause of a wildfire?

To predict the cause of a wildfire, the following steps are taken in order: 1) Feature that will be used, 2) Determining which model to train, 3) Training and evaluating the model.

Features:
FIRE_SIZE
STATE
LATITUDE
LONGITUDE
CONT_DOY
CONT_TIME
STAT_CAUSE_DESCR

Models:
DecisionTreeClassifier
ExtraTreeClassifier
BaggingClassifier
RandomForestClassifier
KNeighborsClassifier
GradientBoostingClassifier

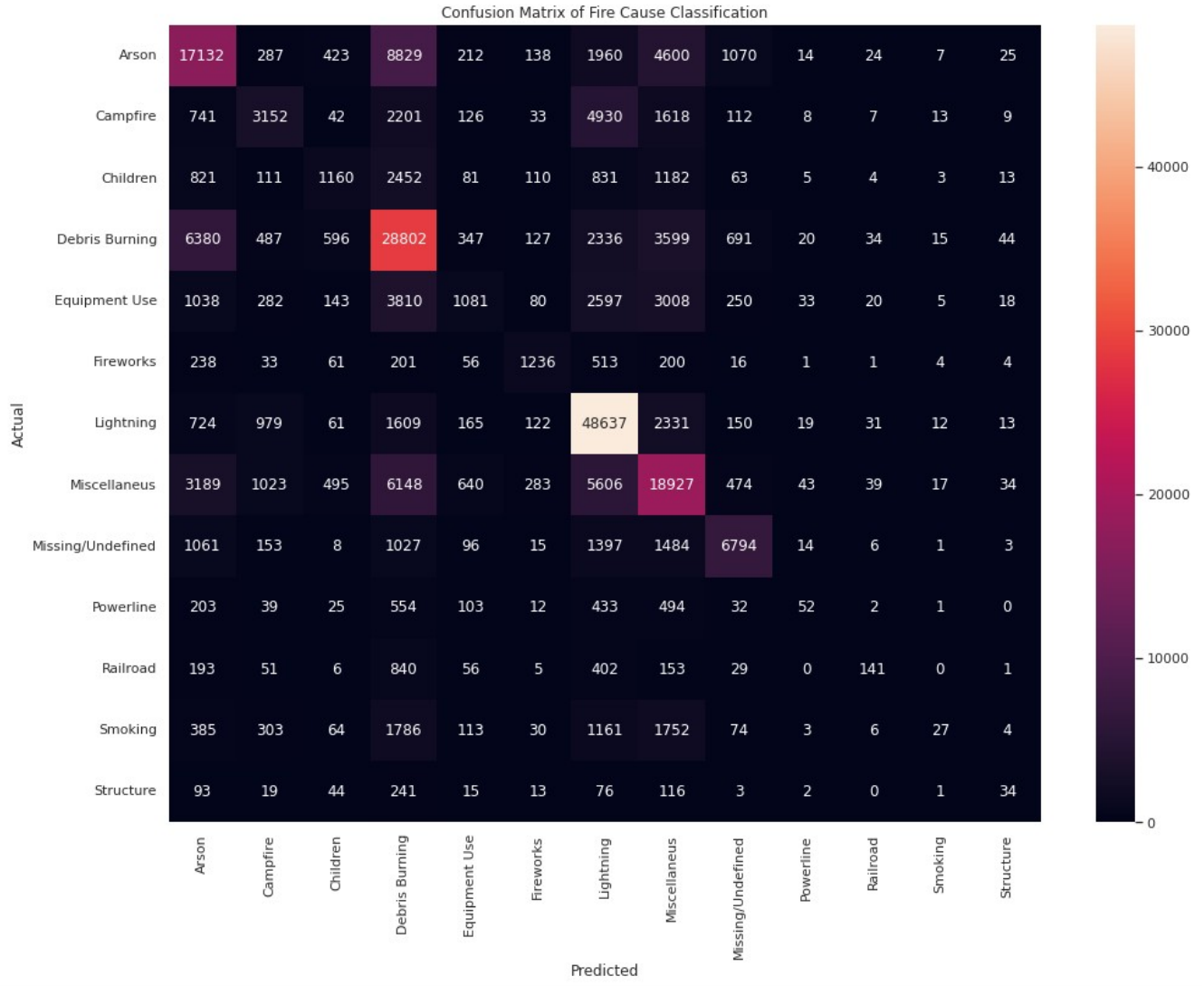
Model Performance Comparison			
	Train ACC	Test ACC	Time Cost (sec.)
Decision Tree	99.9%	47.5%	5.37
Extra Tree	99.9%	44.1%	0.59
Bagging	98.2%	55.8%	36.09
Random Forest	99.9%	58.9%	136.33
KNeighbors	63.1%	50.1%	0.85
Gradient Boosting	52.8%	52.3%	7621.54

From the table above, the most promising model is Gradient Boosting with 52.8% of training accuracy and 52.3% of testing accuracy. Decision Tree, Extra Tree, Bagging and Random Forest seems to overfit heavily. The most closer model in terms of performance is Kneighbors. To train a model that predicts the cause of a wildfire, we will choose Gradient Boosting. If the project can’t afford a time cost, we can choose Kneighbors over Gradient Boosting.

Training and Evaluating:

Gradient Boosting Classifier Accuracy and Time Cost			
	Train ACC	Test ACC	Time Cost (sec.)
Gradient Boosting (n_estimators=600)	57.0%	56.0%	7903

From the table, it is obvious that the trained Gradient Model with *n_estimators=600* is better than the previous trained models. To evaluate the model further, we will look at the confusion matrix and the classification report.



Gradient Boosting Classifier Confusion Matrix Report				
	TP	TN	FP	FN
0 - Arson	17132	177191	15066	17589
1 - Campfire	3152	210219	3767	9840
2 - Children	1160	218174	1968	5676
3 – Debris Burning	28802	153802	29698	14676
4 -Equipment Use	1081	212603	2010	11284
5 - Fireworks	1236	223446	968	1328
6 - Lightning	48637	149883	22242	6216
7 - Miscellaneus	18927	169523	20537	17991
8 - Missing/Undefined	6794	211955	2964	5265
9 - Powerline	52	224866	162	1898
10 - Railroad	141	224927	174	1736
11 - Smoking	27	221191	79	5681
12 – Structure	34	226153	168	623

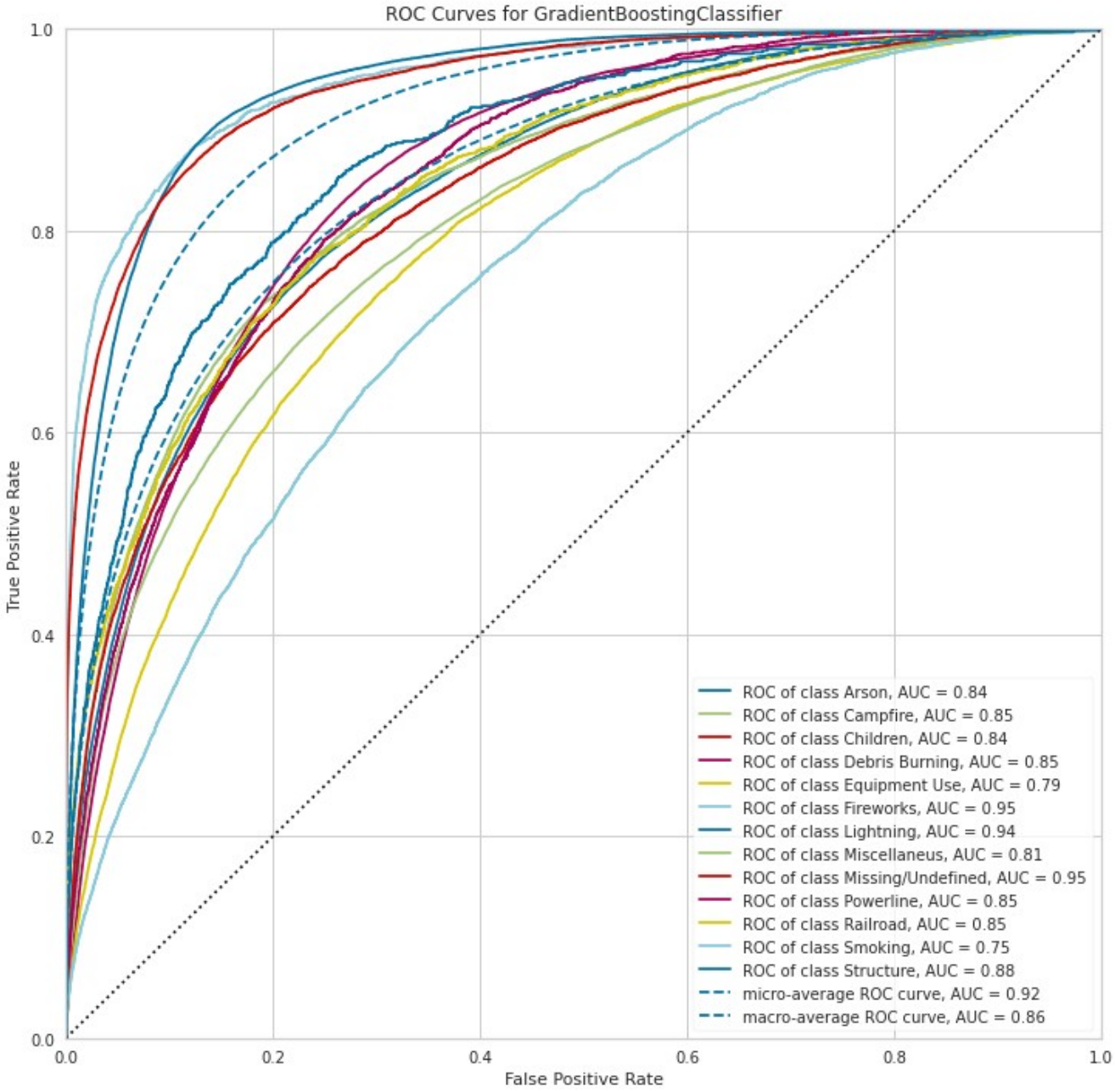
From the analysis of confusion matrix, the model seems like successful only at predicting *Arson, Lightning, Missing/Undefined* classes. Now let's ook at to a more detailed table which examines sensitivity, specificity, presicion, recall and f1-score.

Gradient Boosting Classifier Classification Report						
	Sensitivity	Specificity	Precision	Recall	F1-Score	Support
0 - Arson	49.3%	92.1%	53%	49%	51%	34721
1 - Campfire	24.2%	98.2%	46%	24%	32%	12992
2 - Children	16.9%	99.1%	37%	17%	23%	6836
3 – Debris Burning	66.2%	83.8%	49%	66%	56%	43478
4 -Equipment Use	8.7%	99%	35%	9%	14%	12365
5 - Fireworks	48.2%	99.5%	56%	48%	52%	2564
6 - Lightning	88.6%	87%	69%	89%	77%	54853
7 - Miscellaneus	51.2%	89.1%	48%	51%	50%	36918
8 - Missing/Undefined	56.3%	98.6%	70%	56%	62%	12059
9 - Powerline	2.6%	99.9%	24%	3%	5%	1950
10 - Railroad	7.5%	99.9%	45%	8%	13%	1877
11 - Smoking	0.4%	99.9%	25%	0%	1%	5708
12 - Structure	5.1%	99.9%	17%	5%	8%	657

The model was pretty unsuccessful on predicting the following wildfire causes: *Arson, Campfire, Children, Equipment Use, Fireworks, Miscellaneus, Missing/Undefined, Powerline, Railroad, Smoking* and *Structure*.

The model was successful only on predicting the following wildfire causes: *Debris Burning* and *Lightning*.

ROC Curve



In conclusion, from looking at the tables and the ROC curve given above ***we can't successfully predict the cause of a wildfire***.