Wildfire Cause Prediction

CMPE255 Project

By:

Garima Chaphekar[014638871]

Saliha Mehboob [012553216]

Vidita Vijaykumar Daga [014488630]

Date: 10/06/2020

Professor: Dr. Jorjeta Jetcheva

SECTION 1: INTRODUCTION

Motivation:

Wildfires, whether natural or caused by humans, are considered among the most dangerous and devastating disasters around the world. They are a colossal problem in the United States (US). Their complexity comes from the fact that they are hard to predict, hard to extinguish and cause enormous financial losses. It can result in significant impacts to humans, either directly through loss of life and destruction to communities, or indirectly through smoke exposure. With growing concerns about wildfire duration and frequency, knowing the cause of the wildfire is increasingly important.

Predicting the source and spread of fires could have considerable benefits for human health and life, the economy, and the environment. This could help identify areas with higher risk - for example, with limited resources, the authorities could choose to focus on monitoring specific areas.

Objective:

The project aims to use a government-funded dataset of wildfire data to generate visualizations for better understanding of wildfire scope, frequency, and causation. Additionally, these visualizations guide the creation of machine learning models to predict the causes of wildfires throughout the United States given user input. It would be immensely useful to understand the causes, and severity of wildfires within the United States to help make recommendations for government aid spending and even to develop initiatives for preventative measures

Along with predicting the cause of the wildfire, we also plan to analyze few points like

- Has wildfire become frequent over time?
- What period of the day/month/year are the fires to be likely started/discovered?
- What counties are the most and least fire prone given the size, location, and date?
- Which cause of wildfire has increased over time?
- Predict the burned area of a forest fire

Approach:

The dataset used is presented in the form of an SQLite database. It was read from the database into a Pandas data frame object. Data cleaning followed by data exploration and visualization was performed. After determining the correlations between parameters from data exploration, various methods for machine learning were pursued. The intention is to create a model that can accurately predict the causes of wildfires given user input. Different models like KNN, Decision Tree, Random Forest, XGBoost, Neural network, Bagging classifier, Logistic regression and Naïve Bayes were employed to train on the data and predict the cause of the wildfire.

Literature Review:

De Bem et. al worked on predicting wildfire vulnerability in Brazil's Federal District dataset [1]. The authors proposed two data mining models namely Logistic Regression and Artificial Neural Network to predict fire occurrence in Brazil district located inside Brazil Cerrado. The presented models are optimized using feature selection techniques. The results of modeling shows that Logistic Regression performed better than ANN in case of burned area whereas in non-burned areas ANN provides better Area under the Curve(AUC).

Many researchers have worked to study the behaviour of wildfire to predict, monitor and prevent using Artificial Intelligence, Big Data techniques, and remote sensors. The remote sensing offers a rich source of satellite images to help monitor the wildfires. Sayad, Younes Oulad et. al combined Big Data, Remote Sensing and Data mining algorithms (ANN and SVM) to predict the occurrence of wildfires using the images from the satellite [2]. The model proposed used the dataset based on remote sensing data. The model is validated using classification metrics, cross validation, regularization, and model comparison.

SECTION 2: SYSTEM DESIGN AND IMPLEMENTATION

Our project goal is to predict the cause of wildfires based on the dataset of the United States. Many researches have been carried out to study the environmental impact of wildfires, economic impact, and wildfire prediction etc. However, the question remains is what causes the wildfire. We have performed the ablative study of the wildfire dataset using different approaches and parameters. The exploratory and visual analysis on wildfires provides useful insights about the data and hence helped us to perform the modeling.

The next section focuses on various machine learning algorithms explored to determine the cause of wildfire in the United States.

ALGORITHMS USED:

We tried several classification models with different preprocessing techniques for predicting the fire cause. Below are the major classification models that gave good accuracy:

1) **Decision Trees and Random Forest :** Both the models are built to predict the cause of the wildfire. In the initial experiments, the models are trained on the stratified data. Hyperparameter tuning is also performed to get the best parameter value. However, the models did not give a good accuracy, which gave a clear indication to tailor the data. The data contains numerous potential causes for the fires. This makes it difficult for accurate predictions due to the many possible outcomes. For simplicity, these categories are condensed into four possible causes: Natural, Accidental, Malicious, and Other. The Natural category includes lightning. The accidental category includes structure, fireworks, power line, railroad, smoking, children, campfire, equipment use, and debris burning. The malicious category includes arson

and the final category, other, includes missing/undefined and miscellaneous. Another grouping method of labels is also done.

2) **XGBoost :** This model is built to predict the cause of the wildfire, specifically for three labels: Lightning, Railroad, Arson. The classes are reduced because with the thirteen classes, the model gave only 55 percent accuracy. Since some causes like campfire and smoking had similar patterns, cause could not be predicted accurately. Hence, we chose to reduce the number of classes and work on the subset of the dataset with only three labels.

3) **Neural Network:** Artificial Neural Network is one of the commonly used models in the domain of wildfires analysis. There are several high level APIs available for building neural networks and Keras is one of them. TensorFlow backend is used with keras. The features considered to predict the cause with Neural Network are Fire size, time and location of the fire.

In order to develop the model, Keras Sequential model is used that provides the stack of layers structure. The first step is to define the sequential model with the number of hidden layers to be used in a model,activation function to be used, number of nodes , input dimensions, and output dimension. The second step is to compile the model providing the loss function and optimizer. The third step includes the model fitting on the dataset where batch size, epochs and learning rate is provided.

An alternative approach is to build a feed forward neural network where two hidden layers are used. The gradient descent is used to learn the weights during training to reduce the error during backward propagation. Tanh activation is applied to each node in the layer.

In addition, we also experimented with KNN, Naive Bayes, Logistic Regression,Bagging Classifier.

TECHNOLOGIES ,TOOLS AND LIBRARIES USED:

- Python 3, Jupyter Notebooks, Scikit-learn, HPC, TensorFlow, Tensor Board, Keras, Ray Tune, Geopy

ARCHITECTURE-RELATED DECISION:

The project focuses on different aspects of wildfires analysis. Multiple models are used with different specifications such as dataset size, features, cause prediction, and preprocessing etc. This helps us to delve deeper into studying and analyzing various dimensions of wildfires cause prediction.As With Neural Network, the model requires the standardization before training and normalization is usually done after the splitting of data into train,test and validation sets.

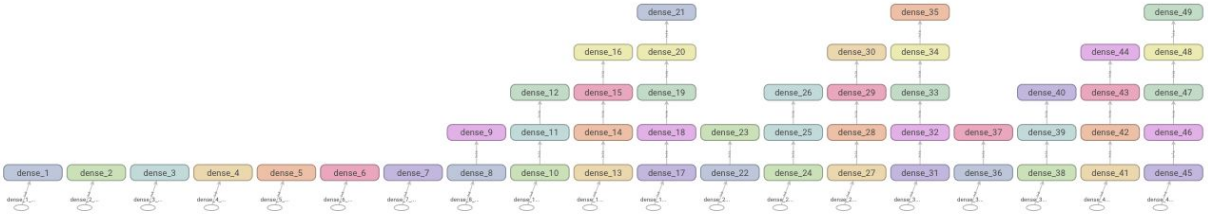


Figure1: Keras ANN Model

SECTION 3: EXPERIMENTS:

DATASETS USED:

The dataset used in this work was found on Kaggle and is titled ***"1.88 Million US Wildfires"***. The data was compiled by Rachael Tatman and released on Kaggle on September 13, 2017. It includes ***1.88 million geo-referenced wildfire records, representing a total of 140 million acres burned during the 24-year period.***

PREPROCESSING:

1) Dimensionality Reduction

The original dataset has 38 columns. The features are reduced to decrease the complexity of the model and to get more accuracy. Principal Component Analysis and Correlation Matrix is used to reduce the features set to important features in all the models. The features : FIRE_YEAR, FIRE_SIZE, STATE, LATITUDE, LONGITUDE, DISCOVERED_DATE, DISCOVERED_TIME, CONT_DATE, CONT_TIME

2) Features addition

It turned out that the date feature for each row is given in Julian date format in the original dataset. So, two new columns are created in the Gregorian date format - Discovery date and contained date are used to get total fire time. Day of week is added based on experiments that certain fire causes like arson occur more on weekends. Month is added based on experiments that certain fire causes like lightning occur more in specific months in certain states.

3) Dataset Sampling

The models are created on original dataset as well as reduced dataset. We created below 2 reduced datasets for 2 different models.

1) For reducing the dataset, stratified sampling is performed to capture balanced data from all classes and across all years.

2) A dataset is created with only 3 fire causes / labels - Lightning, Smoking, Arson

3) Since the model with 13 classes did not give good accuracy**,** in addition to label grouping we also created a model based on a reduced number of classes. 3 classes are chosen : Lightening, Smoking, Arson for classification model.

4) Label grouping

There are a total of 13 fire causes. The number of classes are reduced by grouping similar causes into 1 category to get better accuracy. For example, Lightening, Campfire, children are grouped into *Natural* , Arson into Malicious, etc.

5) Missing Values imputation

Below columns are imputed for missing values:

Discovery Time - 0.33 missing values. Values are replaced by mean of all time values on the same discovery date.

Contained Date and Contained Time: The strategy adopted to fill the missing values for both the columns is that the rows with similar values for columns like DISCOVERY_DATE, FIRE_SIZE_CLASS and STAT_CAUSE_DESCR tend to have the same contained date and time. Groupby functionality is used to implement the same.

County: Python Geopy API is used to impute the missing counties. The Geopy provides a geolocator to determine the county based on the longitude and latitude. The dataset has gps location available that helped us to use the geopy package to find the county.

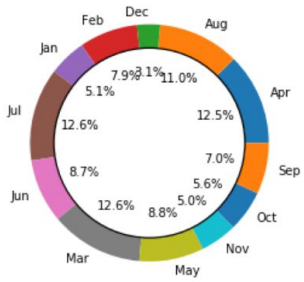
METHODOLOGY FOLLOWED:

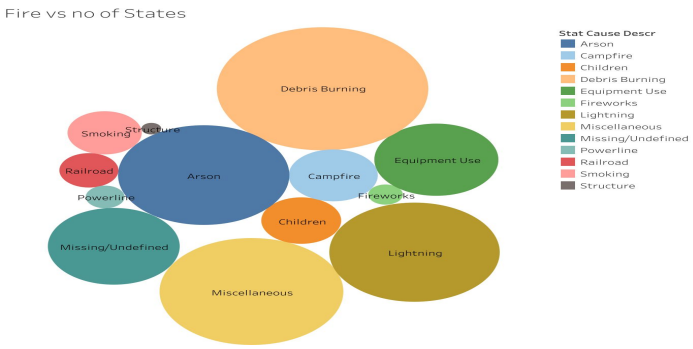
Cross Fold Validation - 10 fold cross validation is used on the stratified samples dataset (with 376K entries) to flag problems like overfitting or selection bias.

GRAPHS:

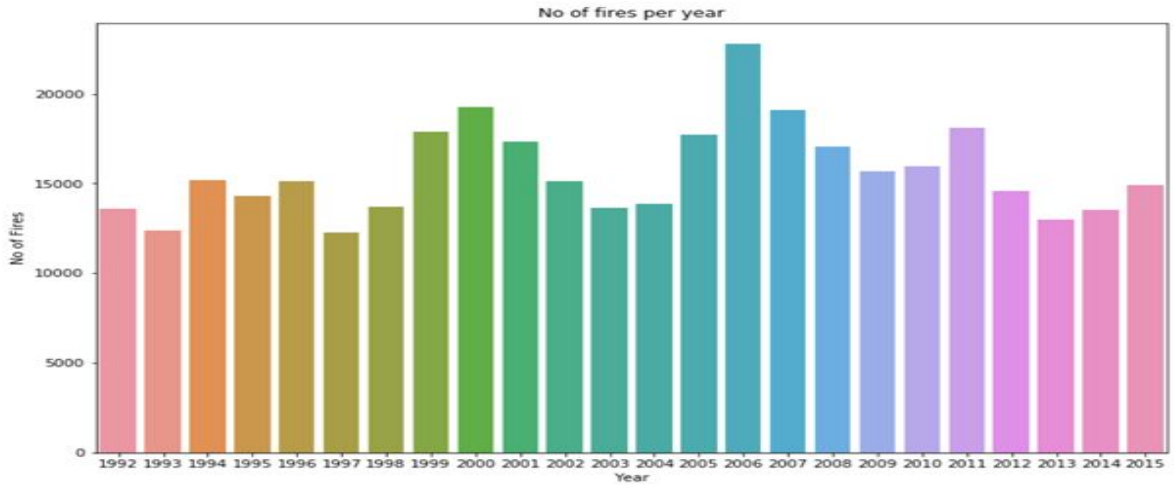
The visualizations were largely created using libraries such as matplotlib, seaborn, geopandas.

A donut chart is made to depict the month wise distribution of fires.

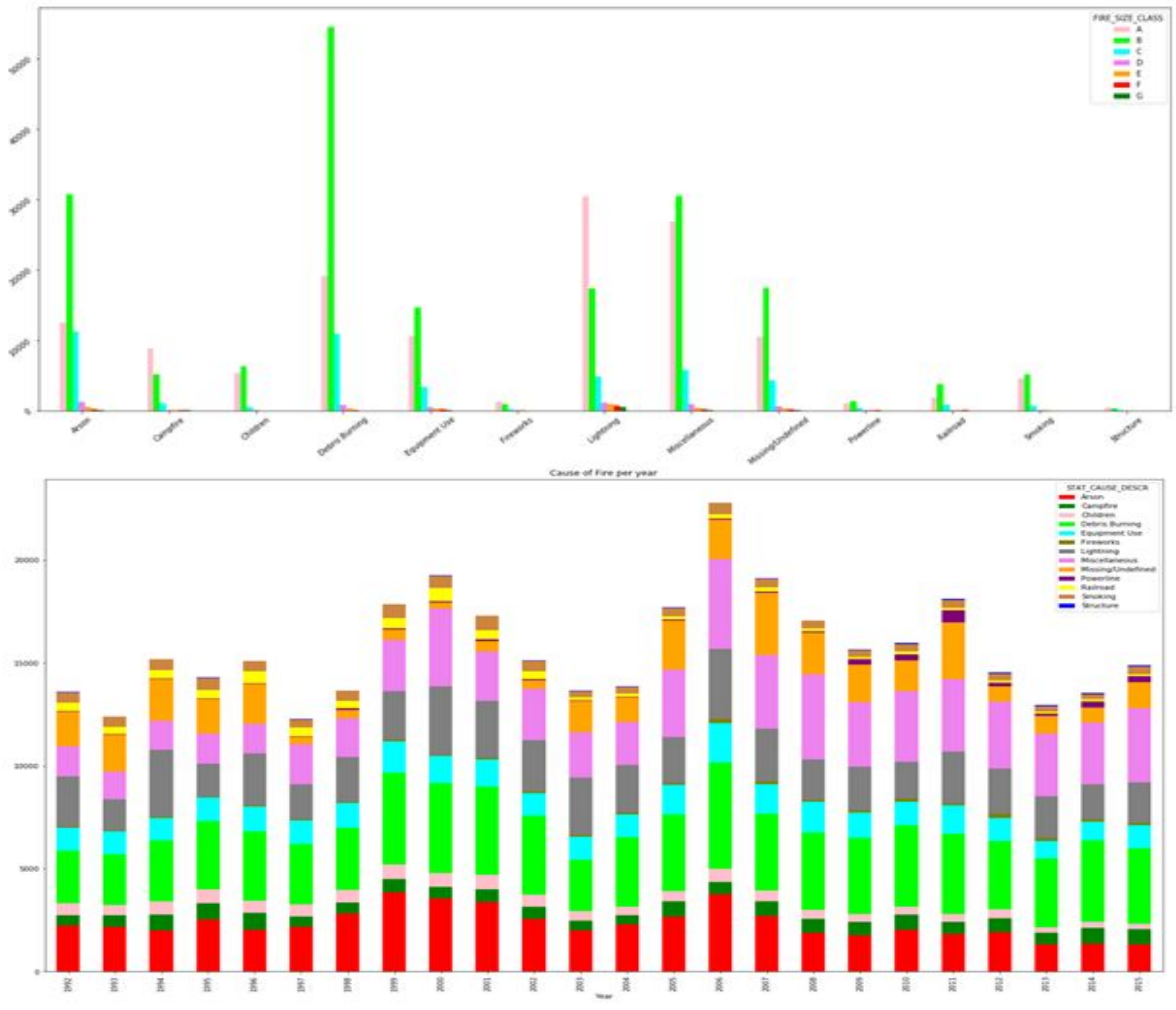




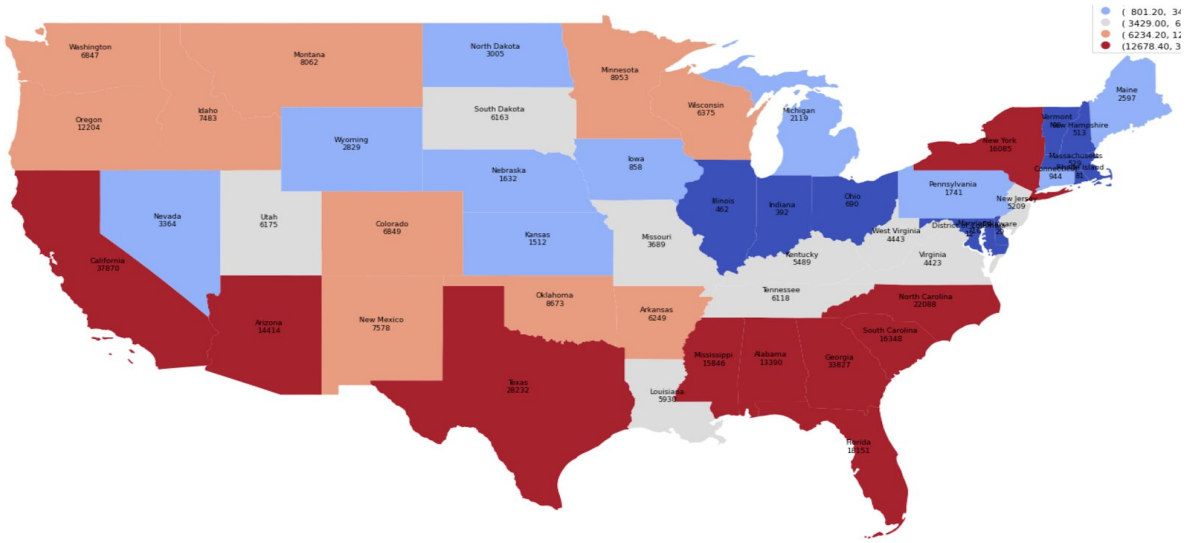
Bar graphs are made to understand the growth in fire cases over the years.



Cause of fire per year and the area affected by each fire cause.



Number of fire incidents across the US



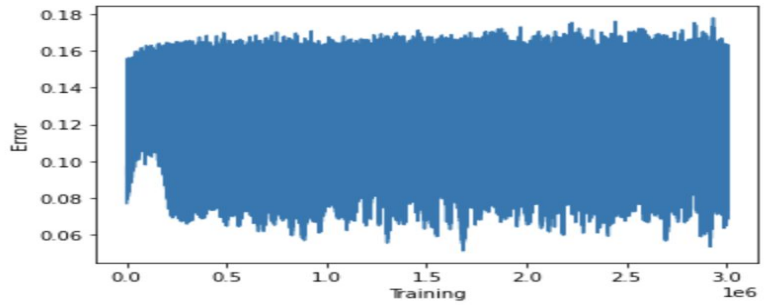
ANALYSIS OF RESULTS:

Below are the results for various approaches tried to predict fire cause:

1) For stratified sampled dataset with ~300K samples and 13 fire causes taken across all the years and fraction of all fire causes, the results are as follows: XGBoost: Training accuracy : 55%, Test accuracy : 50%, Decision Trees and Random Forest: Accuracy is approximately 47% for both models.

Neural Network:

```
no of epochs =  10
```



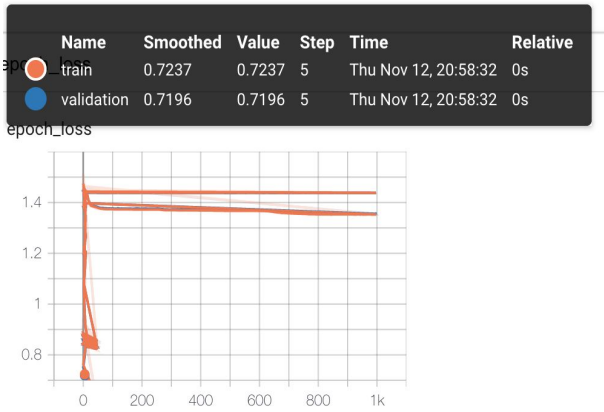

```
Training Accuracy 86.86%
Test Accuracy 86.89%
```

2) Neural Network with complete dataset of 1.8 million and classes grouped to result into small number of labels:

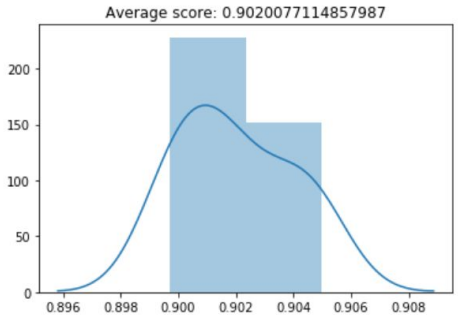
Test Results:

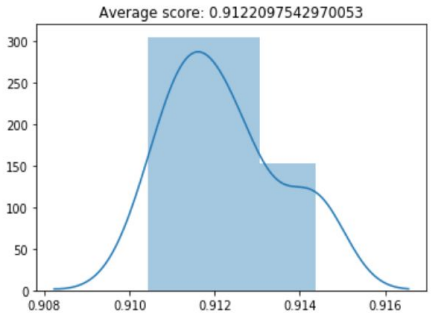
Activation Func	Test Loss	Test Accuracy	
Relu	0.83	0.713	
Tanh	0.82	0.723	

Training-Validation Results:



3) Decision Trees and Random forest with 376K dataset with original labels and also grouped labels: The first method has 5 labels which are ['Lightning', 'Arson', 'Powerline/Railroad/Children/Structure/EquipmentUse', 'Campfire/Fireworks/Smoking/DebrisBurning', 'Miscellaneous/Missing/Undefined']. Accuracy increased to 56% and 66% for decision tree and random forest respectively for the first method. In the second method, the categories are condensed into four possible causes: Natural, Accidental, Malicious, and Other. The accuracy for decision tree and random forest substantially increased to around 90%.





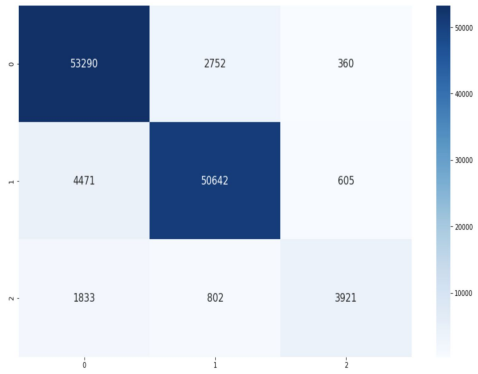
	precision	recall	f1-score	support
1	0.72	0.65	0.69	11033
4	0.94	0.96	0.95	64188
accuracy			0.91	75221
macro avg	0.83	0.81	0.82	75221
weighted avg	0.91	0.91	0.91	75221

	precision	recall	f1-score	support
1	0.66	0.69	0.67	11033
4	0.95	0.94	0.94	64188
accuracy			0.90	75221
macro avg	0.80	0.81	0.81	75221
weighted avg	0.90	0.90	0.90	75221

4) XGBoost with 600K dataset that includes samples with only three causes: Lightning, Arson and Railroad

Training accuracy: 91%, Test accuracy: 90 %

	precision	recall	f1-score	support
Arson	0.89	0.94	0.92	56402
Lightning	0.93	0.91	0.92	55718
Railroad	0.80	0.60	0.69	6556
accuracy			0.91	118676
macro avg	0.88	0.82	0.84	118676
weighted avg	0.91	0.91	0.91	118676



SECTION 4: DISCUSSION AND CONCLUSION:

DECISIONS MADE:

1) It was important to come up with a proper subset of the dataset. We chose to take a subset with samples across all the years and 15 percent of all the fire causes to get a balanced dataset.

2) We chose to build multiple models so that we can compare and learn the models better.

DIFFICULTIES FACED:

1) With thirteen classes, we could achieve only 50 % accuracy with all the models. Hence, we chose to work with less number of labels by reducing and grouping the labels.

2) The sparseness of the dataset was one of the challenges we faced while working on the project. Initial decisions include exploring subsets of data to avoid computational complexity. Later the whole dataset was also incorporated for some models.

THINGS THAT WORKED WELL:

Working with different classification models and different samples of the dataset worked out well because we could compare the results of all approaches and learn from that.

THINGS THAT DIDN'T WORK WELL:

It was difficult to train the data on the models because of the number of labels and presence of string categorical values. Experiments with one hot encoding to convert the string categorical values lead to substantial increase in the number of features which made the training process for models even more difficult.

FUTURE WORK:

The dataset can be used to analyze the cause of fire and predict the occurrence of fire in future. Further, it can also be used to analyze the major causes of fire in every state. This can help the authorized departments to take precautionary measures to reduce the incidences in future.

CONCLUSION:

In this project, we implemented various machine learning algorithms to predict the cause of fire. We were able to optimize the algorithm using several techniques. Moreover, we performed exploratory and visual analysis on the dataset. We evaluated the model based on classification metrics, accuracy and error.

References:

[1] De Bem, Pablo Pozzobon, De Carvalho Júnior, Osmar Abílio, Matricardi, Eraldo Aparecido Trondoli, Guimarães, Renato Fontes, & Gomes, Roberto Arnaldo Trancoso. (2019). Predicting wildfire vulnerability using logistic regression and artificial neural networks: A case study in Brazil's Federal District. *International Journal of Wildland Fire, 28*(1), 35.

[2] Sayad, Younes Oulad, Mousannif, Hajar, & Al Moatassime, Hassan. (2019). Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal, 104*, 130-146.

SECTION 5 : PROJECT PLAN / TASK DISTRIBUTION:

Garima Chaphekar	1. Dataset selection, preprocessing, imputation for missing values. 2. Visualization 3. Decision Tree and Random forest modelling 4. Experiments with other models like KNN, Bagging classifier, Naive Bayes . 5. Documentation
Saliha Mehboob	1. Data Preprocessing 2. Visualization 3. Modeling ● ANN (Keras, Tensor flow, Feedforward) ● Logistic Regression (Scikit learn, keras) 4. Evaluation 5. Documentation Github: *https://github.com/salihasjsu/cmpe255Project/tree/saliha_work*
Vidita Vijaykumar Daga	1. Dataset sampling, preprocessing, imputation for missing values, visualization for fires across states on US map 2. Class grouping and class reduction strategies 3. XGBoost classifier on sampled dataset and with class reduction 4. Experiment with KNN classifier 5. Documentation
Github : https://github.com/salihasjsu/cmpe255Project	