EasyChair Preprint

№ 7605

Forest Fire Area Estimate of an Ongoing Fire

Nuzhat Shaikh, Pratik Temkar, Prem Kulkarni, Sakshi Chaudhari and Karan Wagh

March 23, 2022

Forest Fire Area Estimate of an Ongoing Fire

Dr. N. F. Shaikh, Pratik Temkar, Prem Kulkarni, Sakshi Chaudhari and Karan Wagh
Modern Education Society's College of Engineering
Pune, Maharashtra, India
nfshaikh@mescoepune.org; pratikstemkar@gmail.com; kulkarniprem04@gmail.com;
sakshee1603@gmail.com; karanwagh17@gmail.com

***Abstract*—The most common hazard in forests is wildfire. These fires present a challenge to the invaluable forest resources and the natural ecosystem which includes plants and animals. It drastically affects the biodiversity and ecology of a region. Wildfires are a bigger problem and they occur anywhere irrespective of the topology. A huge number of fire points were detected in India in the last three years. The increase in the rate of fires is approximately doubling every three years. There are multiple causes of forest fires including natural and manmade. Natural ones include soil erosion, change in climate, global warming, low humidity, high atmospheric pressures, dryness, etc. Taking into account the statistics from last year, India witnessed a 43 percent increase in the number of wildfires in the last decade(2009-2021), witnessing a 125% spike in such fires in a span of just 2 years from 2015 to 2017. Taking into consideration this serious issue we have developed a model which will help us predict the forest fires area when the forest fire is at an early stage of ignition and take appropriate measures. It provides two types of inputs, one via form and the other through SMS. Further, satellite data is integrated and the prediction is done. Machine Learning Algorithms along with GIS data is used to perform predictions. The results of the prediction are displayed on the map at respective coordinates. Once the spread is determined the nearby areas can be evacuated and the lives of many can be saved. Also, appropriate measures can be taken to control the spread of the fire.**

***Index Terms*—**
- **Forest Fire Prediction**
- **Machine Learning**
- **Decision Trees**
- **Gaussian Naive Bayes**
- **K-Nearest Neighbours**
- **Random Forest**

I. Introduction

MILLIONS of acres of forest are burned every year. These forest fires have a significant influence on flora loss, air pollution, and, most importantly, human life. In many circumstances, the authorities lack a fire warning system as well as an alert system for sending and receiving warning messages. As a result, warnings to the public and rescuers are frequently issued too late. As a result, the goal of this project is to create a fire alert system with more functionality for monitoring and detecting forest fires. These capabilities include the ability to gather data from the forest in order to analyze it and detect fires in their early phases. Furthermore, the massive acres of land that have been burned makes it exceedingly improbable that vegetation will grow on this land again. The badly burned soil becomes water-resistant, there is no plant left to hold the soil in place, and the earth can no longer absorb any more water, resulting in a decrease in groundwater level. To add to the gravity of the situation, when this soil is washed into rivers, it pollutes the water. When you consider how much vegetation is burned, you can't overlook the emissions that arise from the process. Forest fires are mentioned in the Global Warming Report 2008 as one of the key causes of global warming due to the massive amounts of greenhouse gases released into the atmosphere. Every year, flames of varied strength and extent ravage enormous regions of the forest. According to forest inventory statistics, 54.40 percent of India's forests are exposed to occasional fires, 7.49 percent to moderately regular fires, and 35.71 percent have not yet been exposed to fires of any significance, which is a major cause of biodiversity loss and environmental degradation. However, by creating a model that can replicate the fire in real-time, we can investigate the pattern and likely fire behavior. As a result, scientists and authorities will have an easier time managing the situation, taking measures, and minimizing the harm.

II. Literature Review

Machine learning models play a significant role in evaluating and predicting the results. Burned areas [1] in the forest were predicted using many methods such as the Radial Basis Function Networks, Fuzzy Logic, Multilayer Perceptron, SVM [2], etc. The results displayed that MLP gave more accurate results.[3] Moreover, Regression, a statistical analysis technique was used to predict forest fires.[4] This technique evaluated the relationship between numerous variables.[5] Linear regression could be single linear regression or multiple linear regression depending on the number of independent variables used in prediction.[6] If there exists a linear relationship between one independent variable and one dependent variable the regression is termed simple linear regression, and multiple is the number of independent variables is greater than two. Ridge regression is used when the independent variables are highly correlated.[7] This technique gave good results only when there was a correlation between variables. Regularization was used to prevent the model from overfitting.[8] It used L2 regularization which minimized the sum of squares of coefficients. In short, it minimized the parameters to reduce the complex nature of the model making it a little more efficient. Lasso regression also gave good results when the coefficients were few. For the three models, i.e. linear regression, ridge regression, and lasso regression two different implementations

were conducted, one including 100% i.e. all the features, and the other including 70% of the features. For the first implementation using 100% of the features, the accuracy percentages for linear, ridge ad lasso regression for the training data set were 100%, 98%, 88%, and for the testing data set they were 100%, 95%, and 81% respectively. For the second implementation, the accuracy percentage on the training data set was 99%, 76%, and 84% on the linear, ridge, and lasso regression respectively. The accuracy percentage on the testing data set was 99%, 79%, and 87% respectively. The results clearly demonstrated that the model using linear regression predicted the best results of all three. Along with machine learning a few Artificial Techniques were also used for Forest fire prediction.[9]

III. Data Set Information

The dataset was collected from Kaggle. It is the wildfire data of the United States of America from the year 1992 to 2015. The dataset consists of around 1 million rows and 39 features in all. Feature selection is dependent on how much is the attribute contributing to the final prediction. To decide if the feature is to taken or not to be taken into consideration an exploratory data analysis is performed while preprocessing the data.

1) OBJECT_ID: ID for each row.
2) FOD_ID: Global unique identifier
3) Shape: Tells the shape of the forest fire.
4) FIPS_NAME: Equivalent entries are represented.
5) FIPS_CODE: Three-digit code from the Federal Information Process Standards.
6) COUNTY: County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.
7) STATE: State is represented by two letter alphabet.
8) OWNER_DESCR: Land is managed by primary owner and that owner reports the incident in his territory.
9) LONGITUDE: Longitude for point of fire.
10) LATITUDE: Latitude for point of fire.
11) FIRE_SIZE_CLASS: The fire class size has the values divided into 7 classes as per area burned.
12) FIRE_SIZE: Determines final area or the perimeter of the fire.
13) CONT_TIME: Fire contained time in HH/MM.
14) CONT_DOY: Day on which the fire occured.
15) CONT_DATE: Date on which fire was contained in MM/DD/YYYY.
16) STAT_CAUSE_DESC: Description of the statistical cause of the fire.
17) STAT_CAUSE_CODE: Code for the statistical cause of the fire.
18) DISCOVERY_TIME: Time of day when the fire was confirmed to exist.
19) DISCOVERY_DOY: Day when the fire was confirmed.
20) DISCOVERY_DATE: Fire confirmation date.
21) FIRE_YEAR: Fire confirmation year.
22) COMPLEX_NAME: Name of the complex or the area under which fire was declared or confirmed.
23) MTBS_FIRE_NAME: MTBS perimeter dataset, name of the fire.
24) MTBS_ID = MTBS perimeter dataset, incident identifier.
25) ICS_209_NAME: Name of the incident, from the ICS-209 report.
26) ICS_209_INCIDENT_NUMBER: Identifier for ICS 209.
27) FIRE_NAME: Incident name from fire report.
28) FIRE_CODE: Code used to track and compile cost information for emergency fire suppression.
29) LOCAL_INCIDENT_ID: Local code for the incident.
30) LOCAL_FIRE_REPORT_ID: Local ID for the fire.
31) SOURCE_REPORTING_UNIT_NAME: Reporting agency unit name for the fire.
32) NWCG_REPORTING_UNIT_ID: Active NWCG Unit Identifier for the unit preparing the fire report.
33) NWCG_REPORTING_UNIT_NAME: Active NWCG Unit Name for the unit preparing the fire report.
34) SOURCE_SYSTEM: Source database identifier.
35) SOURCE_SYSTEM_TYPE: Type of source, federal, intragency or nonfederal.
36) SOURCE_REPORTING_UNIT: Code for the agency unit preparing the fire report.
37) FPA_ID: Unique ID to track back to the original record.

IV. Exploratory Data Analysis

Out of the 39 features, some of them include the identification numbers which are provided by different agencies. These identification numbers do not provide any relevant data related to the wildfires and therefore, can be omitted. The class label is then analyzed. The fire area is divided into seven different classes depending on the area that is damaged.

- 1/A: 0-0.25 acres
- 2/B: 0.26-9.9 acres
- 3/C: 10.0-99.9 acres
- 4/D: 100-299 acres
- 5/E: 300-999 acres
- 6/F: 1000-4999 acres
- 7/G 5000+ acres

After performing the first step of analyzing the dataset, it is observed that there are a total of 39 columns out of which 12 are numeric data types (8 floats + 4 ints) and, the rest are all objects which need to be converted into numeric forms for feeding into the machine learning models. Next, all the features were analyzed individually. For each feature the following steps were conducted:

- Unique values of the feature
- Length of the unique values
- Number of NULL entries
- If the data type is an object converting it to numeric form
- Graphical analysis to check the data distribution[10]

For feature 1, FIRE_CLASS_SIZE, the graphical representation showed that the dataset was highly imbalanced. Class 2 consisted of the maximum count of forest fires and class 7 had the minimum count. The 2nd feature was the OBJECT_ID which can be omitted in the final

dataset s it is just an identifier. The next two columns, FOD_ID, FPA_ID can also be omitted as they are also IDs. For the feature SOURCE_SYSTEM, a violin plot and box plot of fire size class and source system was generated which displayed that the Median value of most source systems is at Fire size class 2. This is an important feature for analysis and needs to be considered. The other features are NWCG_REPORTING_UNIT_NAME, NWCG_REPORTING_UNIT_ID, SYSTEM_REPORTING_UNIT, SOURCE_REPORTING_UNIT, LOCAL_FIRE_REPORT_ID, LOCAL_INCIDENT_ID, FIRE_CODE, FIRE_NAME, ICS_209_INCIDENT_NUMBER, ICS_209_NAME, MTBC_ID, MTBS_FIRE_NAME, COMPLEX_NAME. These features are identification numbers and they do not provide any information about the fire points hence cannot be used for the prediction. The next feature is FIRE_YEAR which is already in integer form and has 0 null values. Hence, no changes need to be done and can be used for analysis directly. DISCOVERY_DATE can be discarded as we can get the month of fire from DISCOVERY_DOY and discard the rest data. For forest fire time in minutes, the classification is done in a time-based manner consisting of 5 intervals. Then it is converted to integer form.

- 0: Null values
- 1: Early Morning (12 am - 6 am)
- 2: Morning (6 am - 12 pm)
- 3: Afternoon (12 pm - 4 pm)
- 4: Evening (4 pm - 8 pm)
- 5: Night (8 pm - 12 am)

The next features are STAT_CAUSE_DESCR and STAT_CAUSE_CODE are different representations of the same feature. STAT_CAUSE_DESCR is human readable and STAT_CAUSE_CODE is machine-readable and hence will help the ML model understand easier. Therefore, we have discarded STAT_CAUSE_DESCR and considered STAT_CAUSE_CODE. CONT_DATE i.e. contained date does not add any value to the dataset and can be omitted. CONT_TOD i.e. containment time of day is maximum for all fires occurring in the early morning. The CONT_TOD is divided into 7 classes. The next feature is FIRE_SIZE which we will be predicting, it's our class label so it is removed from the dataset. The last two features are the LATITUDE and LONGITUDE values. These values are needed to point to the exact location of the fire on the map. To group the nearby forests together the values are rounded off and flooring is applied. Hence, these two features are necessary for prediction. At the end, feature selection is done so as to get good results. [11]

V. ML Algorithms Used

The data we are dealing with is having geospatial features like the latitude and longitude so we cannot use normal linear classifier to train our model (due to highly non-linear scaled data), one way is to drop these features and move ahead, but we need these features to map the predicted fire on the worldmap, the other way as described in this article is that we should prefer non-linear models like the tree based which will not only preserve those features but also will not help us not to modify the input model PREDICTED_CLASS every time.[12] We have classified our output feature in 7 classes(A,B,C,D..) which corresponds to 7 different range of Areas(A- 0-0.25 acres,B- 0.26-9.9 acres, C- 10.0-99.9 acres,..) The Algorithms that we are using without the hypertuning are:

1) K-Nearest Neighbours
2) Gaussian Naive Bayes
3) Random Forest
4) Decision Tree

A. K-Nearest Neighbours

We create a scatter plot of input features and label all of the PREDICTED_CLASS data points. Now, when we enter a new data point, we must classify it into which PREDICTED CLASS it belongs. To do so, we must choose the number of neighbors K by trial and error (K=5). We need to calculate the euclidean distance between all of those K neighbors. Calculate the total number of data points in each category; the category with the highest total is where our new data point will be placed.

B. Gaussian Naive Bayes

Calculate the probability of all the PREDICTED_CLASS, Here the values of the input features are continuous in nature that's why we can't find probability separately, rather calculate the Mean and Variance of all input features given the PREDICTED_CLASS. Now when we enter the new instance which is to be classified, Calculate the posterior probability of the PREDICTED CLASS using the Gaussian distribution equation. The PREDICTED_CLASS with the highest posterior probability is your required output.

C. Random Forest

As our data set is too complex, Random Forest is the best opt classifier algorithm because it contains many decision trees which take the average to improve the accuracy of classification in which PREDICTED_CLASS the data point will lie. Random Forest is more accurate than a decision tree classifier because it has more decision trees. It also averts the issue of overfitting.

D. Decision Tree

The decision tree classifier creates the classification model. Each node in the tree represents a test on an PREDICTED_CLASS, and each branch descending from that node represents one of the attribute's possible values.Each leaf represents one of the PREDICTED_CLASS labels. The training set's instances are classified by navigating them from the root of the tree to a leaf, based on the results of the tests along the way and then once when we reach the leaf node we get the final Area associated with the PREDICTED_CLASS.[13]

Algorithm used	Train data accuracy	Test data accuracy	Mean abs. error	Mean absolute % error	Remarks from observation
KNN	0.7149	0.6044	0.4707	25.57	Can be used if hyper params tuning is done
Gaussian NB	0.5316	0.5308	0.8041	53.20	KNN performs better than Naive bayes
Decision tree classifier	0.9529	0.5569	0.5514	32.22	Model overfitting. Huge diff in train and test accuracy.
Random forest classifier	0.9529	0.6179	0.4543	25.36	Overfitting but can be overcomed using hyper params tuning.

Fig. 1. A comparison table displaying accuracy's of 4 algorithms used.

VI. Result

After comparing all the MAPE and MAE scores of all the algorithms, we can conclude that KNN performs better than Gaussian Naive Bayes because the accuracy of KNN is better and the MAE score is also less when compared to Gaussian Naive Bayes. After Trying a few Tree-based models (Decision Tree Classifier) the value of MAPE and MAE scores were greater than the KNN model but there was a huge difference in the accuracy of the Decision Tree Classifier which lead to the case of overfitting in the training dataset, but it can be controlled by proper hyperparameter tuning, When we trained the model using Random Forest Classifier the difference between the training and testing data accuracy was less compared to the Decision Tree Classifier which means the existing error can be overcome by adjusting the class-weights and few other params, later we will store the pickle file of Random Forest Classifier to avoid recomputing the model each time.

Algorithm used	Mean abs. error	Mean absolute % error	Remarks from observation
KNN	0.4383	23.95	Values decreased by 0.01%
Decision Tree	0.4640	35.90	MAPE reduced by 4%
Random forest classifier	0.4657	24.99	MAPE reduced by 4%

Fig. 2. Final accuracy of top performing algorithms

Using the pickle file of the Random Forest Classifier we now can remove all the unnecessary feature columns (eg - OBJECT', 'FOD_ID', 'FPA_ID', 'NWCG_REPORTING_UNIT_ID', 'NWCG_REPORTING_UNIT_NAME', 'SOURCE_REPORTING_UNIT' .and more) and add new dataset of AVG_TEMP, FOREST_AREA and AVG_PREP. And start performing Feature Engineering on the final dataset. Mean Absolute Percentage Error (MAPE) is the most common error for Forecasting. In our study the MAPE value comes out to be 23.42% which means for the remaining 77% of the time, the model is predicting the right fire size class.

VII. Conclusion

The necessity for building a technology that can accurately identify forest fires in a timely manner is urgent. The system understudy's examination reveals that it has the potential to be implemented and manufactured. Satellite-based forest fire detection is a practical technique that can be used as a backup or supplement to conventional fire monitoring methods. Satellite-based fire detection can play an important role in forest fire detection in the early stages of the fire season when fire surveillance flights have not yet begun. Fires in unoccupied places can also be detected by satellites. The most intriguing and unique element of this system is its capacity to ingest and process diverse instrument data, using cellular automation. The pattern of spread of the fire can be studied and by performing the required analysis appropriate measures can be taken to control the disaster.

References

[1] G.-M. W. M.-P. M. Boubeta M, Lombardía MJ, ""Burned area prediction with semiparametric models," Ph.D. dissertation, 2016.

[2] B. M. Adam Stanford-Moore, "Wildfire Burned Area Prediction," Ph.D. dissertation, 2020.

[3] C. Y. S. P.-F.-G. E. R. J. Coffield SR, Graff CA, ""Machine learning to predict final fire size at the time of ignition," Ph.D. dissertation, 2019.

[4] H. T. T. R. James G, Witten D, ""An introduction to statistical learning," Ph.D. dissertation, 2013.

[5] A. M. Elshewey and A. A. Elsonbaty, "Forest Fires Detection Using Machine Learning Techniques," Ph.D. dissertation, 2020.

[6] V. Z. Aleksandar P, Silvana P, """Multiple linear regression model for predicting bidding price," Ph.D. dissertation, 2016.

[7] d. S. A. Pereira JM, Basto M, ""The logistic lasso and ridge regression in predicting corporate failure," Ph.D. dissertation, 2016.

[8] R. Singh, "Predicting Wildfire using Data Mining," Ph.D. dissertation, 2016.

[9] A. P. Mauro Castelli, Leonardo Vanneschi, "Predicting Burned Areas of Forest Fires: an Artificial Intelligence Approach," Ph.D. dissertation, 2015.

[10] J. OLIEHOEK, "Wildfire visualization: heat maps ," Ph.D. dissertation, 2018.

[11] J. W. S. Y. S. Cai, J.; Luo, ""Feature selection in machine learning," Ph.D. dissertation, 2018.

[12] M. M. Carmine Maffei, "Predicting forest fires burned area and rate of spread from pre-fire multispectral satellite measurements," Ph.D. dissertation, 2019.

[13] M. A. N. A.-A. Binh Thai Pham, Abolfazl Jaafari, ""Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction," Ph.D. dissertation, 2020.