Wildfire Cause Prediction

Ganesh Thanu
gthanu@ncsu.edu

Arun Gaonkar
agaonka@ncsu.edu

Nikhil Prashant Patil
npatil@ncsu.edu

1 Introduction

Forests are integral part of terrestrial ecosystem and it has major impact on the total biomass of the earth. Each year, in US alone, around 60,000 wildfires can be seen and it causes impeccable damage to the earth. Although Wildfires are a natural occurrence, it is becoming extreme and widespread. Hotter, drier weather caused by climate change or natural events like lightning can cause a fire in the forest. Human activities like smoking, unattended campfires, uncontrolled fireworks are also a major contributors to wildfire.

Forest fire analysis and cause prediction measures have become increasingly important. Technologies for finding the cause represents an essential tool to predict future forest fire risks, back up the forest fire monitoring, activate prevention and security measures.

In this project, a Machine Learning model is proposed for the prediction of the cause of forest fires. The model can be used with Data Analysis techniques to forecast the cause of a wildfire using the dataset, which comprises geographic records of forest fires and causes for over 20 years. To build the predictive model, the parameters being considered are forest fire size, geography, burn duration, number of fires, etc.

2 Method

2.1 Data Preprocessing

The original Dataset from FPA-FOD contains 1,880,465 fire instances and distributed over 130 attributes. Most of the attributes are related to Federal and State agencies which is not relevant for cause predictions. After filtering, we selected 12 attributes for further processing.

Original dataset contained Fire Discovery Date and End Date in the Julian format and that is converted to Gregorian Date format.

2.2 Exploratory Data Analysis

We have 13 categories of Fire cause.

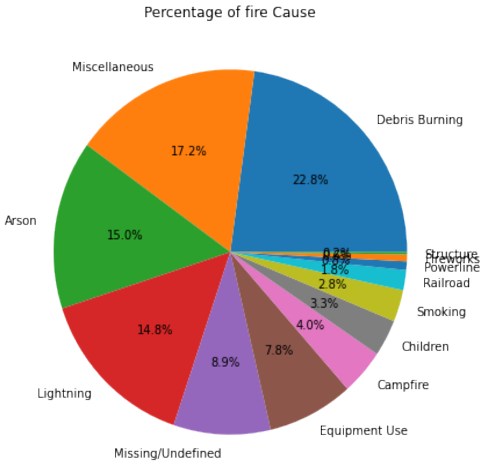
Debris burning, Arson, Lightning are the major causes of wildfire accounting for more than 50% of total wildfires, from Fig 1.

There is no significant difference in wildfire count yearwise, from Fig 2.

There is significant difference between wildfire count from October-January and February-September. It might be because of weather, since weather is cold from October-January which is not conducive for wildfire, , from Fig 3.

Fire count varies by state to state. CA, GA and TX has seen most number of wildfires, usually drier states have been under more Wild fires, , from Fig 4.

FireSize varies based on the reason for the firecause, from Fig 5. Children activites have larger burn size than other activites.



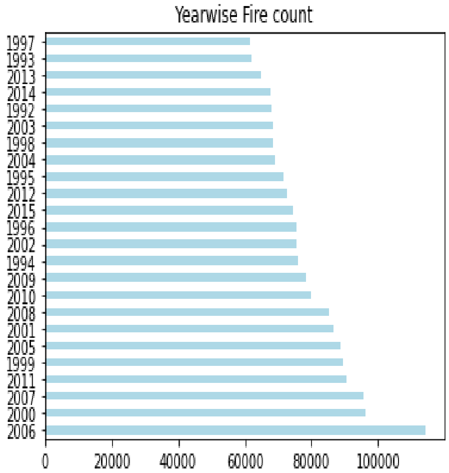
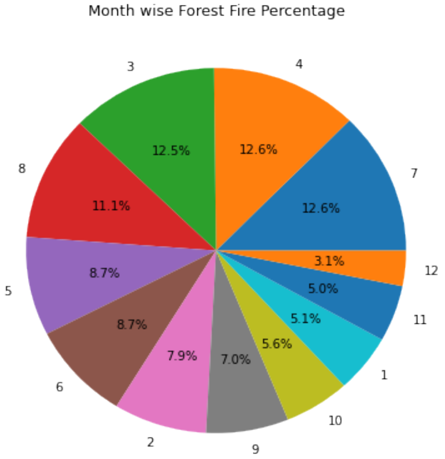


Figure 1: Percentage of fire cause

Figure 2: Yearwise Fire Count



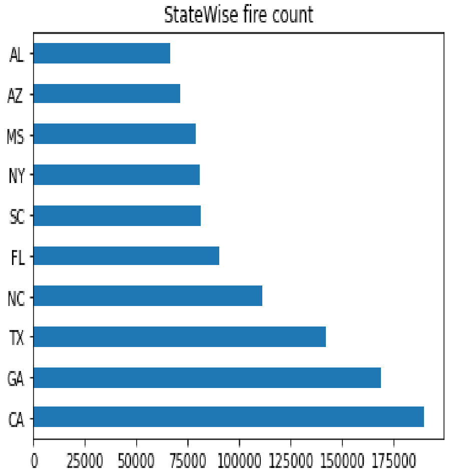
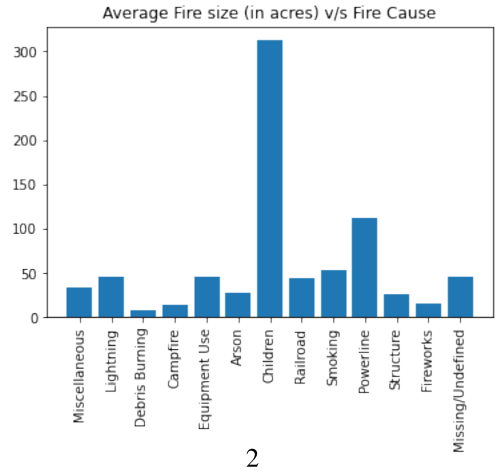


Figure 3: Monthwise Fire Percentage

Figure 4: Statewise Fire Count

Figure 5: Average Fire Size (in acres) vs Fire Causes



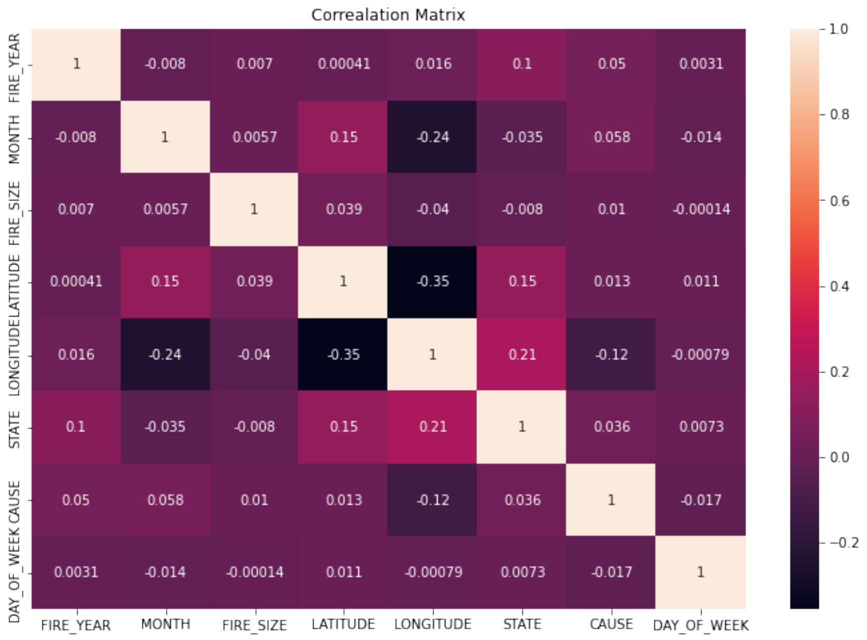


Figure 6: Correlation Matrix

From Fig 6, it can be inferred that, there is no significant correlation between columns.

3 Experimental Setup

3.1 Dataset

The forest fire dataset *1.88 Million US Wildfires* [1] obtained from Kaggle [2] is used for building the model. This database contains geospatial records of wildfires that occurred in the US from 1992 to 2015. The wildfire records are taken from the reporting systems of federal, state, and local fire organizations. This database is referred to as *Fire Program Analysis Fire-Occurrence Database* (FPA-FOD).[3]

3.2 Libraries

The following python libraries are used in Google Colaboratory (Google Colab)-

1. sqlite3
2. numpy
3. pandas
4. matplotlib
5. seaborn
6. sklearn

[1] Short, Karen C. 2017. Spatial wildfire occurrence data for the United States, 1992-2015 [FPAFOD20170508]. 4th Edition. Fort Collins, CO: Forest Service Research Data Archive. https://doi.org/10.2737/RDS-2013-0009.4

[2] https://www.kaggle.com/rtatman/188-million-us-wildfires

[3] http://regclim.coas.oregonstate.edu/FireStarts/fpa-fod_RODBC_01.html

4 Results

We tested decision tree, random forest algorithms for classification. We got better testing accuracy for random forest(58%) than the decision tree(46%). As model was classifying data into 13 labels, we tried to group them in 4 categories as below:

natural = [‘Lightning’]

accidental = [‘Structure’,‘Fireworks’,‘Powerline’,‘Railroad’,‘Smoking’,‘Children’,‘Campfire’,‘Equipment Use’,‘Debris Burning’]

malicious = [‘Arson’]

other = [‘Missing/Undefined’,‘Miscellaneous’]

After grouping them in 4 categories, testing accuracy improved significantly from 58% to 70%.

Figure 7 shows the confusion matrix.

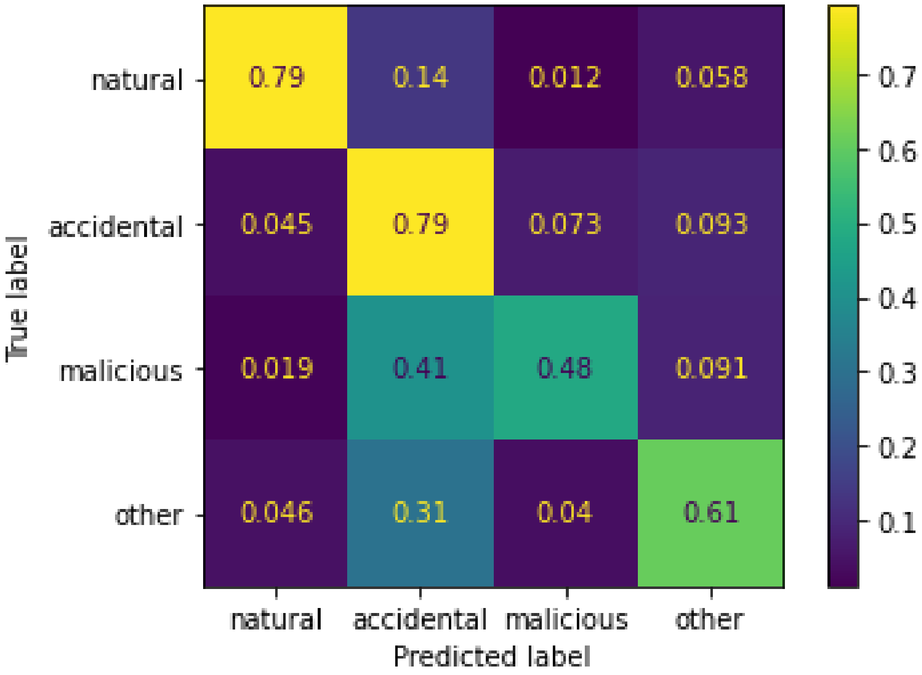


Figure 7: Confusion Matrix

5 Conclusion

We now have Random Forest Classifier as our baseline model.

We are planning to train the model with other classical ML approaches to check for better accuracy. Further we planned to train the model on Neural Networks, LSTM or SVM.

6 References

[1] J. Brownlee, “How to develop a random forest ensemble in Python,” Machine Learning Mastery, 26-Apr-2021. [Online]. Available: `https://machinelearningmastery.com/random-forest-ensemble-in-python/`.

[2] P. Cortez and A. Morais, A data mining approach to predict forest fires using meteorological data, Proceedings of the 13th Portugese Conference on Artificial Intelligence, pp. 512-523, 2007.