**Case Study – A Data Mining Approach to Predict Forest Fires Using Linear Regression**

**Abstract**:

Forest fires are a major environmental issue, creating economical and ecological damage while endangering human lives. Fast detection is a key element for controlling such phenomenon. To achieve this, one alternative is to use automatic tools based on local sensors, such as provided by meteorological stations. In effect, meteorological conditions (e.g. temperature, wind) are known to influence forest fires and several fire indexes, such as the forest Fire Weather Index (FWI), use such data. In this work, we explore a Data Mining (DM) approach to predict the burned area of forest fires.

**1. Introduction:**

Forecasting the forest fires (also called wildfires) now becomes the crucial issue now a days which affects the forestation, and do ecological and environmental damage which in turn affects the humans and livestock directly and indirectly.

Humans suffers the most from this phenomenon. National economies are strongly linked and heavily influenced by the wildfires. The key setback for predicting the area is its uncertainty. This particular feature is undesirable for the firefighting responsiveness by the fire fighters however, it is unavoidable. This uncertainty will remain and cannot be ruled out however, an attempt can be made to minimize its gravity.

Forest Fires Forecasting is one of the solutions in this process. Each year around millions of hectares(ha) of forest land are destroyed all around the world.Portugal is highly affected by the forest fires [13].

The 2003 and 2005 fire seasons were especially dramatic, affecting 4.6% and 3.1% of the territory with 21 and 18 human deaths and vast livestock.Early detection is the key element for successful firefighting and firefighting preparedness.

**2. Background Study**

**Some of the Popular Prediction Model to Solve the Problem:**

1. Support Vector Machines (SVM)
2. Random Forests
3. four distinct feature selection setups (using spatial
4. temporal
5. FWI components and weather attributes)

were tested on recent real-world data collected from the northeast region of Portugal. The best configuration uses a SVM and four meteorological inputs (i.e. temperature, relative humidity, rain and wind) and it is capable of predicting the burned area of small fires, which are more frequent. Such knowledge is particularly useful for improving firefighting resource management.

**Forest Fire Data:**

Our study is considering the fire data from Montesinho Natural Park situated in Trans-os-Montes northeast region of Portugal

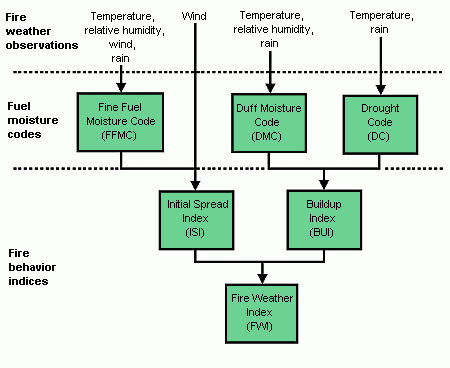
This park contains a high flora and fauna diversity having Mediterranean Climate, the average annual temperature is with in the range 8º C to 12º C. The data used in our model was collected from January 2000 to December 2003.

The data contains 517 entries. It has 12 variables all of which has been recorded for three years.

The Forest Fire Weather Index is the Canadian System for rating fire danger and it includes six components.

1. Fine Fuel Moisture Code (FFMC),
2. Duff Moisture Code (DMC),
3. Drought Code (DC),
4. Initial Spread Index (ISI),
5. Buildup Index (BUI)
6. FWI.

**The Fire Weather Index structure:**

****

**3. DESCRIPTION OF DATASET ATTRIBUTES**

1. X=x-axis coordinate (from 1 to 9)
2. Y=y-axis coordinate (from 1 to 9)
3. Month=Month of the Year(January to December)
4. FFMC=FFMC Code (Fine Fuel Moisture Code)
5. DMC=DMC Code (Duff Moisture Code)
6. DC=DC Code (Drought Code)
7. ISI=ISI Code (Initial Spread Index)
8. Temp=Outside temperature (in C)
9. RH=Outside relative humidity (in %)
10. Wind=Outside wind speed(in km/h)
11. Rain=Outside rain(in mm/m2 )
12. Area=Total Burned area(in ha)

**Attribute information:**

1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9

2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9

3. month - month of the year: "jan" to "dec"

4. day - day of the week: "mon" to "sun"

5. FFMC - FFMC index from the FWI system: 18.7 to 96.20

6. DMC - DMC index from the FWI system: 1.1 to 291.3

7. DC - DC index from the FWI system: 7.9 to 860.6

8. ISI - ISI index from the FWI system: 0.0 to 56.10

9. temp - temperature in Celsius degrees: 2.2 to 33.30

10. RH - relative humidity in %: 15.0 to 100

11. wind - wind speed in km/h: 0.40 to 9.40

12. rain - outside rain in mm/m2 : 0.0 to 6.4

13. area - the burned area of the forest (in ha): 0.00 to 1090.84

(this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

**4. Linear Regression**

Linear regression most widely used statistical technique.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

linear regression equation ***Y = a + bX***

where ***X*** is the explanatory variable and ***Y*** is the dependent variable.

The slope of the line is ***b***, and ***a*** is the intercept (the value of ***y*** when ***x*** = 0).

**Simple Linear Regression:**

1. Linear regression most widely used statistical technique.
2. Simple linear regression is a linear regression model with single explanatory variable.
3. 1dependent variable,1 independent variable.
4. X variable sometimes called the independent variable and Y variable dependent variable.
5. Simple linear regression plots one independent variable X against one independent variable Y.
6. The regression analysis the independent variable is usually called the predictor variable and dependent variable is called criterion variable**.**

Formula:**Yi=β0+β1Xi+έi**

Yi=dependent variable, β0=intercept, β1=slope coefficient, Xi=independent variable, έi=random error

**Source code:**

a <- read.csv("forestfires.csv")

> View(a)

> b <- lm(area~temp,data=a)

> print(b)

**Out put:**

Call:

lm(formula = area ~ temp, data = a)

Coefficients:

(Intercept) temp

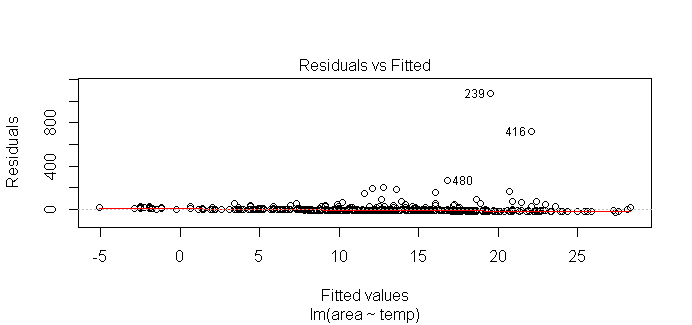
-7.414 1.073

**Source code:**

plot(b)

residual vs fitted

**out put:**

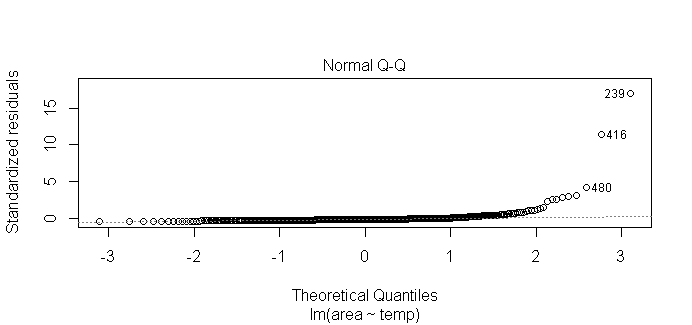


**Source code:**

plot(b)

Normal Q-Q

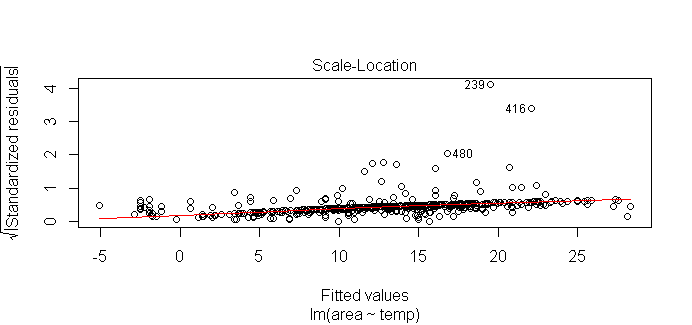
**Out put:**

 **Source code:**

plot(b)

Scale location

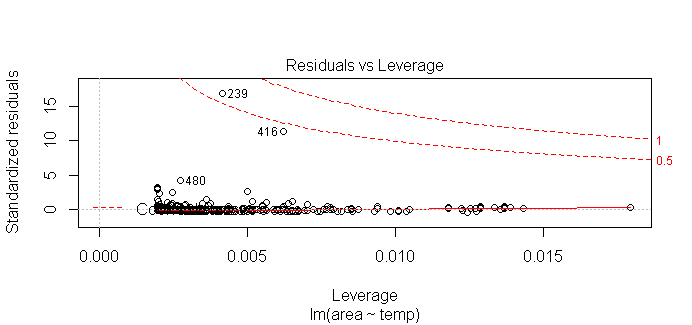
**Out put:**

 **Source code:**

plot(b)

Residual vs Leaverage

**Out put:**



**Source code:**

names(b)

**out put:**

[1] "coefficients" "residuals" "effects" "rank"

[5] "fitted.values" "assign" "qr" "df.residual"

[9] "xlevels" "call" "terms" "model"

**Source code:**

> summary(b)

**out put:**

Call:

lm(formula = area ~ temp, data = a)

Residuals:

Min 1Q Median 3Q Max

-27.34 -14.68 -10.39 -3.42 1071.33

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.4138 9.4996 -0.780 0.4355

temp 1.0726 0.4808 2.231 0.0261 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.41 on 515 degrees of freedom

Multiple R-squared: 0.009573, Adjusted R-squared: 0.00765

F-statistic: 4.978 on 1 and 515 DF, p-value: 0.0261

**Source code:**

> residuals(b)

**Out put:**

1 2 3 4 5

-1.381795e+00 -1.189355e+01 -8.246612e+00 -1.489058e+00 -4.814203e+00

6 7 8 9 10

-1.639858e+01 -1.843657e+01 -1.167269e+00 -6.637670e+00 -1.704216e+01

11 12 13 14 15

-1.167902e+01 -1.328796e+01 -1.082092e+01 -1.543322e+01 -2.090362e+01

16 17 18 19 20

-1.714942e+01 -8.782926e+00 -1.049913e+01 -9.641028e+00 -2.561685e+00

21 22 23 24 25

-1.221533e+01 -1.307344e+01 -1.511143e+01 -1.350249e+01 -1.800752e+01

26 27 28 29 30

-1.007008e+01 -1.296617e+01 -1.339522e+01 -2.497960e+01 -1.704216e+01

31 32 33 34 35

-1.983099e+01 -4.599678e+00 -1.468238e+01 -1.157176e+01 -1.532595e+01

Source code:

**effects(b)**

**out put:**

(Intercept) temp

-2.921170e+02 1.414808e+02 -7.845086e+00 -5.241153e-01 -4.126498e+00

-1.667673e+01 -1.888465e+01 -1.754977e-01 -6.101998e+00 -1.737397e+01

-1.156367e+01 -1.330676e+01 -1.063403e+01 -1.563088e+01 -2.155738e+01

-1.749017e+01 -8.426116e+00 -1.028541e+01 -9.355763e+00 -1.686174e+00

-1.214470e+01 -1.307435e+01 -1.528226e+01 -1.353917e+01 -1.841982e+01

-9.820586e+00 -1.295815e+01 -1.342297e+01 -2.597320e+01 -1.737397e+01

-1.159850e+01 -2.318426e+01 -1.674426e+01 3.796188e+01 -4.354675e+00

-2.062773e+01 -4.591321e+00

attr(,"assign")

[1] 0 1

attr(,"class")

[1] "coef"

**Source code:**

coef(b)

**out put:**

(Intercept) temp

-7.413752 1.072628

**Source code:**

> confint(b)

**Out put:**

2.5 % 97.5 %

(Intercept) -26.0765078 11.249005

temp 0.1281498 2.017105

**source code:**

> predict(b,data.frame(temp=(c(5,10,15))),interval="confidence")

**Out put:**

fit lwr upr

1 -2.050613 -16.266830 12.16560

2 3.312525 -6.712697 13.33775

3 8.675663 2.079358 15.27197

**Source code:**

|  |
| --- |
| predict(b,data.frame(temp=(c(5,10,15))),interval="prediction")  **out put**:  fit lwr upr  1 -2.050613 -127.4369 123.3356  2 3.312525 -121.6679 128.2930  3 8.675663 -116.0766 133.4279  **Accuracy:**  overall average accuracy of 88.12%. |
|  |
| |  | | --- | |  | |

**5. Conclusion:**

The study shows that the Forest Fires Forecasting can be predicted with reasonable accuracy with the help of proposed simple linear regression model which otherwise have unreliability with traditional methods. It also reveals that the simple linear regression with multiple back propagation can be used to model any other time series forecasting.