Predictive Model of Spatial Distribution of Forest Fires Driving Factors: A Case Study in Portugal, Spain

Report Info:
Brendan B. Pham
University of Minnesota - Twin Cities
Created 11/04/2023

Keywords:
Wildfire Feature Selection
Neural Networks

ABSTRACT

The ability to predict spatial distribution of Forest Fires is an important tool for forest fire management services that could prove helpful in preventing the volatility of forest fires in the United States. Wildfire poses an escalating threat to ecosystems, resources, and communities in high risk prone areas. A comprehensive exploration of several regressive machine learning models, to primary use in the wildfire forest protection services toolbox. Feature deterministic values are important in determining factors in volatile wildfires, so determining importance in factors such as relative humidity or temperature can be distinct between a low area burn or high area burn. Preliminary findings reveal distinct patterns in the importance of environmental factors in predicting power for wildfire areas burned. Neural Networks and logistic regression random forest models exhibit varying degrees of accuracy and predictive power. The comparative analysis sheds light on the strengths and limitations of each model. In order to train models with relative similarity to the climate of the United States, Forest Service Incident Management Team in Colorado, requires fast detection, which is important to controlling such an event. A case study in Portugal's Montesinho Natural Park (Northeastern region), simulates the various climate conditions and area of fire similar to most locations at risk from wildfires.

1 INTRODUCTION

Forest Fires otherwise known as Wildfires pose multifaceted threats to ecosystems, communities, and resources. The challenges of managing wildfires compounds from the volatility of forest fires, forest fire services conducively states that wildfires are an unpredictable force. A phenomenon can get out of control easily, the initial spread can grow to a state that is no longer controllable simply by the sheer size and violence of a forest fire. US forest policy, initiated U.S. After 1910 August a terrifying firestorm set YellowStone park ablaze into destroying Three Million Acres of land and leaving Brimstone. 90 deaths and Towns Destroyed, forest policy has evolved into quick actionary responses that were effective, but today wildfires are getting increasingly more catastrophic due climate changes large dense forests

Wildfire has an adverse human impact, some immediate threats of loss of lives, destruction of communities, and displacement of said communities. The further impacts of wildfire destruction stems to destruction of ecosystems; altering the biodiversity, landscape transformation, and polluting the air, and destruction of foundations; wildfire causes substantial economic costs; volatile wildfire can cause substantial economic efforts from services.

Artificial Intelligence tools offer a detectable, scalable, transformative, and highly beneficial approach to wildfire management. The ability to interpret vast amounts of data, analyze complex patterns, and provide rapid decisions is a critical advantage in detecting early signs of differing levels of volatile wildfires. Furthermore, the integration of Artificial Intelligence tools can enhance speed and accuracy of decision making that could prove valuable to enhancing the firefighters emergency response efforts. Utilization can improve the resilience and preparedness in the face of the everchanging unpredictable force of forest fires.

Development of models in this research, we have decided to utilize Neural Networks, emerging from the complex worm of the human brain inspired by the use of biological neurons. The neural network model was designed to perform intelligent systems of learning. The systems of learning use neural networks to adjust internal parameters, improving performance of a model over time. The network of neurons mimicking biological neurons, autonomously learns training, and with the interconnected woven thought cells create an output. Neural networks excel in multiple tasks: complex data processing, non-linear relationships, image and speech recognition, natural language processing, and predictive modeling.

Neural Networks is a multifaceted tool that can be applied to various statistical problems across different domains. Neural network can be utilized ins data/statistical use cases such as:

1. **Classification Problems** - neural networks proves valuable in classifying data into distinct categories
2. **Regression Problems** - In this case Neural Networks can predict numerical values
3. **Pattern Recognition** - Recognizes Complex patterns such as speech and facial recognition
4. **Time Series Analysis** - forecast patterns in time series data such as stock prices
5. **Natural Language Processing** - Widely used for language translation and sentiment analysis

The powerful and multipurpose Neural Network model could prove useful in our case study on wildfire forest management, the case study requires a regression model analysis.

Throughout the research paper, we will conduct model comparative analysis between modern standard practices of random forest, and neural networks. The associated reason we choose random forest, renown for its ensemble learning methods and resilience to overfitting, can prove effective in handling complex data and predicting the area of wildfire burned.The research will embark on determining the model with the most accurate and powerful model in predicting the wildfire area burnt. We all also conduct feature analysis, contributing to what features are highly significant in the area of wildfire burned.

2 METHODS

The data set was obtained from the center for machine learning repository at University of California, Irvine. The dataset has 517 burn events and 13 variables that were collected from January 2000 to December 2003. The data was collected in the Northeast region of Portugal of Montesinho National Park. Data aims to predict the area of wildfire burned highlighted red

Variable Name	Description	Range of Values
Area	The Burned area by wildfire (in ha)	0.0 - 1090.8
X	X - axis spatial coordinate	1 - 9
Y	Y- axis spatial coordinate	2 - 9
Month	Month of the year	Jan - Dec
Day	Day of the week	Mon - Sun
FFMC	Fine Fuel Moisture Code Index	18.7 - 96.2
DMC	Duff Moisture Code Index	1.1 - 291.3
DC	Drought Code Index	8.9 - 860.6
ISI	Initial Spread Index	0.0 - 56.10
Temp	Temperature (in Celsius)	2.2 - 33.3
RH	Relative humidity in %	15.0 - 100.0
Wind	Wind Speed in km/h	0.4 - 9.4
Rain	Outside rain in mm	0.0 - 6.4

Data Preparation: Data normalization process was conducted on the 13 properties, 4 of which (X,Y,Month, and Day) were not normalized. The remaining features were aggregated into a new dataset given the mapping from standard normalization technique. To establish a normalized base for linear values, employing 'max-min' normalization: marketed through this equation below

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

3 METHOD

Random Forest Regression: Supervised machine learning algorithm that is used for this purpose of the project to predict two cases utilizing ensemble learning methods. Random forest trains multiple decision trees, and random sampling with replacement. The decision trees are averaged across all predictions to produce a final prediction. Averaging models has the benefit of reducing overfitting. Random forest has become a staple machine learning model for data scientists. This will be used as a base against the primary model in question: neural networks. The method of random forest algorithm "is a highly flexible machine learning algorithm whose basic unit is a decision tree by integrating multiple trees into on through ensemble learning" (**Gao, C., Lin, H., & Hu, H.**) Some assumptions that are required for use of random forest models; the data does not have to be normal, linear, and homoscedasticity. Overall, a random forest can handle unbalanced data. Under Random Forest Regression the key concepts are:

1. **Decision Tree Equation:** for regression, the prediction of each leaf node is the average of the target values in that leafs
2. **Random Forest Prediction For Regression:** A measure from the output of each leave tells us how much the prediction is away from the original target using mean square error. The aggregation of predictions of each leaf averages the individual trees predictions

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^{2}$$

3. **Bootstrapping:** To minimize the variance, bootstrapping random selects N samples with replacement from the original dataset. Once all decision trees are grown, predictions are made of each tree of out of bag samples

Neural Networks: Artificial neural networks employ nodes with no form loops. During the data flow, nodes will receive data, travel through a weighted matrix of hidden layers where it will transform the input based on functions, and predict the output layer (Han, 2018). Neural networks are a fundamental component of deep learning models, a subfield of machine learning models. Deep learning involves training deep neural networks to automatically learn the data flow representation. These equations and concepts represent the core concepts using and training a neural network model (**Kartalopoulos,1997)**:

1. **Neurons:** Receives input , perform a computation, and produces an output
2. **Layers:** input layers; receives the initial data, hidden layer; intermediate layers that transforms data; Output layer; produces the final output
    a. Activation Function: forward propagation utilizing output of neuron

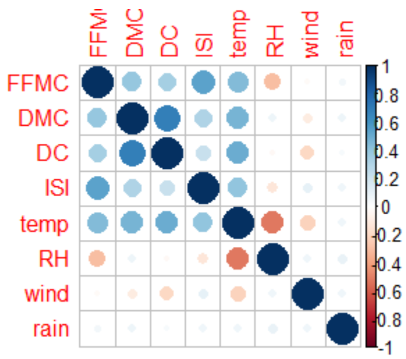
$$Output \ = \ F(Weight\ x \ + \ b) \quad \text{Activation } \sigma(z) \ = \ \frac{1}{1+e^{-z}}$$

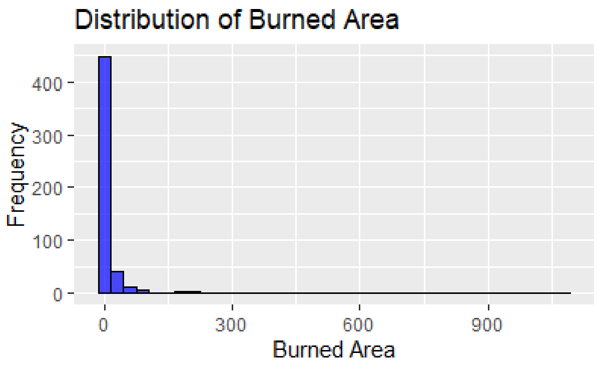
3. **Weights and biases**; each connection of neurons has n associated weight; represents the the strength of connection

Statistically to address the data, which data is fundamentally significant to wildfire prediction, does climate or conditions of meteorology affect wildfire distribution. Meteorology stems from the ground; features like fine fuel moisture code, drought code, and fine fuel moisture code are meteorological features in the dataset. Climate features include wind, relative humidity, and rain index. These grouping of variables can determine how volatile a wildfire burn can be.

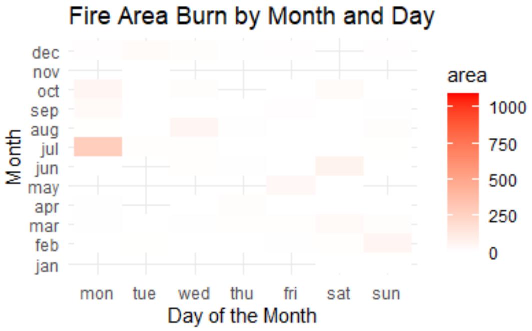
4 RESULT

Below are some preliminary data exploration. Meteorology data features show high multicollinearity; features such as DMC and DC, FFMC and ISI, show potential issues in multicollinearity. Another potential pitfall of the data is the left skewed data in burned area, as most of the distribution of burned area falls into 0-100 (ha), could impact models biases

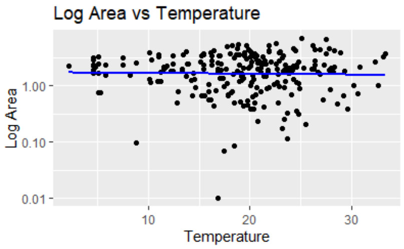


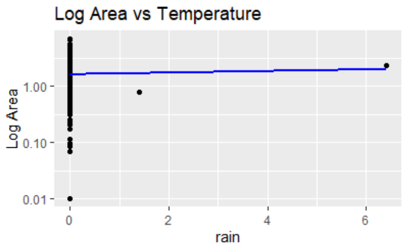


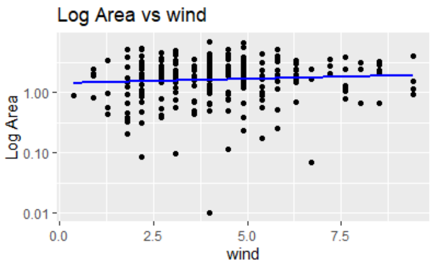
Below is a heatmap in response of the target area burned by month and day, months like July in the summer experiences a higher wildfire distribution of area burned (in ha).

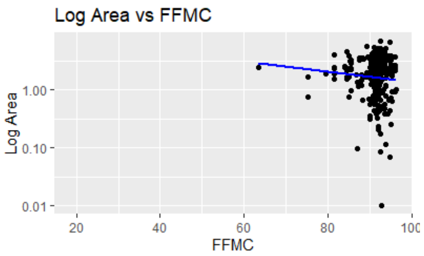


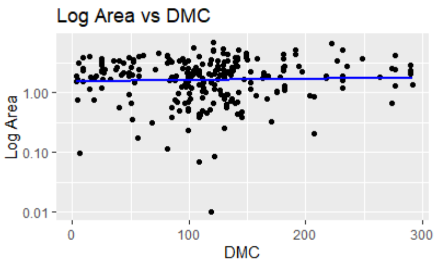
Below are some scatter plot of Log Area against meteorological and climate features

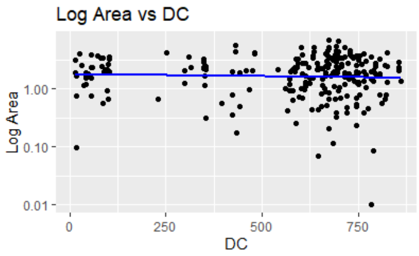


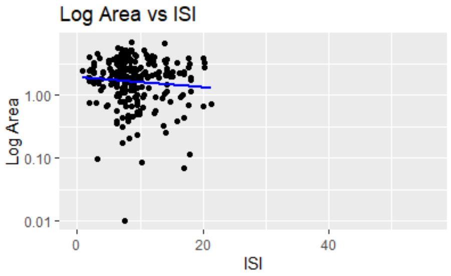


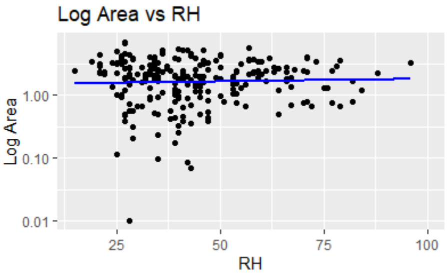








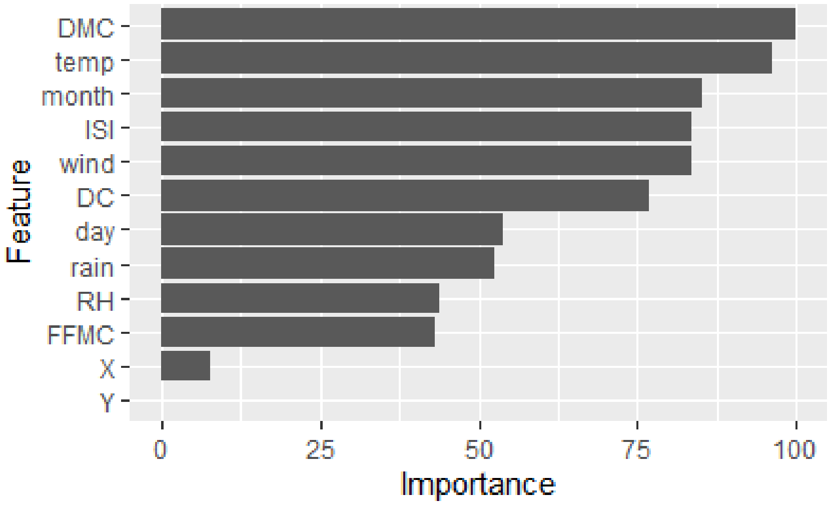




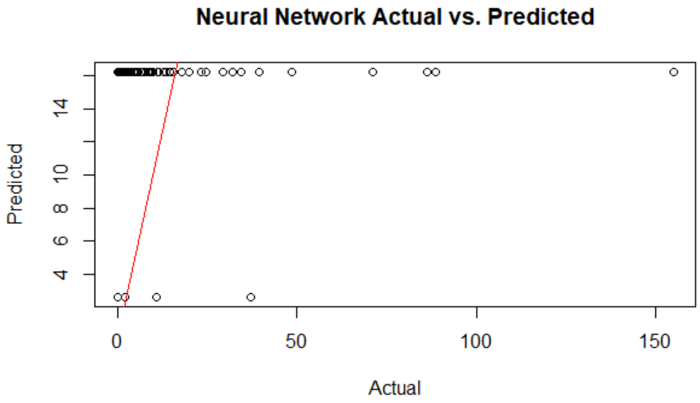
Utilizing log area, we can see that features like rain and wind, climate features show low relationship between features and log area.

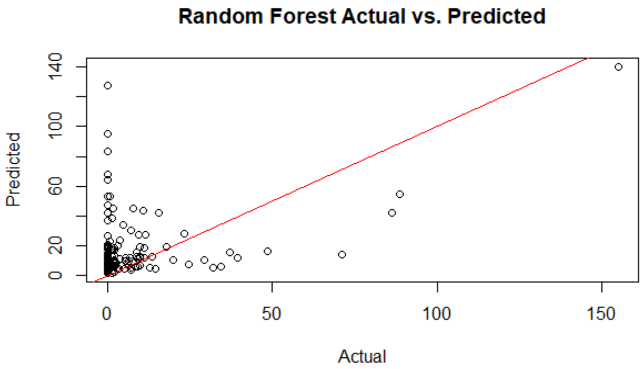
Variable Importance

Based on the scores generated by the model it is apparent that the most influential feature is DMC (duff moisture code), followed by temperature, month. Although Duff moisture code is a meteorological feature it seems that features of climate prove to take most of the higher ranking importance. Whereas features such as FFMC and DC stay in the lower half of the rankings. Based on feature selection it is best to ignore X and Y.



Model Metric Comparison





To compare regression models we will utilize mean square error, mean absolute percentage error, mean absolute error (average of absolute differences), and root mean square error which is more interpretable than the other metrics in model comparison. Below is a scatter plot of the models actual vs predicted values highlighted in black. Where random forest is closer to actual in the scatter plot:

Model Metrics

Model	MSE	MAPE	MAE	RMSE
Artificial Neural Network	407.4038	Inf	14.9607	20.1843
Random Forest	503.5503	Inf	15.0210	23.4557

From model metrics artificial neural networks incurs a lower MSE, MAE, and RMSE in the final model metrics

5 DISCUSSION AND SUMMARY

Addressing the research question in hand, what model is most flexible in addressing the fire area burned. If we were to employ the model, from research it is shown that the artificial model shows higher accuracy and predictive power. Neural network was highly adaptive, the only issue with the neural network was the ability to prove the importance of each variable and lower time to deploy the output. Random forest had the ability to employ feature selection/importance in determining whether meteorological or climate proves more significant in predicting wildfire area burned. Another concept to look at between models is the time it takes to process the data, under the neural network there was difficulty in interpreting the data under a lower step max, forcing the hyperparameter to a higher stepmax causes time to process. Another concept was ranking the importance between meteorological or climate features in predicting wildfire area burn. In the random forest model, Duff moisture code was the most significant feature, although temperature and wind ranked higher than meteorological data. A lower duff moisture code has higher significance in predicting area of fire burn.

To other models being used, naive, svm, and multiple regression it appears random forest has the fastest processing time to predict forest fires in a timely manner. Some potential limitations and suggestions for future research are data matching variables, feature selection sensitivity analysis of meteorological data vs climate data. Another method could be comparing classification models, determining lower, medium, and higher risk of volatile wildfire burn events. Most importantly for the speed of wildfires it is important to consider timing of the models

References

- Gao, C., Lin, H., & Hu, H. (2023). Forest-Fire-Risk Prediction Based on Random Forest and Backpropagation Neural Network of Heihe Area in Heilongjiang Province, China. Forests, 14(2), 170. MDPI AG. Retrieved from http://dx.doi.org/10.3390/f14020170

- Han, S. H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dementia and neurocognitive disorders, 17(3), 83–89. https://doi.org/10.12779/dnd.2018.17.3.83

- Kartalopoulos, S. V., & Kartakapoulos, S. V. (1997). Understanding neural networks and fuzzy logic: basic concepts and applications. Wiley-IEEE Press.

- Wood,D (2021). Prediction and Data Mining of burned areas of forest fires: optimized data matching and mining algorithm provides valuable insights, Science Direct,https://www.sciencedirect.com/science/article/pii/S2589721721000118

- Cortez,Paulo and Morais,Anbal. (2008). Forest Fires. UCI Machine Learning Repository. https://doi.org/10.24432/C5D88D.