Big Data driven Wildfire Risk Analysis and Prediction

Aditee Vasant Jadhav
Dept. of Computer Engineering
San Jose State University
San Jose, USA
aditeevasant.jadhav@sjsu.edu

Parineeta Kishore patil
Dept. of Computer Engineering
San Jose State University
San Jose, USA
parineetakishore.patil@sjsu.edu

Sai Kumar Reddy
Dept. of Computer Engineering
San Jose State University
San Jose, USA
saikumarreddy.sandannagari@sjsu.edu

Sheshank Makkapati
Dept. of Computer Engineering
San Jose State University
San Jose, USA
sheshank.makkapati@sjsu.edu

***Abstract*—The increase in number and intensity of wildfire incidents in California has affected social, ecological and economical systems adversely. This had induced the need of machine learning techniques to learn from historical data and dig out patterns and behavior of wildfire. Such analysis of data is useful for building up strategies to prevent such incidents. In our project, we proposed and evaluated use of four models against different types of datasets to yield results of wildfire risk assessment and prediction. The four models used are- Logistic Regression, Decision Tree, Random Forest and Adaboost. We analyzed datasets using k-fold cross validation techniques to improve performance of models used. Also, we analyzed datasets to dig out significant features and the knowledge hidden under those features. Our models performed well to yield accuracy up to 85%.**

Keywords—wildfire, logistic regression, decision tree, random forest, adaboost, k-fold cross validation

I. INTRODUCTION

In last few years, the changes in climate has affected the patterns and features of wildfire incidents observed all over the world. The increase in wildfire incidents in California in the terms of magnitude, frequency, scale and time duration has elevated the risk to social and ecological systems. The rise in temperature that induces changes in water availability along with changes in climate, results into intensification of wildfire incidents. Many other factors affect the behavioral patterns of wildfire [1]. In order to adapt the management of land and other resources, it is important to gain better insights into the patterns and spread of wildfire incidents in California. The risk and high cost imposed on social and economic system due to uncontrolled widespread of wildfires has motivated us to perform wildfire risk analysis rather than focusing on its classification after the incident of wildfire has occurred. In this project we focus on developing methods for better understanding and analysis of wildfire risks using historical wildfire data, weather data and severity data depicting patterns of burnt areas and causes of wildfire. We input these types of data over years from 2014 to 2017 to four models: Logistic regression, decision tree, Adaboost and Random forest. Our output of models and analysis highlights on potential causes of wildfire that play significant role in determining future strategies to avoid wildfire risks and also it helps us getting comparative study among our models for different types of datasets.

The study of wildfire patterns and behavior before performing risk assessment on our dataset, we studied and collected theories about wildfire and then compared our results with theories. In this project, we successfully analyzed datasets and conclusions from our models almost matched with study done prior to risk analysis. The brief description of our theories can be given as below.

Wildfire commonly is the fire which usually occurs in the forests and damages the natural vegetation. These fires are divided based on the places they occur. They are divided as bush fires, forest fires, desert fires, peat fires etc. Once the wildfire occurs it is not easy to stop them and the fire spreads very quickly causing huge amount of destruction and leading to death [2]. In 2018, California suffered $400 billion in damage. It cost the California fire department $1 billion. These statistics clearly show the worst effects of the wildfire. So there are various causes for the occurrence of wildfire. The main causes are:

- Shifting cultivation: Agricultural practices such as slash and burn or shifting agriculture cause the intentional wildfire by humans. This is done to create a clear land of vegetation and can be used for cultivation. The land loses the fertility after continuously cultivating on the same land so this shifting cultivation is performed and people shift to new places and perform cultivation. This is one of the important reasons of wildfire occurrence.
- Human Activity: Most of the fires in United States, about almost 85% of the fires are caused by humans. They may be because of the carelessness or due the voluntary involvement in causing the fire. The reasons may be unattended campfires , the burning of debris and equipment use. The intentional acts of arson is also the main reason for the wildfire caused by humans.
- Cigarette Butts: The carelessness of the cigarette smoker can be the reason for he loss of many lives and loss of vegetation. Cast away cigarette butts that are unattended are the main cause for the wildfire occurrence. It is the responsibility of every smoker to ensure that their negligence doesn't cause the wildfire and damage the vegetation and lives.
- Arson: Many wildfires have occurred due to intentional acts of the people to gain something. Such an act is called Arson. For example, people might set fire to their own property to gain insurance. This kind of fire is uncontrollable and can cause huge damage.
- Fireworks: Fireworks displays must be held at safe locations. Often, an irresponsible act by an amateur

person can lead to devastating consequences. If there is even a little chance that fireworks might start a wildfire at a particular venue, the activity must be avoided.

- Campfires: Camping comes under the fun activity which causes wildfire. If such fires are not properly dealt, then they may be dangerous. Thus, it is better if the camping is done far away from the forests. After the need for the fire is over, it must be completely put off with water.
- Burning Debris: Burning of waste could often be the reason for the wildfire when it is supported by strong winds. It is important to ensure that fire sparks do not fly out and land elsewhere where fire can erupt again.
- Machinery or Equipment Generated Fires: There is also a possibility of fire occurrence due to machinery or automobiles located near forest land. These fires spread quickly and damage large areas. Plane crashes and automobile fires can also cause damaging effects.
- Irresponsible Logging Activities: Clearance of forest land for logging purposes can encourage the dominance of flammable gases. Also, logging roads that have been abandoned might be populated by such vegetation and act as fire corridors. So logging also contributes for the wildfire in this way.

Natural causes of Wildfires: Nature also plays a major role in triggering the wildfire. Some of the natural causes of such fire include :

- Dry Lightning: Dry lightning is one of the common natural causes of wildfires. There is no precipitation during this type of lightning. When lightning strikes a tree, it can produce a spark that can start a fire.
- Dry Climate/Drought: Dry climatic conditions or drought can lead to wildfires. During this time, there is little moisture in the air and on the ground, and the vegetation also dries up. Thus, the conditions are ideal for starting a fire.
- Volcanic Eruptions: Volcanic eruptions are highly destructive. One of the major destructions caused by them is wildfire. Burning lava from a volcano can reach forests to burn down everything. Wildfires caused by volcanic eruptions are thus of catastrophic nature.

However, there are many reasons for the occurrence of Wildfire, there is a necessity of Wildfire prediction and risk analysis. The prediction and risk analysis will be one of the key solutions to all these natural and human causes of the wild fire. So a Machine Learning fire prediction model is used to do the risk analysis and prediction. The prediction is done considering the text data and image data. The risk analysis is done considering the different fire occurrences data from 2014 to 2017.

The remaining report is organized as follows: Section II gives description of roles assigned in project team. Section III describes literature survey. Section IV presents training and testing data preparation and data preprocessing used in project. In Section V, we explain selected models and justification for their selection in details. Experimental results, comparison of results among models and lessons learned are discussed in Section VI, VII and VIII respectively.

II. Project team and roles

In our project team, there are four members and all contributed towards planning project experiments, collection of datasets and writing report. We divided datasets to perform different analysis among team members. Each team member collected wildfire historical data for different years and performed analysis for respective data using four different models: Logistic regression, Decision tree, Adaboost and Random forest. Each team member also performed k-fold cross validation analysis on one model each. Finally we compared results obtained for each team member to conclude wildfire risk analysis using data from year 2014 to 2017.

III. Literature survey

Most of the previous work in wildfire detection and risk analysis has focused on burned area and fire severity post fire incident has happened. The related work in the field of wildfire detection and prevention can be described in details as follows:

Caroline Famiglietti et.al. [1] proposed a system that takes into account precipitation, temperature, evaporation, soil moisture, wind speed to characterize the patterns of wildfire and predicts wildfire risk into certain areas using three models: logistic regression, gradient boosted decision trees and multilayer perceptron. The experiments were carried out on weather dataset for dedicated regions. Michael A. Madaio [2] proposed and constructed predictive model for fire risk which categorizes fire risk based on previous incidents of fire for the given property over a certain period of time and predicts likelihood of fire for given location. The major part of proposed system included experiments based on historical data of fire incidents.

Christopher D. O'Connor [3] suggested predictions based model for wildfire detection which uses boosted logistic regression model that focuses on areas which can be used for controlling the widespread of wildfires. The experiments were carried out on satellite data to prove the hypothesis. Mauro Castelli et.al. [4] proposed a model that works on meteorological data and geometric semantic genetic programming that focuses on finding best fitness function for predicting fires using burned areas data.

Mahsa Salehi et. Al. [5] built a unsupervised dynamic wildfire danger prediction model that addresses the aforementioned challenge using context-based anomaly detection (CBFR) techniques. The model is trained over weather data for period of observed wildfires in California county. Stojanova Daniela et.al. [6] proposed a better learning model that uses predictive models based on the forest structure GIS (geographical information system), the weather prediction model - Aladin and MODIS. They applied logistic regression and decision trees (J48) as well as random forests, bagging and boosting of decision trees. This model is trained using satellite data.

Cortez Paulo et.al. [7] proposed a system that learns using meteorological data and predicts fires. The proposed model uses Support Vector Machines, Random Forests, and four distinct feature selection setups such as using spatial, temporal, FWI components and weather attributes. Anupam Mittal et.al. [8] proposed a fire detection model that trains on sensor data and detects fire incidents using different machine learning techniques like Support Vector Machine, Artificial Neural Network, Decision Tree, Feed Forward Neural Network..

IV. DATA PREPARATION

A. Training and testing data with samples

Our dataset consists of historical wildfire data, fire incidents severity of burns data and weather data. We collected historical fire data over years from 2014 to 2017. The dataset was split into training dataset and testing dataset by randomly selecting data instances in the ratio 70:30 as training data : testing data. The overall number of data instances for historical wildfire data are about 36k whereas severity burns data has 21k data instances.

The number of features in historical data are 10, while severity burns data contains 15 features. The features in all datasets belong to one of types: numerical values, categorical data, continuous data values and object data types. The metadata and data samples can be shown as:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21673 entries, 0 to 21672
Data columns (total 10 columns):
OBJECTID        21673 non-null int64
ACRES           21673 non-null int64
FIRE_ID         21673 non-null object
FIRE_NAME       21673 non-null object
YEAR            21673 non-null int64
STARTMONTH      21673 non-null int64
STARTDAY        21673 non-null int64
SHAPE_Length    21673 non-null float64
SHAPE_Area      21673 non-null float64
FIRE_TYPE       21673 non-null int64
dtypes: float64(2), int64(6), object(2)
memory usage: 1.7+ MB
```

Fig. 1 Metadata for dataset

B. Data preprocessing and validation

Our dataset contains many features and multiple class labels. The data types vary from numerical type to categorical type. However, some features contain object data type which may not be processed by some machine learning models. Due to such reasons, we need to perform data preprocessing steps. In our dataset, some data types were of type object, which we converted into NumPy array format as a part of data preprocessing phase. Also some feature contained information that does not contribute towards final results, hence we dropped such feature from dataset.

V. MODELS USED

A. Logistic Regression

Logistic Regression is one of the analyses which perform regression in order to use when the output variable is binary. Logistic regression is a predictive analysis as most of the regression analyses are . Logistic regression is used in describing the data and to explain the relationship between one output binary variable and one or more input or independent variables. Logistic Regressions are not always easy to interpret, it sometimes becomes a daunting task. The statistics tool is the one which allows us to interpret.

Logistic regression major assumptions:

- The output variable must be dichotomous i.e. Binary in nature (e.g., yes vs. no).
- There must be no outliers in the data, outliers are the ones which are misclassified in a classification analysis. These outliers can be assessed by changing the continuous predictors to standardized scores, and removing the values below and above a particular value.
- There must be no huge correlation among the predictors. Because Multicollinearity may also affect the performance of the Logistic regression model. The correlation matrix among the predictors can be used for this. The major task of Logistic regression is estimating the log odds of an event.

One of the solutions for classification is logistic regression. The logistic regression model uses the logistic function to fit the output of a linear equation between 0 and 1 instead of fitting a straight line or hyperplane. For classification, we use probabilities between 0 and 1, so the right side of the equation is converted into the logistic function. While selecting a model for the logistic regression, the most important thing is to consider the model fit. Addition of input variables to a logistic regression model will always result in the increase in the amount of variance explained in the log odds (typically expressed as R^2). However, adding more and more variables to the model can result in overfitting. This overfitting reduces the generalizability of the model beyond the data on which the model is fit.

Justification for selecting Logistic regression: Logistic regression is superior classifier on categorical data when compared to linear regression. It uses a cross-entropy error function instead of least squares which is used in Linear regression. Therefore, it isn't that sensitive to outliers and also doesn't be harsh to correct data points like least-squares method does. All these advantages over linear Regression makes Logistic Regression one of our models to predict wildfire using weather data. These advantages of the Logistic regression i.e. the superiority over the Linear Regression model makes this model best fit for our data being used.

B. Decision Tree

Decision tree regression is a process of regression , which iterates and splits the data into many branches, and the splitting process continues to make the data into smaller groups as the we move up each branch. Initially, training dataset are supposed to be one group. The algorithm then starts partitioning for every possible binary split in every feature. The split that minimizes the sum of squared deviations from the mean is selected to categorize data in two separate partitions. This process is applied to each new branch and is continued until every node reaches a threshold value of minimum node size and becomes a end node. As tree is obtained from the training data set, a fully grown regression tree may suffer from overfitting problem which yields poor performance. The regression tree is generally

easy to understand and it can also be modelled automatically for given interactions. This allows the model to develop an internal representation of relationships between input data features. Also regression tree is known for handling missing data easily. Whereas other models such as Logistic Regression, Linear Regression, Random forests may fail to handle missing data. Logistic regression also does not give the efficient results if target variable in non-discrete. So regression tree which uses continuous target variable is used to give efficient results.

Justification for selecting Decision tree: Decision tree is a type of machine learning algorithm that is used primarily to classify dataset when the target variable involves continuous values instead of discrete values. The dataset used in this project involved data instances for which target variable is continuous. Also dataset may involve data points for which values may be missing. Regression tree is known for yielding better performance even in scenario of missing values case. Also it will generate easy to understand representation of complex dataset used for this project. So for the given dataset we use regression tree for prediction of fire.

C. Adaboost

Adaboost classifier helps the weak classifier algorithm to form strong classifier. A single algorithm may be poor at classifying the objects properly. But the combination of multiple classifiers along with selection of training set at every iteration and giving proper amount of weight in final voting i.e. the last node, can have good accuracy score for overall classifier. The main purposes of using Adaboost is: The algorithm is retrained iteratively by selecting the training set based on accuracy of previous training. So the accuracy of the previous training plays a key role in this process. The necessity of each trained classifier at any iteration depends on the achieved accuracy.

Justification for selecting Adaboost: We choose Adaboost technique to boost the performance of the classifiers we have already used on the data. So this model act as a conjunction with different models to improve performance. Though Adaboost is sensitive to outliers, it sometimes is very less vulnerable to overfitting issue. As this model also avoids the overfitting issue, it is used for our project in order to reiterate and boost the performance of all the earlier used models.

D. Random Forest

Random Forest is a popular and easy to use machine learning algorithm that gives great predictions. It can be used for classification as well as regression purposes. It is a supervised learning algorithm that contains multiple Decision Trees and combines them to get a more accurate and stable result. It is an ensemble of decision trees. Random forests are the combination of different decision trees. This Random forest can be mainly used for classifying data.

Steps involved in Random Forest pseudocode are as follows:

- Select “N” features randomly from total features which are assumed to be K, where N << K.
- Choosing the best split point is important and then calculating node “d” among the “N” is done.
- Split the node into child nodes using the best split.
- Repeat the above 3 steps until “l” number of nodes has been reached.
- Repeat all the above steps “n” number of times to create “n” number of trees.

The number of default hyperparameters associated with Random Forest is not very high, it is considerably low. These are easy to understand as the parameters are straightforward,. Also, Random Forest can avoid the problem of “overfitting”. Overall, it is a simple and flexible tool with certain limitations. Random Forest is a flexible, easy to use machine learning algorithm that produces result with great accuracy. It is simple and can be used for both regression and classification. Hence, it is one of the most used algorithms. Random forest creates multiple number of decision trees and merges them together to achieve more accurate results. Also, it is easy to measure the relative importance of each feature on the prediction. Hence, random forest is used extensively in various fields such as stock market, banking, medicine and e-commerce.

Justification for selecting Random forest: Random forest is a very handy and easy to use algorithm. The number of hyperparameters used in this algorithm is not that high. Also, they are straightforward and simple to understand. In case of random forest, the probability of occurrence of overfitting issue can be reduced by using enough number of trees in the analysis. Also, these algorithms are fast to train. Overall, random forest is a fast, flexible and highly accurate tool which is suitable for the given dataset.

VI. EXPERIMENTAL RESULTS

The results obtained after applying classification models over dataset are:

The output summary for each model is obtained as follows:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1625
1	0.00	0.00	0.00	1036
2	0.58	1.00	0.73	3777
3	0.00	0.00	0.00	64
micro avg	0.58	0.58	0.58	6502
macro avg	0.15	0.25	0.18	6502
weighted avg	0.34	0.58	0.43	6502

Accuracy of logistic regression model is : 0.5808981851737927

Fig. 2 Performance of Logistic Regression model

	precision	recall	f1-score	support
0	0.77	0.79	0.78	1625
1	0.68	0.68	0.68	1036
2	0.85	0.83	0.84	3777
3	0.10	0.14	0.12	64
micro avg	0.79	0.79	0.79	6502
macro avg	0.60	0.61	0.60	6502
weighted avg	0.79	0.79	0.79	6502

Accuracy of decision tree model is : 0.7882190095355275

Fig. 3 Performance of Decision tree model

	precision	recall	f1-score	support
0	0.67	0.60	0.63	1625
1	0.47	0.53	0.50	1036
2	0.81	0.82	0.81	3777
3	0.16	0.09	0.12	64
micro avg	0.71	0.71	0.71	6502
macro avg	0.53	0.51	0.52	6502
weighted avg	0.71	0.71	0.71	6502

Accuracy of adaboost model is : 0.7116271916333435

Fig. 4 Performance of Adaboost model

	precision	recall	f1-score	support
0	0.85	0.81	0.83	1625
1	0.84	0.70	0.76	1036
2	0.85	0.92	0.88	3777
3	0.62	0.08	0.14	64
micro avg	0.85	0.85	0.85	6502
macro avg	0.79	0.63	0.66	6502
weighted avg	0.85	0.85	0.85	6502

Accuracy of random forest model is : 0.8503537373115965

Fig. 5 Performance of Random Forest model

K-Fold cross validation: The K fold cross validation in this case is used to compare all the models used. The different models used on our dataset are being compared using the k-fold cross validation technique. K fold cross validation is used evaluate machine learning models by resampling. So, the resampling process takes place when K fold cross validation is performed. The variation of the value of K with respect to all the machine learning models used gives the above results.

We evaluated our models against different values of k for k-fold cross validation while training phase of models. The value of k ranges from 5 to 50. The results obtained by performing k-fold cross validation for different values of k over our models can be graphically demonstrated as:

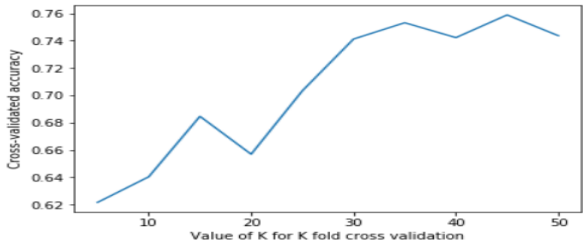


Fig. 6 K-fold for Decision Tree model

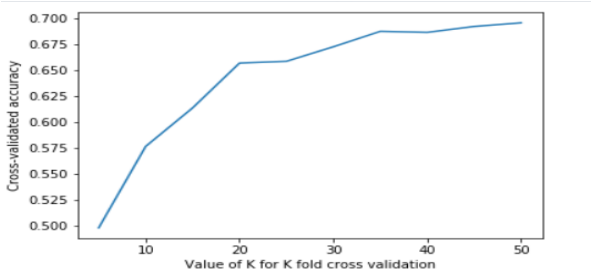


Fig. 7 K-fold for Adaboost model

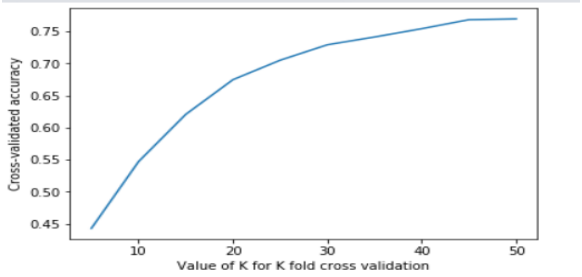


Fig. 8 K-fold for Random Forest model

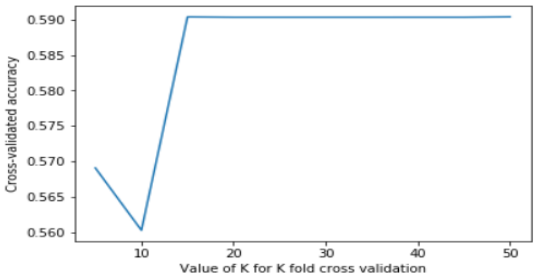


Fig. 9 K-fold for Logistic Regression model

Statistical analysis: The characteristics wildfire prediction results depend on two factors: cause of wildfire and detection method for wildfire. We collected list of all possible causes of wildfire for data of each year. Also, we computed the number of fire incidents observed for each of listed cause. The important causes of wildfire are Arson, Equipment use, debris and campfire.

The analysis of wildfire causes for each can be represented in statistical as well as graphical format as follows:

Miscellaneous	1945
Lightning	1412
Equipment Use	500
Arson	446
Debris	220
Campfire	175
Smoking	144
Vehicle	135
Powerline	123
Playing with fire	75
Escaped Prescribed Burn	37
Railroad	32
Structure	8
Non-Firefighter Training	7
Aircraft	5
Firefighter Training	4
Illegal Alien Campfire	4

Name: CAUSE, dtype: int64

14 - Unknown / Unidentified	6218
9 - Miscellaneous	1957
1 - Lightning	1475
2 - Equipment Use	508
7 - Arson	457
5 - Debris	224
4 - Campfire	180
3 - Smoking	144
10 - Vehicle	139
11 - Powerline	126
8 - Playing with fire	77
18 - Escaped Prescribed Burn	39
6 - Railroad	32
<Null>	13
15 - Structure	8
13 - Non-Firefighter Training	7
16 - Aircraft	6
12 - Firefighter Training	4
19 - Illegal Alien Campfire	4

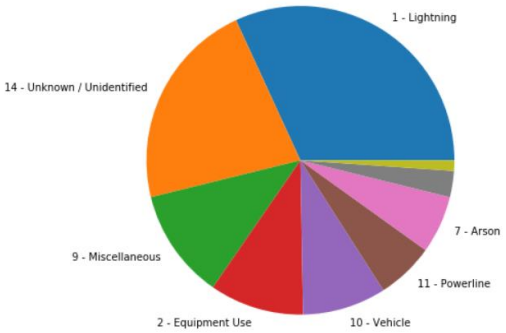
Name: CAUSE, dtype: int64

14 - Unknown / Unidentified	6327
9 - Miscellaneous	1975
1 - Lightning	1485
2 - Equipment Use	519
7 - Arson	466
5 - Debris	226
4 - Campfire	182
10 - Vehicle	148
3 - Smoking	145
11 - Powerline	129
8 - Playing with fire	78
18 - Escaped Prescribed Burn	40
6 - Railroad	33
<Null>	11
15 - Structure	9
16 - Aircraft	7
13 - Non-Firefighter Training	7
12 - Firefighter Training	4
19 - Illegal Alien Campfire	4

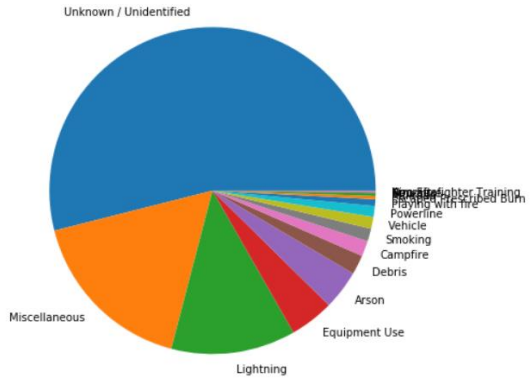
Name: CAUSE, dtype: int64

1 - Lightning	58
14 - Unknown / Unidentified	40
9 - Miscellaneous	21
2 - Equipment Use	18
10 - Vehicle	16
11 - Powerline	11
7 - Arson	11
5 - Debris	5
4 - Campfire	2

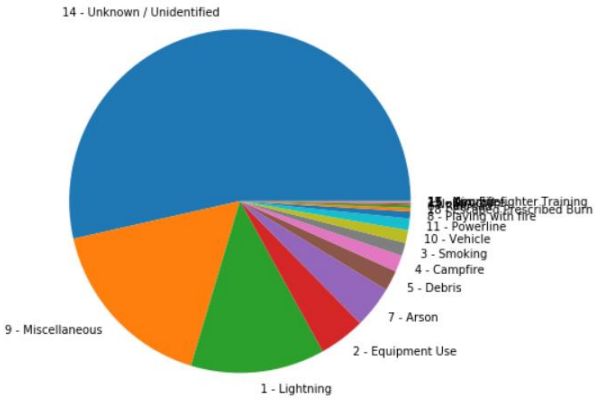
Name: CAUSE, dtype: int64



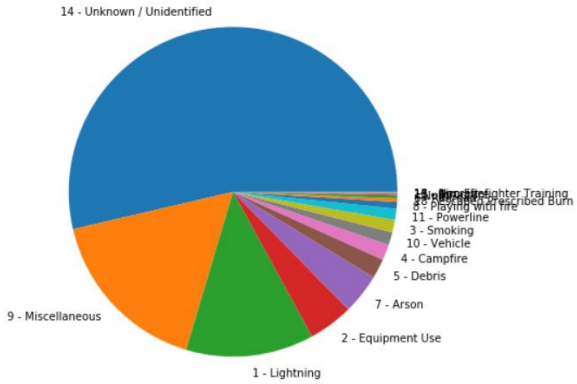
Distribution by cause-Year 2017



Distribution by cause-Year 2014



Distribution by cause-Year 2015



Distribution by cause-Year 2016

To dig more into details, we computed area burnt due to wildfire induced by each cause. After rigorous analysis of data over years, we conclude that some causes behind wildfire may induce lesser number of incidents, however the severity and area burnt due to them is high. Thus, while building strategy for wildfire prevention, the cause of wildfire as well as area and severity of burnt areas for each cause should be analyzed deeply with equal weightage.

CAUSE	GIS ACRES
Aircraft	6779.046400
Arson	4102.137128
Campfire	6684.979314
Debris	2538.800750
Equipment Use	3336.077120
Escaped Prescribed Burn	1432.523432
Firefighter Training	723.627000
Illegal Alien Campfire	235.789500
Lightning	4358.916533
Miscellaneous	3426.099812
Non-Firefighter Training	1580.487286
Playing with fire	1493.932920
Powerline	2095.967171
Railroad	2867.139688
Smoking	2447.040986
Structure	921.300250
Unknown / Unidentified	1865.653559
Vehicle	1449.117844

CAUSE	GIS_ACRES
1 - Lightning	4565.722109
10 - Vehicle	1376.119417
11 - Powerline	2069.046421
12 - Firefighter Training	723.625975
13 - Non-Firefighter Training	1580.487757
14 - Unknown / Unidentified	1888.582182
15 - Structure	921.300238
16 - Aircraft	5699.175667
18 - Escaped Prescribed Burn	1375.800846
19 - Illegal Alien Campfire	235.789225
2 - Equipment Use	3309.897749
3 - Smoking	2447.041819
4 - Campfire	6634.182877
5 - Debris	2496.410058
6 - Railroad	2867.140853
7 - Arson	4032.465676
8 - Playing with fire	1531.910621
9 - Miscellaneous	3470.238511
<Null>	3814.782231

CAUSE	GIS_ACRES
1 - Lightning	4546.653735
10 - Vehicle	1632.806195
11 - Powerline	2028.048067
12 - Firefighter Training	723.625975
13 - Non-Firefighter Training	1580.487757
14 - Unknown / Unidentified	1883.832158
15 - Structure	848.515556
16 - Aircraft	4930.116343
18 - Escaped Prescribed Burn	1351.131550
19 - Illegal Alien Campfire	235.789225
2 - Equipment Use	3324.904691
3 - Smoking	2431.070361
4 - Campfire	7301.258804
5 - Debris	2502.544664
6 - Railroad	2791.761509
7 - Arson	4061.283422
8 - Playing with fire	1519.207873
9 - Miscellaneous	3503.883578
<Null>	4486.085455

CAUSE	GIS_ACRES
1 - Lightning	7618.599107
10 - Vehicle	10828.515450
11 - Powerline	1271.504718
14 - Unknown / Unidentified	6698.855010
2 - Equipment Use	1534.036822
4 - Campfire	310.937000
5 - Debris	2274.384440
7 - Arson	4555.280709
9 - Miscellaneous	20416.709005

We listed out wildfire event detection methods from historical data and computed number of incidents detected by each method and area of burns covered by each method. Wildfire detection method plays important role in analyzing wildfire patterns. The visualization and statistical report of detection method analysis can be represented as:

<Null>	8446
8 - Unknown	1489
1 - GPS Ground	583
7 - Mixed Collection Methods	348
2 - GPS Air	342
6 - Hand Drawn	250
3 - Infrared	76
4 - Other Imagery	64
5 - Photo Interpretation	20

Name: C_METHOD, dtype: int64

<Null>	8443
8 - Unknown	1496
1 - GPS Ground	622
2 - GPS Air	356
7 - Mixed Collection Methods	351
6 - Hand Drawn	334
3 - Infrared	91
4 - Other Imagery	81
5 - Photo Interpretation	21

Name: C_METHOD, dtype: int64

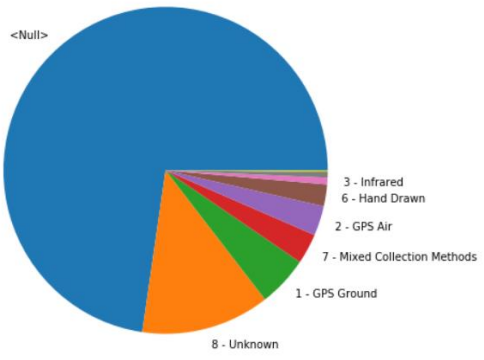
1 - GPS Ground	84
8 - Unknown	24
7 - Mixed Collection Methods	23
3 - Infrared	21
2 - GPS Air	15
6 - Hand Drawn	14
5 - Photo Interpretation	1

Name: C_METHOD, dtype: int64

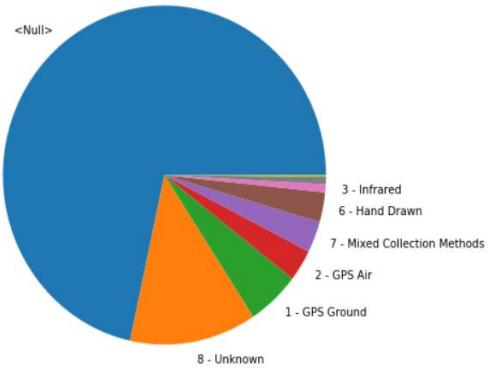
C_METHOD	GIS_ACRES
1 - GPS Ground	1089.226607
2 - GPS Air	4567.629006
3 - Infrared	19124.211342
4 - Other Imagery	1274.868681
5 - Photo Interpretation	5793.389725
6 - Hand Drawn	1964.983388
7 - Mixed Collection Methods	8091.025197
8 - Unknown	4510.257253
<Null>	2113.372745

C_METHOD	GIS_ACRES
1 - GPS Ground	1341.645408
2 - GPS Air	4577.072916
3 - Infrared	17121.438162
4 - Other Imagery	1762.808579
5 - Photo Interpretation	5543.280514
6 - Hand Drawn	1844.305678
7 - Mixed Collection Methods	8024.885956
8 - Unknown	4524.208567
<Null>	2113.970483

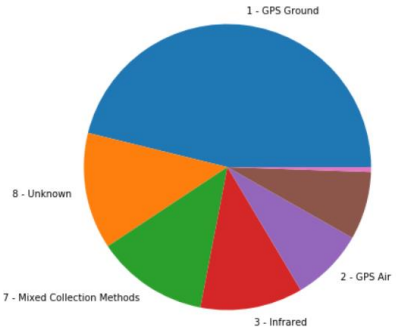
C_METHOD	GIS_ACRES
1 - GPS Ground	2201.596702
2 - GPS Air	5990.061253
3 - Infrared	21736.644448
5 - Photo Interpretation	702.306600
6 - Hand Drawn	5887.836114
7 - Mixed Collection Methods	20097.916404
8 - Unknown	5787.604954



Distribution by method-year 2015



Distribution by method-year 2016



Distribution by method-year 2017

VII. Comparison of Results Among Models

Decision tree regressor after performing on the considered dataset has given the accuracy of 78%. The accuracy obtained at this level is compared with the accuracy of the other models such that they can be compared. The Random forest gives more accuracy when compared with the Decision tree model. The accuracy in this case is about 85%. These results show that the Random forest model is performing effectively on the given data. The adaboost is basically used to boost the performance of the decision tree mainly. The adaboost model used increases the number of iterations and gives the better accuracy. The adaboost model considered for our data gives the accuracy of about 71%.

VIII. Experience and Lessons Learned

The statistical analysis demonstrates that there are instances where the occurred fire many times due to a particular cause, but the destruction caused by them is not that high. Whereas, there might be few causes of wildfire such as Lightning which do not occur so frequently but causes a huge damage.The results of k-fold cross validation experiments show that for every model, as we recursively apply k-fold cross validation, the performance of models go on increasing and after certain point the performance becomes stable and outstands over less trained models. Some models show exponential curve of the performances with increase in value of k. However, some of the models start losing performance after certain point and again increase performance accuracy up to certain value. Some models show fluctuating rate of accuracy for given range of k. This concludes that performance of models when trained using k-fold cross validation depends upon fact whether data overfits or under fits the models. If data fits properly, the performance graph goes on increasing in exponential curve.

CONCLUSION

Our work gives demonstration of the potential of machine learning models for better understanding of vulnerabilities leading to disastrous wildfire events in California. We used four models of machine learning: Logistic regression, Decision tree, Adaboost and Random Forest to different types of datasets collected for events of wildfires in California observed over years 2014 to 2017. We applied our models and analysis over historical wildfire data, severity depicting fire incidents data and weather data. We performed different experiments over k-fold cross validation techniques for our models and induced analytical results. In general, we were able to yield performance of models about 60-85% accuracy. This work can be extended by assisting learning models with real world data collected from sensors located remotely.

REFERENCES

[1] Caroline Famiglietti, Natan Holtzman, Jake Campolo, "Satellite-Based Prediction of Fire Risk in Northern California", 2018.

[2] Michael A. Madaio, "Predictive Modeling of Building Fire Risk", 2018.

[3] Connor Christopher D. O', Calkin David E., Thompson Matthew P., "An empirical machine learning method for predicting potential fire control locations for pre-fire planning and operational fire management",International Journal of Wildland Fire 26, 2017, pp.587-597.

[4] Castelli Mauro, Vanneschi, Leonardo, Popovic Ales, "Predicting Burned Areas of Forest Fires: An Artificial Intelligence Approach", Fire Ecology, 2015, pp.106-118.

[5] Salehi Mahsa, Rusu Laura, Lynar Timothy, Phan Anna, "Dynamic and Robust Wildfire Risk Prediction System: An Unsupervised Approach", 2016, pp. 245-254.

[6] Stojanova Daniela, Panov Pance, Kobler Andrej, Dzeroski Saso, Taskova, Katerina, "Learning to predict forest fires with different data mining techniques", 2006.

[7] Cortez Paulo, Morais A, "A Data Mining Approach to Predict Forest Fires using Meteorological Data", 2007.

[8] Anupam Mittal, Geetika Sharma, Ruchi Agarwal, "Forest Fire Detection Through Various Machine Learning Techniques using Mobile Agent in WSN", International Research Journal of Engineering and Technology, Vol. 3, Issue 6, June 2016, pp. 702-706.