Machine Learning for Forest Fire Prediction

1st Sean Seneviratne
Dept. of Electrical and Computer Engineering
Stevens Institute of Technology
ssenevir@stevens.edu

***Abstract*—In recent decades, significant emphasis has been placed on data-driven technology to battle the ravaging effects of climate change and environmental degradation. One of the most virulent forces is that of forest fires that ravage through much of the world's rain-forestry and in recent, almost 4 million acres of land in California. Forest fires cause severe environmental impacts that often endanger the livelihoods of communities in the area through the effects of impure breathing oxygen and massive droughts. The objective of this project is to use existing environmental data to model the "at-risk" area's of a particular region and proactively predict which regions are likely to go up in smoke. By leveraging the vast variety of open-source machine learning and data analysis libraries in Python, we hope to build a baseline predictive modeler to fit onto geographical data and model risk.**

***Index Terms*—climate change, data analysis, machine learning, Python**

I. Introduction

Forest fires are one of the world's largest contributors to environmental damage. They have caused considerable losses to ecologies, societies and economies worldwide. In light of recent alarm regarding irrevocable climate damage, quantitative models and analyses have been done using existing climate data to help understand how forest fires begin and spread. Our CPE695 project hopes to develop machine learning models that will use geographical attributes of the California regions as well as past fire history data to create a forest fire susceptibility map. We implemented and optimized three different types of ML models to predict the severity of fire. Between the K-Nearest Neighbors approach, Random Forest approach and Support Vector Machine approach - we found that the Random Forest had the most consistent and accurate predictions compared to the other two. In this report we will explain our methodology and results.

II. Related Work

A. Forest Fire Prediction in Wunnan Province, China

The first study was done by a Chinese team trying to model forest fire susceptibility in the Wunnan Province of China by using a Convolutional Neural Network. Guoli Zhang, Ming Wang and Kai Lu of the Beijing Normal University published a paper in September 2019 analyzing techniques to best model forest fires and to help better prevent and manage them. They proposed a spatial prediction model for forest fire susceptibility using a convolutional neural network. The team analyzed 8 years of forest fire locations from 2002 to 2010 and found 14 factors to be the most influential in the cause of them. The Beijing Normal University team collected forest fire data of the Yunnan Province in southwestern China. One of the first things they did to better visualize their data was something called forest fire inventory that we believe will be one of the first steps of our project. A forest fire inventory map is compiled using the influencing factors of a region including historical fire reports and satellite imagery. From this preliminary analysis they were able to notice that from 2002 to 2010, 7675 fires had occurred, 58% of which occurred in the spring time which helped show a pattern of seasonality.

The dataset used for the study was sourced from a couple of places including NASA, Google Cloud and NCAR (National Center for Atmospheric Research) primarily in the data types of NetCDF, Vector and Raster which are common types for atmospheric data. Our group has become familiar with unpacking and storing similar data formats to use with Python.

B. Forest Fire Prediction in Pu Mat National Park, Vietnam

The second study our team sourced from was a machine learning exercise done on a fire dataset from Vietnam. The studied evaluated Bayes Network (BN), Naive Bayes, Decision Tree and Multivariate Logisitic Regression methods to evaluate elevation, slope degree and aspect and how they correlate to fire occurence. Similarly to the Chinese team, the Vietnamese team conducted a feature selection analysis of the explanatory variables and then trained different machine learning models to predict fire occurence in Pu Mat National Park.

III. Solutions

A. Dataset

The dataset that was used is based on the research done from Demtrios Gatziolis and Bob McGaughey who are a part of the United States Department of Agriculture. Their team composed a dataset of 188 explanatory variables that was taken by LiDar sensors mounted on airplanes that flew over the area of Archie Creek, Oregon. The forest line in this area is connected to the California forest line and often the occurence of fires in both these regions are correlated.

The explanatory variables included several different types of metrics regarding the Archie Creek forest region in Oregon. These variables describe the landscape of the forest in elevation, canopy-relief ratio, mean height of tree's and more. It was intimidating to look at 30,000 data points over 188 variables and understand which were the most important to our target variable which was fire-severity. We were advised by Demtrios

Gatziolis to hone our attention in on 8 explanatory variables that measured the highest correlation to fire severity.

- 1st cover gt2m : percent of area under tree crowns
- elev-canopy-relief-ratio : heterogeneity in canopy height
- elev-CV-2m+ : Coefficient of variation in vegetation height
- elev-P20-2m+ : 20th percentile in vegetation height
- elev-P90-2m+ : 90th percentile in vegetation height
- elev-quadratic-mean : highest return heterogeneity
- int-ave-2m+ : mean intensity of returns above 2 meters
- severity-class : fire-severity classification (1-4)

1st_cover_gt2m	elev_canopy_relief_ratio	elev_CV_2m+	elev_P20_2m+	elev_P90_2m+
5.19	0.26	0.21	2.25	3.53
8.27	0.20	0.21	2.22	3.45
7.97	0.23	0.19	2.21	3.42
11.76	0.19	0.22	2.20	3.47
14.03	0.25	0.18	2.21	3.32

Fig. 1. The first 5 variables are captured in this snippet of a pandas dataframe. Each variable pertains to elevation-related metrics of the forest terrain

elev_quadratic_mean	FIRST_RETURNS_elev_canopy_relief_ratio	int_ave_2m+
2.82	0.26	5994.10
2.75	0.20	6973.51
2.72	0.23	7440.64
2.74	0.19	7303.92
2.68	0.25	6948.74

Fig. 2. The last 3 explanatory variables that we were told to focus in on by our forest expert advisor

Quite unfamiliar with forest terminology, it was difficult for us to truly understand the nature of the variables and how they affected our target. Even so, we tried to learn them as much as we could to make thoughtful interpretations, analyses and adjustments.

The target variable we were trying to predict was known as Fire Severity which is broken into 4 classes. Unchanged(1), Low(2), Moderate(3), and High(4). Our goal was to use our data set to as accurately predict these severity classes given a pattern in the explanatory variables.

B. Univariate Analysis

Our first task in demystifying our data was to understand the distributions. Was the data set balanced or imbalanced? We discovered that there were more ”High” severity data points as there were any other variable on our set.

Unsure if this would cause a bias in our training model, we sought to halve the number of Class 4 data points to even out the distribution. We kept the raw data set in consideration however assuming the more data our model ingested, the better predictions it would be able to make.

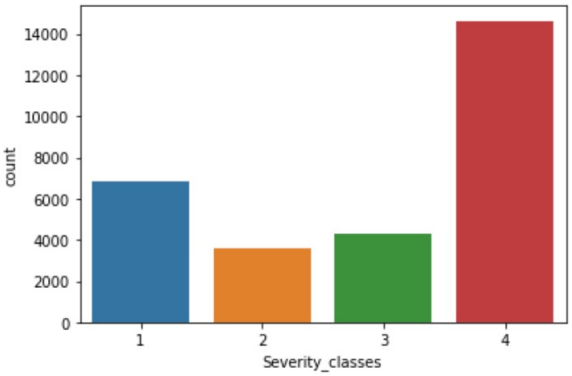

Fig. 3. An imbalanced distribution, with severity class 4 outweighing the other attributes

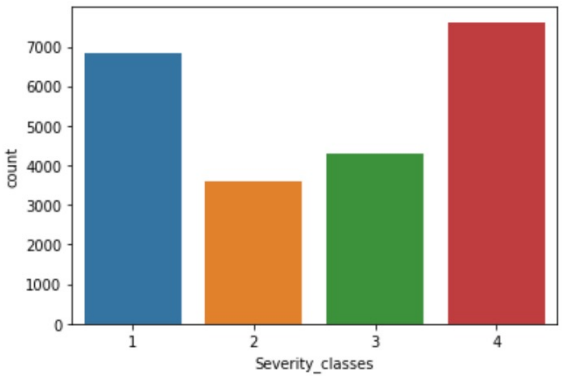

Fig. 4. A more even balanced data set

C. Feature Selection

After understanding our data set it was important to understand how our explanatory variables (features) mapped to our target variable. We implemented a Random Forest regression and a Decision Tree Classifier to understand feature importance to weed out the less important ones.

```
Feature: 0, Score: 0.17691
Feature: 1, Score: 0.08348
Feature: 2, Score: 0.08250
Feature: 3, Score: 0.14202
Feature: 4, Score: 0.13658
Feature: 5, Score: 0.11815
Feature: 6, Score: 0.08471
Feature: 7, Score: 0.17565
```

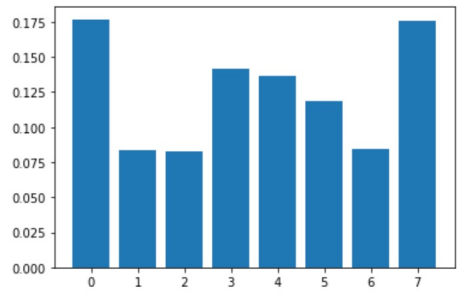

Fig. 5. Feature importance analysis using a Decision Tree Classifier

```
Feature: 0, Score: 0.19915
Feature: 1, Score: 0.06949
Feature: 2, Score: 0.08160
Feature: 3, Score: 0.13421
Feature: 4, Score: 0.13024
Feature: 5, Score: 0.11412
Feature: 6, Score: 0.07900
Feature: 7, Score: 0.19219
```

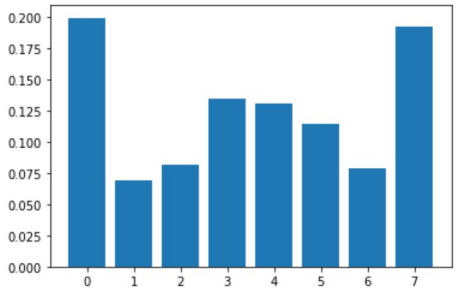

Fig. 6. Feature importance analysis using a Random Forest Regressor

We used these feature significance percentages to boil down our features to 5 variables that exceeded 10% importance.

```
Feature: 0, Score: 0.23157
Feature: 1, Score: 0.19062
Feature: 2, Score: 0.18658
Feature: 3, Score: 0.16612
Feature: 4, Score: 0.22512
```

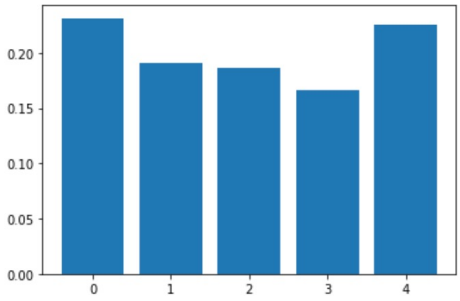

Fig. 7. Feature importance analysis using a Random Forest Regressor

D. Machine Learning Models

The next question was to understand which models would be most beneficial in classifying severity classes. Three models that we chose for implementation were k-nearest-neighbors, random forest and support vector machines.

K-NN is a pattern recognition algorithm that uses training data sets to find the k closest relatives in future examples. It help us us find which class the new input belongs to when k nearest neighbors are chosen and distance is calculated between them.

Random Forest is an expansion of the decision tree classifier which constructs a real world decision tree with training data then fits new data within one of the trees as a 'random forest'. Random forest averages the data points and connects the input to the nearest tree on the data scale. It is a commonly used model in classification techniques because it adds randomness to the model while creating the trees. Rather than searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This wide diversity in creation results in a generally better model.

Support Vector Machines are another commonly used model in classification problems. It uses algorithms to train and classify data within degrees of polarity. SVM assigns a hyperplane that best separates the data it is given to distinguish the classes.

Each of these machine learning models have been studied to be the most effective in understanding relationships for classification problems which is why we picked them.

E. Implementation

Our baseline models for the Random Forest, SVM and k-nn showed to be somewhat ineffective with RMSE's in the range of 1 to 2. This is significant because of 4 possibile classifications, an average error of even 1 tells us that the model incorrectly classifies by one categorical unit. Our approach in tuning our parameters was to understand which would lower our RMSE.

The KNN model parameters that we tuned were

- leaf-size: the larger the leaf size, the closer neighbors the algorithm picks. This is because the trees main purpose is to reduce the number of candidates for its neighbors.
- nearest-neighbors: Number of neighbors to use for kneighbor inquiries
- p values: power parameter for Minkowski metric. When p=1, this is equivalent to using the manhattan-distance whereas p=2 is euclidean distance. These are different methods in calculating distances between points.

These parameters were entered into a hyperparameter dictionary and then tested using GridSearch with 5 cross validations.

The Random Forest parameters that we chosen were

- n-estimators : Number of trees in the random forest
- max-features : Number of features to consider when looking for best split.
- max-depth : Maximum depth of the tree
- min-samples-split : Minimum number of samples to split an internal node
- min-samples-leaf : Minimum number of samples required to be at a leaf node
- bootstrap : Decides whether bootstrap samples are used. If False, the whole dataset is used to build each tree

These parameters were sampled using the RandomSearchCV with 100 iterations and 3 cross validations.

F. Comparisons

Each model, from baseline to optimized version, was judged based on a confusion matrix, classification report and RMSE values. We established a baseline of model accuracy in this first figure.

Base Models Head	**Accuracy Metrics**			
	Precision	***Recall***	***F1***	***RMSE***
KNN	0.40	0.49	0.42	1.62
SVM	0.35	0.34	0.33	1.79
RF	0.61	0.63	0.62	1.21

Tuned Models Head	Accuracy Metrics			
	Precision	***Recall***	***F1***	***RMSE***
KNN	0.42	0.47	0.44	1.59
SVM	0.37	0.34	0.36	1.65
RF	0.68	0.65	0.63	1.18

Of these models we found Random Forest to be the most accurate of the 3 with an RMSE of 1.2.

Tuning the model parameters did slightly improve the performance of our machine learning models but none to the degree we would have liked. The best RMSE we were able to reach was 1.18 with the optimized Random Forest model. The KNN model did not show significant model progress but it was interesting to understand the training behavior. We found that the algorithm reached a training asymptote early on in the training. We are continuing to understand why that is considering knn seemed a likely candidate of success in preliminary training.

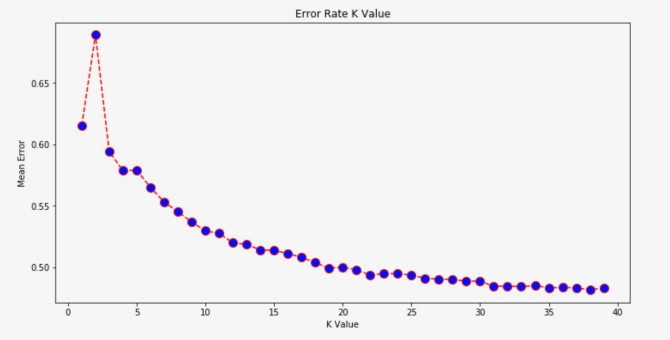

Fig. 8. Error Rate K Value graph based on first 40 training samples

The goal was to reach an RMSE of sub 1 at the very least and within 0.5 to be considered succesful which we were not able to accomplish. We hope to continue optimizing this approach and yield better results in coming months.

IV. Future Research

The next step of this research project would be to use a greater set of parameters to test over a higher cross validation. The time it took to tune just these three models using GridSearch and RandomSearch amounted to a total of 6 hours. Scaling this project onto a cloud computing platform like Google Compute or AWS with a GPU would likely give us better results. The next phase of our project is to build a neural network implementation using an artificial neural network and a convolutional neural network to use deep learning techniques to find a better model. A number of studies have been done in the field of employing deep learning techniques to similar environmental classification problems incorporating spatial-awareness and image classification so there is hope.

Conclusions

Although our techniques did not yield as successful results as we might have expected, it is important to understand that this field of study is relatively small compared to other ML exercises and research teams are dedicating lots of resources to the topic. A more in-depth analysis of the data would perhaps tell us of other patterns in the data that we could use to optimize our approach. The next step in this process would be to do a final iteration of ML approaches then scale to a neural network approach using an artificial neural network or convolutional neural network similar to the studies that we found. Since there is a high level of dimensionality in the data perhaps a longer tuning period would have helped us reach a better result.

Of the three models we selected we found the most success with the Random Forest Model nearing an RMSE of 1 which the closet we got by far. Parameter tuning increased the accuracy of all three models but it was limited to the computing resources and time restraints. Given more time and perhaps a GPU we believe that a smaller RMSE is definitely within reach.

All in all we were grateful to have the opportunity to work on a real-world problem using intermediate machine learning techniques to try and solve it. We could have easily chosen a less complex project where successful results were trivial but it was more interesting to understand a real-world application. We appreciate you taking the time to read our report and critique our methods and welcome any feedback for us to do a better job with this problem.

References

[1] Guoli Zhang, Ming Wang, Kai Liu “Forest Fire Susceptibility Modeling using a Convolutional Neural Network for Yunnan Province of China,” Key Laboratory of Earth Surface Processes and Resource Ecology / Academy of Disaster Reduction and Emergency Management, Faculty of Geographical Science, 19 September 2019

[2] Omid Ghorbanzadeh, Khalil Valizadeh Kamran, Thomas Blaschke, Jagannath Aryal, Amin Naboureh, Jamshid Einali, Jinhu Bian. ”Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches”, Department of Geoinformatics–Z-GIS, University of Salzburg, 5020 Salzburg, Austria , Department of Remote Sensing and GIS, University of Tabriz, Tabriz 5166616471, Iran , Discipline of Geography and Spatial Sciences, University of Tasmania, Hobart 7005, Australia, Chinese Academy of Sciences, Chengdu 610041, China, University of Chinese Academy of Sciences, Beijing 100049, China ,University of Zanjan, Zanjan 45371-38791, Iran. 28 July 2019

[3] Binh Thai Pham 1OrcID,Abolfazl Jaafari,Mohammadtaghi Avand ,Nadhir Al-Ansari,Tran Dinh Du 5,Hoang Phan Hai Yen 6,Tran Van Phong,Duy Huu Nguyen,Hiep Van Le ,Davood Mafi-Gholami,Indra Prakash 11,Hoang Thi Thuy and Tran Thi Tuyen. ”Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction” University of Transport Technology, Research Institute of Forests and Rangelands, Agricultural Research, Education, and Extension Organization, Department of Watershed Management Engineering, College of Natural Resources, Tarbiat Modares University. 17 June 2020