PREDICTION OF FOREST FIRE AREA

Class: ISM6136.002.F21

Dr. Kiran Garimella

University of South Florida

Final Project Group 7

AMITOJ SINGH LOTEY
PRUDHVI RAJ MADISETTY
CHARAN KUMAR SULAGIRI

Background of the problem

Forest fires (also known as wildfires) are a serious environmental hazard that result in forest degradation, economic and ecological devastation, as well as human suffering. Forest fires can be caused by a variety of factors, including human activity such as open burning, incorrectly discarded cigarettes, power line failure, and camping fires on windy or dry days and lightning. Although high winds are not a direct cause of fires, they do contribute to their spread.

Consequences of Wildfires are devastating with their increased frequency. Wildfires destroy animal and plant habitats. It has an impact on the diverse flora and fauna, resulting in ecosystem and biodiversity loss. It can cause soils to lose their nutrients, death of animals, and burning of trees and plants. If the fire is not put out quickly, it may result in the extinction of certain species in the area. People's health is affected as dust and smoke could cause respiratory problems. Wildfires contribute to increased carbon dioxide levels in the atmosphere. As a result, the greenhouse effect becomes stronger, and accelerates climate change. Fires have a negative impact on the economy of the country. To put out a single fire, authorities must invest in logistical support, trucks, phosphate fertilizer, and water-dropping planes.

Motivation for solving the problem

There is a saying 'Prevention is better than cure'. Because we have limited manpower to tackle the forest fires, distribution of the effort to control it becomes of great importance in such scenarios. We can use prediction models and data mining to limit the spread of forest fires before they can become a disaster. We can save the remaining forest area without getting affected from the forest fires. This approach would help save firefighters' time and energy, while also protecting the atmosphere and saving the environment.

Description of the dataset

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 (The Fine Fuel Moisture Code) - FFMC index from the FWI system: 18.7 to 96.20

6. DMC (Duff Moisture Code)- DMC index from the FWI system(Fire Weather Index (**FWI**)): 1.1 to 291.3

7. DC (Drought Code) - DC index from the FWI system: 7.9 to 860.6

8. ISI (The Initial Spread Index) - 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 output variable.

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0
8	6	sep	tue	91	129.5	692.6	7	13.1	63	5.4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	22.8	40	4	0	0
7	5	sep	sat	92.5	88	698.6	7.1	17.8	51	7.2	0	0

The Forest Fire Weather Index (FWI) is a Canadian system for measuring fire hazard, and it consists of the elements listed below. Drought Code (DC), Initial Spread Index, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Fine Fuel Moisture Code (FFMC). FFMC represents moisture content of surface litter that promotes ignition and fire propagation, whereas DMC and DC denote the moisture content of shallow and deep organic layers which increases the fire intensity. The Initial Spread Index (ISI)is a score that refers to the spread of a fire. The higher the value, the more severe the burning conditions. Relative humidity, wind, temperature, and rain are some of the additional climatic parameters that might be considered. The higher the temperature, wind, and relative humidity, the more likely a fire will spread. The greater the possibility of rain, the lower the chance of fire. All of the current elements are independent variables (IV), and the dependent variable (DV) here is the area of the forest that will burn. We may use data mining prediction models like Linear regression to predict forest fires by taking all of these aspects into consideration.

Solution methodology and evaluation metrics

Algorithms used: - Linear Regression, Decision Tree Regression

Column	Variable Type	Cause
X-axis	Categorical	Explanatory
Y-axis	Categorical	Explanatory
Month	Categorical	Explanatory
Day	Categorical	Explanatory
FFMC	Continuous	Explanatory
DMC	Continuous	Explanatory
DC	Continuous	Explanatory
ISI	Continuous	Explanatory
Temp	Continuous	Explanatory
RH	Continuous	Explanatory
Wind	Continuous	Explanatory
Rain	Continuous	Explanatory
Area	Discrete	Response

In predicting the area forest fire will cover, not all the metrics explained in the above dataset come into play. And also we will convert the independent variables month and day to categorical variables by using as.factor function. We have 517 records in our dataset which will be used to train our model and understand the accuracy of it.

Comparison between two algorithms

Data Preprocessing: - We can remove the variables X and Y since all the other meteorological variables already include this information. So, we don't need it. The dependent variable for our algorithm is **area**. The area variable in the dataset is highly right skewed, so to remove the skewness, we applied the log (area + 1) transformation

```
library(dplyr)

library(rio)

library(party)
library(rpart)

data <- read.csv("C:\\Users\\amito\\Downloads\\forestfires.csv",header = T)

data$area=log(data$area + 1)
data$month=as.factor(data$month)
data$day=as.factor(data$day)
```

Implementing Linear Regression Algorithm: -

```
set.seed(1234)

dt<-sample (2, nrow(data), replace = TRUE, prob=c (0.8,0.2))

train<-data[dt==1,]
validate<-data[dt==2,]

model = lm(area~.-month-day,data = train)
summary(model)
```

- Residual standard error: 1.382 on 489 degrees of freedom
- Multiple R-squared: 0.07426, Adjusted R-squared: 0.02315
- F-statistic: 1.453 on 27 and 489 DF, p-value: 0.06765

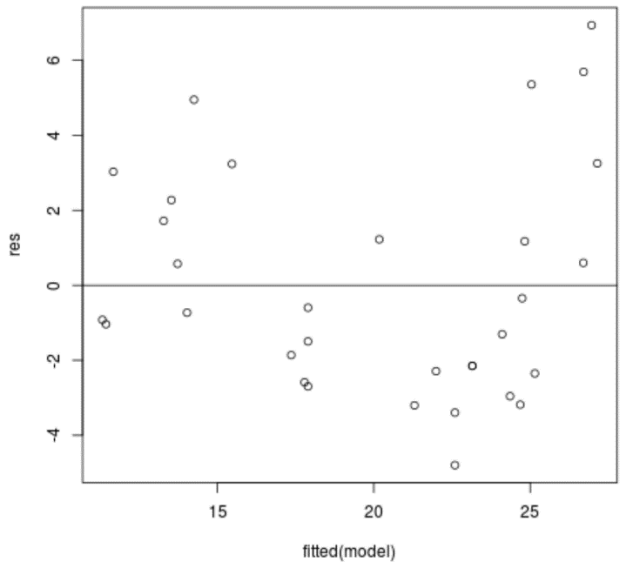
```
predict(model,validate)
```

Mean Absolute Error	1.23912
Root Mean Squared Error	1.619865
Relative Absolute Error	1.023707
Relative Squared Error	1.199358
Coefficient of Determination	-0.199358

As we can see linear regression did not work well with our model, we achieved an RMSE of 1.62 and Adjusted R squared of 0.07.

Upon plotting the residuals, we get to know that the residuals follow a sine curve

```
res <- resid(model)
plot(fitted(model), res)
```



For such datasets, decision tree regression works a lot better as it better fits the sin curve

Implementing Decision Tree Regression: -

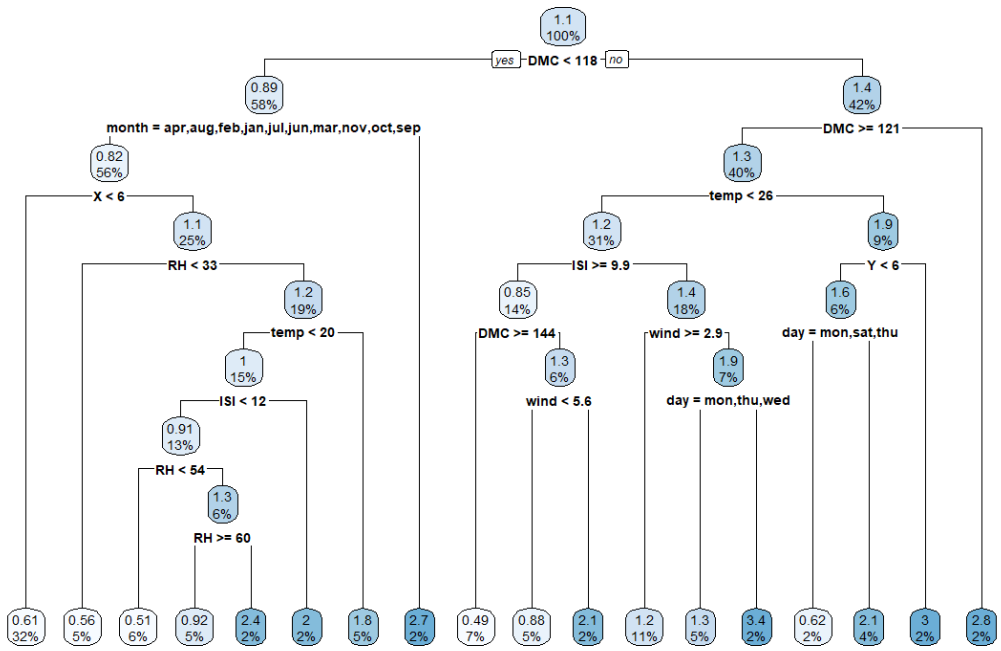
```
set.seed(1234)

dt1<-sample (2, nrow(data), replace = TRUE, prob=c (0.8,0.2))

train1<-data[dt1==1,]
validate1<-data[dt1==2,]

model <- train(
area ~ .,
data = train1,
method = 'rpart2'
)
model
```

CART
419 samples
12 predictor
Pre-processing: centered (27), scaled (27)
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 419, 419, 419, 419, 419, 419, ...
Resampling results across tuning parameters:

maxdepth RMSE Rsquared MAE
1 0.878234 0.528826124 1.181407
2 0.938456 0.539814424 1.179835
3 0.845693 0.519526733 1.187688

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was maxdepth = 3.



Model Summary:-

Algorithm	RMSE	Adjusted R squared
Linear Regression	1.62	0.07
Decision Tree Regression	0.85	0.42

With Decision Tree Regression, we achieved an RMSE of 0.85 which is much better than the previous model which had RMSE of 1.62. Also, Adjusted R squared value increased from 0.07 to 0.42. So, we can use this model with confidence to predict the area a forest fire will cover.

Conclusion and Further Recommendations: -

The problem that we are tackling is a very serious issue, and can help save millions of dollars in tackling and managing forest fires as efforts can be concentrated on reducing the spread of fire that might engulf a large amount of area. After exploring the data and modeling, here are our findings and recommendations that will help in predicting the forest fire area :-

- The data that we had was not balanced and also was heavily right skewed for the response variable which is area. So a better data set could have yielded much better results from the ones that we got.
- The data was limited to a certain geological region (Portugal in our case). A more diverse dataset could have added more value to our findings as it would have represented a diversity of geological conditions.
- The RMSE and adjusted R squared value is good enough to recommend the use of the model as the model usually predicted the larger forest fires correctly which was the problem we went out to solve. Predicting smaller forest fires as larger forest fires is still an issue and can be removed by further tweaking of parameters for Decision Tree Regression.