MACHINE LEARNING APPROACHES FOR WILDFIRE ANALYSIS

National Research University Higher School of Economics
Moscow, Russia

MOHAMMED BILAL ANSARI

SUPERVISOR: DMITRY I. IGNATOV

Faculty of Computer Science

ABSTRACT

Environmental, social, and economic causatums from wildfires have been continuously increasing around the world over the past decade. These fires not only devastated forest and grassland but also detrimentally impacted wildfire habitat, water quality & supply, tourism, and property values. In the past few years, a number of research studies have been conducted to monitor, predict and prevent wildfires using several Artificial Intelligence techniques such as Machine Learning, Deep Learning, Big data, and Remote Sensing. In this paper, we proposed the wildfire classification and prediction system to classify the wildfires into elven different types based on the data on temperature anomalies from satellites and geographical data using the CatBoost classifier. Quality metric - multi-class ROC-AUC has been considered to evaluate the performance of the system. The proposed system achieved high performance on the test set.

Keywords: - Machine Learning, Wildfire, Classification, CatBoost, Environmental features, Geographical features.

I. INTRODUCTION

Nowadays, global warming or climate change has become an extremely serious challenge which requires a deep concern to mitigate the impact of it. Due to climate change, the forest ecosystem faces unexpected disturbances such as wildfires and storms. Wildfires are uncontrolled fires, swiftly spreading, and raging immensely colossal flames highly propagated with wind action that can wipe out a wide range of forest or vegetation land area within a very small fraction of time which is hazardous and inimical to the environment of the planet and also to human lives. Wildfires are

not only caused by natural phenomena but also by human activities like burning debris, unattended campfires, arson, and fireworks.

Every year around 30–46 million km^2 of the global terrestrial land surface burned (approx. 4% the global land surface) [1] but the direct repercussion of it on a particular person is limited and ergo does not draw much attention globally. Sometimes, the media react to the enormity of wildfires and tend to report the pernicious impacts of it and short term effects, with the concentration on singular focus. It issues a spark debate about forest management with the intent to minimize the environmental and gregarious damages, in particular when hamlets, villages, and infrastructure are affected [2, 3]. In recent years, it has been observed the immense increase in a wildfire in terms of frequency and size at various places around the world even where we typically don't expect extensive wildfires. Climate change is not the only ignition factor to decide the size and destructiveness due to fire. Humans are increasingly intruding into wildland areas to perform different kinds of activities which result in a high probability of fires and destruction especially by parching forests and making them more susceptible to blazing.

Russia has the largest forested region in the world and the northern forests occupy around 45 percent of the country area. A predominance of the wildfires transpires in inaccessible areas. According to a news report, Russia is warming at a 2.5 times faster rate compared to the rest of the planet and Siberia is experiencing an extremely early start of fire season this year. In Siberia and other parts of the nation, wildfires were 10 times much more intensive in the month of May as compared to this time last year due to climate catastrophe. Wildfires are not restricted to a specific continent/region or environment.

Every year, the US reports 67,000 wildfires incidents and 7.0 million acres burned-area on an average. The government spends around $900 million for wildfire suppression annually. The cost of fighting wildfires has increased in recent decades along with their severity and the destruction they have caused.

In the era of Artificial Intelligence and Big Data analysis, people are contributing in different ways like building robots, drones, and IoT based systems to control and prevent the wildfires and detecting them at the early stage to avoid situations out of control. Majority of wildfires control systems can be categorized into two types i.e. automatic wildfire recognition through image processing by collecting data using stationary field cameras or satellites and another system based on measurement and monitoring of weather parameters such as temperature, humidity using sensors or satellite.

Most of the existing wildfire detection systems are based on Wireless Sensor Networks(WSN) in which sensors are deployed to monitor specific environmental conditions such as temperature, fire, and send those data to the base station. The base station processes the received data and detects the wildfire using a machine learning model.

In this paper, we proposed the wildfire classification and prediction system to classify the wildfires into elven different types. Sberbank wildfire labeled dataset has been used to train the model. First, we performed the exploratory data analysis(EDA) on the data to understand the context needed to build the model and then generated the features using NCEP environmental data and geographical data. Performed the hyperparameters tuning on CatBoost classifier and selected the most significant features which help to describe the problem in the best way. Finally, we trained the model on the tuned hyperparameters and the most significant features.

II. Literature Review

Several approaches or systems have been proposed by many researchers and engineers around the globe to predict and obviate wildfires using different, different techniques and strategies based on Machine Learning, Big Data, and Remote Sensing. Some of the following approaches are discussed as below: -

The socioeconomic factors like demographic growth, urban growth, and human intervention play a vital role as explanatory variables in predicting the high/low fire-affected area [4]. In this paper, the performance of multiple algorithms like Random Forest, Boosting Regression Trees have been compared with the Logistic Regression. Boosting Regression Trees algorithm turns out to be most adequate compared to the others but doesn't seem to be much promising. It considers very few explanatory variables to make the prediction, which makes the model quite less complex.

Support vector machine and Neural network seem to achieve high accuracy in predicting the wildfires using features collected from the satellite images [5]. The author expressed the will to consider the weather data for future work to strengthen the model. Occurrence, growth, and spread of wildfires highly depend on the weather, it impacts on the extensiveness and movement of fire. Weather data such as air temperature, humidity, and wind speed have a high influence on the wildfire.

In [6], wildfires have been classified into crown and surface types by acoustic emission spectrum of fire using Wireless Sensor Networks(WSN). Fire acoustic data is accumulated using WSN and subsequent analysis of it is performed in the central

processing node. The author suggested comparing the recorded acoustic emission spectrum with a set of spectra samples from the typical fire emissions in a database. The major problem of WSN based systems in wildfire detection is the high number of sensors needed to cover the extensive forest area.

III. Exploratory Data Analysis (EDA)

EDA is a significant step that is needed to consider before diving into machine learning or any statistical modeling because it gives the direction to understand the problem profoundly and provides the context needed to build a suitable model for our problem and also helps to interpret its result correctly.

Let us perform the EDA on our wildfire data set, fires have been classified into eleven different types. Below is the distribution of a dataset over different types of fires. We can limpidly observe the unequal distribution of classes(fire types) within a dataset which makes it highly imbalanced. Among all the fires incident, peat fire incident has been reported least number of times which seems to be a very rare event. Most of the fire incidents recorded have been uncontrolled in the dataset.

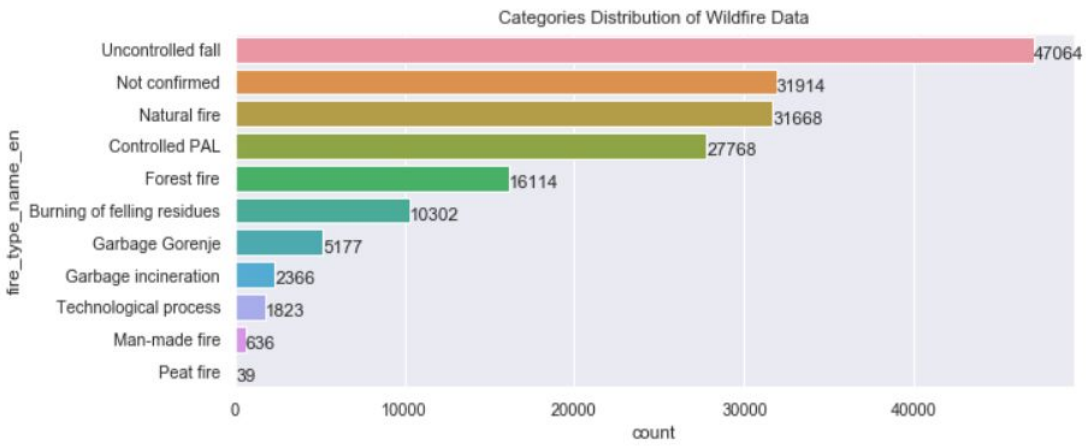

Figure 1. Distribution of classes or fire types

Analyzing the data based on year and month, 2014 has been the most hardest hit year by massive wildfires in the period of eight years from 2012 to 2019. As per the dataset, 20.32% of fire incidents have been recorded in the year 2014. On inspecting the monthly distribution of data, most of the fire incidents have been recorded in the month of March, April and May that signifies fire incidents are more likely to happen during the summer due to lack of atmospheric humidity at

high-temperature forest trees become littered with dry senescent leaves and twinges that burst into flames with the slightest ignition source.

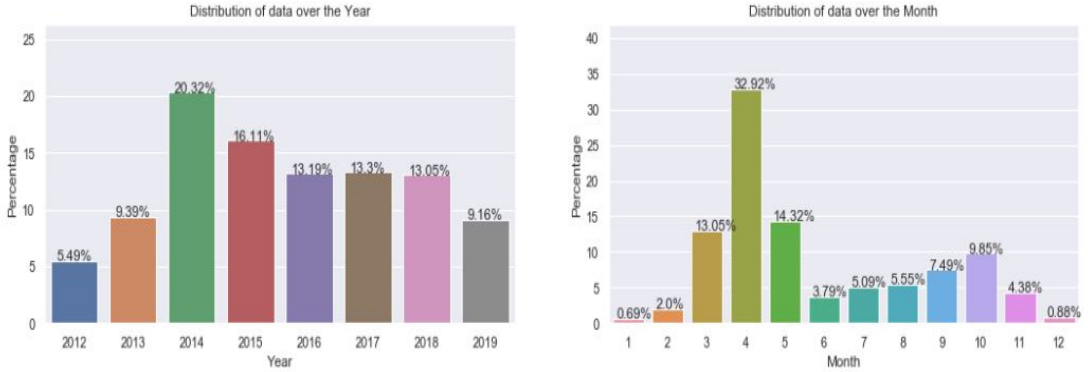

Figure 2. Distribution of data over years and months

Distribution of different types of wildfires over the geographical area. Most of the instances in our dataset come from Russian wildfire incidents. Russia has five different types of vegetation regions i.e. Temperate grassland, desert & dry shrub, coniferous forest, deciduous & mixed forest, and tundra [7]. Temperate grassland and desert & dry shrub regions of Russia have been affected by natural fire, man-made fire, and garbage incineration. Forest fires have been reported in the deciduous & mixed forests region and coniferous forests region which cover most of the Russian land. Virtually no wildfire incident has been reported in the Tundra region.

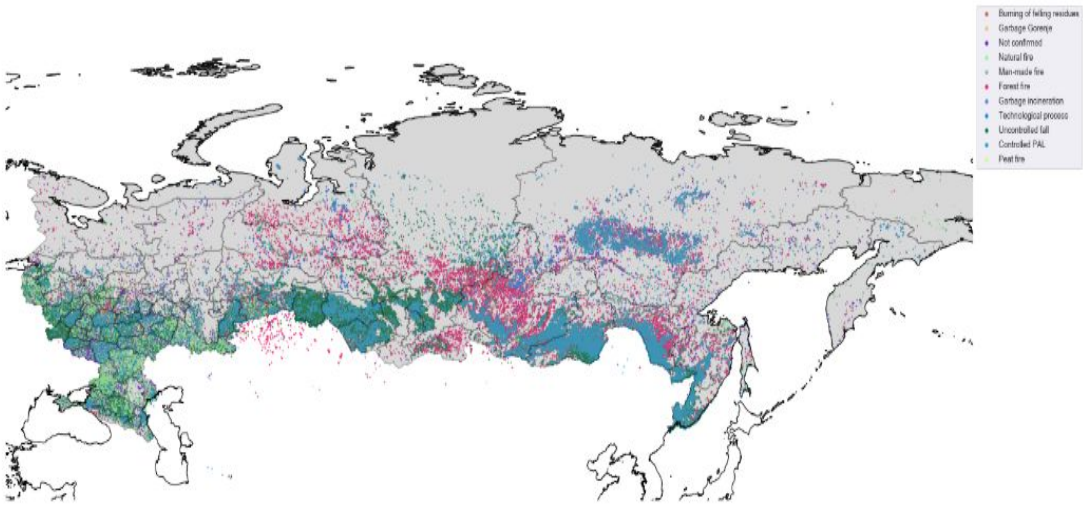

Figure 3. Distribution of data over the geographical region

IV. Data Preprocessing

Data preprocessing plays a vital role in ameliorating the quality of the raw experimental data and it is an extremely consequential step in machine learning as the quality of data and the useful information that can be derived from it directly affects the learning ability of the model. That's why, it is extremely crucial that we preprocess raw data before feeding it into the model.

Our wildfire dataset [8] consists of fields like fire_id, date, latitude, longitude, fire_type, and fire_type_name for each reported fire incident but this information is not sufficient to predict the fire type. Wildfires highly depend on weather so we decided to use the NCEP(National Centers for Environmental Prediction) data to generate the environmental features for our wildfire data. Geographical location of fire incidents is also equally important in predicting the wildfire and categorizing into fire type.

1. Environmental Features

We used the NCEP data to generate the set of environmental explanatory variables which can be helpful in predicting the wildfire and type of it. There are many weather or environmental parameters that can affect the occurrence, growth, and spread of wildfires. We considered the following weather parameters like air temperature, geopotential height, relative humidity, east wind or zonal wind, north wind or meridional wind, tropopause pressure, tropopause temperature, the surface potential temperature at different pressure levels and also average over one week, two weeks and three weeks of these parameters have been taken into the account.

Let's see how these weather parameters influence wildfire, high air temperature makes the trees sensitive toward catching fire which results in an extensive wildfire and creates a lot of destruction. Geopotential height is an approximate height above sea level of a pressure level, according to many studies wildfire activities are closely linked with low geopotential height. When relative humidity is low in the atmosphere that means there is a low amount of water vapor in the air than expected to saturate the air. In case of low relative humidity, the air takes moisture from the dead forest trees and makes them more drier which results in the massive wildfire. The wind has the most prominent and strongest impact on wildfire behavior due to the fanning effect on the fire. It supplies additional oxygen to fire, which results in the forest burning more rapidly and removes the surface fuel moisture, which increases the drying of the fuel. It also pushes flames, sparks, and firebrands into new fuel to move faster across the forest land. The tropopause defines

the boundary between the troposphere and the stratosphere in the atmosphere. Tropopause pressure is strongly correlated with a distinct geographical location.

Wind speed and it's direction has been computed using zonal wind(uwnd) and meridional wind(vwnd) components as

$$wind\_speed = (uwnd^2 + vwnd^2)$$

$$wind\_direction = tan^{-1}(vwnd/uwnd)$$

We observed in the EDA that most fire incidents have been reported in the summer. So, we decided to consider the season(i.e. winter, autumn, summer, spring) also as an explanatory variable to predict the wildfire and it's type.

2. Geographical Features

Geographical parameters are valuable in understanding fire risk and also determining the wildfire and its type. We generated geographical features like land type, district, distance from fire incident to the natural forest, field, city, etc. for each fire incident in the wildfire dataset. Type of the burning area is influential to determine the fire type that's why we considered land type which describes whether the burning area is forest or field. Federal subject, district, and its population where the fire incident transpired are also considered. We took into account how far are the fire incident areas from natural forest, forest, field, and city and computed these geographical features using unsupervised learner 'NearestNeighbors'. A number of forest, field, and cities within the specified radius of fire incident have been computed and also the number of fire incidents reported in the same month last year within the specified radius of fire incident so that model can learn the distribution of fire incident over the month. These geographical features help the machine learning model to learn better about the fire incident and type of it because some geographical areas are more vulnerable to fire and others are less.

V. Model Pipeline

Now we will build the model to predict wildfire and it's type using the Catboost classifier. Catboost is the gradient boosted decision tree algorithm developed by Yandex researchers. It is used for self-driving cars, recommendation systems, and weather prediction tasks at Yandex. It is an open-source library and can be used by anyone. It provides state-of-the-art results on a wide range of datasets, not only for datasets with categorical features, and has high performance in comparison to the other boosting algorithms. It doesn't require extensive hyper-parameter tuning and

builds more generalized models. Although, it has multiple hyperparameters like learning rate, tree depth, and the number of trees to perform the tuning.

We trained the Catboost classifier using the preprocessed and newly generated environmental and geographical features and performed the hyper-parameter tuning on it. Using the feature importance, we decided to consider the population and week of the year as golden features for our Catboost model. After running the Catboost model on different sets of hyper-parameters, we found the optimal values for hyper-parameters. We set the hyperparameter eval_metric and loss_function of CatBoost classifier as 'MultiClass' and L2 regularization term of the cost function to 5 and trained over 6000 iterations.

Let's put everything in a single box using a pipeline to handle the model workflow. The machine learning model pipeline consists of several components or steps which provide flexibility, cost control, and speed, etc. It includes data preparation and the CatBoost-model as components.

Step-1: Generate Environmental Features	Step-2: Generate Geographical Features	Step-3: Drop Features	Step-4: Fill missing values	Step-5: Catboost Classifier

Figure 4. CatBoost model pipeline

We feed the date, latitude, longitude of fire incident from our wildfire dataset into the model pipeline. In the first step, the model pipeline generates the environmental features like air temperature, geopotential height, relative humidity, east wind or zonal wind, north wind or meridional wind etc. using NCEP data for the fed latitude, longitude and date and then passes generated features to the second step. In the second step, it generates the geographical features like land type, district, distance from fire incident to the nature forest, field, city etc. In the third step, we drop the irrelevant features from the set of generated features which have less importance in predicting the wildfire types. We defined the fourth step, to fill the missing values in the explanatory variables using simple imputer fill as zero. In the final step, we predict the type of wildfire like forest fire, man made fire, peat fire etc. using all important explanatory variables fed and generated during the multiple steps of the pipeline.

VI. Experimental Result

We used the Sberbank wildfire labeled dataset [8] to train and test the CatBoost model. It has 174871 instances, 11 fire types or classes and the following attributes fire_id, date, latitude, longitude, fire_type, fire_type_name. We splitted the dataset into train, val, test sets in the ratio of 80%, 10%, and 10% respectively. To generate the environmental features, NCEP data [9] has been used and geographic coordinate data for generating the geographical features.

Performance of our proposed system for wildfire classification has been evaluated using multi-class ROC_AUC quality matrix. Our system achieved the high ROC_AUC score on most of the classes and low ROC_AUC score on few. We got highest performance on fire type_1 i.e. ROC_AUC 0.94 and lowest performance on fire type_10 i.e. ROC_AUC 0.69.

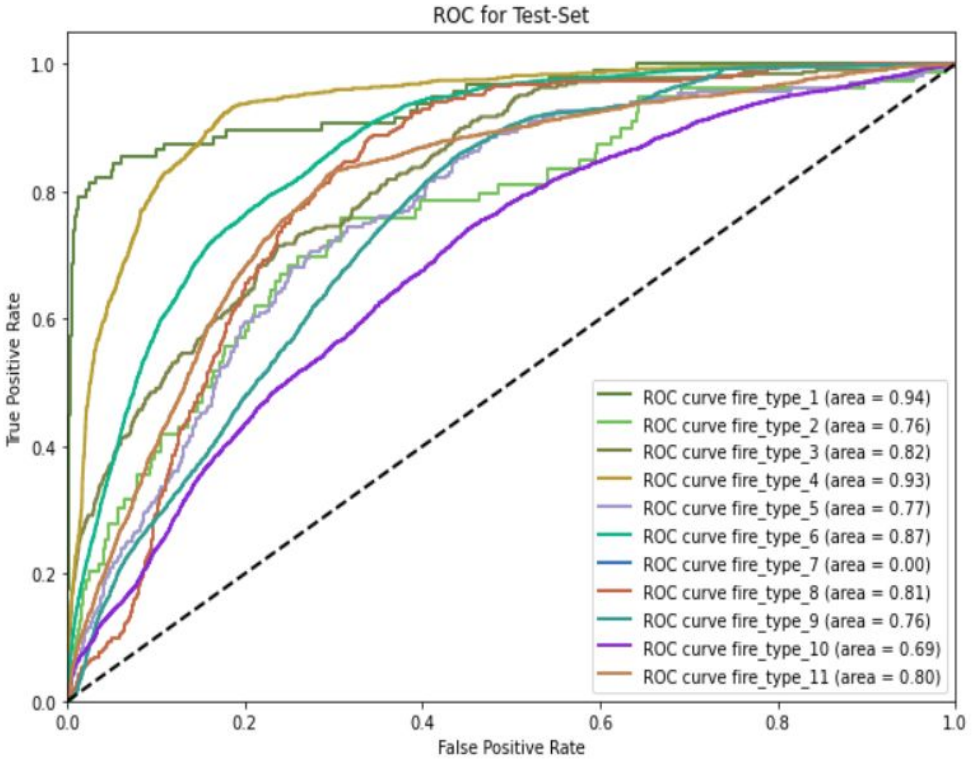

Figure 5. ROC AUC

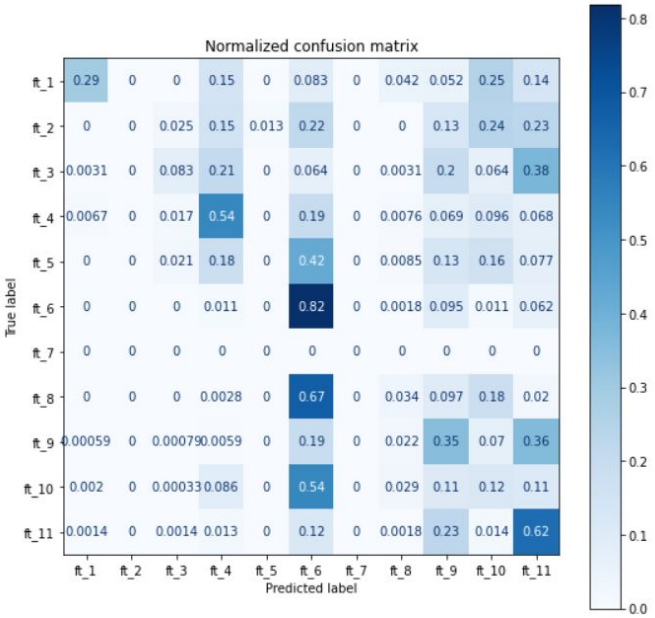

Figure 6. Normalized Confusion matrix

VII. Conclusion

In this paper, a wildfire classification and prediction system based on CatBoost classifier has been proposed using the data on temperature anomalies from satellites and geographical data. Our system achieved a high ROC_AUC score on most of the classes and a low on rest of the classes. In future, we can make our proposed system more robust by considering the more sophisticated features which will turn out into high model efficiency.

References:

[1] Randerson JT, Chen Y, Van Der Werf GR, Rogers BM, Morton DC. 2012. Global burned area and biomass burning emissions from small fires. J. Geophys. Res. Biogeosci. 117, G04012 (10.1029/2012JG002128).

[2] Paveglio T, Norton T, Carroll MS. 2011. Fanning theflames? Media coverage during wildfire events and its relation to broader societal understandings of the hazard. Hum. Ecol. Rev. 18, 41–52.

[3] Varela E, Jacobsen JB, Soliño M. 2014. Understanding the heterogeneity of social preferences for fire prevention management. Ecol. Econ. 106, 91–104. (10.1016/j.ecolecon.2014.07.014).

[4] Rodrigues M, de la Riva J. 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environ Model Software. 57:192–201.

[5] Sayad YO, Mousannif H, Al Moatassime H. Predictive modeling of wildfires: A new dataset and machine learning approach. Fire Saf J. 2019;104:130–146.

[6] Khamukhin, A.A.; Demin, A.Y.; Sonkin, D.M.; Bertoldo, S.; Perona, G.; Kretova, V. An algorithm of the wildfire classification by its acoustic emission spectrum using Wireless Sensor Networks. J. Phys. Conf. Ser. 2017, 803, 1–6.

[7] https://slideplayer.com/slide/6041727/

[8] https://wildfire.sberbank.ai/competition

[9] https://www.ncep.noaa.gov/