BUSINESS ANALYTICS REPORT

Group 24

William Zhou, Sosan Thomas & Janaka Cooray

Table of Contents

Executive Summary

Purpose:

The purpose of this report is to identify factors that can predict the occurrence of forest fires by conducting an analysis of the various weather conditions. It also recommends measures to combat them. The analysis is conducted by using different testing methods and models.

Key Descriptive Statistics:

The data contains 15 features, including 3 classification factors: month, region and classes. Numeric variables consist of 4 weather variables and the other 6 variables calculated from weather variables. To avoid overfitting, we only take fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC) and initial spread index (ISI) and the 4 key weather variables to use in further model analysis.

Main Result and Inferences:

With outcomes from multiple machine learning models, the prediction outcome from the random forest model is best fit. It indicated that the velocity of wind spread (Ws), relative humidity (RH), Drought Code (DC), Initial Spread Index (ISI) are significantly contributing to the occurrence of forest fires in the two distinct regions of Algeria.

Recommendations and conclusion:

By gathering insights of outcomes from models we have undergone, we recommend Algerian authorities to enhance their weather monitoring capabilities by deploying more sensors and forecasting devices, particularly during the summer months. This would enable better tracking of temperature, wind, relative humidity, and precipitation, crucial for fire risk assessment. Secondly, we propose utilizing this weather data to calculate the Fire Weather Index (FWI), which quantifies fire intensity potential. When FWI reaches the hazardous level of 5, prompt action should be taken, including deploying additional fire control and rescue forces to prepare for and mitigate potential disasters.

Introduction

Forests play a crucial role in maintaining the ecological balance of the earth. However, they are also susceptible to the catastrophic impact of forest fires, which can result in devastating consequences for both ecosystems and human life. Forest fires are a common disturbance in many forest systems in the world and particularly in the Mediterranean region (Gonçalves & Sousa, 2017). The risk of forest fires is particularly high in the countries in the Mediterranean basin, including Algeria, Morocco, Portugal, Spain, and France. Our dataset comprises 244 instances, combining data from two distinct regions of Algeria: The Bejaia region, situated in the northeast, and the Sidi Bel-abbes region, located in the northwest. This report focuses on the analysis of Algerian forest fire data and aims to develop a predictive model that can identify the likelihood of fires. Weather conditions play a key role in influencing forest fires. Variables such as wind speed (Ws), relative humidity (RH), temperature (Temp), and precipitation (Rain) are key factors that can affect the risk of fire outbreaks. These variables, as well as others like the Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI), and the Fire Weather Index (FWI), collectively form the foundation of our analysis.

Method

The Algerian Forest fire data only contains data within 4 months, aiming to produce an accurate fitting model, the analysis had been divided into three parts, data decluttering, correlation analysis and fitting model using random forest. Before that, a few assumptions were proposed.

Assumptions:

To identify which variables can predict the best whether a forest fire is going to happen or not, the dependent variable “Classes” was chosen which is a factor with two levels: “fire” and “not fire”. Through prior research, it was found that the 6-fire weather index (FWI) components are calculated by 4 key factors (Xie. H, 2017). Among these factors, fine fuel moisture code (FFMC) directly relates to the ignition and spread of fire and relates with the four main variables. Therefore, our objective is to assess whether FFMC will be the most significant observation when predicting the occurrence of a forest fire.

Correlation Analysis:

To check the correlation between variables, all the numeric variables were declared in the first instance in our data by using “as.numeric” including the 4 main weather variables. As from the data description of this data set, variable BUI and FWI is calculated using DC, DMC, ISI, so to avoid overfitting in later analysis, we only use "FFMC", "DMC", "DC", "ISI", "Temperature", "Rain", "RH","Ws" as numeric variables. Then a correlation matrix was produced using ggpairs function from GGally package. From the matrix it was found that DMC and DC have a strong linear relationship as well as FFMC and ISI.

Data Segmentation:

As the original data been divided into two parts by region, firstly, the data was re-joined by creating a new 2-levels factor “Region”. Then the training set was generated with 90% of the new data while the testing

set with 10% against the factor “Classes”. Function “set seed” was used for generating random numbers to make the results repeatable, by set seed (12222).

Model Prediction:

In this part, the first attempt was to use glm( ) and stepwise regression to generate LR model, but none of the coefficients of the logistic regression results are significant enough. Then decision tree model was used, and to avoid over-fitting random forest model was chosen to generate final predictioni model which is the best fit model. All the final prediction models are examined with AUC-value and confusion matrices (Appendix 2, 3, 4, 5, 6).

Descriptive Statistics and Preliminary Correlation Analysis

Descriptive Statistics:

To examine the factors and numeric variables in this data set, “Classes”, “month” and “Region” were taken as factors, and the numeric variables are mentioned earlier above. Then the first plotted output was the correlation matrix (Figure 1) with point plot, density plot and correlation-value to check the correlation between pairwise variables. Also, for loops were used to plot combined density plots and bar charts using all 8 variables and set “Classes” as fill, to examine, factor “Classes” and numeric variable “Temperature” were used to build a cross-table by running chisq.test. The result was consistent with our density plot (Figure 2).

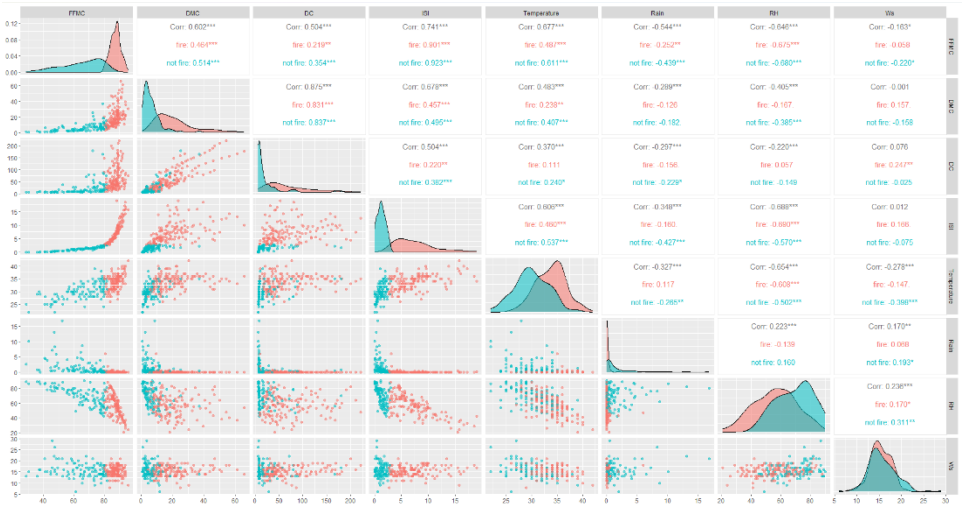

Figure 1: Correlation matrix

Preliminary Correlation Analysis:

From the correlation matrix, it shows that DMC and DC have a strong linear relationship as well as FFMC and ISI which their absolute correlation values larger than 0.8 and both are statistically significant. Also, from the density plot it indicates that when FFMC is larger than 80, the count of forest fire rises rapidly.

Finally, before going to analysis part, a scatter plot was generated using all numeric variables to check if there were any possibility curves exist for future logistic regression model, which turned out that none of the numeric is good enough to generate LR model.

```
> TemperatureTable <- xtabs(~Classes + Temperature, data = ForestfireData)
> print(TemperatureTable)
            Temperature
Classes      22 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 42
  fire        0  0  0  1  1  3  5  7 16 12 15 16 22 20  9  2  5  3  1
  not.fire    2  3  6  4  7 12 13 15  9  9  8  8  7  1  0  1  1  0  0
> summary(TemperatureTable)
Call: xtabs(formula = ~Classes + Temperature, data = ForestfireData)
Number of cases in table: 244
Number of factors: 2
Test for independence of all factors:
        Chisq = 74.38, df = 18, p-value = 8.083e-09
        Chi-squared approximation may be incorrect
```

Figure 2: Output of cross table

Analytics

To select the best fit model, the analytics started with logistic regression model. Although from the preliminary analysis result, it could be concluded that an LR model may not fit enough, an LR model for all numeric variables (model_all) using formula = Classes ~., were generated, then the possibility curve (Figure 3) of the prediction model of modell_all was plotted. The result (Appendix 1) was that none of the coefficients from the logistic regression results were significant using Classes as independent variables.

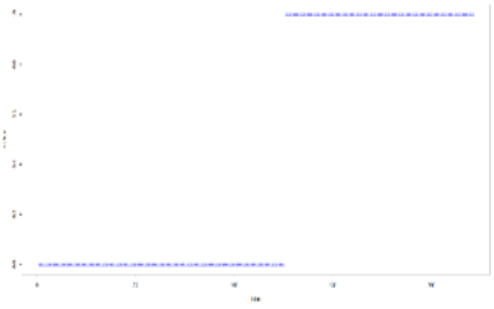

Figure 3: possibility curve of prediction model

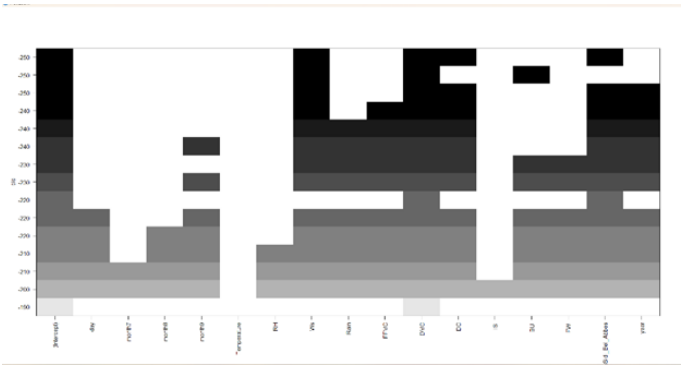

Figure 4: Plot of model0

For double confirmation and to avoid exhaustive enumeration, a model (model0) using stepwise regression to filter out influential variables was produced, and the stepwise regression model was plotted out (Figure 4). The conclusion is that DMC, DC and WS are influential against Classes and among these three the most significant variable is DMC. The prediction model using DMC and Region only (model2) reached an AUC-value of 0.946 with test data.

As for comparison experiment, a decision tree model(modelDT2) using same formula: Classes ~ DMC + Region (Figure 5) was generated and the AUC-value of modelDT2's prediction is 0.900. Also, a decision tree model using all numeric variables(modelall) was generated (Figure 6), which have only one node and

only returns with one judgement FFMC >= 80, the AUC-value of its prediction is 0.950 (Figure 7). However, to avoid over-fitting, a random forest algorithm was chosen to build the final prediction model.

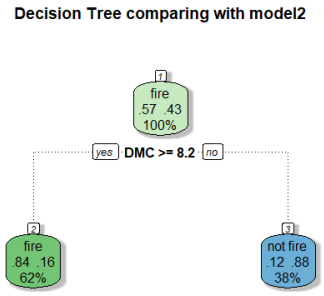

Figure 5：Decision Tree model 2

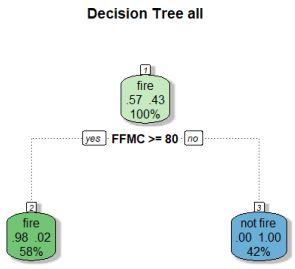

Figure 6: Decision Tree all

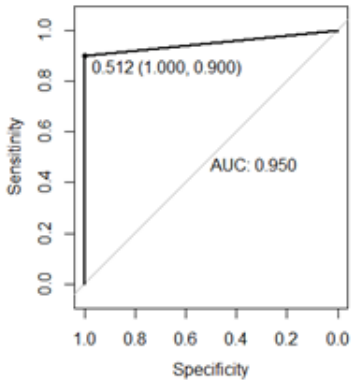

Figure 7: ROC curve of modelall

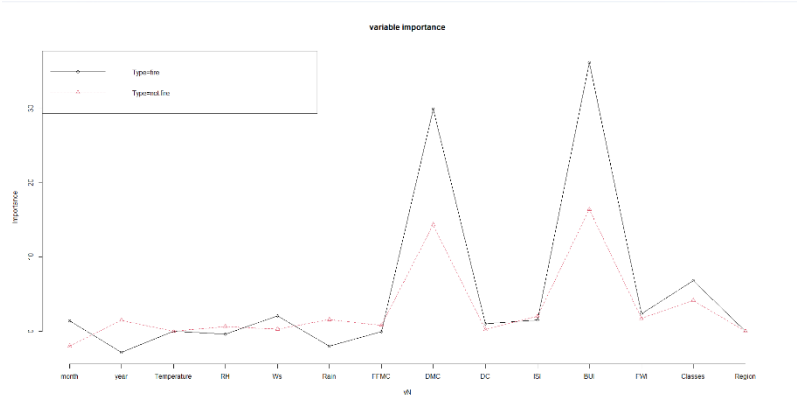

Figure 8: The importance of variables

By library R package “randomForest”, Algerian_rf1 model was created using R function randomForest(). Firstly, the importance of all variables was examined, which the result been attached in appendix 2, and was plotted out in line chart (Figure 8) and bar charts, the bar charts are according to mean decrease accuracy and mean decrease gini. Then by using the top 5 most important variables: “Ws”, “RH”, “DC”, “ISI” and “month”, the final prediction model: predictionRF_2 (Appendix 7) was generated. The AUC-value of this model is 0.992 with kappa = 0.9105, and from the confusion matrix, the accuracy is 95.7% while 95% confidence interval is 78.1% to 99.9%, the complete output is in appendix 3. The comparison table of all three models is as follows (Table1), which concluded random forest model was best fit.

Models	**Used variables**	**AUC- value**
LR Model	DC (Drought Variable)	0.946
Decision Tree Model	All 8 numeric variables	0.95
Random Forest Model	DC (Drought Code), Ws (Wind speed), RH(Relative Humidity), ISI (Initial Spread Index), and month	0.992

Table 1: Comparison of the three models

Recommendation and Conclusion

Recommendation 1:

Algerian authorities should focus on employing more sensors and other forecasting devices to monitor weather conditions such as the temperature, wind, relative humidity, and level of precipitation during the summer months.

Recommendation 2:

The information on weather conditions can be utilized to calculate the Fire Weather Index (FWI), which is a numeric estimation of fire intensity potential. When FWI reaches the hazardous levels of 5, the Algerian government can act on it immediately by sending more fire control and rescue team personnel to prepare for the disaster and for further rescue activities.

Conclusion:

In conclusion, forest fires pose serious threats to life and can cause large scale losses, urging us to act immediately. This analysis is significant to various stakeholders. It draws on the urgent need for collaborative effort across various fields such as AI, remote sensing, meteorology, and geology to minimize the impact of forest fires on human lives through the timely, effective, and precise distribution of information.

Government agencies responsible for disaster management and firefighting can benefit from the predictive models developed here to allocate resources effectively and plan emergency responses. Organizations involved in the management of forests can utilize these findings to introduce land management practices and reduce fire-related risks. Environmental conservation groups can leverage this data to advocate for policies that safeguard ecosystems. This will also be beneficial to businesses and communities to understand and prepare for the disaster beforehand.

To effectively combat forest fires, it is crucial to be a step ahead. By constantly optimizing the model, not only can we predict the occurrences of forest fires, but also measure their intensity. This helps in timely intervention by deploying firefighting resources swiftly and efficiently.

Appendices:

Appendix 1: Summary of LRmodel_all

```
> summary(model_all)

Call:
glm(formula = Classes ~ ., family = binomial(link = "logit"),
    data = Algerian_train, control = list(maxit = 100))

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.189e+03  8.558e+06       0        1
day                  2.060e+00  6.581e+04       0        1
month7               3.615e+01  2.492e+05       0        1
month8               7.007e+01  6.520e+05       0        1
month9               2.596e+01  1.203e+06       0        1
year                        NA         NA      NA       NA
Temperature          4.485e+00  2.240e+05       0        1
RH                  -1.945e+00  1.817e+04       0        1
Ws                  -2.737e+00  6.661e+04       0        1
Rain                -5.110e+01  1.895e+05       0        1
FFMC                -2.697e+01  9.474e+04       0        1
DMC                  8.043e+00  1.079e+05       0        1
DC                  -1.824e+00  2.042e+04       0        1
ISI                  1.953e+01  1.363e+05       0        1
BUI                  1.274e+00  9.139e+04       0        1
FWI                 -3.269e+01  2.532e+05       0        1
RegionSidi_Bel_Abbes -1.828e+01  7.806e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.0255e+02  on 220  degrees of freedom
Residual deviance: 5.8612e-10  on 205  degrees of freedom
AIC: 32

Number of Fisher Scoring iterations: 30
```

Appendix 2: Importance of variables

```
> importance(Algerian_rf1,type = 1)
            MeanDecreaseAccuracy
day                    0.6495293
month                 -0.7711648
year                   0.0000000
Temperature            1.5017285
RH                     3.1221300
Ws                    -1.5094316
Rain                   1.5983242
FFMC                  32.0631920
DMC                    1.7821717
DC                     3.1596217
ISI                   35.3963186
BUI                    3.3925478
FWI                    8.6725294
Region                 1.3395332
```

Appendix 3: Output of confusion matrix of predictionRF_2

```
> predictionRF2_cf
Confusion Matrix and Statistics

          Reference
Prediction fire not fire
  fire       13        1
  not fire    0        9

               Accuracy : 0.9565
                 95% CI : (0.7805, 0.9989)
    No Information Rate : 0.5652
    P-Value [Acc > NIR] : 3.738e-05

                  Kappa : 0.9105

 Mcnemar's Test P-Value : 0.001787

            Sensitivity : 0.9926
            Specificity : 0.9000
         Pos Pred Value : 0.9286
         Neg Pred Value : 1.0000
             Prevalence : 0.5652
         Detection Rate : 0.5652
   Detection Prevalence : 0.6087
      Balanced Accuracy : 0.9500

       'Positive' Class : fire
```

Appendix 4: Output of confusion matrix of predictionDT_all

```
> predictionDTall

Confusion Matrix and Statistics

          Reference
Prediction fire not fire
  fire       13        1
  not fire    0        9

               Accuracy : 0.9565
                 95% CI : (0.7805, 0.9989)
    No Information Rate : 0.5652
    P-Value [Acc > NIR] : 3.738e-05

                  Kappa : 0.9105

 Mcnemar's Test P-Value : 1

            Sensitivity : 1.0000
            Specificity : 0.9000
         Pos Pred Value : 0.9286
         Neg Pred Value : 1.0000
```

```
             Prevalence : 0.5652
         Detection Rate : 0.5652
   Detection Prevalence : 0.6087
      Balanced Accuracy : 0.9500

       'Positive' Class : fire
```

Appendix 5: Output of confusion matrix of predictionDT2

```
> predictionDT2

Confusion Matrix and Statistics

          Reference
Prediction fire not fire
  fire       13        2
  not fire    0        8

               Accuracy : 0.913
                 95% CI : (0.7196, 0.9893)
    No Information Rate : 0.5652
    P-Value [Acc > NIR] : 0.0003367

                  Kappa : 0.8189

 Mcnemar's Test P-Value : 0.4795001

            Sensitivity : 1.0000
            Specificity : 0.8000
         Pos Pred Value : 0.8667
         Neg Pred Value : 1.0000
             Prevalence : 0.5652
         Detection Rate : 0.5652
   Detection Prevalence : 0.6522
      Balanced Accuracy : 0.9000

       'Positive' Class : fire
```

Appendix 6: Output of confusion matrix of predictionDT1

```
> predictionDT1

Confusion Matrix and Statistics

          Reference
Prediction fire not fire
  fire       12        2
  not fire    1        8

               Accuracy : 0.8696
                 95% CI : (0.6641, 0.9722)
    No Information Rate : 0.5652
    P-Value [Acc > NIR] : 0.001949

                  Kappa : 0.7315

 Mcnemar's Test P-Value : 1.000000
```

```
            Sensitivity : 0.9231
            Specificity : 0.8000
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.8889
             Prevalence : 0.5652
         Detection Rate : 0.5217
   Detection Prevalence : 0.6087
      Balanced Accuracy : 0.8615

       'Positive' Class : fire
```

Appendix 7: Output of Algerian_rf2

```
> print(Algerian_rf2)

Call:
 randomForest(formula = Classes ~ month + RH + DC + Ws + ISI,      data =
Algerian_train, mtry = 10, importance = TRUE, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 1.81%
Confusion matrix:
         fire not fire class.error
fire      122        3  0.02400000
not fire    1       95  0.01041667
```

Reference List

Gonçalves, A. C., & Sousa, A. M. O. (2017). The fire in the Mediterranean Region: A case study of forest fires in Portugal. Mediterranean Identities - Environment, Society, Culture. https://doi.org/10.5772/intechopen.69410

Huang. J. X., (2017). ForestFires. Github. https://github.com/axelhuang24/ForestFires