Leveraging Data Mining Techniques to Predict Wildfires

Project Report

By

Abiodun Timothy Olaoye
Department of Mechanical Engineering, MIT

26th of March 2018

Executive Summary

Accurate prediction and classification of wildfire incidents are critical to mitigating damage by providing guidance for resource allocation to first responders. When these incidents occur information on the extent of damage is not immediately available but meteorological data are usually retrievable hence, the need for data-driven algorithms to predict the damage. Three supervised learning methods are employed in this project namely linear regression, k-nearest neighbors and neural network.

The neural network model gives the best predictive performance for this application while the multiple linear regression model has the worst performance.

For the classification procedure, it should be noted that, classifying severe occurrences is more important than non-severe ones. Logistic regression model (cutoff value = 0.5) fails to accurately classify any severe occurrence, while the KNN model (cutoff = 0.5) performs better at classifying severe occurrences than the other models. Although the neural network model gives the best overall accuracy, the KNN model is adjudged the best of the three classifiers for this application.

Finally, the neural network model is the most computationally expensive using two hidden layers with three nodes per layer in the predictive model and a single hidden layer with four nodes in the classifier model.

Background

Wildfires also known as forest fires are destructive occurrences which pose ecological, economic and environmental challenges while being fatal in extreme cases. Furthermore, weather conditions such as temperature, humidity and wind impact the severity of wildfire outbreaks. However, the amount of damage may be mitigated through quick responses by fire-fighting officers and adequate provision of necessary fire-fighting resources.

Hence, the main objective of this study is to predict the burned area of moderate wildfires from meteorological predictors using supervised data mining techniques such as linear regression and k-nearest neighbors. Attempt will be made to classify occurrences into severe and non-severe events depending on size of the burned area. This would enable concerned municipal authorities to properly allocate resources based on estimated damage. The dataset is for northeast region of Portugal[1].

Methodology

The dataset consists of 517 records with 12 predictors and an outcome variable (burned area). Domain knowledge reveals that some of the predictors are well correlated hence, the need for variable selection. Furthermore, logarithmic transformation of the burned area data is required due to the positive skewness of the variable[2].

Hence, the data mining tasks associated with the proposed work include identification of significant predictors, scale transformation, application of supervised techniques such as linear regression, k- nearest neighbors and logistic regression classifier to predict target burned area and classify severity of occurrences.

To implement the classification procedure, a new categorical variable will be created from the burned area data. Furthermore, two dummy variables will be used to represent the categorical variable as follows: *non-severe* – 0; *severe* – 1; Cataloging of the burned area data into severe or non-severe occurrence (using some cut off value) is mainly for academic purpose to demonstrate the classification procedure and not based on industry standard or government policy.

The different data mining techniques employed will be assessed using their predictive performance on the validation set (30% of observations). Machine learning phenomena such as overfitting of the training data will be discussed.

Preprocessing

Although the obtained data has been preprocessed (records with missing values removed) [2], only 4 meteorological variables (temp, relative humidity, wind speed and rain) out of 12 predictors are employed in building the different models.

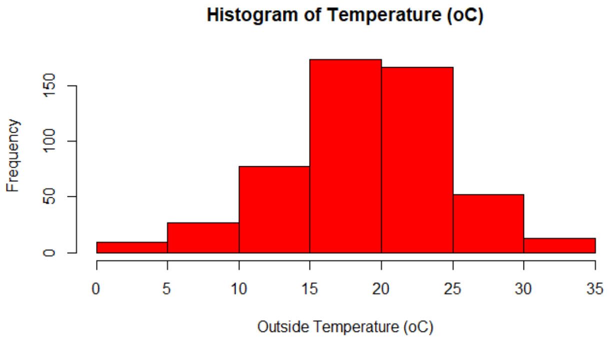
Summary statistics

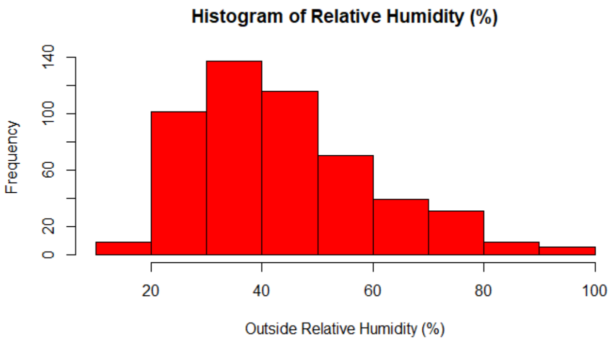
The table below shows the fundamental statistical measures for all variables in dataset. Only the outcome variable (burned area) has a median value significantly different from the mean. This indicates skewness of the variable hence, a new variable area.log is introduced which is a logarithmic transformation of the original burned area data.

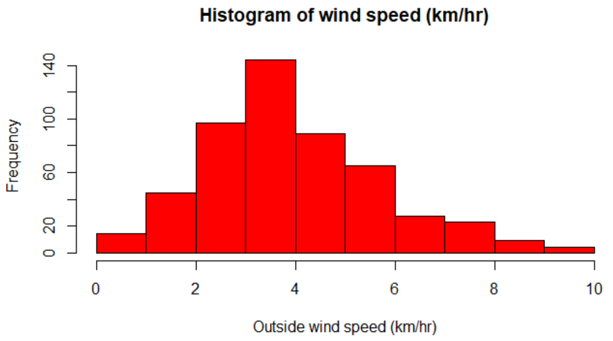
Table 1. Summary statistics of variables in forestfires dataset

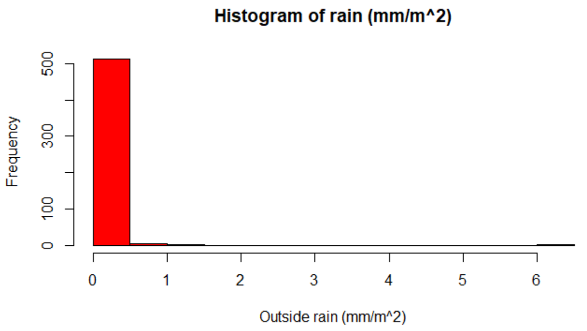
	mean	median	min	max	SD
X	4.669246	4	1	9	2.313778
Y	4.299807	4	2	9	1.2299
FFMC	90.64468	91.6	18.7	96.2	5.520111
DMC	110.8723	108.3	1.1	291.3	64.04648
DC	547.94	664.2	7.9	860.6	248.0662
ISI	9.021663	8.4	0	56.1	4.559477
temp (^{o}C)	**18.88917**	**19.3**	**2.2**	**33.3**	**5.806625**
RH (%)	**44.2882**	**42**	**15**	**100**	**16.31747**
wind (km/h)	**4.017602**	**4**	**0.4**	**9.4**	**1.791653**
rain (mm/m2)	**0.021663**	**0**	**0**	**6.4**	**0.295959**
area (ha)	12.84729	0.52	0	1090.84	63.65582
area.log ($\log_{10}$(1+area))	**0.482512**	**0.181844**	**0**	**3.038159**	**0.607333**

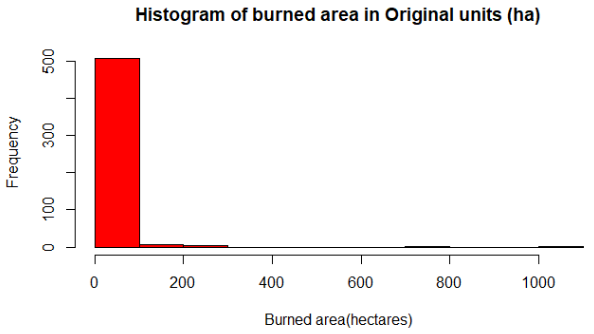
Next, the histograms of the 4 selected predictors and outcome variable (original and logarithmic transformation) are shown.

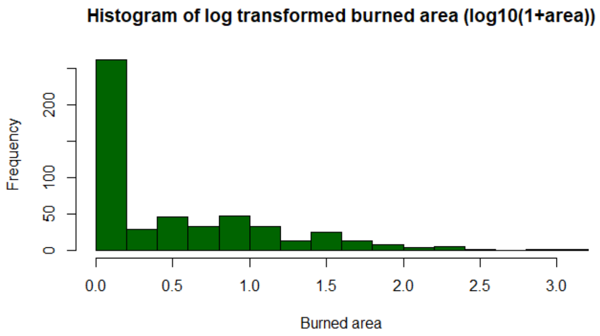












Predictive Models

Three different predictive models are employed namely, linear regression, K-nearest neighbors and neural network. The predictive performance of each model is assessed using the mean absolute error (MAE) and root mean square error (RMSE). The model with the smaller of these values is adjudged relatively better for this problem.

Multiple Linear Regression

This method utilizes a linear relationship between the predictors and the outcome variable. Hence, accuracy may be quite low for problems involving more complex relationship between the predictors and the outcome variables.

Table 2. Multiple linear regression coefficients

(Intercept)	temp	RH	wind
2.22E-16	0.099504	0.001959	0.067791

k-Nearest neighbors

This method makes predictions based on the *k* nearest neighbors to the new data. A value of k which is too high may result in not sufficiently capturing the relationship between the predictors and outcome variable. However, overfitting of the data may result from selecting a value of *k* too low. It is therefore important to carry out an experiment choose an optimum value of *k*.

Table 3. MAE of different *k* values for training and validation set

k	MAE (training)	MAE (validation)
1	0.037803452	0.595216835
2	0.327792940	0.573947360
3	0.376077723	0.572695024
4	0.426398657	0.548456433
5	0.434309089	0.548021343
6	0.435646423	0.493307486
7	0.463263847	0.475399317
8	**0.478267525**	**0.466625811**
9	0.480382831	0.470002102
10	0.486622340	0.468171989
11	0.484920274	0.468171989
12	0.486123493	0.468171989
13	0.486123493	0.468171989
14	0.486123493	0.468171989
15	0.486123493	0.468171989

The KNN model is therefore assessed using *k* = 8 from the table above.

Neural Network Model

This model computes the relationship between predictors and outcome variable by employing weights attached to neurons connecting the nodes of different layers (input, hidden and output layers). The process of improving predictive performance by adjusting the weights is known as learning. This method can handle predictor-outcome variable relationships more complex than linear as required by linear regression method for example. Overfitting may occur if the number of hidden layers or nodes per hidden layer is too large. Hence, the need to carry out some experiment to select the optimal parameters for this problem.

From Table 4, a single layer with 4 nodes gives about the same predictive performance as two layers with three nodes per layer. However, the later is chosen to assess the neural network model since it gave the optimal performance.

Table 4. MAE values for training and validation set of (a) single and (b) two hidden layers models

(a)

Single Hidden Layer		
Nodes per layer	MAE (training)	MAE (validation)
1	0.349804052	0.347684162
2	0.348032659	0.355816719
3	0.350940074	0.367940592
4	**0.350828982**	**0.354082932**
5	0.353785090	0.406676183
6	0.351156327	0.354734780
7	0.349719607	0.363750527
8	0.351118831	0.358331334
9	0.347635293	0.402013844
10	0.350812829	0.361513760

(b)

Two Hidden Layers		
Nodes per layer	MAE (training)	MAE (validation)
1	0.349667138	0.354109814
2	0.348632823	0.374228340
3	**0.350366965**	**0.350786340**
4	0.348146353	0.383656125
5	0.349841389	0.355431960
6	0.350332304	0.546403498
7	0.350478904	0.351782113
8	0.361726999	0.519847135
9	0.349809459	0.514949325
10	0.352762556	0.529591278

Comparison of the Predictive Models

The RMSE and MAE values of the transformed burned area for the three selected supervised learning tools are used to assess their predictive performance.

The optimal neural network model chosen above gives the best prediction for this application followed by the KNN model. This may be as a result of the neural network model's ability to better represent complex relationships between the meteorological predictors and the burned area compared to other models employed.

Table 5. Comparison of predictive models based on RMSE and MAE values

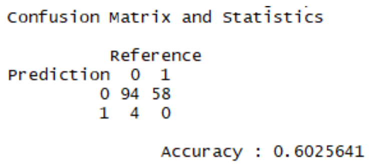
Method	RMSE	MAE
Multiple Linear Regression	0.988612	0.819527
K-Nearest Neighbors	0.759208	0.466626
Neural Network	**0.391865**	**0.350786**

Classification Models

For allocating resources for fire occurrences, it is important to be able to classify incidents as non-severe or severe in addition to predicting the burned area value. After the logarithmic transformation of the burned area data, any burned area value less than 0.5 (corresponds to approximately 2 hectares of burned area) is classified as non-severe or severe otherwise. Logistic regression, k-nearest neighbors and neural network models are employed.

Logistic Regression

This method is like linear regression as it employs a model which defines the relationship between predictors and outcome variable but classify a new record into one of the classes (in this case non-severe or severe) based on some computed propensities.



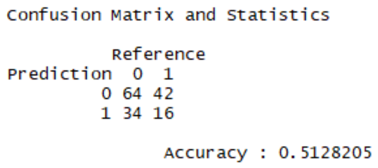
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 94 58
         1  4  0

               Accuracy : 0.6025641
```

K-Nearest Neighbor Classifier

Similar idea as KNN predictor by identifying records closest to the new record. Closeness may be measured by different distance measures including Euclidean distance. Also, the new record is assigned to the majority class among the *k* nearest neighbors. The optimal value of *k* employed in the predictive model is used here.



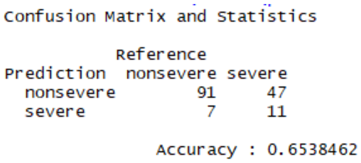
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 64 42
         1 34 16

               Accuracy : 0.5128205
```

Neural Network Classifier

Here, the neural network model is used to compute the propensity of a new record belonging to one class rather than another class. A single hidden layer with four nodes per layer is employed.



```
Confusion Matrix and Statistics

           Reference
Prediction  nonsevere severe
  nonsevere        91     47
  severe            7     11

               Accuracy : 0.6538462
```

Comparing the classifiers

From the confusion matrices for each of the model, neural network has the overall accuracy best accuracy but the KNN model (cutoff = 0.5) is better at classifying the severe occurrences than the other models. It is noteworthy that, misclassifying a severe event is costlier than misclassifying a non-severe one. Hence, the KNN model is the best classifier for this application.

Conclusion

Machine-learning algorithms namely multiple linear regression, k-nearest neighbor and neural network have been applied to a set of meteorological predictors with the aim of predicting the extent of damage and subsequent classification as either severe or non-severe occurrence.

The neural network model has the best predictive performance and gave better overall accuracy for classification of the wildfire occurrences than other models considered. However, the KNN model is better at classifying severe occurrences which is more important for this application. Hence, the KNN classifier is adjudged the best classifier for this problem.

References

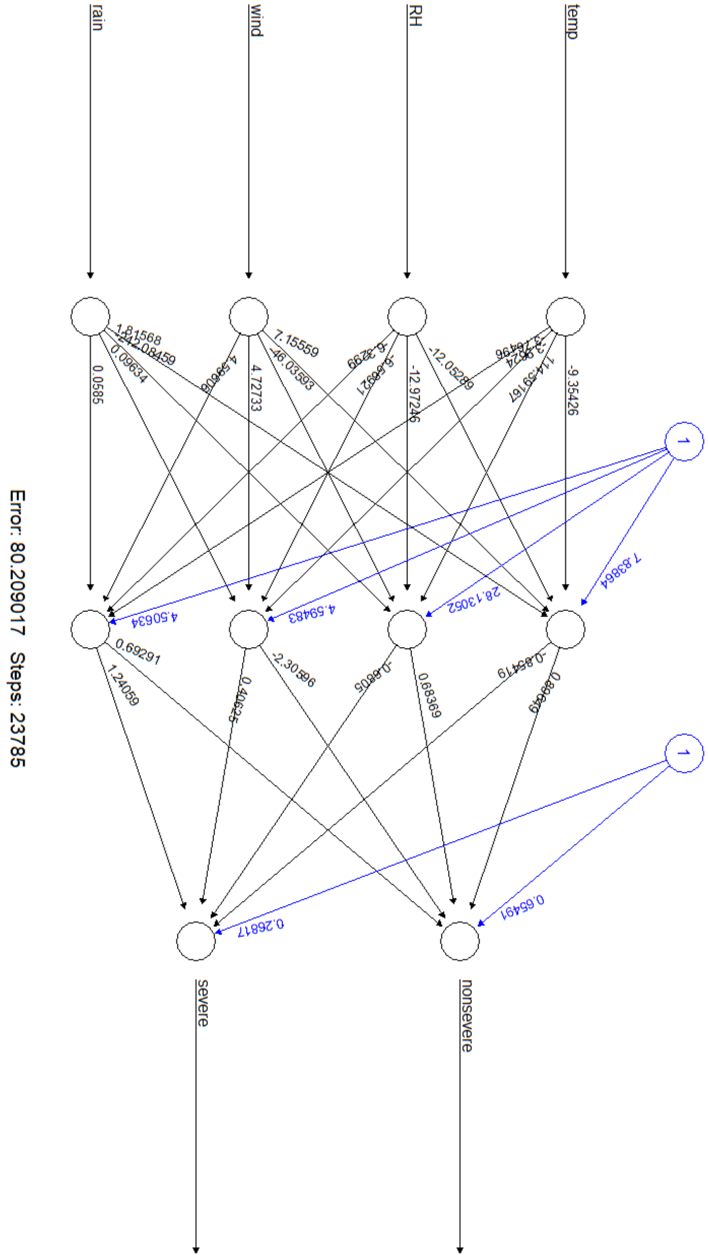
[1] UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems, http://archive.ics.uci.edu/ml/datasets/Forest+Fires; Source: Paulo Cortez, pcortez '@' dsi.uminho.pt, Department of Information Systems, University of Minho, Portugal. Aníbal Morais, araimorais '@' gmail.com, Department of Information Systems, University of Minho, Portugal.

[2] P. Cortez and A. Morais, A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9.

[3] G. Shmueli et. al., Data Mining for Business Analytics (Concepts, Techniques, and Applications in R), 1st edition, © 2018 John Wiley & Sons, Inc.

Appendix

Neural Network Classifier for predicting severity of Wildfires from Meteorological Data


rain
wind
RH
temp
1.81568
-242.08459
0.09634
0.0585
7.15559
-46.03593
4.59606
4.72733
-6.3299
-6.60821
-12.97246
-12.05289
-3.76496
-3.924
114.59167
-9.35426
1
7.8364
28.13052
4.59483
4.50634
0.69291
1.24059
-2.30596
0.40625
-0.6605
0.68369
-0.85419
0.88649
1
0.65491
0.26817
severe
nonsevere
Error: 80.209017 Steps: 23785