Utilizing Supervised Machine Learning to predict the cause of fires in the US: report

Problem Statement: the primary goal of this project is to determine if we can train a model that can predict the cause of fires in the United States.

Background:

In the United States, wildfires are frequent every year, especially during summer. Wildfires can be natural and non-natural. Natural wildfires are vital to maintaining the ecosystems' health. Indeed, many plants and animals rely on natural fires to complete the cycle of life. Non-natural wildfires are those caused by factors external to nature—for instance, campfires, smoking, arson, among others. When non-natural causes initiate a wildfire, it can hurt the ecosystems affected by the fire.

On the other hand, wildfires can also have an impact on our economy. In fact, in 2015, wildfires caused more than USD 14 billion in real estate damage(https://www.usfa.fema.gov/data/statistics/#tab-4). Therefore, a better understanding of the causes of fires is essential to reduce the number of wildfires initiated by non-natural factors.

The ideas exposed above give us the setting to bring in all the tools of Machine Learning to address this task. We have analyzed a dataset of fires in the US for a period of more than 20 years to train a model that can predict the cause of fires.

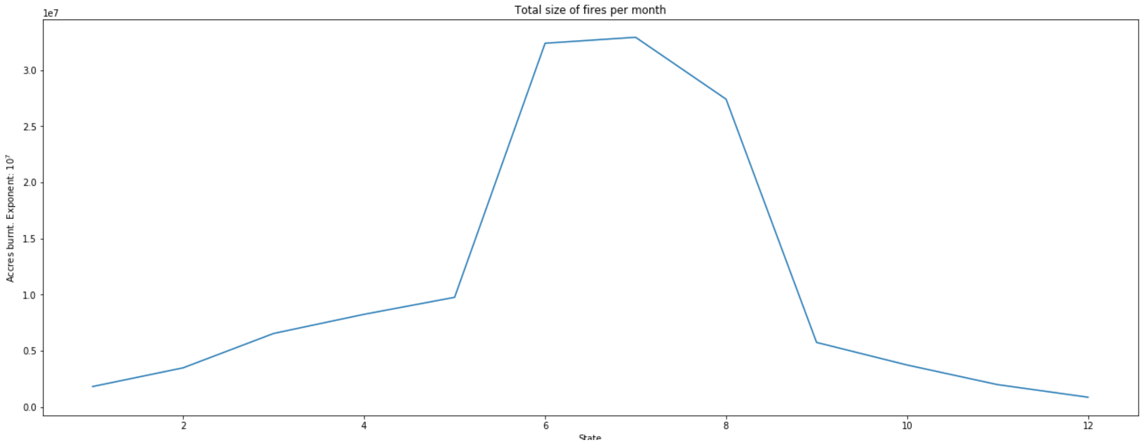
Source of Data: The data is available in Kaggle: https://www.kaggle.com/rtatman/188-million-us-wildfires

Preprocessing: the first step of the analysis consists of reading in the data that comes in SQLite format. I explored the tables contained in the data, and I chose the table called 'Fires'. This table has about 1.87 million datapoints and contains information like the location of fires, size, among others. My machine was able to read in only the 1.8 million observations.

The next step is to look for NANs, duplicated rows, or columns. For the NANs, I eliminated columns with more than 25% on NANs. In the case of duplicated rows, I also removed them. The duplicated columns required more analysis than the duplicated rows. Indeed, some columns contained the same information but in different formats. For instance, the year fires took place is the same as the discovery year of those fires. In such cases, I removed one of those columns. There were some columns whose information I considered irrelevant to make the analysis. For instance, I dropped the columns with the reporting agency information, as I believe the models trained should be independent of who reported a fire.

I unsuccessfully tried to calculate VIF and the correlation matrix, due to the size of this dataset. Therefore, I addressed multicollinearity via PCA.

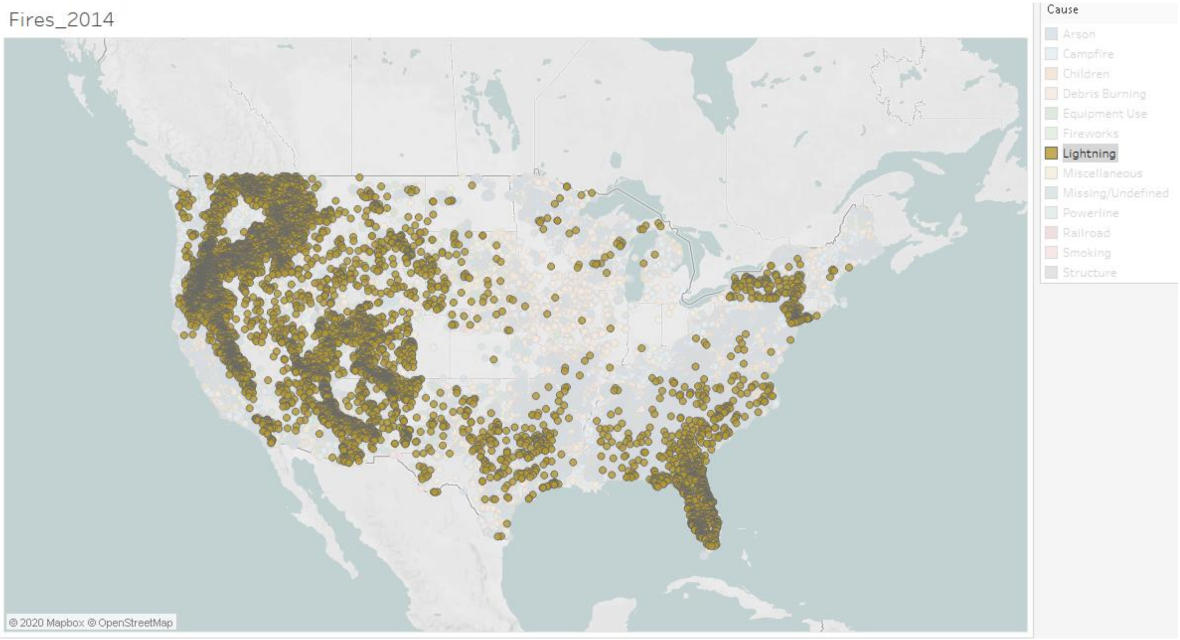
I also created some visualizations to understand the behaviour of fires from different perspectives. For example, the image below shows the number of fire occurrences per month.



After doing cleaning the data and doing an EDA, I defined the dependent and independent variables. One conclusion drawn from the EDA is that we are dealing with imbalanced data.

Since we have 1.8M observations, I worked with a subsample of the data that allowed me to execute the next steps. After making a random subsampling, I downsampled or (inclusive or) oversampled the data to balance the number of observations per category.

I did the preprocessing work in two Jupyter Notebooks that are attached. The first notebook is called USAfires_Cleaning, and it contains the code used to extract the data from SQLite format to a pandas data frame format, as well as part of the cleaning process. The second one, called USAFires_Cleaning_2, contains the last part of the cleaning process, as well as some visualizations and the data balancing process. I also used Tableau to create visuals of the distribution of fires in the USA geography. These latter visuals are in the attached Tableau book. As an example of these, we can see the distribution of fires caused by Lightning in 2014 below.



Modelling:

I run three models: Decision Tree, Random Forest and KNN. I run each model independently. I also used PCA together with those models via Pipelines. The pipeline was also used to optimize the parameters. To evaluate those models, I calculated the accuracies, confusion matrices, precisions, and recalls. When possible, I computed the feature importance. I run the models mentioned above with two rebalanced datasets: one that gives us a constant number of members in each category (among the cause of fires) and another one that combined downsampling and oversampling. In both cases, the accuracies were similar, but the latter approach gave us the highest accuracy. The model that gave us the best performance in terms of accuracy is Random Forest without PCA. The accuracy was about 45%

I also run the models aforementioned in my local machine with a small sample of the dataset, and then I run the models in the cloud with a more significant portion of the data. The accuracies improved from about 42% to about 45%. Attached are the two notebooks where I run those models. The notebook run in the cloud is called Sage, and the other one is called USAFires_Modelling. The code in both notebooks is mostly the same.

Conclusions: we can use supervised machine learning to train a Random Forest model that predicts the cause of fires with an accuracy of 45%. This accuracy is acceptable, given the scope of the problem.

Businesses applications: we can use a Random Forest model to obtain an initial possible cause of a wildfire. Such a model could give hints to conduct preliminary investigations, and it could reduce the amount of time, effort and resources invested in determining what the actual cause of a fire is. Indeed, to prevent future non-natural wildfires, we need to understand what was the spark that started the fire.

Next Steps.

First, I would like to train the models mentioned above with a bigger subsample of the original dataset. In this way, I could improve the accuracy of our models.

Second, I want to incorporate information to the data set to retrain the models. For instance, I think human activity near fire locations could have a potential predictive power.

Next, I want to raise awareness among the people on the impact of wildfires in our lives and our economy. I want to make a website or an application that collects, shows fires' location, predicted cause, size, area affected, in real-time.

Finally, another goal is to analyze data in other places of the world where wildfires are frequent and are negatively impacting the people's or ecosystems' life. For instance, wildfires should not be regular in the Amazon rainforest. Unfortunately, fires are happening every year, and their occurrences are increasing.